Nonparametric Covariance Estimation for Longitudinal Data via Tensor Product Splines

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

List of Tables

List of Figures

Figure **Page**

Chapter 1: Introduction

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 with high dimensional data has has recently gained growing interest; it is generally recognized that there are two primary hurdles responsible for the difficulty in covariance estimation: the instability of sample covariance matrices in high dimensions and a positive-definite constraint we wish estimates to obey.

Prevalent technological advances in industry and many areas of science make high dimensional longitudinal and functional data a common occurrence, arising in numerous areas including medicine, public health, biology, and environmental science with specific applications including fMRI, spectroscopic imaging, gene microarrays among many others, presenting a need for effective covariance estimation in the challenging situation where parameter dimensionality p is possibly much larger than the number of observations, N.

We consider two types of potentially high dimensional data: the first is the case of functional data or times series data, where each observation corresponds to a curve sampled densely at a fine grid of time points; in this case, it is typical that the number of time points is larger than the number of observations. The second is the case of sparse longitudinal data where measurement times may be almost unique yet sparsely distributed within the observed time range for each individual in the study. In this case, the nature of the high dimensionality may not be a consequence of having more measurements per subject than the number of subjects themselves, but rather because when pooled across subjects, the total number of unique observed time points is greater than the number of individuals. Several approaches have been taken in effort to overcome the issue of high dimensionality in covariance estimation. Regularization improves stability of covariance estimates in high dimensions, particularly in the case where the parameter dimensionality p is much larger than the number of observations N. Regularization of the covariance matrix and its Cholesky decomposition has been explored extensively through various approaches including banding, tapering, kernel smoothing, penalized likelihood, and penalized regression; see Pourahmadi [2011] for a comprehensive overview.

To overcome the hurdle of enforcing covariance estimates to be positive definite, several have considered modeling various matrix decompositions including variance-correlation decomposition, spectral decomposition, and Cholesky decomposition. The Cholesky decomposition has received particular attention, as it which allows for a statistically meaningful interpretation as well as an unconstrained parameterization of elements of the covariance matrix. This parameterization allows for estimation to be accomplished as simply as in least squares regression. If we assume that the data follow an autoregressive process with (possibly) heteroskedastic errors, then the two matrices comprising the Cholesky decomposition, the Cholesky factor (which diagonalizes the

covariance matrix) and diagonal matrix itself, hold the autoregressive coefficients and the error variances, respectively.

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. We view the response as a stochastic process with corresponding continuous covariance function and the generalized autoregressive parameters as the evaluation of a continuous bivariate function at the pairs of observed time points rather than specifying a finite set of observations to be multivariate normal and estimating the covariance matrix. This is advantageous because it is unlikely that we are only interested in the covariance between pairs of observed design points, so it is reasonable to approach covariance estimation in a way that allows us to obtain an estimate of the covariance between two measurements at any pair of time points within the time interval of interest.

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. **?**, 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.

The remainder of the chapter 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 for reparameterizing elements of the covariance matrix, translating covariance estimation into a generalized linear modeling problem.

Estimation of the covariance matrix is fundamental to the analysis of multivariate data, and the most commonly used estimator is the sample covariance matrix, S. While it is both positive-definite and an unbiased estimator of Σ, it is unstable large dimension M. 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, $N >> M$ (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]

The sample covariance matrix S, which is used in virtually all multivariate techniques, is both unbiased and positive-definite. The flexible estimator is also computationally convenient, however it is neither parsimonious nor, in high dimensions, a stable estimator. Given a sample of size N $Y_1, \ldots, Y_N$, from an M-dimensional Normal distribution with mean μ and covariance matrix Σ, the sample covariance matrix

$$S = (N-1)^{-1} \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right) \left(Y_i - \bar{Y}\right)' \tag{1.1}$$

is a straightforward estimator of the $\frac{M(M+1)}{2}$ parameters of the unstructured covariance matrix Σ. The number of parameters of $\Sigma = (\sigma_{ij})$ grows quadratically in the dimension M, and the parameters must satisfy the positive-definiteness constraint

$$v'\Sigma v = \sum_{i,j=1}^{M} v_i v_j \sigma_{ij} \geq 0 \tag{1.2}$$

for all $v \in \Re^M$. The challenge presented by these hurdles have motivated a growing body of research in statistics and its areas of application aimed at effectively estimating covariance matrices.

Our review of work in this area focuses on developments made from two connected perspectives: regularization or sparsity in covariance matrices for high-dimensional data, and generalized linear models (GLM) or parsimony and use of covariates in low dimensions. A recurring technique in both perspectives is the reduction of covariance estimation to estimating a single of sequence of regression. The generalized linear model (GLM) framework McCullagh and Nelder [1989] merges numerous seemingly disconnected approaches to model the mean of a distribution, and can

accommodate many types of including normal, probit, logistic and Poisson regressions, survival data, and log-linear models for contingency tables. The key to the power of the GLM paradigm is the use of a link function to induce unconstrained reparameterization for the mean of a distribution, and hence the ability to reduce the dimension of the parameter space via modeling the covariate effect additively by increasing the number of parameters gradually one at a time corresponding to inclusion of each covariate. The extension of the GLM has lead to large class 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]. An analogous framework for modeling covariance matrices facilitates further developments in covariance estimation from the Bayesian, nonparametric and other paradigms.

1.1 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. Typical choices are simple models which depend on a small number of parameters such as compound symmetry and autoregressive models of order k, where k is small.We will review a selection of modeled frequently encountered in the applied statistics literature in sections to follow. This approach is attractive because it is computationally inexpensive, and software packages implementing fitting procedures for a growing number of simple models are readily accessible. The compound symmetric model was at one time a very popular choice for parametric covariance structure, specifying

$$\sigma_{_{ij}} = \begin{cases} \rho, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \tag{1.3}$$

where σ_{ij} denotes the (i, j) element of Σ. With only two parameters to be estimated, this model is highly parsimonious, but has received less attention with the development of models that allow for heterogeneous variances and non-constant correlation.

The first order autoregressive model for response variable y_t associated with measurement time t specifies

$$y_t = \begin{cases} \mu_t + \epsilon_t, & t = 1, \\ \mu_t + \rho\left(y_{t-1} - \mu_{t-1}\right) + \epsilon_t, & t > 1, \end{cases} \tag{1.4}$$

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 $t = 2, \ldots, M$. The corresponding dependence components of the covariance structure are monotonically decreasing in $l = |i - j|$; specifically,

$$\sigma_{ij} = \begin{cases} \rho^{|i-j|}, & i \neq j, \\ \sigma^2, & i = j, \end{cases} \tag{1.5}$$

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 $|i - j|$.

The response specification for q^{th} order moving average model is given by

$$y_t = \sum_{j=0}^{q} \theta_j \epsilon_{t-j}, \tag{1.6}$$

where the $\{\epsilon_t\}$ are independently and identically distributed mean zero Normal random variables with variance σ^2. This model corresponds to covariance structures with elements given by

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

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

$$\sigma_{ij} = \begin{cases} \rho_{i-j} & |i-j| \leq q, \\ 0 & |i-j| > q, \\ \sigma^2 & i = j, \end{cases} \tag{1.7}$$

or covariance matrix of the form

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \ldots & m_{p-1} \\ m_1 & m_0 & m_1 & \ldots & m_{p-2} \\ m_2 & m_1 & m_0 & \ldots & m_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p-1} & m_{p-2} & m_{p-3} & \ldots & m_0 \end{bmatrix}, \tag{1.8}$$

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:

$$y_t = \mu_t + \sum_{j=1}^{t} \epsilon_j, \quad t = 1, \ldots, M, \tag{1.9}$$

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:

$$\sigma_{ij} = \begin{cases} \sqrt{i/j} & i \neq j \\ j\sigma_\epsilon^2 & i = j, \end{cases} \tag{1.10}$$

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

$$y_i = X_i\beta + Z_i\gamma_i + \epsilon_i, \quad i = 1, \ldots, M, \tag{1.11}$$

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 1.5. 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

$$\sigma_{ij} = \begin{cases} \frac{\sigma_{00}+\sigma_{01}(i+j)+\sigma_{11}ij}{\sqrt{\sigma^2+\sigma_{00}+2i\sigma_{01}+\sigma_{11}i^2}\sqrt{\sigma^2+\sigma_{00}+2j\sigma_{01}+j^2\sigma_{11}}} & i \neq j \\ \sigma^2 + \sigma_{00} + 2\sigma_{01}j + \sigma_{11}j^2 & i = j, \end{cases} \tag{1.12}$$

These models are can permit variance and covariances which exhibit several kinds 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 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 example, see Jennrich and Schluchter [1986] for additional models for repeated measures data. While these models are computationally attractive and 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. To strike a balance between the variability of the sample covariance matrix and the bias of the estimated structured covariance matrix, it is prudent to rely on the data to formulate structures for the unknown underlying dependence in the data.

1.2 Shrinking the sample covariance matrix

1.2.1 Shrinking the spectrum and the correlation matrix

Stein [1975] observed that the sample covariance matrix systematically distorts the eigenstructure of Σ, especially when M 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

$$\hat{\Sigma} = \Sigma(S) = P\Phi(\lambda)P', \tag{1.13}$$

where $\lambda = (\lambda_1, \ldots, \lambda_M)'$, $\lambda_1 > \cdots > \lambda_M$ 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 $\Phi(\lambda) = diag(\phi_1, \ldots, \phi_M)$ 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_M 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

$$\alpha_j(\lambda) = N - M + 2\lambda_j \sum_{i \neq j} \frac{1}{\lambda_j - \lambda_i}. \tag{1.14}$$

The Stein estimators ϕ_j differ from the sample eigenvalues when are nearly equal and N/M 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 $\phi_1 < \cdots < \phi_M$.

1.2.2 Ledoit-Wolf shrinkage estimator

The estimator proposed by Ledoit and Wolf [2004] is motivated by the fact that the sample covariance matrix is unbiased but has high variance - the risk associated with S is considerable when $M >> N$, 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 1.1, 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

$$\hat{\Sigma} = \alpha_1 I + \alpha_2 S, \tag{1.15}$$

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

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

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

$$\begin{aligned} \mu &= \text{tr}\left(\Sigma\right)/M, \\ \alpha^2 &= ||\Sigma - \mu I||^2, \\ \beta^2 &= ||S - \Sigma||^2, \\ \delta^2 &= ||S - \mu I||^2. \end{aligned} \tag{1.16}$$

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]).

1.2.3 Elementwise shrinkage

A broad class of estimators that aim to stabilize the sample covariance matrix do so by applying shrinkage elementwise to the same covariance matrix. Shrinking the elements of the sample covariance matrix has been approached in a multitude of ways, including banding, tapering, and thresholding. These estimators are computationally inexpensive, with the exception of cross validation necessary for smoothing parameter selection. The tradeoff accompanying the ease of

computation is that, because transformations of sample estimates are elementwise, the resulting estimators are not guaranteed to be positive definite.

Tapering and banding the sample covariance matrix

The sample covariance matrix is unstable when the dimension of the data M is larger than the sample size N, and even when the sample size is larger than the dimension of the data many entries of the sample covariance matrix $S = (s_{ij})$ could be small. Setting certain entries to zero is one approach to reducing parameter dimension to stabilize estimates. In time series analysis, one observes a sample size of $N = 1$: the data is a single, long realization. Assuming stationarity of the process reduces the number of distinct parameters of the $M \times M$ covariance matrix Σ from $M(M+1)/2$ to M, which could be large yet. 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, $0 < k < M$, the k-banded sample covariance matrix is given by

$$B_k(S) = \left[s_{ij} 1(|i-j| \leq k)\right] \tag{1.17}$$

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, $t = 1, \ldots, M$ 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

$$\hat{\Sigma} = R * S \tag{1.18}$$

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

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

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 1.3.4.

When N, M, and k are large, asymptotic analysis of banding estimators is available. 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 $\log M/N \to 0$, 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:

$$S^{\omega} = \left[\omega_{ij}^{k} s_{ij}\right], \tag{1.19}$$

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 M, 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].

Thresholding the sample covariance matrix

When both N and M 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. A generalized thresholding estimator $\int_\lambda (z)$ is the solution to

$$\int_\lambda (z) = \arg\min_\sigma \left[\frac{1}{2} (\sigma - z)^2 + J_\lambda (\sigma) \right] \tag{1.20}$$

For detailed discussion of the connection between penalty functions and the resulting thresholding rules, see Antoniadis and Fan [2001]. Soft thresholding results from minimizing 1.20 using the lasso penalty, $J_\lambda = \lambda|\sigma|$, which corresponds to thresholding rule

$$\int_\lambda (\sigma) = \text{sign}(\sigma) (\sigma - \lambda)_+ . \tag{1.21}$$

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

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

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. 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 2. 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 1.20,

$$||\hat{\Sigma}^{\lambda} - \Sigma||^2 = \left(\sum_{i,j} \left(\hat{\sigma}_{ij}^{\lambda} - \sigma_{ij} \right)^2 \right)^{1/2} \tag{1.22}$$

If Σ were available, one would choose the value of the tuning parameter λ which minimizes **??**. 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

$$\text{CV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(-k)} - \tilde{\Sigma}^{(k)}||_F^2, \tag{1.23}$$

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:

$$\text{rCV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(k)} - \tilde{\Sigma}^{(-k)}||_F^2, \tag{1.24}$$

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 $\mathcal{D}_k$.

1.3 Matrix decompositions

The most methodic and successful approaches to covariance modeling is to decompose the covariance matrix into its variance and dependence components. The following section demonstrates the role of multiple matrix parameterizations in removing the positive definite constraint that poses a challenge in most covariance estimation settings.

1.3.1 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

$$\Sigma = DRD, \tag{1.25}$$

where $D = \text{diag}\left(\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{MM}}\right)$ denotes the diagonal matrix with diagonal entries equal to tje 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 eithering iteratively fixing one sequence of parameters to estimate the other. Moreover, one set of parameters may be more important than the others in some applications; the dynamic correlation model presented in Engle's (2002) is actually motivated by the fact that variances (volatilities) of individual assets are more important than their time-varying correlations.

While the diagonal entries of D are constrained to be nonnegative, their logarithms are unconstrained. However, the correlation matrix R is positive-definite constrained to have unit diagonal entires and off-diagonal entries to be less than or equal to 1 in absolute value. Because of these constraints, the variance-correlation decomposition does not lend to modeling its components with the use of covariates.

1.3.2 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 $\Sigma^{-1} =$

(σ^{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]).

Graphical models are a common way of representing the conditional independence structure in 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

$$\Sigma^{-1} = \left(\sigma^{ij}\right) = \tilde{D}\tilde{R}\tilde{D} \tag{1.26}$$

can be interpretted as certain coefficients of a regression model. 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 M 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 $M - 2$ remaining variables is given by

$$\tilde{\rho}_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}. \tag{1.27}$$

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

$$\tilde{d}_{ii}^2 = \frac{1}{\sigma^{ii}}. \tag{1.28}$$

To connect these parameters to those of a regression model, consider partitioning random vector $Y = (y_1, \ldots, y_M)'$ into two components $(Y_1', Y_2')'$ of dimensions M_1 and M_2, and similarly partitioning its covariance and precision matrices:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{1.29}$$

Let $\Phi_{2|1}$ denote the $M_2 \times M_1$ 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:

$$\begin{aligned} \Phi_{2|1} &= \Sigma_{21}\Sigma_{11}^{-1} \\ &= -\left(\Sigma^{22}\right)^{-1}\Sigma^{21} \end{aligned} \tag{1.30}$$

$$\begin{aligned} Cov\left(e_{2|1}\right) &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \\ &= \Sigma_{22|1} = \left(\Sigma^{22}\right)^{-1}. \end{aligned} \tag{1.31}$$

If we let $M_2 = 1$, then one can establish the relationship between elements of the inverse covariance matrix and these regression coefficients and conditional covariances. When $Y_1 = Y_{-(i)} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_M)'$ 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 $Var\left(\epsilon_i^*\right) = {d^*}_i^2$. Then

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

where (1.31) and (1.32) give

$$\begin{aligned}\beta_{ij} &= -\frac{\sigma^{ij}}{\sigma^{ii}}, \quad j \neq i \\ d^{*2}_{i} &= Var\left(y_i|y_j\right) = \frac{1}{\sigma_{ii}}, \quad j \neq i, \;\; i = 1, \ldots, M\end{aligned} \tag{1.32}$$

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 $M_2 = 2$. Letting $Y_2 = (y_i, y_j)'$, $i \neq j$ and $Y_1 = Y_{-(ij)}$ contain the remaining $M - 2$ variables, the covariance of (y_i, y_j) after removing the linear effects of $\{y_k : k \neq i, j\}$ is given by

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

The regression coefficients (1.32) can be written in terms of the partial correlation between y_i and y_j:

$$\rho^{*}_{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}}\sigma^{ij}}. \tag{1.33}$$

Rewriting the β_{ij}, we have

$$\beta_{ij} = \rho^{*}_{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}, \tag{1.34}$$

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.

1.3.3 The spectral decomposition

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

$$\Sigma = P\Lambda P' = \sum_{i=1}^{M} \lambda_i e_i e_i', \tag{1.35}$$

where Λ is a diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_M$, 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 thevariances and coefficients of the M principal components. The matrix P is constrained by its orthogonality, its use within the framework of GLM or alongside covariates in an effort 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:

$$\log \Sigma = P \log \Lambda P' = \sum_{i=1}^{M} \log(\lambda_i)\, e_i e_i', \tag{1.36}$$

This decomposition is particularly interesting because it highlights a tradeoff between the requirements for unconstrained parameterization of covariance matrices and the statistical interpretability of the corresponding parameters. The components of the matrix logarithm, $\log \lambda_i$, are free, but lack any relevant statistical interpretability. We further discuss the log-linear GLM for covariance matrices in Section 1.4.2 .

1.3.4 The Cholesky decomposition

The Cholesky decomposition of a positive-definite matrix has the form

$$\Sigma = CC', \tag{1.37}$$

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]. But by transforming C to unit lower-triangular matrices, statistically interpreting of the diagonal entries of C and the resulting unit lower-triangular matrix is much easier. To do this, one must simply divide the i^{th} column of C by its i^{th} diagonal element c_{ii}. Letting $D^{1/2} = diag\,(c_{11}, \ldots, c_{MM})$, the standard Cholesky decomposition 1.37 can be written

$$\Sigma = CD^{-1/2}D^{1/2}D^{1/2}D^{-1/2}C' = LDL', \tag{1.38}$$

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:

$$D = T\Sigma T', \quad \Sigma^{-1} = T'D^{-1}T, \tag{1.39}$$

where $T = L^{-1}$. Like P as in the spectral decomposition, the lower triangular matrix T diagonalizes Σ. However, the Cholesky decomposition is perhaps more attractive since unlike the entries of the orthogonal matrix of the spectral decomposition, the entries of T are unconstrained, and furthermore, have a specific statistical interpretation.

Like the variance-correlation decomposition of the inverse covariance matrix 1.26, the Cholesky factor T and diagonal matrix D can be constructed using components of a regression model. Consider regressing y_t on its predecessors $y_1, \ldots, y_{t-1}$. Let $Y = (y_1, \ldots, y_M)'$ denote a mean zero

random vector with positive definite covariance matrix Σ, and let $\hat{y}_t$ be the linear least-squares predictor of y_t based on previous measurements $y_{t-1}, \ldots, y_1$. Let ϵ_t denote the corresponding prediction residual having variance $\sigma_t^2 = Var(\epsilon_t)$. Standard regression machinery gives us that there exist unique scalars ϕ_{tj} so that

$$y_t = \sum_{j=1}^{t-1} \phi_{t,j} y_j + \sigma_t \epsilon_t, \quad t = 2, \ldots, M \tag{1.40}$$

where

$$\epsilon_t = \begin{cases} y_t - \hat{y}_t, & t > 1 \\ y_t, & t = 1 \end{cases}$$

are i.i.d. mean zero random variables with unit variance. The connection between the Cholesky decomposition and the autoregressive model (1.3.4) 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 $\epsilon = (y_1, \ldots, y_M)'$ denote the vector of uncorrelated prediction residuals with

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

Then model (1.3.4) can be written in vector form $\epsilon = TY$, where the (t, j) entry of T is $-\phi_{tj}$, and the (t, t) entry of D is the t^{th} prediction variance $\sigma_t^2 = var(\epsilon_t)$.

$$\begin{bmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \vdots & & & \ddots & \\ -\phi_{m1} & -\phi_{m2} & \cdots & -\phi_{m,m-1} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \ddots \\ y_m \end{bmatrix} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \ddots \\ \epsilon_m \end{bmatrix} \tag{1.41}$$

Table 1.1 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 $M-1$ 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 (1.39) 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 1.1: *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_{m1}	ϕ_{m2}	$\cdots$	$\cdots$	$\phi_{m,m-1}$	1
σ_1^2	σ_1^2	$\cdots$	$\cdots$	σ_{m-1}^2	σ_m^2

1.4 Generalized linear models for covariances

Modeling covariance matrices in a systematic, data-driven manner is impeded by the positive-definiteness constraint and high-dimensionality; however, similar (albeit simpler) hurdles in modeling the mean vector μ of the distribution of a random vector $Y = (y_1, \ldots, y_M)'$ has been successfully handled in the context of regression analysis. The resulting techniques have lead to the

framework of generalized linear models (GLM), which enjoys a rich and extensive theoretical foundation. The success of GLMs is in most part due to the use of a link function $g(\cdot)$ and a linear predictor $g(\cdot) = X\beta$, which induces an unconstrained parameterization and reduces the parameter space dimension simultaneously. Since the covariance matrix of a random vector Y , defined by $\Sigma = E(Y - \mu)(Y - \mu)$, is a mean-like parameter, one would like to exploit the idea of GLM along with the experience and progress in fitting the mixed-effects and time series models in developing a systematic, data-based procedure for covariance matrices.

Approaches to modeling covariances with the explicit use covariates has been extensively explored in the time series literature, while the implicit use of covariates for covariance modeling has been the focus of many in the area of variance components; see Klein [1997] and Searle et al. [2009]. Time series techniques based on spectral and Cholesky decompositions provide the necessary tools for handling the cumbersome positivedefiniteness constraint on a stationary covariance matrix or covariance function. In the GLM setting, simply applying a link function componentwise to the potentially constrained mean vector μ permits its unconstrained estimation. Unfortunately employing the same precise approach to covariance matrices isn't viable since positive-definiteness is a simultaneous constraint on all entries of a matrix. Successfully modeling a general covariance structure almost necessitates decomposing a covariance matrix into its "variance" and "dependence" components because of its inherent complicated structure. The three major methods for performing such decompositions include the variance-correlation decomposition, the spectral decomposition, and the Cholesky decomposition. Section 1.3.4 touched on the attractive properties of the latter that lead to advantages over the other two covariance parameterizations.

1.4.1 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:

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

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

$$\Sigma = \sum_{i=1}^{q} \alpha_q U_q \tag{1.42}$$

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 (**??**), since one can decompose every covariance matrix as

$$\Sigma = \sum_{i=1}^{M}\sum_{j=1}^{M} \sigma_{ij} U_{ij}, \tag{1.43}$$

where U_{ij} is an $M \times M$ matrix with a 1 in the (i, j) position, and zeros everywhere else. The linear model (1.42) can be viewed as modeling the link-transformed covariance $g(\Sigma) = \sum_{i=1}^{q} \alpha_q U_q$, where $g(\cdot)$ is the identity link. Despite the convenience of parameterization, the positive definite constraint (1.2) 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 (1.42) and the covariance parameterization given in (1.43) 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

$$\sigma_{ij} = 0 \quad \text{for } |i-j| > q, \tag{1.44}$$

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

$$\sigma^{ij} = 0 \quad \text{for } |i-j| > p. \tag{1.45}$$

Of course, there are the nonstationary analogues to these classes of models, some of which were discussed in Section **??**. We will review others which are related to antedependence models and Gaussian graphical models. Random variables $y_1, \ldots, y_M$, which correspond to observation times $t_1, \ldots, t_M$, 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 (1.45). 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.

1.4.2 Log-linear covariance models

The constraint on the α_is in (1.42) 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

$$\Sigma = P\Lambda P', \tag{1.46}$$

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:

$$g(\Sigma) = \log \Sigma = \sum_{i=1}^{q} \alpha_i U_i, \tag{1.47}$$

where the U_is are as before in 1.42 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, $\log \Sigma = \text{diag}(\sigma_{11}, \ldots, \sigma_M M)$, and model 1.47 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.

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 1.48 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 1.39 can be written as the entries of

$$g(\Sigma) = 2I - T - T' + \log D. \tag{1.48}$$

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 1.42 and those of $\log D$ as in, say, model 1.47, letting

$$\begin{aligned} \phi_{tj} &= x_{tj}'\beta, \\ \log \sigma_t^2 &= z_t'\gamma, \end{aligned} \tag{1.49}$$

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

$$\begin{aligned} z'_{jk} &= \left(1, t_j - t_k, (t_j - t_k)^2, \ldots, (t_j - t_k)^{p-1}\right)' \\ z'_i &= \left(1, t, \ldots, t^{q-1}\right)' \end{aligned} \tag{1.50}$$

Pourahmadi [1999], Pourahmadi [2000], and Pan and [2006] prescribe methods for identifying models such as model 1.49 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, and optimization is achieved by solving these 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 $M \times M$ Toeplitz matrix Σ is a matrix with elements σ_{ij} such that $\sigma_{ij} = \sigma_{|i-j|}$ i.e. a matrix of the form (1.8), having entries which are constant on each subdiagonal.

