Nonparametric Covariance Estimation for Longitudinal Data via Penalized Tensor Product Splines

Tayler A. Blake* Yoonkyung Lee†

May 17, 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

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

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

(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 citetpourahmadi2011covariance 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. The autoregressive coefficients are often referred to in the literature as the *generalized autoregressive parameters*, or *GARPs*, and the error variances are often called the *innovation variances*, or *IVs*.

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.

citetchen2011efficient, citetpourahmadi1999joint, and citetpourahmadi2002dynamic 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.

citethuang2007estimation follow the hueristic argument presented in citetpourahmadi1999joint 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 softly enforce monotonicity in l by penalizing order restriction as in the work of citettibshirani2011nearly.

The rest of the paper is organized as follows: Section 2 summarizes the general penalized estimation approach and introduces the transformed design coordinates and penalties for stationarity and non-monotonicity. Section 3 presents a detailed discussion of optimization and computational issues. Section 4 presents a simulation study and a real example to examine the performance of our methods as well as others. Section 5 concludes with discussion and future work.

2 Covariance estimation: a review

Parametric models are frequently used for modeling covariance structure in the longitudinal data setting. Simple models which depend on a small number of parameters are commonly found in the literature such as those corresponding to compound symmetry and autoregressive models of order

k, where k is small. However, model misspecification can lead to considerably biased estimates. Alternately, several have proposed applying nonparametric methods directly to elements of the sample covariance matrix or a function of the sample covariance matrix. Diggle and Verbyla (1998) introduced a nonparametric estimator obtained by kernel smoothing the sample variogram and squared residuals. Yao, Mueller, and Wang 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. [REVIEW 2009 WU AND POURAHMADI METHOD: banding the sample covariance matrix. Under the assumption of short range dependency, they show that their estimator converges to the true covariance matrix for a broad class of nonlinear processes.] The estimates yielded by these approaches, however, are not guaranteed to be positive definite.

To satisfy the positive-definiteness constraint, methods have been developed and applied to certain reparameterizations of the covariance structure. Chiu, Leonard, and Tsui modeled the matrix logarithm of the covariance matrix. Early nonparametric work using the spectral decomposition of the covariance matrix included that of Rice and Silverman (1991) which discussed smoothing and smoothing parameter choice for eigenfunction estimation for regularly-spaced data. Staniswalis and Lee (1998) extended kernel-based smoothing of eigenfunctions to functional data observed on irregular grids. However, when the data are sparse in the sense that there are few repeated within-subject measurements and measurement times are quite different from subject-to-subject, approximation of the functional principal component scores defined by the Karhunen-Loeve expansion of the stochastic process by usual integration is unsatisfactory and requires numerical quadrature. Many have explored regression-based approaches using the Spectral decomposition, framing principal components analysis as a least-squares optimization problem. Among many others, Zou, Hastie and Tibshirani (2006) imposed penalties on regression coefficients to induce sparse loadings. [REVIEW THE METHODS OF HUANG, KAUFMAN, YAO HERE]

Leveraging regression techniques for covariance estimation has recently received much attention. [OMIT THIS; INCLUDE BRIEF SUMMARIES OF THEIR TECHNIQUES BELOW. including citetbickel2008regularized and citethuang2006covariance] have proposed nonparametric estimators of a specific covariance matrix (or its inverse) rather than the parameters of a covariance function.

[DISCUSS THE CHOLESKY PARAMETERIZATION OF THE INVERSE COVARIANCE MATRIX. REFERENCE THE EARLY INFLUENTIAL WORKS]

Recently, many have considered a modified Cholesky decomposition (MCD) of the inverse of the covariance matrix. This decomposition also ensures positive-definite covariance estimates, and, unlike the Spectral decomposition whose parameters follow an orthogonality constraint, the entries in the MCD of the covariance matrix are unconstrained and have an attractive statistical interpretation as particular regression coefficients and variances. One drawback we might note, however, is that the interpretation of the regression model induced by the MCD assumes a

[DISCUSS THE CHOLESKY PARAMETERIZATION OF THE INVERSE COVARIANCE MATRIX. REFERENCE THE Liao, Park, Hannig, and Kang (2015) paper, discuss their estimators and the competitors they investigated.]

citetyao2005functional do not utilize the Cholesky parameterization, and their estimates are

not guaranteed to be positive definite. We combine the advantages of bivariate smoothing as in citetyao2005functional with the added utility of the Cholesky parameterization in citethuang2007estimation; in doing so, we present a flexible and coherent approach to covariance estimation, while simultaneously we ensuring positive definiteness of estimates.Rather than shrinking element of the Cholesky factor to zero after a particular value of l, we choose to softly enforce monotonicity in l by using a hinge penalty as in the work of citettibshirani2011nearly.

3 Covariance Estimation via Bivariate Smoothing

Assume that we have measurements on individual i, denoted $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{i,n_i})^T$, at times $t_{i1} < t_{i2} < \cdots < t_{i,n_i}$, with $i = 1, 2, \ldots, N$. Observed time points may be individual-specific and not necessarily on a regular grid. To present a comprehensive overview our estimation procedure, we begin with the representation of the inverse covariance matrix, Σ^{-1}, in terms of its Cholesky decomposition (see citetpourahmadi2007cholesky for a detailed discussion). Decomposing the precision matrix in such a way allows for both an unconstrained parameterization and statistically meaningful interpretation of covariance parameters. For any positive definite matrix Σ, there exists a unique unit lower triangular matrix T with diagonal entries equal to 1 which diagonalizes Σ:

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

The entries of T and D are easily interpretable if we consider regressing y_{ij} on its predecessors:

$$y_{ij} = \sum_{k=1}^{j-1} \phi_{ijk} y_{ik} + \sigma_{ij}\epsilon_{ij} \tag{1}$$

for $j = 2, \ldots, n_i$; we define $y_{i1} = \epsilon_{i1}$. Standard regression theory gives us that if $\{\phi_{ijk}\}$ are the coefficients of the linear least squares predictor of y_{ij} based on its predecessors, then the prediction residuals $\boldsymbol{e}_i = (e_{i1}, e_{i2}, \ldots, e_{i,n_i})^T$ have diagonal covariance. Let T_i be the unit lower triangular matrix with jk^{th} below-diagonal entry given by $-\phi_{ijk}$. Let $\boldsymbol{Y}_i$ denote the random vector giving rise to observed data $\boldsymbol{y}_i$, and let $\boldsymbol{\epsilon}_i$ denote the associated vector of random errors. Then we may write the model (**??**) as follows:

$$\boldsymbol{\epsilon}_i = T_i \boldsymbol{Y}_i \tag{2}$$

We assume $\boldsymbol{Y}_i$ centered to have mean $\mathbf{0}$ with covariance matrix Σ_i. Let D_i be the diagonal matrix with $\{\sigma_{ij}\}$ down the diagonal, and taking covariances on both sides of (**??**),

$$D_i = T_i \Sigma_i T_i^T$$

and immediately, we have that $\Sigma_i^{-1} = T_i^T D_i^{-1} T_i$. The regression coefficients $\{\phi_{ijk}\}$ are referred to as the *generalized autoregressive parameters* (GARPs), and the $\{\sigma_{ij}\}$ are referred to as the *innovation variances* (IVs.)

Rather than a vector of longitudinal data points, we view the random vectors $\boldsymbol{Y}_i$ and $\boldsymbol{\epsilon}_i$ as discrete renditions of the stochastic processes: $Y(t)$ and $\epsilon(t)$. We assume $Y(t)$ has corresponding covariance function $G(s,t)$ and that $\epsilon(s)$ follows a zero mean Gaussian white noise process with unit variance. It is reasonable to assume that if $\boldsymbol{Y}$ is reasonably well-behaved, then $G(s,t)$ satisfies some smoothness conditions, where smoothness is defined in terms of square integrability of certain derivatives. We view the entries of Σ_i as values of G evaluated at the distinct pairs of observed time points on individual i.

Additionally, we treat the elements of the precision matrix Σ_i^{-1} as the values of the smooth function, $\gamma(s,t)$ evaluated at observed time points. If we consider the Cholesky decomposition of Σ^{-1}, it is natural to extent the same notion to the elements of T_i and D_i: view the GARPs $\{\phi_{ijk}\}$ and innovation variances as the evaluation of the smooth functions $\phi(s,t)$ and $\sigma^2(t)$ at observed time points and interpret $\phi_{ijk} = \phi(t_{ij}, t_{ik})$ and $\sigma_{ij}^2 = \sigma^2(t_{ij})$.

Analogous to Pourahmadi's model (**??**), we model the continuous time process as follows:

$$y(t_{ij}) = \sum_{k=1}^{j-1} \phi(t_{ij}, t_{ik})\, y(t_{ik}) + \sigma(t_{ij})\, \epsilon(t_i) \quad i = 1, \ldots, N, \tag{3}$$

It is advantageous to estimate the smooth function $\gamma(s,t)$ rather than a covariance matrix at a predetermined set of pairs of observed time points since observed time points may be unevenly spaced and vary from individual to individual. Several approaches to function estimation have been utilized in this setting; citewu2003nonparametric, for example, used locally weighted polynomials to smooth down the sub-diagonals of T. citehuang2007estimation smoothed the sub-diagonals of T using univariate smoothing splines. Within our formulation, the task of estimating a covariance matrix is equivalent to estimating the function $\phi(s,t)$ using bivariate smoothing. For ease of exposition, we assume that $\sigma^2(t)$ is fixed and known. Like other nonparametric situations, we make no assumption about the functional form of ϕ other than that ϕ is smooth, with smoothness defined in terms of square integrability of certain derivatives and let ϕ belong to a reproducing kernel Hilbert space, $\mathcal{H}$.

Pooling the observed time points across subjects, we let $\mathcal{T}$ denote the set of all unique observed time points $\mathcal{T} = \bigcup_{i=1}^{N} \bigcup_{j=1}^{n_i} \{t_{ij}\}$ and order them so that the elements of this set are given by $t_1 < t_2 < \cdots < t_p$, $||\mathcal{T}|| = p$. Let $\boldsymbol{Y} = \left(Y_{t_1}, Y_{t_2}, \ldots, Y_{t_p}\right)^T$ denote the vector of random variables corresponding to the process Y at each of the unique

of pooled observations, and let $Cov(\boldsymbol{Y}) = \Sigma$ where the ij^{th} element of the Σ is given by $\Sigma_{ij} = Cov\left(y_{t_i}, y_{t_j}\right)$.

