Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

June 16, 2017

Abstract

With high dimensional longitudinal and functional data becoming much more common, there is a strong need for methods of estimating large covariance matrices. Estimation is made difficult by the instability of sample covariance matrices in high dimensions and a positive-definite constraint we desire to impose on estimates. A Cholesky decomposition of the covariance matrix allows for parameter estimation via unconstrained optimization as well as a statistically meaningful interpretation of the parameter estimates. Regularization improves stability of covariance estimates in high dimensions, as well as in the case where functional data are sparse and individual curves are sampled at different and possibly unequally spaced time points. By viewing the entries of the covariance matrix as the evaluation of a continuous bivariate function at the pairs of observed time points, we treat covariance estimation as bivariate smoothing.

Within regularization framework, we propose novel covariance penalties which are designed to yield natural null models presented in the literature for stationarity or short-term dependence. These penalties are expressed in terms of variation in continuous time lag and its orthogonal complement. We present numerical results and data analysis to illustrate the utility of the proposed method.

keywords: non-parametric, covariance, longitudinal data, functional data, splines, reproducing kernel Hilbert space

1 Introduction

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. Estimation of population covariance matrices from samples

*The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201
†The Ohio State University, 1958 Neil Avenue, Columbus, OH 43201

of multivariate data has been important for methods 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 establishing independence or conditional independence through graphical models, constructing discriminant functions as in linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) for the classification of Gaussian data, building confidence intervals for component means and contrasts, and constructing a low-dimensional representation of data via principal components analysis (PCA). One may note that the last two techniques require an estimate of the covariance matrix, and the first two require estimation of the inverse.

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 is possibly much larger than the number of observations. Additional difficulty due to constraints required to yield positive definite estimates make covariance estimation a potentially complex optimization problem. Further, most existing approaches to covariance estimation require data to be sampled at regular grid (time) points, with subjects sharing a set of common observation points. However, in many practical situations, data are irregularly sampled, and subjects may share few common observation times, and methods are needed to accommodate for data collected in this way.

To address the challenge of enforcing positive definiteness, 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.

It is well known that the sample covariance matrix is unstable in high dimensions, and there is an extensive existing body of work addressing the issue of high dimensionality in the context of covariance estimation. See Pourahmadi [2011] for a survey of approaches to covariance estimation from the generalized linear modeling and regularization perspectives. However, much of this work addresses high dimensionality arising from functional or times series data sampled on a dense, regular grid. With such data, it is typical that the number of time points is larger than the number of observations. Few have addressed the challenges posed by 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, 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. Incomplete and unbalanced data arise when measurement schedules with targeted time points which are not necessarily equally spaced or if there is missing data. Sparse longitudinal data arise when the measurement schedule has arbitrary or almost unique time points for every individual. A given time point may have very few individuals with corresponding measurements.

We sidestep both issues of high dimensionality and irregularly sampled data by viewing the

response as a stochastic process having continuous covariance function. Recent work outlines the use of function estimation for smoothing elements of the covariance matrix, including Wu and Pourahmadi [2003],Huang et al. [2007]. To our knowledge, however, no previous work has applied smoothing to both dimensions of the Cholesky factor; we model the generalized autoregressive parameters using tensor product splines. Viewing covariance modeling as bivariate function estimation both accommodates irregularly sampled curved and permits interpolation and extrapolation of the covariance function between two measurements at any pair of time points within the time interval of interest rather than at observed pairs of time points only. The Cholesky decomposition enables covariance estimation through the estimation of a varying coefficient model. A transformation of the design point axes allows for an ANOVA-like decomposition of the coefficient function into two components, corresponding to the lag between time points and an additive component. Through this general framework, we can easily impose penalties on fitted functions to yield natural null models presented in the literature.

2 Cholesky Decomposition of Σ

To present a comprehensive overview our estimation procedure, we begin with the representation of the inverse covariance matrix, $\Omega = \Sigma^{-1}$, in terms of its Cholesky decomposition (see Pourahmadi [2007] for a detailed discussion.) In the section to follow, we will demonstrate that this parameterization of the precision matrix is particularly attractive due to both the computational advantages as well as the convenient modeling interpretation it permits. For any positive definite matrix Σ, there exists a unique unit lower triangular matrix T with diagonal entries equal to 1 which diagonalizes Σ:

$$T\Sigma T^T = D$$

If we assume that the data having covariance matrix Σ follow an autoregressive model, then the entries of the Cholesky factor T and D enjoy a useful interpretation. Let $Y = (Y_1, Y_2, \ldots, Y_m)^T$ be defined on a probability space with some probability measure $\mathcal{P}$ corresponding to the multivariate Normal distribution with mean 0 and covariance Σ, and let $Y_1, Y_2, \ldots, Y_m$ have associated measurement times

$$t_1 < t_2 < \cdots < t_m.$$

Consider regressing Y_j on its predecessors:

$$Y_j = \sum_{k=1}^{j-1} \phi_{jk} Y_k + \sigma_j e_j, \qquad j = 2, \ldots, m, \tag{1}$$

where we define $y_1 = e_1$. Standard regression theory gives us that if $\{\phi_{jk}\}$ are the coefficients of the linear least squares predictor of y_j based on its predecessors, then the residuals $e = (e_1, e_2, \ldots, e_m)^T$ have diagonal covariance. Let T denote the $m \times m$ matrix with elements

$$T_{jk} = \begin{cases} -\phi_{jk} & j > k \\ 1 & j = k \\ 0 & otherwise, \end{cases}$$

for $j, k = 1, \ldots, m$. Then in matrix notation, model 1 may then be written

$$e = TY, \tag{2}$$

Taking covariances on both sides of 2, we have

$$D = T\Sigma T^T \tag{3}$$

An attractive feature of this reparameterisation is that, regardless of the modelling approach, the estimated covariance matrix is guaranteed to be positive definite. The unconstrained regression coefficients $\{\phi_{jk}\}$ are referred to as the *generalized autoregressive parameters* (GARPs). The $\{\sigma_j^2\}$ are called the *innovation variances* (IVs.) Unconstrained estimation of the $\{\sigma_k^2\}$ is achieved by log transformation; we leave these details for section 2. Expressing the precision matrix in terms of the GARPs and IVs, we have

$$\Omega = \Sigma^{-1} = T^T D^{-1} T. \tag{4}$$

Rather than estimating a specific covariance matrix for data observed on a fixed, regular grid, we aim to estimate a smooth covariance function. This accomodates data which may consist of observations on multiple subjects measured at potentially unequally spaced and individual-specific times. In estimation of the means μ of p-vectors of i.i.d. variables, the Gaussian white noise model [9] is the appropriate infinite-dimensional model into which all objects of interest are embedded. In estimation of matrices, a natural analogue is the space B(l2,l2), which we write as B, of bounded linear operators from l2 to l2. These can be represented as matrices [cite *Regularized estimation of large covariance matrices by Bickel and Levina - section 4.*]

Rather than m-dimensional vectors, consider Y and e as the values of the stochastic processes $Y(t)$ and $e(t)$ at the set of observation times. We assume that $Y(t)$ is equipped with covariance function $G(s,t)$, and

$$e(s) \sim \mathcal{WN}(0,1)$$

is a zero mean Gaussian white noise process with unit variance. We assume that $G(s,t)$ satisfies some smoothness conditions, where smoothness is defined in terms of square integrability of certain derivatives. [TODO: clean up statement about smoothness of covariance function; integrability of covariance function of a stochastic process?] The entries of Σ, then, correspond to G evaluated at the distinct pairs of observed time points. Similarly, we treat the elements of the precision matrix Ω as the values of some smooth function, $\omega(s,t)$ evaluated at observed pairs of time points.

Extending this perspective to the elements of D and the elements of the Cholesky factor T leads us to the varying coefficient (VC) models first introduced by Hastie and Tibshirani. The procedures presented by Fan and Zhang [2000] and Huang et al. [2002] utilize varying coefficient models for modeling the mean of longitudinal data; parameterizing the covariance matrix according to 3 allows us to exploit these models in covariance estimation for such data as well. A

generalization of traditional linear regression models, varying coefficient models offer more flexibility than their static analogues by allowing the effect of covariates to change smoothly with the value other variables. Both regressors and response variables are assumed to vary according to an *indexing variable*, which is particularly attractive because this permits interpolation of regressors and response variables at values of this indexing variable where there is either missing data of only a single observation and slope estimation is not feasible. Replacing $\{\phi_{jk}\}$ and $\{\sigma_j\}$ with smooth functions, we model

$$y(t_j) = \sum_{k=1}^{j-1} \phi(t_j, t_k)\, y(t_k) + \sigma(t_j)\, \epsilon(t_j) \qquad j = 1, \ldots, m, \tag{5}$$

for $t_1 < t_2 < \cdots < t_m$.

We represent the varying coefficient function and the innovation variances using tensor product smoothing splines and penalized tensor product B-splines alongside penalties to induce simplicity in ϕ and σ^2 to produce final covariance estimates exhibiting the desired null structure. For ease of exposition, we first focus our attention on the estimation of ϕ assume that $\sigma^2(t)$ is fixed and known; we will later propose an iterative procedure for simultaneous estimation of σ^2 and ϕ. Recasting the problem as the estimation of model 5 allows us access to the existing set of tools developed in the bivariate smoothing literature; our approach provides a flexible, comprehensive framework for covariance estimation.