The estimated covariance matrix may be considerably biased when the specified parametric model is far from the truth. To avoid model misspecification, many have alternatively proposed nonparametric and semiparametric techniques approaches to estimation. When the data $Y_1, \ldots, Y_N$

are a random sample of M-dimensional vectors from a mean zero multivariate normal population with common covariance matrix Σ parameterized as $D = T'\Sigma T$, 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

$$\begin{aligned} -2\ell\left(Y_1, \ldots, Y_N, \Sigma\right) &= \sum_{i=1}^{N} \left(\log|\Sigma| + Y_i'\Sigma^{-1}Y_i'\right) \\ &= N\log|D| + N\text{tr}\Sigma^{-1}S \\ &= N\log|D| + N\text{tr}D^{-1}TST', \end{aligned} \tag{1.51}$$

where $S = N^{-1}\sum_{i=1}^{N} Y_i Y_i'$. The negative log likelihood (1.51) 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 (1.3.4). A similar procedure by Dahlhaus et al. [1997] uses varying coefficient regression models for each subdiagonal of T:

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

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 (1.3.4), 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 1.2 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 2, 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 1.2: *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}

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 2 presents a reproducing kernel Hilbert space framework for the estimation of both ϕ and σ^2. In Chapter 3, we discuss an alternative representation the varying coefficient function using the penalized B-splines of Eilers and Marx [1996]. We properties of the P-splines that establish their connection to the usual spline penalty on the second derivative and demonstrate how their simple construction allows for extremely flexible regularization.

Chapter 2: Modeling the Cholesky decomposition via smoothing spline ANOVA models

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

$$y(t_i) = \sum_{j=1}^{i-1} \tilde{\phi}(t_i, t_j)\, y(t_j) + \sigma(t_j)\, \epsilon(t_j) \quad i = 1, \ldots, M, \tag{2.1}$$

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 **??**. To facilitate the penalty specification that achieves this, we consider modeling the varying coefficient function which takes inputs

$$\begin{aligned} l &= t - s \\ m &= \frac{t+s}{2}, \end{aligned} \tag{2.2}$$

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 $Cov\left(y\left(t_i\right), y\left(t_j\right)\right) = G\left(t_i, t_j\right) = G\left(||t_i - t_j||\right)$ can then be written as

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

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

$$\phi(l, m) = \phi\left(s - t, \frac{1}{2}(s+t)\right) = \tilde{\phi}(t, s). \tag{2.3}$$

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 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

$$-2\ell(Y, \Sigma) = \log|\Sigma| + Y'\Sigma Y \tag{2.4}$$

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

$$\begin{aligned} -2\ell(Y, \Sigma) &= \log|D| + Y'T'D^{-1}TY \\ &= \sum_{i=1}^{m} \log\sigma_i^2 + \sum_{i=1}^{m} \frac{\epsilon_i^2}{\sigma_i^2}, \end{aligned} \tag{2.5}$$

where

$$\epsilon_i = \begin{cases} y(t_1), & i = 1, \\ y(t_i) - \sum_{j=1}^{i-1} \phi(\boldsymbol{v}_{ij}) y_j, & i = 2, \ldots, M, \end{cases} \tag{2.6}$$

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

$$\begin{aligned} y(t_{ij}) &= \sum_{k=1}^{j-1} \tilde{\phi}(t_{ij}, t_{ik}) y(t_{ik}) + \sigma(t_{ij}) \epsilon(t_{ij}) \\ &= \sum_{k=1}^{j-1} \phi(\boldsymbol{v}_{ijk}) y(t_{ik}) + \sigma(t_{ij}) \epsilon(t_{ij}) \end{aligned} \tag{2.7}$$

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

$$-2\ell\left(Y_1, \ldots, Y_N, \phi, \sigma^2\right) = \sum_{i=1}^{N} \sum_{j=1}^{m_i} \log \sigma_{ij}^2 + \sum_{i=1}^{N} \sum_{j=1}^{m_i} \frac{\epsilon_{ij}^2}{\sigma_{ij}^2}, \tag{2.8}$$

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 2.8. Employing regularization, we take ϕ, σ^2 to minimize

$$-2\ell\left(Y_1, \ldots, Y_N, \phi, \sigma^2\right) + \lambda J(\phi) + \breve{\lambda} \breve{J}\left(\sigma^2\right), \tag{2.9}$$

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 2.8 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

$$-2\ell_\phi + \lambda J(\phi) = \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} \phi(\boldsymbol{v}_{ijk})\, y_{ik} \right)^2 + \lambda J(\phi), \tag{2.10}$$

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 2.10 and setting $\phi = \phi^*$, we update our estimate of σ^2 by minimizing

$$-2\ell_{\sigma^2} + \breve{\lambda}\breve{J}\left(\sigma^2\right) = \sum_{i=1}^{N} \sum_{j=2}^{m_i} \log \sigma_{ij}^2 + \sum_{i=1}^{N} \sum_{j=1}^{m_i} \sigma_{ij}^{-2} r_{ij}^{*\,2} + \breve{\lambda}\breve{J}\left(\sigma^2\right), \tag{2.11}$$

where the $\left\{r_{ij}^{*\,2} = \left(y_{ij} - \sum_{k<j} \phi^*(\boldsymbol{v}_{ijk})\, y_{ik}\right)\right\}$ 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.

The remainder of the chapter is reserved for presenting two functional representations of (ϕ, σ^2). The first leverages the rich theoretical foundation of reproducing kernel Hilbert space techniques for function estimation. This framework has been studied extensively for the problem of estimating a function nonparametrically (see Aronszajn [1950], Wahba [1990], and Berlinet and Thomas-Agnan [2011] for detailed examinations), but to our knowledge has received little attention in the context of covariance models. We use a smoothing spline ANOVA decomposition of the varying coefficient function ϕ to construct a flexible class of covariance models while simultaneously maintaining interpretability. The second approach is based on the penalized B-splines, or

P-splines, of Eilers and Marx [1996]; these models exhibit many of the attractive numerical properties of the basis functions on which they are built. 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.

2.1 Smoothing spline representation of ϕ, σ

2.1.1 An RKHS framework for estimating ϕ

This section presents a method for regularized estimation of the varying coefficient function ϕ using a reproducing kernel Hilbert space (RKHS) framework. To do so, we first must establish some notation and review the relevant mathematical details of the surrounding framework. 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 $v \in \mathcal{V}$ 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 $\{a_1, \ldots, a_{n_1}\}$, $\left\{a_1, \ldots, a'_{n_2}\right\} \in \Re$,

$$\langle \sum_{i=1}^{n_1} a_i Q\left(\cdot, \boldsymbol{v}_i\right), \sum_{j=1}^{n_2} a_j' Q\left(\cdot, \breve{\boldsymbol{v}}_j\right) \rangle_{\mathcal{H}}. \tag{2.12}$$

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 $\psi\left(\boldsymbol{v}_{ij}\right)$ as the inner product of itself with the reproducing kernel:

$$\psi_{ij}\left(\boldsymbol{v}\right) = \langle \psi_{ij}, Q\boldsymbol{v} \rangle = L_{ij} Q\boldsymbol{v} = L_{ij(\cdot)} Q\left(\boldsymbol{v}, \cdot\right) \tag{2.13}$$

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 $\psi_{ij}\left(\boldsymbol{v}\right)$ by applying L_{ij} to $Q\left(\boldsymbol{v}, \boldsymbol{v}^*\right)$, considered as a function of $\boldsymbol{v}^*$. Wahba [1990] established an explicit form for the minimizer of the penalized sums of squares

$$-2\ell_\phi + \lambda J\left(\phi\right) = \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} \phi\left(\boldsymbol{v}_{ijk}\right) y_{ik} \right)^2 + \lambda J\left(\phi\right), \tag{2.14}$$

which can now be written

$$-2\ell_\phi + \lambda J\left(\phi\right) = \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma_{ij}^{-2} \left(y_{ij} - \sum_{k<j} \left(L_{ijk}\phi\right) y_{ik} \right)^2 + \lambda ||P_J \phi||_{\mathcal{H}}^2, \tag{2.15}$$

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

Theorem 2.1.1. *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 2.15 is given by*

$$\phi_\lambda = \sum_{\nu=1}^{d_0} d_\nu \eta_n u + \sum_{\boldsymbol{v}_i \in V} c_i \xi_i, \tag{2.16}$$

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 the Appendix, Chapter 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

$$\begin{aligned}\mathcal{H} &= [\mathcal{H}_{0[1]} \oplus \mathcal{H}_{1[1]}] \otimes [\mathcal{H}_{0[2]} \oplus \mathcal{H}_{1[2]}] \\ &= [\mathcal{H}_{0[1]} \otimes \mathcal{H}_{0[2]}] \oplus [\mathcal{H}_{0[1]} \otimes \mathcal{H}_{1[2]}] \oplus [\mathcal{H}_{1[1]} \otimes \mathcal{H}_{0[2]}] \oplus [\mathcal{H}_{1[1]} \otimes \mathcal{H}_{1[2]}]\end{aligned} \tag{2.17}$$

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

$$\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2, \tag{2.18}$$

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

Example 2.1.1. 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

$$\begin{aligned}\langle f, g \rangle &= \langle f, g \rangle_0 + \langle f, g \rangle_1 \\ &= \sum_{\nu=0}^{m_i - 1} M_\nu f M_\nu g + \int_0^1 f^{(m_i)}(v)\, g^{(m_i)}(v)\, dv, \quad i = 1, 2\end{aligned} \tag{2.19}$$

where the order i differential operator M_ν is defined $M_\nu \phi = \int_0^1 \phi^{(m)}(v)\, dv$, $\nu = 1, \ldots, m_i$, $i = 1, 2$. 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

$$Q_{[i]_v}'(v) = k_{m_i}(v)\, k_{m_i}(v') + (-1)^{m_i - 1} k_{2m_i}(v' - v) \tag{2.20}$$

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

$$\{\phi : \phi'' \in \mathcal{L}_2[0,1]\} = \{\phi : \phi \propto 1\} \oplus \{\phi : \phi \propto k_1\} \oplus \left\{\phi : \int_0^1 \phi dv = \int_0^1 \phi' dv = 0,\; \phi'' \in \mathcal{L}_2[0,1]\right\} \tag{2.21}$$

$$= \mathcal{H}_{00} \oplus \mathcal{H}_{01} \oplus \mathcal{H}_1, \tag{2.22}$$

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

$$Q_{00}(v, v') = 1 \tag{2.23}$$

$$Q_{01}(v, v') = k_1(v) \, k_1(v') \tag{2.24}$$

$$Q_1(v, v') = k_2(v) \, k_2(v') - k_4(v - v') \, . \tag{2.25}$$

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 2.1 and Table 2.2, 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 2.1: *Construction of the tensor product cubic spline subspace from marginal subspaces* $\mathcal{H}_{[1]}$, $\mathcal{H}_{[2]}$

Table 2.2: *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

$$\mathcal{H} = \bigoplus_{\beta=0}^{g} \mathcal{H}_\beta \tag{2.26}$$

and

$$J(\phi) = \sum_{\beta=0}^{g} \theta_\beta^{-1} J_\beta(\phi_\beta), \tag{2.27}$$

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 2.7, and stack these on top of one another to obtain

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix}, \tag{2.28}$$

which has dimension $n_y \times |V|$. Then the solution ϕ_λ minimizing 2.15 is the solution to the minimization problem

$$||D^{-1/2}\left(Y - X\left(Bd + Qc\right)\right)||^2 + \lambda c' Q c \tag{2.29}$$

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

Example 2.1.2. 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 $t_1, \ldots, t_M$,

$$X_i = \begin{bmatrix} y_{i,t_1} & 0 & 0 & 0 & & \ldots & 0 \\ 0 & y_{i,t_1} & y_{i,t_2} & 0 & 0 & \ldots & 0 \\ \vdots & & & & & & \\ 0 & \ldots & 0 & \ldots & y_{i,t_1} & \ldots & y_{i,t_{M-1}} \end{bmatrix} \tag{2.30}$$

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

Example 2.1.3. 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:

$$X_1 = \begin{bmatrix} 0 & y_{1,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{1,1} & 0 & y_{1,2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & y_{1,1} & y_{1,2} & y_{1,3} \end{bmatrix} \tag{2.31}$$

and similarly, for subject 2:

$$X_2 = \begin{bmatrix} y_{2,1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & y_{2,1} & y_{2,2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & y_{2,1} & y_{2,2} & y_{2,3} & 0 & 0 & 0 \end{bmatrix} \tag{2.32}$$

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

$$\frac{\partial}{\partial c}\left[-2\ell_{\phi}+\lambda J\left(\phi\right)\right]=QX'D^{-1}\left[X\left(Bd+Qc\right)-Y\right]+\lambda Qc=0$$
$$\iff X'D^{-1}X\Big[Bd+Qc\Big]+\lambda c=X'D^{-1}Y \tag{2.33}$$

$$\frac{\partial}{\partial d}\left[-2\ell_{\phi}+\lambda J\left(\phi\right)\right]=B'X'D^{-1}\left[X\left(Bd+Qc\right)-Y\right]=0$$
$$\iff -\lambda B'c=0 \tag{2.34}$$

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

$$Y=X\Big[Bd+\Big(Q+\lambda\left(X'D^{-1}X\right)^{-1}\Big)c\Big] \tag{2.35}$$
$$B'c=0 \tag{2.36}$$

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

$$-2\ell_{\lambda}\left(c,d\right)+\lambda J\left(\phi\right)=\Big[\tilde{Y}-\tilde{B}d-\tilde{Q}c\Big]'\Big[\tilde{Y}-\tilde{B}d-\tilde{Q}c\Big]+\lambda c'Qc. \tag{2.37}$$

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

$$\begin{aligned}\tilde{B}'\tilde{B}d+\tilde{B}'\tilde{Q}c&=\tilde{B}'\tilde{Y}\\ \tilde{Q}'\tilde{B}d+\tilde{Q}'\tilde{Q}c+\lambda Qc&=\tilde{Q}'\tilde{Y},\end{aligned} \tag{2.38}$$

Some algebra yields that this is equivalent to solving the system

$$\begin{bmatrix}\tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q}\\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q}+\lambda Q\end{bmatrix}\begin{bmatrix}d\\ c\end{bmatrix}=\begin{bmatrix}\tilde{B}'\tilde{Y}\\ \tilde{Q}'\tilde{Y}\end{bmatrix} \tag{2.39}$$

Fixing smoothing parameters λ and θ_β (hidden in Q and $\tilde{Q}$ if present), assuming that $\tilde{Q}$ is full column rank, 2.39 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

$$\begin{bmatrix} \tilde{B}'\tilde{B} & \tilde{B}'\tilde{Q} \\ \tilde{Q}'\tilde{B} & \tilde{Q}'\tilde{Q}+\lambda Q \end{bmatrix} = \begin{bmatrix} C_1' & 0 \\ C_2' & C_3' \end{bmatrix} \begin{bmatrix} C_1 & C_2 \\ 0 & C_3 \end{bmatrix} \tag{2.40}$$

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

$$\tilde{C}_3 = \begin{bmatrix} H_1 & H_2 \\ 0 & \delta I \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} C_1 & C_2 \\ 0 & \tilde{C}_3 \end{bmatrix}; \tag{2.41}$$

then

$$\tilde{C}^{-1} = \begin{bmatrix} C_1^{-1} & -C_1^{-1}C_2\tilde{C}_3^{-1} \\ 0 & \tilde{C}_3^{-1} \end{bmatrix}. \tag{2.42}$$

Premultiplying 2.40 by $\tilde{C}^{-T}$, straightforward algebra gives

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{C}_3^{-T}C_3^TC_3\tilde{C}_3^{-1} \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c} \end{bmatrix} = \begin{bmatrix} C_1^{-T}\tilde{B}'\tilde{Y} \\ \tilde{C}_3^{-T}\tilde{Q}'\left(I - \tilde{B}\left(\tilde{B}'\tilde{B}\right)^{-1}\tilde{B}'\right)\tilde{Y} \end{bmatrix} \tag{2.43}$$

where $\left(\tilde{d}' \;\; \tilde{c}'\right)' = \tilde{C}'\,(d \;\; c)'$. 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

$$\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{d} \\ \tilde{c}_1 \\ \tilde{c}_2 \end{bmatrix} = \begin{bmatrix} * \\ * \\ 0 \end{bmatrix}, \tag{2.44}$$

which can be solved, but with c_2 arbitrary. One may perform the Cholesky decomposition of 2.39 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

$$\widehat{\tilde{Y}} = \tilde{B} d + \tilde{Q} c = \begin{bmatrix} \tilde{B} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1} \tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}' \end{bmatrix} \tilde{Y} = \tilde{A}(\lambda, \boldsymbol{\theta}) \tilde{Y}. \tag{2.45}$$

where

$$\begin{aligned}\tilde{A}(\lambda, \boldsymbol{\theta}) &= \begin{bmatrix} \tilde{B} & \tilde{Q} \end{bmatrix} \tilde{C}^{-1} \tilde{C}^{-T} \begin{bmatrix} \tilde{B}' \\ \tilde{Q}' \end{bmatrix} \\ &= G + (I - G) \tilde{Q} \left[\tilde{Q}' (I - G) \tilde{Q} + \lambda Q \right]^{-1} \tilde{Q}' (I - G),\end{aligned} \tag{2.46}$$

for

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

2.1.2 Smoothing parameter selection

By varying smoothing parameters λ and θ_β, the minimizer ϕ_λ of 2.39 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

$$\Phi = D^{-1/2}X\left[Bd + Qc\right]. \tag{2.47}$$

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

$$\begin{aligned} L(\lambda) &= \sum_{i=1}^{N}\sum_{j=1}^{m_i}\left(\hat{\tilde{y}}_{ij} - E\left[\tilde{y}_{ij}\right]\right)^2 \\ &= ||\tilde{Y} - \tilde{\mu}||^2 \end{aligned} \tag{2.48}$$

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

$$L(\lambda) = \mu' \left(I - \tilde{A}\right)^2 \mu - 2\mu' \left(I - \tilde{A}\right)^2 \tilde{A}\tilde{\epsilon} + \tilde{\epsilon}' \tilde{A}^2 \tilde{\epsilon} \tag{2.49}$$

Define the unbiased risk estimate

$$U(\lambda) = \frac{1}{N}\tilde{Y}' \left(I - \tilde{A}\right)^2 \tilde{Y} + \frac{2}{N}\text{tr}\tilde{A} \tag{2.50}$$

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

$$\begin{aligned} U(\lambda) &= \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y}\right)' \left(\tilde{Y} - \mu + \mu - \tilde{A}\tilde{Y}\right) + 2\text{tr}\tilde{A} \\ &= \left(\tilde{A}\tilde{Y} - \mu\right)' \left(\tilde{A}\tilde{Y} - \mu\right) + \tilde{\epsilon}'\tilde{\epsilon} + 2\tilde{\epsilon}' \left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right) \end{aligned} \tag{2.51}$$

This gives

$$U(\lambda) - L(\lambda) - \tilde{\epsilon}'\tilde{\epsilon} = 2\tilde{\epsilon}' \left(I - \tilde{A}\right)\mu - 2\left(\tilde{\epsilon}'\tilde{A}\tilde{\epsilon} - \text{tr}\tilde{A}\right), \tag{2.52}$$

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

$$V_{loso}(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right)'\left(\tilde{Y}_i - \widehat{\tilde{\mu}}_i^{[-i]}\right) \tag{2.53}$$

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:

$$\begin{aligned} MSPE &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - \widehat{\tilde{\mu}}_i||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N} E\left[||\tilde{Y}_i^* - D_i^{-1/2}W_i\Phi^* + D_i^{-1/2}W_i\Phi^* - D_i^{-1/2}W_i\hat{\Phi}^*||^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\tilde{\mu}_i - \widehat{\tilde{\mu}}_i^{[-i]}||^2\right]\right\} \end{aligned} \tag{2.54}$$

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

$$E\left[V_{loso}(\lambda)\right] = \frac{1}{N}\sum_{i=1}^{N}\left\{m_i + E\left[||\widehat{\tilde{\mu}}_i - \tilde{\mu}_i||^2\right]\right\}. \tag{2.55}$$

When N is large, we expect that $\widehat{\tilde{\mu}}_i$ should be close to $\widehat{\tilde{\mu}}_i^{[-i]}$, so $E\left[V_{loso}(\lambda)\right]$ 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 2.1.2 (Shortcut formula for LosoCV)**.** *The LosoCV score satisfies the following identity:*

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

where $\tilde{A}_{ii}$ is the diagonal block of smoothing matrix $\tilde{A}$ 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. 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}:

$$V^*_{loso}(\lambda) = \frac{1}{N}||\left(I - \tilde{A}\right)\tilde{Y}||^2 + \frac{2}{N}\sum_{i=1}^{N}\hat{\tilde{e}}_i'\tilde{A}_{ii}\hat{\tilde{e}}_i, \tag{2.56}$$

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

2.1.3 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 2.46 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 **??**, 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 a straightforward task.

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 $\tilde{A}\left(\lambda|\boldsymbol{\theta}^*\right) = \tilde{A}\left(\lambda, \exp\left(\boldsymbol{\kappa}^*\right)\right)$.
Minimize $U\left(\lambda|\boldsymbol{\kappa}^*\right) = \tilde{Y}'\left(I - \tilde{A}\right)^2 \tilde{Y} + 2\text{tr}\tilde{A}$
Set $U_* := \min_{\lambda} Y\left(\lambda|\boldsymbol{\kappa}^*\right)$
if $U^* > U_{-}$ **then**
Set $\Delta\boldsymbol{\kappa} := \Delta\boldsymbol{\kappa}/2$
Go to (1).
else
Continue
end if
Evaluate gradient $\mathbf{g} = \left(\partial/\partial\boldsymbol{\kappa}\right) U\left(\boldsymbol{\kappa}|\lambda\right)$
Evaluate Hessian $H = \left(\partial^2/\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}'\right) U\left(\boldsymbol{\kappa}|\lambda\right)$.
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 $\tilde{H} = \text{diag}\left(\boldsymbol{\epsilon}\right)$ 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

$$(\lambda_1, \ldots, \lambda_q) = \left(\lambda\theta_1^{-1}, \ldots, \lambda\theta_q^{-1}\right), \tag{2.57}$$

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.

2.1.4 An RKHS framework for estimating $\log \sigma^2$

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 2.1.1. Fixing $\phi = \phi^*$ for given estimate ϕ^*, the negative log likelihood of the data $Y_1, \ldots, Y_N$ is satisfies

$$-\ell\left(Y_1, \ldots, Y_N, \phi, \sigma^2\right) = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\log\sigma_{ij}^2 + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{\epsilon_{ij}^2}{\sigma_{ij}^2}; \tag{2.58}$$

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

$$\text{RSS}(t) = \sum_{i,j:t_{ij}=t}\left(y_{ij} - \sum_{k<j}\phi_{ijk}y_{ik}\right)^2 \tag{2.59}$$

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 2.58 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

$$-\ell(z, \mu, \alpha) \propto \alpha\left[\frac{z}{\mu} + \log\mu\right] \tag{2.60}$$

$$= \alpha\left[ze^{-\eta} + \eta\right], \tag{2.61}$$

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

$$-\ell\left(Z_1,\ldots,Z_N,\phi,\sigma^2\right)=\sum_{i=1}^{N}\sum_{j=1}^{m_i}\log\mu_{ij}+\sum_{i=1}^{N}\sum_{j=1}^{m_i}\frac{z_{ij}}{\mu_{ij}}, \tag{2.62}$$

which we can see coincides with a Gamma dsitribution with scale parameter $\alpha = 2$. Smoothing spline ANOVA models for exponential familes 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 $\eta(t) = \log\sigma^2(t)$:

$$-\ell\left(Z_1,\ldots,Z_N,\phi,\sigma^2\right)+=\sum_{i=1}^{N}\sum_{j=1}^{m_i}\eta_{ij}+\sum_{i=1}^{N}\sum_{j=1}^{m_i}z_{ij}e^{-\eta_{ij}}+\lambda J(\eta), \tag{2.63}$$

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 (2.27), 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 (2.63) 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

$$\eta(t)=\sum_{\nu=1}^{d_0}d_\nu\kappa_\nu(t)+\sum_{i=1}^{|\mathcal{T}|}c_iQ_J(t,t_i), \tag{2.64}$$

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

$$
\begin{aligned}
L(\eta) &= -\sum_{i=1}^{N}\sum_{j=1}^{m_i}\left[z_{ij}\eta(t_{ij}) - b(\eta(t_{ij}))\right] \\
&= -\sum_{i=1}^{N}\sum_{j=1}^{m_i}\left[z_{ij}\eta(t_{ij}) - b(\eta(t_{ij}))\right]
\end{aligned}
\tag{2.65}
$$

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 $\eta_\nu(v_i)$ is full column rank, so that $L(f)$ is strictly convex in $\mathcal{H}$ and the minimizer of (2.63) uniquely exists. See Wahba et al. [1995].