γ is defined through ϕ and σ, which we also assume to be smooth functions.

$$y_{t_i} = \sum_{j=1}^{i-1} \phi_{t_i t_j} y_{t_j} + \sigma_{t_j} \epsilon_{t_j} \tag{4}$$

where $\boldsymbol{\phi}_{t_i} = \left(\phi_{t_i t_1}, \phi_{t_i t_2}, \ldots, \phi_{t_i, t_{i-1}}\right)^T$ is the coefficient vector corresponding to the best linear predictor of y_{t_i} based on its predecessors; Let T be the $p \times p$ lower triangular matrix with unit

diagonal and $(ij)^{th}$ element $-\phi_{ij}$, $i > j$ and D be the diagonal matrix with diagonal entries $\sigma_1^2, \sigma_2^2, \ldots, \sigma_p^2$. Using this notation, we may write
and letting $\mathrm{L} = T^{-1}$, the modified Cholesky decomposition of Σ^{-1} and Σ are given by

$$\Sigma^{-1} = T^T D^{-1} T, \quad \Sigma = \mathrm{L} D \mathrm{L}^T$$

Analogous to Pourahmadi's model (**??**), we model the continuous time process as follows:

$$y(t_i) = \sum_{j=1}^{i-1} \phi(t_i, t_j)\, y(t_j) + \sigma(t_i)\, \epsilon(t_i) \qquad i = 1, \ldots, n \tag{5}$$

The task of estimating a covariance matrix becomes the task of estimating the bivariate function $\phi(s,t)$ given noisy, discrete, and possibly unevenly spaced observations $Y_i = (Y(t_{i1}), Y(t_{i2}), \ldots, Y(t_{in_i}))^T$, $i = 1, \ldots, N$. The entries of the covariance matrix are viewed as the evaluation of this bivariate function at the unique observed pairs of time points. Like other nonparametric situations, we make no assumption about the functional form of ϕ other than that ϕ is smooth, with smoothness defined in terms of square integrability of certain derivatives.

Along with citethuang2006covariance, citetlevina2008sparse, and citetpourahmadi2000maximum we consider the normal log-likelihood as a loss function, though it is important to note that the derivation of the Cholesky decomposition did not rely on any distributional assumption on $\boldsymbol{\epsilon}$. Under the Gaussian assumption on $\epsilon(t)$, the negative log-likelihood of the data $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$ up to a constant is given by

$$-2L(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N, \Phi) = \sum_{i=1}^{N} \sum_{j=2}^{n_i} \sigma(t_j)^{-2} \left(y(t_{ij}) - \sum_{k=1}^{j-1} \phi(t_{ij}, t_{ik})\, y(t_{ik}) \right)^2 \tag{6}$$

We define our estimator $\hat{\phi}(s,t)$ to be the minimizer of the penalized log-likelihood.

$$\hat{\phi} = \arg\min_{\phi} \left(-2L + \lambda_1 J_1(\phi) \right) \tag{7}$$

The first term in (**??**) discourages the lack of fit of ϕ to the data; J_1 is a penalty functional, and λ_1 is the smoothing parameter which controls the tradeoff between the lack of fit and amount of regularization imposed on $\hat{\phi}$ through J_1. J_1 denotes the penalty assigned to the amount of "non-stationarity" in $\phi(s,t)$, or rather, any functional component that cannot be described in terms of the difference between the two argument values, $s - t$, $s \geq t$.

3.1 Parsimonious precision structures

Many have specified parsimonious parametric models for ϕ_{ijk} to overcome the issue of dimensionality. A commonly utilized approach in previous work is to model $\phi_{ijk} = z_{ijk}^T \gamma$ where z_{ijk} is a vector of powers of time differences and γ is a vector of unknown "dependence" parameters to be estimated. citetchen2011efficient, citetlin2009robust, citetpan2003modelling, and citetpourahmadi1999joint let

$$z_{ijk}^T = \left(1, t_{ij} - t_{ik}, (t_{ij} - t_{ik})^2, \ldots, (t_{ij} - t_{ik})^{q-1}\right) \tag{8}$$

Modeling the covariance in such a way is reduces a potentially high dimensional problem to something much more computationally feasible; if we model 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 lag only. Modeling ϕ^* in such a way is equivalent to specifying a Toeplitz structure for Σ. A $p \times p$ Toeplitz matrix M is a matrix with elements m_{ij} such that $m_{ij} = m_{|i-j|}$ i.e. a matrix of the form

$$M = \begin{bmatrix} m_0 & m_1 & m_2 & \dots & m_{p-1} \\ m_1 & m_0 & m_1 & \dots & m_{p-2} \\ m_2 & m_1 & m_0 & \dots & m_{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_{p-1} & m_{p-2} & m_{p-3} & \dots & m_0 \end{bmatrix} \tag{9}$$

To shrink ϕ toward the continuous analogue of these models (**??**), it is useful to consider transforming the pairs of time points and estimating the re-parameterized coefficient function. The rotated pairs of time points become $l = s - t$, $m = \frac{1}{2}(s + t)$; re-expressing ϕ in terms of these new arguments, our goal is to estimate

$$\phi^*(l, m) = \phi^*\left(s - t, \frac{1}{2}(s + t)\right) = \phi(s, t) \tag{10}$$

When ϕ^* corresponds to the simple models of the form (**??**), the bivariate function may be written in terms of only its first argument. Writing $\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m$ as a tensor product of two Hilbert spaces for each of l and m, the ANOVA decomposition of $\phi^*(l, m)$ as presented in citetgu2002smoothing is given by

$$\phi^*(l, m) = \mu^* + \phi_1^*(l) + \phi_2^*(m) + \phi_{12}^*(l, m) \tag{11}$$

We let $\mathcal{H}_l = \mathcal{H}_m = W_2(0, 1)$ where W_2 denotes the second-order Sobolev space:

$$W_2(0, 1) = \{f : \quad f, f' \text{absolutely continuous}, \int_0^1 \left(f^{(2)}\right)^2 dt < \infty\}$$

The penalty functional J_1 induces a decomposition of $\mathcal{H}$ as follows: $\mathcal{H}_l = \mathcal{H}_l^0 \oplus \mathcal{H}_l^1$ and $\mathcal{H}_m = \mathcal{H}_m^0 \oplus \mathcal{H}_m^1$ where let $\mathcal{H}_l^0 = \{1\} \oplus \{k_1\}$, $\mathcal{H}_m^0 = \{1\}$, and where $\{k_r\}$ denotes the subspace spanned by k_r. $\mathcal{H}_l^1$ and $\mathcal{H}_m^1$ are the subspaces orthogonal to $\mathcal{H}_l^0$ and $\mathcal{H}_m^0$, respectively:

$$\mathcal{H}_l^1 = \{\phi_1^* : \int_0^1 \phi_1^{*(\nu)}(l)\, dl = 0, \quad \nu = 0, 1\}$$

$$\mathcal{H}_m^1 = \{\phi_2^* : \int_0^1 \phi_2^*(m)\, dm = 0\}$$

Using the properties of tensor product spaces, we may decompose $\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1$ where

$$\begin{aligned} \mathcal{H}^0 &= \{1\} \oplus \{k_1\} \\ \mathcal{H}^1 &= \mathcal{H}_l^1 \oplus \mathcal{H}_m^1 \oplus \left[\{k_1\} \otimes \mathcal{H}_m^1\right] \oplus \left[\mathcal{H}_l^1 \otimes \mathcal{H}_m^1\right] \end{aligned}$$

We may write the penalty functional in terms of the projection of $\phi^* \in \mathcal{H}$ onto the penalized space of functions, $\mathcal{H}_1$:

$$\begin{aligned} \lambda_1 J_1(\phi) &= \lambda_1 \left(||P_1\phi_1^*||^2 + ||P_1\phi_2^*||^2 + ||P_1\phi_{12}^*||^2 \right) && (12) \\ &= \lambda_1 \left(||\phi_1^{*\prime\prime}||^2 + ||\phi_2^*||^2 + ||\phi_{12}^*||^2 \right) && (13) \end{aligned}$$

$P_1\phi^*$ denotes the projection of $\phi^* \in \mathcal{H}$ onto $\mathcal{H}_1$. To find the solution $\hat{\phi}^*$ which is the stage-wise minimizer of (**??**): we first set $\lambda_2 = 0$ and find $\tilde{\phi}^*$ which minimizes (**??**):

$$-2L + \lambda_1 J_1(\phi^*) = \sum_{i=1}^{N} \sum_{j=2}^{p_i} \sigma(t_j)^{-2} \left(y(t_{ij}) - \sum_{k=1}^{j-1} \phi(t_{ij}, t_{ik})\, y \right)^2 + \lambda_1 \left(||P_1\phi_1^*||^2 + ||P_1\phi_2^*||^2 + ||P_1\phi_{12}^*||^2 \right) \tag{14}$$

Let $\tilde{\lambda}_1$ denote the value of λ_1 corresponding to the minimizer $\tilde{\phi}^*$. Observe that large values of λ_1 (i.e. when $\lambda_1 \to \infty$) forces a parametric model in the null space of $J_1(\phi^*) = \{\phi^* : \phi^*(l,m) = \alpha + \beta l\}$, the set of ϕ^* to which J_1 assigns zero penalty. By construction, these nulls models correspond to modeling ϕ^* as linear functions of lag, so that as $\lambda_1 \to \infty$, the minimizer of (**??**) corresponds to those models previously proposed (**??**) for the case where $q = 2$.

Before we discuss the precise details of finding the minimizer of (**??**), we introduce some notation: let $\mathcal{H}$ be endowed with inner product $\langle f, g \rangle$, and define the reproducing kernel for $\mathcal{H}_1$

$$\begin{aligned} R_l^1(l, l') &= k_2(l)\, k_2(l') - k_4([l - l']) \\ R_m^1(m, m') &= k_1(m)\, k_1(m') + k_2(m)\, k_2(m') - k_4([m - m']) \\ R^1((l,m),(l',m')) &= \left[k_1(l)\, k_1(l') + R_l^1(l, l') \right] R_m^1(m, m') \end{aligned} \tag{15}$$

where $k_\nu = B_\nu/\nu!$ are scaled Bernoulli polynomials satisfying $B_0(x) = 1$, $\frac{d}{dx} B_j(x) = j B_{j-1}(x)$, and where $[\alpha]$ is the fractional part of α. Then we have the following result:

Theorem 3.1. *Let $p = \sum_{i=1}^{N} \binom{n_i}{2}$ be the total number of distinct 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 $k_j((l,m)_i)$ with rank $r = 2$. Then, the unique minimizer of the penalized likelihood* (**??**)*, $\phi^* \in \mathcal{H}$ is of the form*

$$\phi^*(l,m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{p} c_i R^1((l,m),(l,m)_i) \tag{16}$$

Proof: 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:

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

Rewriting the data fit functional, we have that

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

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

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

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

$$\phi^*(l,m) = d_0 + d_1 k_1(l) + \sum_{i=1}^{N_{\phi^*}} c_i R_1((l,m),(l_i,m_i)) \tag{17}$$

This result can be seen as a special case of the representer theorem given in citekimeldorf1971some and gives us that the minimizer ϕ^* lies within a finite dimensional space despite the optimization being carried out over an infinite dimensional space, $\mathcal{H}$.