3 Penalized Maximum Likelihood Estimation of ϕ

We employ maximum likelihood for the estimation of the varying coefficient function $\phi(t, s)$ and the innovation variance function $\sigma(t)$, though neither the derivation the form of model 1 nor model 5 via the Cholesky decomposition rely on any assumptions about the distribution of Y. Fixing σ_j^2, for a sample of N i.i.d. observations $Y_1, Y_2, \ldots, Y_N$ from a multivariate Gaussian distribution, the negative log-likelihood as a function of ϕ_{jk} corresponds to the usual error sums of squares and is proportional to

$$-2L(y_1, y_2, \ldots, y_N, \Phi) \propto \sum_{i=1}^{N} \sum_{j=2}^{m_i} \sigma(t_j)^{-2} \left(y_{ij} - \sum_{k=1}^{j-1} \phi(t_{ij}, t_{ik})\, y_{ik} \right)^2 \tag{6}$$

where

$$y_i = (y_{i1}, y_{i2}, \ldots, y_{i,m_i}), \quad i = 1, \ldots, N$$

denotes the vector of observations for subject i with corresponding measurement times

$$t_{i1} < t_{i2} < \cdots < t_{i,m_i}.$$

The form of the likelihood of $y_1, \ldots, y_N$ indicates that we allow both the number of measurements as well as the observation times to varying across subjects. The $\{t_{ij}\}$ need not be evenly-spaced within or across individuals.

In the case that subjects share a common set of observation times $t_1 < \cdots < t_m$, it is well known that the MLE for Σ, $S = \sum_{i=1}^{N} y_i y_i^T$ is highly unstable in high dimensions, a condition that is potentially worsened when one or more subjects has at least one observation time that is unique from the set of observation times common across subjects. To mitigate instability due to high dimensionality and simultaneously permit the estimation of $\phi(\cdot, \cdot)$ as a smooth bivariate function, we obtain a covariance estimator by applying bivariate smoothing of the elements of the Cholesky factor.

4 Smoothing spline representation of ϕ

To impose structure on the estimated varying coefficient function, we augment the negative log-likelihood **??** with penalty functional, which discourages the flexibility of the fitted function. We take the estimator of ϕ to minimize

$$-2L + \lambda J_\phi(\phi). \tag{7}$$

The first term in 7 discourages the lack of fit of ϕ to the data, and λ is a smoothing parameter which controls the tradeoff between the lack of fit and amount of regularization imposed on the fitted function through the penalty, J_ϕ. Since ϕ explicitly defines an inverse covariance function, imposing specific types of structure on ϕ is of particular interest; covariance models for longitudinal or time series data are commonly defined in terms of lag, or in the continuous case, the difference between two measurement times. By transforming the $s - t$ input axis, we reparameterize ϕ and express the coefficient function in terms of

$$\begin{aligned} l &= s - t \\ m &= \frac{1}{2}(s + t). \end{aligned}$$

Writing ϕ in terms of the rotation gives the reparameterized coefficient function

$$\phi^*(l, m) = \phi^*\left(s - t, \frac{1}{2}(s + t)\right) = \phi(s, t). \tag{8}$$

We define our estimator $\hat{\phi}^*$ as the minimizer of

$$-2L + \lambda^* J_{\phi^*}(\phi^*). \tag{9}$$

We consider models that capture the marginal effects of l and m, as well as interaction between the two directions. We first consider the smoothing spline ANOVA decomposition of Gu [2002], modeling

$$\phi^*(l, m) = \mu + \phi_l(l) + \phi_m(m) + \phi_{lm}(l, m). \tag{10}$$

As in Gu [2002], Craven and Wahba [1978],[more Wahba citations here], we consider functions ϕ^* belonging to a reproducing kernel Hilbert space (r.k.h.s.), $\mathcal{H}$. We equip each l and m

with corresponding univariate Hilbert spaces, $\mathcal{H}_l$ and $\mathcal{H}_m$, choosing to let $\mathcal{H}_l$ correspond to the second-order Sobolev space $W_2(0,1)$ and $\mathcal{H}_m$ to the first-order Sobolev space $W_1(0,1)$, where

$$W_m(0,1) = \{f: \quad f, f' \text{absolutely continuous}, \int_0^1 \left(f^{(m)}\right)^2 dt < \infty\}.$$

for $m = 1, 2$. Each space $\mathcal{H}_l$, $\mathcal{H}_m$ is endowed with inner product product

$$\langle f, g \rangle = \sum_{\nu=0}^{m-1} \left(\int_0^1 f^{(\nu)}(x)\, dx\right)\left(\int_0^1 g^{(\nu)}(x)\, dx\right) + \int_0^1 f^{(m)} g^{(m)} dx \tag{11}$$

The space of bivariate functions $\mathcal{H}$ can be constructed from the tensor product of the univariate function spaces for l and m:

$$\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m.$$

Several have proposed methods for applying regularization of Cholesky decomposition including banding, tapering, kernel smoothing, penalized likelihood, and penalized regression. See [ ] Within the function estimation paradigm, a number of approaches to estimate the coefficient function $\phi(\cdot,\cdot)$ have been proposed including See Wu and Pourahmadi [2003], Huang et al. [2007] . Common techniques for inducing structure to produce simple and stable covariance estimates include shrinking estimated functions or the elements of the covariance matrix itself so that the resulting dependency structure corresponds to parsimonious covariance models frequently adopted in the time series and longitudinal data literature. [CITE PAPERS PROPOSING PARSIMONIOUS MODELS FOR phi ij] The ANOVA model in 10 allows us to easily specify penalties J that encourage estimates to adhere to the structure of these models. [cite some general time series/longitudinal sources] When ϕ^* corresponds to the simple models of the form (**??**), the bivariate function may be written in terms of only its first argument. . .

The penalty functional J induces a decomposition of $\mathcal{H}$ as a direct sum of two subspaces:

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

where $\mathcal{H}_0$ denotes the null space of J, spanned by $\tau_1, \tau_2, \ldots \tau_M$, and $\mathcal{H}_1$ is the subspace orthogonal to $\mathcal{H}_0$. Let $P_1\phi^*$ denote the projection of ϕ^* onto the penalized space $\mathcal{H}_1$. We can express J in terms of the projection of $\phi^* \in \mathcal{H}$ onto $\mathcal{H}_1$:

$$\begin{aligned} J(\phi) &= ||P_1\phi^*||^2 \\ &= ||P_1\phi_l||^2 + ||P_1\phi_m||^2 + ||P_1\phi_{lm}||^2 \end{aligned} \tag{12}$$

The decomposition of $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ can be characterized by the decompositions of $\mathcal{H}_l$ and $\mathcal{H}_m$ induced by J:

$$\begin{aligned} \mathcal{H}_l &= \mathcal{H}_{l0} \oplus \mathcal{H}_{l1} \\ \mathcal{H}_m &= \mathcal{H}_{m0} \oplus \mathcal{H}_{m1} \end{aligned} \tag{13}$$

As $\lambda \to \infty$, the penalty term dominates the objective function in 9, forcing the minimizer to adopt the functional form of the $\mathcal{H}_0$. The parameterization in 10 allows us to easily construct penalties to so that for large values of λ, the fitted function will correspond to [cite the simple parametric and semiparametric models of Pourahmadi, Wu, etc as well as the null models proposed by others utilizing smoothing methods]. We consider specification of the penalty so that the null space excludes any functions ϕ^* which are non-constant in m, letting

$$\mathcal{H}_{m0} = \{1\} \tag{14}$$

Additionally, we let ϕ^* which are linear in lag l to incur zero penalty, letting

$$\mathcal{H}_{l0} = \{1\} \oplus \{k_1\}, \tag{15}$$

where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials satisfying

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

The penalized spaces $\mathcal{H}_{l1}$, $\mathcal{H}_{m1}$, defined as the subspaces orthogonal to $\mathcal{H}_{l0}$ and $\mathcal{H}_{m0}$ respectively, satisfy

$$\begin{aligned} \mathcal{H}_{l1} &= \{\phi_l : \int_0^1 \phi_l{}^{(\nu)}(l)\, dl = 0, \;\; \nu = 0, 1\} \\ \mathcal{H}_{m1} &= \{\phi_m : \int_0^1 \phi_m(m)\, dm = 0\} \end{aligned}$$

Using the properties of tensor product spaces, we may write $\mathcal{H}_0$ and $\mathcal{H}_1$ in terms of the elements defining the marginal subspaces:

$$\begin{aligned} \mathcal{H}_0 &= \{1\} \oplus \{k_1\} \\ \mathcal{H}_1 &= \mathcal{H}_{l1} \oplus \mathcal{H}_{m1} \oplus [\{k_1\} \otimes \mathcal{H}_{m1}] \oplus [\mathcal{H}_{l1} \otimes \mathcal{H}_{m1}.] \end{aligned}$$

Table 1: Tensor product space $\mathcal{H}$

	$\{1\}$	$\{k_1(l)\}$	$\{\mathcal{H}_{l1}\}$
$\{1\}$	$\{1\}$	$\{k_1(l)\}$	$\{\mathcal{H}_{l1}\}$
$\{\mathcal{H}_{m1}\}$	$\{\mathcal{H}_{m1}\}$	$\{\mathcal{H}_{m1}\} \otimes \{k_1(l)\}$	$\{\mathcal{H}_{m1}\} \otimes \{\mathcal{H}_{l1}\}$

The subspaces of $W_1[0,1] \otimes W_2[0,1]$ by the tensor product of the marginal subspaces of $\mathcal{H}_l$, $\mathcal{H}_m$.

