Chapter 3 outlines a cohesive nonparametric approach to modeling $\phi(s,t)$ as a smooth bivariate function via P-splines, or penalized B-spline basis expansions. Eilers and Marx were the first to propose the method, which generalizes the work of O'Sullivan, which utilized derivative-based penalties to achieve smoothness in the fitted function. Their approach stands on two fundamental building blocks: B-splines and difference penalties. In the sections to follow, we will review the definition of the B-spline basis and derive the essential properties necessary for motivating the application of difference penalties on the spline coefficients. We will show that the simple difference penalties are easy to compute and achieve nearly the same construction as when utilizing the more computationally demanding derivative-based penalties, and demonstrate the effect of the roughness measure on the estimated functional coefficients.

1 A representation for piecewise polynomial functions

Let $\xi = \{\xi_1 < \xi_2 < \cdots < \xi_{l+1}\}$ be a set of strictly increasing series of points, and let k be a positive integer. Further, let $P_1, \ldots, P_l$ denote a sequence of l polynomials of order k. Then the correponding piecewise polynomial (pp) function of order k is defined as follows:

$$f(x) = P_i(x) \text{ if } \xi_i < x < \xi_{i+1}$$

for $i = 1, \ldots, l$. $\{\xi\}$ are known as the breakpoints of f. At the interior breakpoints, $\xi_2, \ldots, \xi_l$, the function value is defined by specifying f to be right continuous; that is,

$$f(\xi_i) = f\left(\xi_i^+\right), \quad i = 2, \ldots, l$$

However, in a sense, without this specification, the function has two values at any interior breakpoint: the value it gets from the polynomial piece to the left of the breakpoint, $f\left(\xi_i^-\right) = P_{i-1}(\xi_i)$, in addition to the value it gets from the polynomial piece to the right of the breakpoint, $f\left(\xi_i^+\right) = P_i(\xi_i)$. To properly define the function, one can specify f to be right-continuous:

$$f(\xi_i) \equiv f\left(\xi_i^+\right) \tag{1}$$

Denote the set of pp functions of order k with breakpoints $\xi = \{\xi_1, \ldots, \xi_{l+1}\}$ by

$$\mathcal{P}_{k,\xi}.$$

$\mathcal{P}_{k,\xi}$ is a linear space having dimension kl, as it consists of l polynomials, each having k polynomial coefficients. The j^{th} derivative of a pp f,

$$D^j f$$

is a pp function of order $k - j$ having the same breakpoint sequence and constructed from the same j^{th} derivatives of the polynomial pieces from which f was constructed. This "definition" dodges much of the complicated discussion of the derivatives of a pp function at its breakpoints and thus must be treated with considerable care in context of the fundamental theorem of calculus.

Proposition 1.1. *A pp function, f satisfies*

$$f(x) - f(a) = \int_a^x (Df)(t)\, dt \quad \text{for all} \quad x$$

if and only if f is a continuous function.

Consider a piecewise constant function f: by the previous definition, its first derivative is identically zero, and is therefore equal to the usual derivative of f if and only if f is constant. The following definition gives the information necessary for a convenient and efficient representation of this class of functions:

Definition 1.1. The *piecewise polynomial (pp) representation* of a function $f \in \mathcal{P}_{k,\xi}$ consists of

I. integers k and l, specifying the order and number of polynomial pieces, respectively,

II. a strictly increasing set of breakpoints, $\xi_1, \xi_2, \ldots, \xi_{l+1}$,

III. and the set of values of the right derivatives at each of the breakpoints,

$$c_{ij} = D^j f\left(\xi_i^+\right), \quad j = 1, \ldots, k; \quad i = 1, \ldots, l$$

2 The truncated power basis and the spaces $\mathcal{P}_{k,\xi,\nu}$

This prerequisite information is merely for the ability to responsibly refer to the set of piecewise polynomial functions and have a shorthand way of doing so. These means enable us to introduce two sets of basis functions: first, the truncated power basis, followed by B-spline basis functions. We will see that both are closely related, with the former having some properties which leave them unattractive for function approximation and thus present the construction of B-splines and how to use them to construct a representation of $\mathcal{P}_k$. In practice, one typically is given some information about an unknown function, g, and the task is to construct a function $f \in \mathcal{P}_{k,\xi}$ which satisfies conditions that g also satisfies, and in addition, has a certain number of continuous derivatives. These conditions define a subspace of $\mathcal{P}_{k,\xi}$, $\mathcal{P}_{k,\xi,\nu}$ for which we will need a corresponding basis.

2.1 Example: histogram smoothing by parabolic splines

For illustrative purposes, consider the task of smoothing a histogram using parabolic splines. Suppose we are given points

$$\tau_1 < \tau_2 < \cdots < \tau_{n+1}$$

and non-negative numbers $h_1, h_2, \ldots, h_n$, with h_i denoting the height of the histogram over the interval (τ_i, τ_{i+1}). The histogram is an approximate representation of some underlying density function, g. Letting $\Delta\tau_i = \tau_{i+1} - \tau_i$, one may interpret $h_i \Delta\tau_i$ as (approximately) equal

to the integral of g over $[\tau_i, \tau_{i+1}]$. One may impose the following interpolation conditions on our smooth function, f:

$$\int_{\tau_i}^{\tau_{i+1}} f(x)\, dx = h_i \Delta\tau_i$$

for $i = 1, \ldots, n$. Let f be a piecewise polynomial of order 3 having continuous first derivative:

$$f \in \mathcal{P}_{3,\xi} \cap \mathcal{C}^{(1)}$$

Choose the breakpoint sequence ξ to coincide with $\tau = \{\tau_1, \ldots, \tau_{n+1}\}$. If g is smooth and vanishes outside its support, $[\tau_1, \tau_{n+1}]$, then

$$g^{(j)}(\tau_1) = g^{(j)}(\tau_{n+1}) = 0,$$

for $j = 0, 1, \ldots, d$, where d characterizes the extent of the smoothness of g, we may also wish to require f to obey two additional interpolation constraints:

$$f(\tau_1) = f(\tau_{n+1}) = 0,$$

giving a total of $n + 2$ interpolation conditions. These, along with the $2(n-1)$ continuity conditions yield a total $3n$ constraints on the $3n$ polynomial coefficients,

$$c_{ji} \equiv D^{j-1} f\left(\xi_i^+\right).$$

These conditions lead to the system of equations:

$$\begin{array}{rrrrrrrrrrr}
c_{11} & & & & & & & & & = & 0 \\
c_{11} + & c_{21}\dfrac{\Delta\tau_1}{2!} + & c_{31}\dfrac{(\Delta\tau_1)^2}{3!} & & & & & & & = & h_1 \\
c_{11} + & c_{21}\Delta\tau_1 + & c_{31}\dfrac{2(\Delta\tau_1)^2}{3!} - & c_{12} & & & & & & = & 0 \\
\vdots & c_{21} + & c_{31}\Delta\tau_1 & - & c_{22} & & & & & = & 0 \\
 & & & c_{12} + & c_{22}\dfrac{\Delta\tau_2}{2} + & c_{32}\dfrac{(\Delta\tau_2)^2}{3!} & & & & = & h_2 \\
 & & & c_{12} + & c_{22}\Delta\tau_2 + & c_{32}\dfrac{(\Delta\tau_2)^2}{2} & \ldots & & = & 0 \\
 & & & & c_{22} + & c_{32}\Delta\tau_2 & \ldots & & = & 0 \\
 & & & & & & & \ddots & & (2)
\end{array}$$

2.2 The subspace, $\mathcal{P}_{k,\xi,\nu}$

One may quickly see that this system is two-thirds homogeneous; that is, for every integral interpolation constraint, we have two continuity constraints that lead to zeros on the right hand side of the equality. To solve 2, the homogeneous equations are solved, leaving a reduced set of equations. To do this, one may construct a set of linearly independent functions $\phi_1, \phi_2, \ldots$ of the same size as the number of interpolation constraints which satisfy the homogeneous equations in 2. The smoother, f, is then constructed within this subspace of $\mathcal{P}_{3,\xi}$ and has form

$$f = \sum_j \alpha_j \phi_j.$$

The $\{\phi_j\}$ span a particular subspace of $\mathcal{P}_{k,\xi}$, which is comprised of functions which satisfy the homogeneous equations in 2. In general, we may characterize these homogeneous equations in terms of a given function having a particular number of continuous derivatives, which may be expressed as follows:

$$\begin{aligned} \mathcal{J}_{ij} f = 0, \quad i &= 2, \ldots, l \\ j &= 1, \ldots, \nu_i \end{aligned} \tag{3}$$

where

$$\mathcal{J}_{ij} f = \lim_{x \to \xi^+} D^{j-1} f(x) - \lim_{x \to \xi^-} D^{j-1} f(x) \tag{4}$$

and $\nu = (\nu_1, \ldots, \nu_l)$ is a vector of non-negative integers. Each ν_i specifies the number of continuity conditions impose on the function at the i^{th} breakpoint, ξ, and $\mathcal{J}_{ij}$ is simply the size of the jump in the $(j-1)^{st}$ derivative at ξ.

2.3 The truncated power functions

As homogeneous conditions specified in 3 are done so in terms of a linear operator applied to the functions in the space, those functions $\{g \in \mathcal{P}_{k,\xi}\}$ constitute a linear subspace, denoted

$$\mathcal{P}_{k,\xi,\nu}$$

Now that the subspace is defined, we need a basis for it. One such candidate set of basis functions is the truncated power basis. Define

$$(x-t)_+ = \max(0, x-t)$$

We may then define the truncated power function as follows:

$$\begin{aligned} (x-t)_+^p &= \left((x-t)_+\right)^p \\ &= \begin{cases} (x-t)^p, & x \geq t \\ 0 & x < t \end{cases} \end{aligned} \tag{5}$$

The function $g(x) = ()_+^p$ is a piecewise polynomial with a single breakpoint at ξ, and is continuous at this breakpoint for $p > 0$. For $p = 0$, it has a jump of size 1 at ξ. Since

$$D(\cdot - \xi)_+^p = p(\cdot - \xi)_+^{p-1},$$

it is clear that g has $p-1$ continuous derivatives. Define the set of linear operators $\{\lambda_{ij}\}$ and corresponding functions $\{\phi_{ij}\}$ as follows:

$$\begin{aligned} \lambda_{ij} f &= \begin{cases} D^j f(\xi) & i = 1 \\ D^j f(\xi^+) - D^j f(\xi^-) & i = 2, \ldots, l \end{cases} \\ \phi_{ij}(x) &= \begin{cases} \frac{(x-\xi_i)^p}{j!} & i = 1 \\ \frac{(x-\xi_i)_+^p}{j!} & i = 2, \ldots, l \end{cases} \end{aligned} \tag{6}$$

for $j = 0, \ldots, k-1$. Per this definition, we have that $\phi_{ij} \in \mathcal{P}_{k,\xi}$ for each $i = 1, \ldots, l$. Further, we have that

$$\lambda_{ij}\phi_{pq} = \begin{cases} 1 & i = p,\ j = q \\ 0 & \text{otherwise} \end{cases},$$

making $\{\phi_{ij}\}$ a set of kl linearly independent functions, and since $\mathcal{P}_{k,\xi}$ has dimension kl, they constitute a basis for the space. We may represent any $g \in \mathcal{P}_{k,\xi}$ in the form

$$g = \sum_{i,j} (\lambda_{ij} g)\, \phi_{ij}$$

Rewriting this expansion in the terms presented in 6, we have that we may express any function in the space as

$$g(x) = \sum_{j=0}^{k-1} \left[D^j g(\xi_1) \frac{(x-\xi_1)^j}{j!} + \sum_{i=2}^{l} \left[D^j g\left(\xi_i^+\right) - D^j g\left(\xi_i^-\right) \right] \frac{(x-\xi_i)_+^j}{j!} \right] \tag{7}$$

From this representation of the function, one can see that the coefficients of the basis functions are explicitly defined in terms of jumps of various derivatives of g at the breakpoints. Thus, enforcing the homogeneous constraints

$$\mathcal{J}_{ij} f = 0, \quad i = 2, \ldots, l; \quad j = 1, \ldots, \nu_i$$

is accomplished simply by restricting our attention to functions of the form 7 for which these coefficients are zero. This implies that the reduced set of basis functions $\{\phi_{ij}\}$, $i = 1, \ldots, l$, $j = \nu_i, \ldots, k-1$ is a basis for the subspace, $\mathcal{P}_{k,\xi,\nu}$. (We let $\nu_1 = 0$.) Then, every function $g^* \in \mathcal{P}_{k,\xi}$ in satisfying the homogeneous equations may be written in exactly one way of the form

$$g^* = \sum_{i=1}^{l} \sum_{j=\nu_i}^{k-1} \alpha_{ij} \phi_{ij} \tag{8}$$

2.3.1 The pitfalls of the truncated power basis

While the truncated power basis initially appears quite attractive for smoothing problems involving piecewise polynomials, they exhibit properties that can lead to poor function representations. In order to discuss these characteristics, we must first introduce the notion of the *condition* of a function representation.

Definition 2.1. Consider a piecewise polynomial representation of a function, p:

$$p = \sum_{i=1}^{n} a_i P_i \tag{9}$$

where $a = (a_1, \ldots, a_n)$ is a coefficient vector and $\{P_i\}$ is a set of polynomial functions defined on a closed interval $[a, b]$. We define the *size* of the polynomial p to be

$$\|p\| = \max_{a \le x \le b} |p(x)|,$$

and similarly, we define the size of the coefficient vector a to be

$$\|a\| = \max_{1 \le i \le n} |a_i|,$$

Then, we can bound the size of the function p:

$$m \|a\| \le \left\| \sum_{i=1}^{n} a_i P_i \right\| \le M \|a\| \tag{10}$$

where

$$m = \min_a \frac{\|\sum_{i=1}^{n} a_i P_i\|}{\|a\|}, \quad \text{and}$$
$$M = \min_a \frac{\|\sum_{i=1}^{n} a_i P_i\|}{\|a\|}$$

The *condition* of the representation of p as written in 9 is given by

$$condition(P_i) = \frac{M}{m}$$

The condition of a function representation quantifies the extent to which a slight change in the coefficient vector will impact the function itself. To see this, rather than the coefficient vector a, consider instead a perturbation of a:

$$a + \delta a$$

The corresponding perturbed polynomial is given by

$$p + \delta p = \sum_{i=1}^{n} (a_i + \delta a_i) P_i$$

By 10, we then have that

$$\frac{m\,\|\delta a\|}{M\,\|a\|} \leq \frac{\|\delta p\|}{\|p\|} \leq \frac{M\,\|\delta a\|}{m\,\|a\|},$$

implying that a relative change of $\frac{\|\delta a\|}{\|a\|}$ to the coefficient vector may result in a relative change to the function p as large as $\frac{M}{m} = \mathit{condition}\,(P_i)$ times the relative change in a (and at least as large as $\mathit{condition}\,(P_i)^{-1}$). Note that the width of the interval

$$\left[\mathit{condition}\,(P_i)^{-1}\,\frac{\|\delta a\|}{\|a\|},\, \mathit{condition}\,(P_i)\,\frac{\|\delta a\|}{\|a\|}\right]$$

is increasing in $\mathit{condition}\,(P_i)$; so large values of the condition of a representation imply that small relative changes in their corresponding coefficients may result in much smaller or much larger relative changes in the function being represented.