3.2 Computation: Shrinkage toward Toeplitz precision structures

4 Shrinkage toward banded inverse structures

The regularization instituted by J_1 determines the complexity of those functional components which extent the stationary models, or rather the set of models that may be written in terms of an overall mean and the main effect of l only. Several other ways of imposing structural simplicity have been explored in previous works; specifically, citepourahmadi1999joint was one of the first to present a hueristic argument that the GARPs, $\phi_{t,t-l}$ should be monotonically decreasing in l. That is, the effect of y_{t-l} on y_t through the autoregressive parameterization should decrease as the distance in time between the two measurements increases. They and others (see citebickel2008regularized, citehuang2007estimation,citelevina2008sparse) enforce this structure by setting $\phi_{t,t-l} = 0$ for $l > K$, or equivalently, setting all off-diagonals of T beyond the K^{th} off-diagonal to 0, where K is chosen using a model selection criterion such as AIC or BIC (see citetwu2003nonparametric.) This regularization is equivalent to banding the Cholesky factor T, or rather, regressing y_t as in (**??**) on only its K immediate predecessors, setting $\phi_{ijk} = 0$ for $j - k > K$. We may show that banding T to only its first K off-diagonals is equivalent to banding Σ^{-1} to its first K off-diagonals, and is a reasonable approach to imposing parsimony in the inverse covariance matrix, as off-diagonal zeros imply conditional independence between y_t and y_{t-l} given the intermediate observations under the assumption of Gaussian likelihood for any $l > K$. For this to become immediate, we need the following two propositions:

Proposition 4.1. *Let $Y = (Y_1, \ldots, Y_n)^T$ have joint distribution with mean vector μ and covariance matrix Σ. The elements of $\Sigma^{-1} = (\sigma^{ij})_{1 \le i,j \le n}$ may be interpreted as partial covariances between the elements of Y.*

Proof: We may easily derive the covariance between two measurements Y_j and Y_k conditional on the remaining measurements $\{Y_l | l \ni j, k\}$
which leads to the corresponding partition of $\Sigma = Cov(Y)$:

Conditional distributions are easily obtained for the multivariate normal distribution; in particular, the covariance corresponding to the joint distribution Y_j and Y_k conditional on $\{Y_{il} | l \neq j, k\}$ is given by

Noting that the inverse of a matrix partitioned as in (**??**) is given by

$$\Sigma^{-1} \equiv \left[\begin{array}{c|c} \Sigma^{11} & \Sigma^{12} \\ \hline \Sigma^{21} & \Sigma^{22} \end{array}\right] = \left[\begin{array}{c|c} \Sigma_{11} - \Sigma_{12}^T \Sigma_{22}^T \Sigma_{12} & -\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right) \\ \hline -\Sigma_{22}^{-1}\Sigma_{21}\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) & \Sigma_{22} - \Sigma_{21}^T \Sigma_{11}^T \Sigma_{21} \end{array}\right] \tag{18}$$

We may quickly see that the covariance of Y_j and Y_k conditional on $\{Y_{il} | l \neq j, k\}$ (**??**) corresponds to the upper left 2×2 sub-matrix of Σ^{-1}, Σ^{11}.

Proposition 4.2. *For any $p \times p$ positive definite matrix Σ^{-1} and modified Cholesky decomposition $\Sigma^{-1} = T^T D^{-1} T$ where T is unit lower triangular, for any column j and $r(j) > j$, $\sigma^{pj} = \cdots = \sigma^{r(j),j} = 0$ if and only if $t_{pj} = \cdots = t_{r(j),j} = 0$ where $\{\sigma^{ij}\}$ and $\{t_{ij}\}$ denote the entries of Σ^{-1} and T, respectively.*

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,

Regularization via banding the inverse covariance matrix is a generalization of tapering; citecai2010optimal define the k-tapered estimator of a $p \times p$ covariance matrix Σ by $\tilde{\Sigma}^{(k)} = \left(\tilde{\sigma}_{ij}^{(k)}\right)_{1 \leq i,j \leq p} = \left(w_{ij}^{(k)} \hat{\sigma}_{ij}\right)_{1 \leq i,j \leq p}$ where $\hat{\Sigma}$ denotes the MLE for Σ and where, for tapering parameter k,

$$w_{ij}^{(k)} = \begin{cases} 1, & |i-j| \leq k/2 \\ 2 - \frac{|i-j|}{k/2}, & k/2 < |i-j| < k \\ 0, & otherwise \end{cases}$$

The k-banded estimator of Σ may be defined in the same manner by specifying

$$\tilde{\sigma}_{ij}^{B(k)} = \hat{\sigma}_{ij} I\left(|i-j| \leq k\right)$$

Rather than tapering the covariance matrix itself, citecai2012estimating have generalized banding the inverse covariance matrix as in citebickel2008regularized, citehuang2007estimation, and

citelevina2008sparse by proposing k-tapered estimators of Σ^{-1}.Rather than tapering elements of a covariance matrix given a specific set of observed time points, citekaufman2008covariance employ tapering to regularize a parametric estimator of a smooth covariance function, defining the tapered covariance function as follows:

$$K_1(x, \theta, \gamma) = K_0(, \theta) K_{taper}(x, \gamma)$$

where x is the distance between two observations, K_0 is the original covariance function which is assumed to be known up to a parameter vector $\theta \in \mathcal{R}^d$, and where K_{taper} is an isotropic correlation function such that $K_{taper}(x, \gamma) = 0$ for $x \geq \gamma$. Rather than tapering the covariance matrix or a covariance function, citecai2012estimating have generalized banding the inverse covariance matrix as in citebickel2008regularized, citehuang2007estimation, and citelevina2008sparse by proposing k-tapered estimators of Σ^{-1}.

Thus, shrinking the Cholesky factor toward a unit diagonal structure is clearly a natural and desirable way to impose sparsity in the estimated inverse covariance matrix. We consider the form of the models belonging to the null space of J_1:

$$\begin{aligned} \mathcal{H}_0 &= \{\phi^* | \phi^* = \mu^* + \phi_1^*(l);\ \phi_l^*(l) = \beta_0 + \beta_1 l\} \\ &= \{\phi^* | \phi^* = d_0 + k_1(l)\} \end{aligned}$$

To further impose simplicity in the structure of the inverse covariance, we consider "banding" the functional components of these of these null models; specifically, if we consider any ϕ^* corresponding to a Toeplitz precision matrix, that is any ϕ^* of the form

$$\phi^*(l, m) = \mu^* + \phi_1^*(l) \tag{19}$$

we propose truncating the stationary functional components: the overall mean and the main effect of l to zero for any $l > l_0$ for some truncation point $l_0 \in (0, 1)$. We consider the class of penalty functions that can be written in terms of an L_p norm of the sum of the overall mean and the functional main effect of l. We follow in the work of citehuang2006covariance and consider penalties which may be written

$$J_{2,(p)} = \sum_{l_i \in \mathcal{L}: l_i > l_0} |\mu^* + \phi_1^*(l_i)|^p \tag{20}$$

where $\mathcal{L}$ denotes the observed values of l, so that any ϕ^* to which J_2 assigns zero penalty is one that inherits nonzero contribution from stationary functional components only for lags $l \leq l_0$. We focus our attention to two important members of this family of penalties: the L_2 penalty and the L_1 penalty, given by

$$\begin{aligned} J_{2,(2)} &= \sum_{l_i \in \mathcal{L}: l_i > l_0} (\mu^* + \phi_1^*(l_i))^2 \quad \text{and} \\ J_{2,(1)} &= \sum_{l_i \in \mathcal{L}: l_i > l_0} |\mu^* + \phi_1^*(l_i)| \end{aligned} \tag{21}$$

respectively. These penalties will induce shrinkage in the autoregressive coefficient function ϕ^* (and hence in the overall inverse covariance function) as in ridge regression and LASSO, respectively. Considering both types of regularization introduced in this section and in the previous, any ϕ^* belonging to the set of models incurring zero penalty from both J_1 and J_2 may be written

$$\phi^*(l,m) = \begin{cases} d_0 + d_1 k_1(l), & l \leq l_0 \\ 0, & l > l_0 \end{cases}$$

We impose this further regularization from a stage-wise approach and define $\hat{\phi}^*$ to be the minimizer of

$$\begin{aligned} & -2L \quad + \hat{\lambda}_1 J_1(\phi^*) + \lambda_2 J_2(\phi^*) && (22) \\ = \; & -2L \quad + \hat{\lambda}_1 J_1(\phi^*) + \sum_{l_i \in \mathcal{L}: l_i > l_0} |\mu^* + \phi_1^*(l_i)| && (23) \end{aligned}$$

where $\hat{\lambda}_1$ is the optimal choice of tuning parameter value as determined by some model selection criterion; thorough discussion is reserved for Section four.

Models chosen with tuning parameter selection determined using GCV have a desirable number of properties (see citewahba1990spline for detailed discussion.)

Minimizing

$$-2L + \hat{\lambda}_1 J_1(\phi^*) + \lambda_2 \sum_{l_i \in \mathcal{L}: l_i > L} |\mu^* + \phi_1^*(l_i)| \tag{24}$$

We introduce two sets of non-negative variables $\{\eta_{+,i}\}$, $\{\eta_{-,i}\}$ so that minimizing (**??**) is equivalent to minimizing

$$-2L + \hat{\lambda}_1 J_1(\phi^*) + \lambda_2 \sum_{l_i \in \mathcal{L}: l_i > l_0} (\eta_{+,i} + \eta_{-,i}) \tag{25}$$

subject to the constraints

$$\begin{aligned} \eta_{+,i},\ \eta_{-,i} &\geq 0 \\ \mu^* + \phi^*(l_i) &\leq \eta_{i,+} \\ -(\mu^* + \phi^*(l_i)) &\leq \eta_{i,-} \end{aligned}$$

for $i = 1, \ldots, ||\mathcal{L}|| - l_0$. Note that the solution $\hat{\phi}^*$, which depends on both λ_1 and λ_2, does not minimize the loss

$$-2L + \lambda_1 J_1(\phi^*) + \lambda_2 J_2(\phi^*)$$

Rather than simultaneously minimizing over the joint parameter space for λ_1 and λ_2: $\mathbb{R}^+ \times \mathbb{R}^+$, we simplify the optimization problem by sequentially minimizing over the parameter space for

λ_1 and, given the optimal value for $\lambda_1 = \hat{\lambda}_1$, the parameter space for λ_2. The choice of l_0 also determines the degree to which we truncate the inverse covariance function, or if we consider the precision matrix, which may be viewed as the discretized analogue to the smooth function, l_0 determines the degree to which we shrink the precision matrix toward a diagonal structure.

