Multidimensional Penalized Signal Regression

Brian D. Marx
Department of Experimental Statistics
Louisiana State University
Baton Rouge, LA 70803 USA
(bmarx@lsu.edu)

Paul H. C. Eilers
Department of Medical Statistics
Leiden University Medical Center
2300 RA, Leiden, The Netherlands
(p.eilers@lumc.nl)

December 1, 2004

Abstract

We propose a general approach to regression on digitized multidimensional signals that can pose severe challenges to standard statistical methods. The main contribution of this work is to build a two-dimensional coefficient surface that allows for interaction across the indexing plane of the regressor array. We aim to use the estimated coefficient surface for reliable (scalar) prediction. We assume that the coefficients are smooth along both indices. We present a rather straight-forward and rich extension of penalized signal regression using penalized B-spline tensor products, where appropriate difference penalties are placed on the rows and columns of the tensor product coefficients. Our methods are grounded in standard penalized regression, thus cross-validation, effective dimension and other diagnostics are accessible. Further the model is easily transplanted into the generalized linear model framework. An illustrative example motivates our proposed methodology and performance comparisons are made to other popular methods.

Keywords: Fluorescence; multivariate calibration; P-splines; signal regression; spectra; tensor product.

1 Introduction

Modern technology and instrumentation routinely delivers rich multidimensional data for which new statistical tools are needed. In many cases, we cannot use standard least squares statistical modelling approaches, for example when using the numerous digitizations from spectra as explanatory variables in a regression setting. The problem is inherently ill-posed: the number of regressors generated for each experimental unit generally far exceeds the number of units collected in the study. Paradoxically, it is as if too much information is the problem source. Figure 1 displays signal regressors (at two different excitation levels) for each of $m = 268$ observations, coming from a sugar processing experiment (Bro, 1999). Actually each "signal" consists of numerous digitizations ($p = 571$) along the emission wavelength axis (275 to 560, by 0.5 nm). If such optical regressors are to be related, e.g. to a chemometric response, then some regularization is needed. Not only is $p \gg m$, but the regressors are highly correlated. Eilers and Marx (1999) provided a practical solution for such a one-dimensional signal regression problem by taking into account the (spatial) indexing axis of the regressors and forcing their estimated coefficients to be smooth (using P-splines). Among many approaches, partial least squares and principal component regression, have been successfully used for signal regression, but ignore spatial information. Eilers and Marx (2003) extended the notion of signal regression to a two-dimensional smooth surface, essentially adjusting the signal coefficient vectors for an additional (scattered) covariate, e.g. temperature. The surface could be viewed as a sequence of smooth signal coefficient vectors that varied smoothly (interacted) with temperature. Thus the signal coefficient vector corresponded to the slice of the surface at a given temperature. In this article, we consider more complicated regressors in the form of multidimensional signals.

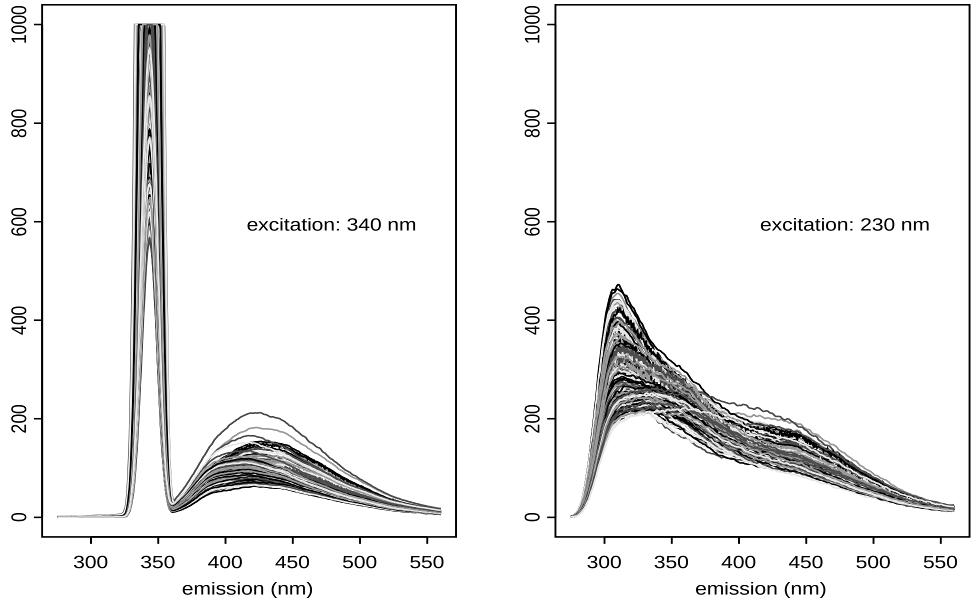

Figure 1: *Signal regressors for sugar experiment, at two different excitation levels.*

Notice that the left and right panels of Figure 1 presents "signals" at *excitation* levels of 340 and 230 nm, respectively, and one could imagine even more, forming a sequence of several "extremely narrow images". Thus a natural question to ask is: What if the signal regressors become fully two-dimensional, and we wish to take into account spatial information in both directions? One could view this problem as multivariate calibration with multidimensional spectra. Other regressor structures could consist of digitized images, even beyond two-dimensions. We met unique challenges from a fluorescence spectroscopy experiment: thus motivating the development of new statistical methodologies that we propose in this article. The data structure is as follows, each observation consists of the data pair: (y_i, X_i), where $i = 1, \ldots, m$. The response y_i is scalar. We assume independence among the responses, with common variance $\text{var}(y) = \sigma^2$. The two-dimensional signal consists of (often thousands of) digitized regressors, X_i, arranged in a $p \times \breve{p}$ array.

The indexing axes, i.e. v and $\breve{v}$, that define the support coordinates of X_i are usually on a regular grid, but the only requirement for our method is that the scatter of digitizations are common for all i. To illustrate an example of X_i, again consider the chemometric sugar processing experiment of Bro (1999) with $m = 268$ experimental units. Figure 2 presents 3997 regressors for one of these sugar units, with a corresponding scalar responses (`Ash=20, Color=10`). Indeed the number of regressors are rich, over ten times greater in number than observations. The regressor support is specified as $v^\star$ (emission) with $p = 571$ channels (275 to 560nm, by 0.5nm) and $\breve{v}^\star$ with $\breve{p} = 7$ channels (230, 240, 255, 290, 305, 325, 340nm). In fact this experiment is a practical implementation of fluorescence spectroscopy: typically there are between 5 to 20 excitation and several hundred emission channels. The excitation channels are chosen so that high energy makes the substance emit light. As Figure 2 also demonstrates, fluorescence spectra take a parallelogram shape due to a physical energy constraint between emission and excitation wavelengths. Originally discovered for fluorine, the spectrum is characteristic of the substance: thus we use it for modelling purposes in the developments that follow.

The main contribution of this article is that, rather than smoothing the signal or image themselves, we only aim for a single *smooth regression coefficient surface* for reliable prediction of the response. What is essential to our generalized linear regression approach is that the coefficient surface is smooth on the signal indexing plane. We make the assumption that such smoothness does not harm prediction. We emphasize that, despite the fact that the amount of detail in a signal can in principle be very large, the information necessary for the regression coefficient surface can be comparatively small. This can be especially true in applications presented, as the dimension of the response vector is very small relative to the number of regressors ($p\breve{p} \gg m$). Unless strong prior information is available, this implies smooth coefficients.

We use similar ingredients as Marx and Eilers (1999) and Eilers and Marx (2003), and provide an extremely practical solution for functional linear models as presented in Ramsay and Silverman (1997, Chapters 10 and 11). We use the entire two-dimensional signal as regressors for model building. The overarching coefficient surface weighs each two-dimensional signal. To regularize, we choose to impose some sensible constraints: ones that take into account the spatial structure of the regressors, while ensuring smoothness in the coefficient surface. We take two steps towards smoothness: (a) We purposely overfit the coefficient surface (not the signal) using two-dimensional tensor product B-splines, making the surface more flexible than needed. (b) We penalize estimation of the surface using difference penalties on each of the rows and columns of the tensor product B-spline coefficients.