When representing a function $f \in \mathcal{P}_{k,\xi}$ as in 7, two issues of concern arise: first, if l is large, the value of the function at a point x can potentially rely on far more than just k of the basis coefficients. Additionally, if the breakpoints ξ are very irregularly spaced, then the truncated power basis can present poor condition, which, in turn, can result in poorly conditioned linear systems (like the specific example given by 2.) Consequently, small perturbations of the basis function coefficients result in disproportional changes in the function.

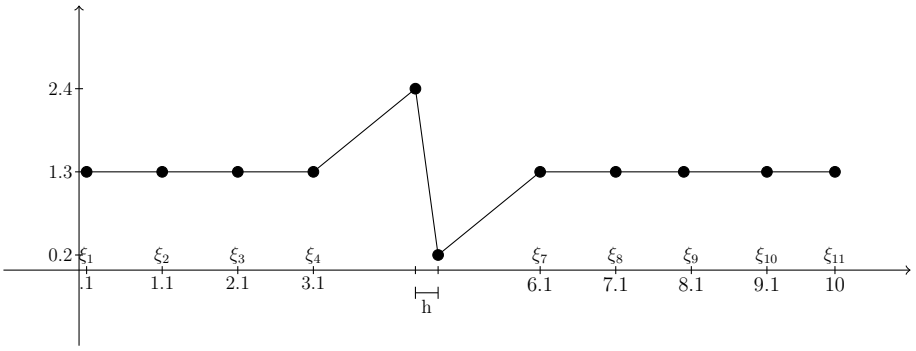


Figure 1: The linear system for the coefficients of the truncated power basis is ill-conditioned for the above piecewise linear function and choice of breakpoints, $\xi_1, \ldots, \xi_{11}$.

Example 2.1. Suppose we wish to construct a function $f \in \mathcal{P}_{k=2,\xi} \cap \mathcal{C}^{(0)}$ satisfying $f\,(\xi_i) = 1.3$ for $i = 1, \ldots, 4, 7, \ldots, 11$, $f\,(\xi_5) = 2.4$, and $f\,(\xi_6) = 0.2$ with breakpoints specified as pictured in Fig 1.

The function f is piecewise linear, so that $f \in \mathcal{P}_{2,\xi,\nu}$ with $\nu_i = 1$, $i = 1, \ldots, 11$. Then, for f is of the form

$$f\,(x) = \alpha + \beta\,(x - \xi_1) + \sum_{i=2}^{l} \alpha_i\,(x - \xi_i)_+$$

one can show that, given the continuity and interpolation constraints, $\alpha = 1.3$, $\beta = f'(\xi_1) = 0$, $\alpha_i = 0$ for $i \notin \{4,5,6,7\}$, and

$$\begin{bmatrix} \alpha_4 \\ \alpha_5 \\ \alpha_6 \\ \alpha_7 \end{bmatrix} = \begin{bmatrix} \frac{1.1}{\Delta\xi_4} \\ -\frac{2.2}{\Delta\xi_5} - \frac{1.1}{\Delta\xi_4} \\ \frac{1.1}{\Delta\xi_6} + \frac{2.2}{\Delta\xi_5} \\ -\frac{1.1}{\Delta\xi_6} \end{bmatrix} = \begin{bmatrix} \frac{1.1}{\Delta\xi_4} \\ -\frac{2.2}{h} - \frac{1.1}{\Delta\xi_4} \\ \frac{1.1}{\Delta\xi_6} + \frac{2.2}{h} \\ -\frac{1.1}{\Delta\xi_6} \end{bmatrix}$$

When $h = \Delta\xi_5 \to 0$, we have

$$\begin{aligned} (x-\xi_5)_+ &\approx (x-\xi_6)_+ \\ \alpha_5 &\approx -\alpha_6 >> 1, \end{aligned}$$

leading to a loss of significance in the evaluation of the function. For example, if we were to choose $\xi_5 = 4.5$, $\xi_6 = 4.8$, making $h = 0.3$, with five significant decimal digit arithmetic, then we get

$$\alpha = 1.30000,\ \beta = 0.00000,\ \alpha_4 = 0.78571,\ \alpha_5 = -8.11905,\ \alpha_6 = 8.17949,\ \alpha_7 = -0.84615$$

and the other $\alpha_i = 0$. Evaluating the corresponding function at $x = 9.6$ yields 1.299987 rather than the correct value of 1.3. This error becomes larger as h approaches 0.

A quick remedy for the problem in 2.1 is to replace the truncated power basis with the set of hat functions:

$$H_i(x) = \begin{cases} (x-\xi_i)/\Delta\xi_{i-1}, & \xi_{i-1} < x \le \xi_i \\ (\xi_{i+1}-x)/\Delta\xi_i, & \xi_i < x \le \xi_{i+1} \\ 0 & otherwise \end{cases}$$

To utilize the hat functions, we augment $\xi_1, \ldots, \xi_12$ with two additional breakpoints: $\xi_0 \le \xi_1$ and $\xi_{12} \ge \xi_{11}$. Then,

$$\begin{aligned} f(x) = 1.3H_1(x) + \cdots + 1.3H_4(x) + 2.4H_5(x) \\ + 0.2H_6(x) + 1.3H_7(x) + \cdots + 1.3H_{11}(x) \end{aligned}$$

Even just using just two decimal digit arithmetic, we have $f(9.5) = 1.3$. Even as $h \to 0$, f is well represented using the hat functions as a basis. More generally speaking, *B-splines*, which are a generalization of these hat functions, overcome the issues that accompany the truncated power basis previously illustrated. The alternative basis is constructed by assembling linear combinations of the truncated power functions, forming a set of basis functions having "small" support: the functions vanish outside a small interval of their domain. In the following section, we will define the k^{th}-order B-splines as scaled *divided differences* of truncated power functions. We will also show that every subspace $\mathcal{P}_{k,\xi,\nu}$ has a basis consisting of these functions, bringing forth a B-spline representation of any pp function.

3 The B-spline representation of piecewise polynomial functions

We will present an introduction to B-splines and their properties in the section to follow. As these basis functions are defined in terms of divided differences of the truncated power functions discussed in the previous section, we must first review the definition and properties of the divided difference operator. While there are many ways of defining divided differences, the following (somewhat nonconstructive definition) is intuitive and adequate for our purposes here.

3.1 The divided difference

Definition 3.1. The k^{th} *divided difference* of a function g at the points $\tau_i, \ldots, \tau_{i+k}$, denoted

$$[\tau_i, \ldots, \tau_{i+k}]\, g,$$

is the leading coefficient of the polynomial of order $k+1$ which *agrees with* g at $\tau_i, \ldots, \tau_{i+k}$.

This leads us to the following definition:

Definition 3.2. Let $\{\tau_i\}_{i=1}^n$ denote a sequence of points. We say that a function f *agrees with* a function g at τ if, for every point τ which occurs m times in $\{\tau_i\}$, f and g agree up to m derivatives at τ; i.e.

$$f^{(i-1)}(\tau) = g^{(i-1)}(\tau), \quad i = 1, \ldots, m$$

3.1.1 Properties of the divided difference

Definition 3.1 yields the following properties of the k^{th} divided difference:

I. Let $\{p_i\}$, $i = 1, 2, \ldots$ denote a sequence of polynomials with p_i having order i. If p_i agrees with a function g at $\tau_1, \ldots, \tau_i$ for $i = k, k+1$, then

$$p_{k+1}(x) = p_k(x) + (x - \tau_1) \cdot \cdots \cdot (x - \tau_k)\, [\tau_1, \ldots, \tau_{k+1}]\, g$$

Proof. Note that $p_{k+1}(x) - p_k(x)$ is a polynomial of order $k+1$ and vanishes at $\tau_1, \ldots, \tau_k$, and, by definition, has $[\tau_1, \ldots, \tau_{k+1}]\, g$ as its leading coefficient. Consequently, $p_{k+1} - p_k$ is of the form

$$p_{k+1}(x) - p_k(x) = c \prod_{j=1}^{k} (x - \tau_j)$$

where

$$c = [\tau_1, \ldots, \tau_{k+1}]\, g$$

□

This tells the reader that interpolating polynomials may be constructed using divided differences by adding the interpolation points one by one, giving the *Newton form* of the n^{th} order polynomial which agrees with g at $\tau_1, \ldots, \tau_n$:

$$p_n(x) = \sum_{i=1}^{n} (x - \tau_1) \times \cdots \times (x - \tau_{i-1}) [\tau_1, \ldots, \tau_{i-1}] g$$

II. $[\tau_i, \ldots, \tau_{i+k}] g$ is symmetric in its arguments $\tau_i, \ldots, \tau_{i+k}$ (since the interpolating polynomial depends only on the points of interpolation and not the order in which they are specified.)

III. $[\tau_i, \ldots, \tau_{i+k}] g$ is a linear operator; if we let

$$f = \alpha g + \beta h$$

for some functions g, h and scalars α, β, then

$$[\tau_i, \ldots, \tau_{i+k}] f = \alpha [\tau_i, \ldots, \tau_{i+k}] g + \beta [\tau_i, \ldots, \tau_{i+k}] h$$

IV. If $f = gh$, then

$$[\tau_i, \ldots, \tau_{i+k}] f = \sum_{j=i}^{i+k} [\tau_i, \ldots, \tau_{i+k}] g [\tau_i, \ldots, \tau_{i+k}] h$$

V. If g is a polynomial of degree $\leq k$, then $[\tau_i, \vdots, \tau_{i+k}] g$ is constant as a function of $\tau_i, \vdots, \tau_{i+k}$. In particular,

$$[\tau_i, \vdots, \tau_{i+k}] g = 0 \quad \forall g \in \mathcal{P}_k$$

The polynomial which agrees with g at $\tau_i, \ldots, \tau_{i+k}$ or any other $k+1$ points must be g itself due to the uniqueness of interpolating polynomials. For more details, we refer the reader to A Practical Guide to Splines, DeBoor.

VI. $[\tau_i, \ldots, \tau_{i+k}] g$ is a continuous function of its $k+1$ arguments in case $g \in \mathcal{C}^{(k)}$ (that is, if g has k continuous derivatives.)

VII. If $g \in \mathcal{C}^{(k)}$, then there exists a point ξ in the smallest interval containing $\tau_i, \ldots, \tau_{i+k}$ such that

$$[\tau_i, \ldots, \tau_{i+k}] g = \frac{g^{(k)}(\xi)}{k!}$$

VIII.

$$[\tau_i, \ldots, \tau_{i+k}] g = \begin{cases} \frac{g^{(k)}(\tau_i)}{k!} & \tau_i = \cdots = \tau_{i+k}, \quad g \in \mathcal{C}^{(k)} \\ \frac{[\tau_i, \ldots, \tau_{r-1}, \tau_{r+1}, \ldots, \tau_{i+k}] g - [\tau_i, \ldots, \tau_{s-1}, \tau_{s+1}, \ldots, \tau_{i+k}] g}{\tau_s - \tau_r} & \tau_r, \tau_s \text{ are any two distinct points in } \{\tau_i, \ldots, \tau_{i+k}\} \end{cases}$$

IX. Let $\tau = \{\tau_i\}$, $i = 1, \ldots, n$ be a nondecreasing sequence, and let g be a "smooth enough" function. Define the restriction of g to the sequence τ to be the vector $g|_\tau = (g_1, \ldots, g_n)^T$, where

$$g_i = g^{(r)}(\tau_i), \quad r = \max\{j : \tau_{i-j} = \tau_i\}$$

For any $1 \leq r \leq s \leq n$, there exist constants $\{d_i\}$, $i = 1, \ldots, n$ which depend only on r, s, and τ and are free of g such that

$$[]\, g = \sum_i = 1^n d_i g_i \tag{11}$$

Proof. The proof is done by induction on $s - r$. The statement is obvious for the case where $s - r = 0$; all $d_i = 0$ except the constant corresponding to $g(\tau_r)$, which is equal to 1. For any $r < s$, if $\tau_r = \cdots = \tau_s$, then

$$[\tau_r, \ldots, \tau_s]\, g = \frac{g^{(s-r)}(\tau_r)}{(s-r)!}$$

so that 11 holds with a single $d_j = \frac{1}{(s-r)!}$ and all other $d_i = 0$. The index $j \leq s$ of the nonzero constant depends only on the number of τ_j with $j < r$ and $\tau_j = \tau_r$.

Assume that 11 holds for $s - r < k$. Consider $s - r = k$ with $\tau_r < \tau_s$. By the induction hypothesis, there are coefficients $\{d_i'\}$ and $\{d_i''\}$ □

In the following section, we will present B-splines as divided differences of the truncated power basis and present some of their properties which come as a result of the properties of the divided difference presented in this section.

4 A B-spline representation for pp functions

Definition 4.1. 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

$$B_{i,k,t}(x) = (t_{i+k} - t_i)\,[t_i, \ldots, t_{i+k}]\,(\cdot - x)_+^{k-1} \tag{12}$$

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.

4.1 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 3.1.1 V,

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

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 3.1.1 VIII, allows us to write

$$B_i(x) = [t_{i+1}, \ldots, t_{i+k}]\,(\cdot - x)_+^{k-1} - [t_i, \ldots, t_{i+k-1}]\,(\cdot - x)_+^{k-1} \tag{13}$$

For $x \in (t_j, t_{j+1})$, by 3.1.1 I,

$$\begin{aligned}
\sum_i B_i(x) &= \sum_{i=j+1-k}^{j} B_i(x) \\
&= \sum_{i=j+1-k}^{j} [t_{i+1}, \ldots, t_{i+k}]\,(\cdot - x)_+^{k-1} - \sum_{i=j+1-k}^{j} [t_i, \ldots, t_{i+k-1}]\,(\cdot - x)_+^{k-1} \\
&= [t_{j+1}, \ldots, t_{j+k}]\,(\cdot - x)_+^{k-1} - [t_{j+1-k}, \ldots, t_j]\,(\cdot - x)_+^{k-1} \\
&= 1 - 0
\end{aligned} \tag{14}$$

The last equality in 14 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 3.1.1 V,

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

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 (3.1.1 IV) 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

$$\begin{aligned}
[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-1} &= [t_i,\ldots,t_{i+k}]\,(t-x)\,(t-x)_+^{k-2} \\
&= \sum_{r=i}^{i+k} [t_i,\ldots,t_{i+r}]\,(t-x)\,[t_r,\ldots,t_{i+k}]\,(t-x)_+^{k-2} \\
&= \Big[\,[t_i]\,(t-x)\,\Big]\Big[\,[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-2}\,\Big] \\
&\qquad + \Big[\,[t_i,t_{i+1}]\,(t-x)\,\Big]\Big[\,[t_{i+1},\ldots,t_{i+k}]\,(t-x)_+^{k-2}\,\Big] \\
&= (t_i-x)\,[t_i,\ldots,t_{i+k}]\,(t-x)_+^{k-2} \\
&\qquad + 1\cdot[t_{i+1},\ldots,t_{i+k}]\,(t-x)_+^{k-2} \qquad (15)
\end{aligned}$$

since $[t_i,\ldots,t_j]\,(\cdot-x)=0$ for $j>i+1$. By 3.1.1 VIII,

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

and we may express 15 as

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

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

$$\frac{B_{i,k}\,(x)}{t_{i+k}-t_i}=\frac{x-t_i}{t_{i+k}-t_i}\frac{B_{i,k-1}\,(x)}{t_{i+k-1}-t_i}+\frac{t_{i+k}-x}{t_{i+k}-t_i}\frac{B_{i+1,k-1}\,(x)}{t_{i+k}-t_{i+1}} \qquad (16)$$

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 16, 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 4.1 I 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 I, II, and III demonstrate that a sequence of B-splines form a *partition of unity*: a set of non-negative functions which sum, pointwise, to one.

