SORT 39 (2) July-December 2015, 149-186

Twenty years of P-splines

Paul H.C. Eilers[1], Brian D. Marx[2] and Maria Durbán[3]

Abstract

P-splines first appeared in the limelight twenty years ago. Since then they have become popular in applications and in theoretical work. The combination of a rich B-spline basis and a simple difference penalty lends itself well to a variety of generalizations, because it is based on regression. In effect, P-splines allow the building of a "backbone" for the "mixing and matching" of a variety of additive smooth structure components, while inviting all sorts of extensions: varying-coefficient effects, signal (functional) regressors, two-dimensional surfaces, non-normal responses, quantile (expectile) modelling, among others. Strong connections with mixed models and Bayesian analysis have been established. We give an overview of many of the central developments during the first two decades of P-splines.

MSC: 41A15, 41A63, 62G05, 62G07, 62J07, 62J12.

Keywords: B-splines, penalty, additive model, mixed model, multidimensional smoothing.

1. Introduction

Twenty years ago, *Statistical Science* published a discussion paper under the title "Flexible smoothing with B-splines and penalties" (Eilers and Marx, 1996). The authors were two statisticians with only a short track record, who finally got a manuscript published that had been rejected by three other journals. They had been trying since 1992 to sell their brainchild P-splines (Eilers and Marx, 1992). Apparently it did have some value, because two decades later the paper has been cited over a thousand times (according to the *Web of Science*, a conservative source), in both theoretical and applied work. By now, P-splines have become an active area of research, so it will be useful, and hopefully interesting, to look back and to sketch what might be ahead.

[1] Erasmus University Medical Centre, Rotterdam, the Netherlands, p.eilers@erasmusmc.nl

[2] Dept. of Experimental Statistics, Louisiana State University, USA, bmarx@lsu.edu

[3] Univ. Carlos III Madrid, Dept of Statistics, Leganés, Spain, mdurban@est-econ.uc3m.es
Received: October 2015

P-splines simplify the work of O'Sullivan (1986). He noticed that if we model a function as a sum of B-splines, the familiar measure of roughness, the integrated squared second derivative, can be expressed as a quadratic function of the coefficients. P-splines go one step further: they use equally-spaced B-splines and discard the derivative completely. Roughness is expressed as the sum of squares of differences of coefficients. Differences are extremely easy to compute and generalization to higher orders is straightforward.

The plan of the paper is as follows. In Section 2 we start with a description of basic P-splines, the combination of a B-spline basis and a penalty on differences of coefficients. The penalty is the essential part, and in Section 3 we present many penalty variations to enforce desired properties of fitted curves. The penalty is tuned by a smoothing parameter; it is attractive to have automatic and data-driven methods to set it. Section 4 presents model diagnostics that can be used for this purpose, emphasizing the important role of the effective model dimension. We present the basics of P-splines in the context of penalized least squares and errors with a normal distribution. For smoothing with non-normal distributions, it is straight-forward to adapt ideas from generalized linear models, as is done in Section 5. There we also lay connections to GAMLSS (generalized additive models for location, scale and shape), where not only the means of conditional distributions are modelled. We will see that P-splines are also attractive for quantile and expectile smoothing. The first step towards multiple dimensions is the generalized additive model (Section 6). Not only can smoothing be used to estimate trends in expected values (and other statistics), but it also can be used to find smooth estimates for regression coefficients that change with time or another additional variable. The prototypical case is the varying-coefficient model (VCM). We discuss the VCM in Section 7, along with other models like signal regression. In modern jargon these are examples of functional data analysis. In Section 8, we take the step to full multidimensional smoothing, using tensor products of B-splines and multiple penalties. In Section 9, we show how all the models from the previous sections can be added to each other and so combined into one structure. Here again the roots in regression pay off.

One can appreciate the penalty as just a powerful tool. Yet it is possible to give it a deeper meaning. In Section 10, P-splines are connected to mixed models. This leads to further insights, as well as to new algorithms for finding reasonable values for the penalty parameters. From the mixed model perspective, it is just a small step to a Bayesian approach, interpreting the penalty as (minus) the logarithm of the prior distribution of the B-spline coefficients. This is the subject of Section 11.

Asymptotics and boosting do not have a natural place in other sections, so we put them together in Section 12, while computational issues and availability of software are discussed in Section 13. We close the paper with a discussion

As far as we know, this is the first review on P-splines. Earlier work by Ruppert et al. (2009) took a broader perspective, on the first five years after appearance of their book (Ruppert et al., 2003). We do not try to be exhaustive. That would be impossible (and boring), given the large number of citations. With the availability of *Google Scholar*

and commercial citation databases such as *Scopus* and the *Web of Science*, anyone can follow the trail through history in detail.

We have done our best, in good faith, to give an overview of the field, but we do not claim that our choice of papers is free from subjectivity. The advent of P-splines has led to formidable developments in smoothing, and we have been actively shaping many of them. We hope that we will not offend any reader by serious omissions.

2. P-spline basics

The two components of P-splines are B-splines and discrete penalties. In this section we briefly review them, starting with the former. We do not go much into technical detail; see Eilers and Marx (2010) for that.

2.1. B-splines

Figure 1 shows four triangles of the same height and width, the middle ones overlapping with their two neighbours. These are linear B-splines, the non-zero parts consisting of two linear segments. Imagine that we scale the triangles by different amounts and add them all up. That would give us a piecewise-linear curve. We can generate many shapes by changing the coefficients, and we can get more or less detail by using more or fewer B-splines. If we indicate the triangles by $B_j(x)$ and if a_1 to a_n are the scaling coefficients, we have $\sum_{j=1}^{n} a_j B_j(x)$ as the formula for the function. This opens the door to fitting data pairs (x_i, y_i) for $1, \ldots, m$. We minimize the sum of squares

$$S = \sum_i (y_i - \sum_j a_j B_j(x_i))^2 = ||\boldsymbol{y} - \boldsymbol{Ba}||^2,$$

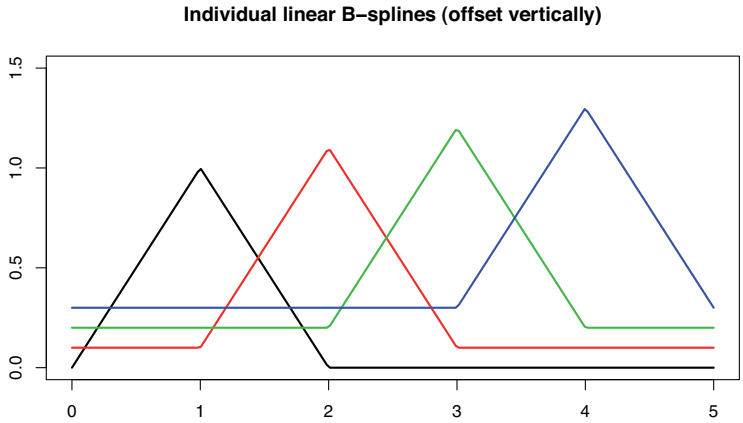

Figure 1: *Linear B-splines illustrated. The individual splines are offset for clarity. In reality the horizontal sections are zero.*

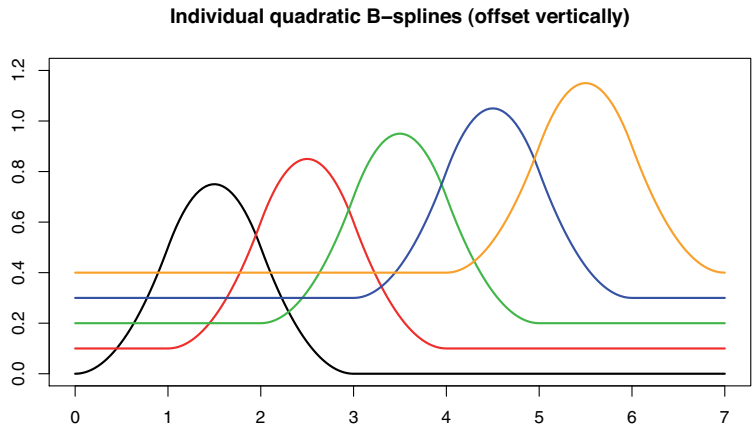

Figure 2: *Quadratic B-splines illustrated. The individual splines are offset for clarity. In reality the horizontal sections are zero.*

where $\boldsymbol{B} = [b_{ij}]$, the so-called basis matrix. This is a standard linear regression problem and the solution is well known: $\hat{\boldsymbol{a}} = (\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{y}$. The flexibility can be tuned by changing the width of the triangles (and hence their number).