The first step provides an initial reduction in parameter estimation through smoothness, as we will see that the tensor product B-splines are driven by relatively few parameters. The overfitting in this step is in the spirit of P-splines (Eilers and Marx, 1996) and is done to circumvent knot selection schemes. The second step ensures further smoothness, regularizing yet allowing general surfaces. We will see that two tuning parameters associated with the row and column penalties, respectively, allow for continuous control over the surface. To provide an idea, Figure 3 displays examples of (coefficient) surfaces using tensor products B-splines. The upper, left panel displays a surface constructed from essentially unpenalized tensor products, whereas the lower, right surface displays the limiting plane resulting from large second order penalties on every row and column of tensor products. The other two panels have a mixture of a low penalty on one axis and a high penalty on the other.

We term our approach presented in this article as *Multidimensional Penalized Signal Regression* (MPSR) and some of its gains include: (a) The entire signal can be used as regressors. (b) The number of highly spatially correlated regressors can far exceed the number of observations. (c) The parameterization (and the effective dimension) of the surface is dramatically reduced; the system of equations is manageable. (d) The candidate surface can be very general (non-additive), yet heavy penalization will yield polynomial surfaces. (e) Since the approach is grounded in standard (penalized) regression, delete-one diagnostics (e.g. cross-validation) are accessible. (f) The approach is easily transplanted to the generalized linear model (e.g. binary response) framework. (g) Since the two-dimensional signals and single estimated coefficient surface (and twice standard error surfaces) have a common indexing plane, potentially important regions can be visually identified.

In the next section, we provide a basic overview of tensor product B-splines. Section 3 presents the motivation and development of the penalty and its implementation to estimate the smooth coefficient surface. Some penalty modifications are briefly discussed. We aim for reliable prediction, and in Section 4 we discuss cross-validation measures that can be used to optimally tune the parameters associated with the penalty. We also discuss effective dimension of the estimated coefficient surface, estimates of variance, and the construction of twice standard error surfaces. In Section 5, the proposed methodology is applied to a sugar process data set and compared to several other popular methods. Our method is extended to the generalized linear model in Section 6. Computational details are briefly presented in Section 7. We close with a discussion.

2 Tensor Product B-splines in a Nutshell

This section shows the basic simplicity of tensor products, provides the impetus towards smooth coefficient surfaces, and defines notation. Eilers and Marx (2003, Section 4) presented *Tensor product B-splines in a nutshell.* A more thorough presentation of the subject can be found in Dierckx (1995, Chapters 1 and 2). Figure 4 displays the essential building block: a bicubic basis function, which is the *tensor product* of the two univariate (cubic) B-splines, say B and $\breve{B}$ (presented on the margins). Again the axes are v and $\breve{v}$ (e.g. emission and excitation wavelength), respectively. The tensor product has a zero (non-zero) value in the corresponding zero (non-zero) univariate B-spline support along v and $\breve{v}$. Eilers and Marx (1996, Section 2) encapsulated the details of (univariate) B-splines, including indexing. We first revisit some specifics of a univariate B-spline basis focusing only on the v axis.

We choose to divide the domain $v_{\min}$ to $v_{\max}$ into n' equal intervals, using $n'+1$ interior knots. Taking each boundary into consideration, a complete basis needs $n'+2q+1$ total knots, where q is the degree of the B-spline. Denote the knots as: $\varphi_1, \ldots, \varphi_{n'+2q+1}$. The total number of B-splines on the axis is $n = n' + q$. For indexing purposes it is convenient

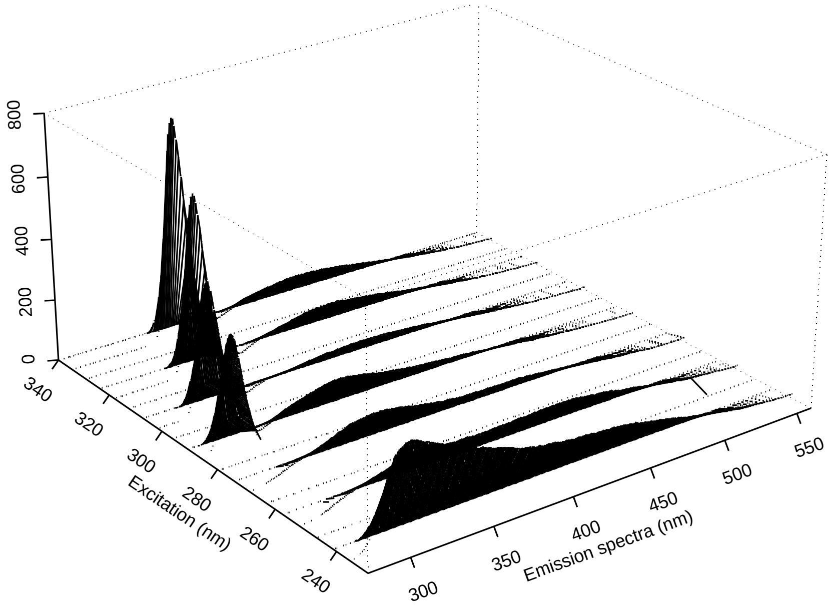

Figure 2: Two-dimensional signal regressor for one observation from sugar process data.

to associate each B-spline, $B_r(v)$ with exactly one of the (first) $r = 1, \ldots,\ n$ knots. The general properties of a B-spline of degree q include: (i) it consists of $q + 1$ polynomial pieces, each of degree q; (ii) the polynomial pieces join at q inner knots; (iii) the joining points have continuous derivatives up to degree $q-1$; (iv) a B-spline of degree q is positive on a domain spanned by $q + 2$ knots (zero elsewhere); (iv) except at the boundaries, it overlaps with $2q$ polynomial pieces of its neighbors; (v) at a given v, exactly $q+1$ B-splines are nonzero. A full basis denoted as B has dimension $m \times n$. Also using equally-spaced knots $\breve{\varphi}$, a similar division of the $\breve{v}$ axis is made to index $\breve{B}_s(\breve{v})$ for $s = 1, \ldots,\ \breve{n}$, possibly using a different $\breve{q}$.

Tensor product B-splines exist in the $v \times \breve{v}$ plane. For our presentation, n ($\breve{n}$) equally-spaced indexing knots are placed on v ($\breve{v}$) to yield a regularly-spaced grid, carving out the plane into subrectangles. The rth-sth single tensor product $B_r(v)\breve{B}_s(\breve{v})$, as presented in Figure 4, is positive in the rectangular region defined by the knots $R = [\varphi_r,\ \varphi_{r+q+2}] \times [\breve{\varphi}_s,\ \breve{\varphi}_{s+\breve{q}+2}]$ or on a support of spanned by $(q + 2) \times (\breve{q} + 2)$ knots. Similar to univariate B-splines, it is convenient to index each tensor product by one of the $n \times \breve{n}$ knot pairs and

$$\begin{aligned} B_r(v)\breve{B}_s(\breve{v}) &> 0 \text{ for all } v,\ \breve{v} \in R \\ &= 0 \text{ for all } v,\ \breve{v} \notin R, \end{aligned} \tag{1}$$

$r = 1,\ \ldots,\ n$ and $s = 1,\ \ldots\ ,\breve{n}$. Figure 5 sparsely displays nine tensor product B-splines, which represents only a portion of a full basis. A graphic of a complete basis would be difficult to appreciate, as the “hills” strongly overlap. Associated with each “hill” in Figure 5, there is an unknown coefficient. A complete tensor product B-splines basis thus has an