Example 4.1. Figure 2 show fives parabolic B-splines corresponding to the set of knots $\{0, 1, 1, 3, 4, 6, 6, 6\}$. It is clear that each spline has compact support and is non-negative on this support. The function values are provided at select domain values so that property II is evident, though it is worth noting that

$$\sum_{i=1}^{6} B_i(x) = 1$$

only on $\left[\frac{1}{6}, 1\right]$. In particular, they do not sum to 1 on $\left(0, \frac{1}{6}\right)$.

Each B_i is piecewise parabolic, with the breakpoints being locations of discontinuity of the function or one of its derivatives. B_5 is discontinuous at $x = 6$, as the knot at 6 is repeated three times in the knot sequence defining B_5: $\{t_5, t_6, t_7, t_8\}$.

B_1, B_2, and B_4 have discontinuous first derivatives since 1 is repeated twice in the knots defining B_1 and B_2, and 6 is repeated twice in the knots defining B_4. The relationship between knot replication and smoothness will be discussed in more detail in the following section.

4.1.1 The Curry-Schoenberg Theorem

There is an extensive body of literature

Definition 4.2. A *spline function of order k with knot sequence t* is any linear combination of B-splines of order k for the knot sequence t. We denote this set of functions by

$$\mathcal{S}_{k,t} = \left\{ \sum_i \alpha_i B_{i,k,t} : \quad \alpha_i \in \Re \right\}$$

Theorem 4.1. *Curry-Schoenberg Theorem For a strictly increasing sequence* $\xi = \{\xi_i\}$, $i = 1, \ldots, l+1$ *and a sequence of non-negative integers* $\nu = \{\nu_j\}$, $j = 2, \ldots, l$ *with* $\nu_j \leq k$ *for all* j, *define*

$$\begin{aligned} n &= k + \sum_{i=2}^{l} (k - \nu_i) = kl - \sum_{i=2}^{l} \nu_i \\ &= \dim(\mathcal{P}_{k,\xi,\nu}) \end{aligned}$$

and let $t = \{t_i\}$, $i = 1, \ldots, n+k$ *be a non-decreasing sequence with*

I.

$$\begin{aligned} t_1 \leq t_2 \leq \cdots \leq t_k \leq \xi_1 \\ \xi_{l+1} \leq t_{n+1} \leq \cdots \leq t_{n+k} \end{aligned}$$

II. ξ_i *occurs exactly* $k - \nu_i$ *times in* t *for* $i = 2, \ldots, l$.

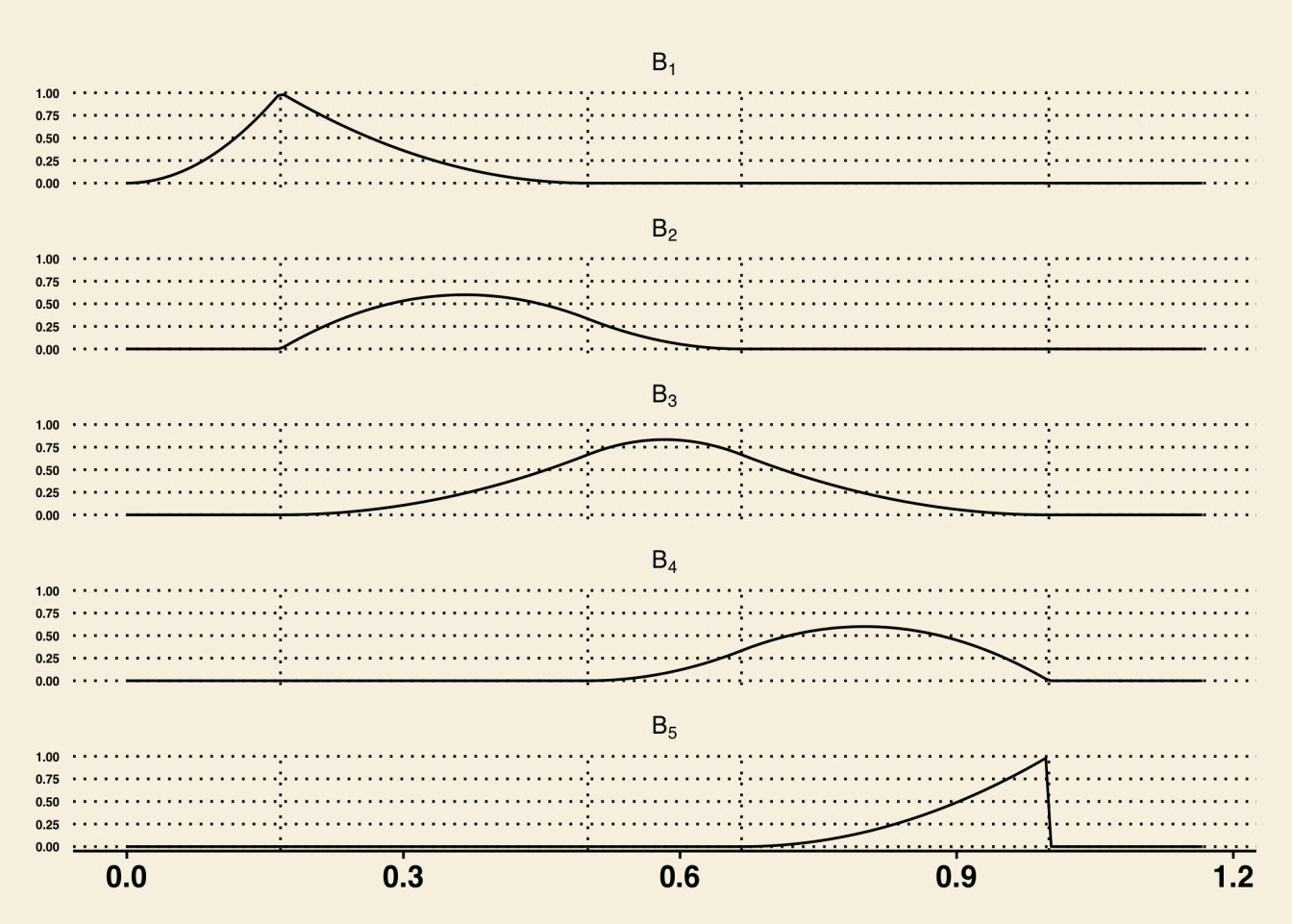


Figure 2: Parabolic B-splines corresponding to knot sequence $\{0, 1, 1, 3, 4, 6, 6, 6\}$, illustrating the connection between knot multiplicity and smoothness.

Then, the sequence of B-splines of order k for the knot sequence t, $B_1, \ldots, B_n$, is a basis for $\mathcal{P}_{k,\xi,\nu}$ on the domain $[t_k, t_{n+1}]$.

Theorem 4.1 gives an explicit prescription for the construction of a B-spline basis for any particular pp space, $\mathcal{P}_{k,\xi,\nu}$, via the specification of t. These functions were named (B for 'basis') for this theorem. The choice of t controls the smoothness of the corresponding basis functions via knot multiplicities; the number of knots at a given breakpoint dictates the amount of smoothness at the breakpoint. Fewer knots placed at a breakpoint leads to more continuity conditions such that

$$[\#\text{ continuity conditions at } \xi] + [\#\text{ knots at } \xi] = k$$

A knot with multiplicity k yields a basis with no continuity conditions at that point, and a point where there are no knots forces k continuity conditions there.

4.1.2 The Curry-Schoenberg Proof

Proof. We begin by showing that $B_i \in \mathcal{P}_{k,\xi,\nu}$ for all i as functions defined on $[t_k, t_{n+1}]$. Each $B_i = B_{i,k,t}$ is defined as the k^{th} divided difference of $(t-x)_+^{k-1}$ at $t_i, \ldots, t_{i+k}$ times a scalar, so by 3.1.1 IX, we can find coefficients $d_i, \ldots, d_{i+k}$ which depend only on $t_i, \ldots, t_{i+k}$ so that for any smooth function g,

$$[t_i, \ldots, t_{i+k}]\, g = \sum_{r=i}^{i+k} d_r g^{(j_r)}(t_r) \tag{17}$$

with

$$j_r = \max\left\{ s : \;\; r - s \geq i \text{ and } t_{r-s} = t_r; \;\; r = i, \ldots, i+k \right\}.$$

Using this, we may write

$$B_i(x) = (t_{i+k} - t_i) \sum_{r=i}^{i+k} d_r (t_r - x)_+^{k-j_r-1} \frac{(k-1)!}{(k-j_r-1)!}. \tag{18}$$

From 18, it follows immediately that B_i is a pp function of order k with breakpoints $t_i, \ldots, t_{i+k}$ (and consequently some of the $\{\xi_i\}$.)

Now, we must establish the number of continuous derivatives of each B_i at each of its breakpoints, ξ_i, $i = 2, \ldots, l$. For any B_i, there cannot be a jump in its s^{th} derivative across ξ_j unless

$$\xi_j = t_r \text{ and } k - 1 - j_r = s$$

for some $r \in \{i, \ldots, i+k\}$. Since

$$j_r = \#\;\; t_r = t_m \;\; i \leq m < r,$$

j_r must be less than $k - \nu_j$, the total number of $\{t_m\}$ coinciding with ξ_j and hence equal to t_r, due to the construction of t. However, we must have $s \geq \nu_j$, and thus

$$D^m B_i\left(\xi_j^+\right) - D^m B_i\left(\xi_j^-\right) = 0, \qquad m = 0, \ldots, \nu_j - 1.$$

So $B_i \in \mathcal{P}_{k,\xi,\nu}$ for all $i = 1, \ldots, n$. We now only need to show that the B_i are linearly independent to complete the proof.

Lemma 4.2. *de Boor, Fix (1973) Define the linear functional λ_i by*

$$\lambda_i f = \sum_{r=0}^{k-1} (-1)^{k-r-1} \psi^{(k-r-1)}(\tau_i)\, D^r f(\tau_i)$$

where

$$\psi(t) = \frac{(t_{i+1} - t) \times \cdots \times (t_{i+1} - t)}{(k-1)!}$$

and where $\tau_i \in (t_i, t_{i+k})$. Then

$$\lambda_i B_j = \delta_{ij} \quad \textit{for all} \quad i, j.$$

Proof. It follows from its definition that $\lambda_i B_j$ is a pp function as a function of τ_i with breakpoints at the $\{t_i\}$. If we assume that $\tau_i \notin \{t_j\}$ for all i, then it is sufficient to show that $\lambda_i B_j = \delta_{ij}$.

By 17 and 18,

$$\lambda_i B_j = (t_{j+k} - t_j) \sum_{r=j}^{j+k} d_r \lambda_i \left[D_s^{j_r} (s - \cdot)_+^{k-1} \right] |_{s=t_r}$$

where $D_s^{j_r}$ denotes the operator for j_r-fold differentiation with respect to s. Now consider

$$\lambda_i (s - \cdot)_+^{k-1};$$

For $s < \tau_i$, $f(x) = (s-x)_+^{k-1}$ vanishes near τ_i, so that

$$\lambda_i f = 0.$$

For $s > \tau_i$, f agrees with $(s-x)^{k-1}$ in a neighborhood of τ_i, while

$$\begin{aligned}
\lambda_i (s-x)^{k-1} &= \sum_{r=0}^{k-1} (-1)^{k-r-1} \psi^{(k-r-1)}(\tau_i) (k-1) \times \cdots \times (k-r) (-1)^r (s-\tau_i)^{k-r-1} \\
&= \sum_{r=0}^{k-1} \frac{(k-1)!}{(k-r-1)!} \psi^{(k-r-1)}(\tau_i) (s-\tau_i)^{k-r-1} (-1)^{k-1} \\
&= (-1)^{k-1} (k-1)! \sum_{r=0}^{k-1} \frac{(s-\tau_i)^{k-r-1}}{(k-r-1)!} \psi^{(k-r-1)}(\tau_i) \qquad (19)
\end{aligned}$$

We recognize the sum in 19 as the truncated Taylor series in s of order k for ψ. Since ψ is itself a pp of order k, the sum must agree with ψ at s due to the uniqueness of interpolating polynomials. This implies that

$$\lambda_i (s - \cdot)^{k-1} = (-1)^{k-1} (k-1)! \psi(s).$$

Accounting for the three cases, we have that

$$\lambda_i (s - \cdot)_+^{k-1} (-1)^{k-1} (k-1)! \psi(s) (s-\tau_i)_+^0$$

since

$$D_s^r \lambda_i (s - \cdot)_+^{k-1} = \lambda_i \left[D_s^r (s - \cdot)_+^{k-1} \right].$$

Now we have that

$$\lambda_i B_j = (t_{j+k} - t_j) (-1)^{k-1} (k-1)! [t_j, \ldots, t_{j+k}] \phi_i$$

where $\phi_i(s) = \psi(s) (s-\tau_i)_+^0$. Taking $[t_j, \ldots, t_{j+k}] \phi_i$ to be the leading coefficient of the order $k+1$ polynomial which agrees with ϕ_i at $t_j, \ldots, t_{j+k}$, if we assume that

$$t_i < \tau_i < t_{i+k}$$

then

I. ϕ_i agrees with ψ at $t_{i+1}, t_{i+2}, \ldots$, and as a polynomial of order $k+1$, ψ has leading coefficient 0, so

$$[t_j, \ldots, t_{j+k}]\, \phi_i = 0, \qquad j = i+1, i+2, \ldots$$

II. ϕ_i agrees with 0 at $t_{i+k-1}, t_{i+k-2}, \ldots$, so that

$$[t_j, \ldots, t_{j+k}]\, \phi_i = 0$$

III. ϕ_i agrees with the $(k+1)^{st}$ order polynomial

$$p(x) = \frac{\psi(x)(x - t_i)}{(t_{i+k} - t_i)}$$

at $t_i, \ldots, t_{i+k}$.

Together, I - III show that

$$\lambda_i B_j = \delta_{ij} \text{ for all } j.$$

□

It is worth noting that we may apply Lemma 4.2 under the assumption that we can find τ_i in the open interval (t_i, t_{i+k}) - that is, if $t_i < t_{i+k}$ for all i. This case is of little interest, however; in the case that $t_i = t_{i+k}$, then $[t_i, t_{i+k}]$ is just a point. From 4.1 I, $B_i(x) = 0$ anywhere outside $[t_i, t_{i+k}]$, so if $t_i = t_{i+k}$, then it follows that B_i is simply the zero function. □

The recipe provided by Theorem 4.1 for constructing the knot vector t such that the resulting $\{B_i\}$ form a basis for $\mathcal{P}_{k,\xi,\nu}$ leaves open the choice for the first k knots and the last k knots. It is conventional, due likely to convenience, to specify

$$t_1 = \cdots = t_k = \xi_1, \quad t_{n+1} = \cdots = t_{n+k} = \xi_{l+1},$$

and by setting $\nu_1 = \nu_{n+1} = 0$, one can include these knots according to the same method of choice of the other knots. This specification imposes no continuity conditions at the endpoints of the interval of interest; this is, however, consistent with the manner in which the B-splines are defined, spanning $\mathcal{P}_{k,\xi,\nu}$ only on $[t_k, t_{n+1}]$.