A piecewise-linear fit to the data may not be pleasing to the eye, nor be suitable for computing derivatives (which would be piecewise-constant). Figure 2 shows quadratic B-splines, each formed by three quadratic segments. The segments join smoothly. In a similar way cubic B-splines can be formed from four cubic segments. The recipe for forming a curve and fitting the coefficients to data stays the same.

The positions at which the B-spline segments join are called the knots. In our illustrations the knots are equally-spaced and so all B-splines have identical shapes. This is not mandatory for general B-splines, but rather it is a deliberate choice for P-splines, as it makes the construction of penalties trivial.

One should take care when computing the B-splines. The upper panel of Figure 3 shows a basis using equally-spaced knots. Note the "incomplete" B-splines at both ends, of which not all segments fall within the domain of x. The lower panel shows a basis as computed by the R function `bs()`. It has so-called multiple knots at both ends and therefore is unsuitable for P-splines. To avoid this, one should specify an enlarged domain, and cut off the splines at both ends, by removing the corresponding columns in the basis matrix. Alternatively, one can use the code that is presented by Eilers and Marx (2010).

2.2. Discrete penalties

With the number of B-splines in the basis we can tune the smoothness of a curve to the data at hand. A smaller number of splines gives a smoother result. However, this

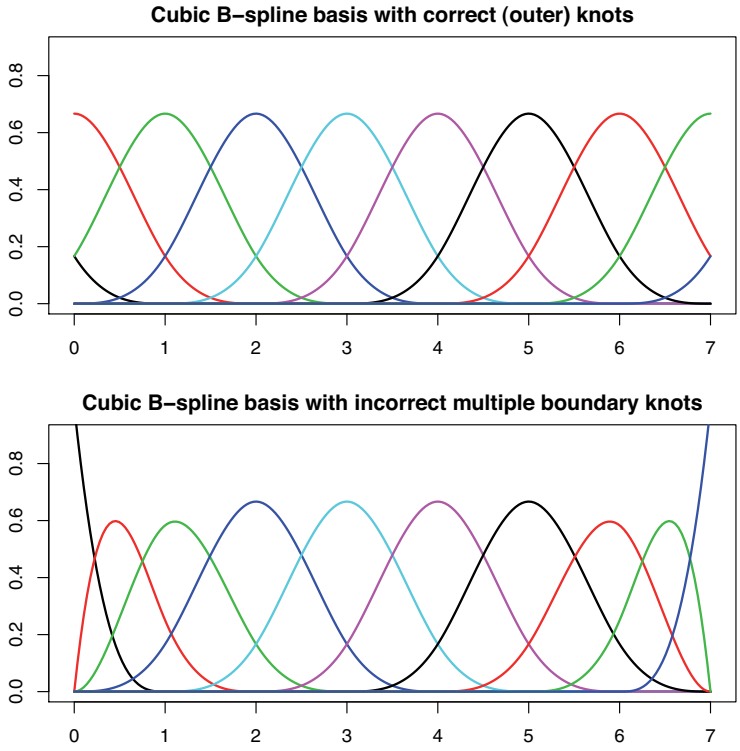

Figure 3: *B-splines bases with different choices of knots. Top: equally spaced knots, the proper basis for P-splines. Bottom: multiple knots at both ends of the domain, which is the result of the R function* `bs()` *and is unsuitable for P-splines.*

is not the only possibility. We can also use a large basis and additionally constrain the coefficients of the B-splines, to achieve as much smoothness as desired. A properly chosen penalty achieves this.

O'Sullivan (1986) had the brilliant idea to take a basis with many B-splines and to use a discrete penalty. The latter was derived from the integrated square of the second derivative of the curve. This was, and still is, an established way to measure roughness of a curve $f(x)$:

$$R = \int_l^u [f''(x)]^2 dx, \tag{1}$$

where l and u indicate the bounds of the domain of x. If $f(x) = \sum_j a_j B_j(x)$, we can derive a (banded) matrix $\boldsymbol{P}$ such that $\boldsymbol{R} = \boldsymbol{a}^\mathsf{T}\boldsymbol{P}\boldsymbol{a}$. The elements of $\boldsymbol{P}$ are computed as integrals of products of second derivatives of neighbouring B-splines.

O'Sullivan proposed to minimize

$$\boldsymbol{Q} = \boldsymbol{S} + \lambda \boldsymbol{R} = S + \lambda \boldsymbol{a}^\mathsf{T}\boldsymbol{P}\boldsymbol{a} = ||\boldsymbol{y} - \boldsymbol{a}||^2 + \lambda \boldsymbol{a}^\mathsf{T}\boldsymbol{P}\boldsymbol{a}, \tag{2}$$

where λ is the parameter that sets the influence of the penalty. The larger the λ, the smoother the result. In the limit the second derivative is forced to be very close to zero and a straight line fit will result. Note that we only have to compute $\boldsymbol{P}$ once. The system to be solved is

$$(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} + \lambda \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{y}. \tag{3}$$

The computation of $\boldsymbol{P}$ is not trivial, and it becomes quite tedious when the third or fourth order derivative is used to measure roughness. Wand and Ormerod (2008) have extended O'Sullivan's idea to higher orders of the derivative. They used a computer algebra system to construct a table of formulas. P-splines circumvent the issue by dropping derivatives and integrals completely. Instead they use a discrete penalty matrix from the start. It is also simple to compute, as it is based on difference formulas. Let $\Delta a_j = a_j - a_{j-1}$, $\Delta^2 a_j = \Delta(\Delta a_j) = a_j - 2a_{j-1} + a_{j-2}$ and in general $\Delta^d a_j = \Delta(\Delta^{d-1} a_j)$. Let $\boldsymbol{D}_d$ be a matrix such that $\boldsymbol{D}_d\boldsymbol{a} = \Delta^d\boldsymbol{a}$. If we replace the penalty by $\lambda||\boldsymbol{D}_d\boldsymbol{a}||^2 = \lambda\boldsymbol{a}^{\mathsf{T}}\boldsymbol{D}_d^{\mathsf{T}}\boldsymbol{D}_d\boldsymbol{a} = \lambda\boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{a}$, we get a similar construction as O'Sullivan's, but with a minimal amount of work. In modern languages like R and Matlab, $\boldsymbol{D}_d$ can be obtained mechanically as the dth order difference of the identity matrix.

It is surprising that nothing is lost by using a simplified penalty. Eilers and Marx (1996) showed how many many useful properties can be proved in a few lines of simple mathematics. Wand and Ormerod (2008) motivate their work by claiming that extrapolation by P-splines goes wrong. They recommended their "O-splines" as a better alternative; see also (Ruppert et al., 2009). In Appendix A we present a small study that lays severe doubt on their conclusion.

