P-spline Varying Coefficient Models for Complex Data

Brian D. Marx

Abstract Although the literature on varying coefficient models (VCMs) is vast, we believe that there remains room to make these models more widely accessible and provide a unified and practical implementation for a variety of complex data settings. The adaptive nature and strength of P-spline VCMs allow a full range of models: from simple to additive structures, from standard to generalized linear models, from one-dimensional coefficient curves to two-dimensional (or higher) coefficient surfaces, among others, including bilinear models and signal regression. As P-spline VCMs are grounded in classical or generalized (penalized) regression, fitting is swift and desirable diagnostics are available. We will see that in higher dimensions, tractability is only ensured if efficient array regression approaches are implemented. We also motivate our approaches through several examples, most notably the German deep drill data, to highlight the breadth and utility of our approach.

1 Introduction

The varying coefficient model (VCM) was first introduced by Hastie & Tibshirani (1993). The main idea of the VCM is to allow regression coefficients to vary smoothly (interact) with another variable, thus generating *coefficient curves*. Such coefficient curves can, for example, reflect slow changes in time, depth, or any other indexing regressor. Hence regression coefficients are no longer necessarily constant. Typically estimation for the varying coefficients usually requires the backfitting algorithm, i.e. cycling through and updating each smooth term successively, until convergence. But backfitting also has drawbacks: no information matrix is being computed, so the computation of standard errors and effective model dimension,

Brian D. Marx
Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803 USA
e-mail: bmarx@lsu.edu

or efficient leave-one-out (LOOCV) cross-validation is not available. Also convergence can be slow.

We have published an efficient fitting algorithm for VCM, based on P-splines (Eilers & Marx, 2002), abbreviated as GLASS (Generalized Linear Additive Smooth Structures). GLASS directly fits all smooths simultaneously of the VCM, without backfitting. In the linear case it converges in one step, and in the generalized linear case it needs only a handful of iterations, similar to the iterative weighted regression for generalized linear models. Standard errors, LOOCV and effective dimension, and diagnostics are readily available at little extra cost. Further optimization is relatively easy and is based on data-driven techniques.

Our GLASS algorithm only considers coefficients that are smooth curves along one dimension (although it allows several of those components). However VCMs can be applied to problems with even richer structure, e.g. coefficients that vary in two or more dimensions and with other additive components in the model. Such data can be generated from modern image or spectral instrument, but can arise naturally from simple tabulations. In principle, using tensor-product P-splines, VCMs can be extended to higher-dimensions, allowing the estimation of (higher dimensional) coefficient surfaces. In theory this is allowed, but in practice one often encounters severe limitations in memory use and computation time. The reason is that large-

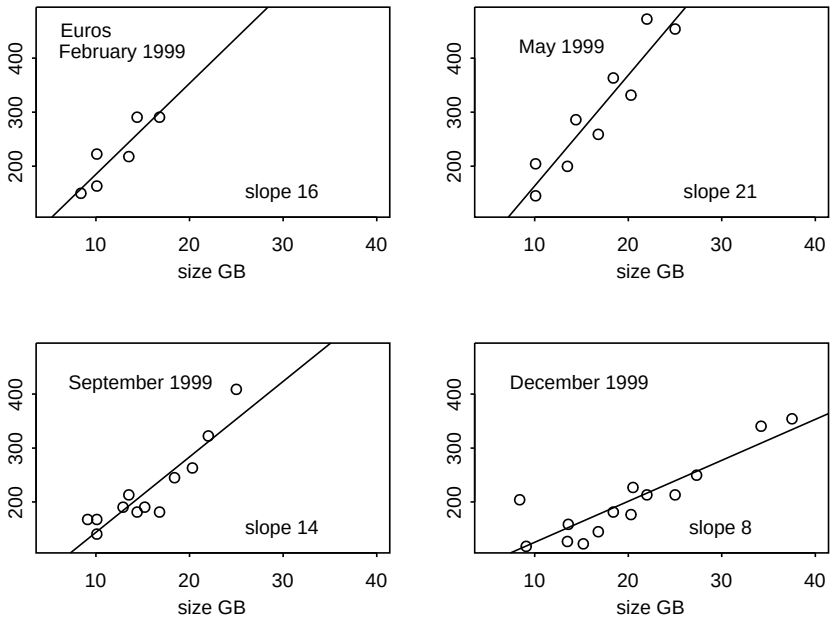

Fig. 1 *IBM hard drives: price (Euro) vs. size (GB), at four different months.*

scale multi-dimensional VCMs need a large basis of tensor products of B-splines. In combination with many observations this can lead to inefficiencies. Consider, as an example, an image of 500×500 pixels, to which one likes to fit a VCM, using a 10 by 10 grid of tensor products of B-splines. The regression basis has 250 thousand rows and 100 columns, or 25 million elements, each taking 200 Mb of memory. With several VCM components storing just the basis can already take on Gigabyte of memory. Computation times to compute inner products will be long. Note that the final system of penalized normal equations is not large with a few hundreds of coefficients. Recently very efficient algorithms have been published for smoothing of multidimensional data arrays with P-splines (Currie, Durbán, & Eilers 2006). They offer improvements of several orders of magnitude in memory use and computation time. With small adaptations, these algorithms can be used for multi-dimensional VCM fitting.

We do not attempt to survey all of the VCM developments. Rather, the major goal of this paper is to provide a unified, accessible, and practical implementation of VCMs using P-splines; one that is conducive to generalizations and tractable in

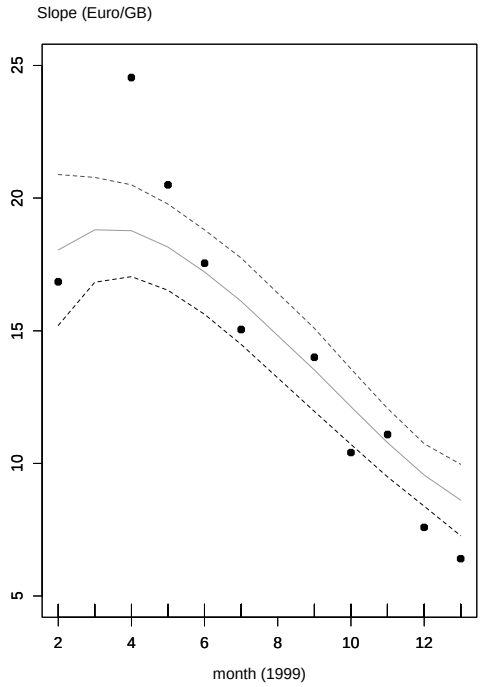

Fig. 2 *IBM: Estimated varying slope, combining monthly data. The individual data points represent the estimated slopes using the data month by month. Note that March and August do not have estimate slopes since they have missing data or one observation.*

a variety of relatively complex settings, such as two and three-dimensional space-varying GLM extensions, all while avoiding backfitting.

Warm-up: An Intuitive Example

We first illustrate the basic structure and mechanics of a VCM through a simple example. Consider the disk data with the triplets $(y_i, x_i, t_i),\ i = 1, \ldots, m$, where the response y_i is the *price* (Euro) of an IBM hard drive, the regressor x_i is its *size* (GB), and t_i is the indexing variable *month* (ranging from February 1999 through January 2000). Figure 1 displays the (x,y) scatterplot separately for four selected months yielding some evidence of a varying (estimated) slope. The VCM combines the monthly data into one model, allowing the slope coefficient to vary smoothly in t. Consider modeling the mean response

$$\mu = x(t) f(t),$$

where $f(t)$ is a smooth slope function. Figure 2 displays the estimated $\hat{f}(t)$ (with twice standard error bands) which strongly suggests that the estimated Euro/GB is decreasing with time. The data points in Figure 2 represent the estimated slopes using the individual monthly data. Note that we are not simply smoothing the points on this graph, but rather borrowing strength from all the monthly data to produce a smooth *coefficient curve*. Such a VCM approach allows for interpolation of Euro/GB for months with missing data (e.g. March and August) or for months with only one observation where slope cannot be estimated. Further we can extrapolate Euro/GB into future months. The details for estimation follow in the coming sections.

2 "Large Scale" VCM, without Backfitting

The German Continental Deep Drill Program (KTB) was an ambitious project with its aim to study the properties and processes of the upper 10 km of the continental crust (`www.icdp-online.de/sites/ktb/`). The actual drill cuttings comprise of 68 variables measured at each of 5922 depth points (having a 1 m median spacing) down to a final depth of 9.1 km.