Definition 4.3. The *B-representation of* $f \in \mathcal{P}_{k,\xi,\nu}$ consists of

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

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

respectively.

II. The knot vector $t = \{t_i\}$, $i = 1, \ldots, n+k$ with elements arranged in increasing order, constructed according to Theorem 4.1, via ξ and ν.

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

Given I, II, and III in 4.3, the function value at $x \in [t_k, t_{n+1}]$ is given by

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

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

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

Example 4.2 (Conversion between function representations)**.** Conversion from the B-representation to the pp-representation is straightforward the availability of stable evaluation and differentiation of the function. The $l+1$ distinct points among $t_k, \ldots, t_{n+1}$ are taken to be $\xi_1, \ldots, \xi_{n+1}$; the polynomial coefficients $\{c_{ji}\}$, $i = 1, \ldots, l$, $j = 1, \ldots, k$ are obtained by calculating $D^{j-1} f\left(\xi_i^+\right)$.

The conversion in the opposite direction, however, is much more difficult because there is no implicit information in the pp-representation about the smoothness of the function at its breakpoints, nor is it possible to numerically derive the minimal knot multiplicity necessary to represent the function as a spline using finite precision arithmetic. However, if it is known that f lies within $\mathcal{P}_{k,\xi,\nu}$ for a certain ν, then one may use Theorem 4.1 to construct t, and then α using Lemma 4.2. If such a representation for f is accessible, then the $\{\tau_i\}$ may be chosen to coincide with one of the $\{\xi_i\}$ so that the necessary derivatives may be extracted directly from the pp-representation.

Consider $f(x) = (x-3)(x-6)(x-9)$. Clearly f is a pp function of order $k = 4$. Suppose we wish to construct the B-representation of f on $[0, 10]$ as a cubic spline using interior knots $1, 2, \ldots, 9$. Then, we take

$$\begin{aligned} t_1 = t_2 = t_3 = t_4 &= 0 \\ t_{4+i} &= i, \qquad\qquad i = 1, \ldots, 9 \\ t_{13+1} = t_{13+2} = t_{13+3} = t_{13+4} &= 10 \end{aligned}$$

Using Lemma 4.2, we may let

$$\tau_i = t_{i+2}$$

for all i, which gives

$$\alpha_i = f(t_{i+2}) + \frac{1}{3}(\Delta t_{i+2} - \Delta t_{i+1}) f'(t_{i+2}) - \frac{1}{3}\Delta t_{i+2}\Delta t_{i+1}\frac{f''(t_{i+2})}{2}$$

4.2 More B-spline Properties

The properties that follow illustrate how a spline $f \in (S)_{k,t}$ is related to its corresponding sequence of B-spline coefficients.

IV. **A recurrence relation for $B_{i,k}$:** The evaluation of a B-spline via its definition of a divided difference amounts to evaluating the function in terms of the truncated power basis. This approach is then plagued by the same computational issues that one hoped to sidestep with the introduction of the B-spline basis in the first place.

Via the derivation of 4.1 III, we showed that the B-splines satisfy the recurrence relation

$$B_{i,1}(x) = \begin{cases} 1, & t_i \le x < t_{i+1} \\ 0, & otherwise \end{cases} \tag{20}$$

$$B_{i,k}(x) = \frac{x - t_i}{t_{i+k-1} - t_i} B_{i,k-1}(x) + \frac{t_{i+k} - x}{t_{i+k} - t_{i+1}} B_{i+1,k-1}(x) \tag{21}$$

so that computing $B_{ik}(x)$ simply requires repeatedly taking positive linear combinations of positive quantities, starting with 20. Using this, one may express a series of B-splines as a series of lower order, but with coefficients that are polynomials in the spline argument. In general, we can write

$$\sum_i \alpha_i B_{i,k}(x) = \sum_i \alpha_i^{[j]}(x) B_{i,k-j}(x) \tag{22a}$$

$$\alpha_i^{[j]}(x) \equiv \begin{cases} \alpha_i, & j = 0 \\ \frac{(x-t_i)\alpha_i^{[j-1]}(x) + \left(t_{i+k-j} - x\right)\alpha_{i-1}^{[j-1]}(x)}{t_{i+k-j} - t_i}, & j \ge 1 \end{cases} \tag{22b}$$

and, in fact, $\alpha_i^{[k-1]}$ is a polynomial of degree $< k$ which agrees with $f = \sum_i \alpha_i B_{i,k}$ on $[t_i, t_{i+1}]$.

One may encounter 20 and 21 presented as the definition of the set of B-splines; properties 4.1 I and III follow immediately from 20 and 21, while II, is shown by induction. However, the smoothness properties of the $\{B_i\}$ presented in Theorem 4.1 are not as obvious from 20 and 21 as is their derivation from the properties of the divided difference, 3.1.1 I and 3.1.1 VI. We might even say that it is unexpected that the combination of $B_{i,k}$ and $B_{i,k-1}$ as given in 21 produces a function which has one more continuous derivative than either of the functions themselves.

V. **Differentiation of $f \in \mathcal{S}_{k,t}$** A B-spline series may be differentiated by differencing the coefficients. (This fact will prove very useful in sections to follow.) Since the derivative of the truncated power function $g(x) = (t - x)_+^{k-1}$ is given by

$$D_x g(x) = D_x (t - x)_+^{k-1} = -(k-1)(t - x)_+^{k-2}$$

and by 13 and 17, we may write the derivative of the i^{th} order-k B-spline as

follows:

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

This allows us to write

$$\begin{aligned} D_x\left[\sum_i \alpha_i B_i(x) \right] &= \sum_i \alpha_i DB_{i,k}(x) \\ &= \sum_i (k-1)\frac{\alpha_i - \alpha_{i-1}}{t_{i+k-1}-t_i} B_{i,k-1}(x) \end{aligned} \tag{23}$$

Note that the limits on the previous summation in 23 are left unspecified; the formula is written for biinfinite sums, and their application to finite sums is accessible after they are written formally as biinfinite sums by augmenting the appropriate zero terms.

However, if we are interested in a particular interval over the domain, say $[t_r, t_s]$, then for $x \in [t_r, t_s]$,

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

since $B_{i,k-1}(x) = 0$ for all $i \notin \{r-k+2,\ldots,s-1\}$ when $t_r \le x \le t_s$. Applying 23 j times gives the j^{th} derivative of $f \in \mathcal{S}_{k,t}$, which has form

$$D^j\left[\sum_i \alpha_i B_{i,k}(x) \right] = \sum_i \alpha_i^{(j+1)} B_{i,k-j} \tag{24a}$$

$$\alpha_i^{(j+1)} \equiv \begin{cases} \alpha_i, & j = 0 \\ \frac{\alpha_i^{(j)} - \alpha_{i-1}^{(j)}}{(t_{i+k-j}-t_i)/(k-j)}, & j \ge 1 \end{cases} \tag{24b}$$

Proof. We proceed by induction on j. We have already shown the case for $j = 1$ in the derivation of 23. Assume that the statement holds for some $j^* > 1$, so that we have

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

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

$$\begin{aligned}
D^{j^*+1}\left[\sum_i \alpha_i B_{i,k}\right] &= \sum_i \frac{\alpha_i^{(j^*)} - \alpha_{i-1}^{(j^*)}}{\left(t_{i+k-j^*} - t_i\right)/\left(k-j^*\right)} DB_{i,k-j^*} \\
&= \sum_i \alpha_i^{(j^*)} DB_{i,k-j^*} \\
&= \sum_i \alpha_i^{(j^*)} \left(k-\left(j^*+1\right)\right)\left[\frac{B_{i,k-(j^*+1)}}{t_{i+k-(j^*+1)} - t_i} - \frac{B_{i+1,k-(j^*+1)}}{t_{i+k-(j^*+1)+1} - t_{i+1}}\right] \\
&= \sum_i \frac{\alpha_i^{(j^*)} - \alpha_{i-1}^{(j^*)}}{\left(t_{i+k-(j^*+1)} - t_i\right)/\left(k-\left(j^*+1\right)\right)} B_{i,k-(j^*+1)} \\
&= \sum_i \alpha_i^{(j^*+1)} B_{i,k-(j^*+1)}
\end{aligned}$$

□

The choice to write $k-j$ as a divisor in the denominator lends to the interpretation of 24a as a difference quotient, with the quantity

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

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

VI. **The condition of the B-spline basis** To characterize the condition of the B-spline basis, we must bound the spline in terms of its B-spline coefficients and, additionally, bound the B-spline coefficients in terms of the spline. The first bound is easy to find; we need only find constants $m > 0$ and M such that

$$m \max_i |\alpha_i| \leq m \max_t |\sum_i \alpha_i B_{i,k}(t)| \leq M \max_i |\alpha_i|$$

holds for any coefficient vector $\alpha = \{\alpha_i\}$. Since $\{B_{i,k}\}$ form a partition of unity, we have

$$|\sum_i \alpha_i B_{i,k}(t)| \leq \sum_i |\alpha_i| B_{i,k}(t) \leq \sum_i \max_i |\alpha_i| B_{i,k}(t) = \max_i |\alpha_i|.$$

So for $M = 1$, the second inequality holds. To show that the first inequality holds, let $f = \sum_i \alpha_i B_{i,k}$. From Theorem 4.1, we know that we may write

$$\alpha_i = \lambda_{ik} f = \sum_{j=1}^{k} \frac{-D^{j-1}\psi_{ik}(\tau)}{(k-1)!} D^{k-j} f(\tau) \tag{25}$$

for some $\tau \in [t_i, t_{i+k})$. To bound this sum in terms of $\max_t |f(t)|$, suppose that

$$\tau \in [t_l, t_{l+1}) \subset [t_i, t_{i+k}).$$

Then for some constant c_k depending on k only, and for any pp function p of order k,

$$|D^j p(\tau)| \leq c_k (\Delta t_l)^{-1} \max_{t_l \leq t \leq t_{l+1}} |p(t)|. \tag{26}$$

For the case in which $\Delta t_l = 1$, the existence of c_k follows from the fact that the space of pp functions of order k is finite dimensional; from this, the statement follows for arbitrary Δt_l via scaling. Since f agrees with some polynomial of degree $< k$ on $[t_l, t_{l+1})$,

$$|D^j f(\tau)| \leq c_k (\Delta t_l)^{-1} \max_{t_i \leq t \leq t_{i+k}} |f(t)| \tag{27}$$

However, $\psi_{ik} = (t_{i+1} - \cdot)(t_{i+2} - \cdot)\ldots(t_{i+k-1} - \cdot)$ is also a polynomial of degree $< k$, and

$$\max_{t_l \leq t \leq t_{l+1}} \psi(t) \leq c_k^* |\Delta t_{l*}|^{k-1} \tag{28}$$

for some c_k^* which depends on k only and $[t_{l^*}, t_{l^*+1})$ a largest interval of that form in $[t_{l^*}, t_{l^*+1})$. Taking $l^* = l$, from 26 with $p = \psi_{ik}$ and from 27 we obtain the bound

$$|D^{\nu-1}\psi(\tau) D^{k-\nu} s(\tau)| \leq (c_k)^2 c_k^* \max_{t_i \leq t \leq t_{i+k}} |f(t)|.$$

If we sum these bounds over ν and divide by $(k-1)!$, we have

$$|\alpha_i| = |\lambda_{ik} f| \leq c \max_{t_i \leq t \leq t_{i+k}} |f(t)|,$$

with c depending only on k.

Properties 4.2 V and VI demonstrate how closely a spline function is modeled by its B-spline coefficients. In particular, the formula for differentiation of a spline function given by V tells us that derivatives of a spline function may be obtained by simply differencing its coefficients. This fact will motivate the difference penalties discussed in section P-spline penalties to follow.

In the following section, we will consider a simple varying-coefficient model with a single, univariate regressor variable, $x(t)$, and demonstrate the mechanics of smoothing the coefficient function using a univariate B-spline basis. Later in the chapter, we will show how one can easily extend univariate B-splines to two dimensions (or of any arbitrary higher dimension) using tensor products and use these to generalize the single dimension varying coefficient model to a two-dimensional varying coefficient model for a bivariate surface. The P-spline setting makes this extension quite straightforward, requiring only a bit of care in constructing the penalty scheme for the basis coefficients.

5 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

$$y(t) = f(t) + \epsilon(t) \tag{29}$$

where ϵ is a zero-mean error process. Modeling the mean function as a q^{th}-order B-spline, we can rewrite 29 as

$$y(t) = \sum_{j=1}^{q+K} \alpha_j B_j(t) + \epsilon(t) \tag{30}$$

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 $m \times (K+q)$ 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 $K+q$ unknown basis coefficients. We take $\hat{\alpha}$ to be the minimizer of

$$S = \sum_{i=1}^{m} \left(y_i - \sum_{j=1}^{K+q} \alpha_j B_j \left(t_i \right) \right)^2$$
$$= \left| y - B\alpha \right|^2 \tag{31}$$

[not sure about the indexing on the basis function coefficients. If there are K basis functions, should I not only need K coefficients?] Differentiating 31 with respect to α and setting equal to zero, we obtain normal equations

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

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

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

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

To extend the varying intercept model 29 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

$$E\left[Y\left(t \right) \right] = \beta_0 \left(t \right) + \beta_1 \left(t \right) x\left(t \right) \tag{32}$$

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 29, requiring only minor adjustments to the design matrix. In matrix notation as described in 31, the mean vector may be written

$$\mu = B\alpha_0 + diag\left\{ x\left(t \right) \right\} B\alpha_1 \tag{33}$$

where $diag\{x(t)\}$ 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 $U = diag\{x(t)\} B$, 33 becomes

$$\mu = [B|U] \left(\alpha_0^T, \alpha_1^T\right)^T \tag{34}$$
$$\equiv Q\alpha \tag{35}$$

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

$$S = |y - Q\alpha|^2 \tag{36}$$

which has explicit solution

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

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

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

and the estimate of the smooth mean function is obtained via

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

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

6 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. [**?**] 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. [**?**] 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

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

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

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

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

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

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

6.1 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. In the following section, we present penalized B-splines, or *P-splines*, as an approach to nonparametric smoothing; they circumvent any complexity associated with constructing such penalty matrices by ommitting derivatives and integrals altogether. Instead, smoothness is imposed via a discrete penalty matrix which is simple to compute, as it is based on finite difference formulas. This approach achieves smoothness in fitted functions in two ways:

I. Use a rich B-spline basis with equally spaced knots to purposefully overfit the smooth coefficient vectors, allowing one to dodge the difficulty of choosing the optimal set of knots.

II. Attach a difference penalty to the goodness of fit measure to prevent overfitting and make a potentially ill-conditioned fitting procedure a well-conditioned one.

Using the properties of B-splines derived in B-spline section, it is relatively straightforward to show that nothing is lost by using the simplified penalty, and that for second order differences, the P-splines of Marx and Eilers 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; the P-spline smoothing framework makes the use of a penalty of any arbitrary order nearly seamless.