2.3. The power of the penalty

A fruitful way of looking at P-splines is to give the coefficients a central position as a skeleton, with the B-splines merely putting "the flesh on the bones." This is illustrated in Figure 4. A smoother sequence of coefficients leads to a smoother curve. The number of splines and coefficients is immaterial, as long as the latter are smooth. The role of the penalty is to make such happen.

The penalty makes interpolation easy (Currie et al., 2004; Eilers and Marx, 2010). At the positions where interpolated values are desired one introduces pseudo-observations with $y = 0$ (or any arbitrary number) and zero weights and solves the system. The true observations get weight 1. One solves

$$(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{B} + \lambda \boldsymbol{P})\hat{\boldsymbol{a}} = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{y}, \tag{4}$$

where $\boldsymbol{W}$ is a diagonal matrix with the weights on the diagonal. Smooth interpolation takes place automatically. Extrapolation can be implemented in the same way, by introducing pseudo-observations outside the domain of the data.

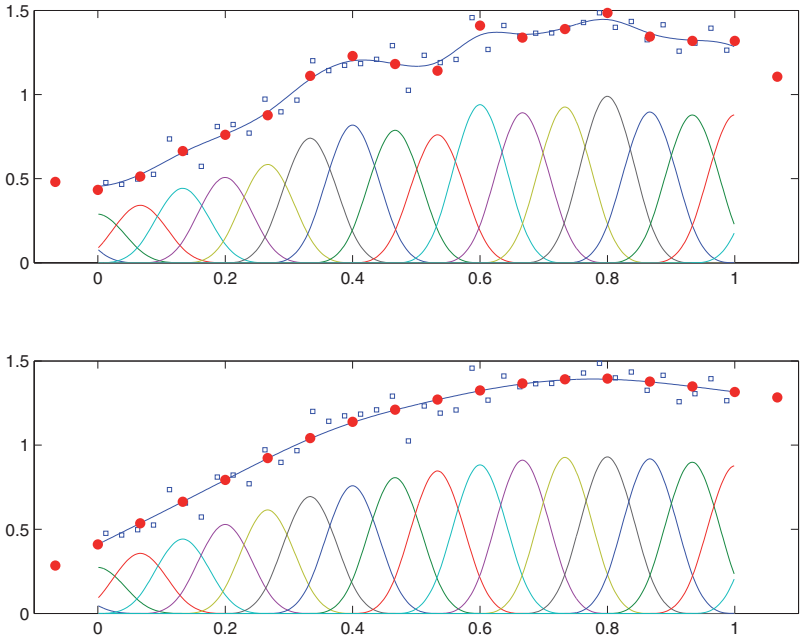

Figure 4: *Illustration of the role of the penalty. The number of B-splines is the same in both panels. In the upper panel the fit to the data (the squares) is more wiggly than in the lower panel, because the penalty is weaker there. The filled circles show the coefficients of the B-splines. Because of a stronger penalty they form a smooth sequence in the lower panel, resulting in a smoother curve fit.*

The number of B-splines can be (much) larger than the number of observations. The penalty makes the fitting procedure well-conditioned. This should be taken literally: even a thousand splines will fit ten observations without problems. Such is the power of the penalty. Figure 5 illustrates this for simulated data. There are 10 data points and 40 (+3) cubic B-splines. Unfortunately, this property of P-splines (and other types of penalized splines) is not generally appreciated. But one simply cannot have too many B-splines. A wise choice is to use 100 of them, unless computational constraints (in large models) come into sight.

We will return to this example in Section 4, after introducing the effective model dimension, and further address this issue of many splines in Appendix B.

2.4. Historical notes

The name P-splines was coined by Eilers and Marx (1996) to cover the combination of B-splines and a discrete difference penalty. It has not always been used with that specific meaning. Ruppert and Carroll (2000) published a paper on smoothing that also used the idea of a rich basis and a discrete penalty. Their basis consists of truncated power functions (TPF), the knots are quantiles of x, and the penalty is on the size of the

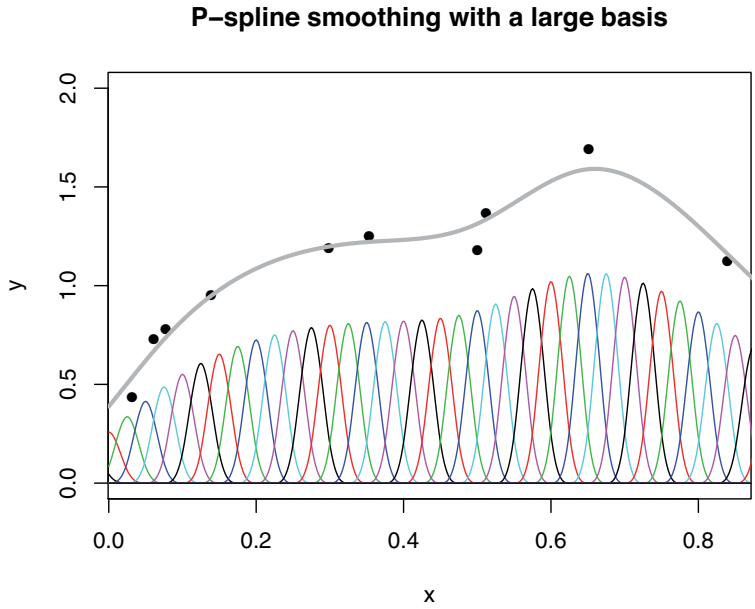

Figure 5: *P-spline smoothing of 10 (simulated) data points with 43 cubic B-splines.*

coefficients. This work has been extended in the book by Ruppert et al. (2003). Some people have called the TPF approach P-splines too. This is confusing and unfortunate because TPF are inferior to the original P-splines; Eilers and Marx (2010) documented their poor numerical condition.

B-splines and TPF are strongly related (Greven, 2008; Eilers and Marx, 2010). Actually B-splines can be computed as differences of TPF, but in the age of single precision floating point numbers it was avoided, for fear of large rounding errors. Eilers and Marx (2010) showed that this no longer holds. P-splines allow to select the degree of the B-splines and the order of the penalty independently. With TPF there is no choice: they imply a difference penalty the order of which is determined by the degree of the TPF.

3. Penalty variations

Standard P-splines use a penalty that is based on repeated differences. Many variations are possible. As stated, the B-spline coefficients form the skeleton of the fit, so if we can find other useful discrete penalties, then we can get curve fits with a variety of desired properties. Eilers and Marx (2010) called them "designer penalties" and they presented several examples. We give a summary here:

- A circular penalty connects the first and last elements of the coefficient vector using differences, making both ends connect smoothly. Combined with a circular

B-spline basis, this is the right tool for fitting periodic data or circular observations, like directions.