For fixed λ and θ_β, which may be hidden in J, the penalized log likelihood (2.63) is convex in η, so that the minimizer can be computed via Newton iteration. Let $u_{ij} = -z_{ij} + b'(\tilde{\eta}(t_{ij})) = -z_{ij} + \tilde{\mu}(t_{ij})$, and $\tilde{\omega}_{ij} = b''(\tilde{\eta}(t_{ij})) = \tilde{v}(t_{ij})$. The quadratic approximation of $-z_{ij}\eta(t_{ij}) + b(\eta(t_{ij}))$ at $\tilde{\eta}(t_{ij})$ is given by

PLACEHOLDER - TO DO

2.1.5 Smoothing parameter selection for exponential familes

The gamma penalized log likelihood (2.64) is non-quadratic, so η_λ must be computed using iteration even for fixed smoothing parameter. Performance-oriented iteration and generalized approximate cross validation (GACV) are the most common approaches to selecting the smoothing

parameter for penalized regression with exponential families. As in our discussion of model selection for ϕ, we omit dependence of any components on the θ_β and only explicitly express dependence on smoothing parameters through λ.

Performance-oriented iteration

A measure of the discrepancy between distributions belonging to an exponential family with density $p_Z(z) = exp\{(y\eta(t) - b(\eta(t)))/a(\phi) + c(y,\phi)\}$ is the Kullback-Leibler distance

$$\begin{aligned} \text{KL}(\eta, \eta_\lambda) &= E_\lambda [Z(\eta - \eta_\lambda) - (b(\eta) - b(\eta_\lambda))]/a(\phi) \\ &= [b'(\eta)(\eta - \eta_\lambda) - (b(\eta) - b(\eta_\lambda))]/a(\phi), \end{aligned} \tag{2.66}$$

which simplifies to

$$-\mu\left(e^{-\eta} - e^{-\tilde{\eta}}\right) - (\eta - \tilde{\eta})$$

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

$$\begin{aligned} \text{SKL}(\eta, \eta_\lambda) &= \text{KL}(\eta, \eta_\lambda) + \text{KL}(\eta_\lambda, \eta) \\ &= (b'(\eta) - b'(\eta_\lambda))(\nu - \nu_\lambda)/a(\phi), \\ &= (\mu - \mu_\lambda)(\nu - \nu_\lambda)/a(\phi), \end{aligned} \tag{2.67}$$

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 $t_{11}, \ldots, t_{1,m_1}, \ldots, t_{N1}, \ldots, t_{N,m_N}$:

$$L(\eta, \eta_\lambda) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_i}(\mu(t_{ij}) - \mu_\lambda(t_{ij}))(\nu(t_{ij}) - \nu_\lambda(t_{ij})), \tag{2.68}$$

which reduces to

$$L(\eta, \eta_\lambda) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_i}(\mu(t_{ij}) - \mu_\lambda(t_{ij}))(\nu(t_{ij}) - \nu_\lambda(t_{ij})), \tag{2.69}$$

for the Gamma distribution. The ideal smoothing parameters are those which minimize (2.69). The performance-oriented iteration operates on a alternative expression of the symmetrized Kullback-Leibler loss. The mean value theorem gives us that (2.69) can be written

$$L_{\omega}\left(\eta,\eta_{\lambda}\right)=L\left(\eta,\eta_{\lambda}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_{i}}\omega^{*}\left(t_{ij}\right)\left(\nu\left(t_{ij}\right)-\nu_{\lambda}\left(t_{ij}\right)\right)^{2}, \tag{2.70}$$

where $\omega^{*}\left(t_{ij}\right)=b''\left(\eta^{*}\left(t_{ij}\right)\right)$ and $\eta^{*}\left(t_{ij}\right)$ is a convex combination of $\eta\left(t_{ij}\right)$ and $\eta_{\lambda}\left(t_{ij}\right)$. 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} ist he $m_{i}\times m_{i}$ diagonal matrix having diagonal entries $\omega^{*}\left(t_{i1}\right),\ldots,\omega^{*}\left(t_{i,m_{i}}\right)$, an unbiased estimate of relative loss is given by

$$U_{\omega}\left(\lambda\right)=L\left(\eta,\eta_{\lambda}\right)=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{N}\sum_{j=1}^{m_{i}}\omega^{*}\left(t_{ij}\right)\left(\nu\left(t_{ij}\right)-\nu_{\lambda}\left(t_{ij}\right)\right)^{2}, \tag{2.71}$$

PLACEHOLDER - TO DO

Chapter 3: Modeling the Cholesky decomposition via penalized B-splines

TO DO: clean this up, omitting all of the non-varying coefficient model discussion; move the pertinant B-spline discussion to the appendix.

3.0.1 A B-spline representation for pp functions

Definition 3.0.1. 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

$$B_{i,k,t}(x) = (t_{i+k} - t_i)[t_i, \ldots, t_{i+k}](\cdot - x)_+^{k-1} \tag{3.1}$$

The placeholder notation, $(\cdot - x)_+^{k-1}$, is used to indicate that the k^{th} divided difference of the 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_i rather than $B_{i,k,t}$ when the spline order and knot sequence can be inferred from surrounding context.

3.0.2 Properties of B-splines

I. $B_i(x)$ has isolated support:

$$B_i(x) = 0, \quad x \notin [t_i, t_{i+k}]$$

To see this, note that if $x \notin [t_i, t_{i+k}]$, then $g(t) = (t-x)_+^{k-1}$ is a polynomial of degree $< k$ on $[t_i, t_{i+k}]$, thus by **?? ??**,

$$[t_i, \ldots, t_{i+k}]\, g = 0.$$

As a result, for a set of B-splines of order k corresponding to the knot sequence t, only k of them are nonzero on $[t_j, t_{j+k}]$: $B_{j-k+1}, B_{j-k+2}, \ldots, B_j$.

II. The i^{th} B-spline of order is defined as the k^{th} divided difference of $(\cdot - x)_+^{k-1}$ times a normalization factor: $(t_{i+k} - t_i)$. This normalization, using **?? ??**, allows us to write

$$B_i(x) = [t_{i+1}, \ldots, t_{i+k}](\cdot - x)_+^{k-1} - [t_i, \ldots, t_{i+k-1}](\cdot - x)_+^{k-1} \tag{3.2}$$

For $x \in (t_j, t_{j+1})$, by **?? ??**,

$$\begin{aligned}
\sum_i B_i(x) &= \sum_{i=j+1-k}^{j} B_i(x) \\
&= \sum_{i=j+1-k}^{j} [t_{i+1}, \ldots, t_{i+k}](\cdot - x)_+^{k-1} - \sum_{i=j+1-k}^{j} [t_i, \ldots, t_{i+k-1}](\cdot - x)_+^{k-1} \\
&= [t_{j+1}, \ldots, t_{j+k}](\cdot - x)_+^{k-1} - [t_{j+1-k}, \ldots, t_j](\cdot - x)_+^{k-1} \\
&= 1 - 0
\end{aligned} \tag{3.3}$$

The last equality in 3.3 is a consequence of the following: for $x \in (t_j, t_{j+1})$, $g(t) = (t - x)_+^{k-1}$ is a $k-1$ degree polynomial with unit leading coefficient on $[t_{j+1}, t_{j+k}]$, so by **?? ??**,

$$[t_{j+1}, \ldots, t_{j+k}]\, g = 1.$$

On $[t_{j+1-k}, t_j]$, g is identically 0, hence $[t_{j+1-k}, \ldots, t_j]\, g = 0$.

III. Each $B_i(x)$ is positive on its support. Applying Leibnitz's formula (**?? ??**) to the product

$$[t_i, \ldots, t_{i+k}](t - x)_+^{k-1} = [t_i, \ldots, t_{i+k}](t - x)(t - x)_+^{k-2},$$

we have

$$\begin{aligned}
[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-1} &= [t_i,\ldots,t_{i+k}]\,(t-x)\,(t-x)_+^{k-2} \\
&= \sum_{r=i}^{i+k} [t_i,\ldots,t_{i+r}]\,(t-x)\,[t_r,\ldots,t_{i+k}]\,(t-x)_+^{k-2} \\
&= \Big[\,[t_i]\,(t-x)\,\Big]\Big[\,[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-2}\,\Big] \\
&\qquad + \Big[\,[t_i,t_{i+1}]\,(t-x)\,\Big]\Big[\,[t_{i+1},\ldots,t_{i+k}]\,(t-x)_+^{k-2}\,\Big] \\
&= (t_i-x)\,[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-2} \\
&\qquad + 1\cdot[t_{i+1},\ldots,t_{i+k}]\,(t-x)_+^{k-2} \qquad (3.4)
\end{aligned}$$

since $[t_i,\ldots,t_j]\,(\cdot-x)=0$ for $j>i+1$. By **?? ??**,

$$(t_i-x)\,[t_i,\ldots,t_{i+k}]\,g=\frac{t_i-x}{t_{i+k}-t_i}\Big[\,[t_{i+1},\ldots,t_{i+k}]\,g-[t_i,\ldots,t_{i+k-1}]\,g\Big],$$

and we may express 3.4 as

$$\begin{aligned}
[t_i,\ldots,t_{i+k}]\,(\cdot-x)_+^{k-1} &= \frac{x-t_i}{t_{i+k}-t_i}\,[t_i,\ldots,t_{i+k-1}]\,(\cdot-x)_+^{k-2} \\
&\quad + \frac{t_{i+k}-x}{t_{i+k}-t_i}\,[t_{i+1},\ldots,t_{i+k}]\,(\cdot-x)_+^{k-2}
\end{aligned}$$

which we can write in terms of the normalized B-spline:

$$\frac{B_{i,k}\,(x)}{t_{i+k}-t_i}=\frac{x-t_i}{t_{i+k}-t_i}\frac{B_{i,k-1}\,(x)}{t_{i+k-1}-t_i}+\frac{t_{i+k}-x}{t_{i+k}-t_i}\frac{B_{i+1,k-1}\,(x)}{t_{i+k}-t_{i+1}} \qquad (3.5)$$

This shows that we can write the i^{th} B-spline of order k as a convex combination of the i^{th} and $(i+1)^{st}$ B-splines of order $k-1$ since

$$\frac{x-t_i}{t_{i+k}-t_i}+\frac{t_{i+k}-x}{t_{i+k}-t_i}=1,$$

and each of these weights are positive for $t_i<x<t_{i+1}$. If

$$B_{j,k-1}\,(x)>0,\quad t_j<x<t_{j+k-1} \text{ for all } j,$$

then by 3.5, we have that

$$B_{i,k}(x) > 0, \qquad t_i < x < t_{i+k}$$

since $B_{j,k-1} = 0$ for $x \notin [t_j, t_{j+k}]$ by 3.0.2 I and by induction over k, starting with the fact that

$$B_j, 1(x) = \begin{cases} 1 & t_j \leq x < t_{j+1} \\ 0 & otherwise \end{cases}$$

Properties I, II, and III demonstrate that a sequence of B-splines form a *partition of unity*: a set of non-negative functions which sum, pointwise, to one.

Definition 3.0.2. The *B-representation of* $f \in \mathcal{P}_{k,\xi,\nu}$ consists of

I. integers k and n specifying the order of f as a pp function and the number of linear parameters,

$$n = kl - \sum_i \nu_i = \dim(\mathcal{P}_{k,\xi,\nu}),$$

respectively.

II. The knot vector $t = \{t_i\}$, $i = 1, \ldots, n+k$ with elements arranged in increasing order, constructed according to Theorem **??**, via ξ and ν.

III. The B-spline coefficients $\alpha = \{\alpha_i\}$, $i = 1, \ldots, n$ for the knot sequence, t.

Given I, II, and III in 3.0.2, the function value at $x \in [t_k, t_{n+1}]$ is given by

$$f(x) = \sum_{i=1}^{n} \alpha_i B_i(x),$$

and in particular, by I, for $x \in [t_j, t_{j+1}]$,

$$f(x) = \sum_{i=j}^{j+k-1} \alpha_i B_i(x).$$

3.0.3 Single-regressor varying coefficient models via B-spline basis expansions

Hastie and Tibshirani [1993] were the first to introduce the varying coefficient model, which supplies a modeling approach which permits interpolation of regressors and response variables which varying according to an *indexing variable* at values of this indexing variable where there is either missing data of only a single observation and slope estimation is not feasible. In the section that follows, we will discuss the approach to smoothing the coefficient vector (and *not* the regressor, $x(t)$) first, for mechanical demonstration of parameterization and estimation of the coefficient function via B-spline basis expansion, at a predetermined set of values of an indexing variable, t (knots), then following the approach of Eilers and Marx by assuming that the number and position of the knots are unknown and using penalized B-splines, or P-splines.

Consider data of the form

$$(x_i, y_i, t_i), \qquad i = 1, \ldots, m$$

where y_i is the response, x_i is the single (univariate) regressor variable, and t_i is an indexing variable. We first consider a simple situation as an introductory warmup for demonstrating the mechanics of the varying coefficient model. Suppose we wish to fit a scatterplot smoother to the points (t_i, y_i) using a B-spline basis expansion. Assume that we can model

$$y(t) = f(t) + \epsilon(t) \tag{3.6}$$

where ϵ is a zero-mean error process. Modeling the mean function as a q^{th}-order B-spline, we can rewrite 3.6 as

$$y(t) = \sum_{j=1}^{K} \alpha_j B_j(t) + \epsilon(t) \tag{3.7}$$

Assume we use K of basis functions in our expansion of f. Let $y = (y_1, \ldots, y_M)^T$, and let B denote the $M \times K$ design matrix with $i - j^{th}$ element given by the j^{th} order-q B-spline evaluated at the i^{th} value of t:

$$b_{ij} = B_j(t_i),$$

$i = 1, \ldots, m$, $j = 1, \ldots, K$. Then in matrix notation, we may write the mean vector

$$\mu = E[Y] = B\alpha$$

where α is the vector of K unknown basis coefficients. We take $\hat{\alpha}$ to be the minimizer of

$$\begin{aligned} S &= \sum_{i=1}^{M} \left(y_i - \sum_{j=1}^{K} \alpha_j B_j(t_i) \right)^2 \\ &= |y - B\alpha|^2 \end{aligned} \tag{3.8}$$

$$B^T B\alpha = B^T y$$

which has explicit solution

$$\hat{\alpha} = \left(B^T B\right)^{-1} B^T y$$

Given $\hat{\alpha}$, one may estimate the response at any new value of t, say t^*, by

$$\hat{y}(t^*) = \sum_{j=1}^{K} \hat{\alpha}_j B_j(t^*).$$

3.0.4 B-spline estimators for varying coefficient models with fixed knots

To extend the varying intercept model 3.6 to accommodate for controlling for another regressor, it is natural to consider the varying coefficient model; the single regressor varying-coefficient (VC)

model extends the classical linear model by allowing the slope coefficient to vary smoothly in the dimension of the indexing variable, t. The single-index varying coefficient model assumes that the mean response is of the form

$$E\left[Y\left(t\right)\right] = \beta_0\left(t\right) + \beta_1\left(t\right)x\left(t\right) \tag{3.9}$$

where $\beta_0\left(t\right)$ is the smooth varying intercept function and $\beta_1\left(t\right)$ is the smooth slope function of interest. This model generalizes the well known simple linear regression model

$$E\left[Y\left(t\right)\right] = \beta_0 + \beta_1 x\left(t\right)$$

by trading the static regression coefficients for smooth coefficient functions which are assumed to varying across an indexing variable, t. This allows for the regressor variable to having modified effect, depending on the value of t. Using a set of predetermined knots along the t axis, the VC model can be fit in a fashion similar to that required for fitting model 3.6, requiring only minor adjustments to the design matrix. In matrix notation as described in 3.8, the mean vector may be written

$$\mu = B\alpha_0 + diag\left\{x\left(t\right)\right\}B\alpha_1 \tag{3.10}$$

where $diag\left\{x\left(t\right)\right\}$ is the $m \times m$ diagonal matrix of regressor measurements which ensures that the varying coefficients are appropriately weighted according to the correct value of x by aligning the regressor function with the corresponding slope value. Letting $X = diag\left\{x\left(t\right)\right\}B$, 3.10 becomes

$$\mu = \left[B|X\right]\left(\alpha_0^T, \alpha_1^T\right)^T \tag{3.11}$$

$$\equiv Q\alpha \tag{3.12}$$

where α is the augmented vector of basis coefficients. Here, the same basis is used for smoothing both the varying intercept as well as the varying slope function; this is feasible because both components varying along the same indexing variable. One can relax this structure and allow each additive term to varying according to its own indexing variable. This, of course, requires a separate B-spline basis for each component. Again using least squares techniques as with the varying intercept-only model, we take $\hat{\alpha}$ to minimize

$$S = |y - Q\alpha|^2 \tag{3.13}$$

which has explicit solution

$$\hat{\alpha} = \left(Q^T Q\right)^{-1} Q^T y.$$

It is worth noting that Q is simply a row scaling of the original B-spline design matrix, B; thus, accommodating a varying slope function equates to the simple basis function regression setting with a modified basis, XB. Using the modified basis functions as covariates, estimation of model the varying coefficient model equates to a multiple regression problem. Each of the estimated smooth components are given by

$$\hat{\beta}_k(t) = B\hat{\alpha}_k, \qquad k = 0, 1$$

and the estimate of the smooth mean function is obtained via

$$\begin{aligned} \hat{\mu} &= Q\hat{\alpha} \\ &= Hy \end{aligned}$$

where $H = Q\left(Q^T Q\right)^{-1} Q^T$ is the "hat" matrix, which is analogous to the smoothing matrix $\tilde{A}$ referred to in the discussion of the smoothing spline estimation of ϕ in Chapter 2. Its use in

smoothing parameter selection and model tuning is similar to the reproducing kernel Hilbert space framework, which we will discuss in the coming sections.

3.0.5 P-spline estimators for regularized estimation of fitted curves

The mechanics in the previous section rely on apriori knowledge of the number and locations of the knots $\{t_j\}$, $j = 1, \ldots, K$. In practice this information is readily available, but has a considerable impact on the behaviour of the estimated coefficient functions, as 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. This choice presents a model selection problem, as too many knots lead to overfitting while too few knots lead to underfitting. Optimal knot placement has been closely examined, with some authors proposing automatic methods for optimizing the number and the positions of the knots (Friedman and Silverman [1989], Stone et al. [1997]). This is a difficult numerical problem requiring 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.

As in Chapter 2, we can append a penalty on the coefficients of the basis functions to the goodness of fit measure, and by optimizing this augmented objective function, we can achieve as much smoothness in the fitted function as desired. O'Sullivan [1986] was the first to propose using a rich B-spline basis and applying a discrete penalty to the spline coefficients. He proposed a penalty on the second derivative to restrict the flexibility of the fitted curve, similar to the penalty pioneered for smoothing splines by Reinsch [1967]. This penalty has become the standard in much of the spline literature; see Eubank [1999], Wahba [1990]. This measure of roughness of a curve is given by

$$J = \int_l^u \left[f''(x)\right]^2 \, dx$$

where l and u are the bounds on the domain of x. Using the properties of B-splines, if $f(x) = \sum_j \beta_j B_j(x)$, one can derive a banded matrix P such that

$$J = \beta' P \beta$$

where $\beta = (\beta_1, \ldots, \beta_n)$, and the i-j^{th} element of P is given by

$$p_{ij} = \int_l^u B_i''(x) B_j''(x) \, dx.$$

He then proposed minimizing

$$\begin{aligned} Q(\beta, \lambda) &= \sum_{i=1}^{m} \left(y_i - \sum_j \beta_j B_j(x_i) \right)^2 + \lambda \int_l^u \left[f''(x)\right]^2 \, dx \\ &= |y - B\beta||^2 + \lambda \beta' P \beta \end{aligned}$$

The computation of P is nontrivial and becomes very tedious when the third and fourth derivative are used as the roughness measure. 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. 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

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

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

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

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

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

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.

Difference penalties

Imposing difference penalties on B-spline basis expansions generalizes and simplifies the approach outlined in the previous section in a way that permits application in any context where regression on B-splines is useful. Penalized B-splines, or *P-splines*, are an alternative an approach to nonparametric smoothing which circumvent any complexity associated with constructing such penalty matrices by ommitting derivatives and integrals altogether. Instead, smoothness is imposed via a discrete penalty matrix based on finite difference formulas which is simple to compute. This approach achieves smoothness in fitted functions in two ways:

I. To avoid the difficulty of choosing the optimal set of knots, use a B-spline basis with a large number of equally spaced knots, purposefully overfitting the smooth coefficient vectors.

II. Augment the goodness of fit measure with a difference penalty to prevent overfitting and accommodate a potentially ill-conditioned fitting procedure.

Using the properties of B-splines derived in B-spline section, it is relatively straightforward to show that the simplified penalty is nearly equivalent to the derivative-based penalty and that for

second order differences, P-splines are very similar to O'Sullivan's approach. In some applications, it can be useful to use differences of a smaller or higher order in the penalty, and the P-spline framework makes the use of a penalty of any arbitrary order nearly seamless.

Consider the varying intercept-only model defined in 3.6 for the regression of M data points (t_i, y_i) on a set of K B-splines, $\{B_j\}$. By letting the number of knots, K, be relatively large, we allow more variation in fitted curve than the data reasonably justify. To make the result less flexible and avoid overfitting, O'Sullivan imposed a penalty on the second derivative of the fitted curve and appended this to the residual sum of squares, giving way to the objective function

$$\sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{K} \beta_j B_j \left(t_i\right) \right\}^2 + \lambda \int_{t_{min}}^{t_{max}} \left\{ \sum_{j=1}^{K} \beta_j B_j'' \left(t\right) \right\}^2 \, dt. \tag{3.14}$$

The integral of the square of the second derivative of a fitted function has become common as a smoothness penalty since the seminal work on smoothing splines by Reinsch [1967], though it is useful to note that there is nothing particularly special about the second derivative. One could easily specify higher or lower order derivatives in smoothness penalties. In the context of smoothing splines, the first derivative leads to simple equations and a piecewise linear fit, while higher derivatives lead to systems of equations with a high bandwidth and a very smooth fit.

Proposed for smoothing curves by Whittaker [1922], difference penalties have been utilized for nearly a century, with more recent applications outlined in Eilers [1991a], Eilers [1991b], and Eilers [1995]. The finite difference penalty is easily introduced into regression equations, making it feasible to evaluate the impact of different orders of the differences on the fitted model. In some applications, it is useful to work with third and fourth order differences, since for high values of λ, the fitted curve approaches a parametric polynomial model. Detailed discussion on the effect of the

smoothing parameter on fitted functions will follow. Let D_d denote the d^{th} order matrix difference operator; that is, $D_d\beta = \Delta^d\beta$, where

$$\begin{aligned}\Delta\alpha_j &= \alpha_j - \alpha_{j-1},\\ \Delta^2\alpha_j &= \Delta\left(\Delta\alpha_j\right) = \alpha_j - 2\alpha_{j-1} + \alpha_{j-2},\end{aligned}$$

and in general,

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

The $(K-d)\times K$ differencing matrix D_d is sparse for reasonably small values of d; for example, D_1 and D_2 for small dimensions are given by

$$D_1 = \begin{bmatrix}-1 & 1 & 0 & 0\\ 0 & -1 & 1 & 0\\ 0 & 0 & -1 & 1\end{bmatrix}; \qquad D_2 = \begin{bmatrix}1 & -2 & 1 & 0\\ 0 & - & -2 & 1\end{bmatrix}$$

Eilers and Marx [1996] propose to base the penalty on (higher-order) finite differences of the coefficients of adjacent B-splines:

$$\lambda|D_d\alpha|^2 = \lambda\alpha' D_d' D_d\alpha = \lambda\alpha' P\alpha,$$

Replacing O'Sullivan's penalty with the difference penalty, we can control the smoothness of the fitted mean function $\mu = \beta_0(t) = B\alpha$ by minimizing

$$S_\lambda = |y - B\alpha|^2 + \lambda|D_d\alpha|^2$$

This approach reduces the dimensionality of the problem to the number of B-splines, K instead of the number of observations, M , as with smoothing splines. The tuning parameter λ permits continuous control over smoothness of the fit. We will demonstrate that the difference penalty is

a good discrete approximation to the integrated square of the k^{th} derivative, and with this penalty, moments of the data are conserved and polynomial regression models occur as limits for large values of λ. We will explore the connection between a penalty on second-order differences of the B-spline coefficients and O'Sullivan's choice of a penalty on the second derivative of the fitted function. However, the difference penalty can be handled mechanically for any order of the differences. O'Sullivan [1986] used third-degree B-splines and the following penalty:

$$h^2 P = \lambda \int_{t_{min}}^{t_{max}} \left\{ \sum_j \alpha_j B''_{j,3}(t) \right\}^2 dt \tag{3.15}$$

From the derivative properties of B-splines, it follows that

$$h^2 P = \lambda \int_{t_{min}}^{t_{max}} \sum_j \sum_k \Delta^2 \alpha_j \Delta^2 \alpha_k B_{j,1}(t) B_{k,1}(t) dt \tag{3.16}$$

Most of the cross products of $B_{j,1}(t)$ and $B_{k,1}(t)$ vanish since B-splines of degree 1 only overlap when j is $k-1$, k, or $k+1$. Thus, we have that

$$\begin{aligned} h^2 P &= \lambda \int_{t_{min}}^{t_{max}} \left[\left\{ \sum_j \Delta^2 \alpha_j B_j(t,1) \right\}^2 + 2 \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} B_j(t,1) B_{j-1}(t,1) \right] dt \\ &= \lambda \left[\sum_j \left(\Delta^2 \alpha_j \right)^2 \int_{t_{min}}^{t_{max}} B_j^2(t,1)\ dt + 2 \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} \right] \end{aligned} \tag{3.17}$$

or

$$\begin{aligned} h^2 P = \lambda \sum_j \left(\Delta^2 \alpha_j \right)^2 \int_{t_{min}}^{t_{max}} B_{j,1}^2(t)\, dt + 2\lambda \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} \\ + \int_{t_{min}}^{t_{max}} B_{j,1}(t) B_{j-1,1}(t)\, dt \end{aligned} \tag{3.18}$$

which can be written as

$$h^2 P = \lambda \left\{ c_1 \sum_j \left(\Delta^2 \alpha_j \right)^2 + c_2 \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} \right\} \tag{3.19}$$

where, for given equidistant knots, c_1 and c_2 are constants given by

$$\begin{aligned} c_1 &= \int_{t_{min}}^{t_{max}} B_{j,1}^2(t)\, dt \\ c_2 &= \int_{t_{min}}^{t_{max}} B_{j,1}(t)\, B_{j-1,1}(t)\, dt \end{aligned} \tag{3.20}$$

O'Sullivan's ridge-like B-spline penalty in Equation 3.15 can be written as a linear combination of a difference penalty (**??**) and the sum of the cross products of neighboring second differences. The second term in Equation 3.19 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 Equation 3.19 is used in the penalty. The additional complexity is due to overlapping B-splines, which quickly increases when using higher order differences and higher order B-splines. The use of a difference penalty allows us to sidestep the difficulty of constructing a procedure for incorporating the penalty in the likelihood equations.