We primarily motivate varying coefficient models through the characterization of cataclastic fault zones, and relating the amount of cataclastic rocks (*CATR*), along varying depth, to other variables. Our response is mean amount of *CATR* (which in previous research has been transformed in either units of natural logarithm (log) or log-odds (logit) transformed volume percent), and our central explanatory variables include: Structural water (H_2O), graphite (C), Al_2O_3, Na_2O (all in units weight percent), and *Thermal Conductivity* (in units $Wm^{-1}K^{-1}$).

The KTB statistical models to date only used a subset of depth range. However we find the P-spline VCM is adaptive enough to incorporate the entire range of 9.1km depth, thereby modelling all data zones simultaneously. The choice of these regressors comes, in part, from existing successful statistical analyses of the KTB data, by e.g. Kauermann & Küchenhoff (2003). These authors modelled the mean and dispersion structure of the amount of cataclastic rocks by focusing on a subset of drill samples ranging from 1000 to 5000 meters, which led to the identification of possible depth breakpoints and potential outliers. Further, Winter et al. (2002) investigated the relationship between the amount of cataclastic rocks to several geological variables using standard regression methods for two specific cataclastic zones within two lithologies: gneiss (1738-2380m) and metabasite (4524-4908m).

It is unrealistic to assume *constant* regression coefficients, along $0-9101$m (e.g. associated with H_2O, C, Al_2O_3, Na_2O, and *Thermal Conductivity*), and a VCM approach can be a reasonable model choice, thus allowing variables to have depth dependent flexible influence on the response.

Section 5 will provide the details, but to give an idea of how slope coefficients can vary positively and negatively along depth, consider Figure 3 that uses a P-spline VCM. The panels also present twice-standard error bands associated with

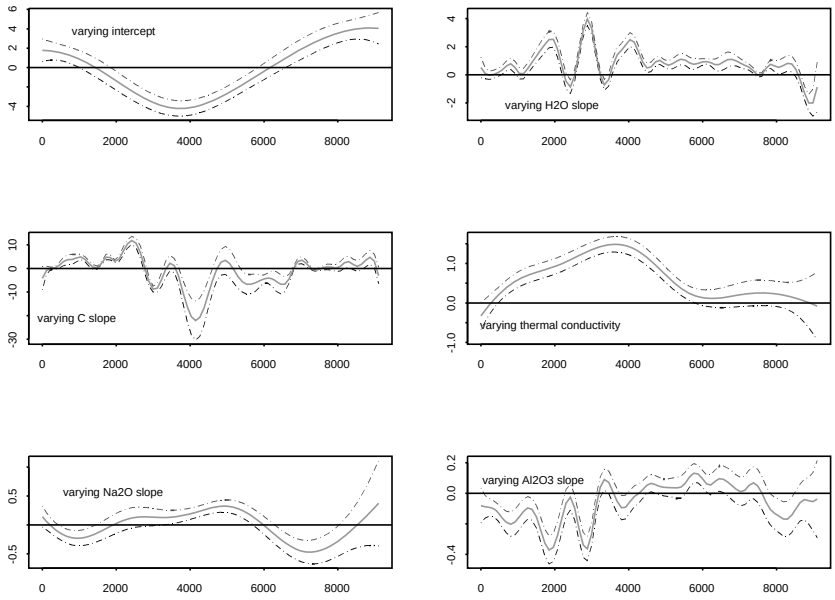

Fig. 3 Using log(*CATR)* as response, varying intercept and varying slopes for H_2O, *C*, *Thermal Conductivity*, Na_2O, Al_2O_3 using cubic ($q=3$) P-splines with 40 equally-spaced knots, $d=3$. Optimal tuning parameters chosen by EM. Twice standard bands are provided.

the varying coefficients. Relative to the zero line we see evidence of reversals or moderating impacts of regressors on *CATR* as depth varies, e.g. *C* appears to have positive, negative, and a near zero impact on *CATR*, e.g., at depths of 2300m, 4000m and greater than 7000m, respectively.

The goodness-of-fit measures associated with P-spline VCM shows promise for applications to the KTB data. For example, the models of Winter et al. (2002) that target specific zones, only using a depth range of several hundred meters, reported R^2 values between 0.57– 0.60. Our VCM approach initially show a 12% – 21% improvement, while using the entire 9.1 km range over all data zones. A more thorough presentation of results is given in Section 7.

3 Notation and Snapshot of a Smoothing Tool: B-splines

We will see in the sections that follow that we initially approach smoothness of the coefficient vector (*not* the explanatory variables), in two ways: (a) by modelling coefficients with a B-splines at predetermined depths (knots), and (b) when the number

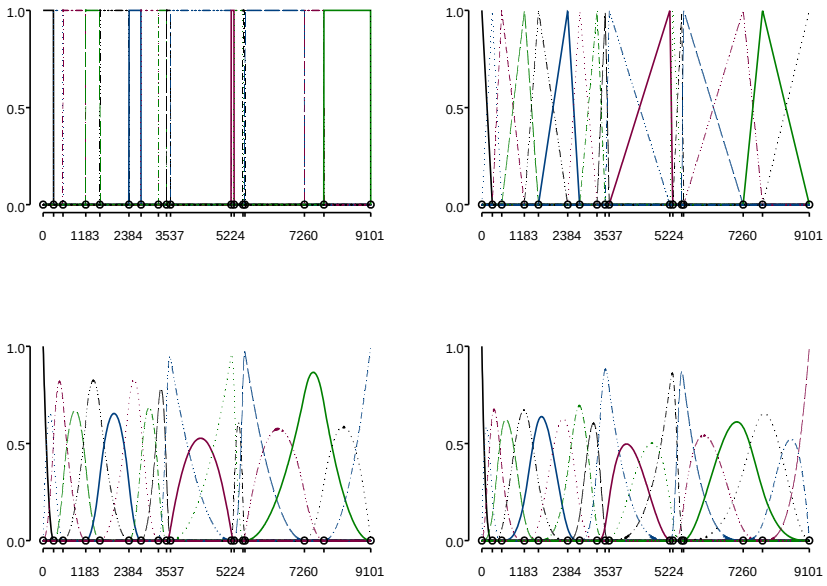

Fig. 4 *B*-spline bases with knots at specific depths: degrees $q = 0, 1, 2, 3$.

and position of knots is assumed not to be known, by using penalized B-splines or P-splines (Eilers & Marx 1996).

3.1 General knot placement

We start with the building block of a complete B-spline basis. The shape of any one B-spline function depends on its degree q. For example, a B-spline takes a constant value (degree $q = 0$), has the form of a triangular density (degree $q = 1$), or can even resemble bell-shaped curves similar to the Gaussian density (e.g. higher degrees $q = 2,\ 3$). A B-spline function has only local support (e.g. in contrast to a Gaussian density). In fact it is constructed from smoothly joining polynomial segments. The positions on the indexing axis, t, where the segments come together, are called the knots. Some general properties of a degree q B-spline include: it consists of $q+1$ polynomial pieces of degree q; the derivatives at the joining points are continuous up to degree $q-1$; the B-spline is positive on the domain spanned by $q+2$ knots, and it is zero elsewhere.

A full B-spline basis is a sequence of B-splines functions along t, each shifted over one knot. Each B-spline is usually indexed by a unique knot, say the leftmost where the B-spline has support. Additional knots must be placed at the boundaries so that each B-spline spans the same number of knots. The knot placement may be general, allowing for unequal spacing. We denote the number of B-splines used in the regression as K, and at any given value of t there are exactly $q+1$ non-zero B-splines, and these values are used to construct the basis matrix B. Given m depths, a $m \times K$ regressor matrix can be constructed. B-spline smoothing is essentially multiple regression. Let $b_{ij} = B_j(t_i)$, $j = 1,\ldots,K$ indicates the value of the jth B-spline function at index t_i, and $B = [b_{ij}]$. The B-spline regressors (and their corresponding parameters) are anonymous in that they do not really have any scientific interpretation: rather predicted values are produced through linear combinations of the basis. We recommend the text by Dierckx (1993) for a nice overview.

Such a basis is well-suited for smoothing of a scatterplot of points (t_i, y_i), $i = 1,\ldots,m$. A smooth mean function can be expressed as $\mu = f(t) = B\alpha$, where B is a $m \times (K+q)$ regressor matrix and α is the unknown B-spline parameters. We minimize

$$S = ||y - B\alpha||^2, \tag{1}$$

with the explicit solution

$$\hat{\alpha} = (B'B)^{-1}B'y. \tag{2}$$

Given $\hat{\alpha}$, the estimated point on the curve at any (new) depth $t^\star$ is $\sum_{j=1}^{K} B_j(t^\star)\hat{\alpha}_j$.