The effect of the tuning parameters λ_1 and λ_2 in addition to l_0 is most easily described if we consider the null models produced from this combination of penalty functions, or rather the form of the solution $\hat{\phi}^*$ as both λ_1 and λ_2 tend to infinity. First, if we let $\lambda_1 \to \infty$, then $\hat{\phi}^*$ may be written as a linear function of l. The precision matrix we may view as the discretized analogue of the corresponding smooth inverse covariance function in this case will be constant down each sub-diagonal, taking form (**??**). Further, letting $\lambda_2 \to \infty$, ϕ^* is truncated to zero for any lag $l > L$, and elements of all subdiagonals of the corresponding precision matrix which lie further than l_0 from the diagonal are set to zero. Notice that if we let $l_0 = \max_i\{l_i\}$, the penalty J_2 is trivial, and no further structure is imposed on the solution $\hat{\phi}^*$ and the second stage of regularization requires no additional computational expense.

4.1 Computation: shrinkage toward banded Toeplitz precision structures via the L_1 penalty and constrained optimization

In this section, we outline the derivation of the dual optimization problem for the optimal autoregressive coefficient function under the L_1 penalty (**??**), though thorough explanation of the theory behind this derivation is omitted to allow for affable discussion.

Define the vector Φ^*_{1,J_2} with elements given by the functional main effect of l, ϕ^*_1 evaluated at the penalized observed values of l, that is, $l_i \in \mathcal{L}$ such that $l_i > l_0$. Let B_l and K_l denote the $(||\mathcal{L}|| - l_0) \times 2$ and $(||\mathcal{L}|| - l_0) \times p$ matrices of basis functions and representers defined such that

$$\mu^* \mathbf{1} + \Phi^*_{1,J_2} = B_l d + K_l c$$

Using the standard machinery of primal-dual formulations in constrained optimization theory, four sets of Lagrange multipliers are introduced for the constraints (**??**): $\alpha_{+,i}$, $\alpha_{-,i}$, $\gamma_{+,i}$, and $\gamma_{+,i}$, $i = 1, \ldots, ||\mathcal{L}|| - l_0$ for $\mu^* + \phi^*(l_i) \leq \eta_{i,+}$, $-(\mu^* + \phi^*(l_i)) \leq \eta_{i,-}$, $\eta_{i,+} \geq 0$, and $\eta_{i,-} \geq 0$, respectively. The Lagrangian primal function is given by

$$\begin{aligned} \mathrm{l}_{\mathcal{P}} = (Y_{(-1)} - Z\Phi)^T D^{-1} (Y_{(-1)} - Z\Phi) \quad &+ \quad \hat{\lambda}_1 c^T K c + \lambda_2 \mathbf{1}^T (\eta_+ + \eta_-) - \boldsymbol{\alpha}_+^T \eta_+ - \boldsymbol{\alpha}_-^T \eta \\ &- \quad \boldsymbol{\gamma}_+^T (\eta_+ - (B_l d + K_l c)) - \boldsymbol{\gamma}_-^T (\eta_- + (B_l d + K_l c)) \end{aligned} \tag{26}$$

where

$$\begin{aligned} \eta_+ &= \left(\eta_{+,1}, \eta_{+,2}, \ldots, \eta_{+,||\mathcal{L}||-l_0}\right)^T \\ \eta_- &= \left(\eta_{-,1}, \eta_{-,2}, \ldots, \eta_{-,||\mathcal{L}||-l_0}\right)^T \\ \boldsymbol{\alpha}_+ &= \left(\alpha_{+,1}, \alpha_{+,2}, \ldots, \alpha_{+,||\mathcal{L}||-l_0}\right)^T \\ \boldsymbol{\alpha}_- &= \left(\alpha_{-,1}, \alpha_{-,2}, \ldots, \alpha_{-,||\mathcal{L}||-l_0}\right)^T \end{aligned}$$

Define

$$\boldsymbol{a} = \begin{bmatrix} d \\ c \end{bmatrix}, \quad H = \begin{bmatrix} B^T Z^T D^{-1} Z B & \mathbf{0} \\ \mathbf{0} & \hat{\lambda}_1 K + K Z^T D^{-1} Z K \end{bmatrix}, \quad \boldsymbol{b} = -2 Y_{(-1)}^T \begin{bmatrix} ZB & \mathbf{0} \\ \mathbf{0} & ZK \end{bmatrix}$$

Then the Lagrangian primal function may be written

$$\begin{aligned} \mathrm{l}_{\mathcal{P}} = -\boldsymbol{b}^T \boldsymbol{a} + \frac{1}{2} \boldsymbol{a}^T H \boldsymbol{a} + \lambda_2 \mathbf{1}^T (\eta_+ + \eta_-) \quad &- \quad \boldsymbol{\alpha}_+^T \eta_+ - \boldsymbol{\alpha}_-^T \eta_- \\ &- \quad \boldsymbol{\gamma}_+^T (\eta_+ - (B_l d + K_l c)) - \boldsymbol{\gamma}_-^T (\eta_- + (B_l d + K_l c)) \end{aligned} \tag{27}$$

with the following constraints

$$\begin{aligned} \frac{\partial \mathrm{l}_{\mathcal{P}}}{\partial \boldsymbol{a}} &= -\boldsymbol{b} + H\boldsymbol{a} + [B_l \; K_L]^T \left(\boldsymbol{\gamma}_+ - \boldsymbol{\gamma}_-\right) = 0 \Leftrightarrow H^{-1} \left[\boldsymbol{b} - [B_l \; K_L]^T \left(\boldsymbol{\gamma}_+ - \boldsymbol{\gamma}_-\right)\right] \\ \frac{\partial \eta_+}{\partial \boldsymbol{a}} &= \lambda_2 \mathbf{1} - \boldsymbol{\alpha}_+ - \boldsymbol{\gamma}_+ = 0 \Leftrightarrow \boldsymbol{\gamma}_+ = \lambda_2 \mathbf{1} - \boldsymbol{\alpha}_+ \\ \frac{\partial \eta_-}{\partial \boldsymbol{a}} &= \lambda_2 \mathbf{1} - \boldsymbol{\alpha}_- - \boldsymbol{\gamma}_- = 0 \Leftrightarrow \boldsymbol{\gamma}_- = \lambda_2 \mathbf{1} - \boldsymbol{\alpha}_- \\ \alpha_{+,i}, \quad \alpha_{-,i} &\geq 0 \text{ and } \gamma_{+,i},\ \gamma_{-,i} \geq 0 \text{ for } i = 1, \ldots, ||\mathcal{L}|| - l_0 \end{aligned}$$

Simplifying the $\mathrm{l}_{\mathcal{P}}$ using the constraints, we have the dual problem of maximizing

$$-\frac{1}{2} \left[\boldsymbol{b} - [B_l \; K_l] (\boldsymbol{\alpha}_- - \boldsymbol{\alpha}_+)\right]^T H^{-1} \left[\boldsymbol{b} - [B_l \; K_l] (\boldsymbol{\alpha}_- - \boldsymbol{\alpha}_+)\right] \tag{28}$$

$$\propto \boldsymbol{b}^T H^{-1} [B_l \; K_l]^T D\boldsymbol{\alpha} - \frac{1}{2} \boldsymbol{\alpha}^T D^T [B_l \; K_l] H^{-1} [B_l \; K_l]^T D\boldsymbol{\alpha} \tag{29}$$

with respect to $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_+^T \; \boldsymbol{\alpha}_-^T\right]^T$ subject to

$$\alpha_{+,i} \leq \lambda_2 \tag{30}$$

$$\alpha_{-,i} \leq \lambda_2 \tag{31}$$

for all $i = 1, \ldots, ||\mathcal{L}|| - l_0$. Note that the dual problem is a quadratic programming problem with non-negative definite matrix given by $Q \equiv D^T [B_l \; K_l] H^{-1} [B_l \; K_l]^T D$. We may also note that the optimization problem (**??**) relies on $||\mathcal{L}|| - l_0$ dual variables, so that the number of observed values of l and choice of l_0 are responsible for determining the size of the problem rather than the parameter dimension. This is a great advantage computationally when the number of penalized observed values of l is relatively small and the dimension of $\boldsymbol{a}$ is large.

5 The truncated power basis and an alternative decomposition of $\mathcal{H}$

The estimation of $\phi^* (l, m)$ 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 stationary 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, making imposing that conditional dependence between observation decay as l and the time between observations increase a reasonable way to institute regularization. This latter notion is instrumental in justifying the family of penalties

$$J_{2,(p)} = \sum_{l_i \in \mathcal{L}: l_i > l_0} |\mu^* + \phi_1^* (l_i)|^p$$

which we may view as a design-driven way of implementing the regularization which may be imposed by the penalty functionals taking the form

$$\begin{aligned} J(\phi^*) &= \int_{l_0}^{1} |\mu^* + \phi_1^* (l)|^p \, dl \\ &= \int_{0}^{1} |\mu^* + \phi_1^* (l)|^p I(l > l_0) \, dl \end{aligned} \tag{32}$$

Previously we decomposed the function space $\mathcal{H}$ according to $J_1 = ||\phi^{*\prime\prime}_1||^2$ in a somewhat traditional sense, but the penalty functionals given by (**??**) motivate a different decomposition of $\mathcal{H}$. The form of (**??**) is significantly different in nature from the penalty discussed in Section 2.1 and those typically encountered in the setting smoothing spline ANOVA models, particularly because (**??**) effects only a subset of the domain for l. Therefore, an appropriate decomposition of the function space into the null space of J and the penalized space should perhaps be formulated in terms of basis functions for the lag component, l with domains which do not include the entire unit interval.