- With second order differences, $a_j - 2a_{j-1} + a_{j-2}$, in the penalty, the fit approaches a straight line when λ is increased. If we change the equation to $a_j - 2\phi a_{j-1} + a_{j-2}$, the limit is a (co)sine with period p such that $\phi = \cos(2\pi/p)$. The phase of the (co)sine is adjusted automatically to minimize the sum of squares of the residuals. For smoothing (and interpolation) of seasonal data (with known period) this harmonic penalty usually is more attractive than the standard one.
- Eilers and Goeman (2004) combined penalties of first and second order to eliminate negative side lobes of the impulse response (as would be the case with only a second order penalty). This guarantees that smoothing of positive data never can lead to negative fitted values.
- As described, the P-spline penalty is quadratic: it uses a sum of squares norm. This leads to a smooth result. Other norms have been used. The sum of absolute values (the L_1 norm) of first order differences allows jumps (Eilers and de Menezes, 2005) between neighbouring coefficients, making it suitable for piecewise constant smoothing. This norm is a natural choice when combined with an L_1 norm on the residuals; standard linear programming software can be used. See also Section 5 for quantile smoothing.
- The jumps that are obtained with the L_1 norm are not really "crisp," but slightly rounded. The reason is that the L_1 norm selects *and* shrinks. Much better results are obtained with the L_0 norm, the number of non-zero coefficients (Rippe et al., 2012b). Although a non-convex objective function results, in practice it can be optimized reliably and quickly by an iteratively updated quadratic penalty.

Other types of penalties can be used to enforce shape constraints. An example is a monotonously increasing curve fit (Bollaerts et al., 2006). A second, asymmetric, penalty $\kappa \sum_j v_j(\Delta a_j)$ is introduced, with $v_j = 1$ when $\Delta a_j < 0$ and $v_j = 0$ otherwise. The value of κ regulates the influence of the penalty. Iterative computation is needed, as one needs to know v to do the smoothing and then to know the solution to determine (update) v. In practice, starting from $v = 0$ works well.

Many variations are possible, to force sign constraints, to ensure (increasing or decreasing) monotonicity, or to require a convex or concave shape. One can also mix and match the asymmetric penalties to implement multiple shape constraints. Eilers (2005) used them for unimodal smoothing, while Eilers and Borgdorff (2007) used them to fit mixtures of log-concave non-parametric densities. This scheme has been extended to two dimensions by Rippe et al. (2012a) and applied to genotyping of SNPs (we discuss multidimensional smoothing in Section 8).

Pya and Wood (2015) took a different approach. They write $\boldsymbol{a} = \boldsymbol{\Sigma}\exp(\boldsymbol{\beta})$ and structure the matrix $\boldsymbol{\Sigma}$ in such a way that $\boldsymbol{a}$ has the desired shape, for any vector $\boldsymbol{\beta}$. For

example $\Sigma_{ij} = I(i \geq j)$, with the indicator function $I(\cdot)$, provides a monotonic increasing function. Patterns for combinations of constraints on first and second derivative are tabulated in their paper.

4. Diagnostics

In contrast to many other smoothers, like kernels, local likelihood, and wavelets, P-splines use a regression model with clearly defined coefficients. Hence we can borrow from regression theory to compute informative properties of the model. What we do not learn is the selection of a good value for the penalty parameter λ. Classical theory only considers the fit of a model to the data and as such is useless for this purpose. Instead we need to measure prediction performance. In this section we look at standard errors, cross-validation, effective dimension, and AIC.

The covariance matrix of the spline coefficients (for fixed λ) is given by

$$\boldsymbol{C}_a = \sigma^2(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}, \tag{5}$$

where σ is the variance of the observation noise ϵ in the model $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}$. The covariance of the fitted values follows as $\check{\boldsymbol{B}}\boldsymbol{C}_a\check{\boldsymbol{B}}^{\mathsf{T}}$, where $\check{\boldsymbol{B}}$ contains the B-spline basis evaluated at any chosen set of values of x.

As it stands, this C_a is not very useful, because we need to know σ. It could be estimated from the residuals, but for that we would need to choose the right value of λ, leading to the proper "degrees of freedom."

Leave-one-out cross-validation (CV) provides a mechanism to determine the predictive power of a P-spline model for any value of λ. Let one observation, y_i, be left out and let the predicted value be indicated by $\hat{y}_{-i}$. By doing this for each observation in turn we can compute the prediction error

$$\mathrm{CV} = \sqrt{\sum_i (y_i - \hat{y}_{-i})^2}. \tag{6}$$

As such, CV is a natural criterion to select λ, through its minimization. Using brute force, the computation of CV is expensive, certainly when the number of observations is large. Fortunately there is an exact shortcut. We have that

$$\hat{\boldsymbol{y}} = \boldsymbol{B}(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}. \tag{7}$$

Commonly $\boldsymbol{H}$ is called the "hat" matrix. One can prove that

$$y_i - \hat{y}_{i-1} = (y_i - \hat{y}_i)/(1 - h_{ii}), \tag{8}$$

and the diagonal of $\boldsymbol{H}$ can be computed quickly. A derivation can be found in Appendix B of Myers (1989). An informal proof goes as follows. Imagine that we change element i of $\boldsymbol{y}$ to get a new vector $\boldsymbol{y}^*$; then $\hat{\boldsymbol{y}}^* = \boldsymbol{H}\boldsymbol{y}^*$. Now it holds that if we set $y_i^* = \hat{y}_{-i}$, then $\hat{y}_i^* = \hat{y}_{-i}$. Hence we have that $\hat{y}_{-i} - \hat{y}_i = h_{ii}(y_{-i} - y_i)$, as $\Delta\hat{y}_i = h_{ii}\Delta y_i$. After adding $y_i - y_i$ to the right part of this equation and rearranging terms we arrive at (8).

The hat matrix also gives us the effective model dimension, if we follow Ye (1998), who proposed

$$\mathrm{ED} = \sum_i \partial\hat{y}_i/\partial y_i = \sum h_{ii}. \tag{9}$$

In fact we can compute the trace of $\boldsymbol{H}$ without actually computing its diagonal, using cyclic permutation:

$$\mathrm{ED} = \mathrm{tr}(\boldsymbol{H}) = \mathrm{tr}[(\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \lambda\boldsymbol{D}^\mathsf{T}\boldsymbol{D})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B}]. \tag{10}$$

A further simplification is possible by noting that

$$(\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} = (\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})^{-1}(\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P} - \boldsymbol{P}) = \boldsymbol{I} - (\boldsymbol{B}^\mathsf{T}\boldsymbol{W}\boldsymbol{B} + \boldsymbol{P})^{-1}\boldsymbol{P}, \tag{11}$$

where $\boldsymbol{P} = \lambda\boldsymbol{D}^\mathsf{T}\boldsymbol{D}$. Harville (1977) presented a similar result for mixed models. In the case of P-splines, the expression is very simple because there are no fixed effects.

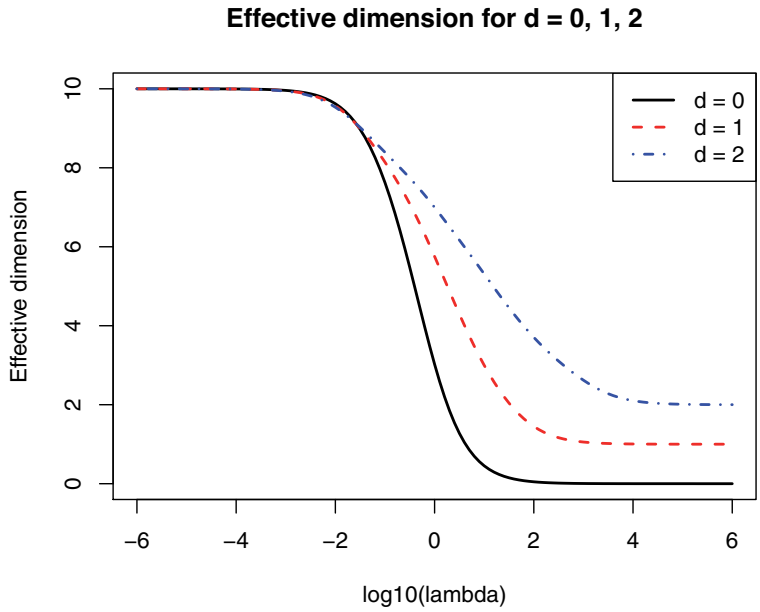

Figure 6: *Changes in the effective model dimension for P-spline smoothing of 10 (simulated) data points with 43 cubic B-splines, for different orders (d) of the differences in the penalty.*

The effective dimension is an excellent way to quantify the complexity of a P-spline model. It summarizes the combined influences of the size of the B-spline basis, the order of the penalty, and the value of the smoothing parameter. The last equation in (11) neatly shows that the effective dimension will always be smaller than n. Actually the effective dimension is always smaller than $\min(m,n)$. An illustration is presented by Figure 6, showing how ED changes with λ for the example with 10 observations and 43 B-splines in Figure 5. For small λ, ED approaches m, while for large λ it approaches d, the order of the differences.