Consider the varying intercept-only model defined in 29 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 the fitted curve to exhibit more variation than the data reasonably justify. Then 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

$$\sum_{i=1}^{m}\left\{y_i - \sum_{j=1}^{K}\beta_j B_j\left(t_i\right)\right\}^2 + \lambda\int_{t_{min}}^{t_{max}}\left\{\sum_{j=1}^{n}\beta_j B_j''\left(t\right)\right\}^2 dt. \tag{37}$$

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. 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

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

[**?**] 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 $\mu = \beta_0\left(t\right) = B\alpha$ 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.

Difference penalties have been utilized for nearly a century, having been a proposed method for smoothing curves by [**?**]; more recent applications are outlined in [**?**], [**?**], and [**?**]. 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. [**?**] used third-degree B-splines and the following penalty:

$$h^2 P = \lambda \int_{t_{min}}^{t_{max}} \left\{ \sum_j \alpha_j B''_{j,3}(t) \right\}^2 dt \tag{38}$$

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

$$h^2 P = \lambda \int_{t_{min}}^{t_{max}} \sum_j \sum_k \Delta^2 \alpha_j \Delta^2 \alpha_k B_{j,1}(t) B_{k,1}(t) \, dt \tag{39}$$

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

$$\begin{aligned} h^2 P &= \lambda \int_{t_{min}}^{t_{max}} \left[\left\{ \sum_j \Delta^2 \alpha_j B_j(t,1) \right\}^2 + 2 \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} B_j(t,1) B_{j-1}(t,1) \right] dt \\ &= \lambda \left[\sum_j \left(\Delta^2 \alpha_j \right)^2 \int_{t_{min}}^{t_{max}} B_j^2(t,1) \; dt + 2 \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} \right] \end{aligned} \tag{40}$$

or

$$\begin{aligned} h^2 P = \lambda \sum_j \left(\Delta^2 \alpha_j \right)^2 \int_{t_{min}}^{t_{max}} B_{j,1}^2(t) \, dt + 2\lambda \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} \\ + \int_{t_{min}}^{t_{max}} B_{j,1}(t) B_{j-1,1}(t) \, dt \end{aligned} \tag{41}$$

which can be written as

$$h^2 P = \lambda \left\{ c_1 \sum_j \left(\Delta^2 \alpha_j \right)^2 + c_2 \sum_j \Delta^2 \alpha_j \Delta^2 \alpha_{j-1} \right\} \tag{42}$$

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

$$\begin{aligned} c_1 &= \int_{t_{min}}^{t_{max}} B_{j,1}^2(t)\,dt \\ c_2 &= \int_{t_{min}}^{t_{max}} B_{j,1}(t)\,B_{j-1,1}(t)\,dt \end{aligned} \tag{43}$$

O'Sullivan's ridge-like B-spline penalty 38 can be written as a linear combination of Marx and Eilers' difference penalty **??** and the sum of the cross products of neighboring second differences. The second term in 42 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 42 is used in the penalty. The added complexity is a consequence of overlapping B-splines, and complexity 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

$$\begin{aligned} S_\lambda &= |y - B\alpha|^2 + \lambda |D_d \alpha|^2 \\ &= (y - B\alpha)^T (y - B\alpha) + \lambda \alpha^T P \alpha \end{aligned} \tag{44}$$

where

$$P = D_d^T D_d$$

and the elements of B are given by $b_{ij} = B_j(t_i)$, as defined in 31. Taking derivatives on both sides of 44 with respect to α gives

$$\begin{aligned} \frac{\partial}{\partial \alpha} S_\lambda &= \frac{\partial}{\partial \alpha} \left(\alpha^T B^T B \alpha - 2 y^T B^T \alpha + \lambda \alpha^T D_k^T D_k \alpha \right) \\ &= 2 B^T B \alpha - 2 B^T y + 2 \lambda D_d^T D_d \alpha \\ &= \left(B^T B + \lambda D_d^T D_d \right) \alpha - B^T y \end{aligned} \tag{45}$$

and setting equal to zero yields normal equations:

$$B^T y = \left(B^T B + \lambda D_d^T D_k \right) \alpha, \tag{46}$$

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

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

When $\lambda = 0$, we have the standard normal equations of linear regression with a B-spline basis, and with $k = 0$ 46 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.

6.1.1 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 32 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 44 must be further modified to accommodate multiple tuning parameters, λ_i, $i = 0, 1$. The objective function then becomes

$$\begin{aligned} S_\lambda^* &= |y - Q\alpha|^2 + \lambda_0|D_{d_0}\alpha_0|^2 + \lambda_1|D_{d_1}\alpha_1|^2 \\ &= |y - Q\alpha|^2 + |\alpha^T P\alpha|^2 \end{aligned} \tag{47}$$

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 47 is given by

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

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

$$H = Q\left(Q^TQ + P\right)^{-1} Q^T. \tag{48}$$

[Figure 3 Need to explain figure 3 here.]

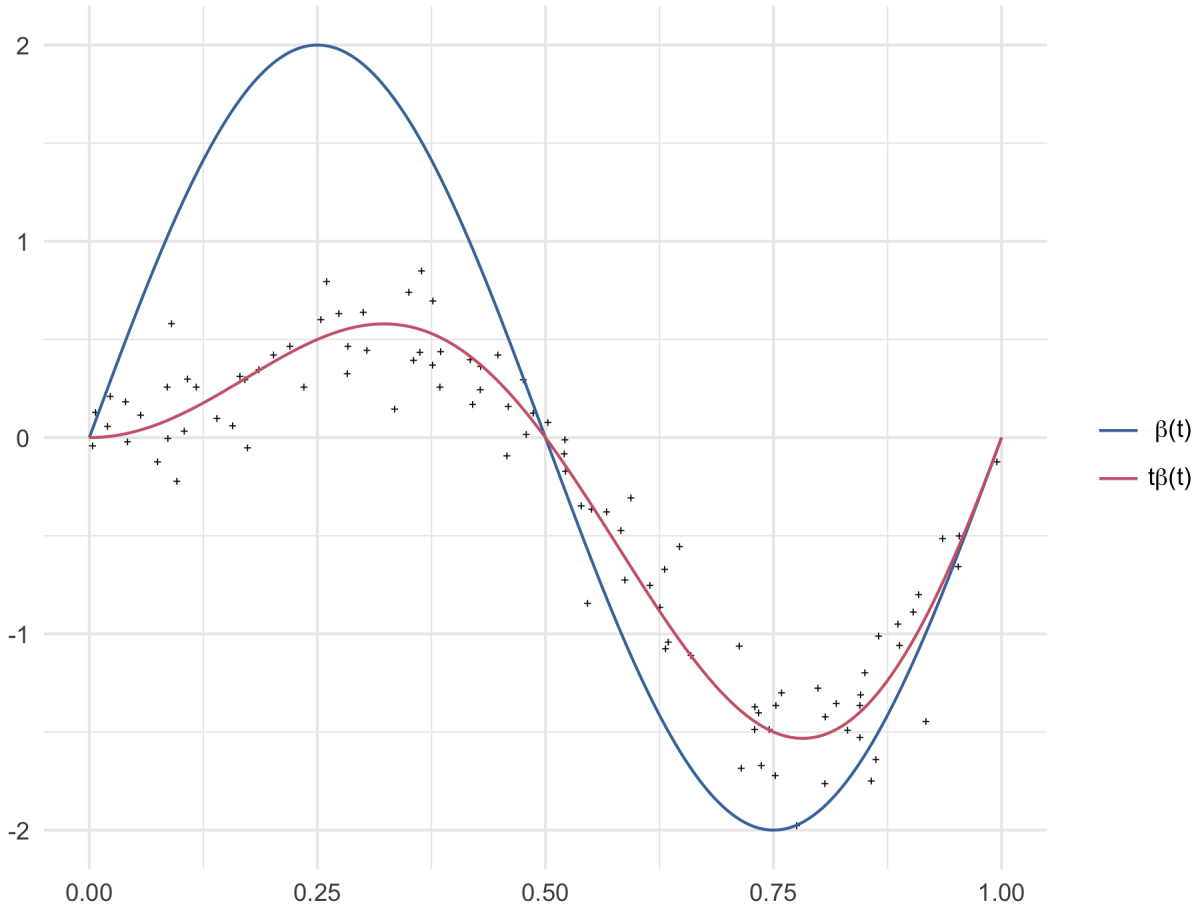
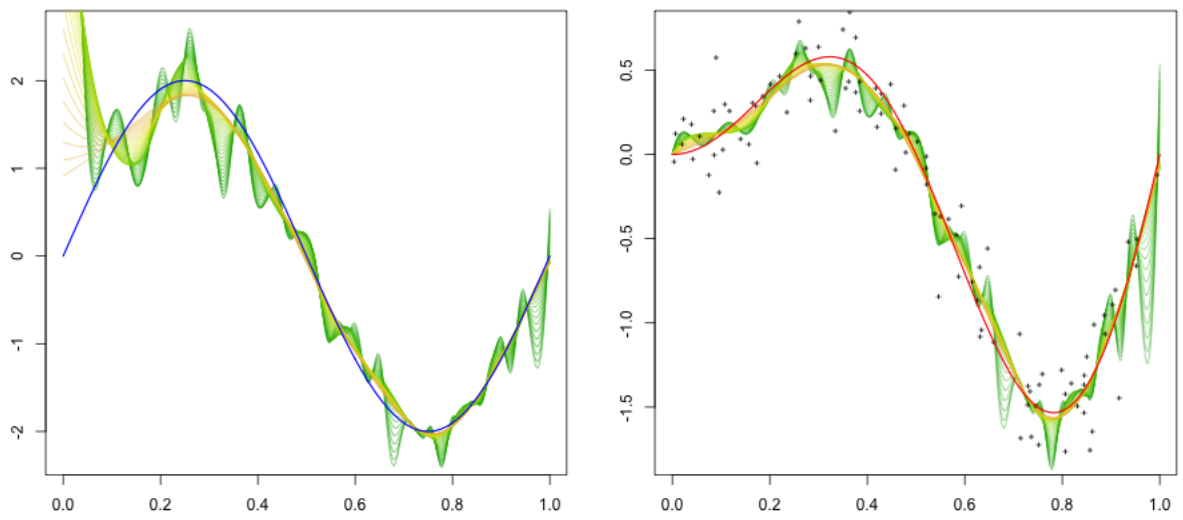


Figure 3: *100 simulated data points where* $y(t) = t\beta(t) + 0.2\epsilon(t)$ *where* ϵ *is a white noise process with unit variance, and* $\beta(t) = 2\sin(2\pi t)$.



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

Figure 4: *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* $k = 3$.

6.2 The impact of the difference penalty

The properties discussed in Section 4.2 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 5. 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.

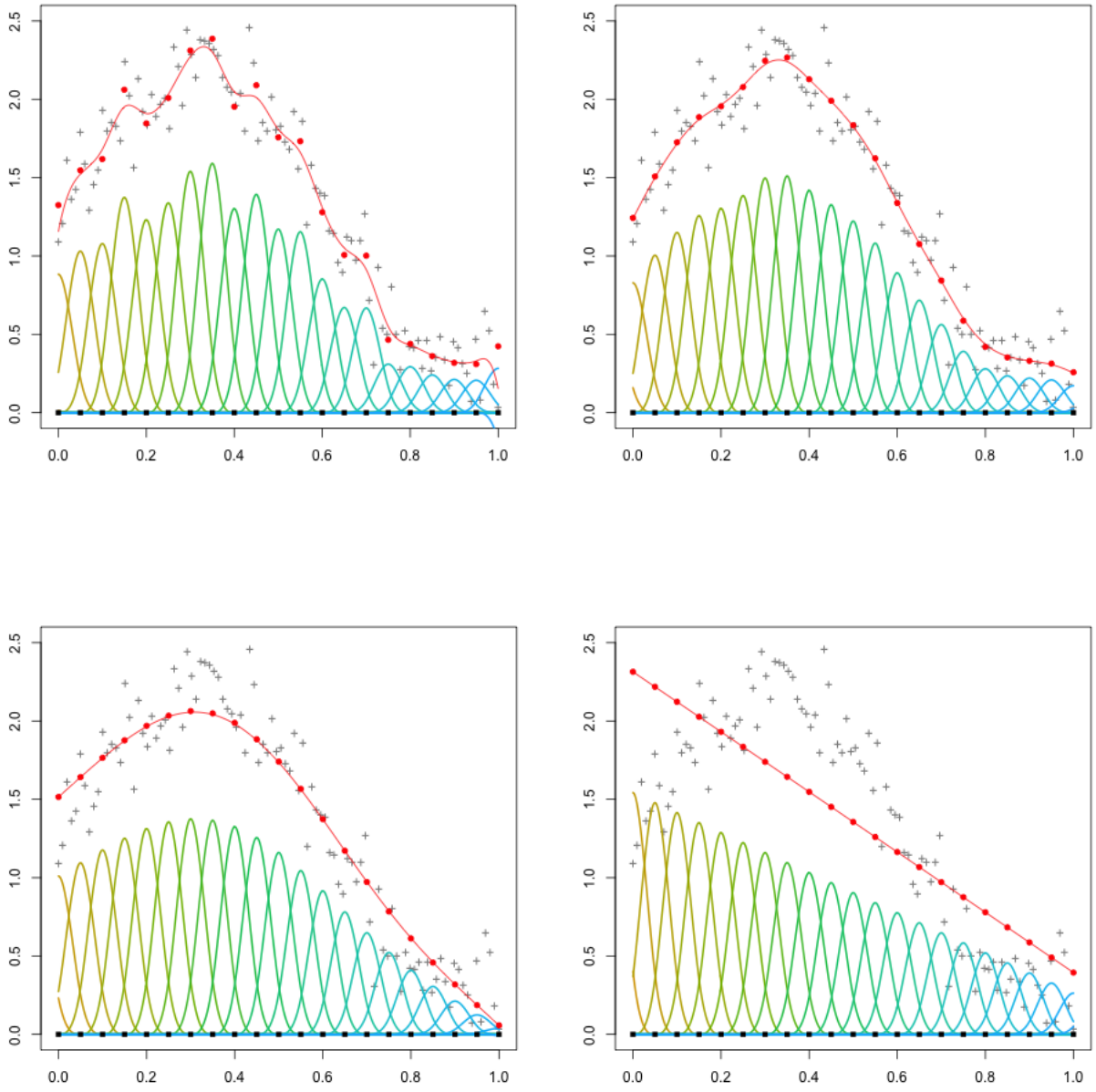


Figure 5: *Illustration of the impact of the second order difference penalty. The number of B-splines used is the same in each plot, with the value of the penalty parameter increasing from left to right and top to bottom across each plot. The fitted curve in the upper left plot is the most "wiggly" of any of the fits, as the penalty plays the weakest roll in the fitted coefficients there. The red circles are the values of each of the B-spline coefficients; as the penalty increases, they form as smoother sequence as we move across the four plots, which results in a smoother fitted function. As the penalty parameter approaches infinity, the fit approaches a linear function as shown in the bottom right plot.*

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.