6 The truncated power basis

Consider a sequence of knots partitioning the unit interval $0 < x_1 < x_2 < \cdots < x_n < 1$; the truncated power functions of degree k, $\{T_{i,k}\}_{i=1}^n$, are given by

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

Polynomial regression splines are widely used in the nonparametric function estimation setting; these are functions are continuous piecewise polynomials where the definition of the function changes at the collection of knot points. A polynomial of degree k has basis

$$\{1, l, \ldots, l^k, (l - l_1)_+^k, \ldots, (l - l_n)_+^k\}$$

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

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

The truncated power basis, as in their use in defining polynomial regression splines, enjoy a particular ease of interpretation, as the coefficient β_{i+k} may be identified as the size of the jump at x_i in the k^{th} derivative of f. This fact is especially useful when tracking change points or, in general, any abrupt changes in the regression curve. If we reflect these basis functions about each of their corresponding knot points and denote these reflections $\{T_{ik}^{-}\}$, then expressing the regularization corresponding to the penalty functionals (**??**) becomes quite natural in terms of the reflected basis functions $(\cdot - l_1)_{-}^{k}, \ldots, (\cdot - l_n)_{-}^{k}$, where $(\alpha)_{-} = \max(-\alpha, 0)$.

Banding the inverse covariance structure becomes quite convenient by simply penalizing the regression coefficients corresponding to the reflected truncated power basis functions with support on any l beyond some banding threshold, l_0.

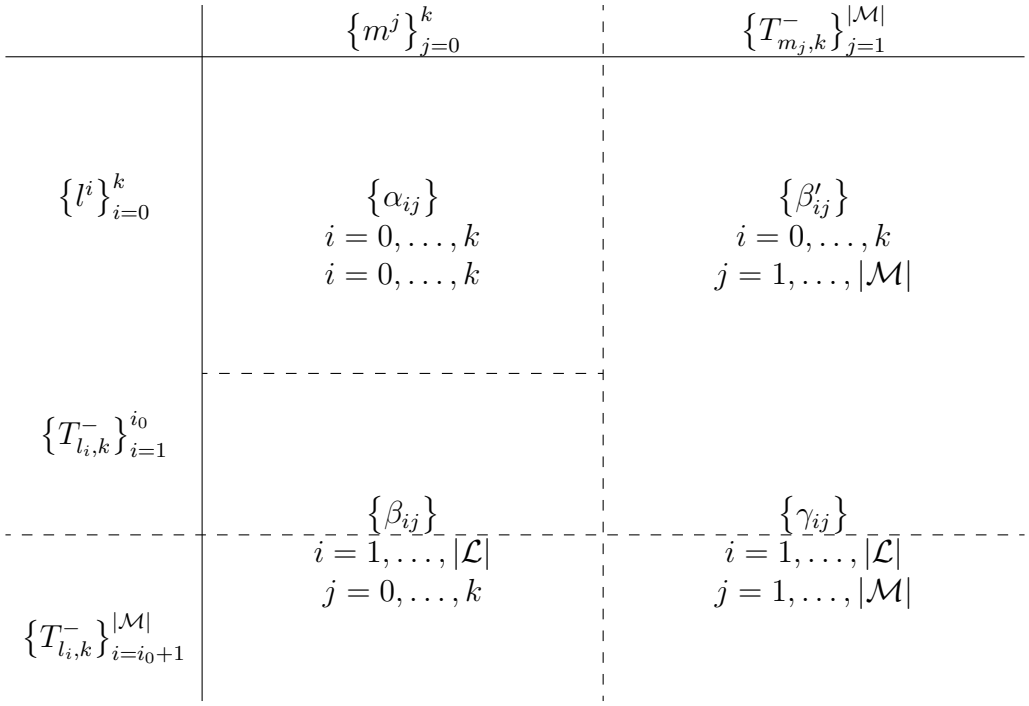
We consider a decomposition of $\mathcal{H}$ in terms of the reflected truncated power basis and corresponding penalty. We maintain the functional decomposition of $\phi^*(l, m) = \mu^* + \phi_1^*(l) + \phi_2^*(m) + \phi_{12}^*(l, m)$ we previously adopted, we preserve our definition of $\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m$ as a tensor product space where $\mathcal{H}_m = \mathcal{H}_m^0 \oplus \mathcal{H}_m^1$. For fixed $l_0 \in (0, 1)$, we decompose $\mathcal{H}_l = \mathcal{H}_l^0 \oplus \mathcal{H}_l^1$ where, for any $f \in \mathcal{H}_l$ we write

$$f(l) = \sum_{j=0}^{k} \beta_j l^j + \sum_{i=1}^{n} \beta_{k+i} T_{i,k}^*(l) \tag{33}$$

so that $\mathcal{H}_l = span\{1, l, \ldots, l^k, T_{1k}^*, \ldots, T_{||\mathcal{L}||}^*, k\}$. We express $\mathcal{H}_m$ as the span of the same basis functions, and as before, we define $\mathcal{H}$ as the tensor product space $\mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m$. The null models previously discussed lead to natural decompositions of $\mathcal{H}_l$ and $\mathcal{H}_m$: to shrink toward banded inverse covariance functions, we decompose $\mathcal{H}_l$ into two subspaces: $span\{T_{ik}^* : l_i < l_0\}$ and $span\{1, l, \ldots, l^k, T_{ik}^* : l_i \geq l_0\}$. For regularization resulting in null models corresponding to inverse Toeplitz structures, only a trivial decomposition of $\mathcal{H}_m$ is required: $span\{0\}$ and $span\{1, m, \ldots, m^k, T_{1k}^*, \ldots, T_{||\mathcal{M}||,k}^*\}$. The elements of

	$\{1\}$	$m \quad \dots \quad m^k$	$T^-_{m_1,k} \quad \dots \quad T^-_{m_{\lvert\mathcal{M}\rvert},k}$
$\{1\}$	$\{1\}$	$\{T^-_{m_j,k}\}_{j=1}^{\lvert\mathcal{M}\rvert}$	$\{m^j\}_{j=1}^{k}$
$\{l\}$ $\{l^2\}$ $\vdots$ $\{l^{k-1}\}$ $\{l^k\}$	$\{l^i\}_{i=1}^{k}$	$\{l^i m^j\}_{i,j=1}^{k}$	$\{l^i T^-_{m_j,k}\}_{i=1,j=1}^{k,\lvert\mathcal{M}\rvert}$
$T^-_{l_1,k}$ $T^-_{l_2,k}$ $\vdots$ $T^-_{l_{i_0-1},k}$ $T^-_{l_{i_0},k}$	$\{T^-_{l_i,k}\}_{i=1}^{i_0}$	$\{m^j T^-_{l_i,k}\}_{1=1,j=1}^{i_0,k}$	$\{T^-_{l_i,k} T^-_{m_j,k}\}_{i=1,j=1}^{i_0,\lvert\mathcal{M}\rvert}$
$T^-_{l_{i_0+1},k}$ $\vdots$ $T^-_{l_{\lvert\mathcal{L}\rvert},k}$	$\{T^-_{l_i,k}\}_{i=i_0+1}^{\lvert\mathcal{L}\rvert}$	$\{m^j T^-_{l_i,k}\}_{i=i_0+1}^{\lvert\mathcal{L}\rvert}$	$\{T^-_{l_i,k} T^-_{m_j,k}\}_{i=i_0+1,j=1}^{\lvert\mathcal{L}\rvert,\lvert\mathcal{M}\rvert}$

	$\left\{m^j\right\}_{j=0}^{k}$	$\left\{T^-_{m_j,k}\right\}_{j=1}^{\lvert\mathcal{M}\rvert}$
$\left\{l^i\right\}_{i=0}^{k}$	$\left\{\alpha_{ij}\right\}$ $i = 0, \ldots, k$ $i = 0, \ldots, k$	$\left\{\beta'_{ij}\right\}$ $i = 0, \ldots, k$ $j = 1, \ldots, \lvert\mathcal{M}\rvert$
$\left\{T^-_{l_i,k}\right\}_{i=1}^{i_0}$ $\left\{T^-_{l_i,k}\right\}_{i=i_0+1}^{\lvert\mathcal{M}\rvert}$	$\left\{\beta_{ij}\right\}$ $i = 1, \ldots, \lvert\mathcal{L}\rvert$ $j = 0, \ldots, k$	$\left\{\gamma_{ij}\right\}$ $i = 1, \ldots, \lvert\mathcal{L}\rvert$ $j = 1, \ldots, \lvert\mathcal{M}\rvert$



So that any $\phi^* \in \mathcal{H} = \mathcal{H}_l \otimes \mathcal{H}_m$ may be written

$$\phi^*(l,m) = \sum_{j=0}^{k} \beta_{lj} l^j + \sum_{j=0}^{k} \beta_{mj} m^j + \sum_{i=0}^{k}\sum_{j=0}^{k} l^i m^j + \sum_{i=1}^{||\mathcal{L}||} \beta_{l,k+i} T^*_{ik}(l) + \sum_{i=1}^{||\mathcal{M}||} \beta_{m,k+i} T^*_{ik}(m)$$
$$\sum_{j=0}^{k}\sum_{i=1}^{||\mathcal{L}||} m^j T^*_{ik}(l) + \sum_{j=0}^{k}\sum_{i=1}^{||\mathcal{M}||} l^j T^*_{ik}(m) + \sum_{i=1}^{||\mathcal{L}||}\sum_{j=1}^{||\mathcal{M}||} T^*_{ik}(l)\, T^*_{jk}(m)$$

7 A B-spline representation for pp functions

Definition 7.1. 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

$$B_{i,k,t}(x) = (t_{i+k} - t_i)\,[t_i, \ldots, t_{i+k}]\,(\cdot - x)_+^{k-1} \tag{34}$$

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.