The fact that ED $< n$ is obvious from the size of the system of penalized likelihood equations. A heuristic argument for ED $< m$ is that $\boldsymbol{B}(\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B}+\lambda\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}\boldsymbol{B}$ is an m by m matrix. It is a hat matrix, having a trace smaller than m. A formal proof is given in Appendix B.

Additionally, the fact that ED $< m$ explains why smoothing with (many) more B-splines than observations works without a problem, for any value of λ. In our experience, many colleagues do not realize this fact. Maybe they fear singularities and stick to small numbers of basis functions.

To estimate σ^2, one divides the sum of squares of the residuals by their effective degrees of freedom, which is the number of observations minus the the effective model dimension: $\hat{\sigma^2} = \sum_i (y_i - \hat{y}_i)^2/(m - \text{ED})$.

Alternatively, one can use Akaike's Information Criterion to choose λ, where AIC $= -2\ell + 2\text{ED}$ and ℓ is the log-likelihood. The beauty of this formula is that it shows the balance between fidelity to the data and complexity of the model.

One should always be careful when using cross-validation or AIC to tune the smoothing parameter. An implicit assumption is that the observations are independent, conditional on their smooth expectations. If this is not the case, as for time series data, the serial correlation will be picked up as a part of the smooth component and severe undersmoothing can occur. One way to approach this problem is to explicitly model the correlation structure of the noise. We return to this subject in Section 10 on mixed models. A recent alternative strategy is the adaptation of the L-curve (Hansen, 1992). It was developed for ridge regression, but can be adapted to difference penalties. See Frasso and Eilers (2015) for examples and a variation, called the V-curve, which is easier to use.

In Section 10 the tuning parameter for the penalty will appear as a ratio of variances, and the effective dimension plays an essential role when estimating them.

5. Generalized linear smoothing and extensions

P-splines are based on linear regression, so it is routine to extend them for smoothing non-normal observations, by borrowing the framework of generalized linear models (GLM). Let $\boldsymbol{y}$ be observed and $\boldsymbol{\mu}$ the vector of expected values. Then the linear predictor $\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{Ba}$ is modelled by B-splines, and a suitable distribution is chosen for y, given $\boldsymbol{\mu}$. The penalty is subtracted from the log-likelihood: $\ell^* = \ell - \lambda\boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}/2$. The penal-

ized likelihood equations result in $\boldsymbol{B}^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{\mu}) = \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}\boldsymbol{a}$. This is a small change from the standard GLM, in which the right-hand side is zero (Dobson and Barnett, 2008).

The equations are non-linear, but penalized maximum likelihood leads to the iterative solution of

$$\hat{\boldsymbol{a}}_{t+1} = (\boldsymbol{B}^{\mathsf{T}}\hat{\boldsymbol{W}}_t\boldsymbol{B} + \lambda \boldsymbol{D}_d^{\mathsf{T}}\boldsymbol{D}_d)^{-1}\boldsymbol{B}^{\mathsf{T}}\hat{\boldsymbol{W}}_t\hat{\boldsymbol{z}}_t \quad \text{with} \quad \boldsymbol{z}_t = \hat{\boldsymbol{\eta}}_t + \hat{\boldsymbol{W}}_t^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}_t), \tag{12}$$

where t denotes the iterate and $\hat{\boldsymbol{W}}_t$ and $\hat{\boldsymbol{\eta}}_t$ denote approximate solutions, while $\hat{\boldsymbol{z}}_t$ is the so-called working variable. The weights in the diagonal matrix $\hat{\boldsymbol{W}}$ depend on the link function and the chosen distribution. For example, the Poisson distribution, with $\eta = \log(\mu)$ has $\hat{w}_{ii} = \hat{\mu}_i$.

A powerful application of generalized linear smoothing with P-splines is density estimation (Eilers and Marx, 1996). A histogram with narrow bins is computed and the counts are smoothed, using the Poisson distribution and the logarithmic link function. There is no danger that bins are chosen too narrow: even if most of them contain only a few counts or zeros good results are obtained. The amount of smoothing is tuned by AIC.

It is essential (for any smoother) that enough bins with zero counts are included at the ends of the observed domain of the data, unless it is known to be bounded (as for intrinsically positive variables).

P-splines conserve moments of distributions up to order $d-1$, where d is the order of the differences in the penalty. This means that, if $d = 3$, the sum, the mean, and the variance of the smooth histogram are equal to those of the raw histogram, whatever the amount of smoothing (Eilers and Marx, 1996). In contrast, kernel smoothers increase variance.

Many variations on this theme have been published. We already mentioned one- and two-dimensional log-concave densities in Section 3.. Kauermann et al. (2013) explored flexible copula density estimation. They modelled the density directly as a sum of tensor products of linear B-splines (we discuss tensor products in Section 8). To reduce the number of coefficients, they used reduced splines, which are similar to nested B-splines (Lee et al., 2013).

Another variation is not to model the logarithm of the counts by a sum of B-splines, but rather the density itself, with constraints on the coefficients (Schellhase and Kauermann, 2012).

Mortality or morbity smoothing is equivalent to discrete density estimation with an offset for exposures. P-splines have found their way into this area, for both one- and two-dimensional tables (Currie et al., 2004; Camarda, 2012); both papers illustrate automatic extrapolation.

The palette of distributions that generalized linear smoothing can use is limited. A very general approach is offered by GAMLSS: generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005). An example is the normal distribution with smoothly varying mean and variance, combined with a (varying) Box-Cox trans-

form of the response variable. Many continuous and discrete distributions can be fitted by the GAMLSS algorithm, also in combination with mixtures, censoring and random components.

Instead of using a parametric distribution, one can estimate smooth conditional quantiles, minimizing an asymmetrically weighted sum of absolute values of the residuals. Bollaerts et al. (2006) combined it with shape constraints to force monotonicity. To avoid crossing of individually estimated smooth quantile curves, Schnabel and Eilers (2013) introduced the quantile sheet, a surface on the domain formed by the explanatory variable and the probability level.

Compared to the explicit solutions of (penalized, weighted) least squares problems, quantile smoothing is a bit less attractive for numerical work as it leads to linear programming or to quadratic programming if quadratic penalties are involved. In contrast, expectiles use asymmetrically weighted sums of squares and lead to simple iterative algorithms (Schnabel and Eilers, 2009). Sobotka and Kneib (2012) extended expectile smoothing to the spatial context, while Sobotka et al. (2013) provide confidence intervals. Schnabel and Eilers (2013) proposed a location-scale model for non-crossing expectile curves.