Define $\hat{\alpha} = (\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_K)$ to be the minimizer of S_λ:

$$S_\lambda = \sum_{i=1}^{m} \left\{ y_i - \sum_{j=1}^{K} \alpha_j B_j(t_i) \right\}^2 + \lambda \sum_{j=d+1}^{K} \left(\Delta^d \alpha_j\right)^2$$

In vector notation, this may be written

$$\begin{aligned} S_\lambda &= |y - B\alpha|^2 + \lambda |D_d \alpha|^2 \\ &= (y - B\alpha)^T (y - B\alpha) + \lambda \alpha^T P \alpha \end{aligned} \tag{3.21}$$

where

$$P = D_d^T D_d$$

and the elements of B are given by $b_{ij} = B_j(t_i)$, as defined in 3.8. Taking derivatives on both sides of 3.21 with respect to α and setting equal to zero yields normal equations:

$$B^T y = \left(B^T B + \lambda D_d^T D_k\right) \alpha, \tag{3.22}$$

which has explicit solution

$$\hat{\alpha} = \left(B'B + \lambda D_d' D_d\right)^{-1} B'y$$

The effective hat matrix is now

$$H_\lambda - B\left(B^T B + \lambda D_k^T D_k\right)^{-1} B'.$$

When $\lambda = 0$, we have the standard normal equations of linear regression with a B-spline basis, and with $k = 0$ 3.22 corresponds to the normal equations under the ridge regression penalty. When $\lambda > 0$, the penalty only influences the main diagonal and k sub-diagonals of the system of equations. The compact support and limited overlap of the B-spline basis functions gives this system a banded structure, though exploiting this structure is of little utility since the number of equations is equal to the number of splines, which is generally moderate by design.

P-splines for single-index VC models

The derivations in the previous section requiring little adjustment for accommodating a regressor and its corresponding varying slope function, as defined in Equation 3.9 with $\mu(t) = Q\alpha$, where

$$Q = [B|diag\{x(t)\} B]$$

but now B holds a rich B-spline basis with equally-spaced knots. If one wishes to allow for differing degrees of smoothing for each of the varying intercept term and the slope function, the P-spline objective function 3.21 must be further modified to accommodate multiple tuning parameters, λ_i, $i = 0, 1$. The objective function then becomes

$$\begin{aligned} S_\lambda^* &= |y - Q\alpha|^2 + \lambda_0 |D_{d_0}\alpha_0|^2 + \lambda_1 |D_{d_1}\alpha_1|^2 \\ &= |y - Q\alpha|^2 + |\alpha^T P \alpha|^2 \end{aligned} \tag{3.23}$$

where the penalty has form $P = \text{block diag}\left(\lambda_0 D_{d_0}^T D_{d_0}, \lambda_1 D_{d_1}^T D_{d_1}\right)$. The minimizer of 3.23 is given by

$$\hat{\alpha} = \left(Q^T Q + P\right)^{-1} Q^T y.$$

The block diagonal structure of the penalty separates the penalization of each individual smooth component. The estimated mean function is then given by

$$\hat{\mu} = Q\hat{\alpha} = Hy$$

where

$$H = Q\left(Q^T Q + P\right)^{-1} Q^T. \tag{3.24}$$

[Figure 3.1 Need to explain figure 3 here.]

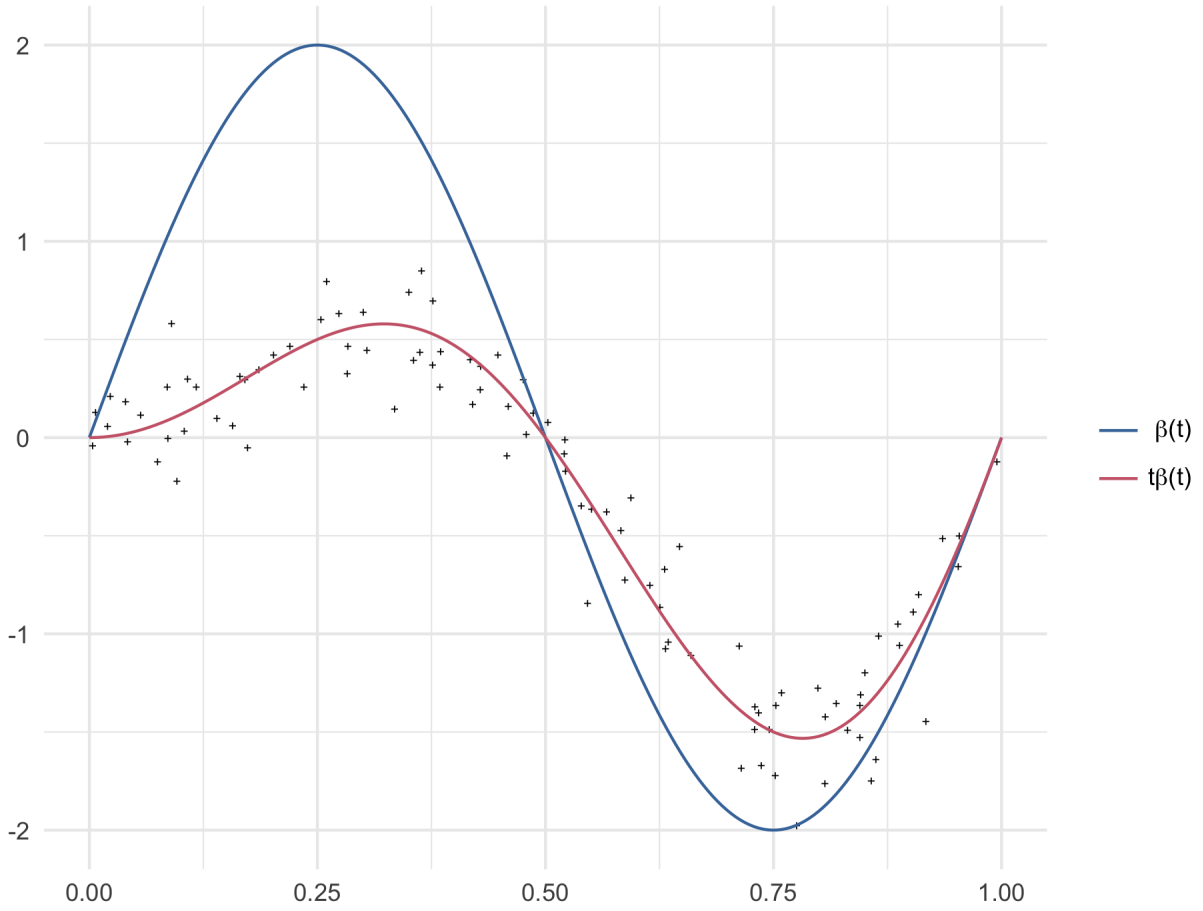
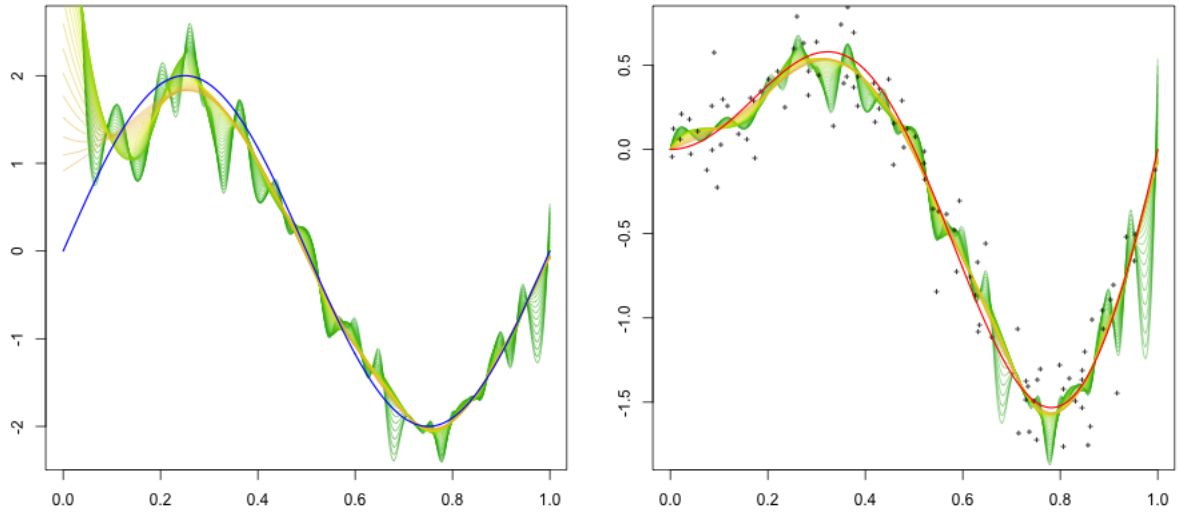


Figure 3.1: *100 simulated data points where* $y(t) = t\beta(t) + 0.2\epsilon(t)$ *where* ϵ *is a white noise process with unit variance, and* $\beta(t) = 2\sin(2\pi t)$.



(a) $\hat{\beta}(t)$ (b) $\hat{\mu}(t) = t\hat{\beta}(t)$

Figure 3.2: *Estimated coefficient function* $\hat{\beta}(t)$ *and mean curve* $\hat{\mu}(t) = t\sin(2\pi t)$ *using a 80 B-splines basis functions of order 5 and a difference penalty of order* $k = 3$.

The properties discussed in Section **??** allude to how controlling the coefficients of a spline $f \in \mathcal{S}_{k,t}$ influences the shape of the overall function. Specifically, the form of the j^{th} derivative provides an avenue of understanding how the differenced B-spline coefficient sequence is related to the volatility of the function on a given interval of its domain. The following figure visually explore the impact of the squared distance on adjacent basis coefficients on the function; a useful way of examining at P-splines is to consider the coefficients as the skeleton of the function, then draping the B-splines over them to put the flesh over the bones. A smoother sequence of coefficients leads to a smoother curve, which is clearly illustrated in Figure 3.3. As long as the coefficient sequence is smooth, the number of basis functions (and coefficients) is unimportant. The penalty ensures the smoothness of the skeleton.

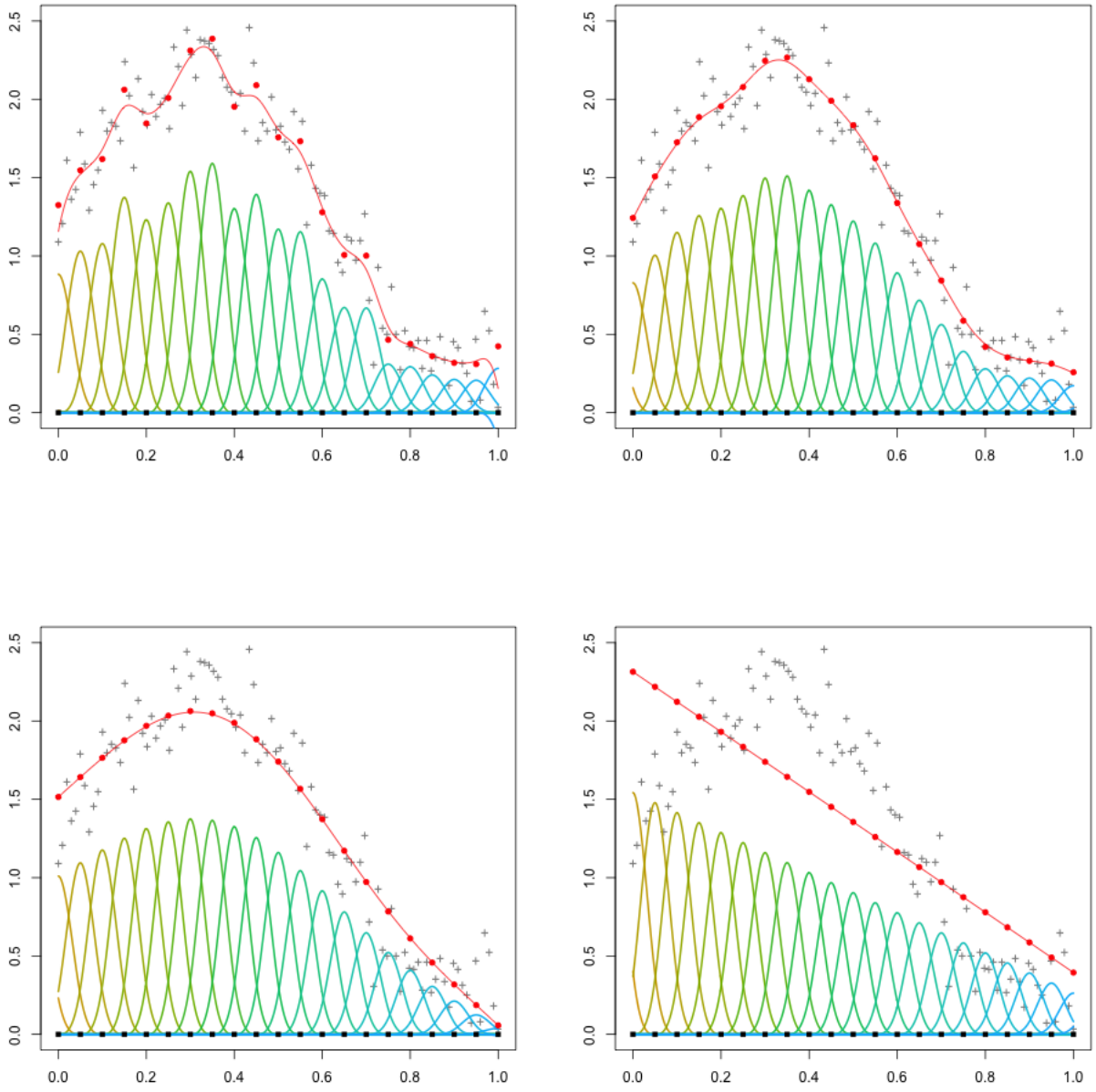


Figure 3.3: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The fitted curve in the upper left plot is the most "wiggly" 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.*

The number of B-splines can be much larger than the number of observations because penalty ensures that the fitting procedure well-conditioned. One could literally use a thousand splines to fit ten observations without problems. Figure **??** illustrates this utility of the penalty for simulated data. There are $m = 10$ observations and $40 + 3$ cubic B-splines. This property of P-splines cannot be overly appreciated, as it allows us to completely circumvent the nontrivial task of the optimal selection of knot placement. But one simply cannot have too many B-splines. Unless computational constraints are of concern, which is possible with large models, it is prudent to use even more. Figure **??** shows how the fitted function changes as the tuning parameter λ is varied in the presence of sparsely sampled data.

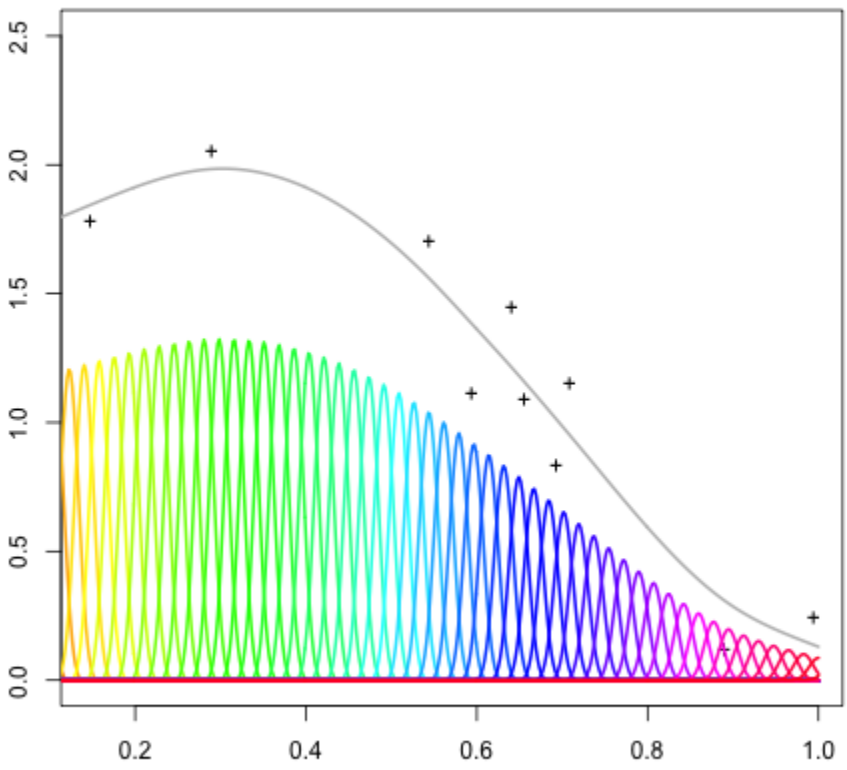


Figure 3.4: P-spline smoothing of 10 observations using 60 B-spline basis functions.

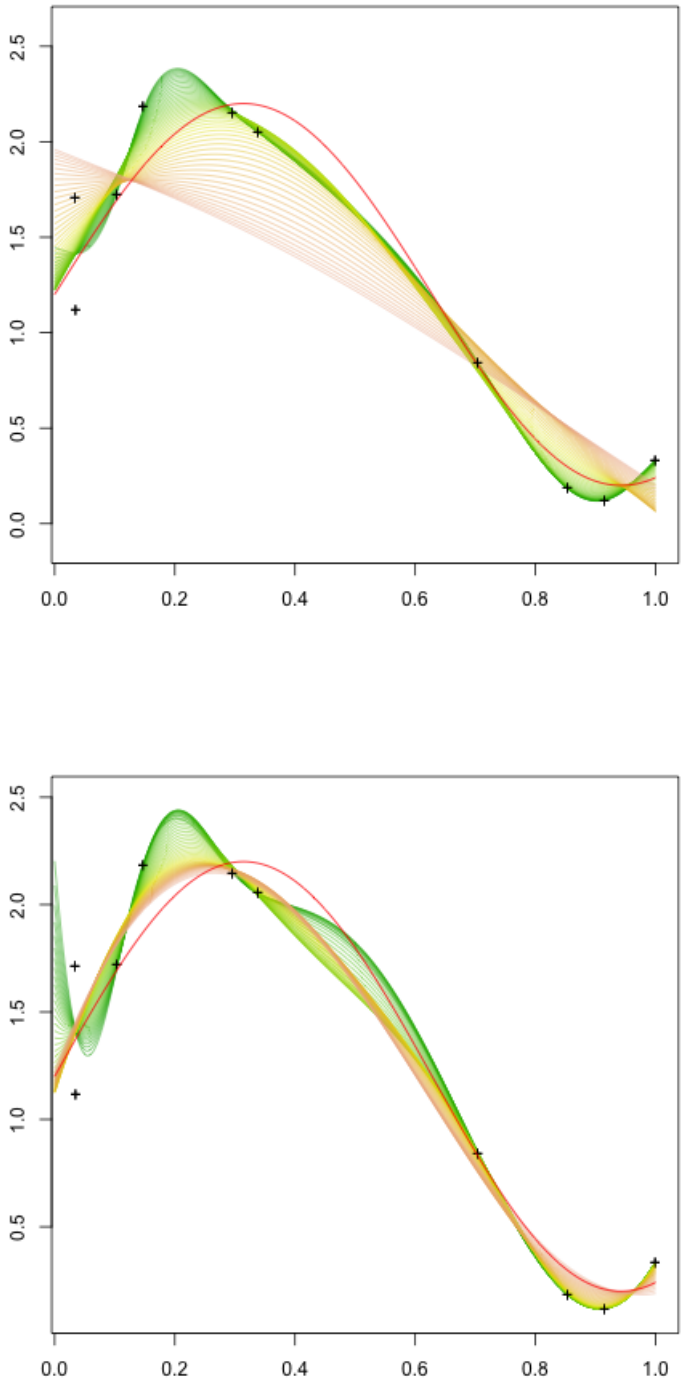


Figure 3.5: Fitted mean curves using a second (top) and third (bottom) order difference penalty for simulated data, sparsely sampled along the indexing variable: $y(t) = 1.2 + \sin(5t) + 0.2\epsilon_t$, where $\epsilon_t \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$. A total of 10 data points were fit using a basis of 60 B-splines of degree $k = 3$.

3.0.6 Properties of P-splines

P-splines exhibit a number of advantageous properties, many of which are due to the inherited properties of the B-spline basis functions.

I. **Boundary effects** P-splines show no boundary effects, as many types of kernel smoothers do. By this, we mean the spreading of a fitted curve or density outside of the (physical) domain of the data, generally accompanied by bending toward zero.

II. **P-splines fit polynomial data exactly.** P-splines can fit polynomial data exactly. Given data (t_i, y_i), if the y_i are a polynomial in t of degree k, then B-splines of degree k or higher will fit the data exactly.

Proof. This statement is equivalent to the claim that given $\xi = \{\xi_i\}$, $i = 1, \ldots, l+1$, and g such that $y(t) = g(t)$, we can find an $f \in \mathcal{P}_{k,\xi} \bigcap \mathcal{C}^{(k-2)}$ which agrees with g at the points $\tau_1 < \cdots < \tau_n$ with $\tau_i \in [\xi_1, \xi_{l+1}]$ for all i, where

$$n = k + l - 1$$

The solution, f is constructed as follows: generate the knot sequence $t = \{t_i\}$ as per the recipe in Theorem **??**:

$$\begin{aligned} t_1 = t_2 = \cdots = t_k &= \xi_1 \\ t_{k+i} &= \xi_{i+1}, \qquad i = 1, \ldots, l-1 \\ t_{n+1} = t_{n+2} = \cdots = t_{n+k} &= \xi_{l+1} \end{aligned}$$

Let $\{B_{ik}\}$, $i = 1, \ldots, n$ be the corresponding sequence of B-splines of order k, which are a basis for $\mathcal{P}_{k,\xi} \bigcap \mathcal{C}^{(k-2)}$ by Theorem **??**. Here, $\mathcal{P}_{k,\xi} \bigcap \mathcal{C}^{(k-2)}$ denotes the space of pp functions with breakpoints ξ having two continuous (global) derivatives. Then,

Schöenberg and Whitney [1953] have shown that there exists exactly one $f \in \mathcal{P}_{k,\xi} \bigcap \mathcal{C}^{(k-2)}$ agreeing with g at $\tau_1, \ldots, \tau_n$ if and only if

$$B_{ik}(\tau_i) \neq 0, \qquad i = 1, \ldots, n.$$

This f has a unique expansion of the form

$$f = \sum_{i=1}^{n} a_i B_{ik}$$

for coefficients $a_i, \ldots, a_n$, which are the solution to the linear system

$$\sum_{j=1}^{n} a_j B_{jk}(\tau_i) = g(\tau_i), \qquad i = 1, \ldots, n.$$

This system has a banded matrix of coefficients since $B_{jk}(\tau_i) \neq 0$ if and only if $\tau_i \in [t_j, t_{j+k}]$. So if $B_{jk}(\tau_i) \neq 0$ and thus $\tau_i \in (t_j, t_{j+k})$, then there are at most k of the j indices such that $B_{jk}(\tau_i)$ is nonzero. And further, each of these indices j must be such that

$$(t_i, t_{i+k}) \bigcap (t_j, t_{j+k}) \neq \emptyset,$$

or such that $|i - j| < k$. At worst, the system corresponds to a banded matrix with $k - 1$ lower and $k - 1$ upper diagonals. □

The same is true for P-splines if the order of the penalty is $k + 1$ or higher, irrespective of the value of λ. Consider imposing a first-order difference penalty and a fit to data y that is constant - a polynomial of degree 0. Since

$$\sum_{j=1}^{n} \hat{\alpha}_j B_j(x_i) = c,$$

we have that

$$\sum_{j=1}^{n} \hat{\alpha}_j B_j'(x) = 0,$$

for all x. From the relationship between differences and derivatives in **?? ??**,

$$0 = \sum_{j=1}^{n} B'_{j,k}(x) = \sum_{j=1}^{n} \Delta\alpha_{j+1} B_{j,k-1}(x),$$

so that we must have $\Delta\alpha_j = 0$ for all j, and

$$\sum_{j=2}^{n} \Delta\alpha_j = 0.$$

This shows that the penalty has no impact on the basis coefficients, and the resulting fit is identical to that when using unpenalized B-splines. Using induction, one can show that this is also true when the relationship between x and y is linear and a second order difference penalty is used, and for any values of the polynomial order and order of the difference penalty.