Table 1 shows how the space of two-dimensional functions is constructed by taking tensor products of each of the subspaces which define the two univariate spaces, $\mathcal{H}_l$ and $\mathcal{H}_m$. One may show that the reproducing kernels for $\mathcal{H}_{l0}$ and $\mathcal{H}_{l1}$ are given by $R_l^0(l,l') = \sum_{\nu=0}^{1} k_\nu(l)\, k_\nu(l')$ and $R_l^1(l,l') = k_2(l)\, k_2(l') - k_4([l-l'])$, respectively, where $[z]$ denotes the integer part of $z \in \mathbb{R}$. The reproducing kernel for the full marginal space $\mathcal{H}_l$ is simply the sum of the reproducing kernels for each of the subspaces:

$$R_l(l,l') = \sum_{\nu=0}^{1} k_\nu(l)\, k_\nu(l') + k_2(l)\, k_2(l') - k_4([l-l']).$$

One can also show that the reproducing kernels for $\mathcal{H}_{m0}$ and $\mathcal{H}_{m1}$ are given by $R_m^0(m,m') = 1$ and $R_m^1(m,m') = k_1(m)\, k_1(m') + k_2(m)\, k_2(m') - k_4([m-m'])$, and similarly, the reproducing kernel R_m for the full marginal space $\mathcal{H}_m$ is taken to be the sum of the kernels for each subspace. The tensor product reproducing kernel for $W_1[0,1] \otimes W_2[0,1]$ is obtained by simply taking the product of each of the reproducing kernels for the univariate function spaces:

$$R((l,l'),(m,m'))$$

Table 2 shows the decomposition of the reproducing kernel for the space of bivariate functions into the product of the reproducing kernels for each of the univariate subspaces.

Table 2: Tensor product reproducing kernel $R((l,m),(l',m'))$

	$\{1\}$	$\{k_1(l)\}$	$\{\mathcal{H}_{l1}\}$
$\{1\}$	$\{1\}$	$k_1(l)\, k_1(l')$	R_1^l
$\{\mathcal{H}_{m1}\}$	$R_m^1(m,m')$	$R_m^1(m,m')\, k_1(l)\, k_1(l')$	$R_m^1(m,m')\, R_l^1(l,l')$

For $\mathcal{H}$ defined as above, our goal is to find $\phi^* \in \mathcal{H}$ that minimizes

$$-2L + \lambda||P_1\phi^*||^2 = \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma(t_j)^{-2}\left(y_{ij} - \sum_{k=1}^{j-1}\phi^*(l_{ijk}, m_{ijk})\, y_{ik}\right)^2 + \lambda||P_1\phi^*||^2 \qquad (16)$$

where (l_{ijk}, m_{ijk}) are the result of transforming subject i's j and k^{th} observation times, t_{ij} and t_{ik}. Let B denote the $p \times M$ matrix with columns correponding to the M basis functions spanning $\mathcal{H}_0$ evaluated at the $p = \sum_i = 1^N \binom{n_i}{2}$ within-subject pairs of observed time points. In spirit of the result on the representation of a univariate $f\mathcal{H}$ given by citekimeldorf1971some, we can similarly show that the minimizer of 16 lies within a finite dimensional space despite the minimization being carried out over an infinite dimensional space, $\mathcal{H}$.

Theorem 4.1. *Let p denote the total number of unique within-subject pairs of design points, and index the transformed pairs $(l,m)_i$, $i = 1,\ldots,p$. Let B be the $p \times 2$ matrix with $(i,j)^{th}$ entry $k_j\left((l,m)_i\right)$ with rank $r = 2$. Then, the unique minimizer of the penalized likelihood (**??**), $\phi^* \in \mathcal{H}$ is of the form*

$$\phi^*(l,m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{p} c_i R^1\left((l,m),(l,m)_i\right) \tag{17}$$

The proof is left to the appendix.

4.1 Demonstration of the smoothing spline fitting procedure: Examples

5 Penalized B-splines for flexible smoothing

Estimating the varying coefficient function ϕ^* is quite different from the usual problem of estimating an arbitrary bivariate function via smoothing. In the case of the latter, we most typically treat both arguments equally in terms of regularization, but in the case of covariance estimation and the generalized coefficient function equal treatment of l and m in terms of penalization perhaps is not the most appropriate approach. The lag component, l, has particularly significant meaning in terms of the covariance function and thus also in terms of ϕ^* and is of considerable more interest than the orthogonal component, m. As discussed in Section 2, we can define an entire class of functional autoregressive models using only the l direction, and additionally, as discussed in Section 3, there is a natural expectation about the functional form of the autoregressive coefficient function (and hence covariance) as a function of l. The use of smoothing splines to estimate ϕ outlined in Section **??** yields smooth null models, but smoothness of the elements of the Cholesky factor alone may not lead to desirable structure in the inverse covariance matrix.

There has been a recent upsurge in both the theoretical analysis and study of the practical application of regularization procedures for large empirical covariance matrices, including Huang et al. [2006], Furrer and Bengtsson [2007], Fan et al. [2008], Ledoit and Wolf [2004]. Furrer and Bengtsson [2007] propose stabilizing the sample covariance matrix by "tapering," or gradually shrinking the off-diagonal elements to zero. d'Aspremont et al. [2008] propose a sparse estimator by applying an L_1 penalty directly to the elements of the covariance matrix. Instead of regularizing the covariance matrix itself, some have opted to regularize its inverse; Wu and Pourahmadi [2003] band the Cholesky decomposition by setting certain diagonals of the Cholesky factor to zero. Huang et al. [2006] and Levina et al. [2008] use L_1 penalties to achieve parsimony of the entries of the Cholesky factor; sparsity of the Cholesky factor, however, does not necessarily imply sparsity in the inverse covariance matrix. Levina et al. [2008] propose banding the Cholesky factor using a nested Lasso penalty which yields sparse estimators of the precision matrix.

These approaches implicitly adopt different notions of sparsity. Like Huang et al. [2006] and Levina et al. [2008], our aim is to regularize the inverse of the covariance matrix through the Cholesky factor.Pourahmadi [1999] presented a hueristic argument that the GARPs, $\phi_{t,t-l}$ should be monotonically decreasing in absolute value as l increases. The following proposition estab-

lishes the relationship between zeros in the Cholesky factor and zeroes in the inverse covariance, connecting the regularization of ϕ^* to the structure in the resulting inverse covariance matrix.

Proposition 5.1. *Let Ω denote a $m \times m$ positive definite matrix with elements ω_{ij} with modified Cholesky decomposition $T^T D^{-1} T$, where T is unit lower triangular. Let t_{ij} denote the ij^{th} element of T. For any column j and row $r(j) > j$, $\omega_{mj} = \cdots = \omega_{r(j),j} = 0$ if and only if $t_{mj} = \cdots = t_{r(j),j} = 0$.*

The proof is left to the appendix. Proposition 5.1 maintains that the modified Cholesky factor T with arbitrary column band lengths corresponds to inverse covariance matrix Ω with the same column band lengths, and hence the inverse covariance matrix is K-banded if and only if its Cholesky factor is K-banded. That is, if ϕ^* is zero or nearly zero for large $|i-j|$, then y_i and y_j are conditionally uncorrelated (or nearly so). This suggests that the effect of y_{t-l} on y_t through model 5 should decrease as the time between the two measurements increases, such that

$$\phi^*(l, m) \approx 0$$

for large l is a reasonable way to institute regularization.

Several have proposed regularization of the inverse covariance matrix by "banding" the Cholesky factor T: setting all elements of T beyond the K^{th} off-diagonal to zero, i.e. setting $\phi_{t,t-l} = 0$ for $l > K$ for some choice of K. (See Pourahmadi [1999], Wu and Pourahmadi [2003], Bickel and Levina [2008], and Huang et al. [2007].) In terms of model 1, this is equivalent to regressing y_t on only its K immediate predecessors, setting the regression coefficient for y_{t-l} to zero for $l > K$. This notion is instrumental in justifying a family of penalties which induces an alternative decomposition of the function space for which the null space of these penalties comprises functions ϕ^* taking nonzero values on a subset of the domain:

$$\mathcal{H}_0(\mathcal{B}) = \{\phi^* : \phi^*(l, m) = 0 \text{ for all } l > l_0\}, \tag{18}$$

which is equivalent to the set of functions having compact support:

$$supp(\phi^*) = (0, l_0].$$

The penalties having nullspace 18 differ significantly from the penalty discussed in Section 2.1 and the derivative-based penalties typically encountered in the smoothing spline ANOVA setting. For $\mathcal{H}_B$ to be non-empty, functions in $\mathcal{H}$ must be constructed from a basis having at least one or more elements with compact support. Consider a sequence of knots partitioning the unit interval $0 < \ell_1 < \ell_2 < \cdots < \ell_n < 1$; the truncated power functions (TPFs) of degree k, $\{T_{i,k}\}_{i=1}^n$, are given by

$$T_{i,k}^+(l) = (l - \ell_i)_+^k = \begin{cases} (l-\ell_i)^{k-1}, & l - \ell_i \geq 0 \\ 0 & l - \ell_i < 0 \end{cases}$$