3.2 Smoothing the KTB data

For the KTB data, $K = 17$ specific knots locations (at depths in meters) are chosen based on prior knowledge of lithologies (Winter et al. 2002), with values: 0, 290, 552, 1183, 1573, 2384, 2718, 3200, 3427, 3537, 5224, 5306, 5554, 5606, 7260, 7800, and 9101 m. The complete B-spline basis (for $q = 0,\ 1,\ 2,\ 3$) using the above knots locations is provided in Figure 4. Using the B-spline bases displayed in Figure 4, Figure 5 displays the estimated smooth mean function for the scatterplot of log(*CATR*) as a function of depth, for various bases degree and the specified $K = 17$ knots.

4 Using B-splines for Varying Coefficient Models

In addition to using smoothing techniques to estimate the mean response, consider broadening the model to control for another regressor, e.g. $x = H_2O$, which itself may also have a varying influence as a function of depth,

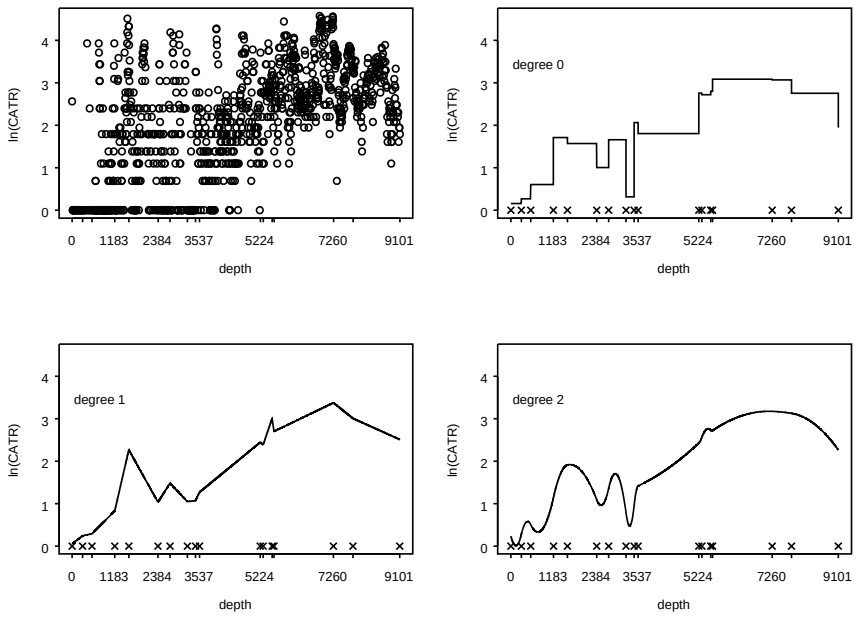

Fig. 5 Scatterplot of log(*CATR)* vs. *depth* and smooth estimated mean functions using B-splines of degree 0, 1, and 2. The "X" symbol indicates knot locations.

$$\mu(t) = \beta_0(t) + x(t)\beta_1(t). \tag{3}$$

This model is a generalization of the simple linear regression model ($\mu = \beta_0 + \beta_1 x$), where the static intercept and slope coefficients (β_0, β_1) are now replaced with coefficients that vary, and thus the regressor has a modified effect, for example depending on depth.

With B-spline smoothing and predetermined knots (along t), backfitting can be avoided and a varying coefficient model can be fit directly. This is clearly illustrated in matrix notation by modelling the mean response in (3),

$$\begin{aligned} \mu &= B\alpha_0 + \text{diag}\{x(t)\}B\alpha_1 \\ &= (B|U)(\alpha_0', \alpha_1')' = Q\alpha, \end{aligned}$$

where the matrix $\text{diag}\{x(t)\}$ aligns the regressors with the appropriate slope value that is also smooth in t, i.e. $\beta_1(t) = B\alpha_1$. Note that the same B basis, built on the t axis, is used for both smooth components. This can be done with data having one natural indexing variable, e.g. as with depth in the KTB data. In general, there can be a different indexing variable for each varying coefficient, thus requiring differing B-spline bases for each additive term. We see that the effective regressors are $Q = (B|U)$, where $U = \text{diag}\{x(t)\}B$, which results in essentially a modest sized "multiple regression" problem. Notice that U boils down to nothing more than a simple row scaling of B. Straightforward least squares techniques similar to (2) are used to estimate the unknown B-spline parameters $\alpha = (\alpha_0, \alpha_1)'$ associated with the smooth intercept and slope. We minimize

$$S = ||y - Q\alpha||^2, \tag{4}$$

with the explicit solution

$$\hat{\alpha} = (Q'Q)^{-1}Q'y. \tag{5}$$

Thus estimated smooth coefficients can be constructed using $B\hat{\alpha}_j$ ($j = 0,\ 1$), and $\hat{\mu} = Hy = Q\hat{\alpha}$, where the "hat" matrix is $H = Q(Q'Q)^{-1}Q'$.

Additive B-spline VCMs

The generalization to (3) follows for p regressors, each having varying slopes,

$$\mu(t) = \beta_0(t_0) + \sum_{j=1}^{p} \beta_j(t_j)x_j(t_j) \tag{6}$$

In matrix notation,

$$\mu = B\alpha_0 + \sum_{j=1}^{p} \text{diag}\{x_j(t_j)\}B_j\alpha_j$$

$$= (B|U_1|\ldots|\,U_p)(\alpha_0', \alpha_1', \ldots \alpha_p')' = R\theta, \tag{7}$$

where generalizations of (4) and (5) follow naturally using R and θ. Notice that B_j is used in (6) to allow the differing indexing variables (t_j) for each regressor, $j = 1, \ldots, p$.

For illustration, Figures 6 and 7 display the fixed knot KTB varying coefficients using B-splines of degree 0 and 3, respectively.

5 P-spline Snapshot: Equally-Spaced Knots & Penalization

The B-spline approach in the previous section required knowledge of the location and number of knots. In general, this information may not be known, and the placement of the proper number of knots is a complex nonlinear optimization problem. Circumventing these decisions, Eilers & Marx (1996) proposed an alternative P-spline smoothing approach, which has two steps to achieve smoothness: (i) Use a

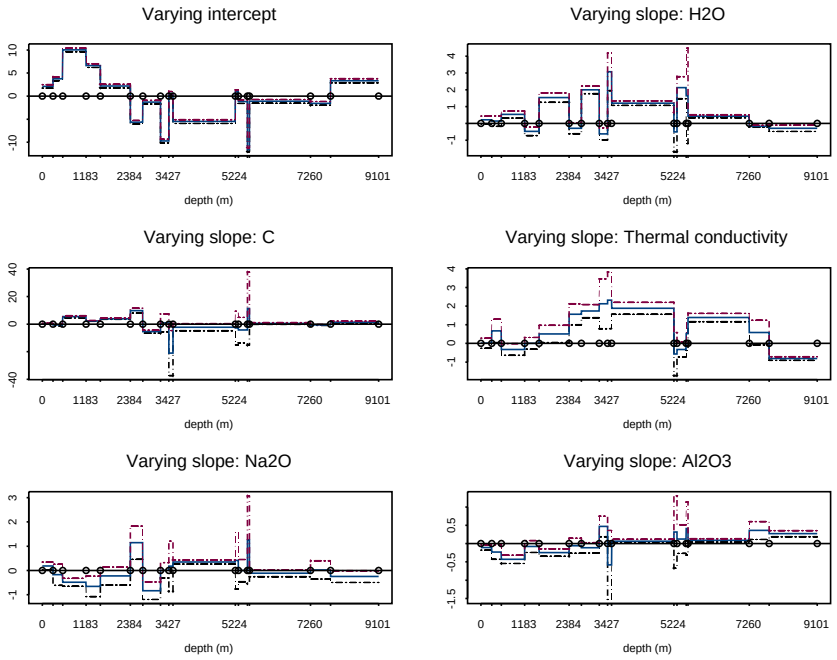

Fig. 6 Using log(*CATR)* as response, varying intercept and varying slopes for H_2O, *C, Thermal Conductivity,* Na_2O, Al_2O_3 using B-spline bases of degree 0. Twice standard bands are provided. Knots locations are indicated by both ticks and circles.