III. **Null models under difference penalties** The limiting P-spline fit approaches a polynomial under strongly enforced smoothing. As $\lambda \to \infty$, 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. To see this, we again need to use the relationship between the differenced coefficient sequence and the derivative of a B-spline as described in **?? ??**. Consider using the second-order difference penalty; when λ is large, the penalty dominates the P-spline objective function defined in 3.21, so that the minimizer α must be such that $\sum_{j=3}^{n} \left(\Delta^2\alpha_j\right)^2$ 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.

IV. **The limiting behaviour of H_λ** The trace of the hat matrix,

$$H_\lambda = B\left(B^T B + \lambda D_k^T D_k\right)^{-1} B^T y$$

(or for H defined for the addition of a varying slope component as in 3.24) approaches k, the order of the differencing operator, as λ increases. We index H with the smoothing parameter to indicate that the elements of H are a function of λ. Let

$$Q_B = B^T B \qquad \text{and} \qquad Q_\lambda = \lambda D^T D. \tag{3.25}$$

Then using properties of the matrix trace, we can write

$$\begin{aligned} \text{tr}\left(H_\lambda\right) &= \text{tr}\left[\left(Q_B + Q_\lambda\right)^{-1} Q_B\right] \\ &= \text{tr}\left[Q_B^{1/2}\left(Q_B + Q_\lambda\right)^{-1} Q_B^{1/2}\right] \\ &= \text{tr}\left[\left(I + Q_B^{-1/2} Q_\lambda Q_B^{-1/2}\right)^{-1}\right] \end{aligned} \tag{3.26}$$

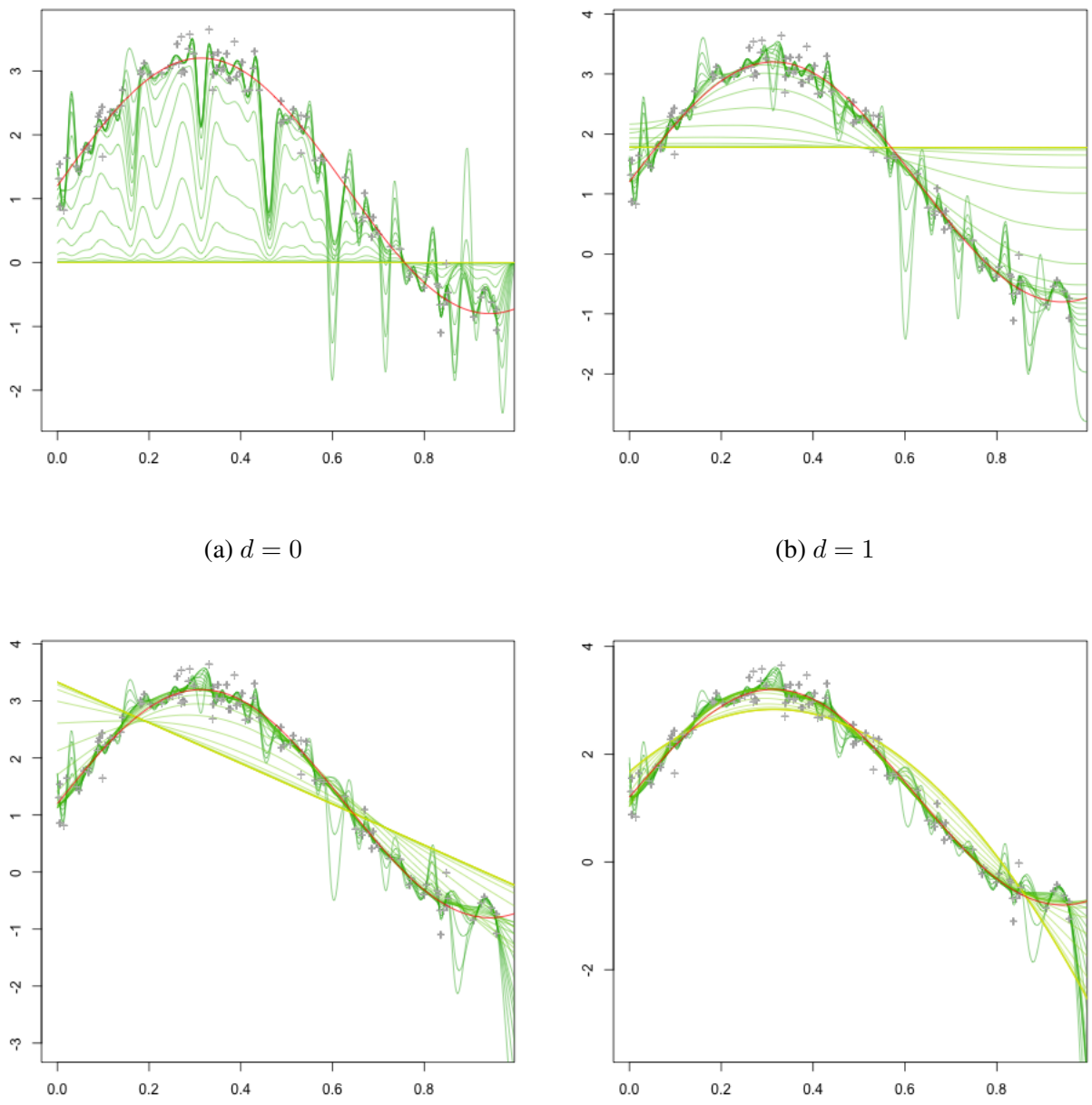
Define $L \equiv Q_B^{-1/2} Q_\lambda Q_B^{-1/2}$. Then

$$\text{tr}\left(H_\lambda\right) = \text{tr}\left[\left(I + \lambda L\right)^{-1}\right] = \sum_{j=1}^{n} \frac{1}{1 + \lambda\gamma_j} \tag{3.27}$$

where γ_j, $j = 1, \ldots, n$ are the eigenvalues of L. Q_λ has exactly k eigenvalues equal to zero, hence L has k zero eigenvalues. For large λ, only the k terms with $\gamma_j = 0$ contribute to the sum which gives the trace of H, so that

$$\lim_{\lambda \to \infty} \text{tr}\left(H\right) = k.$$

The previous derivations hold regardless of whether we are fitting the varying intercept-only model, with $\mu\left(t\right) = \beta_0\left(t\right)$ or accommodating a varying slope for a regressor by specifying $\mu\left(t\right) = \beta_0\left(t\right) + \beta_1\left(t\right) x\left(t\right)$. The inspection of the hat matrix H is a prelude to the following section, where we will discuss how to use the properties of H to tune the smoothing parameter for optimal model selection. We will later show that extension of these results can be extended in a rather straightforward manner to the case that is of our particular interest: when the smooth slope function is a two-dimensional surface rather than a curve.



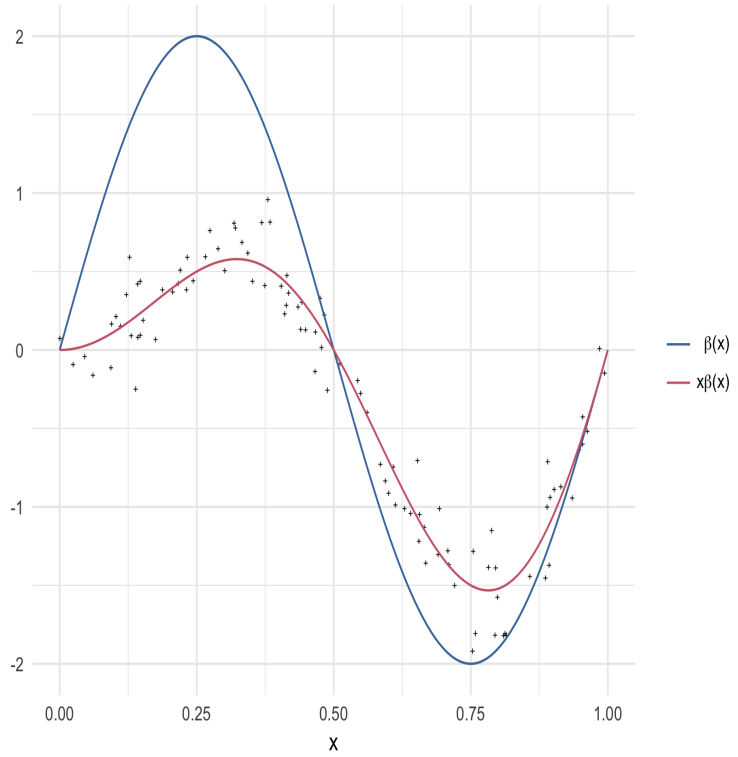
(a) $d = 0$

(b) $d = 1$

(c) $d = 2$

(d) $d = 3$

Figure 3.6: *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* $|D_0\alpha|^2 = \sum_{i=1}^{n} \alpha_i^2$*, analogous to ridge regression using the B-spline basis as regression covariates. The fitted curves approach polynomials of degree* $d - 1$ *as* $\lambda \to \infty$*, as discussed in 3.0.6 III.*



3.0.7 The reguarlized MLE for ϕ via tensor product P-splines

To extend the P-spline framework to allow estimation of a bivariate function, ϕ, we simply need to equip the l and m axes each with a B-spline basis. A basis for the varying coefficient function is constructed taking the tensor product of the two marginal bases. Let

$$B_1(l), \ldots, B_K(l) \text{ and } B_1(m), \ldots, B_L(m)$$

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 while we have chosen not to distinguish between $\{B_k\}$ and $\{B_l\}$ for the sake of brevity, 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, and hence entirely different knot sequences since P-splines rely on bases with equally spaced knots. The tensor

product basis functions

$$T_{jk}(l,m) = B_j(l)\, B_k(m)$$

carve the l-m domain into rectangles. Figure 3.8 shows a thinned tensor product basis $\{T_{kl}\}$; 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 3.8 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 by

$$\phi(l,m) = \sum_{i=1}^{K}\sum_{j=1}^{L} \theta_{ij} B_i(l)\, B_j(m), \tag{3.28}$$

and the function evaluated at the observed (l_{ijk}, m_{ijk}) may be written

$$\boldsymbol{\phi} = B_m \Theta B_l'$$

where Θ denotes the $K \times L$ matrix of tensor product coefficients, with elements θ_{ij}.

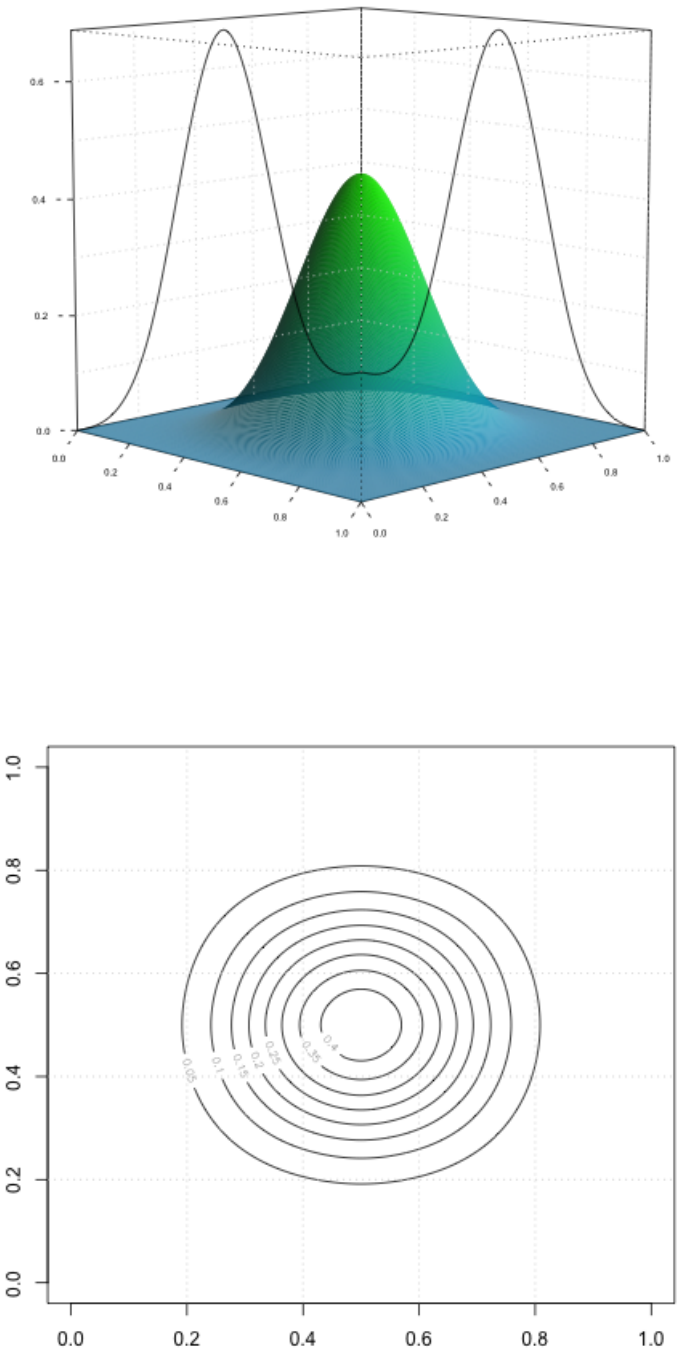


Figure 3.7: Tensor product of two cubic B-splines

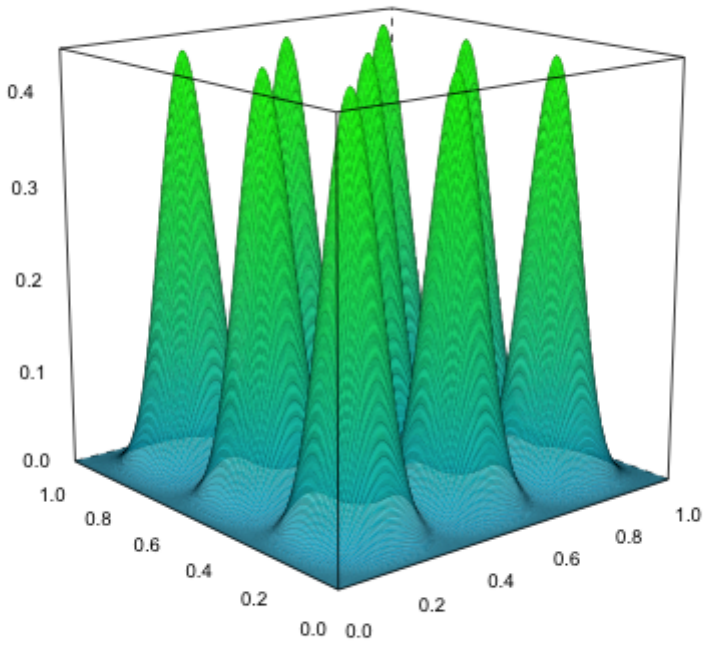


Figure 3.8: A subset of a full bivariate basis of cubic B-splines

3.0.8 Regularization with difference penalties

The minimizer of **??** honors the fidelity to the data, so to balance the complexity of the fitted function with the goodness of fit to the data, we can append a penalty to the negative log likelihood to control the fitted function. 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. O'Sullivan [1986] was the first to propose using a rich B-spline basis and using a penalty to restrict the flexibility of the fitted curve, like Wahba [1990]

applying a penalty to the second derivative of the fitted curve:

$$J = \int_0^1 \left[f''(l)\right]^2 \, dx.$$

For a B-spline of the form

$$f(x) = \sum_{j=1}^{n} \theta_i B_j(x),$$

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

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

where $\theta = (\theta_1, \ldots, \theta_n)$. The i-j^{th} element of P is given by

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

In some applications, it is useful to work with third and fourth order differences, since for large values of λ, the fitted curve approaches a parametric polynomial model. This may be of particular interest in the context of estimating the elements of the Cholesky factor, as many have proposed simple parametric functions of lag only for ϕ, such as low order polynomials. See Pourahmadi [1999]. However, with the use of higher order derivatives, the computation of P is nontrivial and becomes very tedious. Eilers and Marx [1996] were the first to propose P-splines, or *penalized B-splines*, as an approach to nonparametric regression. P-splines circumvent complexity associated with constructing such penalty matrices by omitting derivatives and integrals altogether, replacing them with finite differences and sums.

Instead, flexibility of the fitted function is controlled by using a discrete penalty matrix based on finite difference formulas. Smoothness of the fitted function is achieved by first using a rich B-spline basis with equally spaced knots to purposefully overfit the smooth coefficient vectors; this eliminates the difficulty of choosing the optimal set of knots. Then by attaching a difference penalty to the goodness of fit measure, one may prevent overfitting and make a potentially ill-conditioned

fitting procedure a well-conditioned one. The finite difference penalty is simple to compute and can be handled mechanically for any order of the differences. Since it is easily introduced into regression equations, it is feasible to evaluate the impact of different orders of the differences on the fitted model. 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 replacing the derivative-based penalty with

$$J_d(f) = \sum_{j=d}^{n} \left(\Delta^d \theta_j\right)^2 \tag{3.29}$$

where $\theta = (\theta_1, \ldots, \theta_n)$. Let D_d denote the matrix difference operator: $D_d\theta = \Delta^d\theta$, where

$$\Delta\theta_j = \theta_j - \theta_{j-1}, \quad \Delta^2\theta_j = \Delta\left(\Delta\theta_j\right) = \theta_j - 2\theta_{j-1} + \theta_{j-2}$$

In general,

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

Then, 3.29 can be written in terms of the squared norm of the difference operator applied to the vector of B-spline coefficients:

$$\begin{aligned} J_d(f) &= ||D_d\theta||^2 \\ &= \theta' P_d \theta \end{aligned} \tag{3.30}$$

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:

$$\int_0^1 \left[f''(x)\right]^2 \, dx = \int_0^1 \left\{\sum_{j=1}^{n} \theta_j B''_{j,3}(l)\right\}^2 \, dl.$$

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

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

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

$$\begin{aligned}
\int_0^1 \left[f''(x)\right]^2 \, dx = & \int_0^1 \left[\left\{ \sum_{j=1}^{n} \Delta^2\theta_j B_j(l,1) \right\}^2 + 2 \sum_{j} \Delta^2\theta_j \Delta^2\theta_{j-1} B_j(l,1) \, B_{j-1}(l,1) \right] dl \\
= & \sum_{j=1}^{n} \left(\Delta^2\theta_j\right)^2 \int_0^1 B_j^2(l,1) \, dl \\
& + 2 \sum_{j=1}^{n} \Delta^2\theta_j \Delta^2\theta_{j-1} \int_0^1 B_j(l,1) \, B_{j-1}(l,1) \, dl
\end{aligned} \tag{3.31}$$

which can be written as

$$\int_0^1 \left[f''(x)\right]^2 \, dx = c_1 \sum_{j=2}^{n} \left(\Delta^2\theta_j\right)^2 + c_2 \sum_{j=3}^{n} \Delta^2\theta_j \Delta^2\theta_{j-1} \tag{3.32}$$

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

$$\begin{aligned}
c_1 &= \int_0^1 B_{j,1}^2(x) \, dx \\
c_2 &= \int_0^1 B_{j,1}(x) \, B_{j-1,1}(x) \, dx.
\end{aligned} \tag{3.33}$$

This gives us that the 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 3.29 and the sum of the cross products of neighboring second differences. The second term in 3.32 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 3.32 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 3.3. 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. 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 **??** illustrates this utility of the penalty for simulated data; there are $m = 10$ observations and 60 cubic B-splines. This property of P-splines cannot be overly appreciated because it frees us from the concern of choosing the optimal set of knots. Unless computational constraints are of concern, which is possible with large models, it is prudent to use even more B-splines. Figure **??** shows how the fitted function changes as the tuning parameter varies when the data are sparsely sampled. P-splines enjoy a number of additional advantageous properties, many of which are inherited from the attractive properties of B-splines. See Eilers and Marx [1996] for a detailed presentation.

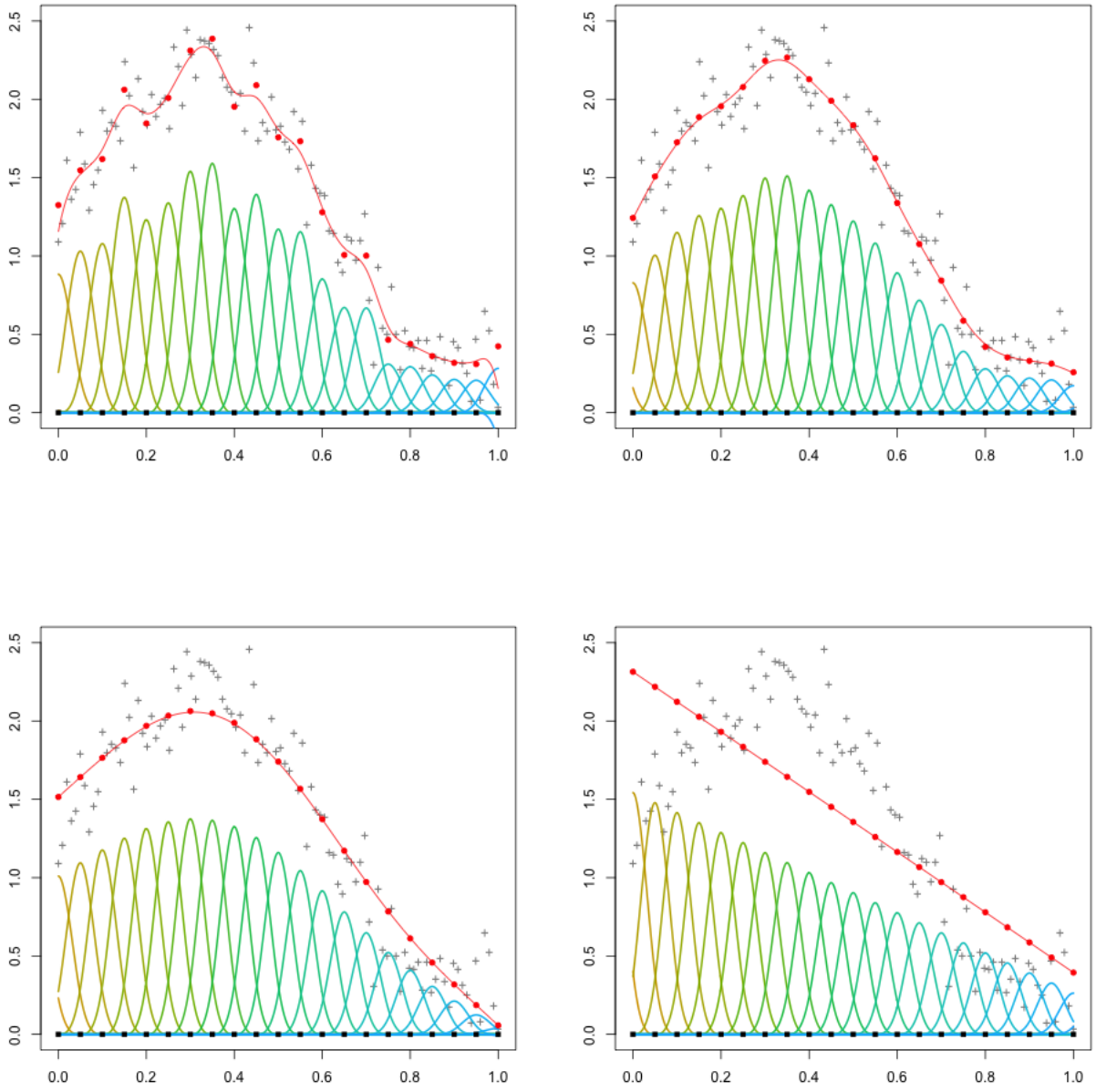


Figure 3.9: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The fitted curve in the upper left plot is the most "wiggly" 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.*

3.0.9 Model selection and tuning parameter estimation

The limiting behaviour of H_λ

The inspection of the hat matrix

$$H_\lambda = WB\left(WB'WB + \lambda_l P_l + \lambda_m P_m\right)^{-1}(WB)' D^{-1}.$$

and its properties are integral for assessing model complexity and selecting the optimal values of the tuning parameters λ_l and λ_m. Summarizing the complexity of a fitted P-spline is far from a trivial task; one must simultaneously consider the value of the smoothing parameter, the number of basis functions in the B-spline basis, as well as the order of the difference penalties. We follow Eilers and Marx [1996] andMarx and Eilers [2005] assess model complexity as discussed in citehastie1990generalized, who proposed to use the trace of the smoother matrix as an approximation to the effective dimensions of linear smoother. The *effective dimension* is easily obtained and combines the effect of all three of these elements:

$$\begin{aligned} \text{ED} &= \text{tr}\left[H_\lambda\right] \\ &= \text{tr}\left[\left[WB\left(WB\right)' D^{-1} WB + \lambda_l P_l + \lambda_m P_m\right]^{-1}(WB)' D^{-1}\right] \end{aligned} \tag{3.34}$$