Polynomial regression splines are continuous piecewise polynomials where the definition of the function changes at the collection of knot points. A polynomial of degree k has basis

$$\{1, l, \ldots, l^k, (l-\ell_1)_+^k, \ldots, (l-\ell_n)_+^k\}$$

A univariate function can be represented as a linear combination of these basis functions:

$$\phi_1 = \sum_{j=0}^{k} \beta_j l^j + \sum_{i=1}^{n} \beta_{k+i} T_{i,k}$$

The truncated power basis, as in their use in defining polynomial regression splines, enjoy a particular ease of interpretation. The coefficient β_{i+k} corresponds to the size of the jump at ℓ_i in the k^{th} derivative of ϕ_1. This fact is especially useful when tracking change points or in general, any abrupt changes in the regression curve, as the decomposition according to 18 demands. Rather than constructing a basis for $\mathcal{H}$, we instead consider representing the set of piecewise polynomial functions using B-splines, which exhibit numerical properties which are preferable to those of the TPFs. Let $\ell = \{\ell_i\}$ denote a non-decreasing sequence. The i^{th} B-spline of order k which corresponds to the knot sequence t is defined by

$$B_{i,k,t}(l) = (\ell_{i+k} - \ell_i)\,[\ell_i, \ldots, \ell_{i+k}]\,(\cdot - l)_+^{k-1}, \tag{19}$$

where $[x_i, \ldots, x_{i+k}]\, f(x)$ denotes the k^{th} order divided difference of the function f at $\{x_i, \ldots, x_{i+k}\}$. The placeholder notation, $(\cdot - l)_+^{k-1}$, is used to indicate that the k^{th} divided difference of the function $g(t) = (t-l)_+^{k-1}$ is obtained by fixing l 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. Using the properties of the divided difference, one can show that $B_i(x)$ has isolated support:

$$B_i(l) = 0, \quad l \notin [\ell_i, \ell_{i+k}]$$

As a result, for a set of B-splines of order k corresponding to the knot sequence ℓ, only k of them are nonzero on $[\ell_j, \ell_{j+k}]$: $B_{j-k+1}, B_{j-k+2}, \ldots, B_j$. See De Boor et al. [1978] for a complete review of B-splines and their properties. A B-spline function is a linear combination of B-spline basis functions:

$$\phi_1(l) = \sum_{i=1}^{n} \beta_{k+i} B_i(l) \tag{20}$$

permits the intuitive expression of a penalty to be used for banding the Cholesky factor. The B-spline basis functions are non-negative on their support; therefore, if a B-spline (as in 20) is zero for $l > l_0$, then the coefficients of the B-splines contributing to that region of the domain are also zero. For functions represented using order k B-spline basis functions $\{B_i\}$, $i = 1, \ldots, n$, one may truncate the function to zero at some truncation point $l_0 \in (0, 1)$ by penalizing the size of the coefficients corresponding to any basis functions having support on $l > l_0$. This naturally leads to penalties of the form

$$J_B = \sum_{i=i^*-k}^{n} |\beta_i|^p \tag{21}$$

where

$$i^* = \max_i \{\ell_i : \ell_i \leq l_0\}$$

is the index of the largest knot which is smaller than the truncation point l_0. Setting $p = 1$ in 21 corresponds to putting a LASSO penalty (see Tibshirani [1996]) on the coefficients contributing to the function value on $l > l_0$, while setting $p = 2$ corresponds to the usual ridge regression setting.

5.1 Regularization with difference penalties

Representing the varying coefficient function as a B-spline lays the foundation of a flexible framework for seamlessly instituting multiple types of regularization on the fitted function, and perhaps encouraging more than one notion of structure at once. As in Section **??**, one may wish for fitted functions to exhibit smoothness on their support. By using a rich B-spline basis along with a discrete difference penalty on the spline coefficients, 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 using a penalty to restrict the flexibility of the fitted curve. Like Wahba [1990] in the smoothing spline literature, he applied a penalty to the second derivative of the fitted curve:

$$J = \int_0^1 \left[\phi_1''(l)\right]^2 \, dl.$$

Using the properties of B-splines, for

$$\phi_1(l) = \sum_{j=1}^{n} \beta_j B_j(l),$$

one can derive a banded matrix P such that

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

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

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

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: 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 which circumvents 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 liberates us from 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.

This is of particular interest within the context of estimating the 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 is a good discrete approximation to the integrated square of the k^{th} derivative, so little is lost by replacing the derivative-based penalty with the following difference penalty:

$$J_d(\phi_1) = \sum_{j=d}^{n} \left(\Delta^d \beta_j\right)^2 \tag{22}$$

where $\beta = (\beta_1, \ldots, \beta_n)$. Let D_d denote the matrix difference operator: $D_d\beta = \Delta^d\beta$, where

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

In general,

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

Then, 22 can be written in terms of the squared norm of the difference operator applied to the vector of B-spline coefficients:

$$\begin{aligned} J_d(\phi_1) &= ||D_d\beta||^2 \\ &= \beta^T P_d \beta \end{aligned} \tag{23}$$

where $P_d = D_d^T D_d$.

By expressing the traditional penalty on the second derivative in terms of the B-spline coefficients, we can examine its connection to a penalty on second-order differences of the third-order B-spline coefficients:

$$\beta^T P \beta = \int_0^1 \left\{ \sum_{j=1}^{n} \beta_j B''_{j,3}(l) \right\}^2 dl. \tag{24}$$

Using the derivative properties of B-splines, this can be written as

$$\beta^T P \beta = \int_0^1 \left[\sum_{j=1}^{n} \sum_{k=1}^{n} \Delta^2\beta_j \Delta^2\beta_k B_{j,1}(l) B_{k,1}(l) \right] dl. \tag{25}$$

Most of the cross products of $B_{j,1}(l)$ and $B_{k,1}(l)$ vanish since B-splines of degree 1 only overlap

when j is $k-1$, k, or $k+1$. Thus, we have that

$$\begin{aligned}\beta^T P\beta &= \int_0^1 \left[\left\{ \sum_{j=1}^{n} \Delta^2\beta_j B_j\left(l,1\right) \right\}^2 + 2\sum_j \Delta^2\beta_j \Delta^2\beta_{j-1} B_j\left(l,1\right) B_{j-1}\left(l,1\right) \right] dl \\ &= \sum_{j=1}^{n} \left(\Delta^2\beta_j\right)^2 \int_0^1 B_j^2\left(l,1\right)\ dl \\ &\qquad + 2\sum_{j=1}^{n} \Delta^2\beta_j \Delta^2\beta_{j-1} \int_0^1 B_j\left(l,1\right) B_{j-1}\left(l,1\right)\ dl \end{aligned} \tag{26}$$

which can be written as

$$\beta^T P\beta = c_1 \sum_{j=2}^{n} \left(\Delta^2\beta_j\right)^2 + c_2 \sum_{j=3}^{n} \Delta^2\beta_j \Delta^2\beta_{j-1} \tag{27}$$

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

$$\begin{aligned} c_1 &= \int_0^1 B_{j,1}^2\left(l\right) dl \\ c_2 &= \int_0^1 B_{j,1}\left(l\right) B_{j-1,1}\left(l\right) dl. \end{aligned} \tag{28}$$

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 22 and the sum of the cross products of neighboring second differences. The second term in 27 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 27 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.

The following figures display the impact of the squared distance on adjacent basis coefficients on the function itself. When examining the relationship between P-spline curves and their coefficients, it is helpful to consider the coefficients as the skeleton of the function, then draping the B-splines over them to "put the flesh on the bones." A smoother sequence of coefficients leads to a smoother curve, which is clearly illustrated in Figure **??**. 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.

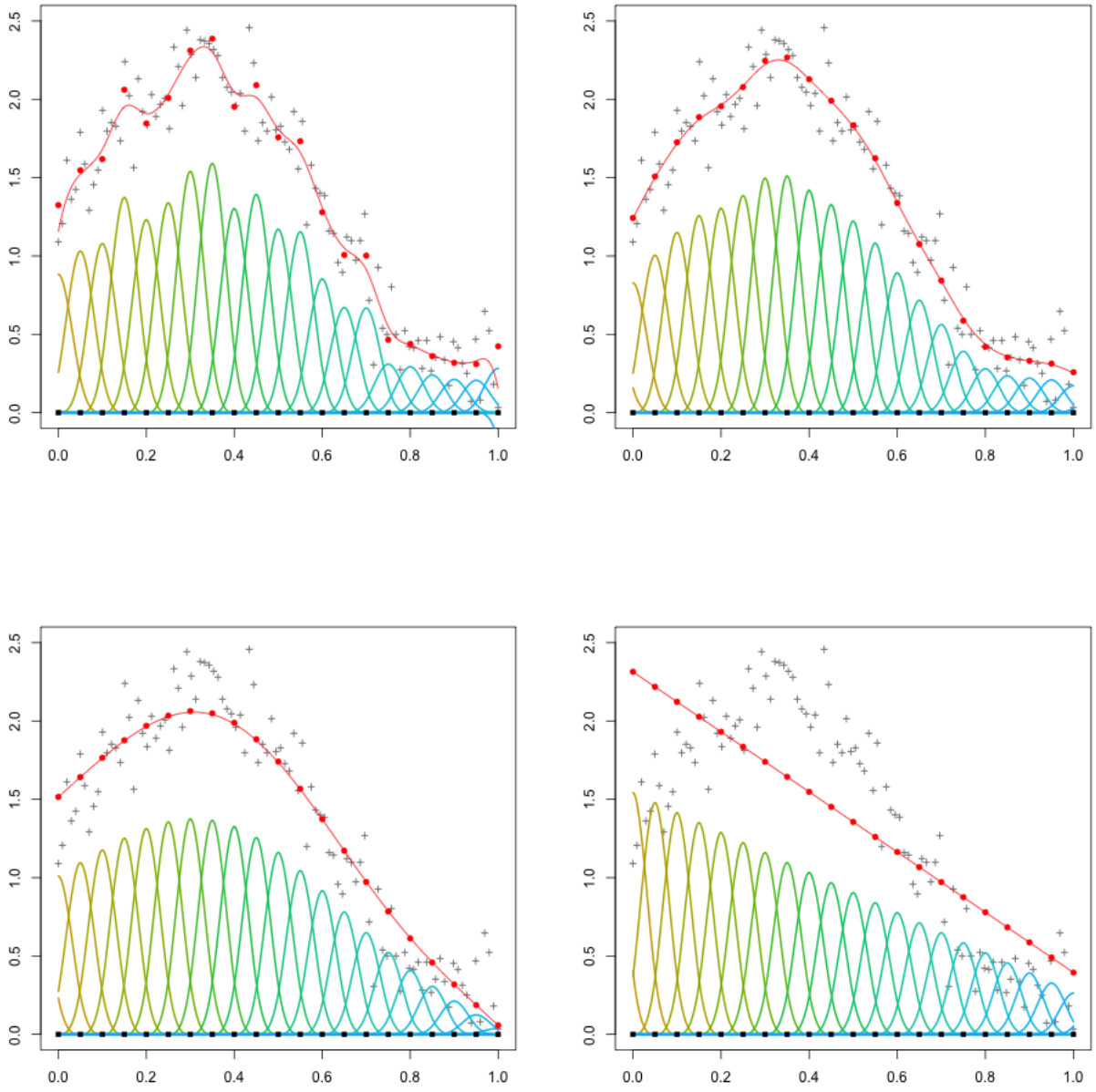


Figure 1: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The fitted curve in the upper left plot is the most "wiggly" of 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 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.