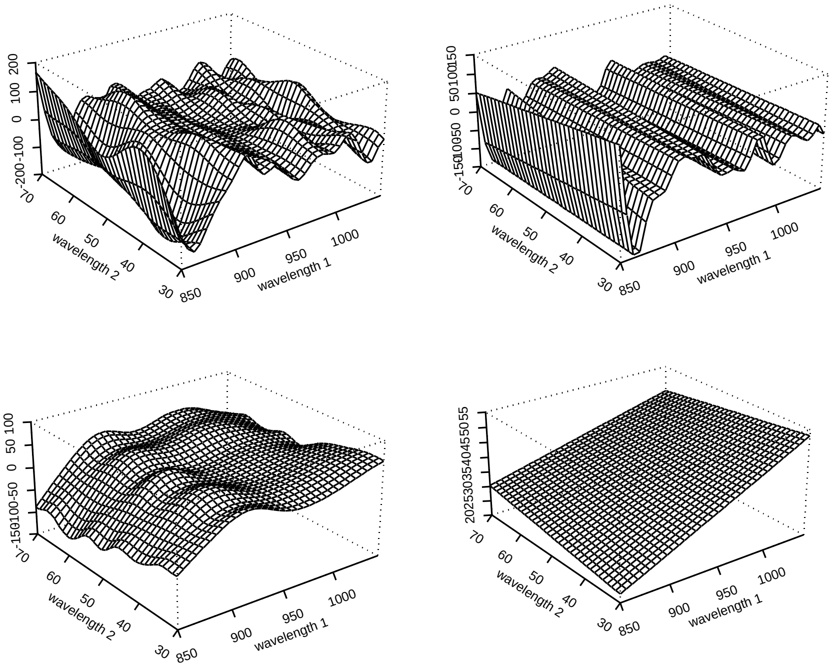

Figure 3: Examples of surfaces that can be generated from tensor products when constraining roughness in two dimensions. Effective dimension (upper, left:74); (upper, right: 34); (lower, left: 23); (lower, right: 5).

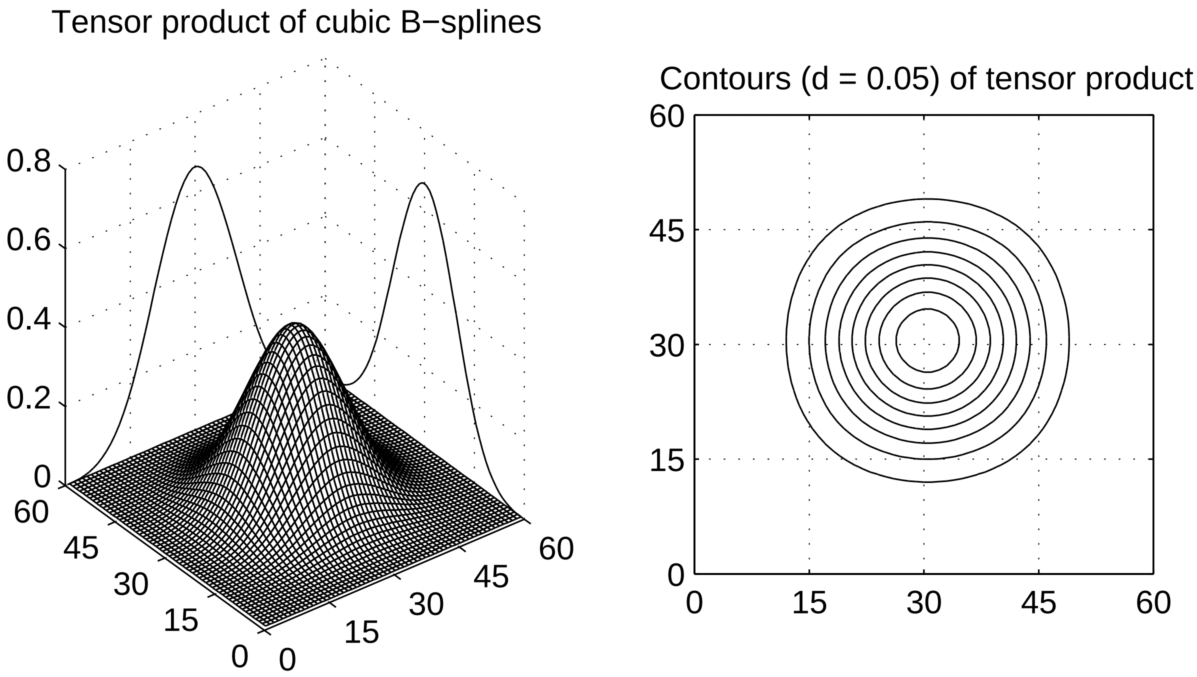

Figure 4: *Tensor product of two cubic B-splines.*

unknown coefficient matrix, denoted by $\Gamma_{n\times\breve{n}} = [\gamma_{rs}]$. For given knot grid, a very flexible surface can be approximated, e.g. at the digitized coordinates. For $j = 1, \cdots, \ p$ and $k = 1, \cdots, \ \breve{p}$,

$$\alpha(v_j^\star, \breve{v}_k^\star) = \sum_{r=1}^{n}\sum_{s=1}^{\breve{n}} B_r(v_j^\star)\breve{B}_s(\breve{v}_k^\star)\gamma_{rs}. \tag{2}$$

The surface is in fact driven by relatively few parameters ($n\breve{n}$), changing Γ changes the surface.

2.1 Unfolding Γ and notation

It is computationally efficient to reexpress the surface in a "unfolded" notation. Before doing so, some further notation is needed. Denote the support coordinate matrix $C = (v^\star \otimes \mathbf{1}_{\breve{p}},\ \mathbf{1}_p \otimes \breve{v}^\star)$ of dimension $p\breve{p} \times 2$. Let the matrix $\mathbf{B}_1^\star$ and $\mathbf{B}_2^\star$ (with respective dimensions $p\breve{p} \times n$ and of $p\breve{p} \times \breve{n}$) be the univariate B-spline basis matrix evaluated at each entry of the first and second column of C, respectively. The unfolded expression at the support coordinates then has the standard multiple regression form

$$\text{vec}\{\alpha(v^\star, \breve{v}^\star)\} = \mathbf{T}^\star\gamma, \tag{3}$$

where $\gamma = \text{vec}(\Gamma)$. Define the matrix

$$\mathbf{T}^\star = (\mathbf{B}_1^\star \otimes \mathbf{1}_n') \odot (\mathbf{1}_{\breve{n}}' \otimes \mathbf{B}_2^\star) \tag{4}$$