When the number of basis functions is significantly smaller than the sample size, it is computationally advantageous to use the cyclic property of the trace:

$$\text{tr}\left[\left[(WB)' D^{-1} WB + \lambda_l P_l + \lambda_m P_m\right]^{-1}(WB)' D^{-1} WB\right],$$

which requires computing the trace of a $KL \times KL$ matrix. The effective dimension approaches $d_l + d_m$, the order of the differencing operator, as λ increases, where d_l and d_m denote the orders of the difference penalties in the l and m directions, respectively. Let

$$Q = (WB)' D^{-1} WB \qquad \text{and} \qquad Q_\lambda = P.$$

Using properties of the matrix trace, we can write

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

Define $L \equiv Q^{-1/2}Q_\lambda Q^{-1/2}$. Then

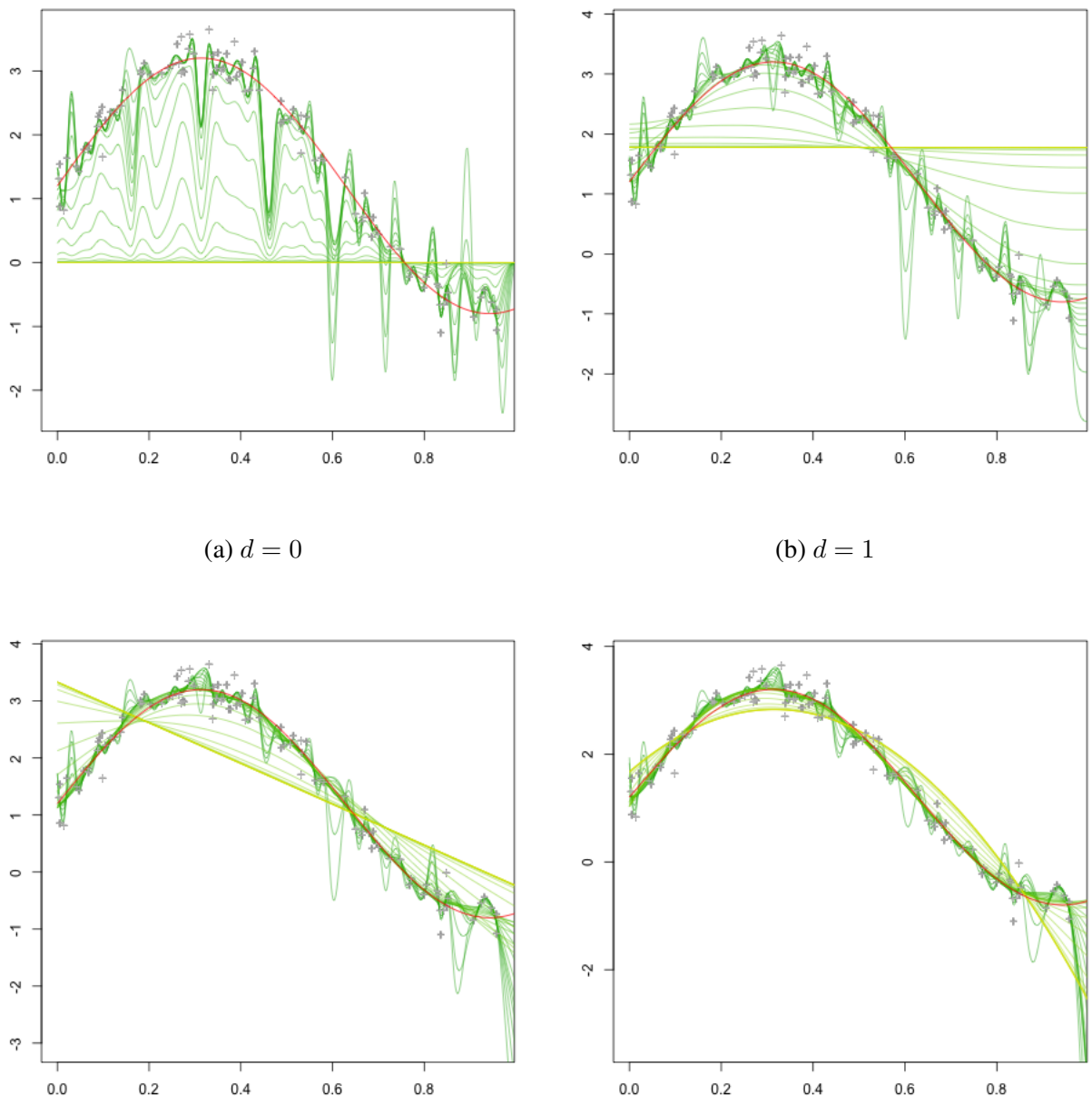
$$\text{tr}\left(H_\lambda\right) = \text{tr}\left[\left(I+\lambda L\right)^{-1}\right] = \sum_{j=1}^{n}\frac{1}{1+\lambda\gamma_j}$$

where γ_j, $j = 1, \ldots, n$ are the eigenvalues of L. Q_λ has exactly $d_l + d_m$ eigenvalues equal to zero. Hence, L has $d_l + d_m$ zero eigenvalues. For large λ, only the $d_l + d_m$ terms with $\gamma_j = 0$ contribute to the sum which gives the trace of H, so that

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

Equation **??** cleanly shows that the effective dimension is always less than n, the number of B-spline used in the regression basis; further, the effective dimension is always smaller than $\min\left(m, n\right)$. A formal proof follows below. This is illustrated in

Figure **??** shows how the effective dimension on a univariate P-spline changes with the smoothing parameter for the ten simulated observations in Figure **??** using 60 B-spline basis functions. For small λ, the effective dimension approaches m. As λ increases, the effective dimension approaches the order of the difference penalty, d. It is worth pointing out here that there are no problems incurred when smoothing with many more B-splines than observations since the effective model dimension is always less than m, for all λ.



(a) $d = 0$

(b) $d = 1$

(c) $d = 2$

(d) $d = 3$

Figure 3.10: *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* $|D_0\alpha|^2 = \sum_{i=1}^{n} \alpha_i^2$*, analogous to ridge regression using the B-spline basis as regression covariates. The fitted curves approach polynomials of degree* $d - 1$ *as* $\lambda \to \infty$*, as discussed in 3.0.6 III.*

Chapter 4: Simulation studies

4.1 Performance benchmarking with complete data

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 4.1. 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 4.1.

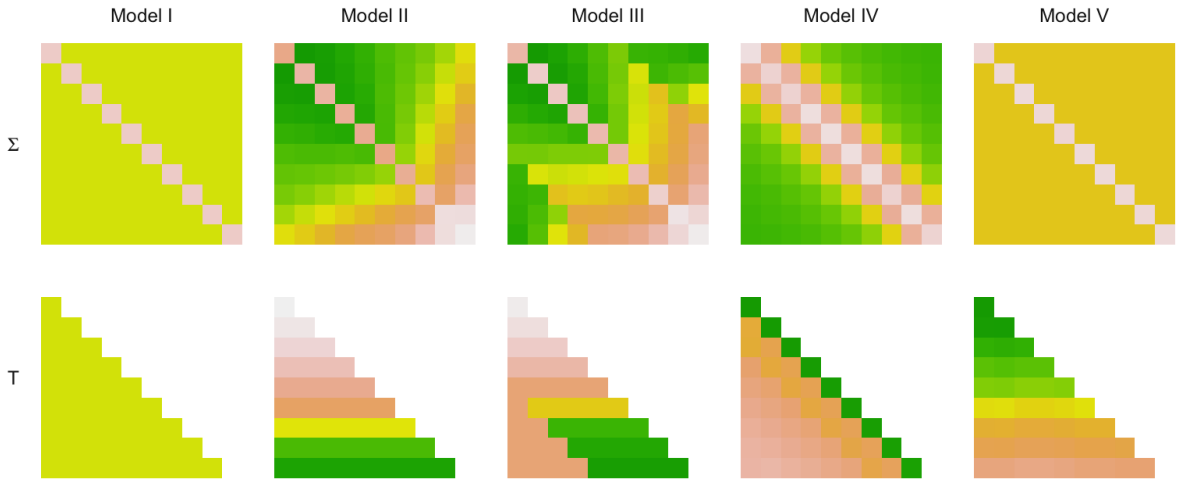


Figure 4.1: *Heatmaps of the true covariance matrices (row 1) under simulation Model I - Model V and their corresponding Cholesky factor T (row 2). The*

Connecting the covariance matrices in the first row of Figure 4.1 with their Cholesky factor in the second row, 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.

For each of the general covariance structures outlined in the simulation study description, data were generated according to multivariate normal distributions with the following covariance matrices:

Table 4.1: *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.*

I: Mutual independence	$\Sigma = \mathrm{I}$	$\phi(t,s) = 0,$ $\sigma^2(t),$
II: Linear varying coefficient function, constant innovation variances	$\Sigma = T^{-1}DT'^{-1}$	$\phi(t,s) = t -$ $\sigma^2(t) = 0.1^2$
III: Banded linear varying coefficient function, constant innovation variances	$\Sigma = T^{-1}DT'^{-1}$	$\phi(t,s) = \{$ $\sigma^2(t) = 0.1^2$
IV:		
V:		

Table 4.2: *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.*

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	
$\Sigma = \sigma^2(\rho \mathrm{J} + (1-\rho)\mathrm{I}),\ \rho = 0.7,\ \sigma^2 = 1$	$\phi_{ts} = -\frac{\rho}{1 + (t-1)\rho}, \quad t = 2, \ldots, M, \quad s = 1, \ldots, t-1$ $\sigma_t^2 = \begin{cases} 1, & t = 1 \\ 1 - \frac{(t-1)\rho^2}{1+(t-1)\rho}, & t = 2, \ldots, M \end{cases}$

I. Mutual independence: $\Sigma = \mathrm{I}$, where

$$\phi(t,s) = 0, \quad 0 \le s < t \le 1,$$
$$\sigma^2(t) = 1, \quad 0 \le t \le 1.$$

II. Linear varying coefficient model with constant innovation variance: $\Sigma^{-1} = T'D^{-1}T$, where

$$\phi(t,s) = t - \frac{1}{2}, \quad 0 \le t \le 1,$$
$$\sigma^2(t) = 0.1^2, \quad 0 \le t \le 1.$$

III. $k_{1/2}$-banded linear varying coefficient model with constant innovation variance: $\Sigma^{-1} = T'D^{-1}T$, where

$$\phi(t,s) = \begin{cases} t - \frac{1}{2}, & t - s \le 0.5 \\ 0, & t - s > 0.5 \end{cases},$$
$$\sigma^2(t) = 0.1^2, \quad 0 \le t \le 1.$$

IV. Rational quadratic covariance: $\Sigma = (\sigma_{ij})$ where

$$Cov(y(t_i), y(t_j)) = \left(1 + \frac{(t_i - t_j)^2}{2\alpha k^2}\right)^{-\alpha}, \tag{4.1}$$

with $k = 0.6$ and $\alpha = 1$.

V. The compound symmetry model: $\Sigma = \sigma^2(\rho \mathrm{J} + (1-\rho)\mathrm{I})$, $\rho = 0.7$, $\sigma^2 = 1$.

$$\phi_{ts} = -\frac{\rho}{1 + (t-1)\rho}, \quad t = 2, \ldots, M, \;\; s = 1, \ldots, t-1$$
$$\sigma_t^2 = \begin{cases} 1, & t = 1 \\ 1 - \frac{(t-1)\rho^2}{1+(t-1)\rho}, & t = 2, \ldots, M \end{cases}$$

For each of the covariance models, we generated a set of observations of sample size $N = 50, 100$ from a multivariate normal distribution, and considered three different values of within-subject sample size $M = 10, 20, 30$.

The estimators were computed with tuning parameters 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. 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 $\Sigma = T^{-1}DT'^{-1}$.

4.1.1 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:

$$\Delta_1\left(\Sigma, \hat{\Sigma}\right) = tr\left(\left(\Sigma^{-1}\hat{\Sigma} - \mathrm{I}\right)^2\right), \tag{4.2}$$

$$\Delta_2\left(\Sigma, \hat{\Sigma}\right) = tr\left(\Sigma^{-1}\hat{\Sigma}\right) - log|\Sigma^{-1}\hat{\Sigma}| - M. \tag{4.3}$$

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 corresponding to the two 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.

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 1.2.3 and 1.3.

TO DO: descriptions of the oracle estimators

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

$$\begin{aligned} z'_{jk} &= \left(1, t_j - t_k, (t_j - t_k)^2, \ldots, (t_j - t_k)^{p-1}\right)', \\ z'_j &= \left(1, t_j, \ldots, t_j^{q-1}\right)'. \end{aligned} \tag{4.4}$$

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

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

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

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

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, these simulations were conducted using complete data with common measurement times across all N subjects.

Figure 4.2 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 4.3, and the decomposition of the $\hat{T}$ corresponding to the smoothing spline ANOVA estimator $\hat{\Sigma}_{SS}$ into functional components is displayed in Figure 4.4

Figure 4.2: *Covariance Model I - Model V 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.*

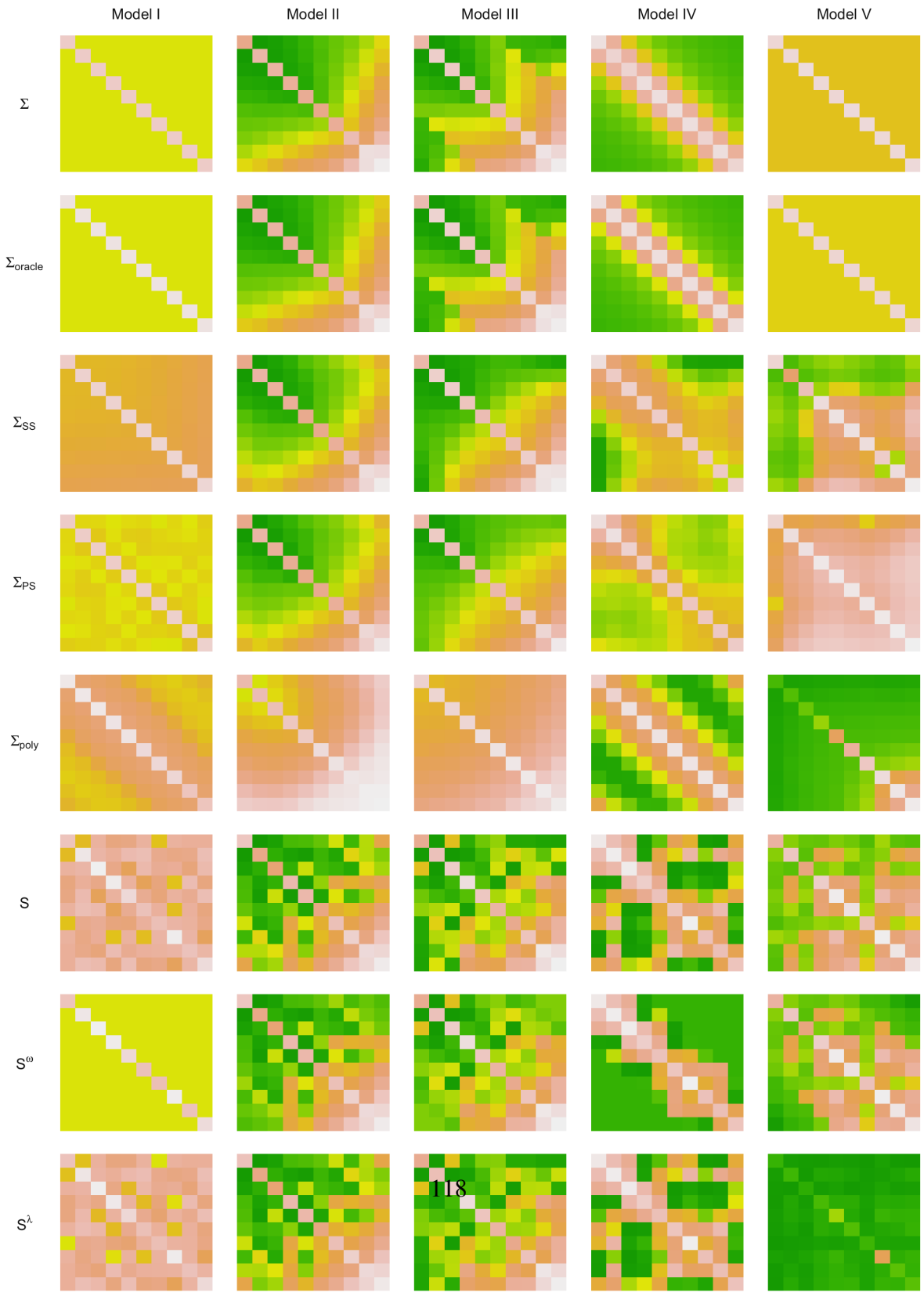


Figure 4.3: *The true lower triangle of Cholesky factor T corresponding to Model I - Model V and estimates of the same surface for estimators based on the modified Cholesky decomposition. The true covariance structure is displayed across the top row.*

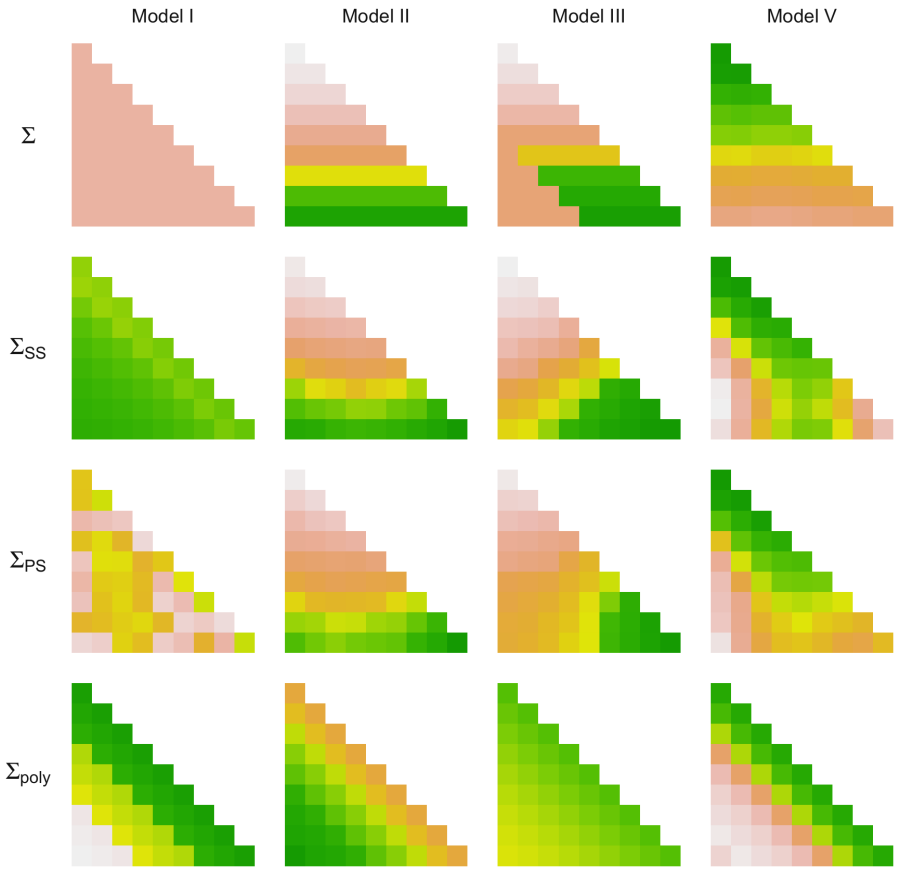
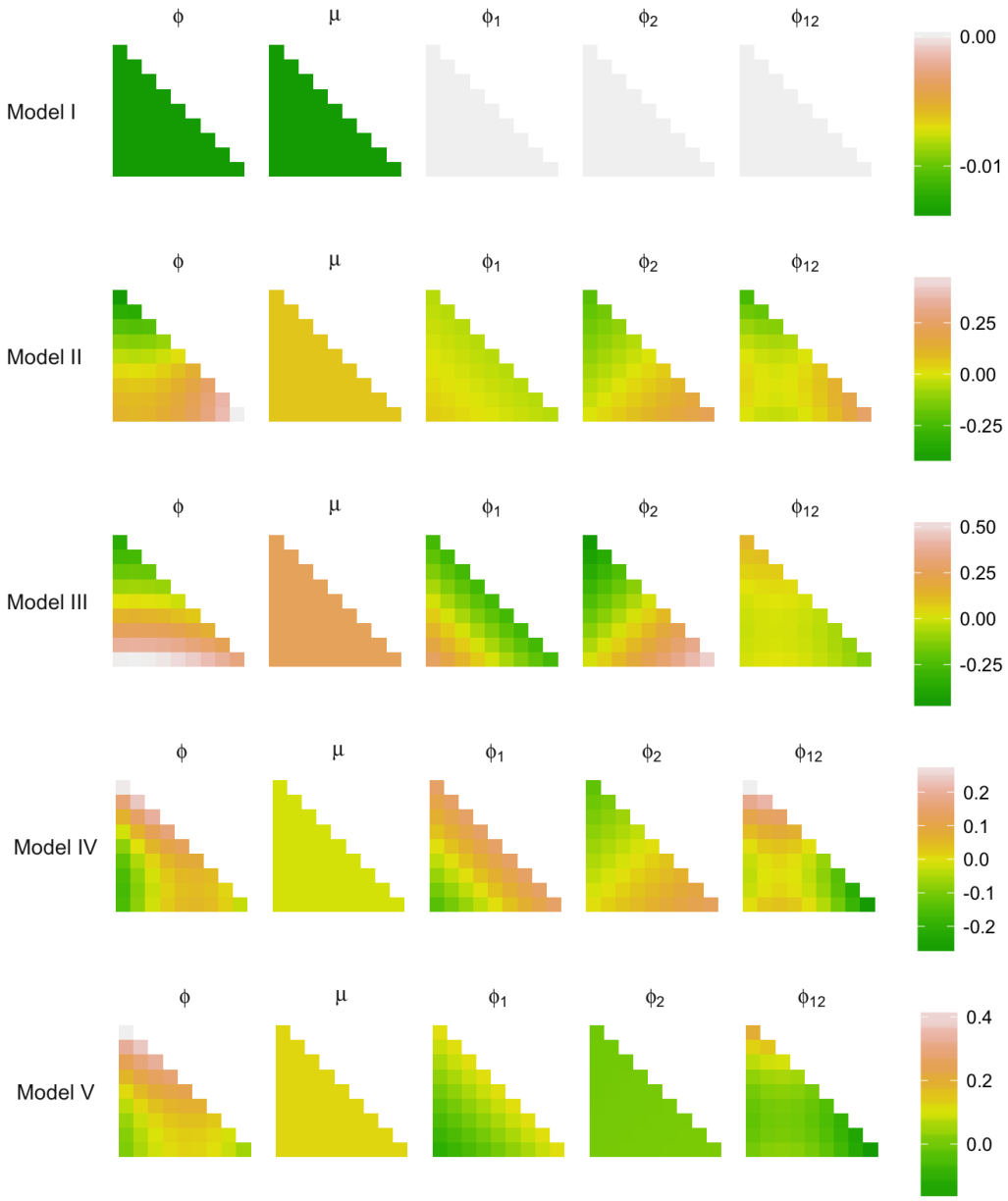


Figure 4.4: *Estimated functional components of the smoothing spline ANOVA decomposition* $\phi = \phi_1 + \phi_2 + \phi_{12}$ *for* $\hat{\Sigma}_{SS}$ *under each simulation model I - V.*



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. The results of the simulations for complete data under entropy loss are presented in Tables 4.3 - 4.7. 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. These are left to the Appendix, Tables A.1-A.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. The P-spline estimator outperforms the smoothing spline estimator under model IV, likely due to the advantage of trivially change the order of the difference penalty. When the difference order is specified so that the generating model belongs to the null space $\mathcal{H}_0$, the search for the optimal set of smoothing parameters is a much easier task.

Table 4.3: *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	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0749	0.1261	0.0135	0.1102	1.2047	0.5369	1.1742
	20	0.0872	0.1713	0.0229	0.1096	4.9850	1.3957	4.7796
	30	0.1102	0.1969	0.0196	0.1127	12.5517	2.8019	11.3175
$N = 100$	10	0.0451	0.0671	0.0105	0.0531	0.5685	0.2045	0.5236
	20	0.0425	0.0965	0.0105	0.0512	2.2831	0.5724	2.1358
	30	0.0431	0.1148	0.0139	0.0472	5.2770	1.2430	4.9126

Table 4.4: *Multivariate normal simulations for model II.*

	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0899	0.3423	0.0581	4.7673	1.2832	1.4644	1.1770
	20	0.0949	1.3640	0.0439	97.2334	5.1665	21.6407	39.3522
	30	0.0811	2.6485	0.0627	1539.6646	12.3582	55.3674	133.9980
$N = 100$	10	0.0457	0.2945	0.0386	4.7911	0.5812	0.8335	0.5628
	20	0.0416	1.2875	0.0269	98.1989	2.3364	10.1841	10.0864
	30	0.0367	2.4365	0.0288	1582.4795	5.2389	33.5207	62.5030

Table 4.5: *Multivariate normal simulations for model III.*

	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.3416	0.1065	0.0619	3.0108	1.2030	1.1460	1.1467
	20	1.1140	0.2555	0.0695	62.7522	4.9824	17.2244	14.9189
	30	2.3215	0.6242	0.0576	1091.1933	12.4792	49.9135	121.7795
$N = 100$	10	0.2904	0.0579	0.0268	3.0383	0.5699	0.5545	0.5371
	20	1.1963	0.2011	0.0275	62.8960	2.2700	11.8274	9.5217
	30	2.2811	0.3845	0.0221	1105.0449	5.2234	29.1693	60.3529

Table 4.6: *Multivariate normal simulations for model IV.*

	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.3422	0.1966	0.0217	0.7144	1.2218	0.7397	1.1921
	20	0.9208	0.3499	0.0286	1.4588	4.9091	1.9786	4.9206
	30	1.5992	0.5100	0.0283	2.2173	12.6114	3.7440	12.1489
$N = 100$	10	0.3047	0.2237	0.0125	0.6958	0.5570	0.3168	0.5515
	20	0.8911	0.3704	0.0105	1.4813	2.2659	0.9365	2.2474
	30	1.5213	0.5282	0.0134	2.2228	5.2106	1.9312	5.2111

Table 4.7: *Multivariate normal simulations for model V.*

	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.2743	0.2464	0.0986	1.2420	1.2023	18.5222	2.9824
	20	0.7526	0.8772	0.2512	2.8557	5.0195	34.6618	13.8690
	30	1.1776	0.9791	0.2641	4.5791	12.3460	46.5437	26.1364
$N = 100$	10	0.2416	0.1722	0.0520	1.1491	0.5821	16.4081	1.7397
	20	0.7286	0.2965	0.0827	2.9080	2.2918	32.5295	5.4649
	30	1.1813	0.4291	0.1799	4.4402	5.2197	39.2914	15.4295

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

$$\mathbf{CV}_F(\lambda) = \arg\min_{\lambda} K^{-1} \sum_{k=1}^{K} ||\hat{\Sigma}^{(-k)} - \tilde{\Sigma}^{(k)}||_F^2, \tag{4.5}$$

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 4.3 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 4.1.1. 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].