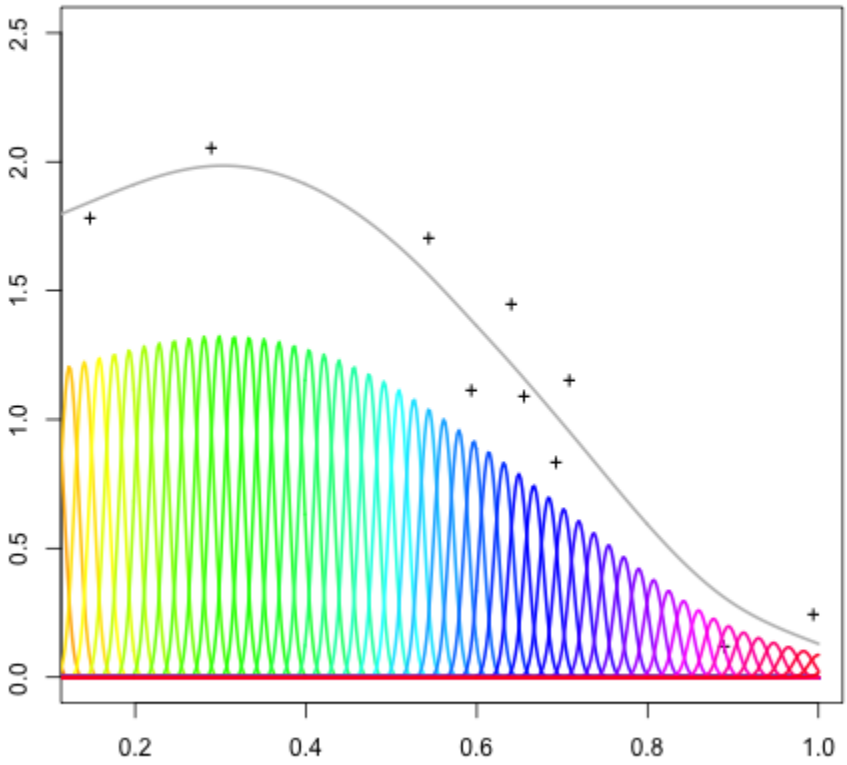


Figure 2: P-spline smoothing of 10 observations using 60 B-spline basis functions.

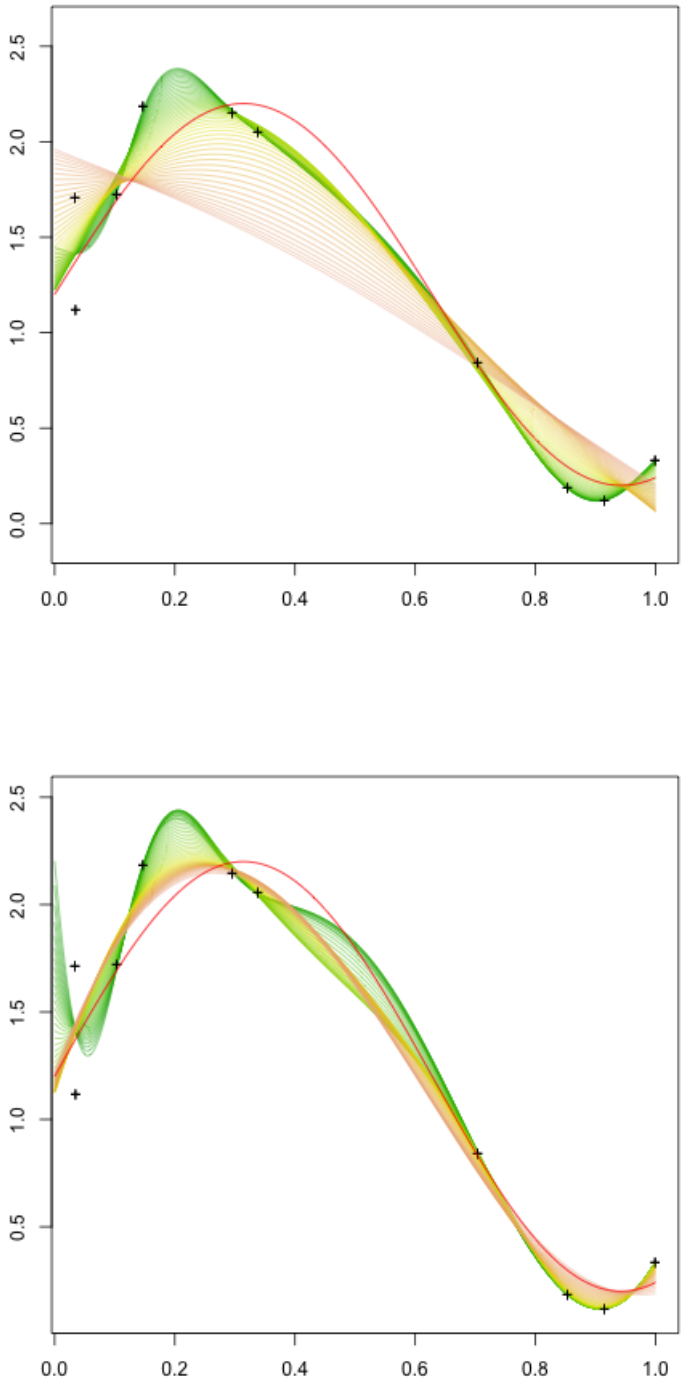


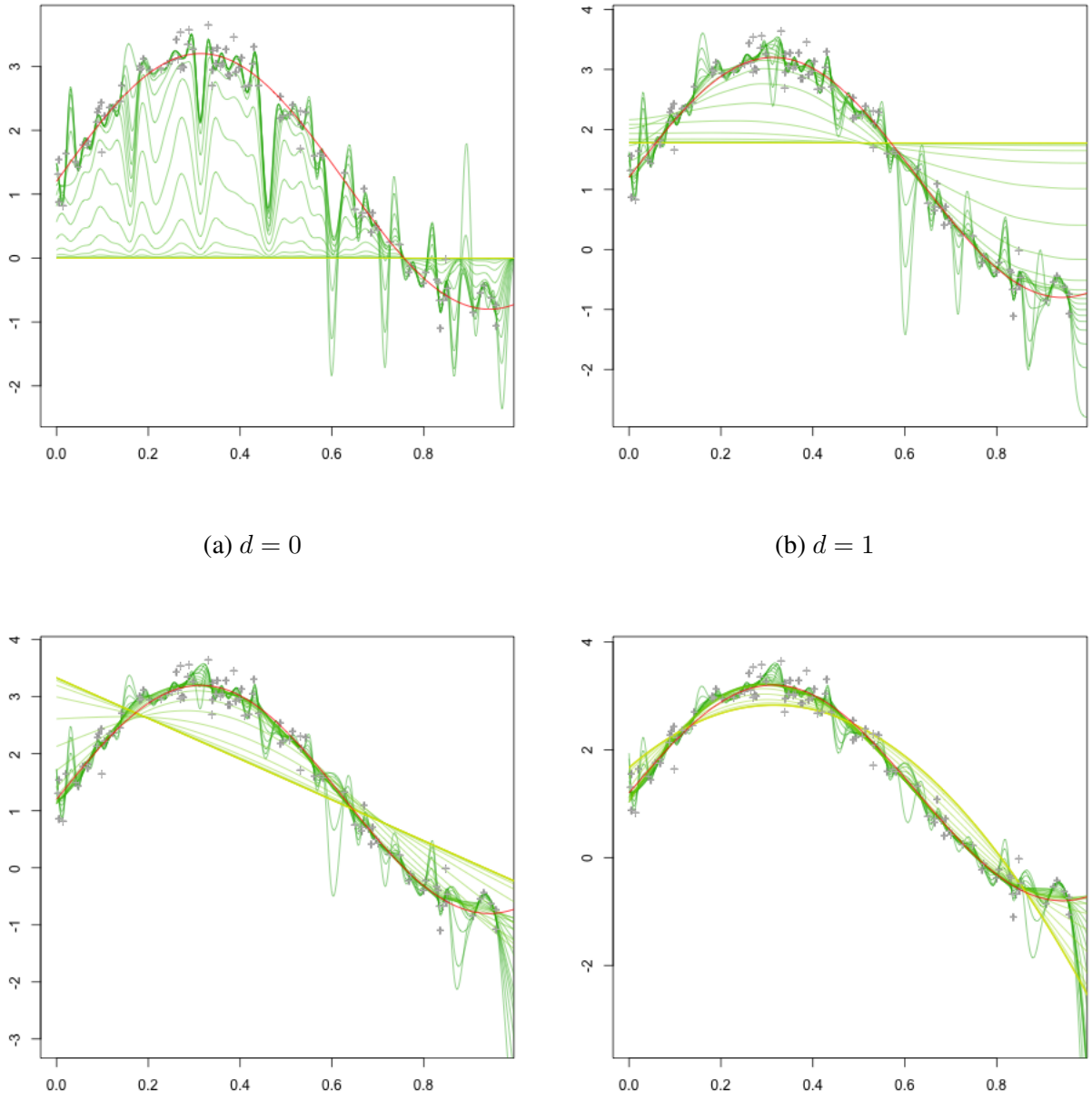
Figure 3: Fitted mean curves using a second (top) and third (bottom) order difference penalty for simulated data, sparsely sampled along the indexing variable: $y(t) = 1.2 + \sin(5t) + 0.2\epsilon_t$, where $\epsilon_t \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$. A total of 10 data points were fit using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

P-splines enjoy many advantageous properties, many of which are inherited from the attractive properties of B-splines. First, unlike many types of kernel smoothers, P-splines show no boundary effects i.e. the spreading of the fitted curve outside of the physical domain of the data, usually while also bending toward zero. See Eilers and Marx [1996] Section 8 for a detailed discussion.