rich regression basis to purposely overfit the smooth coefficient vector with a modest number of (equally-spaced) B-splines. (ii) Ensuring further and the proper amount of smoothness through a difference penalty on adjacent B-spline coefficients. The main idea is that smoothness is driven by the amplitudes of α, and discouraging estimates of α that have erratic adjacent (neighboring) behavior can be sensible. A non-negative tuning parameter regularizes the influence of the penalty, with large (small) values leading to heavy (light) smoothing. For one smooth term, we now minimize

$$S^{\star} = ||y - B\alpha||^2 + \lambda ||D_d \alpha||^2. \tag{8}$$

The matrix D constructs dth order differences of α:

$$D_d \alpha = \Delta^d \alpha. \tag{9}$$

The first difference of α, $\Delta^1 \alpha$ is the vector with elements $\alpha_{j+1} - \alpha_j$, for $j = 1, \ldots, K-1$. By repeating this computation on $\Delta\alpha$, we arrive at higher differences like $\Delta^2 \alpha = \{(\alpha_{j+2} - \alpha_{j+1}) - (\alpha_{j+1} - \alpha_j)\}$ and $\Delta^3 \alpha$. The $(n-1) \times n$ matrix D_1 is sparse, with $d_{j,j} = -1$ and $d_{j,j+1} = 1$ and all other elements zero. Examples of D_1

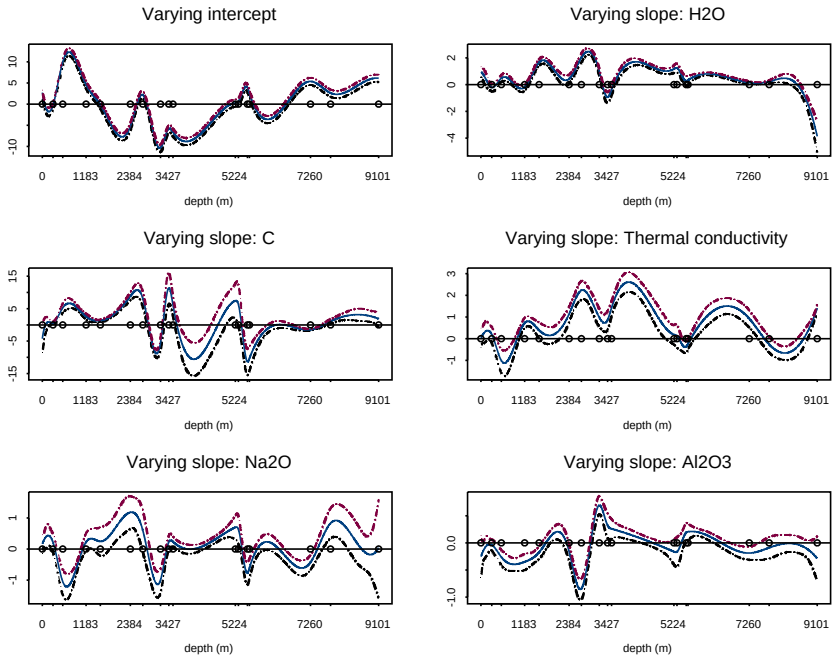

Fig. 7 Using log(*CATR)* as response, varying intercept and varying slopes for H_2O, *C, Thermal Conductivity,* Na_2O, Al_2O_3 using B-spline bases of degree 3. Twice standard bands are provided. Knots locations are indicated by both ticks and circles.

and D_2 of small dimension look like

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

Actually, the number of equally-spaced knots does not matter much provided that enough are chosen to ensure more flexibility than needed: the penalty further smoothes with continuous control. The solution of (8) is

$$\hat{\alpha}_\lambda = (B'B + \lambda D_d' D_d)^{-1} B'y, \tag{10}$$

and the "effective" hat matrix is now given by

$$H_\lambda = B(B'B + \lambda D_d' D_d)^{-1} B'. \tag{11}$$

5.1 P-splines for Additive VCMs

When considering more regressor terms and in a VCM context, the model is as outlined in (6) with $\mu(t) = R\theta$, but now B is a rich basis using equally-spaced knots. The P-spline objective function in (8) must be further modified to allow differing flexibility across the p regressors, i.e. a separate λ is allowed for each term. We now have

$$S^\star = ||y - R\theta||^2 + \sum_{j=0}^{p} \lambda_j ||D_d \alpha_j||^2, \tag{12}$$

with a solution

$$\hat{\theta} = (R'R + P)^{-1} R'y,$$

where the penalty takes on the form $P = \text{block diag}(\lambda_0 D_d' D_d, \ldots, \lambda_p D_d' D_d)$. The block diagonal structure breaks linkage of penalization from one smooth term to the next one. Note that (12) uses a common penalty order d, but there is nothing prohibitive from allowing some terms to have different d. Thus

$$\hat{\mu} = R\hat{\theta} = Hy,$$

where $H = R(R'R + P)^{-1} R'$. Borrowing from Hastie & Tibshirani (1990), the effective dimension of the fitted smooth P-spline model is approximately trace(H). By noting the lower dimension and invariance of the trace of cyclical permutated matrices, effective dimension (ED) can be found efficiently using

$$\text{ED}(\lambda) = \text{trace}\{(R'R + P)^{-1} R'R\}. \tag{13}$$

The effective dimension of each smooth term is the trace of the portion of diagonal terms of H corresponding to each term.

5.2 Standard Error Bands

For fixed λ, twice standard error bands can be constructed relatively easily, and can be used as an approximate inferential tool, for example to identify potentially important depth windows that may relate each regressor to the response. We have

$$\text{var}(\hat{\theta}) = (R'R+P)^{-1}R'\sigma^2 IR(R'R+P)^{-1} = \sigma^2(R'R+P)^{-1}R'R(R'R+P)^{-1}.$$

Thus the covariance matrix associated with the jth smooth component is

$$C_j = \sigma^2 B_j\{(R'R+P)^{-1}R'R(R'R+P)^{-1}\}_j B_j',$$

where $\{\cdot\}_j$ denotes the diagonal block associated with the jth component. The square root of the diagonal elements of C_j are used for error bands, as used in Figure 3. Setting $\lambda = 0$ yields the standard error bands for unpenalized B-splines, as presented in Figures 6 and 7.

6 Optimally Tuning P-splines

For B-spline models, apriori information is essential: The amount of smoothing is determined by the size of the B-spline basis and thus implicitly by the number and position of knots. The smaller the number of knots, the smoother the curve. For P-spline models where R only contains a few smooth terms, cross-validation measures or information criteria can be monitored by varying λ in a systematic way over a grid, and the "optimal" values for the λ vector can be chosen as the one that minimizes, e.g., LOOCV. Although this prediction oriented approach for choosing λ is tractable for low dimensions, it can become computationally taxing and unwieldy, e.g. in our KTB application with six smooth terms. We investigate an alternative estimation-maximization (E-M) approach based on viewing P-splines as mixed models, based on the work of Schall (1991), which appears very promising.

First we consider only one smooth term and then propose a generalized algorithm. Using a mixed model with random α, the log-likelihood, l, can be expressed as

$$-2l = m\log\sigma + n\log\tau + \frac{\|y - B\alpha\|^2}{\sigma^2} + \frac{\|D\alpha\|^2}{\tau^2}, \tag{14}$$

where the $\text{var}(\alpha) = \tau^2$ is the variance of the random effects and $\text{var}(\varepsilon) = \sigma^2$ is the variance of the random error. Maximizing (14) results in the system of equations

$$\left(B'B + \frac{\sigma^2}{\tau^2}D'D\right)\alpha = B'y,$$

and hence we can view $\lambda = \sigma^2/\tau^2$ as a ratio of variances. We also have, under expectation, that

$$E(\|y - B\hat{\alpha}\|^2) \approx (m - ED) \times \sigma^2$$
$$E(\|D\hat{\alpha}\|^2) \approx ED \times \tau^2, \tag{15}$$

where ED is the approximate effective dimension of the fit. Using (15), we can get a current estimate $\hat{\sigma}^2$ and $\hat{\tau}^2$ from fit. An updated fit can be made using updated $\hat{\sigma}^2/\hat{\tau}^2$, until convergence. We propose a generalized estimation-maximization (E-M) algorithm for the p-dimensional varying coefficient model $\mu = R\theta$:

`Algorithm E-M P-spline to optimize` λ

1. Initializations:
   - Generously choose knots K (use 40 as default).
   - Initialize $\lambda,\ j = 1, \ldots, p$ (use 10^{-5} as default)
   - Choose B-spline basis degree q (cubic as default)
   - Choose penalty order d (use 3 as default)
   - Construct Penalty $P = \text{blockdiag}(\lambda_0 D'D, \ldots, \lambda_p D'D)$
   - $\hat{\theta} = (R'R + P)^{-1}R'y$
2. Cycle until $\Delta\lambda$ small
3. For $j = 0$ to p
   a. Compute the $ED_j = \text{trace}\{H\}_j$ (jth smooth diagonals in H)
   b. Estimation (E-step):
      i. $\hat{\sigma}^2 = \frac{\|y - R\hat{\theta}\|^2}{m - \sum_{j=0}^{p} \text{ED}_j}$
      ii. $\hat{\tau}_j^2 = \frac{\|D\hat{\theta}\|^2}{\text{ED}_j}$
      iii. $\hat{\lambda}_j = \frac{\hat{\sigma}^2}{\hat{\tau}_j^2}$
   c. Maximization (M-step):
      i. $P = \text{blockdiag}(\hat{\lambda}_0 D'D, \ldots \hat{\lambda}_p D'D)$
      ii. $\hat{\theta} = (R'R + P)^{-1}R'y$
4. Fit with converged vector $\hat{\lambda}$

`end algorithm`

Cross-validation Prediction Performance

A leave-one-out cross-validation measure can be computed swiftly, only requiring the diagonal elements of the "hat" matrix, h_{ii}, and the residuals $y - \hat{\mu} = y - R\hat{\theta}$. Note

$$y_i - \hat{\mu}_{-i} = (y_i - \hat{\mu}_i)/(1 - h_{ii}), \tag{16}$$

$\hat{\mu} = B(B'B)^{-1}B'y = Hy$, and $\hat{\mu}_{-i}$ is the fitted value for y_i that would be obtained if the model were estimated with y_i left out. It follows that $h_{ii} = r_i'(R'R + P)^{-1}r_i$, where r_i' indicates the ith row of R. Hence the diagonal elements of H and the cross-validation residuals can be computed with little additional work. We can define

Table 1 Preliminary goodness-of-fit and cross-validation, by VCM degree.

Method	Basis q	Penalty d	Knots K	Eff. Dim	LOOCV	R^2
E-M P-spline	3	3	equally 40	155.8	0.737	0.704
E-M P-spline	0	1	equally 40	206.1	0.755	0.691
B-spline	3	-	fixed 17	120	0.773	0.683
B-spline	2	-	fixed 17	120	0.765	0.685
B-spline	1	-	fixed 17	120	0.779	0.670
B-spline	0	-	fixed 17	120	0.800	0.647

$$\text{LOOCV} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{\mu}_{-i})^2} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i - \hat{\mu}_i}{1-h_{ii}}\right)^2},$$

and this result holds in the unpenalized setting by simply setting all $\lambda = 0$ in H.

7 More KTB Results

The P-spline approach was fit using $K = 40$ knots for each of the six smooth components and corresponding difference penalties of order $d = 3$. Summary results are presented in Table 1, for various approaches. For the case $d = 3$, the optimal tuning parameters were chosen using the E-M algorithm above, which converged in 69 cycles, and yielded optimal $\lambda = (570, 0.038, 0.0023, 908, 1252, 6.63)$, respectively. Figure 3 presents the corresponding E-M based estimated smooth coefficients. The convergence criterion was $\max_j\{\Delta\lambda_j/\lambda_j\} < 10^{-8}$. The overall effective dimension of the P-spline fit was ED $= 155.8$. Notice that as λ increases, then ED decreases. When comparing P-splines (Figure 3) to the B-spline approach with unequally-spaced $K = 17$ knots (Figures 6 and 7), we find some general differences. First, the optimal overall ED is higher with P-splines (155.8), when compared to that of each B-spline ED (120), since each B-spline term has an ED=20. Further, some of the P-spline smooth terms need much less ED, e.g. intercept (11.8), *Thermal Conductivity* (15.0), and Na_2O (14.5), whereas other P-spline terms require considerably more ED, e.g. H_2O (40.6), C (34.4), and Al_2O_3 (39.6). We find that the general patterns of negative, positive, and moderate smooth coefficients is preserved from Figures 6 and 7, as a function of depth. However, the P-spline coefficients are smoothed in some cases, and sharpened in others. This P-spline approach required no prior knowledge of depth knots, and yields a very competitive model with an $R^2 = 0.704$– a considerable improvement over previously reported models. The CV value is 0.737, which is the lowest among models presented. Thus the P-VCM model for the KTB data experiences both increase in R^2 and reduction in CV error.

8 Extending P-VCM into the Generalized Linear Model

When responses are non-Normal, e.g. binary outcomes or Poisson counts, the P-spline varying coefficient model extends naturally into the generalized linear model (GLM) framework,

$$g(\mu(t)) = \beta_0(t_0) + \sum_{j=1}^{p} \beta_j(t_j)x_j(t_j) \tag{17}$$

In matrix notation,

$$\begin{aligned} g(\mu) &= B\alpha_0 + \sum_{j=1}^{p} \text{diag}\{x_j(t_j)\}B_j\alpha_j \\ &= (B|U_1|\ldots|\,U_p)(\alpha_0', \alpha_1', \ldots \alpha_p')' = R\theta, \end{aligned} \tag{18}$$

where the subscript j (on both t and B) highlights that differing indexing variables are allowed for each regressor. The GLM allows a (monotone) link function $g(\cdot)$ and requires independent observations from any member of the exponential family of distribution with $\mu = E(Y)$. The specifics of the GLM are well documented and tabulated, e.g. in Fahrmeir & Tutz (2001, Chapter 2).

The penalized objective function for the GLM is now

$$l^\star = l(\theta) - \sum_{j=0}^{p} \lambda_j ||D_d\alpha_j||^2, \tag{19}$$

where $l(\theta)$ is the log-likelihood function, which is a function of θ since $\mu = h(R\theta)$. The inverse link function is denoted as $h(\cdot)$ (with derivative $h'(\cdot)$). We now maximize $l^\star$ and find above that the penalty terms are now subtracted from $l(\theta)$, thus discouraging roughness of any varying coefficient vector. Fisher's scoring algorithm results in the iterative solution

$$\tilde{\theta}_{c+1} = (R'\tilde{V}_cR + P)^{-1}R'\tilde{V}_c\tilde{z}_c,$$

where again the penalty takes on the form $P = \text{block diag}(\lambda_0 D_d'D_d, \ldots, \lambda_p D_d'D_d)$, and $V = \text{diag}\{h'(R\theta)/\text{var}(y)\}$, $z = (y-\mu)/h'(R\theta) + R\theta$ are the usual GLM diagonal weight matrix and "working" dependent variable, respectively, at the current iteration c. Upon convergence, $\hat{\mu} = h(R\hat{\theta}) = h(\hat{H}y)$, with $\hat{H} = R(R'\hat{V}R + P)^{-1}R'\hat{V}$, and approximate effective dimension $ED \approx \text{trace}\{R'\hat{V}R(R'\hat{V}R + P)^{-1}\}$.

Polio Example with Poisson Counts

We apply P-VCM models to the discrete count time series data of monthly polio incidences in the United States (reported to the U.S. Center of Disease Control) during the years 1970 through 1987. The data are taken from Zeger (1988) and

further analyzed by Eilers & Marx (2002). The monthly mean count is modeled with a penalized GLM with a Poisson response and log link function. We choose a model that allows a varying intercept, as well as varying slopes for the cosine and sine regressors (each with both annual and semi-annual harmonics),

$$\log(\mu(t)) = f_0(t) + \sum_{k=1}^{2} \{ f_{1k}(t)\cos(k\omega t) + f_{2k}(t)\sin(k\omega t)\}, \tag{20}$$

where $\omega = 2\pi/12$ for the index $t = 1, \ldots, 216$. In matrix notation, we have

$$\log(\mu) = B\alpha_0 + \sum_{k=1}^{2} \{C_k B\alpha_{ck} + S_k B\alpha_{sk}\} = R\theta, \tag{21}$$

where $R = (B \mid C_1B \mid C_2B \mid S_1B \mid S_2B)$ and θ is the corresponding vector of augmented α's. The C and S are diagonal cosine and sine matrices that repeated cycle through the months (1 through 12) using the appropriate harmonic. Since the index is common for all regressors, we conveniently choose to use a common (cubic) basis B. Figure 8 displays the varying harmonic effects. We used 13 equally-spaced knots and a second order penalty for each term. Related to the work of Schall (1991),