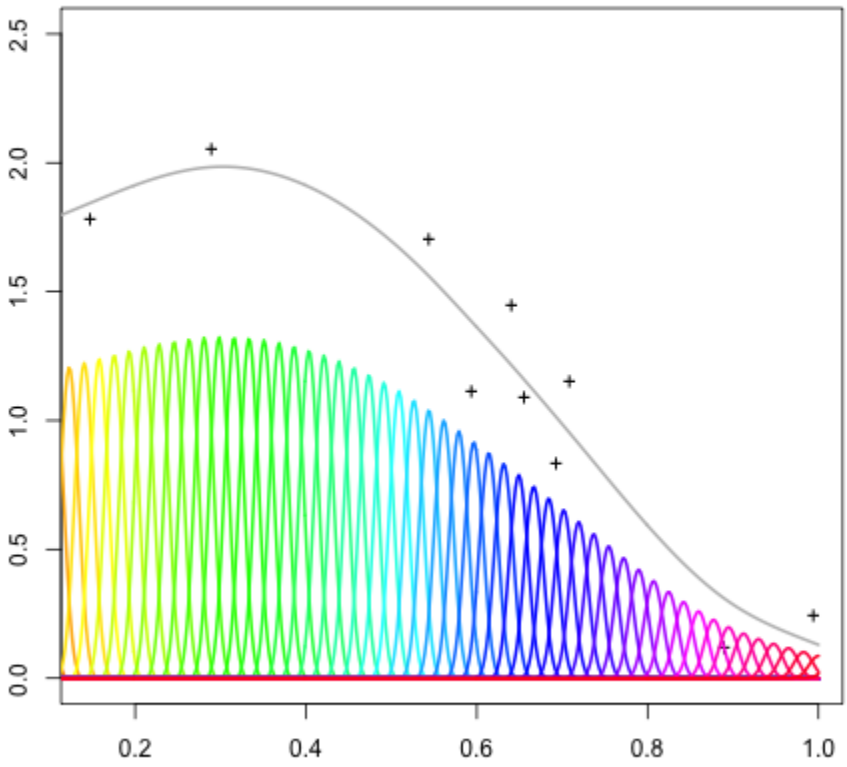


Figure 6: P-spline smoothing of 10 observations using 60 B-spline basis functions.

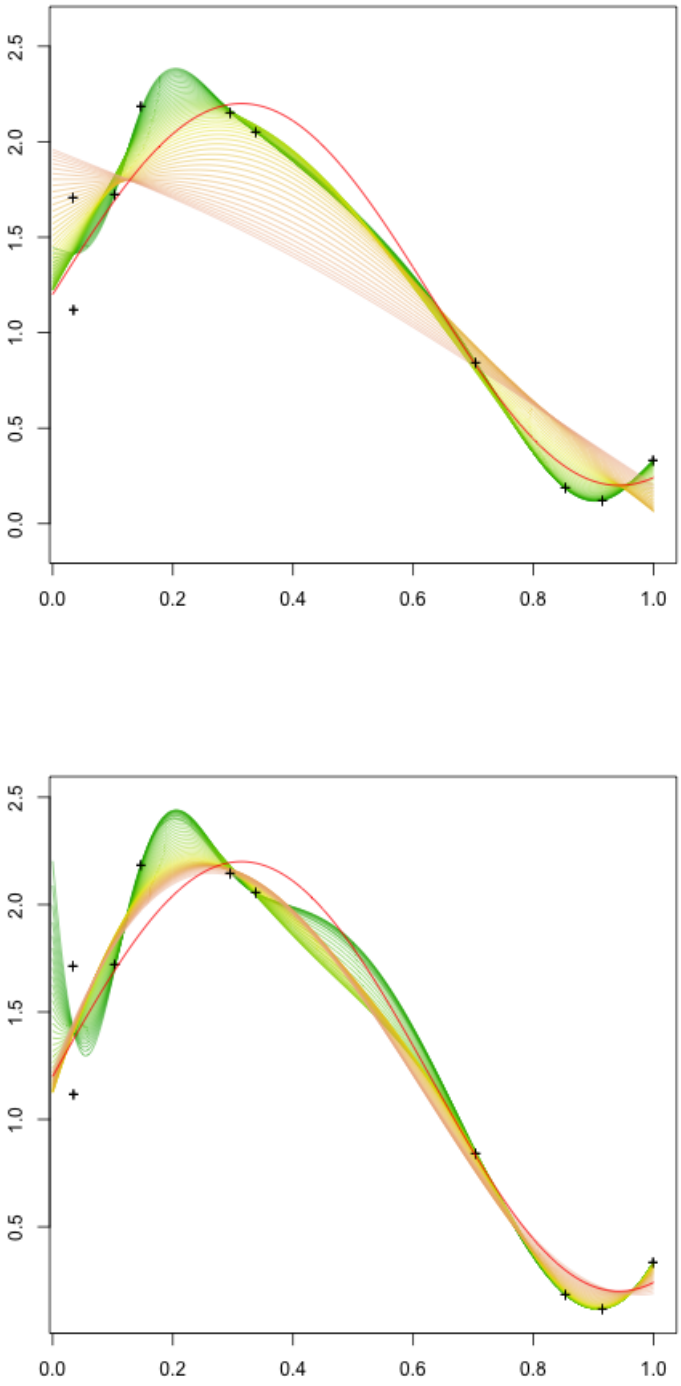


Figure 7: Fitted mean curves using a second (top) and third (bottom) order difference penalty for simulated data, sparsely sampled along the indexing variable: $y(t) = 1.2 + \sin(5t) + 0.2\epsilon_t$, where $\epsilon_t \overset{i.i.d.}{\sim} \mathrm{N}(0, 1)$. A total of 10 data points were fit using a basis of 60 B-splines of degree $k = 3$.

6.3 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 4.1:

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

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 4.1. 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, [**?**] 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 4.2 V,

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

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 4.2 V. Consider using the second-order difference penalty; when λ is large, the penalty dominates the P-spline objective function defined in 44, 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,

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

(or for H defined for the addition of a varying slope component as in 48) approaches k, the order of the differencing operator, as λ increases. We index H with the smoothing parameter to indicate that the elements of H are a function of λ. Let

$$Q_B = B^T B \qquad \text{and} \qquad Q_\lambda = \lambda D^T D. \tag{49}$$

Then using properties of the matrix trace, we can write

$$\begin{aligned} \operatorname{tr}(H_\lambda) &= \operatorname{tr}\left[\left(Q_B + Q_\lambda\right)^{-1} Q_B\right] \\ &= \operatorname{tr}\left[Q_B^{1/2}\left(Q_B + Q_\lambda\right)^{-1} Q_B^{1/2}\right] \\ &= \operatorname{tr}\left[\left(I + Q_B^{-1/2} Q_\lambda Q_B^{-1/2}\right)^{-1}\right] \end{aligned} \tag{50}$$

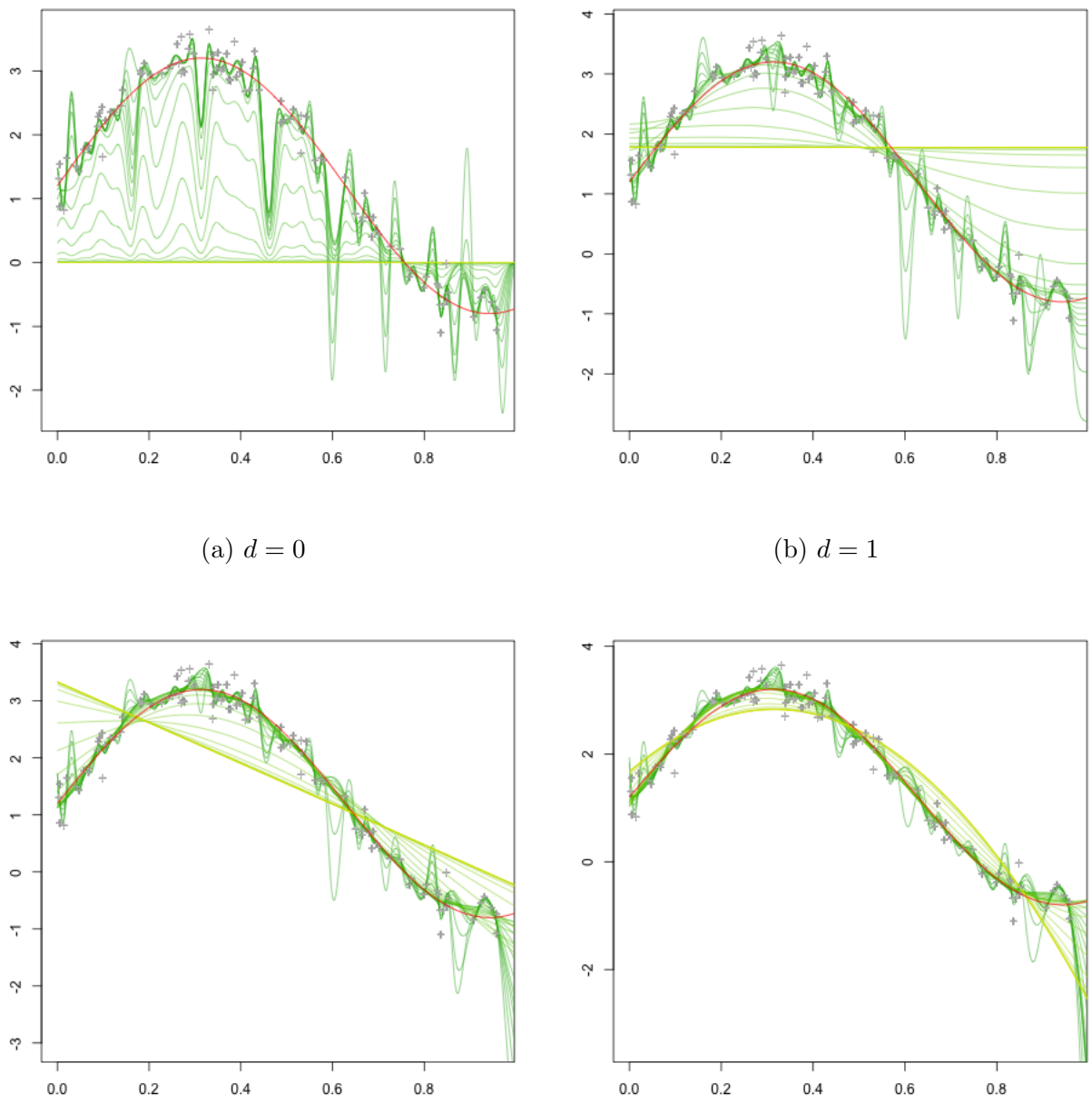
Define $L \equiv Q_B^{-1/2} Q_\lambda Q_B^{-1/2}$. Then

$$\operatorname{tr}(H_\lambda) = \operatorname{tr}\left[\left(I + \lambda L\right)^{-1}\right] = \sum_{j=1}^{n} \frac{1}{1+\lambda\gamma_j} \tag{51}$$

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

$$\lim_{\lambda\to\infty} \operatorname{tr}(H) = k.$$

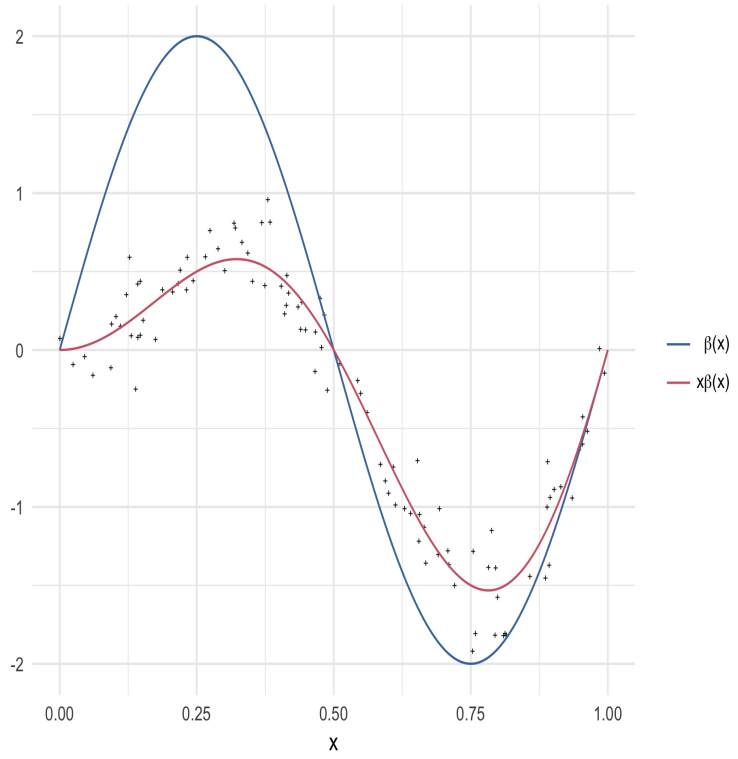
The previous derivations hold regardless of whether we are fitting the varying intercept-only model, with $\mu(t) = \beta_0(t)$ or accommodating a varying slope for a regressor by specifying $\mu(t) = \beta_0(t) + \beta_1(t)\, x(t)$. The inspection of the hat matrix H is a prelude to the following section, 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 8: *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 6.3 III.*



6.4 Optimal tuning and model selection

A major problem involved with any smoothing technique is the choice of the optimal amount of smoothing. Unlike many smoothing approaches such as kernel smoothers, wavelets, and local likelihood methods, the estimation mechanics of P-splines are identical to those used in standard regression models. P-splines are defined by clearly defined coefficients, so we can exploit the surplus of regression theory to compute informative properties of the model. A natural approach to evaluating a model is through inference about the estimated model parameters. This requires the covariance matrix of the estimated B-spline coefficients, which (for fixed λ) is given by

[I'm pretty sure that the last equality follows from taking the QR decomposition of B and writing the solution $\hat{\alpha}$ in terms of Q_1 and R. (I'll figure this out later.)]

$$
\begin{aligned}
\Sigma_\alpha &\equiv Cov\left(\hat{\alpha}\right) \\
&= Cov\left(\left(B^TB + \lambda D^TD\right)^{-1} B^Ty\right) \\
&= \left[\left(B^TB + \lambda D^TD\right)^{-1} B^T\right] Cov\left(y\right) \left[\left(B^TB + \lambda D^TD\right)^{-1} B^T\right]^T \\
&= \sigma^2 \left(B^TB + \lambda D^TD\right)^{-1} B^TB \left(B^TB + \lambda D^TD\right)^{-1} \\
&= \sigma^2 \left(B^TB + \lambda D^TD\right)^{-1},
\end{aligned}
$$

where σ^2 is the error variance under the model $y = B\alpha + \epsilon$. This is of little use, however, as it requires knowing the value of σ. One could estimate it from the residuals if we knew the appropriate value of the tuning parameter, so as to divide by the correct "degrees of freedom." We have illustrated how we can influence the smoothness of the estimated curve via λ, so we must now establish a way to choose the "optimal" value for it. We consider three avenues for model selection criteria. Efficient computation of cross-validation and either the Akaike information criterion or Bayesian information criterion relies on the hat matrix H for fast and simple calculation; in particular, we will discuss the how to calculate both measures using the trace of the smoothing matrix. Alternatively, one can view P-splines as a mixed model; in this context, the penalty parameter can be viewed as a ratio of variances. [**?**] presents an generalization of the Expectation-Maximization (EM) algorithm to optimize for λ.