P-splines can fit polynomial data exactly. Given data (l_i, y_i), if the y_i are a polynomial in l of degree k, then B-splines of degree k or higher will fit the data exactly. The same is true for P-splines if the order of the penalty is $k + 1$ or higher, irrespective of the value of λ. An informal proof is left to the appendix.

Under a difference penalty of order k, the fitted function will approach a polynomial of degree $k - 1$ for large values of λ as long as the degree of the B-splines is greater than or equal to k. This can be shown by once again considering the relationship between the derivatives of a B-spline fit and the differences in neighboring coefficients; detailed discussion is left to the appendix. Figure 4 visually demonstrates this property by examining the behavior of the fitted function as the tuning parameter varies under a difference penalty of order $d = 0, 1, 2, 3$.

We will leverage several properties of the smoothing matrix, H, for model evaluation and selection, so focus on the linearized smoothing problem solved at each value of the penalty parameter, λ, is prudent. The trace of H approaches the order of the differencing operator, k for larges values of λ. The trace of the smoothing matrix is a useful measure of effective model dimension, so understanding its limiting behavior is important.



(a) $d = 0$ (b) $d = 1$

(c) $d = 2$ (d) $d = 3$

Figure 4: *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* $\lambda \to \infty$*.*

5.2 Multidimensional smoothing with B-splines

P-splines can be extended naturally to higher dimensions by constructing a regression basis from the tensor product of the one-dimensional B-splines bases for each dimension. Figure 5 displays the building block of the foundation on which multidimensional P-splines is built: a B-spline tensor product basis $\left\{T_{kl}\right\}$ function. If we equip l and m each with a B-spline basis, we can a basis for the varying coefficient function ϕ in 8 by taking the tensor product of the two marginal bases. To regularize the fitted function, the only other modification necessary for the extension into two dimensions is the addition of a difference penalty for each variable l and m. Let $B_1(l), \ldots, B_K(l)$ 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 6 shows a thinned tensor product basis $\{T_{kl}\}$l; 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 6 is associated with an unknown coefficient α_{ij} which determines the height of the hill. For a given knot grid, we can approximate a surface by

$$\phi^*(l, m) = \sum_{i=1}^{K} \sum_{j=1}^{L} \alpha_{ij} B_i(l) B_j(m). \tag{29}$$

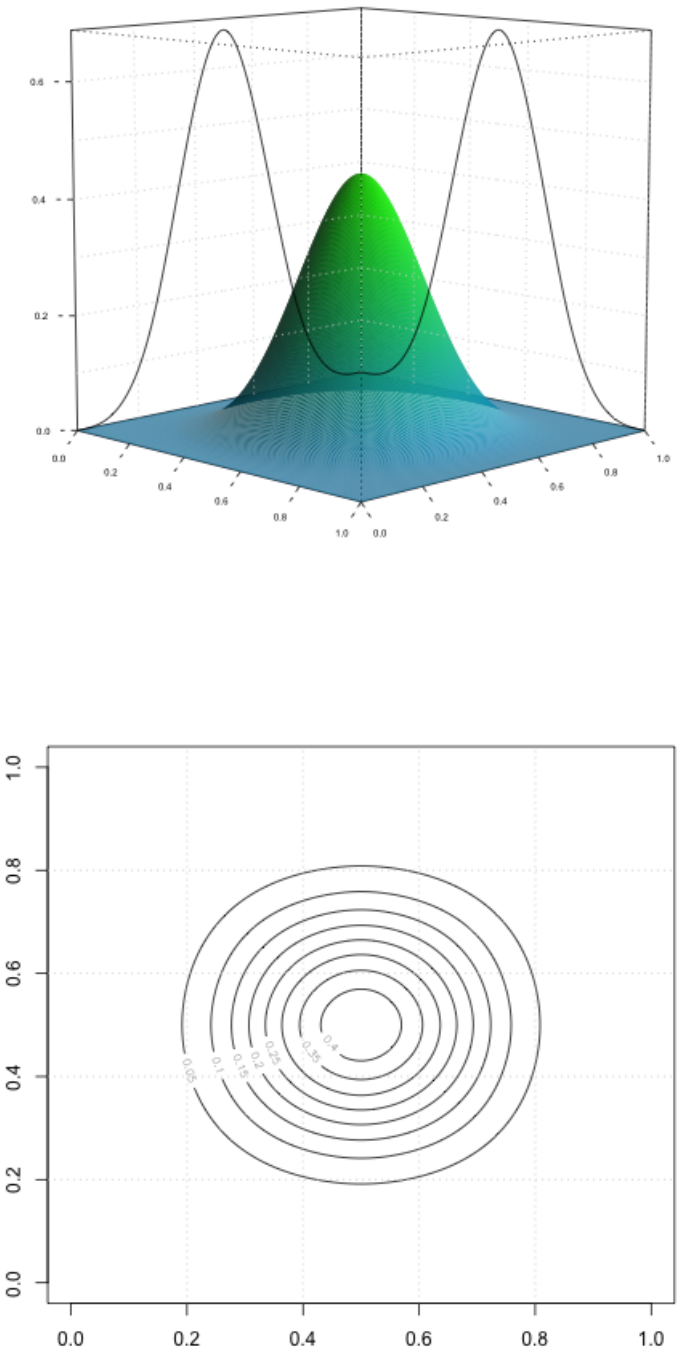


Figure 5: Tensor product of two cubic B-splines

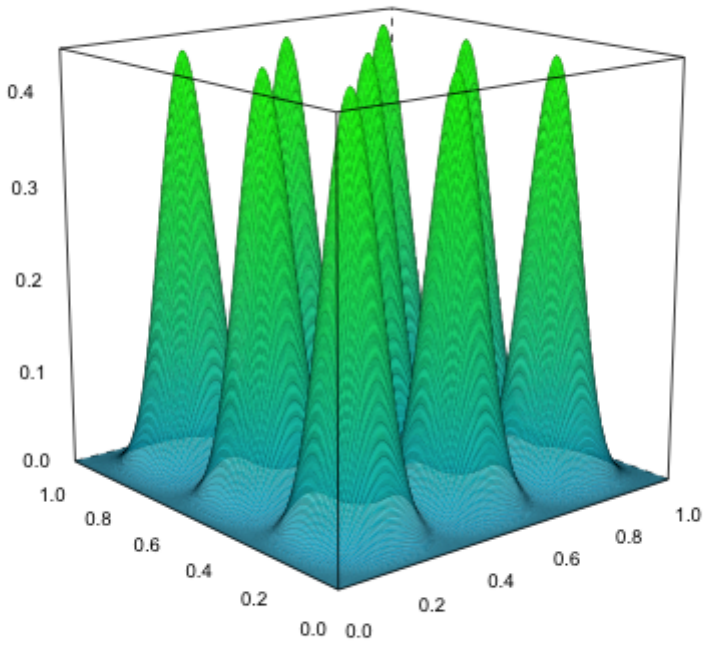


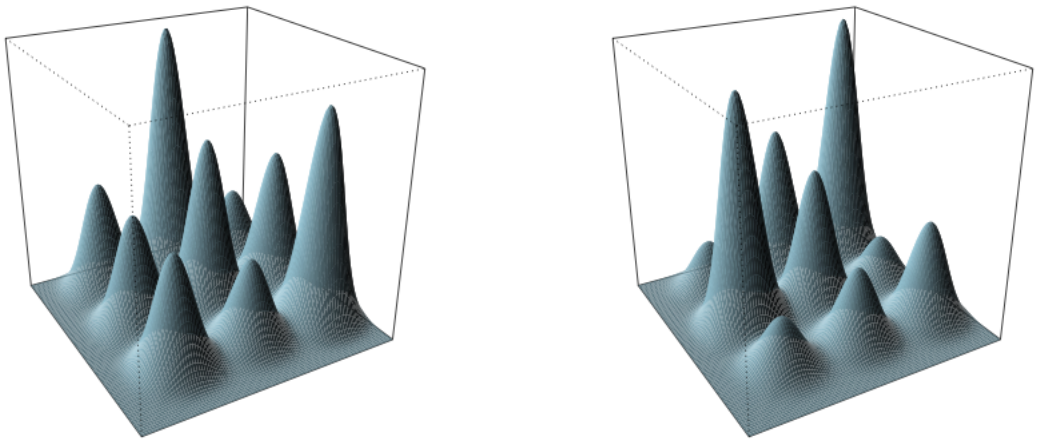
Figure 6: A subset of a full bivariate basis of cubic B-splines

Let p denote the total number of unique within-subject pairs of design points, and index the transformed coordinate pairs $(l, m)_i$, $i = 1, \ldots, p$. Let B_l and B_m denote the B-spline bases along the l and m coordinates of dimension $p \times K$ and $p \times L$, respectively, and let A be the unknown matrix of basis coefficients with elements $[\alpha_{ij}]$. To control the flexibility of the fitted function as with univariate P-splines discussed in Section 5.1, we can apply difference penalties to the rows and columns of A to smooth in the l and m directions respectively, using the penalty

$$\lambda_l \sum_k |D_d \alpha_{k\cdot}|^2 + \lambda_m \sum_l |D_{\bar{d}} \alpha_{\cdot l}|^2. \tag{30}$$

where $\alpha_{k\cdot}$ and $\alpha_{\cdot l}$ denote the k^{th} row and l^{th} column of A, respectively. The first term in30 imposes a difference penalty of order d_l on the rows of the coefficient matrix while the second term places a difference penalty (of possible different order d_m) on the columns. We give each direction its own smoothing parameter to permit anisotropic smoothing; however, one could opt to use a single smoothing parameter for both directions and dodge the added work of optimizing the amount of smoothing with two separate parameters. Figure **??** shows a potential result of heavy column

penalization (left) and heavy row penalization (right) under a second order difference penalty on each row and each column for large values of λ_l and λ_m. The figure demonstrates that the limiting behaviour of each row and column is linear, but the resulting surface may exhibit slope reversals from one row (column) to the next.