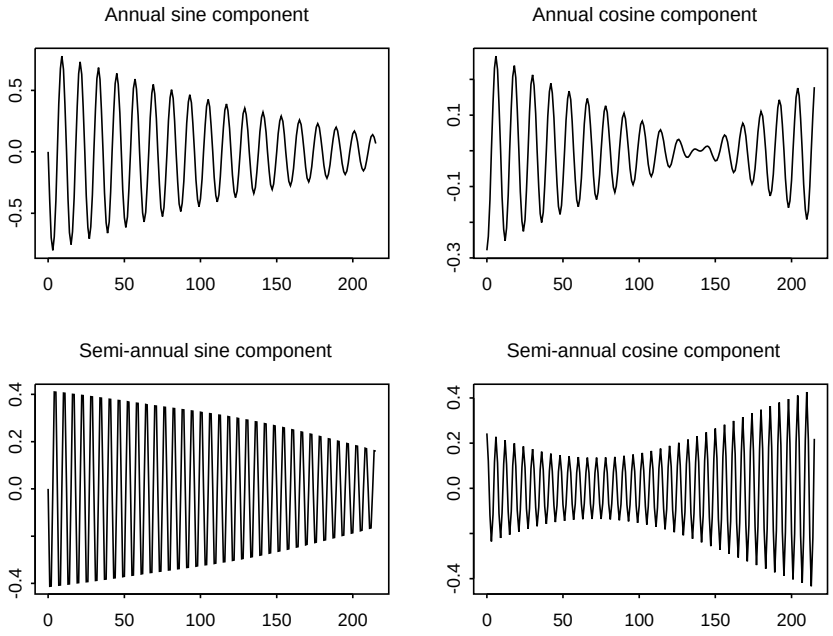

Fig. 8 Polio example: the annual and semi-annual varying cosine and sine effects.

the optimal values of λ are also found using the E-M algorithm found in Section 6 (with small modification): 1. The estimation of the scale parameter is fixed to be one (step 3.b.i), and 2. Although no backfitting is performed, the maximization step is now the iterative method of scoring (step 3.c.ii). The estimate effective dimension is approximately 6.5 for the intercept, and 1.5 for each of the sine and cosine terms.

9 Two-dimensional Varying Coefficient Models

An advantage of the P-spline approach to varying coefficient modeling is its ability to adapt to a variety of extensions with relatively little complication. We will see that it is rather straightforward to extend to an additive two-dimensional varying coefficient model in a generalized linear model setting. Such an approach requires P-VCM to use a tensor product B-spline basis and to use some care in constructing a sensible penalty scheme for the coefficients of this basis. In this way P-VCM remains nothing more than a moderately (generalized) penalized regression problem. Consider the tensor product basis provided in Figure 9. The basis is sparsely presented to give an impression of its structure; a full basis would have severe overlapping "mountains". Corresponding to each basis, there is an array of coefficients $\Theta = [\theta_{kl}]$, $k = 1,\ldots,\ K$ and $l = 1,\ldots,\ L$ (one for each mountain), and these are the drivers of the two-dimensional varying coefficient surfaces. To avoid the difficult issue of optimal knot placement, P-VCM again takes two steps: (i) Use a rich $K \times L (< 1000)$ gridded tensor product basis that provides more flexibility than needed. (ii) Attach difference penalties on each row and on each column of θ with only one tuning parameter for rows and another one for columns. Figure 10 gives an idea of strong penalization of the coefficients.

9.1 Mechanics of 2D-VCM through Example

Figure 11 (top panel) displays log death counts resulting from respiratory disease for U.S. females. The image plot is actually 25,440 cells resulting from the cross-classification of age by monthly time intervals. Details of the data, as well as a thorough modeling presentation can be found in Eilers et al. (2008). The lower panel of Figure 11 display the marginal death count over time, which exhibits a strong and varying seasonal cyclical behavior. Consider the Poisson regression with a log link function

$$\log(\mu_{at}) = \nu_{at} + f_{at}\cos(\omega t) + g_{at}\sin(\omega t) = \eta_{at}, \tag{22}$$

with counts Y_{at} and $\mu_{at} = E(Y_{at})$. For simplicity, we suppress any offset term. The index $a = 1,\ldots,A$ refers to regressor *age* $(44-96)$, whereas year and month are combined to create a variable *time*, indexed by $t = 1,\ldots,T$ $(1-480)$. Annual cyclical behavior in the counts is modeled using the periodic sine and cosine regressors,

with period 2π ($\omega = 2\pi/12$). More harmonics can be added as needed. The two regressors are only indexed with t since the cyclical behavior is only assumed to be associated with *time*. The parameters $v,\ f,\ g$ are indexed by both $(a,\ t)$ and are the smooth (two-dimensional) varying coefficient surfaces for the intercept and slopes for the sine and cosine regressors, respectively.

To express each of the intercept, sine, and cosine varying coefficients smoothly, it is perhaps natural to work with a vectorized form of Θ denoted as $\theta_u = \text{vec}(\Theta_u)$, $u = 0,1,2$. A "flattened" tensor product B-spline basis B can be formed of dimension $AT \times KL$, such that $\text{vec}(s) = B\theta_0$, $\text{vec}(f) = B\theta_1$, and $\text{vec}(g) = B\theta_2$. Each row of B designates one of the AT cell counts, and the columns contain the evaluations of each of the KL basis at that cell location. In matrix terms, (22) can be reexpressed as

$$\begin{aligned} \text{vec}\{\log(\mu)\} &= B\theta_0 + \text{diag}\left[\cos(\omega t)\right] B\theta_1 + \text{diag}\left[\sin(\omega t)\right] B\theta_2 \\ &= B\theta_0 + U_1\theta_1 + U_2\theta_2 \\ &= M\theta, \end{aligned} \tag{23}$$

where $M = [B|U_1|U_2]$ and $\theta' = (\theta_0', \theta_1', \theta_2')$ are the augmented bases and tensor product coefficients, respectively. The diagonalization of the regressors in (23) ensures that the each level of the regressor is weighted by its proper level of the varying coefficient. We now find (23) to be a standard Poisson regression model with effective

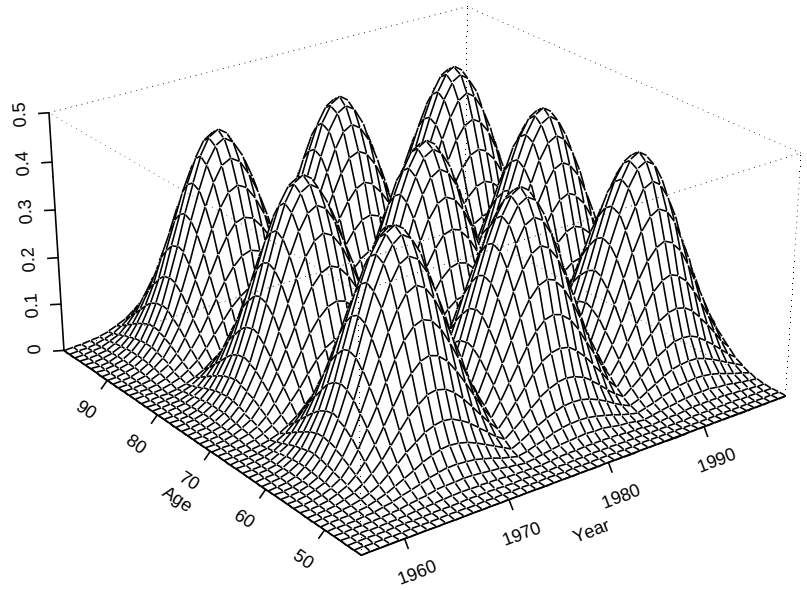

Fig. 9 A sparse portion of a tensor product B-spline basis.

regressors M of dimension $AT \times KL$ and unknown coefficients θ. The dimension of estimation is now reduced from initially $3 \times AT$ to $3 \times KL$.

9.2 VCMs and Penalties as Arrays

Consider the univariate basis: Let $B = [b_{tk}]$ ($\breve{B} = [\breve{b}_{al}]$) be the $T \times K$ ($A \times L$) B-spline basis on the time (age) domain. Denote $\mathscr{A}$, $\mathscr{B}$, and $\mathscr{C}$ as the $K \times L$ matrices of the tensor product coefficients for $V = [v_{ta}]$, $F = [f_{ta}]$, and $G = [g_{ta}]$ respectively. We can rewrite (23) as

$$\begin{aligned} \log(M) &= V + CF + SG \\ &= B\mathscr{A}\breve{B}' + CB\mathscr{B}\breve{B}' + SB\mathscr{C}\breve{B}', \end{aligned} \tag{24}$$

where $M = [\mu_{ta}]$ and C and S represent the (co)sine diagonal matrices defined in (23), and again any offset term is suppressed.