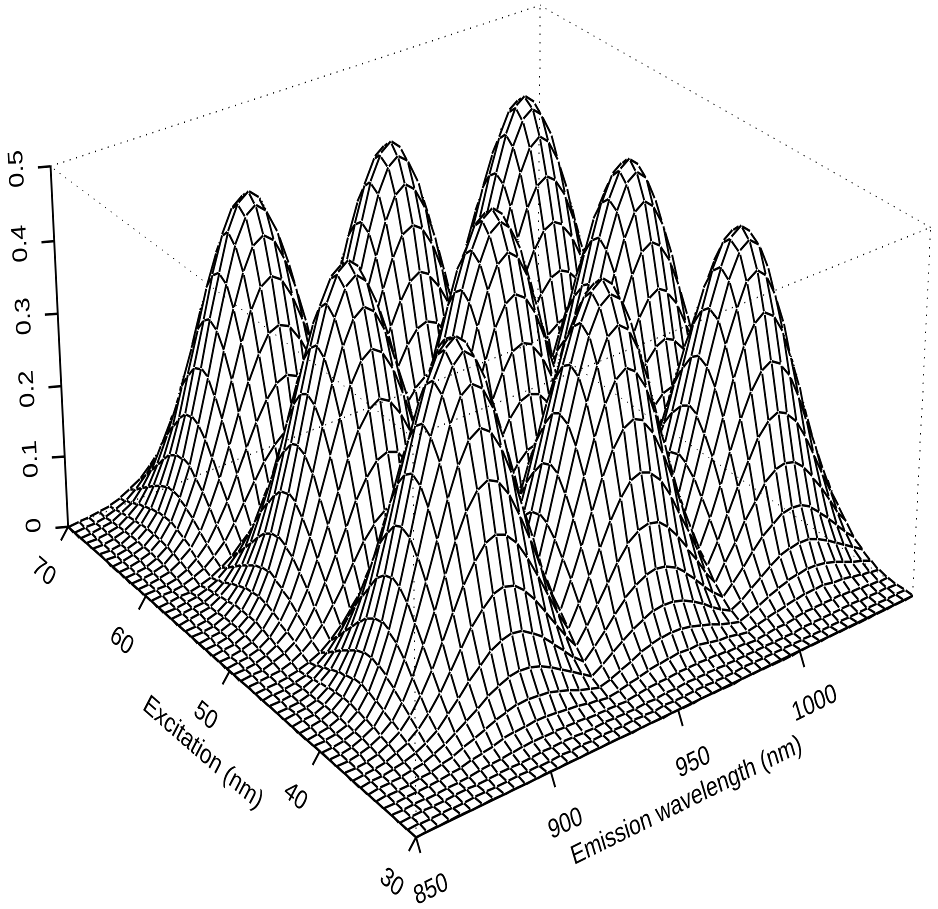

Figure 5: *Landscape of nine cubic B-spline tensor products, a portion of a full basis*

of dimension $p\breve{p} \times n\breve{n}$. The symbols $\otimes$ and $\odot$ denote Kronecker product and elementwise multiplication of matrices, respectively. Penalized estimation of γ and its use with two-dimensional signal regressors are topics discussed in the next section.

3 Penalized Two-Dimensional Coefficient Surfaces

Given the ith regressor matrix $X_i = [x_{ijk}]$ of dimension $p \times \breve{p}$, signal regressor support matrix C, and coefficient surface $\alpha(v, \breve{v})$, express the mean

$$\mu_i = \sum_{j=1}^{p} \sum_{k=1}^{\breve{p}} x_{ijk} \alpha(v_j^\star, \breve{v}_k^\star), \tag{5}$$

where $i = 1, \ldots,\ m;\ j = 1,\ \ldots,\ p;\ k = 1, \ldots,\ \breve{p}$. Using tensor product B-splines, (2) can be substituted into (5) yielding

$$\mu_i = \sum_{j=1}^{p}\sum_{k=1}^{\breve{p}} x_{ijk} \sum_{r=1}^{n}\sum_{s=1}^{\breve{n}} B_r(v_j^\star)\breve{B}_s(\breve{v}_k^\star)\gamma_{rs} = \mathbf{x}_i'\mathbf{T}^\star\gamma, \tag{6}$$

where $\mathbf{x}_i' = \text{vec}(X_i)$. We aim to find a practical solution to minimize

$$Q(\gamma) = |y - \mathbf{X}\mathbf{T}^\star\gamma|^2 = |y - \mathbf{M}\gamma|^2,$$

where $\mathbf{X}$ is the $m \times p\breve{p}$ matrix of vectorized signals and $\mathbf{M} = \mathbf{X}\mathbf{T}^\star$. The use of tensor product B-splines does reduce the dimension of estimation, but there are still $n\breve{n}$ unknown parameters. For even moderately complex surfaces, ill-posed estimation problems can arise, as it may be necessary to increase the number on knots on the grid to allow enough flexibility and/or there may be regions without signal data leading to portions of $\mathbf{X}^\star$ with zeros.

3.1 Implementing the penalty

In the spirit of P-splines, discrete roughness or difference penalties are imposed on γ. Although we implement penalization on the vector form of coefficients γ, the motivation and mechanics of penalization is perhaps best seen through the matrix of coefficients Γ. In fact a separate difference penalty is assigned to each of its rows and each of its columns. The penalties have structure to effectively break the linkage in the penalty from row to row or from column to column. The objective function is now modified, using penalties, to minimize

$$\begin{aligned} Q_P(\gamma) &= \text{Residual SS} + \text{Row Penalty} + \text{Column Penalty} \\ &= \sum_{i=1}^{m}(y_i - \mathbf{x}_i'\mathbf{T}^\star\gamma)^2 + \lambda_1 \sum_{r=1}^{n} \gamma_{r\bullet} D_d' D_d \gamma_{r\bullet}' + \lambda_2 \sum_{s=1}^{\breve{n}} \gamma_{\bullet s}' D_{\breve{d}}' D_{\breve{d}} \gamma_{\bullet s} \\ &= |y - \mathbf{M}\gamma|^2 + \lambda_1 |P_1\gamma|^2 + \lambda_2 |P_2\gamma|^2. \end{aligned} \tag{7}$$

The penalty has two parts: the first (second) puts a difference penalty on the rows (columns) of Γ, where $\gamma_{r\bullet}$ ($\gamma_{\bullet s}$) denotes the rth row (the sth column) of Γ. The penalties are compactly represented using Kronecker products and matrix notation: $P_1 = (D_d' D_d) \otimes I_{\breve{n}}$ and $P_2 = I_n \otimes (D_{\breve{d}}' D_{\breve{d}})$, where I denotes the identity matrix and d denotes the order of the difference penalty. The dimensions of P_1 and P_2 are fixed with $\breve{n}(n - d) \times n\breve{n}$ and $n(\breve{n} - \breve{d}) \times n\breve{n}$, respectively. The order of the penalties (d, $\breve{d}$) are in principle additional hyper-parameters, but in practice are usually fixed by the user. Some guidelines are given below in Section 4.2. The non-negative λs essentially provide continuous control over smoothness. We see from (7) that there are two separate λs to weigh the penalty, one associated with rows and one associated with columns of Γ. Yet the influence or weight of the penalty is the same for each row, the same for each column, but are allowed to differ from rows to columns. Figure 6 displays a possible scenario resulting from strong row (top panel) and strong column (bottom panel) penalization using a second order penalty on each row and column with large λ_1 and λ_2. Notice that the limiting behavior for each row

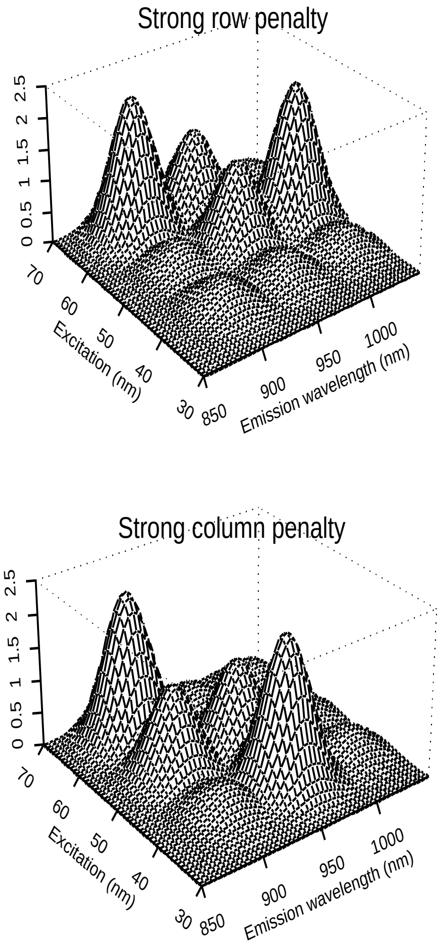

Figure 6: *Nine cubic B-spline tensor products, with a strong linear row penalty (upper panel) and a strong linear column penalty (lower panel)*

and column is linear, but reversals of slopes are possible from one row (or column) to the next.