7.1 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

$$B_i(x) = [t_{i+1}, \ldots, t_{i+k}]\,(\cdot - x)_+^{k-1} - [t_i, \ldots, t_{i+k-1}]\,(\cdot - x)_+^{k-1} \tag{35}$$

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

$$\begin{aligned}
\sum_i B_i(x) &= \sum_{i=j+1-k}^{j} B_i(x) \\
&= \sum_{i=j+1-k}^{j} [t_{i+1}, \ldots, t_{i+k}]\,(\cdot - x)_+^{k-1} - \sum_{i=j+1-k}^{j} [t_i, \ldots, t_{i+k-1}]\,(\cdot - x)_+^{k-1} \\
&= [t_{j+1}, \ldots, t_{j+k}]\,(\cdot - x)_+^{k-1} - [t_{j+1-k}, \ldots, t_j]\,(\cdot - x)_+^{k-1} \\
&= 1 - 0
\end{aligned} \tag{36}$$

The last equality in **??** 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

$$\begin{aligned}
[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-1} &= [t_i,\ldots,t_{i+k}]\,(t-x)\,(t-x)_+^{k-2} \\
&= \sum_{r=i}^{i+k} [t_i,\ldots,t_{i+r}]\,(t-x)\,[t_r,\ldots,t_{i+k}]\,(t-x)_+^{k-2} \\
&= \Big[\,[t_i]\,(t-x)\,\Big]\Big[\,[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-2}\,\Big] \\
&\qquad + \Big[\,[t_i,t_{i+1}]\,(t-x)\,\Big]\Big[\,[t_{i+1},\ldots,t_{i+k}]\,(t-x)_+^{k-2}\,\Big] \\
&= (t_i-x)\,[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-2} \\
&\qquad + 1\cdot[t_{i+1},\ldots,t_{i+k}]\,(t-x)_+^{k-2} \qquad (37)
\end{aligned}$$

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 **??** as

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

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

$$\frac{B_{i,k}\,(x)}{t_{i+k}-t_i} = \frac{x-t_i}{t_{i+k}-t_i}\frac{B_{i,k-1}\,(x)}{t_{i+k-1}-t_i} + \frac{t_{i+k}-x}{t_{i+k}-t_i}\frac{B_{i+1,k-1}\,(x)}{t_{i+k}-t_{i+1}} \qquad (38)$$

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 **??**, 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 **?? ??** and by induction over k, starting with the fact that

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

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

Definition 7.2. 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,

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

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 **??**, **??**, and **??** in **??**, 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 **??**, for $x \in [t_j, t_{j+1}]$,

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

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

Hastie and Tibshirani 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

$$y(t) = f(t) + \epsilon(t) \tag{39}$$

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

$$y(t) = \sum_{j=1}^{K} \alpha_j B_j(t) + \epsilon(t) \tag{40}$$

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

$$\begin{aligned} S &= \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{K} \alpha_j B_j(t_i) \right)^2 \\ &= |y - B\alpha|^2 \end{aligned} \tag{41}$$

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

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

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

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

To extend the varying intercept model **??** 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

$$E\left[Y\left(t\right)\right] = \beta_0\left(t\right) + \beta_1\left(t\right)x\left(t\right) \tag{42}$$

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 **??**, requiring only minor adjustments to the design matrix. In matrix notation as described in **??**, the mean vector may be written

$$\mu = B\alpha_0 + diag\left\{x\left(t\right)\right\}B\alpha_1 \tag{43}$$

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 $U = diag\left\{x\left(t\right)\right\}B$, **??** becomes

$$\mu = \left[B|U\right]\left(\alpha_0^T, \alpha_1^T\right)^T \tag{44}$$

$$\equiv Q\alpha \tag{45}$$

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

$$S = \left|y - Q\alpha\right|^2 \tag{46}$$

which has explicit solution

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

It is of interest to notice 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, UB. 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

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

and the estimate of the smooth mean function is obtained via

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

where $H = Q\left(Q^TQ\right)^{-1}Q^T$ is the "hat" matrix. This will be discussed in further detail in later sections on smoothing parameter selection and model tuning.

9 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; Kooperberg and Stone, 1991,1992). 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 smoothing spline chapter, 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. citeo1986statistical 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 (1988), Wahba (1990) and Green and Silverman (1994). 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. citewand2008semiparametric 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

$$\begin{aligned} \omega = \Bigg(& \frac{1}{6}(\Delta\phi)_1, \frac{4}{6}(\Delta\phi)_1, \frac{1}{6}(\Delta\phi)_1, \frac{1}{6}(\Delta\phi)_2, \frac{4}{6}(\Delta\phi)_2, \\ & \frac{1}{6}(\Delta\phi)_2, \ldots, \frac{1}{6}(\Delta\phi)_{n+7}, \frac{4}{6}(\Delta\phi)_{n+7}, \frac{1}{6}(\Delta\phi)_{n+7} \Bigg) \end{aligned}$$

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.

9.1 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 accomodate 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 **??** 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

$$\sum_{i=1}^{m}\left\{y_i - \sum_{j=1}^{K}\beta_j B_j\left(t_i\right)\right\}^2 + \lambda\int_{t_{min}}^{t_{max}}\left\{\sum_{j=1}^{K}\beta_j B_j''\left(t\right)\right\}^2 dt. \tag{47}$$

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 citewhittaker1922new, difference penalties have been utilized for nearly a century, with more recent applications outlined in citeeilers1991penalized, citeeilers1991nonparametric, and citeeilers1995indirect. 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 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

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

citeeilers1996flexible 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. citeo1986statistical used third-degree B-splines and the following penalty:

$$h^2 P = \lambda \int_{t_{min}}^{t_{max}} \left\{ \sum_j \alpha_j B''_{j,3}(t) \right\}^2 dt \tag{48}$$

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

$$h^2 P = \lambda \int_{t_{min}}^{t_{max}} \sum_j \sum_k \Delta^2\alpha_j \Delta^2\alpha_k B_{j,1}(t) B_{k,1}(t)\, dt \tag{49}$$

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

$$\begin{aligned} h^2 P &= \lambda \int_{t_{min}}^{t_{max}} \left[\left\{ \sum_j \Delta^2\alpha_j B_j(t,1) \right\}^2 + 2\sum_j \Delta^2\alpha_j \Delta^2\alpha_{j-1} B_j(t,1) B_{j-1}(t,1) \right] dt \\ &= \lambda \left[\sum_j \left(\Delta^2\alpha_j\right)^2 \int_{t_{min}}^{t_{max}} B_j^2(t,1)\ dt + 2\sum_j \Delta^2\alpha_j \Delta^2\alpha_{j-1} \right] \end{aligned} \tag{50}$$

or

$$\begin{aligned} h^2 P = \lambda \sum_j \left(\Delta^2\alpha_j\right)^2 \int_{t_{min}}^{t_{max}} B_{j,1}^2(t)\, dt + 2\lambda \sum_j \Delta^2\alpha_j \Delta^2\alpha_{j-1} \\ + \int_{t_{min}}^{t_{max}} B_{j,1}(t) B_{j-1,1}(t)\, dt \end{aligned} \tag{51}$$

which can be written as

$$h^2 P = \lambda \left\{ c_1 \sum_j \left(\Delta^2 \alpha_j\right)^2 + c_2 \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} \right\} \tag{52}$$

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

$$\begin{aligned} c_1 &= \int_{t_{min}}^{t_{max}} B_{j,1}^2 (t)\, dt \\ c_2 &= \int_{t_{min}}^{t_{max}} B_{j,1} (t)\, B_{j-1,1} (t)\, dt \end{aligned} \tag{53}$$

O'Sullivan's ridge-like B-spline penalty in Equation **??** 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 **??** 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 **??** 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

$$\begin{aligned} S_\lambda &= |y - B\alpha|^2 + \lambda |D_d \alpha|^2 \\ &= (y - B\alpha)^T (y - B\alpha) + \lambda \alpha^T P \alpha \end{aligned} \tag{54}$$

where

$$P = D_d^T D_d$$

and the elements of B are given by $b_{ij} = B_j (t_i)$, as defined in **??**. Taking derivatives on both sides of **??** with respect to α gives

$$\begin{aligned} \frac{\partial}{\partial \alpha} S_\lambda &= \frac{\partial}{\partial \alpha} \left(\alpha^T B^T B \alpha - 2 y^T B^T \alpha + \lambda \alpha^T D_k^T D_k \alpha \right) \\ &= 2 B^T B \alpha - 2 B^T y + 2 \lambda D_d^T D_d \alpha \\ &= \left(B^T B + \lambda D_d^T D_d \right) \alpha - B^T y \end{aligned} \tag{55}$$

and setting equal to zero yields normal equations:

$$B^T y = \left(B^T B + \lambda D_d^T D_k \right) \alpha, \tag{56}$$

which has explicit solution

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

The effective hat matrix is now

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

When $\lambda = 0$, we have the standard normal equations of linear regression with a B-spline basis, and with $k = 0$ **??** 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.

9.1.1 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 **??** 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 **??** must be further modified to accommodate multiple tuning parameters, λ_i, $i = 0, 1$. The objective function then becomes

$$\begin{aligned} S_\lambda^* &= |y - Q\alpha|^2 + \lambda_0 |D_{d_0}\alpha_0|^2 + \lambda_1 |D_{d_1}\alpha_1|^2 \\ &= |y - Q\alpha|^2 + |\alpha^T P \alpha|^2 \end{aligned} \tag{57}$$

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 **??** 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

$$H = Q\left(Q^T Q + P\right)^{-1} Q^T. \tag{58}$$

[Figure **??** Need to explain figure 3 here.]