4.1.2 Performance with irregularly sampled data

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

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

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 4.2. 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 0.05, 0.07, and 0.09.

Estimated risk under entropy loss is given in Tables 4.8 - 4.12. Risk estimates under quadratic loss echo in sentiment and are left to the Appendix. 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 the identity matrix and for the non-truncated linear varying coefficient GARP model, we observe little change in estimated entropy risk for within subject sample sizes $M = 10$ and $M = 20$ with downsampling as compared to the estimated risk for both sample sizes in the complete data case. Neither model selection criteria appear to perform better than the other, which suggests 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. When the data are missing

completely at random, the performance of the estimator remains fairly stable when 5%, 7%, and 9% of the data are missing.

Table 4.8: *Model 1: Entropy risk estimates and corresponding standard errors for the MCD smoothing spline ANOVA estimator via 100 simulated multivariate normal sample of size $N = 50$ when 5%, 7%, and 9% of the data are missing. Risk is reported for the estimator constructed using the unbiased risk estimate and leave-one-subject-out cross validation are used for smoothing parameter selection.*

M	% missing	$\Delta_2(\hat{\Sigma}_{SS}^{U})$		$\Delta_2(\hat{\Sigma}_{SS}^{V^*})$	
10	0.00	0.05948765	(0.0040)	0.07355005	(0.0053)
	0.05	0.07526588	(0.0064)	0.06118984	(0.0044)
	0.07	0.07013988	(0.0047)	0.07533206	(0.0060)
	0.09	0.08318940	(0.0056)	0.06423959	(0.0051)
15	0.00	0.08753938	(0.0073)	0.09178231	(0.0073)
	0.05	0.08081929	(0.0052)	0.09203281	(0.0063)
	0.07	0.08584911	(0.0091)	0.08708120	(0.0072)
	0.09	0.09838363	(0.0086)	0.08148412	(0.0072)
20	0.00	0.08742261	(0.0077)	0.08676581	(0.0066)
	0.05	0.08524652	(0.0069)	0.07007607	(0.0056)
	0.07	0.07940546	(0.0061)	0.07639569	(0.0062)
	0.09	0.07629685	(0.0050)	0.08846639	(0.0063)

Table 4.9: *Model 2: Entropy risk estimates and corresponding standard errors.*

M	% missing	$\Delta_2(\hat{\Sigma}_{SS}^U)$		$\Delta_2(\hat{\Sigma}_{SS}^{V^*})$	
10	0.00	0.3476005	(0.0080)	0.3216708	(0.0069)
	0.05	0.3232649	(0.0067)	0.3505707	(0.0114)
	0.07	0.3225538	(0.0063)	0.3300389	(0.0067)
	0.09	0.3347380	(0.0072)	0.3303566	(0.0056)
15	0.00	0.6365878	(0.0095)	0.6808690	(0.0232)
	0.05	0.6424703	(0.0130)	0.6241518	(0.0105)
	0.07	0.6433583	(0.0123)	0.6362110	(0.0084)
	0.09	0.6621593	(0.0223)	0.6396626	(0.0145)
20	0.00	1.1147692	(0.0190)	1.1078676	(0.0089)
	0.05	1.1084520	(0.0101)	1.1111880	(0.0107)
	0.07	1.1221157	(0.0088)	1.1406596	(0.0117)
	0.09	1.0973568	(0.0093)	1.1105763	(0.0098)

Table 4.10: *Model 3: Entropy risk estimates and corresponding standard errors.*

M	% missing	$\Delta_2(\hat{\Sigma}_{SS}^U)$		$\Delta_2(\hat{\Sigma}_{SS}^{V^*})$	
10	0.00	0.07386597	(0.0070)	0.07983551	(0.0072)
	0.05	0.08774246	(0.0070)	0.08925510	(0.0076)
	0.07	0.07721772	(0.0069)	0.07732465	(0.0060)
	0.09	0.06827196	(0.0058)	0.06552676	(0.0058)
15	0.00	0.10355612	(0.0135)	0.11376299	(0.0164)
	0.05	0.08529862	(0.0066)	0.09477951	(0.0080)
	0.07	0.09054799	(0.0085)	0.08137240	(0.0068)
	0.09	0.08621508	(0.0079)	0.07994681	(0.0070)
20	0.00	0.07030612	(0.0048)	0.07252679	(0.0057)
	0.05	0.08540962	(0.0071)	0.08498065	(0.0069)
	0.07	0.07861326	(0.0058)	0.07307411	(0.0060)
	0.09	0.07978378	(0.0069)	0.07698547	(0.0074)

Table 4.11: *Model 4: Entropy risk estimates and corresponding standard errors.*

M	% missing	$\Delta_2(\hat{\Sigma}^U_{SS})$		$\Delta_2(\hat{\Sigma}^{V^*}_{SS})$	
10	0.00	0.3311119	(0.0057)	0.3330848	(0.0053)
	0.05	0.3370160	(0.0065)	0.3255688	(0.0054)
	0.07	0.3313550	(0.0059)	0.3255281	(0.0055)
	0.09	0.3323215	(0.0045)	0.3187247	(0.0047)
15	0.00	0.6325137	(0.0107)	0.6252928	(0.0091)
	0.05	0.6297716	(0.0096)	0.6144335	(0.0076)
	0.07	0.6254772	(0.0087)	0.6241278	(0.0076)
	0.09	0.6121933	(0.0084)	0.6351603	(0.0095)
20	0.00	0.9334111	(0.0104)	0.9082255	(0.0041)
	0.05	0.9247592	(0.0079)	0.9319686	(0.0071)
	0.07	0.9308491	(0.0104)	0.9129209	(0.0055)
	0.09	0.9141808	(0.0073)	0.9212852	(0.0053)

Table 4.12: *Model 5: Entropy risk estimates and corresponding standard errors.*

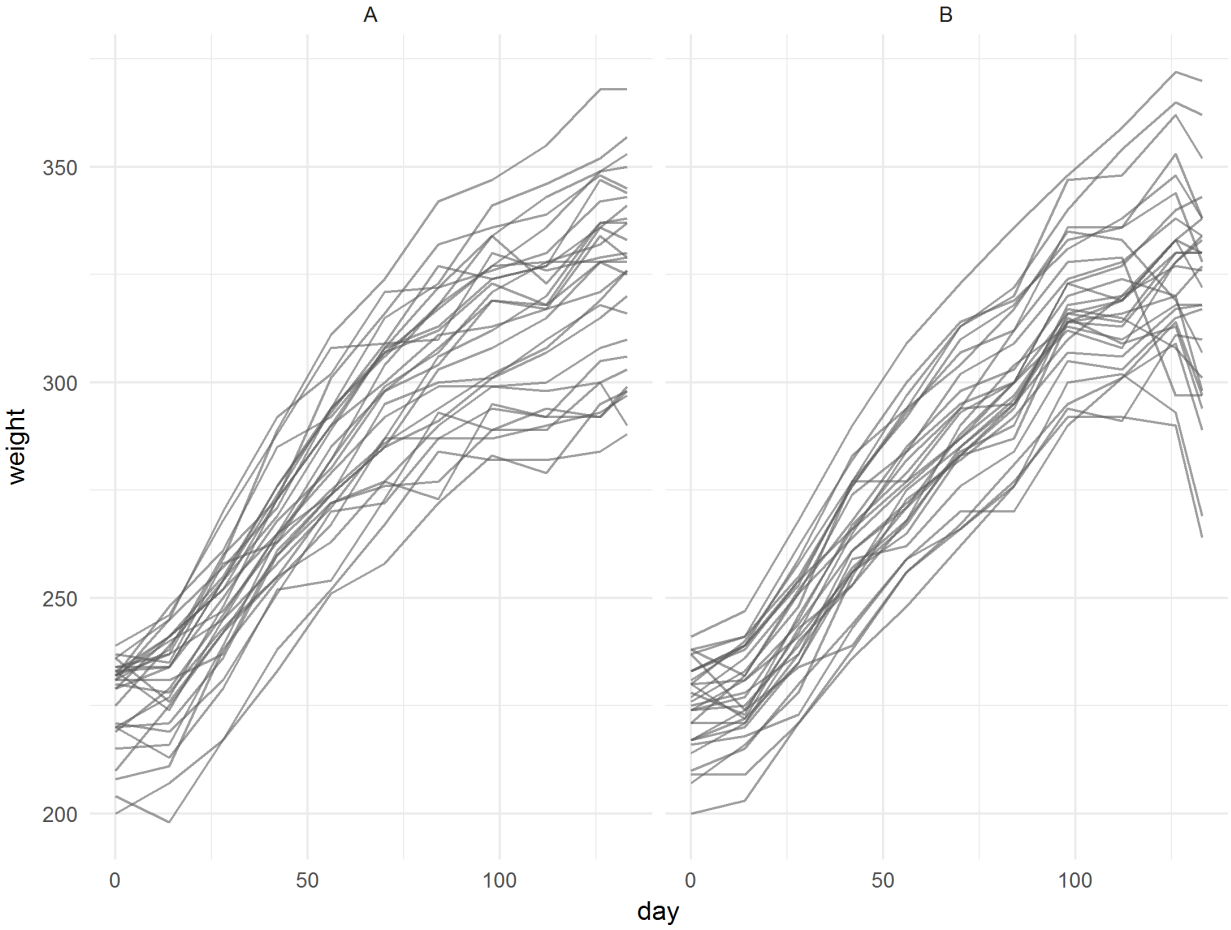
M	% missing	$\Delta_2(\hat{\Sigma}^U_{SS})$		$\Delta_2(\hat{\Sigma}^{V^*}_{SS})$	
10	0.00	0.2774130	(0.0057)	0.2841491	(0.0061)
	0.05	0.2911470	(0.0063)	0.2764965	(0.0059)
	0.07	0.2756559	(0.0060)	0.2721747	(0.0044)
	0.09	0.2820917	(0.0057)	0.2806499	(0.0056)
15	0.00	0.5132241	(0.0064)	0.5171960	(0.0067)
	0.05	0.5210875	(0.0081)	0.5124293	(0.0068)
	0.07	0.5239758	(0.0056)	0.5171861	(0.0059)
	0.09	0.5103058	(0.0049)	0.5151156	(0.0050)
20	0.00	0.7626041	(0.0106)	0.7536784	(0.0059)
	0.05	0.7490586	(0.0094)	0.7692628	(0.0073)
	0.07	0.7577880	(0.0068)	0.7575319	(0.0050)
	0.09	0.7662940	(0.0101)	0.7578350	(0.0057)

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. The measurement times were common across animals and were rescaled to $t = 1, 2, \ldots, 10, 10.5$. The longitudinal dataset is balanced, as there were no missing observations for any of the experimental units. Observed weights are shown in Figure **??**.

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], Pourahmadi [1999], and Pan and [2006], 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 profile plot in Figure **??** of the weights for units in

Figure 4.5: Subject-specific weight curves over time for treatment groups A and B.



treatment group A shows a clear upward trend in weights; variances appear to increase over time, suggesting that the covariance structure is nonstationary.

Before modeling the covariance structure, it is necessary to construct an adequate estimate of the mean weight trajectories. After centering the data using the fitted mean, the residuals serve

as the data reserved for estimating the functions defining the Cholesky factor and innovation variances. To account for any between-subject variability, we adopt an approach akin to the dynamical conditionally linear mixed model proposed by Pourahmadi and Daniels; see Pourahmadi and Daniels [2002] for a detailed discussion. We model

$$r\left(t_{ij}\right)=\sum_{k<j}\phi\left(t_{ij},t_{ik}\right)r\left(t_{ij}\right)+\epsilon\left(t_{ij}\right) \tag{4.6}$$

where

$$r\left(t_{ij}\right)=y\left(t_{ij}\right)-\left(f\left(t_{ij}\right)+\alpha_{i}\right). \tag{4.7}$$

The subject-specific random effects $\{\alpha_i\}$ are assumed to be mutually independent and independent of $\epsilon\left(t_{ij}\right)$ for all $i,\ j$, with

$$\alpha_{i}\sim N\left(0,\sigma_{\alpha}^{2}\right).$$

We assume that the subject trajectories share a common mean function $f\in$

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

Figure **??** displays the he observed weight trajectories over time. Figure **??** shows the corresponding fitted mean curves.

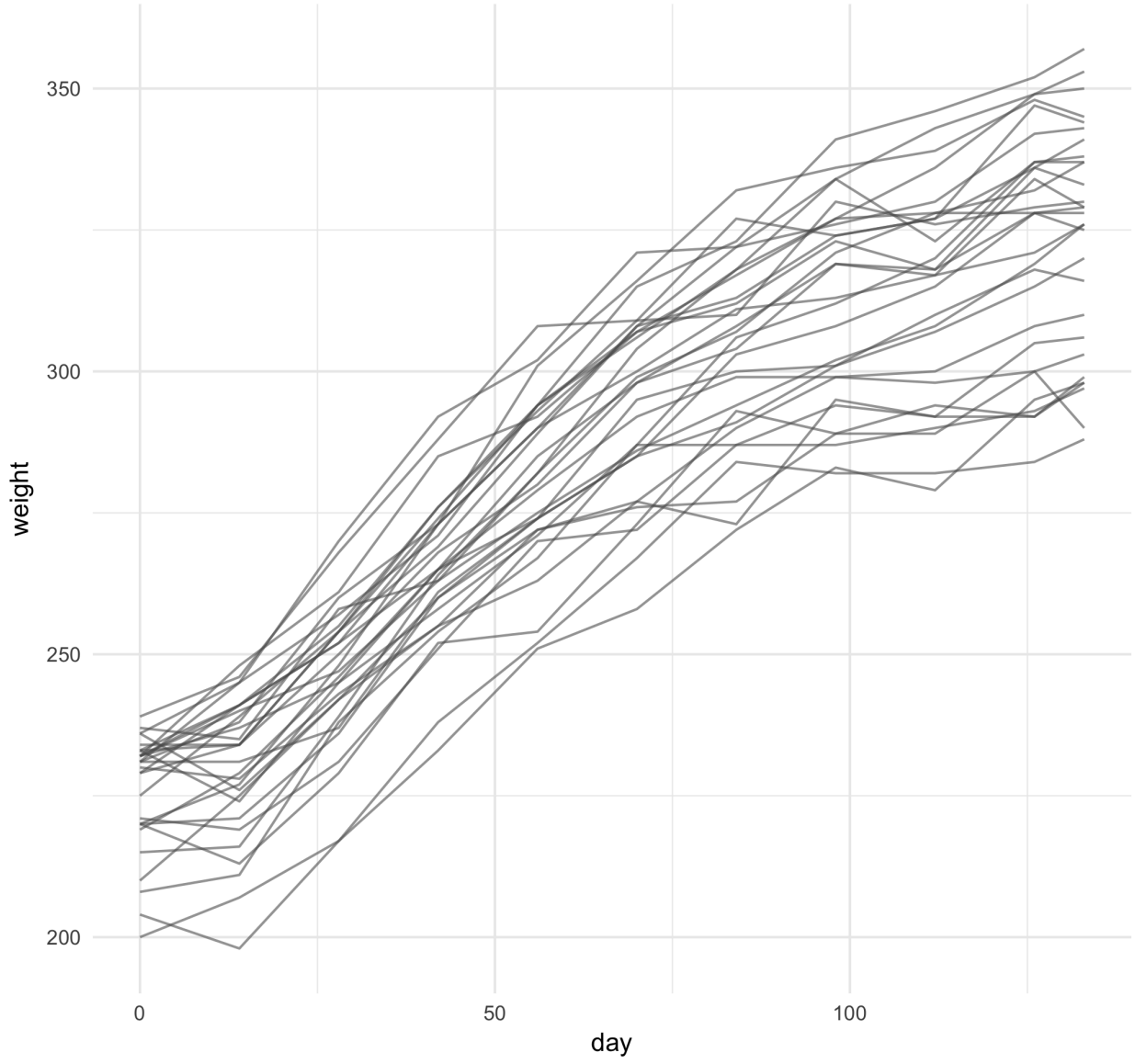
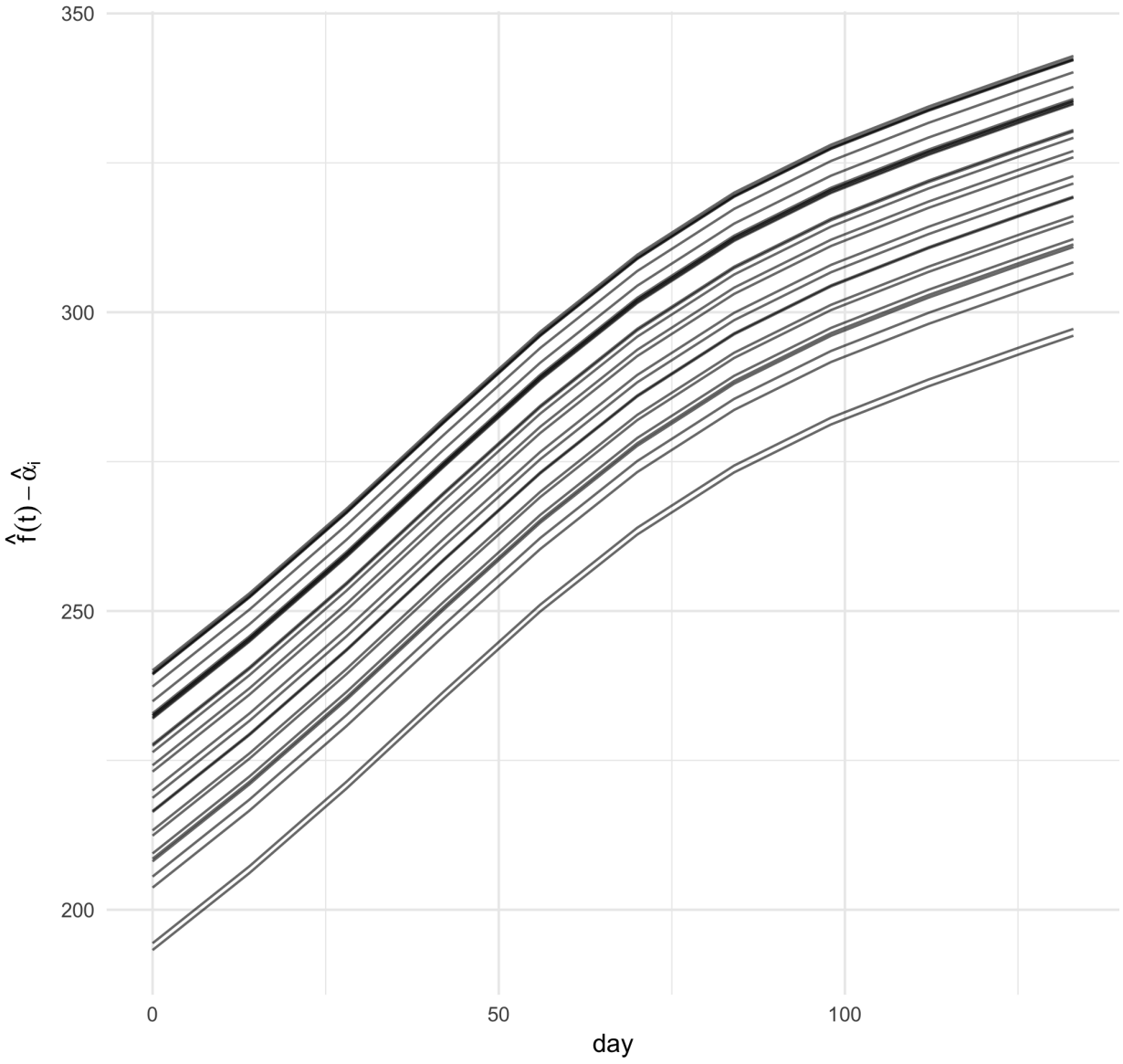


Figure 4.6: Weight trajectories over the observation period for experimental units in treatment group A.

Figure 4.7: Fitted mean weight curve for cattle in treatment group A.

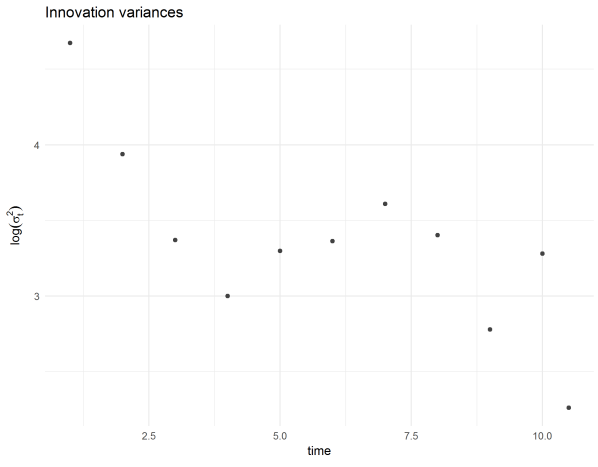


The nonstationarity suggested in Figure **??** is also supported by the sample correlations given in Table **??**; correlations within the subdiagonals are not constant and increase over time, a secondary indication that a stationary covariance is not appropriate for the data. Table **??** gives the

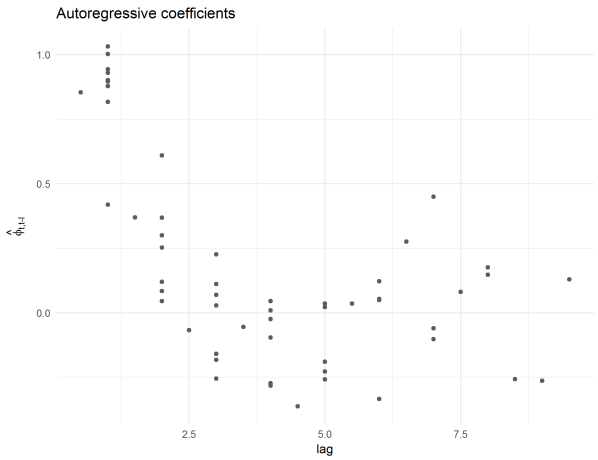
sample generalised autoregressive parameters and the innovation variances, which are plotted in Figure 4.8b and Figure 4.8a respectively.

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

Table 4.13: Cattle data: treatment group A sample correlations.



(a) Sample estimates of innovation variances σ_t^2 obtained by applying the modified Cholesky decomposition to the sample covariance matrix.



(b) Sample estimates of the generalized autoregressive parameters ϕ_{ij} obtained by applying the modified Cholesky decomposition to the sample covariance

		s											
		1	2	3	4	5	6	7	8	9	10	10.5	
	1	1											4.673	
	2	1.00	⋱										3.939	
	3	0.04	0.90										3.370	
	4	-0.25	0.25	0.88									3.000	
	5	-0.02	0.07	0.12	0.90								3.299	
t	6	0.04	-0.28	0.11	0.37	0.82							3.363	$\log(\sigma_t^2)$
	7	0.12	-0.23	0.04	-0.16	0.08	1.03						3.610	
	8	-0.06	0.05	0.02	-0.27	0.23	0.61	0.42					3.403	
	9	0.18	-0.10	0.05	-0.26	-0.10	0.03	0.30	0.93				2.780	
	10	-0.26	0.15	0.45	-0.33	-0.19	0.01	-0.18	0.37	0.94	⋱		3.280	
	10.5	0.13	-0.26	0.08	0.28	0.04	-0.36	-0.05	-0.07	0.37	0.85	1	2.262	

Table 4.14: Cattle data: treatment group A sample generalized autoregressive parameters (below the main diagonal) and log sample innovation variances (rightmost column.)