When analysing counts with a generalized linear model, often the Poisson distribution is assumed, with $\boldsymbol{\mu} = \exp(\boldsymbol{\eta})$ for the expected values. When counts are grouped or aggregated, the composite link model (CLM) of Thompson and Baker (1981) is more appropriate. It states that $\boldsymbol{\mu} = \boldsymbol{C}\exp(\boldsymbol{\eta})$, where the matrix $\boldsymbol{C}$ encodes the aggregation or mixing pattern. In the penalized CLM, a smooth structure for $\boldsymbol{\eta}$ is modelled with P-splines (Eilers, 2007). It is a powerful model for grouped counts (Lambert and Eilers, 2009; Lambert, 2011; Rizzi et al., 2015), but it has also found application in misclassification and digit preference (Camarda et al., 2008; Azmon et al., 2014). de Rooi et al. (2014) used it to remove artifacts in X-ray diffraction scans.

6. Generalized additive models

The generalized additive model (GAM) constructs the linear predictor as a sum of smooth terms, each based on a different covariate (Hastie and Tibshirani, 1990). The model is $\boldsymbol{\eta} = \sum_j f_j(\boldsymbol{x}_j)$; it can be interpreted as a multidimensional smoother without interactions.

The GAM with P-splines, or P-GAM, was proposed by Marx and Eilers (1998). We illustrate the main idea in two dimensions. Let

$$\boldsymbol{\eta} = f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) = [\boldsymbol{B}_1|\boldsymbol{B}_2] \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \end{bmatrix} = \boldsymbol{B}\boldsymbol{a}. \tag{13}$$

By combining the two bases into one matrix and chaining the coefficients in one vector we are back in a standard regression setting.

The roughness penalties are $\lambda_1||\boldsymbol{D}_1\boldsymbol{a}_1||^2$ and $\lambda_2||\boldsymbol{D}_2\boldsymbol{a}_2||^2$ (where the indices here refer to the variables, not to the order of the differences), leading to two penalty matrices $\boldsymbol{P}_1 = \lambda_1\boldsymbol{D}_1^{\mathsf{T}}\boldsymbol{D}_1$ and $\boldsymbol{P}_2 = \lambda_2\boldsymbol{D}_2^{\mathsf{T}}\boldsymbol{D}_2$, which can be combined in the block-diagonal matrix P. The resulting penalized likelihood equations are $\boldsymbol{B}^{\mathsf{T}}(\boldsymbol{y}-\boldsymbol{\mu}) = \boldsymbol{P}\boldsymbol{a}$, which have exactly the same form as those for generalized linear smoothing. The weighted regression equations follow immediately. The same is true for the covariance matrix of the estimated coefficients, cross-validation, and the effective dimension.

Originally, backfitting was used for GAMs (Hastie and Tibshirani, 1990), updating each component function in turn, using any type of smoother. Convergence can be slow and diagnostics are hard or impossible to obtain. Direct fitting by P-splines does not have these disadvantages.

As presented the model is unidentifiable, because an arbitrary upward shift of $f_1(\boldsymbol{x}_1)$ can be compensated by an equal shift downward of $f_2(\boldsymbol{x}_2)$. A solution is to introduce an (unpenalized) intercept and to constrain each component to have a zero average.

The P-GAM has multiple smoothing parameters, so optimization of AIC, say, by a simple grid search involves much work. Heim et al. (2007) proposed a searching strategy that cycles over one-dimensional grid searches. As a more principled approach, Wood (2004) presented algorithms for numerical optimization in cross-validation. His book (Wood, 2006a) contains a wealth of material on GAMs. See also Section 13 for the `mgcv` software.

In Section 10 we will present Schall's algorithm for variance estimation. It is attractive for tuning multiple penalty parameters.

7. Smooth regression coefficients

In the preceding sections P-splines were used to model expected values of observations. There is another class of models in which the goal is to model regression coefficients as a curve or surface. In this section we discuss varying coefficient models (Hastie and Tibshirani, 1993), penalized signal regression (Marx and Eilers, 1999), and generalizations. In modern jargon these are all cases of functional data analysis (Ramsay and Silverman, 2003).

Varying coefficient models (VCM) were first introduced by Hastie and Tibshirani (1993). They allow regression coefficients to interact with another variable by varying smoothly. The simplest form is $E[y(t)] = \mu(t) = \beta(t)x(t)$, where $\boldsymbol{y}$ and $\boldsymbol{x}$ are observed and $\boldsymbol{\beta}$ is to be estimated and forced to change slowly with t. The model assumes that $\boldsymbol{y}$ is proportional to $\boldsymbol{x}$, with a varying slope of the regression line. If we introduce a B-spline basis B and write $\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{a}$, we get $\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{B}\boldsymbol{a}$, where $\boldsymbol{X} = \mathrm{diag}(\boldsymbol{x})$. With a difference penalty on $\boldsymbol{a}$ we have the familiar P-spline structure, with only a modified basis $\boldsymbol{X}\boldsymbol{B}$. A varying offset can be added: $E[y(t)] = \mu(t) = \beta(t)x(t) + \beta_0(t)$. This has the form of an additive model. Building $\boldsymbol{\beta}_0$ with P-splines we effectively get a P-GAM.

This simple VCM can be extended by adding more additive or varying-coefficient terms. For non-normal data we model the linear predictor and choose a proper response distribution.

VCM with P-splines were proposed by Eilers and Marx (2002). Lu et al. (2008) studied them too and presented a Newton-Raphson procedure to minimize the cross-validation error. Andriyana et al. (2014) brought quantile regression into VCMs using P-splines. Kauermann (2005b) and Kauermann and Khomski (2006) developed P-spline survival and hazard models, respectively, to accommodate varying-coefficients. Wang et al. (2014) used VCMs for longitudinal data (with errors in variables) with Bayesian P-splines. Heim et al. (2007) used a 3D VCM in brain imaging.

Modulation models for seasonal data are an interesting application of the VCM (Eilers et al., 2008; Marx et al., 2010). The amplitude of a sinusoidal (or more complex) waveform is made to vary slowly over time. This assumes that the period is known. If that is not the case, or when it is not constant, it is possible to estimate both varying amplitude and phase of a sine wave (Eilers, 2009).

In a VCM, $\boldsymbol{y}$ and $\boldsymbol{x}$ are parallel vectors given at the same sampling positions in time or space. In penalized signal regression (PSR) we have a set of $\boldsymbol{x}$ vectors and corresponding scalars in $\boldsymbol{y}$ and the goal is to predict the latter. If the $\boldsymbol{x}$ vectors form the rows of a matrix $\boldsymbol{X}$, we have linear regression $E(\boldsymbol{y}) = \boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\beta}$. The problem is ill-posed, because $\boldsymbol{X}$ has many more columns then rows. Take for example optical spectra that have been measured with many hundreds of wavelengths. The elements of $\boldsymbol{y}$ are known concentrations of a substance. Because the columns of $\boldsymbol{X}$ are ordered, it makes sense to force $\boldsymbol{\beta}$ to be smooth, by putting a difference penalty on it, thereby making the problem well-posed (Marx and Eilers, 1999).

In principle there is no need to introduce P-splines, by writing $\boldsymbol{\beta} = \boldsymbol{B}\boldsymbol{a}$ and putting the penalty on $\boldsymbol{a}$, but it reduces the computational load when $\boldsymbol{X}$ has many columns. Effectively we get penalized regression on the basis $\boldsymbol{U} = \boldsymbol{X}\boldsymbol{B}$. After this step the machinery for cross-validation, standard errors and effective dimension becomes available. Notice that $\boldsymbol{a}$ is forced to be smooth, but $\boldsymbol{\mu}$ does not have to be smooth at all. Also not the rows of $\boldsymbol{X}$ are smoothed, but the regression coefficients.