An example of a first and second order penalty matrix D for small $n = \breve{n} = 3$ (for one row or column of Γ) looks like

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

Each row (column) of Γ gets such a banded D matrix. Hence the complete column penalty has the contrast structure

$$P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes D_1 = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{bmatrix},$$

whereas the corresponding complete row penalty takes the form

$$P_2 = D_1 \otimes \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ & & & & & & & & \\ 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 \end{bmatrix}.$$

Of course the order of the penalty can vary from P_1 to P_2.

The explicit P-spline solution for (7) is

$$\hat{\gamma} = (\mathbf{M}'\mathbf{M} + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2)^{-1} \mathbf{M}' y. \tag{8}$$

We see from (8) that the system of equations remains $n\breve{n}$ even as the resolution of the two-dimensional signal, $p\breve{p}$, increase dramatically. The predicted values are $\hat{y} = \mathbf{M}\hat{\gamma}$. The effective "hat" matrix is

$$H = [h_{ii'}] = \mathbf{M}(\mathbf{M}'\mathbf{M} + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2)^{-1} \mathbf{M}'.$$

3.2 Modifications for intercept term or additional penalization

The model can be modified to include other factors or covariates, and the objective (7) could have additional (polynomial) penalties. In the examples to follow, we include an intercept term α_0, and an additional overall ridge penalty $\lambda_0|\gamma|^2$, $\lambda_0 > 0$. Thus the modified P-spline solution becomes

$$(\hat{\alpha}_0, \hat{\gamma}')' = (\tilde{\mathbf{M}}'\tilde{\mathbf{M}} + \lambda_1 \tilde{P}_1' \tilde{P}_1 + \lambda_2 \tilde{P}_2' \tilde{P}_2 + \lambda_0 \tilde{I})^{-1} \tilde{\mathbf{M}}' y,$$

with $\tilde{\mathbf{M}} = (1_m|\mathbf{M})$, $\tilde{P} = (0|P)$, and $\tilde{I} = \text{diag}(0, I_{n\breve{n}})$. The zero vector in $\tilde{P}$ and $\tilde{I}$ ensures an unpenalized intercept.

4 Optimization of the penalty

We aim for reliable prediction. The P-spline coefficient surface model is driven by the non-negative penalty regularization parameters $(\lambda_1, \lambda_2, \lambda_0)$, thus we have continuous control over smoothness. Cross-validation can be an effective, prediction oriented, approach to specify the penalty parameters. Since we use penalized least squares, we highlight the simplicity of delete-one cross-validation. We minimize:

$$\mathrm{CV}(\lambda_1, \lambda_2, \lambda_0) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_{-i})^2},$$

where $\hat{y}_{-i}$ is the predicted value for ith data point using a model trained without the ith data point. For the model using all m observations, delete-one cross-validation standard error of prediction can be quickly calculated:

$$\mathrm{CV}(\lambda_1, \lambda_2, \lambda_0) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}\right)^2}, \tag{9}$$

since both $\hat{y}$ and the diagonals of H are easily obtained. See Myers (1990) for a nice presentation of delete-one CV. The optimal $(\lambda_1, \lambda_2, \lambda_0)$ can be found, e.g. using a grid search to minimize CV. Possibly there are other penalty terms and their corresponding parameters which would lead to a higher dimensional grid search. Other optimization approaches can be taken, such as 10-fold cross-validation or validation through an independent data set.

4.1 Effective dimension, estimated variance, standard error surfaces

The effective dimension of the estimated coefficient surface can be approximated by

$$\text{eff dim}(\hat{\alpha}) = \mathrm{trace}(H),$$

(Hastie and Tibshirani, 1990) and the error variance component estimated by

$$\hat{\sigma}^2 = \frac{|y - \hat{y}|^2}{m - \mathrm{trace}(H)},$$

where $m - \mathrm{trace}(H)$ approximates the residual degrees of freedom. Using a permuted form of H, $\mathrm{trace}(H) = \mathrm{trace}\{\mathbf{M}'\mathbf{M}(\mathbf{M}'\mathbf{M} + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2 + \lambda_0 I)^{-1}\}$ is computed more efficiently. For given λ_1 and λ_2, the sandwich estimator for the estimated P-spline coefficients is expressed as:

$$\mathrm{var}(\hat{\gamma}) \approx \hat{\sigma}^2(\mathbf{M}'\mathbf{M} + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2 + \lambda_0 I)^{-1}\mathbf{M}'\mathbf{M}(\mathbf{M}'\mathbf{M} + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2 + \lambda_0 I)^{-1}.$$

A translation back to the variance of the estimated coefficient surface, $\hat{\alpha} = \mathbf{T}^{\star}\hat{\gamma}$, results in

$$\mathrm{var}(\hat{\alpha}) = \mathbf{T}^{\star}\mathrm{var}(\hat{\gamma})\mathbf{T}^{\star\prime}. \tag{10}$$

Twice standard error surfaces can be constructed using ±2 times the square root of diagonals of $\mathrm{var}(\hat{\alpha})$.

4.2 MPSR recipe

We suggest the following MPSR recipe:

1. Place a generous number of tensor product B-splines on a rectangular grid on the $v \times \breve{v}$ plane;

2. For computational efficiency attempt to keep $n\breve{n} < 1000$;

3. Use row and column penalty order of $d = 2$ or 3;

4. Measure performance using cross-validation or an information criterion;

5. Vary $(\lambda_1, \lambda_2, \lambda_0)$ on a logarithmic grid;

6. Find minimum or maximum of performance criterion;

7. Report $\hat{\gamma}$ as the compact form of the estimated coefficient surface.

5 Example: sugar process data

We analyze the $m = 268$ observation, $p\breve{p} = 3997$ variable, two response variable (Ash content, Color) sugar fluorescence data set taken from Bro (1999). Sugar was sampled continuously during 8 hours to make a mean sample representative for one "shift". Samples were taken during 3 months of operation in the late autumn from a sugar plant in Scandinavia. The sugar was sampled directly from the final unit operation of the process. The data and a detailed description can be found at www.models.kvl.dk.

Figure 2 (from Section 1) presented the structure of the regressors for the first observation of the sugar fluorescence data set. Recall that the two-dimensional signal consists of $571 \times 7 = 3997$ regressors, with support: $v^\star$ (emission) with $p = 571$ channels (275 to 560nm, by 0.5nm) and $\breve{v}^\star$ with $\breve{p} = 7$ channels (230, 240, 255, 290, 305, 325, 340nm). In all analyses to follow the $m \times p\breve{p}$ regressor matrix was mean centered, but not scaled. All MPSR models presented included an intercept term, and used $n = 10$ (excitation), $\breve{n} = 25$ (emission).

The response (quality parameter) considered is Ash content (determined by conductivity and given in units of percentage), which measures the amount of inorganic impurities in the refined sugar. The refined sugar was dissolved in non-buffered water (2.25 g/15 ml) and the solution was measured in a 10×10mm curvette on a PE LS50B spectrofluorometer.

A prediction performance study for these data was conducted for the MPSR method, so that the results are directly comparable to the previous study of Eriksson et al. (2000, Their Table 2). These authors looked at the prediction performance of a variety of methods, including (combinations of): partial least squares (PLS), orthogonal signal correction (OSC), multiplicative signal correction (MSC), standard normal variate transform (SNV), wavelet models (Daubechies-4, Symmlet-8, Coiflet-2; wavelet or discrete packet transform). All wavelet methods also used PLS and OSC1. Like these authors, we first sorted the observations in descending order according to the Ash (to suppress the extent of autocorrelation among samples). Also, two subsets comprising all the odd-numbered and all the even-numbered observations were created. Training models were then established on the odd-numbered observations, followed by performing external predictions for the even-numbered samples, and visa versa. Thus our reported prediction performance refers to external prediction.