Penalties are now applied to both rows and columns of $\mathscr{A}$ and $\mathscr{B}$. Denote the (second order) difference penalty matrices D and $\breve{D}$ with dimensions $(K-2) \times K$ and $(L-2) \times L$, respectively. Recall Figure 10 that provides a visualization of strong row and column penalization.The penalty is defined as $P = P_{\mathscr{A}} + P_{\mathscr{B}} + P_{\mathscr{C}}$, with the first term having the form $P_{\mathscr{A}} = \{\lambda_1 ||D\mathscr{A}||_F + \breve{\lambda}_1 ||\mathscr{A}\breve{D}'||_F\}$ with the other naturally following for $\mathscr{B}$ and $\mathscr{C}$. We denote $||\cdot||_F$ as the Frobenius norm, or the sum of the squares of all elements. The first portion of the penalty is equivalently

$$P_{\mathscr{A}} = \text{vec}(\mathscr{A})'[\lambda_1(I_L \otimes D'D) + \breve{\lambda}_1(\breve{D}'\breve{D} \otimes I_K)]\text{vec}(\mathscr{A}),$$

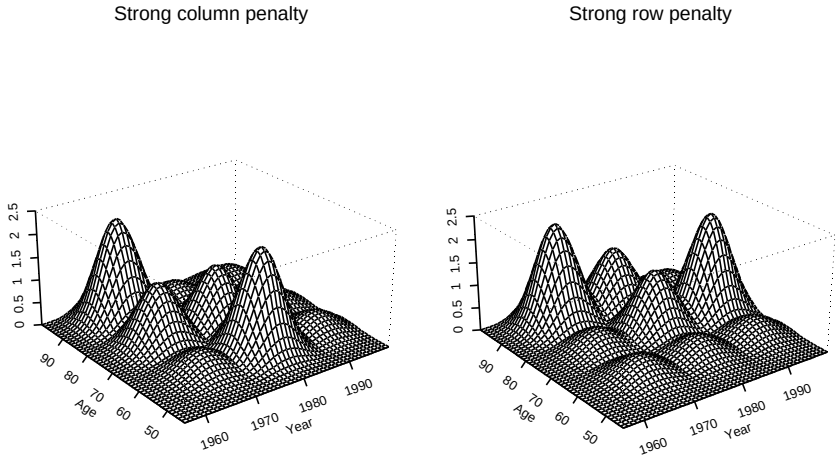

Fig. 10 A sparse portion of a strongly penalized tensor product B-spline basis.

where I is the identity matrix. The tensor product coefficients, $\mathscr{A}$, $\mathscr{B}$ and $\mathscr{C}$ are found by maximizing the penalized Poisson log-likelihood function

$$l^{\star}(\mathscr{A},\mathscr{B}) = l(\mathscr{A},\mathscr{B}) - \frac{1}{2}P. \tag{25}$$

Optimization of the tuning parameters (six in this case) can be found using efficient clever searches, in a greedy way, over the λ space to minimize, e.g. AIC or QIC. Also an extension to the E-M algorithm is possible. Figure 12 presents optimal results based on QIC for the respiratory data using 13×13 equally-spaced tensor products and a second order penalty on rows and columns for each component.

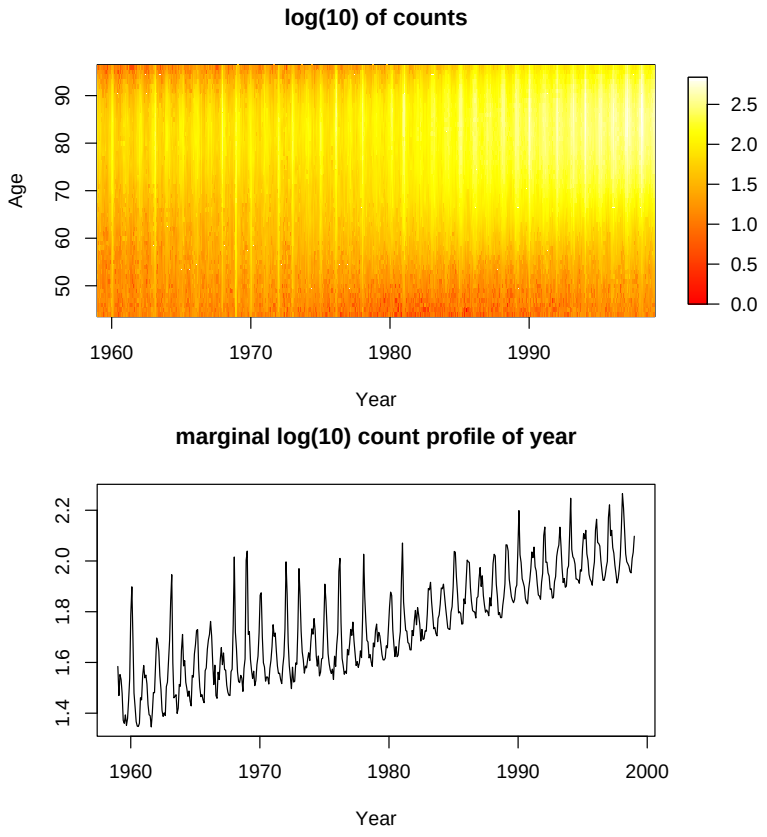

Fig. 11 Raw counts of female respiratory deaths in U.S. for ages 44–96 during 1959-1999 (top) and the marginal plot of time trend (bottom)

9.3 Efficient computation using array regression

The array algorithm can be found in Currie et al. (2006). Without loss of generality, using only the first term in (24), the normal equations can be expresses as

$$(\breve{B} \otimes B)' W (\breve{B} \otimes B) \hat{\alpha} = Q\hat{\alpha} = (\breve{B} \otimes B)' W y, \tag{26}$$

where W is a diagonal weight matrix and $y = \text{vec}(Y)$. With the dimension of $\breve{B} \otimes B$ is $AT \times KL$, and can require much of memory space. Also, but perhaps less obvious, the multiplications and sums that lead to the elements of Q are rather fine-grained and waste an enormous amount of processing time. The problem is compounded when considering all terms in (24). Both problems are eliminated with by rearranging the computations.

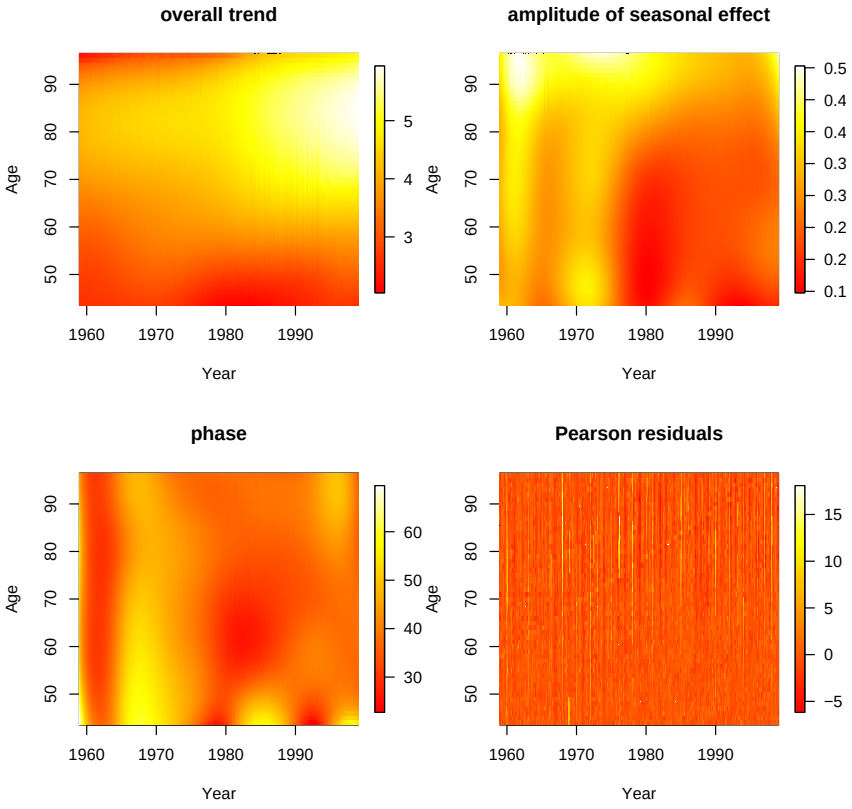

Fig. 12 Fit for female respiratory deaths: varying intercept (trend) (top, left); varying amplitude of (co)sine (top, right); varying phase in months of (co)sine (bottom, left); Pearson residuals (bottom, right)

Let $R = B\square B$ indicate the row-wise tensor product of B with itself. Hence R has T rows and K^2 columns and each row of R is the tensor product of the corresponding row of B with itself. One can show that the elements of

$$G = (B\square B)'W(\check{B}\square\check{B})$$

have a one-to-one correspondence to the elements of Q. Of course they are arranged differently, because Q has dimensions $KL \times KL$ and G dimensions $K^2 \times L^2$. However, it is easy to rearrange the elements of G to get Q. Three steps are needed: 1) re-dimension G to a four-dimensional $K \times K \times L \times L$ array; 2) permute the second and third dimension; 3) re-dimension to a $KL \times KL$ matrix.