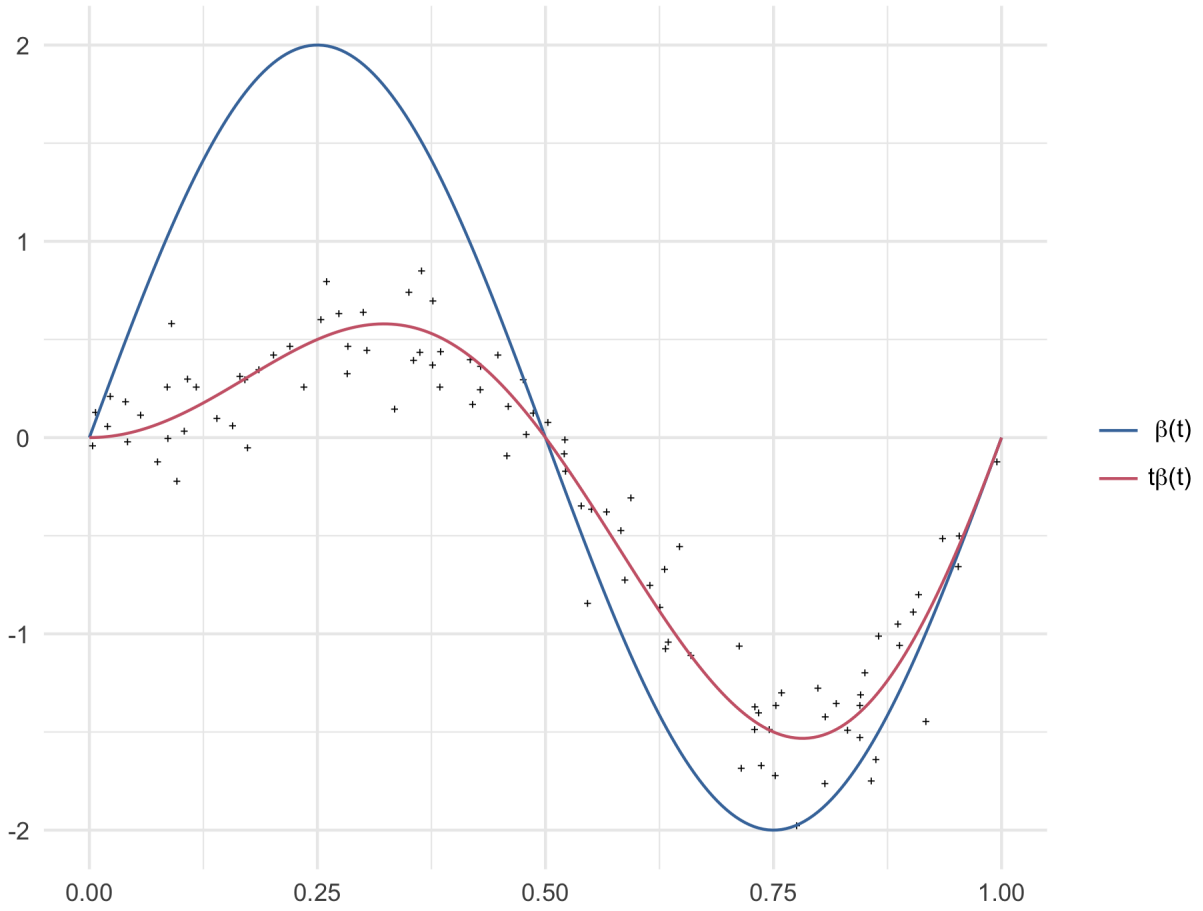
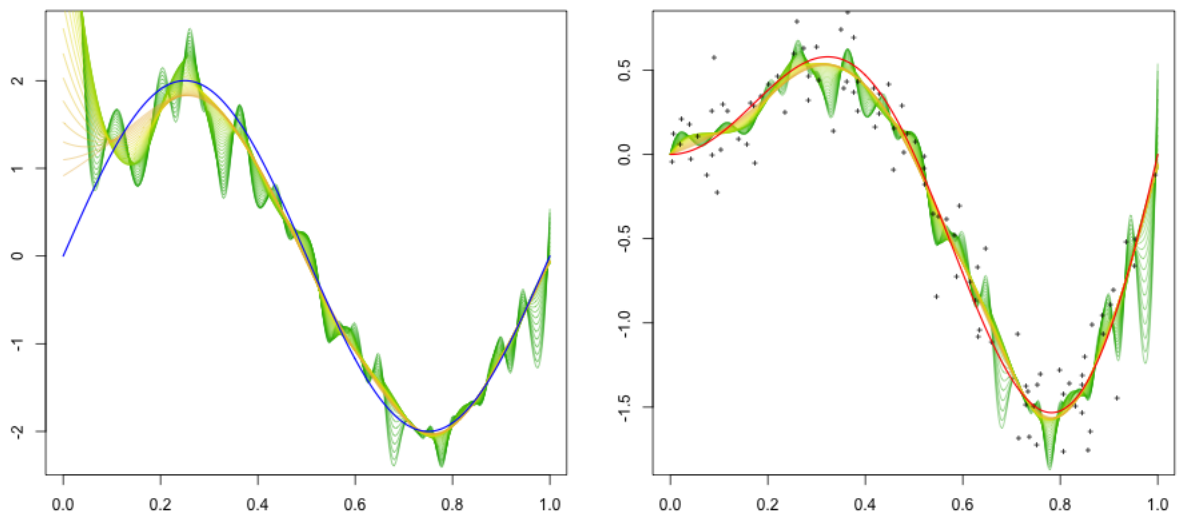


Figure 1: *100 simulated data points where* $y(t) = t\beta(t) + 0.2\epsilon(t)$ *where* ϵ *is a white noise process with unit variance, and* $\beta(t) = 2\sin(2\pi t)$.



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

Figure 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 **??**. 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.

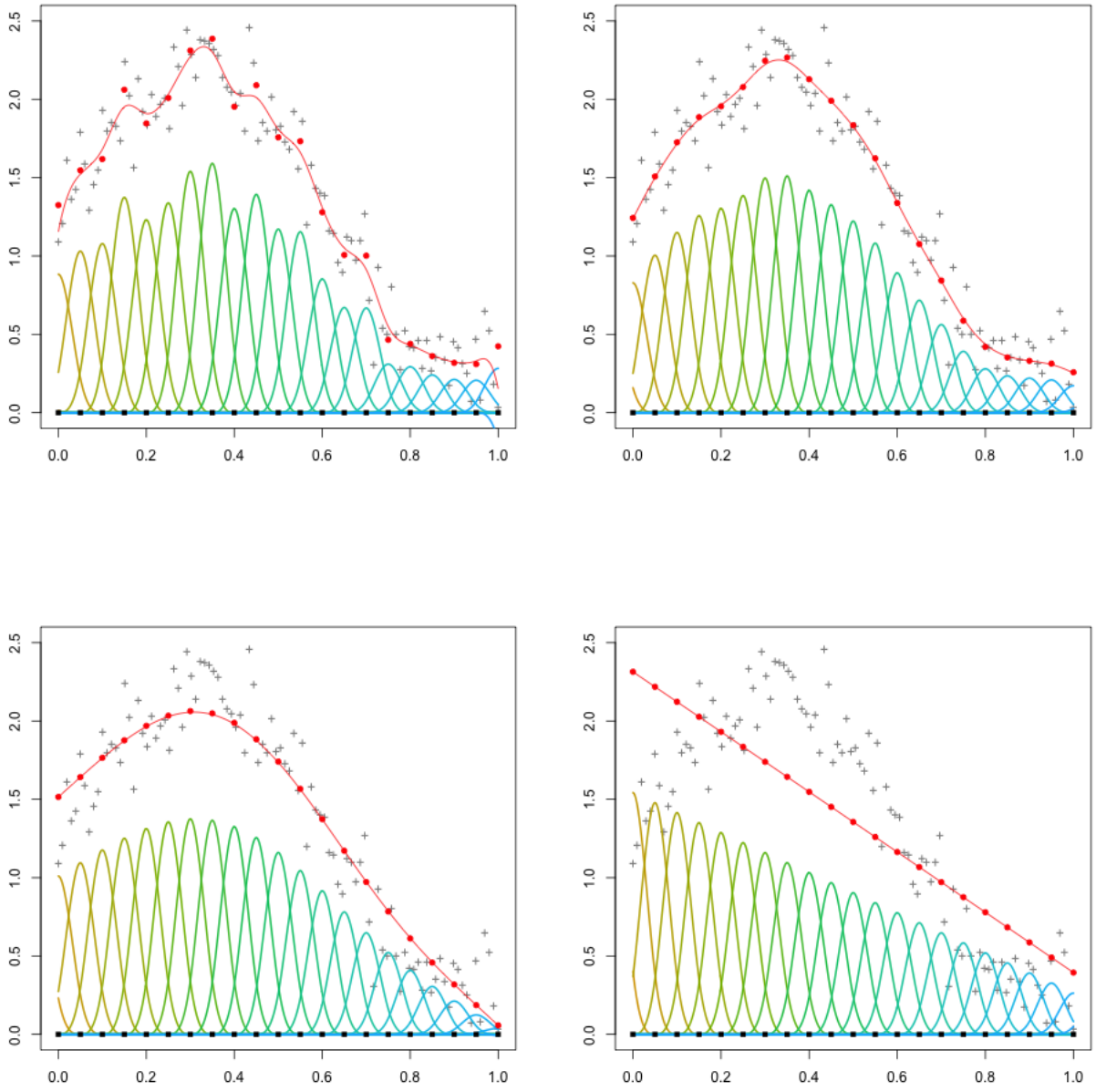


Figure 3: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The fitted curve in the upper left plot is the most "wiggly" of 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.

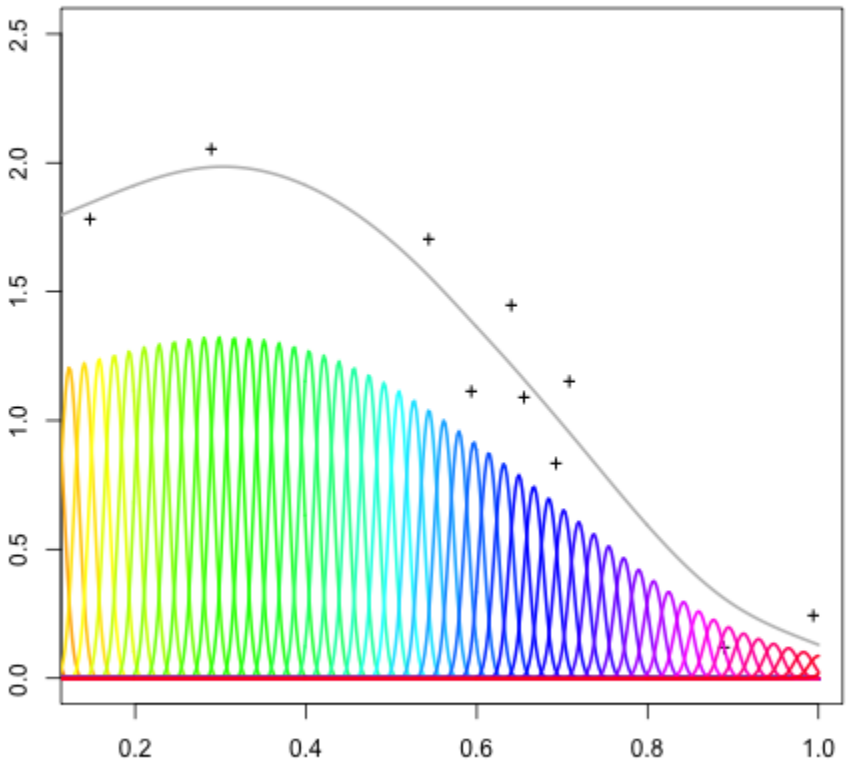


Figure 4: P-spline smoothing of 10 observations using 60 B-spline basis functions.