Table 1. MPSR model summary of external RMSEP for various penalty orders. Caption:(o) training set (prediction set) used odd (even) numbered observations, (e) training set (prediction set) used even (odd) numbered observations.

Penalty order	RMSEP (o)	RMSEP (e)	(λ_1, λ_2) (o)	(λ_1, λ_2) (e)
$d, \breve{d} = 4$	1.84	1.68	(177, 1e7)	(10, 5.62e3)
$d, \breve{d} = 3$	1.83	1.66	(1.77e3, 1e4)	(3.16, 1e4)
$d, \breve{d} = 3$+ridge	1.82	1.62	(1e3, 5.6e3, 177)	(1e-4, 5.6e3, 177)
$d, \breve{d} = 2$	1.83	1.65	(1e-8, 1e5)	(3.16, 1e4)
$d, \breve{d} = 1$	1.84	1.68	(177, 3.16e4)	(177, 3.16e3)

Table 1 presents the root mean square error of prediction (RMSEP) for the external prediction set, using optimal MPSR models with several orders of the penalty. Table 2 summarizes comparable RMSEP values for various models used by Eriksson et al. Effective dimensions of the MPSR estimated coefficient surfaces ranged from (29, 38) (o) and (45, 52) (e). We find MPSR to be a competitor. For the even (e) training model/ odd (o) validation, MPSR's RMSEP is lower than values reported for (PLS, MSC, SNV), and is within less than 5% of the values reported for all the wavelet methods and OSC1. The lowest MPSR RMSEP=1.62 corresponds to the model with $d = \breve{d} = 3$ and a ridge penalty, yet all other penalty orders found similar values (approximately within 3%). For the odd (o) training model/ even (e) validation, RMSEP values were higher for all methods. MPSR's RMSEP was about 1.83 and still competitive, performing better than (MSC, SNV), but approximately 7% higher when compared to any wavelet method. Interestingly, PLS and OSC1 were the best performers with a RMSEP value of 1.66 in this case. To give an idea of the range of `Ash` and MPSR performance, Figure 7 displays the scatterplots of the observed vs. predicted `Ash` content for both the (o) and (e) optimally trained MPSR models.

Perhaps a more honest way of reporting external prediction performance is to first optimize the training model based on CV, e.g. delete-one CV or some other measure – then given this optimally trained model, measure the external prediction performance. For the MPSR models, this would translate to first choosing the training $(\lambda_1, \lambda_2, \lambda_0)$ based on minimizing delete-one (training) CV. For the mentioned wavelet methods, one could do delete-one (training) CV for all combinations of PLS order, OSC order, coefficient extraction, and so on. One of the beauties of our MPSR method is that delete-one CV can be computed extremely swiftly, as outlined in Section 4. Table 3 reports external RMSEP values for MPSR models with various orders, but first optimizing the λs using (9).

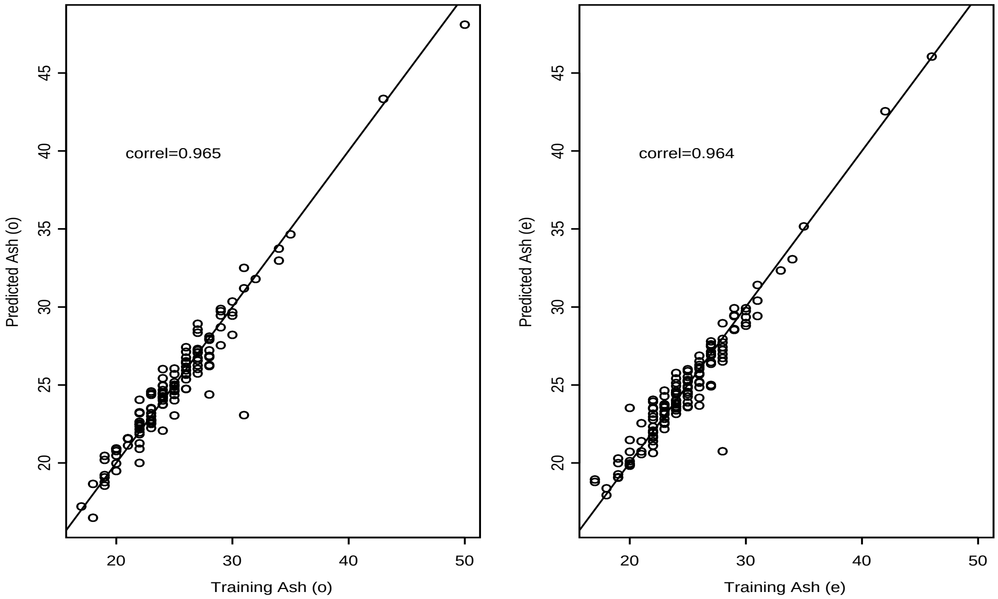

Figure 7: *Scatterplots of observed vs. predicted* `Ash` *content, for both odd and even trained MPSR (optimal) models with* $d = \breve{d} = 3$.

*Table 2. Summary of RMSEP for several other models (*Eriksson et al.*, 2000) . Numbers directly comparable to Table 1 above. Caption:(o) training set (prediction set) used odd (even) numbered observations, (e) training set (prediction set) used even (odd) numbered observations.*

Method	RMSEP (o)	RMSEP (e)	PLS comp.	Coeffs
PLS	1.66	1.99	3	-
OCS1	1.66	1.63	2	-
MSC	2.05	2.10	6	-
SNV	2.03	2.07	6	-
OSC1 Dau4 DWT	1.69	1.60	2	189/183
OSC1 Dau4 WPT	1.69	1.60	2	246/228
OSC1 Sym8 DWT	1.69	1.60	2	167/173
OSC1 Sym8 WPT	1.69	1.60	2	197/218
OSC1 Coi2 DWT	1.69	1.60	2	164/170
OSC1 Coi2 WPT	1.69	1.60	2	200/223
Dau4 DWT	1.66	1.99	3	115/116

Table 3. MPSR model summary of external RMSEP for various penalty orders. The optimal penalty weights are also given and determined by minimizing delete-one CV of the training set. Caption:(o) training set (prediction set) used odd (even) numbered observa-

tions, (e) training set (prediction set) used even (odd) numbered observations.

Penalty order	RMSEP (o)	RMSEP (e)	(λ_1, λ_2) (o)	(λ_1, λ_2) (e)
$d, \breve{d} = 4$	1.98	1.75	(1e6, 1.77e4)	(1e-7, 1e7)
$d, \breve{d} = 3$	1.92	1.78	(1e7, 5.62e5)	(1e7, 562)
$d, \breve{d} = 3$+ridge	1.84	1.99	(1e-7, 5.6e4, 1e4)	(3.2e5, 1e-8, 3.2e5)
$d, \breve{d} = 2$	1.92	2.02	(5.62e4, 5.62e4)	(1.77e5, 5.62e6)
$d, \breve{d} = 1$	1.88	1.98	(5.62e3 , 5.62e4)	(1.78e4, 1e6)

A nice feature of MPSR is that the spatial information of the regressors is built into the model. Thus one can actually see the relative magnitudes for various regions of the coefficient indexing plane. Figure 8 displays one of the optimal coefficient surface for even training ($d = \breve{d} = 3$) and the response `Ash`. One can clearly see its interactive features. This MPSR model used 250 cubic tensor products, with $n = 25$ along emission and $\breve{n} = 10$ along excitation, and the optimal penalty parameters were $\lambda_1 = 3.16$ and $\lambda_2 = 10^4$ ($d = \breve{d} = 3$), reducing the effective dimension of the estimated surface to approximately 48.1. If desired, approximate twice-standard error surfaces can also be constructed using (10). As a tool to help identify important regions of the image for prediction, the optimal surface can be scaled by inverse standard error to provide a t-like surface, as presented in Figure 9.