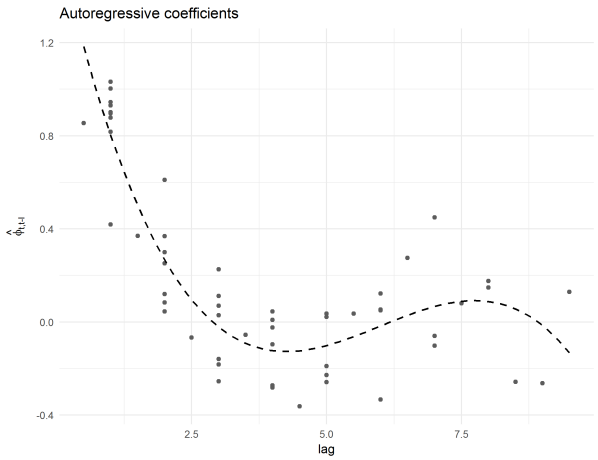
The analysis of Pourahmadi [1999] concluded that the regressogram (Figure 4.8b) and variogram (Figure **??**) suggest that both sample generalised autoregressive parameters and the logarithms of the innovation variances can be characterized in terms of cubic functions of the lag only. They model

$$\begin{aligned} \phi_{ts} &= x_{ts}'\gamma, \\ \log\left(\sigma_t^2\right) &= z_t'\xi, \end{aligned} \tag{4.8}$$

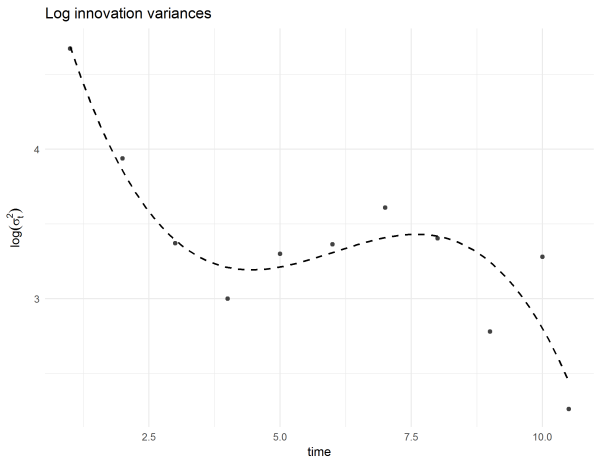
for $t = 2, \ldots, 11$ where

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

They estimate of γ and ξ via maximum likelihood. Figure **??** shows the estimated cubic polynomials corresponding to Model 4.8.



(a) Cubic polynomial fitted to the sample regressogram for the cattle data from treatment group A.



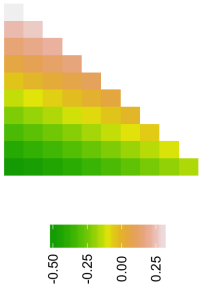
(b) Cubic polynomial fitted to the sample variogram for the cattle data from treatment group A.

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 $\phi_{t,t-l}$ show consistent behaviour over l for each value of t, indicating a lack of a strong functional component of m. To facilitate a making this modeling decision in an entirely data-driven manner, we let $\phi \in \mathcal{H} = \mathcal{H}_{[1]} \otimes \mathcal{H}_{[2]}$, where

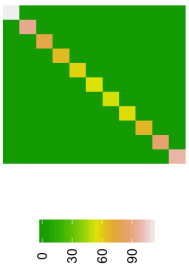
$$\mathcal{H}_{[1]} = \left\{ \phi : \ddot{\phi} = 0 \right\} \oplus \left\{ \phi : \phi(0) = \dot{\phi}(0) = 0; \int_0^1 \ddot{\phi}^2 \, dx < \infty \right\}$$

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

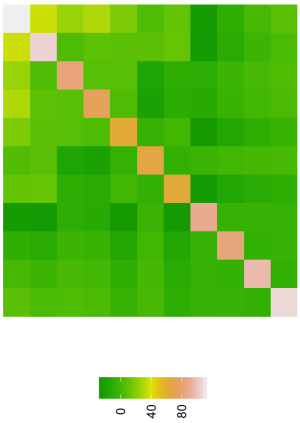
This decomposition leads to a null space comprised of functions of l only, which is attractive because it coincides with the modeling assumptions made by ϕ Pan and Pan [2017], Huang et al. [2006], and Wu and Pourahmadi [2003] for the same data set. Figure **??** and Figure 4.11 show the estimated Cholesky surface $\phi(t,s)$ and innovation variance function $\sigma^2(t)$ evaluated at $t = 1, 2, \ldots, 10, 10.5$ and the corresponding pairs of observation times (t,s), $1 \leq s < t \leq 10.5$.



(a) Estimated Cholesky factor $\hat{T}$



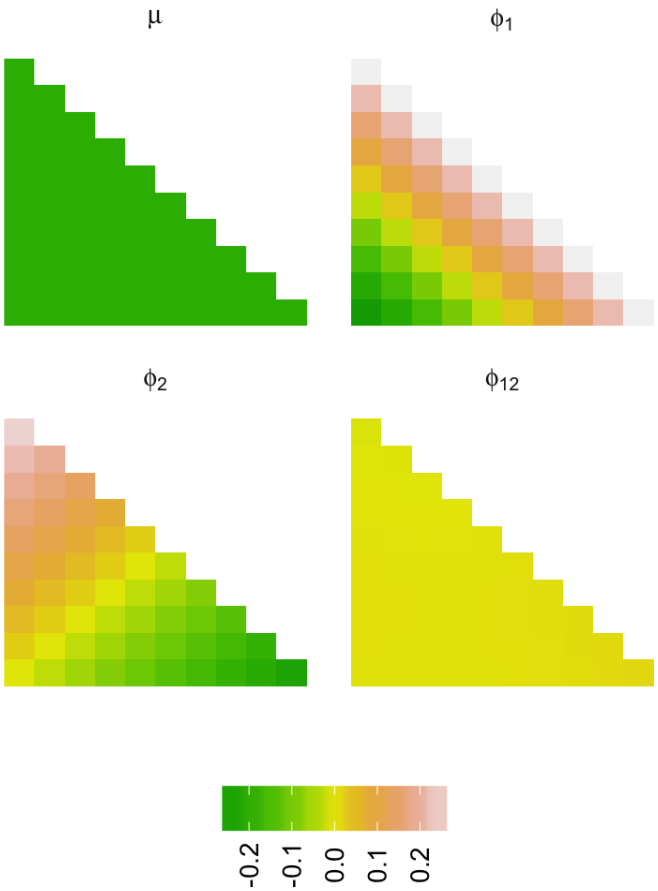
(b) Estimated innovation variances $\hat{D} = diag\left(\sigma^2\left(t_1\right), \ldots, \sigma^2\left(t_{11}\right)\right)$



(c) Estimated covariance matrix $\hat{\Sigma} = \hat{T}^{-1}\hat{D}\hat{T}'^{-1}$

Figure 4.10: Components of the fitted modified Cholesky decomposition for the cattle weight data.

Figure 4.11: Components of the SSANOVA decomposition of the estimated generalized autoregressive coefficient function ϕ evaluated on the grid defined by the observed time points.



Evaluating the normal likelihood at the fitted model gives $\hat{\ell} = -818.5323$.

Appendix A: Appendix

A.1 Chapter 2: Smoothing Spline ANOVA models

Proof of Theorem 2.1.1. 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

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

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

$$\begin{aligned}
\xi_i(\boldsymbol{v}) &= \langle \xi_i, Q\boldsymbol{v} \rangle_{\mathcal{H}} \\
&= \langle P_J \psi_i, Q\boldsymbol{v} \rangle_{\mathcal{H}} = \langle \psi_i, P_J Q\boldsymbol{v} \rangle_{\mathcal{H}} \\
&= \langle \psi_i, Q_J \boldsymbol{v} \rangle_{\mathcal{H}} \\
&= L_i Q_J \boldsymbol{v}.
\end{aligned}$$

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

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

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 (2.15) 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 $||\rho||^2 = 0$.

□

A.2 Chapter 4: Simulation studies

A.2.1 Quadratic risk estimates for simulation study 1

Table A.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	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0015	0.0052	0.0267	0.0912	0.3901	0.3864	0.3874
	20	0.0010	0.0043	0.0459	0.0757	0.8371	0.7710	0.7716
	30	0.0026	0.0036	0.0386	0.1109	1.2857	1.1937	1.2074
$N = 100$	10	0.0005	0.0010	0.0209	0.0426	0.2116	0.1676	0.1720
	20	0.0003	0.0011	0.0212	0.0376	0.4255	0.3902	0.3970
	30	0.0002	0.0011	0.0276	0.0313	0.5984	0.5790	0.5842

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

	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0483	0.0623	0.0792	7.0137	0.6269	0.8108	0.5770
	20	0.7972	1.2456	0.4317	852.2787	2.7659	30.8197	36.1492
	30	6.7921	12.8700	7.2129	96997.8508	21.0228	365.0301	1804.9695
$N = 100$	10	0.0254	0.0525	0.0580	7.0482	0.2683	0.4351	0.2665
	20	0.2877	0.8153	0.2625	861.3937	1.3347	5.5170	7.3283
	30	2.7399	6.9793	3.6619	101509.5641	8.4769	66.9461	420.2973

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

	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0656	0.0665	0.0697	3.4849	0.4977	0.6678	0.5858
	20	1.0095	0.9146	0.4706	426.0848	2.0716	4.8213	8.4099
	30	10.8782	8.1124	5.3699	50613.5638	16.5536	779.2829	1181.3770
$N = 100$	10	0.0486	0.0363	0.0328	3.5437	0.2437	0.2929	0.2791
	20	0.6260	0.3783	0.1958	416.1285	1.0193	1.5353	5.1553
	30	5.9367	3.4576	2.2121	50821.3671	7.9582	14.2394	253.4296

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

	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.0153	0.0196	0.0053	0.2575	0.4420	0.4628	0.4620
	20	0.0450	0.0154	0.0073	0.4384	0.7951	0.9184	0.9177
	30	0.0893	0.0189	0.0072	0.6539	1.3363	1.3014	1.3013
$N = 100$	10	0.0112	0.0186	0.0031	0.2098	0.2136	0.2299	0.2295
	20	0.0420	0.0143	0.0027	0.4877	0.4509	0.4311	0.4307
	30	0.0792	0.0181	0.0035	0.6616	0.6263	0.6598	0.6589

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

N	M	$\hat{\Sigma}_{SS}$	$\hat{\Sigma}_{PS}$	$\hat{\Sigma}_{oracle}$	$\hat{\Sigma}_{poly}$	S	S^{ω}	S^{λ}
$N = 50$	10	0.3659	0.2456	0.1610	1.3738	0.8484	1.6174	0.8963
	20	1.0146	0.8206	0.5236	2.8419	1.7324	3.0233	1.6375
	30	1.5352	1.1507	0.4632	4.1877	2.5484	5.1546	2.6727
$N = 100$	10	0.3091	0.2678	0.0813	1.2439	0.4175	1.0431	0.4922
	20	0.9734	0.4111	0.1522	2.7280	0.7896	2.1932	0.8461
	30	1.6032	0.7701	0.3656	3.8905	1.2577	3.5722	1.3270

A.2.2 Quadratic risk estimates for simulation study 2

Table A.6: Model 1: Quadratic risk estimates and corresponding standard errors for the MCD smoothing spline ANOVA estimator via 100 simulated multivariate normal sample of size $N = 50$ when 5%, 7%, and 9% of the data are missing. Risk is reported for the estimator constructed using the unbiased risk estimate and leave-one-subject-out cross validation are used for smoothing parameter selection.

M	% missing	$\Delta_1(\hat{\Sigma}_{SS}^{U})$		$\Delta_1(\hat{\Sigma}_{SS}^{V^*})$	
10	0.00	0.0009265743	(1e-040)	0.0013929180	(2e-040)
	0.05	0.0017013396	(3e-040)	0.0011392156	(2e-040)
	0.07	0.0013180691	(2e-040)	0.0014179527	(2e-040)
	0.09	0.0019051797	(2e-040)	0.0010759292	(2e-040)
15	0.00	0.0014140333	(2e-040)	0.0015083986	(2e-040)
	0.05	0.0010297818	(1e-040)	0.0014716107	(2e-040)
	0.07	0.0015350683	(3e-040)	0.0013068047	(2e-040)
	0.09	0.0018375709	(3e-040)	0.0012133312	(2e-040)
20	0.00	0.0009757769	(2e-040)	0.0008901316	(1e-040)
	0.05	0.0009224688	(1e-040)	0.0005839169	(1e-040)
	0.07	0.0007680194	(1e-040)	0.0006763406	(1e-040)
	0.09	0.0007406254	(1e-040)	0.0009270686	(1e-040)

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

M	% missing	$\Delta_1(\hat{\Sigma}_{SS}^{U})$		$\Delta_1(\hat{\Sigma}_{SS}^{V^*})$	
10	0.00	0.0667855	(0.0051)	0.0506111	(0.0033)
	0.05	0.0556528	(0.0033)	0.0606226	(0.0040)
	0.07	0.0559094	(0.0034)	0.0582334	(0.0036)
	0.09	0.0567996	(0.0037)	0.0569597	(0.0034)
15	0.00	0.1816735	(0.0130)	0.2353608	(0.0188)
	0.05	0.1842752	(0.0119)	0.2162470	(0.0164)
	0.07	0.2225236	(0.0162)	0.2039186	(0.0147)
	0.09	0.2914073	(0.0283)	0.2239119	(0.0149)
20	0.00	0.7825688	(0.0560)	0.6797717	(0.0393)
	0.05	1.0196188	(0.0949)	0.9072404	(0.0694)
	0.07	0.7106319	(0.0453)	0.8460652	(0.0643)
	0.09	0.7537178	(0.0580)	0.8142586	(0.0618)

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

M	% missing	$\Delta_1(\hat{\Sigma}_{SS}^{U})$		$\Delta_1(\hat{\Sigma}_{SS}^{V^*})$	
10	0.00	0.0219924	(0.0021)	0.0474410	(0.0059)
	0.05	0.0413327	(0.0043)	0.0390815	(0.0045)
	0.07	0.0404115	(0.0044)	0.0333281	(0.0034)
	0.09	0.0332642	(0.0036)	0.0285832	(0.0029)
15	0.00	0.1493227	(0.0192)	0.1897041	(0.0240)
	0.05	0.1327721	(0.0209)	0.1893346	(0.0233)
	0.07	0.1939995	(0.0304)	0.1499919	(0.0201)
	0.09	0.1614249	(0.0218)	0.1244167	(0.0158)
20	0.00	0.4121676	(0.0503)	0.3912354	(0.0507)
	0.05	0.5088004	(0.0731)	0.4557829	(0.0573)
	0.07	0.4321571	(0.0619)	0.3652646	(0.0415)
	0.09	0.3996033	(0.0477)	0.3571157	(0.0569)

Table A.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	0.00	0.01413389	(6e-040)	0.01423523	(5e-040)
	0.05	0.01474523	(7e-040)	0.01340456	(5e-040)
	0.07	0.01410561	(7e-040)	0.01320081	(6e-040)
	0.09	0.01415638	(5e-040)	0.01265559	(4e-040)
15	0.00	0.03044151	(0.0012)	0.02958915	(0.0010)
	0.05	0.02932117	(0.0010)	0.02854528	(8e-040)
	0.07	0.02950376	(0.0010)	0.02879928	(7e-040)
	0.09	0.02820952	(9e-040)	0.03032463	(0.0010)
20	0.00	0.04691149	(0.0012)	0.04356297	(4e-040)
	0.05	0.04540770	(8e-040)	0.04662982	(9e-040)
	0.07	0.04637991	(0.0011)	0.04443208	(6e-040)
	0.09	0.04474288	(8e-040)	0.04526569	(6e-040)

Table A.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	0.00	0.3476117	(0.0084)	0.3650368	(0.0087)
	0.05	0.3746030	(0.0101)	0.3469890	(0.0097)
	0.07	0.3609025	(0.0089)	0.3636064	(0.0080)
	0.09	0.3606898	(0.0096)	0.3833337	(0.0118)
15	0.00	0.6769442	(0.0101)	0.6775040	(0.0097)
	0.05	0.6932373	(0.0115)	0.6763392	(0.0079)
	0.07	0.6603759	(0.0082)	0.6872979	(0.0109)
	0.09	0.6686860	(0.0096)	0.6784823	(0.0113)
20	0.00	1.0350332	(0.0187)	1.0076594	(0.0087)
	0.05	0.9928734	(0.0146)	1.0019491	(0.0098)
	0.07	0.9719834	(0.0138)	1.0194282	(0.0101)
	0.09	1.0136409	(0.0145)	1.0186249	(0.0100)

A.2.3 Comprehensive tables for study 1

Σ	N	M	$\hat{\Sigma}^{ure}_{SS}$		$\hat{\Sigma}^{ure}_{PS}$		$\hat{\Sigma}_{oracle}$		$\hat{\Sigma}_{poly}$		S		S^{ω}		S^{λ}	
I		10	0.0749	(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	0.0872	(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	0.0899	(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	0.0949	(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)	1539.665	(39.7267)	12.3582	(0.1070)	55.3674	(3.8362)	133.9980	(19.200
	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)	1582.479	(36.0484)	5.2389	(0.0475)	33.5207	(0.9390)	62.5030	(14.779
III	50	10	0.3416	(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	1.1140	(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)	1091.193	(31.2219)	12.4792	(0.1182)	49.9135	(7.7026)	121.7795	(18.397
	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)	1105.045	(21.8998)	5.2234	(0.0462)	29.1693	(0.6585)	60.3529	(14.247
IV	50	10	0.3422	(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	0.9208	(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	0.2743	(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	0.7526	(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 A.11: Risk estimates under entropy loss and corresponding standard errors based on 100 Monte Carlo simulations.

Σ	N	M	$\hat{\Sigma}^{ure}_{SS}$		$\hat{\Sigma}^{ure}_{PS}$		$\hat{\Sigma}_{oracle}$		$\hat{\Sigma}_{poly}$		S		S^{ω}		S^{λ}	
I	50	10	0.0015	(3e-040)	0.0052	(0.0010)	0.0267	(0.0045)	0.0912	(0.0103)	0.3901	(0.0247)	0.3864	(0.0221)	0.3874	(0.0
	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	(0.0
	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	(0.0
	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	(0.0
	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	(0.0
	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	(0.0
II	50	10	0.0483	(0.0070)	0.0623	(0.0043)	0.0792	(0.0083)	7.0137	(0.3452)	0.6269	(0.0363)	0.8108	(0.0690)	0.5770	(0.0
	50	20	0.7972	(0.1388)	1.2456	(0.1778)	0.4317	(0.0809)	852.279	(38.431)	2.7659	(0.2037)	30.820	(15.7299)	36.1492	(9.3
	50	30	6.7921	(1.5850)	12.8700	(1.4200)	7.2129	(1.2710)	1997.851	(55.87)	21.0228	(2.2821)	365.030	(18.7437)	1804.970	(43
	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	(0.0
	100	20	0.2877	(0.0477)	0.8153	(0.1501)	0.2625	(0.0377)	861.394	(34.1825)	1.3347	(0.1086)	5.5170	(0.6241)	7.3283	(1.4
	100	30	2.7399	(0.4745)	6.9793	(0.9114)	3.6619	(0.7715)	1509.564	(53.587)	8.4769	(0.7058)	66.9461	(6.0353)	420.297	(11
III	50	10	0.0656	(0.0053)	0.0665	(0.0033)	0.0697	(0.0102)	3.4849	(0.2297)	0.4977	(0.0265)	0.6678	(0.0645)	0.5858	(0.0
	50	20	1.0095	(0.1420)	0.9146	(0.1113)	0.4706	(0.0731)	426.085	(26.445)	2.0716	(0.1360)	4.8213	(1.1130)	8.4099	(1.3
	50	30	10.8782	(1.1771)	8.1124	(1.2342)	5.3699	(0.8475)	5613.564	(112.439)	16.5536	(1.8098)	779.283	(14.9847)	1181.377	(32
	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	(0.0
	100	20	0.6260	(0.0200)	0.3783	(0.0823)	0.1958	(0.0308)	416.129	(12.8666)	1.0193	(0.0701)	1.5353	(0.1560)	5.1553	(1.0
	100	30	5.9367	(0.7791)	3.4576	(0.7345)	2.2121	(0.3658)	4821.367	(85.815)	7.9582	(0.8381)	14.239	(1.7202)	253.430	(75
IV	50	10	0.0153	(0.0010)	0.0196	(0.0039)	0.0053	(0.0012)	0.2575	(0.0340)	0.4420	(0.0293)	0.4628	(0.0365)	0.4620	(0.0
	50	20	0.0450	(6e-040)	0.0154	(0.0024)	0.0073	(0.0012)	0.4384	(0.0416)	0.7951	(0.0447)	0.9184	(0.0397)	0.9177	(0.0
	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	(0.0
	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	(0.0
	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	(0.0
	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	(0.0
V	50	10	0.3659	(0.0123)	0.2456	(0.0206)	0.1610	(0.0332)	1.3738	(0.0999)	0.8484	(0.0549)	1.6174	(0.1133)	0.8963	(0.0
	50	20	1.0146	(0.0102)	0.8206	(0.0213)	0.5236	(0.1373)	2.8419	(0.1751)	1.7324	(0.0802)	3.0233	(0.1872)	1.6375	(0.0
	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	(0.1
	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	(0.0
	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	(0.0
	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	(0.0

Table A.12: Risk estimates under quadratic loss and corresponding standard errors based on 100 Monte Carlo simulations.

Bibliography

T. W. Anderson, editor. *An Introduction to Multivariate Statistical Analysis*. Wiley, 1984.

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

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

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

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

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

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

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

Raymond J Carroll and David Ruppert. *Transformation and weighting in regression*, volume 30. CRC Press, 1988.

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

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

Tom YM Chiu, Tom Leonard, and Kam-Wah Tsui. The matrix-logarithmic covariance model. *Journal of the American Statistical Association*, 91(433):198–210, 1996.

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

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

Arthur P Dempster. Covariance selection. *Biometrics*, pages 157–175, 1972.

John E Dennis Jr and Robert B Schnabel. *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, 1996.

Dipak K Dey, Sujit K Ghosh, and Bani K Mallick. *Generalized linear models: A Bayesian perspective*. CRC Press, 2000.

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

Paul HC Eilers. Penalized regression in action: Estimating pollution roses from daily averages. *Environmetrics*, 2(1):25–47, 1991a.

Paul HC Eilers. Indirect observations, composite link models and penalized likelihood. In *Statistical Modelling*, pages 91–98. Springer, 1995.

Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical science*, pages 89–102, 1996.

PHC Eilers. Nonparametric density estimation with grouped observations. *Statistica neerlandica*, 45(3):255–269, 1991b.

Randall L Eubank. *Nonparametric regression and spline smoothing*. CRC press, 1999.

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

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

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

Philip E Gill, Walter Murray, and Margaret H Wright. Practical optimization. 1981.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Chong Gu. Smoothing spline anova models, 2002.

Chong Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

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

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

Trevor Hastie and Robert Tibshirani. *Generalized additive models*. Wiley Online Library, 1990.

Trevor Hastie and Robert Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 757–796, 1993.

Jianhua Z Huang, Naiping Liu, Mohsen Pourahmadi, and Linxu Liu. Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, pages 85–98, 2006.

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

Robert I Jennrich and Mark D Schluchter. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, pages 805–820, 1986.

Michael G Kenward. A method for comparing profiles of repeated measurements. *Applied Statistics*, pages 296–308, 1987.

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

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

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

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

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

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

Shang P Lin. A monte carlo comparison of four estimators for a covariance matrix. *Multivariate Analysis*, 6:411–429, 1985.

Brian D Marx and Paul HC Eilers. Multidimensional penalized signal regression. *Technometrics*, 47(1):13–22, 2005.

P. McCullagh and J.A. Nelder. *Generalized linear models*. Chapman and Hall, London, 2nd edition, 1989.

Charles E McCulloch and John M Neuhaus. *Generalized linear mixed models*. Wiley Online Library, 2001.

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

Hoh Suk Noh and Byeong U Park. Sparse varying coefficient models for longitudinal data. *Statistica Sinica*, pages 1183–1202, 2010.

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

Jianxin Pan and Gilbert . Regression models for covariance structures in longitudinal studies. *Statistical Modelling*, 6(1):43–57, 2006.

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

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

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 2012.

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

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

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

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

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

James O Ramsay. *Functional data analysis*. Wiley Online Library, 2006.

James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*. Springer, 2007.

Christian H Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183, 1967.

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

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

IJ Schöenberg and A Whitney. On polya frequency functions. *Trans. Amer. Math. Soc*, 74:246–259, 1953.

Shayle R Searle, George Casella, and Charles E McCulloch. *Variance components*, volume 391. John Wiley & Sons, 2009.

Damla Şentürk and Hans-Georg Müller. Generalized varying coefficient models for longitudinal data. *Biometrika*, 95(3):653–666, 2008.

Damla Şentürk, Lorien S Dalrymple, Sandra M Mohammed, George A Kaysen, and Danh V Nguyen. Modeling time-varying effects with generalized and unsynchronized longitudinal data. *Statistics in medicine*, 32(17):2971–2987, 2013.

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

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

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

Charles J Stone, Mark H Hansen, Charles Kooperberg, Young K Truong, et al. Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *The Annals of Statistics*, 25(4):1371–1470, 1997.

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

Grace Wahba. *Spline models for observational data*, volume 59. Siam, 1990.

Grace Wahba, Yuedong Wang, Chong Gu, Ronald Klein, and Barbara Klein. Smoothing spline anova for exponential families, with application to the wisconsin epidemiological study of diabetic retinopathy. *The Annals of Statistics*, pages 1865–1895, 1995.

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

Binhuan Wang. *CVTuningCov: Regularized Estimators of Covariance Matrices with CV Tuning*, 2014. URL `https://CRAN.R-project.org/package=CVTuningCov`. R package version 1.0.

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

Edmund T Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922.

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

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

Wei Biao Wu and Mohsen Pourahmadi. Banding sample autocovariance matrices of stationary processes. *Statistica Sinica*, pages 1755–1768, 2009.

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

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

Ruoyong Yang and James O Berger. Estimation of a covariance matrix using the reference prior. *The Annals of Statistics*, pages 1195–1211, 1994.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

Scott L Zeger and Peter J Diggle. Semiparametric models for longitudinal data with application to cd4 cell numbers in hiv seroconverters. *Biometrics*, pages 689–699, 1994.

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

Dale L Zimmerman and Vicente Núñez-Antón. Structured antedependence models for longitudinal data. In *Modelling longitudinal and spatially correlated data*, pages 63–76. Springer, 1997.