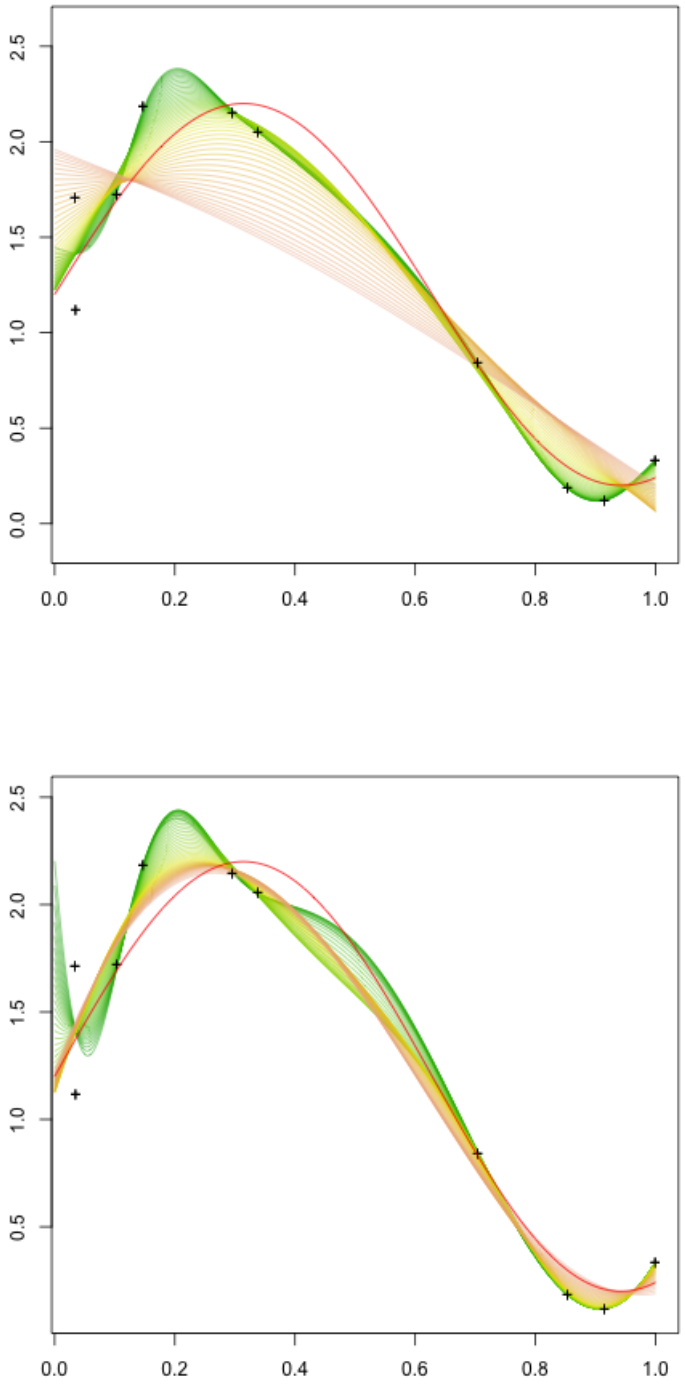


Figure 5: Fitted mean curves using a second (top) and third (bottom) order difference penalty for simulated data, sparsely sampled along the indexing variable: $y(t) = 1.2 + \sin(5t) + 0.2\epsilon_t$, where $\epsilon_t \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$. A total of 10 data points were fit using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

9.2 Properties of P-splines

P-splines enjoy many advantageous properties, many due in part to the inherited properties of the B-spline basis functions on which a generous portion of their foundation is constructed.

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 **??**:

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

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, citeschoenberg1953polya 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 **??**, so that the minimizer α must be such that $\sum_{j=3}^{n} (\Delta^2\alpha_j)^2$ is close to zero. Consequently, each of the individual second differences must also be nearly zero, and thus the second derivative of the fitted function must be close to zero over the entire domain.

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

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

(or for H defined for the addition of a varying slope component as in **??**) 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

$$Q_B = B^TB \qquad \text{and} \qquad Q_\lambda = \lambda D^T D. \tag{59}$$

Then using properties of the matrix trace, we can write

$$\begin{aligned} \text{tr}\left(H_\lambda\right) &= \text{tr}\left[\left(Q_B + Q_\lambda\right)^{-1} Q_B\right] \\ &= \text{tr}\left[Q_B^{1/2}\left(Q_B + Q_\lambda\right)^{-1} Q_B^{1/2}\right] \\ &= \text{tr}\left[\left(I + Q_B^{-1/2} Q_\lambda Q_B^{-1/2}\right)^{-1}\right] \end{aligned} \tag{60}$$

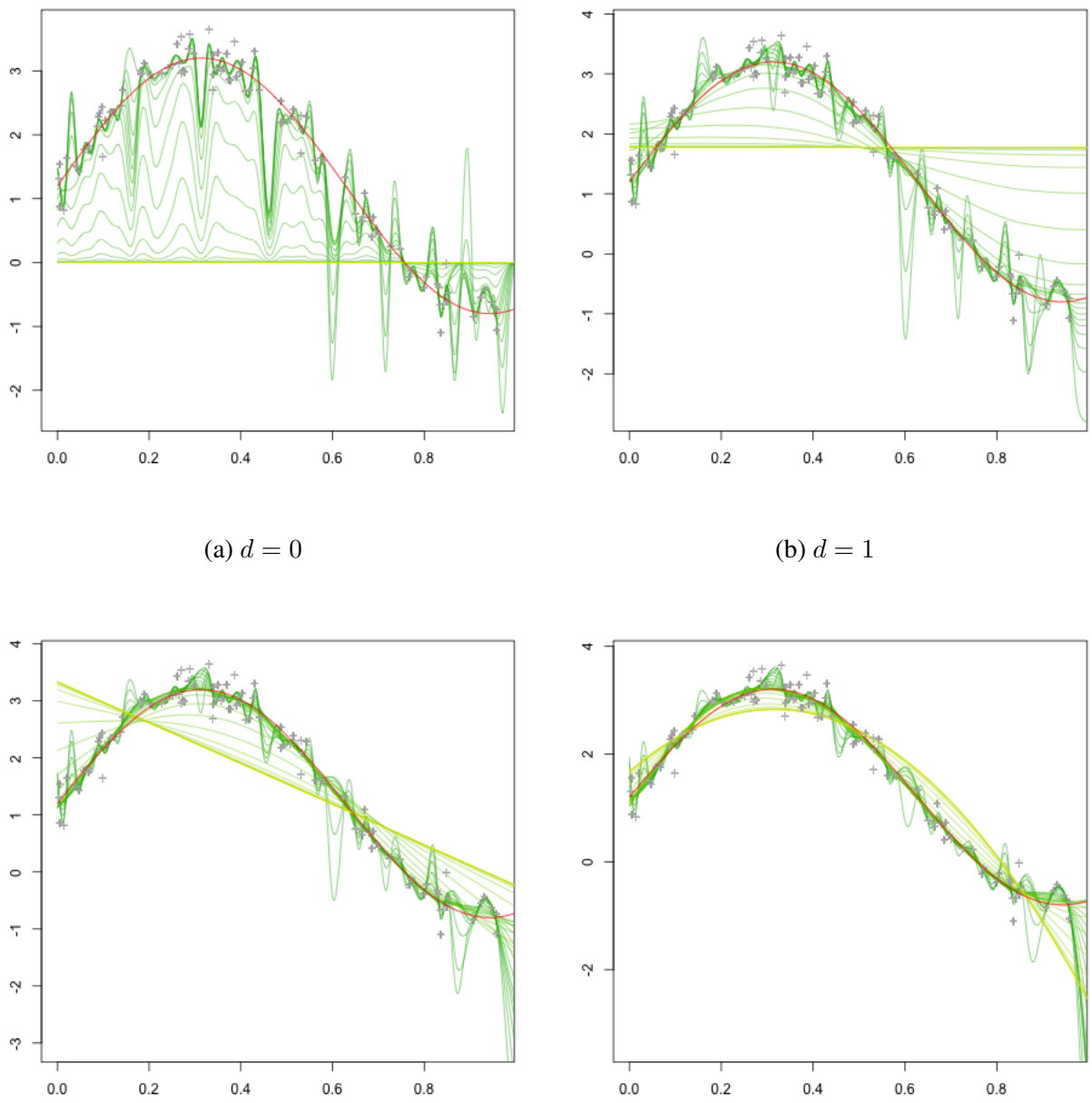
Define $L \equiv Q_B^{-1/2} Q_\lambda Q_B^{-1/2}$. Then

$$\text{tr}\left(H_\lambda\right) = \text{tr}\left[\left(I + \lambda L\right)^{-1}\right] = \sum_{j=1}^{n} \frac{1}{1 + \lambda\gamma_j} \tag{61}$$

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

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

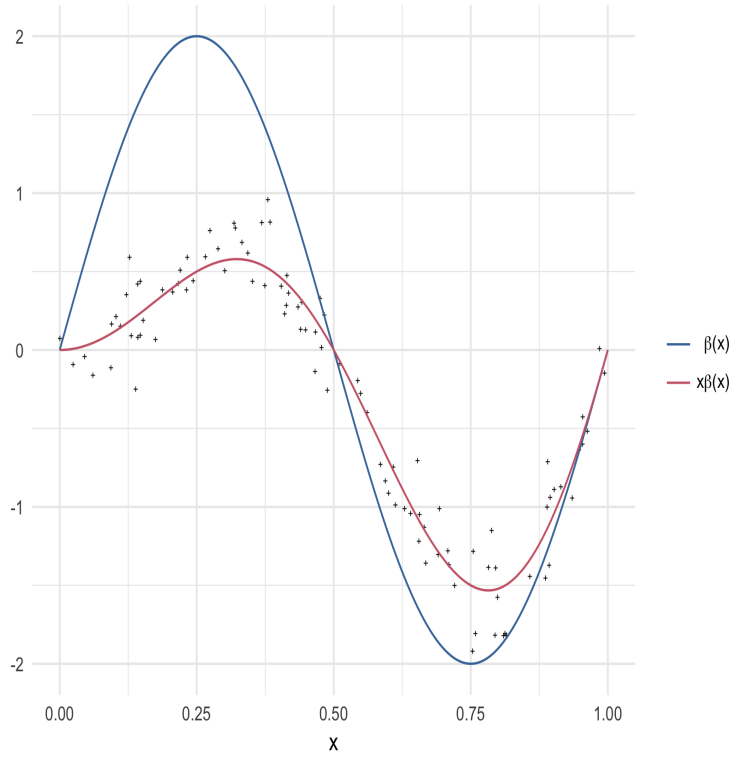
The previous derivations hold regardless of whether we are fitting the varying intercept-only model, with $\mu\left(t\right) = \beta_0\left(t\right)$ or accommodating a varying slope for a regressor by specifying $\mu\left(t\right) = \beta_0\left(t\right) + \beta_1\left(t\right) x\left(t\right)$. The 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 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* **?? ??**.


2
1
0
-1
-2
0.00
0.25
0.50
0.75
1.00
x
β(x)
xβ(x)