6 Extensions to the generalized linear model

Since the coefficient surface model is grounded in (penalized) least squares regression techniques, the methodology can be easily transplanted into the generalized linear model (GLM). A penalized scoring algorithm is used, e.g. with binomial or Poisson responses. The mean response is now modelled using through a (monotone) inverse link function h,

$$\mu = h(\mathbf{X}\mathbf{T}^{\star}\gamma) = h(\mathbf{M}\gamma) = \eta.$$

The objective is now to maximize log-likelihood function, $l(\gamma)$, subject to penalization

$$l^{\star}(\gamma) = l(\gamma) - \lambda_1|P_1\gamma|^2 - \lambda_2|P_2\gamma|^2.$$

This results in the iterative solution

$$\hat{\gamma}_{t+1} = (\mathbf{M}'\hat{W}_t\mathbf{M} + \lambda_1 P_1' P_1 + \lambda_2 P_2' P_2)^{-1}\mathbf{M}'\hat{W}_t\hat{z}_t,$$

where the weight matrix $W = \text{diag}\{h'(\eta)\}/\text{var}(Y)$, and the working vector is $z = (y - \mu)/h(\eta) + \eta$. All of the specifics of the GLM can be found in McCullagh and Nelder (1989).

In choosing optimal penalty parameters for GLMs, it is often easier to minimize an information criterion (e.g. Aikaike's AIC) than it is to use cross validation measures. For example, we could seek an optimal (λ_1, λ_2) by minimizing

$$\text{AIC} = \text{deviance}(y; \hat{\gamma}) + 2\text{trace}(H),$$

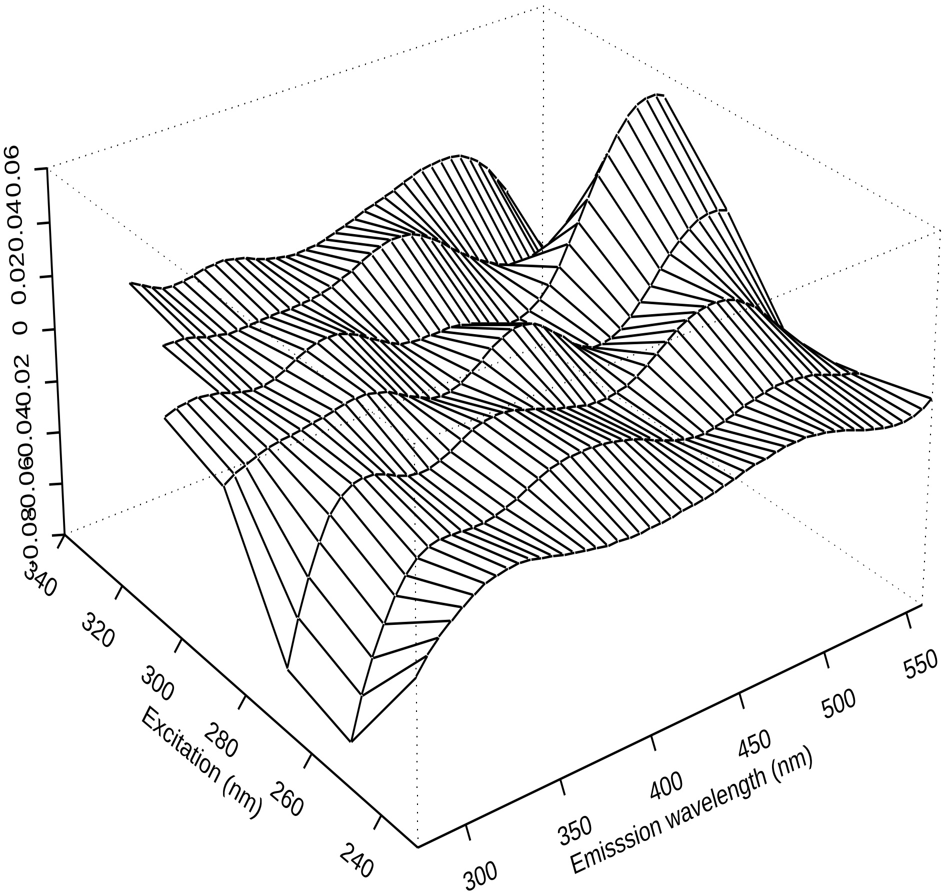

Figure 8: *Optimal estimated coefficient surface (e) for sugar data with eff dim=48.1 and* $d = \breve{d} = 3$.

a simple function of goodness-of-fit and effective dimension. Upon convergence

$$\text{trace}(H) = \text{trace}\{\mathbf{M}'\hat{W}\mathbf{M}(\mathbf{M}'\hat{W}\mathbf{M} + \lambda_1 P_1'P_1 + \lambda_2 P_2'P_2)^{-1}\}.$$

Other criteria can also be used. For example, in logistic regression maximizing the percent correct classification of a validation data set can be a useful approach to determine optimal penalty parameters. Similar to the presentation in Section 3.2, penalties or an intercept term can also be added in this generalized MPSR setting.

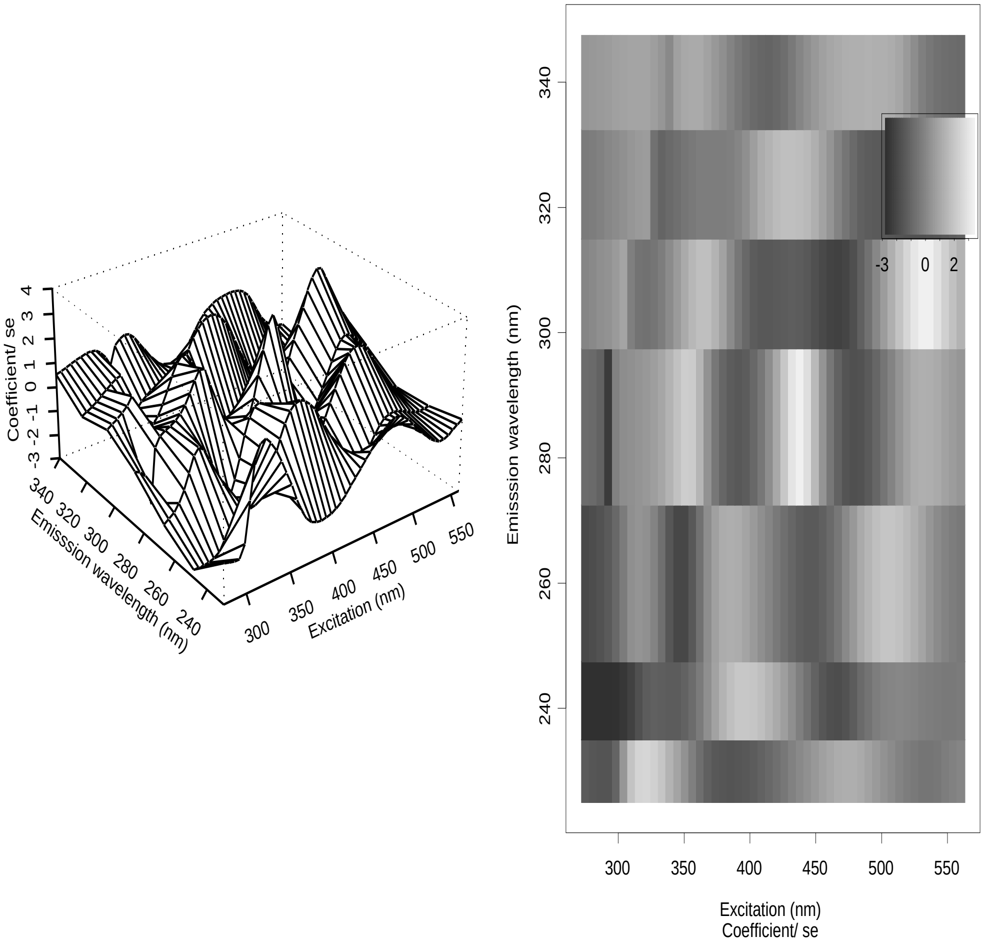

Figure 9: *Surface of t-like statistics, i.e. the above optimal estimated coefficient, divided by its standard error.*