Li and Marx (2008) proposed signal sharpening to enhance external prediction by incorporating PLS weights.

An extensive review of functional regression was presented by Morris (2015).

The standard PSR model implicitly assumes the identity link function. However, it can be bent through $\boldsymbol{\mu} = f(\boldsymbol{X}\boldsymbol{\beta}) = f(\boldsymbol{U}\boldsymbol{a})$, where $f(\cdot)$ is unknown. We call this model single-index signal regression (SISR), which is closely related to projection pursuit (Eilers et al., 2009). To estimate f, a second B-spline basis and corresponding coefficients are introduced. The domain is that of $\boldsymbol{U}\boldsymbol{a}$, and $\boldsymbol{a}$ has to be standardized (e.g. mean zero and variance 1) to make the model identifiable. For given coefficients, the derivative of $f(\boldsymbol{U}\boldsymbol{a})$ can be computed and inserted in a Taylor expansion. Using that, $\boldsymbol{a}$ and the coefficients for f are updated in turn until convergence.

P-splines have been implemented in other types of single-index models, e.g. see Yu and Ruppert (2002) and Lu et al. (2006). Leitenstorfer and Tutz (2011) used boosting and Antoniadis et al. (2004) used a Bayesian approach.

In the next section, we review the tensor product fundamentals that enable PSR extensions into two-dimensions. For example, Eilers and Marx (2003) and Marx and Eilers (2005) extended PSR to allow interaction with a discrete variable and to the two-dimensional case where each $\boldsymbol{x}$ is not a vector but a matrix. In these models there is no notion of time. When each element of $\boldsymbol{y}$ is not a scalar but a time series, as is $\boldsymbol{x}$, the historical functional linear model (HFLM) assumes that in principle all previous $\boldsymbol{x}$ can influence the elements of $\boldsymbol{y}$ (Malfait and Ramsay, 2003). Harezlak et al. (2007) introduced P-spline technology for the HFLM.

A mirror image of the HFLM is the interest term structure, estimating the expected future course of interest rates; see Jarrow et al. (2004) and Krivobokova et al. (2006).

Additionally, Marx et al. (2011) extended SISR to two dimensions, whereas Marx (2015) presented a hybrid varying-coefficient single-index model. In SISR, a weighted sum of $x(t)$ is formed and transformed. McLean et al. (2014) went one step further: $E(y_i) = \mu_i = \int F(x_i(t),t)dt$. This can be interpreted as first transforming x (with a different function for each t) and then adding the results, or "transform and add" in contrast to "add and transform".

8. Multi-dimensional smoothing

A natural extension of P-splines to higher dimensions is to form tensor products of one-dimensional B-spline bases and to add difference penalties along each dimension (Eilers and Marx, 2003). Figure 7 illustrates this idea, showing one tensor product $T_{jk}(x,y) = B_j(x)\check{B}_k(y)$. Figure 8 illustrates a "thinned" section of a tensor product basis; for clarity not all tensor products are shown. A matrix of coefficients determines the height of each "mountain": $\boldsymbol{A} = [a_{kl}]$, $k = 1,\ldots,\ n$ and $l = 1,\ldots,\ \check{n}$. The situation is completely analogous to Figure 4, but extended to two dimensions. The roughness of the elements of $\boldsymbol{A}$ determines how smooth the surface will be. To tune roughness, each column and each row of $\boldsymbol{A}$ is penalized.

One can choose to use one penalty parameter for both directions, (isotropic smoothing), or separate ones (anisotropic smoothing). In the latter case optimizing the amount of smoothing generates much more work. Many useful properties of one-dimensional P-splines carry over to higher dimensions. Weighting of (missing) observations and interpolation and extrapolation work well. Effective model dimension and fast cross-validation are available. They can also be used as a building block in smooth structures (see the next section).

Technically, multidimensional P-splines are challenging. The main issue is that, to be able to estimate A with the usual matrix-vector operations, we need to write it as a

vector and to put the tensor products in a proper basis matrix. With careful organization of the computations this can be solved elegantly (Eilers et al., 2006).

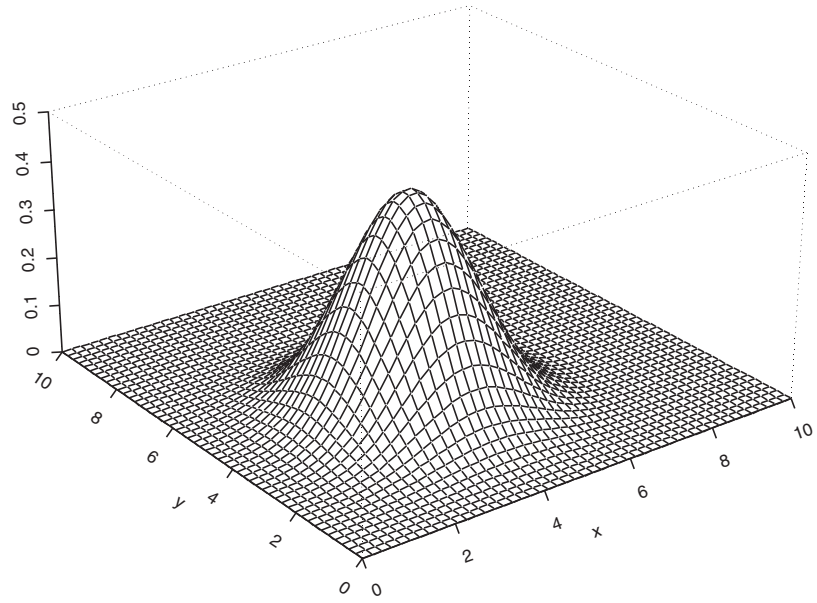

Figure 7: *The tensor product building block.*

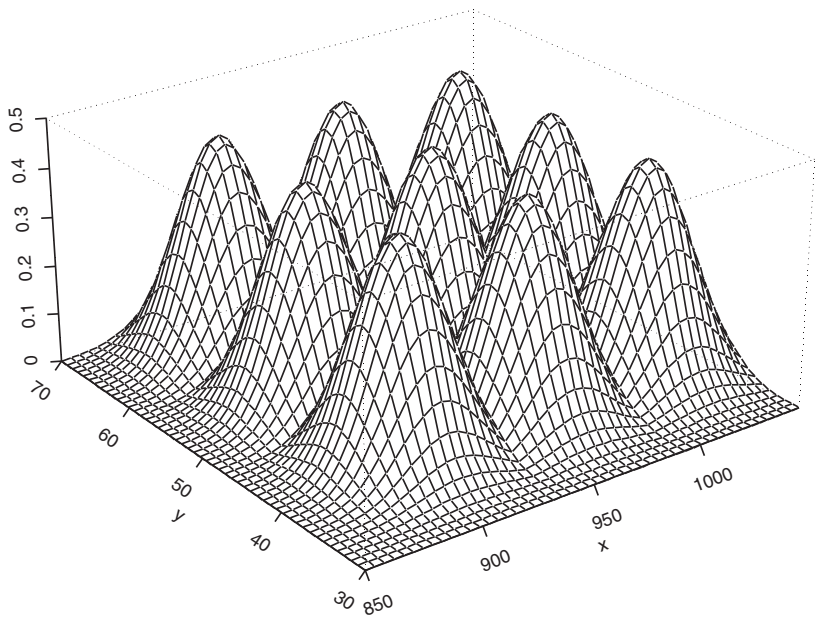

Figure 8: *Sparse portion of a complete tensor product B-spline basis.*

A natural application of multidimensional P-splines is the smoothing of data on a grid. For larger grids the demands on memory and computation time can become too large and special algorithms are needed. See Section 13 for details.