6.4.1 Cross validation and information criteria

The predictive power of a P-spline model for a given value of the smoothing parameter λ can be assessed by its cross validation prediction error. Cross validation and its variants have long been utilized for smoothing parameter selection in spline models, and their properties have been studied extensively. A short list of supplemental references include [**?**], [**?**], [**?**], [**?**], and [**?**]. There are a number of ways to calculate a measure of cross validated prediction error; we first focus on the leave-one-out method. Let $\hat{y}_{-i}$ denote the predicted value for the observation y_i when y_i itself is removed from the data we use to fit the model. We can calculate these predictions for each observation in the data set to obtain the prediction error:

$$
\mathrm{CV}\left(\lambda\right) = \frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_{-i}\right)^2, \tag{52}
$$

Brute force calculation of 52 is generally impractical, especially if the number of observations is large. However, this labor can be sidestepped using the following fact:

$$
\hat{y} = Hy = B\left(B^TB + \lambda D_d^TD_d\right)^{-1} B^Ty
$$

One can show that

$$y_i - \hat{y}_{-i} = (y_i - \hat{y}_i) / (1 - h_{ii}), \tag{53}$$

and the diagonal elements of H can be calculated quickly. An informal proof of is as follows: suppose that we change the i^{th} element of y, obtaining a new response vector y^*. Then $\hat{y}^* = Hy^*$.

Since

$$\hat{y}_i = \sum_j h_{ij} y_j \qquad \text{and} \qquad \hat{y}_{-i} = \sum_j h_{ij} y_j^*,$$

then

$$\hat{y}_i - \hat{y}_{-i} = \sum_j h_{ij} (y_i - y_i^*) = h_{ii} (y_i - y_i^*).$$

For generalized cross-validation [**?**], we compute

$$\text{GCV}(\lambda) = \sum_{i=1}^{m} \frac{(y_i - \hat{y}_i)^2}{\left(m - \sum_{j=1}^{m} h_{jj}\right)^2} \tag{54}$$

The difference between both quantities is generally small; the best λ is the value that minimizes $\text{CV}(\lambda)$ or $\text{GCV}(\lambda)$. The smoothing matrix H plays an integral role of calculating all of the model selection criteria we will discuss; within the context of both AIC and cross validation, the diagonal elements of H, h_{ii} are of particular importance. Closely related to the problem of model selection is the ability to quantify model complexity. In classical regression theory, the degrees of freedom are clearly defined as the number of variables included in the model. [**?**] and[**?**] refer to this measure of model complexity as the model's *effective dimension* ED; they follow [**?**], who discuss the effective dimensions of linear smoother and propose to use the trace of the smoother matrix as an approximation.

This approach to approximating the effective model dimension is also consistent with [**?**], who constructed a generalization of the concept of a model's degrees of freedom using the idea that the degrees of freedom can also be interpreted as the sum of the sensitivity of each fitted value with respect to the corresponding observed value. For smoothing matrix H, the predicted response values are given by $\hat{y} = Hy$. Writing

$$\frac{\partial \hat{y}_i}{\partial y_i} = \frac{\partial}{\partial y_i} \sum_j h_{ij} y_j = h_{ii},$$

we see that the latter interpretation of the effective model dimension reduces to calculating the trace of the hat matrix. Thus we take the effective dimension to be

$$\begin{aligned} \text{ED}(\lambda) &= \text{tr}(H) \\ &= \text{tr}\left[B \left(B^T B + \lambda D^T D\right)^{-1} B^T\right], \end{aligned} \tag{55}$$

Summarizing the complexity of a fitted P-spline is far from a trivial task; one must simultaneously consider the value of the smoothing parameter, the number of basis functions in the B-spline basis, as well as the order of the difference penalty. The effective dimension is easily obtained and combines the effect of all three of these elements. When the number of basis functions is significantly smaller than the sample size, it is computationally advantageous to use the cyclic property of the matrix trace:

$$\text{tr}\left[\left(B^TB + \lambda D^TD\right)^{-1} B^TB\right] = \text{tr}\left[\left(B^TB + \lambda D_d^TD_d\right)^{-1} B^TB\right]. \tag{56}$$

The matrix in Equation 56 has dimension $n \times n$, so when the number of basis functions is smaller than the number of observed grid points, it is computationally advantageous to use the latter formulation. A further simplification will prove useful for deriving properties of the effective dimension of the P-spline smoother:

$$\begin{aligned}\left(B^TB + \lambda D^TD\right)^{-1} B^TB &= \left(B^TB + \lambda D^TD\right)^{-1} \left(B^TB + \lambda D^TD - \lambda D^TD\right) \\ &= I - \lambda \left(B^TB + \lambda D^TD\right)^{-1} D^TD \end{aligned} \tag{57}$$

Equation 57 cleanly shows that the effective dimension is always less than n, the number of B-spline used in the regression basis; further, the effective dimension is always smaller than $\min(m, n)$. A formal proof follows below. This is illustrated in Figure **??**, which shows how the effective dimension changes with the smoothing parameter for the ten simulated observations in Figure **??** using 60 B-spline basis functions. For small λ, the effective dimension approaches m. As λ increases, the effective dimension approaches the order of the difference penalty, d. It is worth pointing out here that there are no problems incurred when smoothing with many more B-splines than observations since the effective model dimension is always less than m, for all λ.

[INSERT PROOF HERE.]

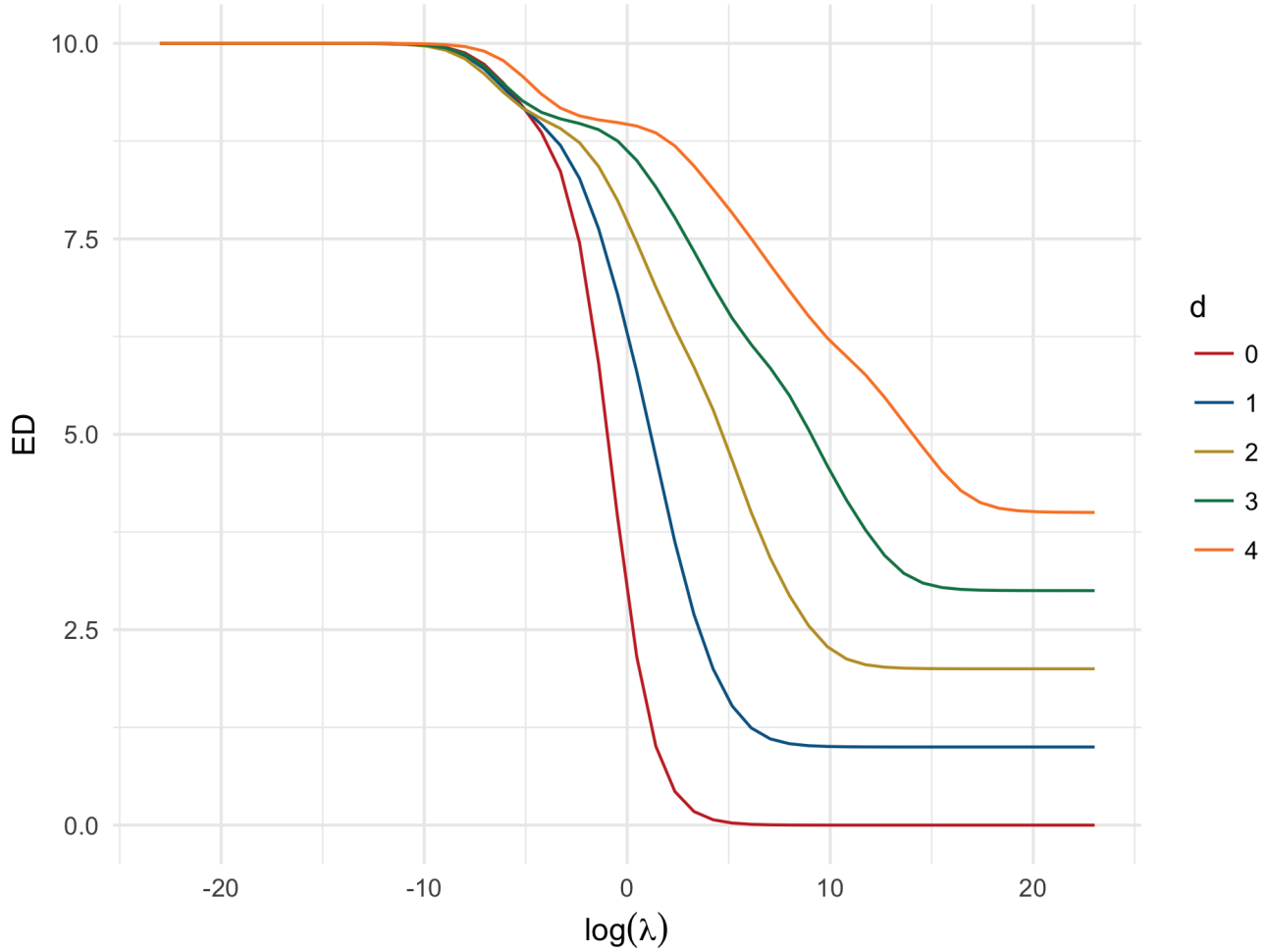


Figure 9: *An illustration of the behaviour of effective model dimension for P-spline smoothing of 10 simulated observations* (x_i, y_i) *using* $n = 70$ *cubic B-splines under various orders of the difference penalty,* d.

An alternative approach to model selection is to minimize the information criterion (IC):

$$\mathrm{IC}\left(\lambda\right) = -2\left(y, \hat{\beta}_\lambda\right) + c\,\dim\left(\hat{\beta}_\lambda\right)$$

Special cases are when $c = 2$, yielding the Akaike information criterion (AIC) and when $c = \log\left(m\right)$, yielding the Bayesian information criterion (BIC). The IC assesses the quality of a model by adjusting the log likelihood of the data under the fitted model to account for some measure of model complexity. In the classical linear model setting, the degrees of freedom serves as the measure of model complexity, is used for obtaining an unbiased estimate of the error variance, and thus, is necessary for comparing the performance of different models. In the case of the normal likelihood, AIC becomes

$$\mathrm{AIC}\left(\lambda\right) = \tag{58}$$

One may use CV to select the optimal value of λ, and then using the corresponding residuals gives a natural choice to use as an estimate of σ^2 for the computation of $\mathrm{AIC}(\lambda)$. It is practical to work with modified versions of $\mathrm{CV}(\lambda)$ and $\mathrm{GCV}(\lambda)$, with values that can be interpreted as estimates of the cross-validation standard deviation:

$$\begin{aligned} \bar{\mathrm{CV}}(\lambda) &= \sqrt{\mathrm{CV}(\lambda)} \\ \bar{\mathrm{GCV}}(\lambda) &= \sqrt{m\mathrm{GCV}(\lambda)} \end{aligned} \tag{59}$$

Alternatively, one might note that an (approximately) unbiased estimate of the error variance is given by the sum of squared residuals divided by the error degrees of freedom, $m - \mathrm{ED}$.

6.4.2 P-splines as mixed models and the E-M algorithm

7 Multidimensional P-spline smoothing

Loop the readers back to the original problem: estimating the $\phi(s,t) = \phi^*(l,m)$, which is the generalized autoregressive surface of interest when we model

$$y(t) = \sum_{s<t} \phi(t,s)\, y(s) + \epsilon(t).$$

We need a way of extending the modeling framework discussed in the previous sections to multidimensional coefficient functions.

Review the previously proposed methods of smoothing down the diagonals of the covariance matrix, including

I. treating each diagonal as p separate regression

   A. which does not permit one to model the smooth in the $s+t$ direction at all

   B. see [?]

   [

II. Will I have already discussed Mohsen Pourahmadi's kernel smoothing technique at this point? If not, point out that kernel smoothing addresses the issue of the equally spaced observations as a necessity, but further elaborate that his method, failing to utilize the modified Cholesky decomposition, does not enjoy the unconstrained parameterization that we do; his methods do not ensure that covariance matrix estimates are positive definite.]

III. using GAMs to allow for each direction to have its own specified functional component

A. which only permits one to model the surface as two smooth main effects: one in the $s-t$ directions and one in the $s+t$ direction. in the $s+t$ direction at all

Review the extensions of univariate P-splines and their corresponding VC models, including generalized additive models and penalized signal regression with $p > 1$ regressors. Demonstrate how these require only simple modifications to the original design framework for estimation.

7.1 Penalized tensor product B-splines for two-dimensional surface approximation

As in the previous chapter with smoothing splines, tensor products permit a nearly seamless extension of P-splines for VC models in which the slope function $\beta_1(\cdot)$ varies along a single indexing dimension to VC models with slope functions which are of two (or more) dimensions. If we equip each dimension, x and v, with its own set of B-spline basis functions, then the only other modification necessary for the extension into two dimensions is the addition of a difference penalty for each augmented indexing variable.

Assume the available data are observed in triplets, (x_i, v_i, y_i), $i = 1, \ldots, m$. We wish to estimate a smooth bivariate function $\mu(x, v)$ which approximates the data well. We assume

$$E[y_i] = \mu(x_i, v_i). \tag{60}$$

Figure 10 displays the building block of the foundation on which multidimensional P-splines is built: B-spline tensor product basis $\{T_{kl}\}$.

The $j-k^{th}$ single tensor product is given by

$$T_{jk}(x, y) = B_j(x)\,\bar{B}_k(y),$$

where $\{B_k\}$ and $\{\bar{B}_l\}$, $k = 1, \ldots, K$, $l = 1, \ldots, L$ denote the B-spline bases for x and v, each having a set of equally spaced knots along their respective domain, carving out the $x - v$ plane into sub-rectangles. (We have chosen to distinguish between $\{B_k\}$ and $\{\bar{B}_l\}$ because one is free to specify a different basis for each dimension either by using different order B-spline or, of course, using different numbers of knots, and hence entirely different knot sequences since P-splines rely on bases with equally spaced knots.)