7 Computational details and software

For computational purposes, an explicit solution for (7) can be found efficiently through data augmentation, i.e. $\hat{\gamma} = (\mathbf{M}_+'\mathbf{M}_+)^{-1}\mathbf{M}_+' y_+$, where

$$\mathbf{M}_+ = \begin{bmatrix} \mathbf{M} \\ \sqrt{\lambda_1}(D_d \otimes I_{\breve{n}}) \\ \sqrt{\lambda_2}(I_n \otimes D_d \\ \sqrt{\lambda_0} I_{n\breve{n}}) \end{bmatrix} \qquad \text{and} \qquad y_+ = \begin{bmatrix} y \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Additional column augmentation can be made to incorporate an intercept term and additional row augmentation could implement another penalty. S-PLUS functions are available at www.stat.lsu.edu/bmarx which use this data augmentation as outlined above. Among other arguments, these functions allow for GLM fitting and twice standard error surfaces in plotting. The two-dimensional signals can be input as stacks of matrices or unfolded in vector form.

8 Discussion

We considered the ill-posed regression problem with a scalar response and rich regressors in the form of a multidimensional signal. Rather than performing variable selection or using the signal at one excitation or additive signal regression, we use the entire two-dimensional signal as regressors while taking advantage of its indexing information. What is essential to our approach is that the coefficient surface (not the signal regressors) can be assumed to be smooth without doing much harm to prediction. Our approach proposed a generous number of (penalized) tensor product B-splines to estimate a smooth coefficient surface for the regressors. This was the first step towards reducing the dimension of parameter estimation. Two penalties, one on the rows and one on the columns, ensured further smoothness. The amount of regularization was determined by prediction oriented criteria.

We emphasize that the goal is not to smooth the image themselves, which might need a (very) large number of tensor products. We only aim at smooth (multi-dimensional) regression coefficients. The amount of detail in the images can in principle be very large, while the information for a regression model is comparatively small, only the relatively small dimension of the response vector y. The effective dimension of the coefficient surface (hyper-volume) has to be (much) smaller than the number of observations. Unless strong prior information is available, this implies smooth coefficients. We realize that smoothness of the coefficients surface is a only one type of constraint, but we find it an extremely practical one. The two-dimensional signal regressors need not be smooth to impose such a constraint; see the points of Marx and Eilers (1999, Section 3.3) with one-dimensional signal regressors.

We summarize several strengths to our approach: a) The number of regressors $p\breve{p}$ can far exceed the number of observations m. (b) The estimated coefficient surface can be very general (non-additive or polynomial), yet with often modest effective dimension. (c) Since the approach is grounded in standard (penalized) regression, the approach is easily transplanted to the generalized linear model. (d) Customized designer penalization is possible.

We stress that our MPSR approach, is not only a competitor, but has some clear advantages (relative to PLS, OSC, MSC, SNV, wavelet methods, or combinations thereof). Unlike other methods, MPSR takes full advantage of the natural spatial information of the signals. MPSR is simple to use: it uses the entire ("raw") signal and works "right out of the box" without needing PLS/ OSC preprocessing, padding, or extraction of wavelet coefficients. The method is intuitive in that you can actually see what is going relative to the spatial indexing plane, i.e. how the coefficient surface is used to contrast the two-dimensional signals for prediction. Further twice standard error surfaces serve to help identify potentially important signal regions. Moreover by virtue of easy computations of delete-one CV for MPSR, optimizing the training model can easily be done– allowing then for a fair assessment of an external validation. Calculations are fast, especially with these models that used for example a modest $n = 10$, $\breve{n} = 25$.

We emphasize that, as presented, tensor products of B-splines are very well suited to multidimensional signal or image regression. Kernel smoothers or local likelihood cannot be used, because they have no parameters to characterize the coefficient surface. In principle thin-plate splines (TPS) could be used, but they have the problem of too many parameters. The fluorescence data matrix has 7 times 571 elements. The number of knots of the TPS would be the same, leading to very large systems of equations. A way out could be to use thin-plate regression splines (Wood, 2003). Wood proposes to approximate a TPS by a limited number of basis functions. Instead of taking such a detour, it seems more effective to start directly with simple basis functions. We also note our approach proposes a more general penalization, as both variants of TPS impose an isotropic penalty, which means that the amount of smoothing is the same for both dimensions.

In our application, the physical background of the measurements dictated the excitation and emission axes of fluorescence, and it was natural to allow different amounts of smoothing. When working with real images, the axes might be relatively arbitrary, only determined by the orientation of the imaging instrument. Then it might be natural to have an isotropic penalty, forcing both penalty parameters to be equal. Unfortunately, isotropic doe not imply rotation invariant: rotation of the axes might lead to different estimates of the coefficient surface. We note that it is possible to extend the penalties with mixed differences. Presently we use penalties that either work exclusively on the rows (columns) of the matrix of coefficients of tensor products. A mixed penalty would work on both rows and columns.

Another attraction of tensor products of B-splines is the straightforward extension to three or more dimensions. (Eilers, Currie and Durbàn, 2004). The same holds for the discrete roughness penalties (of any order). In contrast, "thin-(hyper)-volume regression" splines with non-isotropic roughness penalties look very challenging.

Although we did not pursue the issue of scattered data, our model can be extended to less-regular observation schemes. The sugar spectra all had the same domain, a rectangle with a missing corner. We can envision applications in which the X data are scattered, with a different scattering pattern for each observation. An example would be estimation of averages or totals from scattered spatial samples.

A penalty is especially important in cases where the signals have sharp boundaries with zero regions, like in these excitation/emission energy constraints. Without the penalty, there is yet another layer of ill-conditioning due to the zeros in the unsupported "triangle". In fact the penalty remedies the problem and effectively extrapolates the unsupported region. Further, we found that our proposed methodology does not just attempt to fulfill minimum regularization requirements, rather it can easily accommodate a natural and dramatic growth in the numbers of regressors, for example as future spectroscopy instrumentation improves resolution. In short, the system of equations remains $n\breve{n}$, even as $p\breve{p}$ becomes very large.

The methods developed can be used beyond fluorescence spectra; they can also be used for medical/grey-scale, or other images. The mathematics allow for higher dimensions, e.g. images in time. Future research could also model image regressors while controlling for other (smooth) covariates, factors, signals, or varying coefficient terms, along the lines of Eilers and Marx (2002).

Acknowledgement

Research supported in part for Brian Marx by NSF Grant DMS-0102131.

References

Bro, R. (1999). Exploratory study of sugar production using fluorescence spectroscopy and multiway analysis. *Chemometrics and Intelligent Laboratory Systems*, **46**, 133-147.

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

Eilers, P.H.C, Currie, I., and Durbàn, M. (2004). Low memory, high speed smoothing on large multidimensional grids. *Computational Statistics and Data Analysis.* To appear.

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

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

Eriksson, L., Trygg, J., Johansson, E., Bro, R., and Wold, S. (2000). Orthogonal signal correction, wavelet analysis, and multivariate calibration of complicated fluorescence data. *Analytica Chimica Acta*, **420**, 181-195.

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

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

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models (2nd edition).* Chapman and Hall, London.

Myers, R.H. (1990). *Classical and Modern Regression with Applications, 2nd ed.*, Duxbury Press, Boston.

Ramsay, J.O. and Silverman, B.W. (1997). *Functional Data Analysis.* Springer, New York.

Wood, S.N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society* B **65**(1), 95-114.