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

Figure 7: *Nine cubic B-spline tensor products with heavy linear column penalty and heavy linear row penalty*

We take the estimator of ϕ^* to be the minimizer of

$$Q(\alpha) = \sum_{i=1}^{N}\sum_{j=2}^{m_i} \sigma(t_j)^{-2} \left\{ y_{ij} - \sum_{k=1}^{j-1} \left(\sum_{r=1}^{L}\sum_{s=1}^{K} \alpha_{rs} B_r(l_{ijk}) B_s(m_{ijk}) \right) y_{ik} \right\}^2 + \lambda_l \sum_{r'=1}^{K} |D_{d_l} \alpha_{r'\cdot}|^2 + \lambda_m \sum_{s'=1}^{L} |D_{d_m} \alpha_{\cdot s'}|^2. \tag{31}$$

It is computationally advantageous to express the coefficient matrix in "unfolded" notation, which allows us to express the varying coefficient function at the observed coordinate grid as in the usual multiple regression form:

$$\text{vec}\{\phi^*(l,m)\} = B\alpha$$

Stacking the columns of A gives the vectorized coefficient matrix $\alpha = \text{vec}(A)$. The $p \times KL$ tensor product basis B is constructed from the tensor product of the marginal B-spline bases defined in Eilers et al. [2006] as the *row-wise Kronecker product* of the individual bases:

$$B = B_l \square B_m = \left(B_m \otimes 1_K^T\right) \odot \left(1_L^T \otimes B_l\right). \tag{32}$$

The operator $\odot$ denotes the element-wise matrix product; 1_K (1_L) denotes the column vector of ones having length K (L.) The operations in 32 construct B such that the i^{th} row of $B_m \square B_l$ is the Kronecker product of the corresponding rows of B_m and B_l. The penalty in 30 can also be compactly expressed:

$$\lambda_l |P_l \alpha|^2 + \lambda_m |P_m \alpha|^2 \tag{33}$$

where $P_l = I_L \otimes D_{d_l}^T D_{d_l}$ and $P_m = D_{d_m}^T D_{d_m} \otimes I_K$. We define the matrix W of historical regressors so that 31 can be written in matrix form as

$$Q(\alpha) = |Y - WB\alpha|^2 + \lambda_l |P_l \alpha|^2 + \lambda_m |P_m \alpha|^2, \tag{34}$$

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

$$\left[(WB)^T WB + \lambda_l P_l + \lambda_m P_m\right] \alpha = WBY \tag{35}$$

From 35, we note that the system of equations depends on basis coefficients remains fixed at KL, even as the number of observations increases. The grid of regression coefficients can be recovered by arranging the elements of $\hat{\alpha}$ into a matrix of L columns having length K. The effective hat matrix is given by

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

from which we can easy calculate deletion residuals and obtain a measure of effective model dimension. We will discuss diagnostics and approaches to model selection in detail in Section **??**.

6 P-spline ANOVA models for ϕ^*

An extension of the tensor product models in 5.2 for the varying coefficient function corresponding to the Cholesky factor T is the class of multidimensional models like in Section **??** based on the decomposition of multidimensional smooth functions as additive terms and interactions. Smoothing spline analysis-of-variance models have been previously proposed by Gu and Wahba [1993], Wahba et al. [1995], and Gu [2002], which gives main effects and interaction terms that can be interpreted as in the classical ANOVA setting. More recently, Lee and Durbán [2011] proposed the use of P-splines within a mixed modeling framework to estimate multidimensional functions as a decomposition into smooth main effects and interactions. They show that identifiability of functional components is ensured by imposing the same constraints as in a factorial design.

As in the smoothing spline formulation, we may decompose

$$\phi^*(l, m) = \mu + \phi_1(l) + \phi_2(m) + \phi_{12}(l, m),$$

where μ is a constant, ϕ_1 and ϕ_2 are the additive univariate functions of l and m, and ϕ_{12} is the two-dimensional interaction smooth function of (l, m). The smooth-ANOVA model is given by

$$E[Y] = WB\theta \tag{36}$$

where W is defined as in 34. The B-spline regression basis B is defined

$$B = [\ 1_p \mid B_l \mid B_m \mid B_{lm}\ ] \tag{37}$$

where B_{lm}, the basis for the l-m interaction term, is defined as in 32. The full regression basis has dimension $p \times (1 + K + L + KL)$; the vector of regression coefficients is given by $\theta = (\mu, \theta_l, \theta_m, \theta_{lm})^T$, where θ_l and θ_m are the $K \times 1$ and $L \times 1$ vectors of coefficients for the main effects and θ_{lm} is the $KL \times 1$ vector of coefficients for the smooth l-m interaction. The $(1 + K + L + KL) \times (1 + K + L + KL)$ penalty matrix is $blockdiag\{0, P_l, P_m, P_{lm}\}$

$$P = \begin{bmatrix} 0 & \cdots & & \\ \vdots & \lambda_l D_{d_l}^T D_{d_l} & & \\ & & \lambda_m D_{d_m}^T D_{d_m} & \\ & & & \tau_l \left(I_L \otimes D_{d_l}^T D_{d_l}\right) + \tau_m \left(D_{d_m}^T D_{d_m} \otimes I_K\right) \end{bmatrix}, \tag{38}$$

with the one-dimensional penalties for each of the additive terms (with corresponding smoothing parameters λ_l and λ_m) and the two-dimensional penalty (with two additional smoothing parameters τ_l and τ_m) for the interaction term making up the blocks. D_{d_l} and D_{d_m} are the difference matrices of order d_l and d_m respectively. With the new basis and penalty, the penalized likelihood 34 becomes

$$Q(\alpha) = (Y - WB\theta)^T (Y - WB\theta) + \theta^T P \theta. \tag{39}$$

The basis B, however, is not of full rank; some of the elements of B_l and B_m are also included in the interaction basis B_{lm}, so $1 + K + L$ columns of B are linearly dependent. This identifiability problem is also reflected in the rank deficiency of the penalty matrix; for penalty matrices $D_l^T D_l$ and $D_m^T D_m$ of order d_l and d_m respectively,

$$\begin{aligned} \text{rank}(P) &= \text{rank}(P_l) + \text{rank}(P_m) + \text{rank}(P_{lm}) \\ &= (K - d_l) + (L - d_m) + (K - d_l)(L - d_m) \end{aligned}$$

Model 36 must be modified to remove this redundancy and ensure identifiability of functional components. Wood [2017] (Chapter 4) proposed the use of the QR decomposition to identify linearly dependent columns of the regression basis numerically, and then removing them from the basis. Alternatively, Lee and Durbán [2011] proposed a mixed model framework that allows for efficient smoothing parameter estimation and mitigates identifiability issues when one imposes the same constraints as in factorial designs. Under the mixed model formulation , the smoothing parameter becomes a ratio of the residual variance and the variance of the random effects, so smoothing parameter selection becomes a matter of estimating variance components.

6.1 Mixed model representation of P-splines

Linear mixed effects models are an extension of the linear regression model to include random effects; the connection between mixed models and smoothing splines has been thoroughly explored. See ,, and , for example. The interest in this particular representation is because it allows for smoothing to be included in a large class of models. Additionally estimation and inference for mixed models is already well established, and the use of mixed model methodology and software well already well developed and widely adopted.

Applying the appropriate transformation to the regression basis B and the penalty P can ensure that the functional components of the smooth-ANOVA model 36 are identifiable. Lee and Durbán [2011] proposed such a transformation described by **?** which reparameterizes the model basis so that the PS-ANOVA model and its corresponding penalty can be represented as a mixed model:

$$Y = W\left(X\beta + Z\alpha\right) + \epsilon, \quad \alpha \sim \mathcal{N}\left(0, G\right), \;\; e \sim \mathcal{N}\left(0, \sigma_\epsilon^2 I\right), \tag{40}$$

where X and Z are the model matrices for the fixed and random effects. $G = \sigma_\alpha^2 \Delta$ holds the variance components of the random effects, α, where Δ is some positive definite matrix. The mixed model representation decomposes the varying coefficient function into an unpenalized part $X\beta$ and penalized smooth term $Z\alpha$. Under the assumption of normality and $i.i.d.$ errors, one can carry out estimation of β and α by restricted maximum likelihood. See **?** or **?** for detailed discussion of estimation of random effects models. To reparameterize 37, one can find an orthogonal transformation, M, such that

$$\begin{aligned} BM &= [\ X \mid Z\ ], \text{ and} \\ M\theta &= X\beta + Z\alpha, \end{aligned}$$

and $[\ X \mid Z\ ]$ has full rank; i.e.

$$\begin{aligned} BMM^T\theta &= [\ X \mid Z\ ]\, M^T\theta \\ &= [\ X \mid Z\ ]\, \mu \end{aligned}$$

where $\mu = \left(\beta^T, \alpha^T\right)^T$. This transformation converts the penalty $\theta^T P \theta$ to the form

$$\theta^T M^T P M \theta = \alpha^T F \alpha,$$

for some block diagonal matrix F. To demonstrate the construction of M, consider the two-dimensional tensor product model with basis given in 32. The transformation matrix is derived from the SVD of the penalty matrix 33 by simultaneously diagonalizing $D_l^T D_l$ and $D_m^T D_m$. Let

$$U_l \Delta_l U_l^T$$

denote the singular value decomposition of $D_l^T D_l$, where U_l is the $K \times K$ matrix of eigenvectors and Δ_l is the diagonal matrix with diagonal entries equal to the corresponding eigenvalues. Exactly d_l of of the eigenvalues are identically zero, so if we let $\tilde{\Delta}_l$ denote the submatrix containing only the positive eigenvalues, then we can write

$$\Delta_l = \text{diag}\left(0_{d_l}, \tilde{\Delta}_l\right).$$

Let

$$U_{l0} = \begin{bmatrix} 1_l^* & u_{l1}^* ... & u_{l,d_l}^* \end{bmatrix}$$

denote the matrix of eigenvectors corresponding to the zero eigenvalues, where $1_l^* = (1, 1, \ldots, 1)^T / \sqrt{K}$ and u_{l1i}^* is the vector having elements $(1^i, 2^i, \ldots, K^i)^T$, centered and scaled to have mean 0 and unit length. Let U_{l1} denote the submatrix of eigenvectors corresponding to the nonzero eigenvalues, $\tilde{\Delta}_l$. Define U_{m0}, U_{m1}, and $\tilde{\Delta}_m$ in the same fashion.