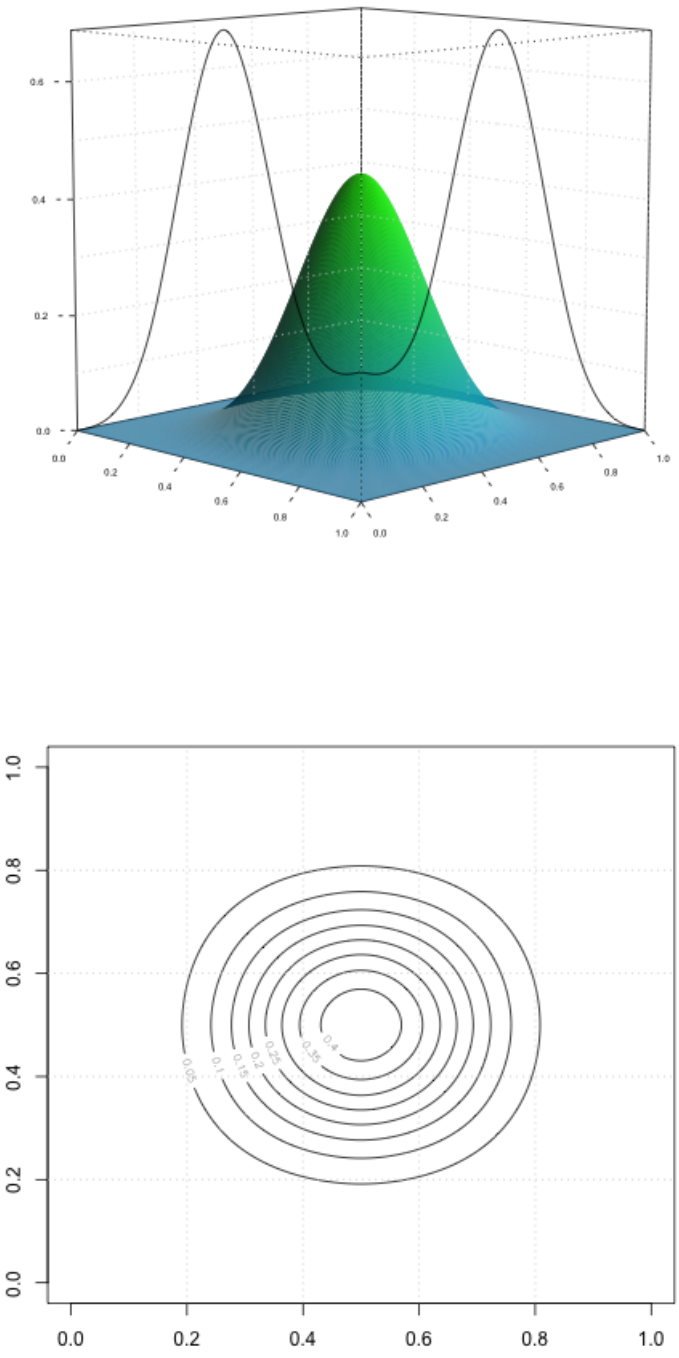


Figure 10: Tensor product of two cubic B-splines

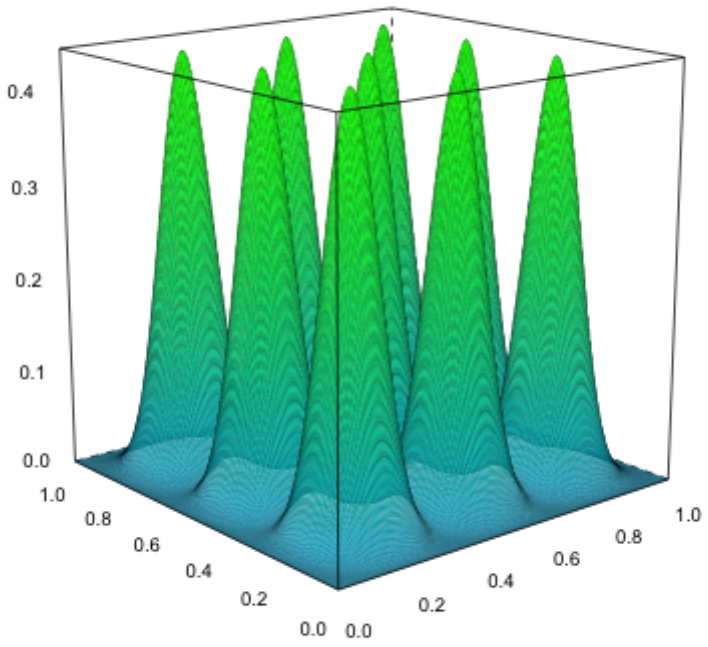


Figure 11: A subset of a full bivariate basis of cubic B-splines

As in one dimension, we may express the mean surface as a tensor product basis expansion:

$$\mu(x, v) = \sum_k \sum_l \alpha_{kl} B_k(x) \bar{B}_l(v), \tag{61}$$

where α_{kl} is the $k - l^{th}$ element of the $K \times L$ coefficient matrix, A. Figure 11 shows a subset of a full tensor product basis $\{T_{kl}\}$. A small portion of the basis is displayed because without omitting many of the basis functions, there would be too much overlap to be able to distinguish one from the rest. [DISCUSS HOW THE PROPERTIES OF UNIVARIATE B-SPLINES TRANSLATE TO THE $x \times v$ PLANE (i.e. compact support, evaluation of the function at any point in the domain depends on only a small number of parameters, non-negative and sum to 1, etc.)]

The relationship between the elements of A and the shape of the fitted surface is entirely analogous to the situation illustrated in Figure 5; the coefficients α_{kl} determine the height of each "mound" corresponding to a single T_{kl}., thus the surface is driven by a relatively small number of total parameters: KL. To control the smoothness of the fitted surface, we

can impose difference penalties on the coefficients, just as in the one-dimensional case, but smoothing in both dimensions requires two penalties: one for row differences of A and one for the column differences of A.

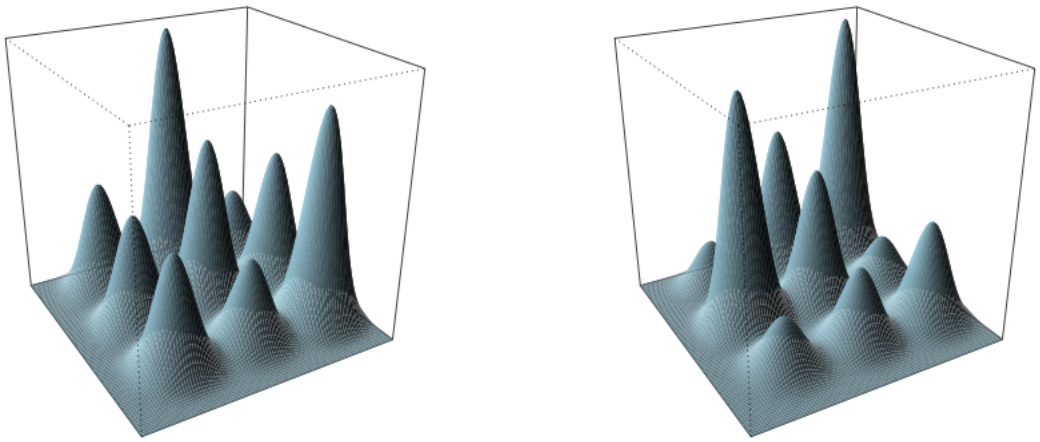
Intuitively, the tensor product B-splines are easily digestible, but to square away the technical details, they do require a bit of careful bookkeeping. As in one dimension, we may consider the sum of squared errors as a goodness of fit function, and in the case where we choose to enforce no smoothness penalty on the fitted surface, the estimate for A can be chosen to minimize

$$S = \sum_{i=1}^{m} \left[y_i - \mu \left(x_i, v_i \right) \right]^2 = \sum_{i=1}^{m} \left[y_i - \sum_k \sum_l \alpha_{kl} B_k \left(x_i \right) \bar{B}_l \left(v_i \right) \right]^2. \tag{62}$$

As before, we can then leverage difference penalties by adding the following penalty term to the goodness of fit measure:

$$\lambda \sum_k |D_d \alpha_{k\cdot}|^2 + \bar{\lambda} \sum_l |D_{\bar{d}} \alpha_{\cdot l}|^2. \tag{63}$$

$\alpha_{k\cdot}$ and $\alpha_{\cdot l}$ denote the k^{th} row and l^{th} column of A, respectively. Thus, the first term in the penalty imposes a difference penalty of order d on the rows of the coefficient matrix while the second term places a difference penalty (of possible different order $\bar{d}$) on the columns. We give each direction its own smoothing parameter to permit anisotropic smoothing; however, one could opt to use a single smoothing parameter for both directions and dodge the added work of optimizing the amount of smoothing with two separate parameters. Figure **??** shows a potential result of heavy column penalization (left) and heavy row penalization (right) under a second order difference penalty on each row and each column for large values of λ and $\bar{\lambda}$. The figure demonstrates that the limiting behaviour of each row and column is linear, but the resulting surface may exhibit slope reversals from one row (column) to the next.



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

Figure 12: *Nine cubic B-spline tensor products with heavy linear column penalty and heavy linear row penalty*

It is possible to derive (penalized) normal equations in terms of the coefficient matrix itself, but this formulation is rather unwieldy. To efficiently minimize the penalized residual sums of squares with respect to the $\{\alpha_{kl}\}$, we can re-express the mean surface in terms of the vectorized coefficient matrix:

$$\mu(x, v) = \mathring{B}\alpha \tag{64}$$

where $\alpha = \text{vec}(A)$ is the flattened coefficient matrix, constructed by stacking the columns of A. $\mathring{B}$ denotes the $m \times KL$ "unfolded" regression basis and is defined in terms of the Kronecker product (denoted $o\times$) and the element-wise multiplication (denoted $\odot$) of two matrices:

$$\mathring{B} = \left[B \otimes 1_L^T\right] \odot \left[1_K^T \otimes \bar{B}\right] \tag{65}$$

where 1_K and 1_L are one vectors of length K and L, respectively. These definitions of $\mathring{B}$ and α, estimation of the coefficient grid can be performed in the usual regression framework with $E[y] = \mathring{B}\alpha$. To write the penalized sums of squares, we must write the penalty 63 on the coefficient matrix in terms of penalties on α. Both the row and column penalties may also be expressed in terms of the Kronecker product of two matrices. The row penalty can be written as $\lambda\alpha^T P\alpha$ where

$$P = I_L \otimes D_d^T D_d.$$

In a similar fashion, the column penalty can be written as $\bar{\lambda}\alpha^T\bar{P}\alpha$, where

$$\bar{P} = D_{\bar{d}}^T D_{\bar{d}} \otimes I_L,$$

and I_K and I_L are identity matrices of dimension $K \times K$ and $L \times L$, respectively. We choose coefficients α to minimize

$$|y - \mathring{B}\alpha|^2 + \lambda\alpha^T P\alpha + \bar{\lambda}\alpha^T\bar{P}\alpha$$

which has solution

$$\hat{\alpha} = \left(\mathring{B}^T\mathring{B} + \lambda D_d^T D_d + \bar{\lambda}D_{\bar{d}}^T D_{\bar{d}}\right)^{-1}\mathring{B}^T y. \tag{66}$$

From 66, notice that the system of equations needed to solve for the basis coefficients remains fixed at KL, even as the number of observations increases. [introduce notation necessary for specifying the observed grid for both the x and v axes.] The grid of regression coefficients can be recovered by arranging the elements of $\hat{\alpha}$ into a matrix of L columns having length K. Just as in one dimension, the effective hat matrix is given by

$$H_\lambda = \mathring{B}\left(\mathring{B}^T\mathring{B} + \lambda D_d^T D_d + \bar{\lambda}D_{\bar{d}}^T D_{\bar{d}}\right)^{-1}\mathring{B}^T,$$

with the calculation of effective dimension and deletion residuals may be calculated just as in Section 6.4 by replacing B with $\mathring{B}$ and replacing $\lambda D_d^T D_d$ with $\lambda D_d^T D_d + \bar{\lambda}D_{\bar{d}}^T D_{\bar{d}}$.

7.2 Diagnostics and model selection for tensor product P-splines

Many of the practical properties of P-splines in one dimension such as accessible cross validation and effective model dimension are inherited in the two dimensional case.

[Discussion of effective model dimension, CV, and AIC should go here, along with figure with perspective plot showing a realistic 2d surface for various values of effective model dimension.]

7.3 Tensor product varying coefficient models

As the extension in one dimension from P-spline smoothing to coefficient smoothing for VC models was nearly seamless, as allowing for the smooth term to become instead a slope function to be multiplied by a fixed regressor, x merely entailed modeling $y = XB\alpha = U\alpha$. One can employ the same modeling framework as in the simple case without covariates with only a modifed basis, $U = XB$. We may employ similar mechanics to extend the set of models discussed in Section 7.1 to models for two-dimensional coefficient surfaces, which we will introduce first in general notation. Later, we will show how the problem of estimating the generalized autoregressive coefficients can be written as a special case.

Given $m \times p$ matrix X of regressors and coefficient surface $\beta(x, v)$, define the mean of the i^{th} observed response as

$$\mu_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta\left(xt_i, v_j\right). \tag{67}$$

One can understand the intuition behind the form of this model if we consider the following VC model with one dimensional slope:

$$\mu_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta\left(v_j\right). \tag{68}$$

[Relate this simplification back to the models for ϕ that involve smoothing down each diagonal of T separately.]

Equation 67 shares the spirit of the one-dimensional model in 68, but uses a slice of the coefficient surface that is specific to the value of x. Approximating the coefficient surface using tensor product B-splines as in the previous section, then we may rewrite the model in Equation 67 as

$$\begin{aligned}
\mu_i - \beta_0 &= \sum_{j=1}^{p} x_{ij}\beta\left(x_i, v_j\right) \\
&= \sum_{j=1}^{p} x_{ij} \sum_{k=1}^{K}\sum_{l=1}^{L} B_k\left(x_i\right)\bar{B}_l\left(v_j\right)\alpha_{kl} \\
&= \sum_{j=1}^{p} x_{ij} \sum_{k=1}^{K}\sum_{l=1}^{L} x_{ij} b_{ik}\bar{b}_{jl}\alpha_{kl} \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L}\left(\sum_{j=1}^{p} x_{ij} b_{ik}\right)\bar{b}_{jl}\alpha_{kl} \\
&= \sum_{k=1}^{K}\sum_{l=1}^{L} u_{ik}\bar{b}_{jl}\alpha_{kl}.
\end{aligned} \tag{69}$$

Adopting the instrumentality of 65, we can express Equation 69 in matrix notation:

$$\mu = \alpha_0 + \mathring{U}\alpha,$$

where $\bar{U}$ is the constructed via the row tensor product of U with $\bar{B}$:

$$\mathring{U} = \left(U \otimes 1_L^T\right) \odot \left(1_K^T \otimes \bar{B}\right). \tag{70}$$

is of dimension $m \times KL$, and where $U = XB$. Penalty matrices $P = I_L \otimes D_d^T D_d$ and $\bar{P} = D_{\bar{d}}^T D_{\bar{d}} \otimes I_L$ may be constructed as in Section 7.1. The solution $\hat{\alpha}^*$ minimizes

$$
\begin{aligned}
S\left(\alpha^*, \lambda, \bar{\lambda}, \lambda_0\right) &= |y - \alpha_0 - \mathring{U}\alpha|^2 + \lambda|P\alpha|^2 + \bar{\lambda}|\bar{P}\alpha|^2 + \lambda_0|\alpha|^2 \\
&= |y - \mathring{U}^*\alpha^*|^2 + \lambda|P^*\alpha^*|^2 + \bar{\lambda}|\bar{P}^*\alpha^*|^2 + \lambda_0|I^*\alpha^*|^2 \qquad (71)
\end{aligned}
$$

where

$$
\begin{aligned}
\alpha^* = \left[\alpha_0, \alpha^T\right]^T, \quad \mathring{U}^* = \left[1_m | \mathring{U}\right], \\
P^* = \left[0_m | P\right], \quad \bar{P}^* = \left[0_m | \bar{P}\right], \\
\text{and} \quad I^* = \operatorname{diag}\left(0, I_{KL}\right).
\end{aligned}
$$

The additional penalty parameter λ_0 imposes an overall ridge penalty on α. We augment U, P, and $\bar{P}$ with zeros so as to not penalize the intercept term α_0. The minimizer $\hat{\alpha}^*$ of Equation 71 is given by

$$
\hat{\alpha}^* = \left[\mathring{U}^{*T} + \lambda P^* + \bar{\lambda}\bar{P} + \lambda_0 I^*\right]^{-1} \mathring{U}^{*T} y.
$$

The predicted response is then given by

$$
\begin{aligned}
\hat{y} &= \mathring{U}^* \left[\hat{\alpha_0}, \hat{\alpha}^T\right]^T \\
&= \mathring{U}^* \left[\mathring{U}^{*T} + \lambda P^* + \bar{\lambda}\bar{P} + \lambda_0 I^*\right]^{-1} \mathring{U}^{*T} y
\end{aligned}
$$