A similar, but simpler computation finds the right side of (26) by computing and rearranging $B'(W \cdot Y)\check{B}$, where $W \cdot Y$ indicates the element-wise product of W and Y. In a generalized additive model or varying-coefficient model with multiple tensor product bases, weighted inner products of the different bases have to be computed using the same scheme as outlined above. Array regression offers very efficient computation with increases in fitting speed (of far more than 10-fold in most cases) when compared to the following unfolded representation. Typically array regression is used when the data are on a regular grid, however it is possible to include a mix of array and other standard regressors.

10 Discussion Toward More Complex VCMs

The adaptive nature and strength of P-splines allows extensions to even more complex models. We have already seen such evidence in this paper by moving from simple to additive P-VCMs, from standard to generalized settings, and from one-dimensional coefficient curves to two-dimensional coefficient surfaces. P-VCMs can also be extended into bilinear models, as presented in Marx et al. (2010). In all cases, P-VCMs further remain grounded in classical or generalized (penalized) regression, allowing swift fitting and desirable diagnostics, e.g. LOOCV.

P-VCMs can be broadened into higher dimensions, e.g. to have three-dimensional varying coefficient surfaces, and with several additive components. Heim et al. (2007) have successfully applied these models to brain imaging applications. Such a model is primarily achieved by broadening the tensor product basis from two to three dimensions and projecting the smooth three-dimensional coefficients onto this lower dimensional space. An additional penalty is needed for the third dimension or layer. In this setting, array regression is of utmost importance due to the formidable dimension of the unfolded design matrix and the number of computations to obtain, e.g., the information matrix. There is nothing prohibitive in P-VCM to consider even higher, e.g. four, dimensional VCM surfaces.

The P-VCM approach also lends itself nicely to high dimensional regression commonly present in chemometric applications, often referred to the multivariate calibration problem. In this setting, the scalar response has digitized “signal” re-

gressors, ones that are ordered and actually ensemble a curve. Marx & Eilers (1999) used P-splines to estimate smooth coefficient curves, but tensor product P-VCMs can allow these smooth coefficient curves to vary over another dimensions. Figure 13 provides an example of how smooth high dimensional coefficient curves can vary over a third variable, temperature. Eilers & Marx (2003) show how to construct such special varying coefficient surfaces, while drawing connections to lower dimensional ribbon models and additive models.

There are various details that will need further investigation. Although it is not always easy to make complete and thorough comparisons across a wide range of other methods under exhaustive settings, it would be interesting to compare the P-VCM approach to Bayesian counterparts (Lang & Brezger 2004), mixed model counterparts (Ruppert, Wand & Carroll 2003) and structural regression approaches (Fahrmeir et al. 2004). Further, we only dampen any effects of serial correlation in data through the use of a varying intercept in the model. In fairness, a more formal investigation of any possible auto-regressive (AR) error structure should be made, e.g. addressing deep drill depth varying covariance similarly to Kauermann & Küchenhoff (2003), $\text{corr}(y_i, y_{i+1}) = \rho(\tilde{t})^{|t_i - t_{i+1}|}$, where ρ is a smooth function in depth and $\tilde{t} = (t_i + t_{i+1})/2$. Additionally, although we extended E-M algorithm of Schall (1991) to optimize tuning parameters in the standard and generalized settings,

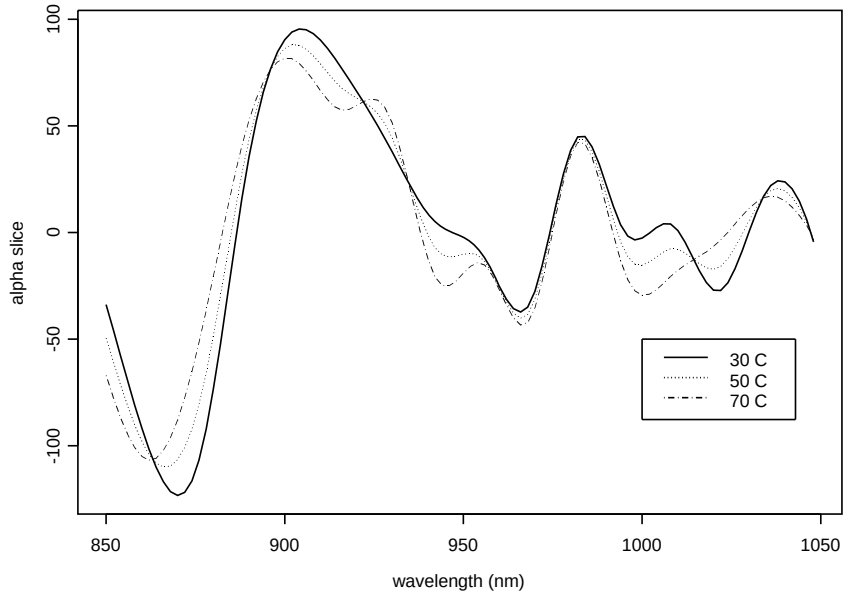

Fig. 13 Various slices of smooth signal regressors that vary over a third variable, temperature

the theory of this approach could be more formally grounded, and the stability of the algorithm should be investigated.

Acknowledgements I would like to thank Paul H.C. Eilers for his generous time and his numerous thought provoking conversations with me that led to a significantly improved presentation.

References

Currie, I. D., Durbán, M. & Eilers, P. H. C. (2006). Generalized linear array models with applications to multidimensional smoothing. *Journal of the Royal Statistical Society: Series B*, **68**(2), 259–280.

Dierckx, P. (1995). *Curve and Surface Fitting with Splines*. Clarendon Press, Oxford.

Eilers, P. H. C., Gampe, J., Marx, B. D. & Rau, R. (2008). Modulation models for seasonal life tables. *Statistics in Medicine*, **27**(17): 3430–3441.

Eilers, P. H. C. & Marx, B. D. (2003). Multivariate calibration with temperature interaction using two-dimensional penalized signal regression. *Chemometrics and Intelligent Laboratory Systems*, **66**, 159–174.

Eilers, P. H. C. & Marx, B. D. (2002). Generalized linear additive smooth structures. *Journal of Computational and Graphical Statistics*, **11**(4), 758–783.

Eilers, P. H. C. & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statistical Science*, **11**, 89–121.

Fahrmeir, L., Kneib, T. & Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica*, **14**, 731–761.

Farhmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models (2nd Edition)*. Springer, New York.

Hastie, T. & Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hastie, T. & Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, B*, **55**, 757–796.

Heim, S., Fahrmeir, L., Eilers, P. H. C., & Marx, B. D.(2007). Space-varying coefficient models for brain imaging. *Computational Statistics and Data Analysis*, **51**, 6212–6228.

Kauermann, G. & Küchenhoff, K. (2003). Modelling data from inside the Earth: local smoothing of mean and dispersion structure in deep drill data. *Statistical Modelling*, **3**, 43–64.

Lang, S. & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, **13**(1), 183–212.

Marx, B. D. & Eilers, P. H. C. (1999). Generalized linear regression on sampled signals and curves: a P-spline approach. *Technometrics*, **41**, 1–13.

Marx, B. D., Eilers, P. H. C., Gampe, J. & Rau, R. (2010). Bilinear modulation models for seasonal tables of counts. *Statistics and Computing*. In press.

Ruppert, D., Wand, M. P. & Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge University Press, New York.

Schall, R. (1991). Estimation in generalized linear models with random effects. *Biometrika*, **78**, 719–727.

Winter, H., Adelhardt, S., Jerak, A. & Küchenhoff, H. (2002). Characteristics of cataclastic shear zones of the ktb deep drill hole by regression analysis of drill cuttings data. *Geophysics Journal International*, **150**, 1–9.

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, **75**, 621–629.