The orthogonal transformation matrix M is defined $M = \left[\; M_0 \; \middle| \; M_1 \; \right]$, where

$$\begin{aligned} M_0 &= \left[\; U_{m0} \otimes U_{l0} \; \right], \text{ and} \\ M_1 &= \left[\; U_{m0} \otimes U_{l1} \; \middle| \; U_{m1} \otimes U_{l0} \; \middle| \; U_{m1} \otimes U_{l1} \; \right]. \end{aligned}$$

The transformed coefficients are given by

$$\beta = M_0^T \theta, \qquad \alpha = M_1^T \theta$$

Using properties of the kronecker product (see Liu [1999]), one can show that the mixed model design matrices X and Z can be written as the tensor product of the marginal bases:

$$\begin{aligned} X &= \left[\; X_m \square X_l \; \right], \text{ and} \\ Z &= \left[\; X_m \square Z_l \; \middle| \; Z_m \square X_l \; \middle| \; Z_m \square Z_l \; \right]. \end{aligned} \tag{41}$$

where

$$\begin{aligned} X_l &= B_l U_{l0}, & Z_l &= B_l U_{l1}, \\ X_m &= B_m U_{m0}, & Z_m &= B_m U_{m1}. \end{aligned}$$

This is convenient because the calculation of M can be avoided in practice. Since β is not penalized, we can replace X_l and X_m with

$$\left[\; 1_p \; \middle| \; l \; \middle| \; \ldots \; \middle| \; l^{d_l - 1} \; \right] \text{ and } \left[\; 1_p \; \middle| \; m \; \middle| \; \ldots \; \middle| \; m^{d_m - 1} \; \right]$$

Making this replacement, the expressions for X and Z can be expanded:

$$X = \left[\; 1_p \; \middle| \; l \; \middle| \; m \; \middle| \; l \square m \; \middle| \; \ldots \; \middle| \; l^{d_l - 1} \square m^{d_m - 1} \; \right], \tag{42}$$

$$Z = \left[\; Z_l \; \middle| \; Z_m \; \middle| \; Z_m \square l \; \middle| \; m \square Z_1 \; \middle| \; \ldots \; \middle| \; Z_m \square l^{d_l - 1} \; \middle| \; m^{d_m - 1} \square Z_1 \; \right], \tag{43}$$

Construction of the fixed effects and mixed effects in this way allows us to express the fitted varying coefficient function ϕ^* as a sum of the main effects and an l-m interaction. After applying the transformation, the penalty becomes

$$F = \begin{bmatrix} 0 & & & \\ & \lambda_l I_L \otimes \tilde{\Delta}_l & & \\ & & \lambda_m \tilde{\Delta}_m \otimes & \\ & & & \lambda_m \tilde{\Delta}_m \otimes + \lambda_l I_L \otimes \tilde{\Delta}_l \end{bmatrix}, \tag{44}$$

where the 0 element of the block diagonal is a matrix of zeros corresponding to the unpenalized fixed effects. The variance components of the mixed model are given by $G = \sigma_\epsilon^2 F^{-1}$. Applying these results to the smooth-ANOVA model 36 with regression basis 37,

7 Appendix

Proof of Theorem 4.1 For simplicity of presentation, relabel the elements of $\mathcal{W}^*$ so that

$$\mathcal{W}^* = \{(l_1, m_1), (l_2, m_2), \ldots, \left(l_{N_{\phi^*}}, m_{N_{\phi^*}}\right)\}.$$

Then we may verify that any $\phi^* \in \mathcal{H}$ can be written

$$\phi^*(l, m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{n} c_i R_1((l, m), (l_i, m_i)) + \rho(l, m)$$

where $\rho \perp \mathcal{H}_0 = \{1\} \oplus \{k_1\},\ span\{R_1((l_i, m_i), \cdot)\}$. We do so by demonstrating that ρ does not improve the first term in (**??**) (the data fit functional) and only adds to the penalty term, $J(\phi^*)$. Consequently, if $\hat{\phi}^*$ is the minimizer of (**??**), then $\rho = 0$. Using the properties of reproducing kernels, we can rewrite ϕ^* as an inner product of itself with R:

$$
\begin{aligned}
\phi^{*}\left(l_j, m_j\right) &= \left\langle R\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \phi^{*}(\cdot,\cdot)\right\rangle \\
&= \Big\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right) + R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot) \\
&\qquad\qquad + \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right) + \rho\left((\cdot,\cdot)\right)\Big\rangle \\
&= \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot)\right\rangle + \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right)\right\rangle \\
&\quad + \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho\left((\cdot,\cdot)\right)\right\rangle + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot)\right\rangle \\
&\quad + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right)\right\rangle + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho\left((\cdot,\cdot)\right)\right\rangle \\
&= \left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), d_0 + d_1 k_1(\cdot)\right\rangle + \left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right)\right\rangle \\
&\quad + \underbrace{\left\langle R_0\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho(\cdot,\cdot)\right\rangle}_{0} + \underbrace{\left\langle R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right), \rho(\cdot,\cdot)\right\rangle}_{0} \\
&= d_0 + d_1 k_1(\cdot) + \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), \left(l_j, m_j\right)\right)
\end{aligned}
$$

Rewriting the data fit functional, we have that

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y\left(t_{ij}\right) - \sum_{k=1}^{j-1} \phi^{*}\left(t_{ij}, t_{ik}\right) y\left(t_{ik}\right)\right)^2 \\
= &\sum_{i=1}^{N}\sum_{j=1}^{n_i} \sigma_{ij}^{-2}\left(y\left(t_{ij}\right) - \sum_{k=1}^{j-1} \left\langle R\left(\left(l_{jk}^{i}, m_{jk}^{i}\right), (\cdot,\cdot)\right), \phi^{*}(\cdot,\cdot)\right\rangle y\left(t_{ik}\right)\right)^2
\end{aligned}
$$

which is free of ρ. Consider the contribution of any nonzero ρ to $J\left(\phi^{*}\right)$:

$$
\begin{aligned}
J\left(\phi^{*}\right) &= ||P_1 \phi^{*}||^2 \\
&= \left\langle \sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right) + \rho(\cdot,\cdot), \sum_{j=1}^{N_{\phi^*}} c_j R_1\left(\left(l_j, m_j\right), (\cdot,\cdot)\right) + \rho(\cdot,\cdot)\right\rangle \\
&= ||\sum_{i=1}^{N_{\phi^*}} c_i R_1\left(\left(l_i, m_i\right), (\cdot,\cdot)\right)||^2 + ||\rho||^2
\end{aligned}
$$

Thus, including ρ in ϕ^{*} only increases the penalty without improving (decreasing) the data fit

functional, so we indeed have that the minimizer of (**??**) has the form

$$\phi^* (l, m) = d_0 + d_1 k_1 (l) + \sum_{i=1}^{N_{\phi^*}} c_i R_1 \left((l, m), (l_i, m_i) \right) \tag{45}$$

Proof: Proposition 5.1

Proof: Using the expression

$$\sigma^{ij} = \sum_{k=i}^{p} d_{ii} t_{ki} t_{kj}$$

it follows immediately that $t_{pj} = \cdots = t_{r(j),j} = 0$ implies that $\sigma^{pj} = \cdots = \sigma^{r(j),j} = 0$.

From citewatkins2004fundamentals, we can show that we can sequentially derive the elements of T and D according to

$$d_{ii} = \sqrt{\sigma^{ii} - \sum_{k=1}^{i-1} t_{ki}^2}$$

$$t_{ij} = \frac{1}{d_{ii}} \left(\sigma^{ij} - \sum_{k=1}^{i-1} t_{ki} t_{kj} \right)$$

We proceed by induction. For the first row of T^T,

References

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

P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31(4):377–403, 1978.

A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, 30(1):56–66, 2008.

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

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

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

J. Fan and J.-T. Zhang. Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62 (2):303–322, 2000.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147(1):186–197, 2008.

R. Furrer and T. Bengtsson. Estimation of high-dimensional prior and posterior covariance matrices in kalman filter variants. *Journal of Multivariate Analysis*, 98(2):227–255, 2007.

C. Gu. Smoothing spline anova models, 2002.

C. Gu and G. Wahba. Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 353–368, 1993.

J. Z. Huang, C. O. Wu, and L. Zhou. Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128, 2002.

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

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

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

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

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

S. Liu. Matrix results on the khatri-rao and tracy-singh products. *Linear Algebra and its Applications*, 289(1-3):267–277, 1999.

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

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

M. Pourahmadi. Cholesky decompositions and estimation of a covariance matrix: orthogonality of variance-correlation parameters. *Biometrika*, pages 1006–1013, 2007.

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

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

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

G. Wahba, Y. Wang, C. Gu, R. Klein, and B. 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.

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

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