Multi-dimensional P-splines are numerically well-behaved, in contrast to truncated power functions. The poor numerical condition of the latter becomes almost insurmountable in higher dimensions. Proponents of TPF have avoided this issue by using radial basis functions (Ruppert et al., 2003; Kammann and Wand, 2003). This is, however, not an attractive scheme: a complicated algorithm is being used for placing the centres of the basis functions.

We emphasize that the set of tensor products does not have to be rectangular, although Figure 8 might give that impression. When dealing with, say, a ring-shaped data domain, we can remove all tensor products that do not overlap with the ring, and reduce the penalty matrix accordingly. This can save much computation, and in the case of a ring also is more realistic, because it prevents the penalty from working across the inner region.

As we showed for one-dimensional smoothing, the number of basis elements, here the tensor products, can be larger than the number of observations without problems, thanks to the penalties.

9. Additive smooth structures

As we have seen for the generalized additive and varying-coefficient model, the use of P-splines leads to a set of (modified) B-spline basis matrices which can be combined side-by-side into one large matrix. The penalties lead to a block-diagonal matrix. This idea extends to other model components like signal regression and tensor products. Standard linear regression and factor terms can be added too. This leads to additive smooth structures. Eilers and Marx (2002) proposed GLASS (generalized linear additive smooth structures), while Brezger and Lang (2006), referring to Fahrmeir et al. (2004), proposed STAR (structured additive regression). Belitz and Lang (2008) introduced simultaneous selection of variables and smoothing parameters in structured additive models.

The geoadditive model has received much attention; it is formed by the addition of one-dimensional smooth components and a two-dimensional spatial trend. Often the spatial component is modelled as a conditional autoregressive model. Brezger and Lang (2006) presented a Bayesian version of GLASS/STAR, also using 2D P-splines for modelling spatial effects in a multinomial logit model for forest health. Fahrmeir and Kneib (2009) further built on Bayesian STAR models by incorporating geoadditive features and Markov random fields, while addressing improper prior distributions. Also considering geoadditive structure, Kneib et al. (2011) expanded and unified Bayesian STAR models to further accommodate high-dimensional covariates.

Hierarchies of curves form a special type of additive smooth structures. For example, in growth data for children we can introduce an overall mean curve and two additional

curves that show the difference between boys and girls. Moreover, we can have a smooth curve per individual child. Durbán et al. (2005) gave an example (using truncated power functions), while Bugli and Lambert (2006) used proper P-splines in a Bayesian context.

10. P-splines as a mixed model

The connection between nonparametric regression and mixed models was first established over 25 years ago by Green (1987) and Speed (1991), but it was not until the late 1990s before it became a "hot" research topic (Wang, 1998; Zhang et al., 1998; Verbyla et al., 1999), partly due to the developments in mixed model software. These initial references were based on the use of smoothing splines. In the penalized spline context, several authors quickly extended the model formulation into a mixed model (Brumback et al., 1999; Coull et al., 2001; Wand, 2003). They used truncated power functions as the regression basis, since these have an intuitive connection with a mixed model. However, as previously mentioned, the numerical properties of TPFs are poor, compared to P-splines. In a short comment, that largely went unnoticed, Eilers (1999) showed how to interpret P-splines as a mixed model. Currie and Durbán (2002) used this approach and extended it to handle heteroscedastic or autocorrelated noise. Work on the general approach for a mixed model representation of smoothers with quadratic penalty was also presented in Fahrmeir et al. (2004).

With $\lambda = \sigma^2/\sigma_a^2$, the minimization problem in (2) is equivalent to:

$$Q^* = ||\boldsymbol{y} - \boldsymbol{B}\boldsymbol{a}||^2/\sigma^2 + \boldsymbol{a}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{a}/\sigma_a^2, \tag{14}$$

with σ_a^2 denoting the variance of the random effects a and σ^2 as the error variance. In fact, this is the minimization criterion in a random effects model of the form:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{a} + \boldsymbol{\epsilon}, \quad \boldsymbol{a} \sim N(0, \sigma_a^2 \boldsymbol{P}^{-1}) \quad \epsilon \sim N(0, \sigma^2 \boldsymbol{I}). \tag{15}$$

As presented, difference penalties of order d do not penalize powers of x up to degree $d-1$. Therefore, $\boldsymbol{P}$ is singular (d eigenvalues are zero), and thus $\boldsymbol{a}$ has a degenerate distribution. One solution is to rewrite the model as $\boldsymbol{B}\boldsymbol{a} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$, such that the d columns of $\boldsymbol{X}$ span the polynomial null space of $\boldsymbol{P}$ and the $(n-d)$ columns of $\boldsymbol{Z}$ span its complement. In this presentation, the random effects u have a non-degenerate distribution. This type of re-parametrization can be done in many ways. Eilers (1999) proposed $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{D}^{\mathsf{T}}(\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}})^{-1}$ (where $\boldsymbol{D}$ is the differencing matrix). A more principled approach (which can be used for any quadratic penalty) was introduced by Currie et al. (2006) and is based on the singular value decomposition of $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathsf{T}}$, yielding $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{U}\boldsymbol{\Sigma}^{-1}$. In either case, the equivalent mixed model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{u} \sim N(0, \sigma_u^2 \boldsymbol{I}) \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I}). \tag{16}$$

Instead of one smoothing parameter, we now have two variances, and we can profit from the stable and efficient algorithms and software that are available for mixed models. Especially in complex models with multiple smooth components, this approach can be more attractive than optimizing cross-validation or AIC. Yet, which approach (based on prediction error or maximum likelihood) is optimal for the selection of the smoothing parameter? Several papers on this subject have appeared along the years, but no unified opinion has been reached: Kauermann (2005a) showed that the ML estimate has the tendency to under-smooth, and prediction error methods give better performance than maximum likelihood based approaches, Gu (2002) also found that ML delivers rougher estimates than GCV, while Ruppert et al. (2003) found, through simulation studies, that REML will produce smoother fits than GCV (simular conclusion was also found in Kohn et al., 1991). Also, Wood (2011) concluded that REML/ML estimation is preferable to GCV for semiparametric GLMs due to its better resistance to over-fitting, less variability in the estimated smoothing parameters, and reduced tendency to having multiple minima. So, it is clear that there is no unique answer to this question, since different scenarios, will yield different conclusions. Moreover, behind the criteria used to select the smoothing parameter, there is, in our opinion, a deeper question: is it fair to use mixed models methodology for estimation and inference, when the mixed model representation of a P-spline could be considered just a "trick" to facilitate parameter estimation? This is a question for which we have no answer; researchers have different (and strong) opinions about the mixed model approach (even the authors of this paper do not always agree on this matter), but the truth is that it has become a revolution that has yielded incredible advances in a very short time. It certainly has helped to make penalized splines "salonfähig": nowadays they are acceptable and even attractive to a large part of the statistical community.

The estimation of the fixed and random effects is based on the maximization of the joint density of $(\boldsymbol{y}, \boldsymbol{u})$ for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ which results in the well-known Henderson's equations (Henderson, 1975):

$$\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{y} \\ \boldsymbol{Z}^\top \boldsymbol{y} \end{bmatrix}, \tag{17}$$

where $\lambda = \sigma^2/\sigma_u^2$. The solution of these equations yields $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$. The variance components σ^2 and σ_u^2 are, in general, estimated by REML (Restricted Maximum Likelihood), see Patterson and Thompson (1971)), and the solutions are obtained by numerical optimization.

Other approaches can be used, and among them, it is worth mentioning the algorithm given by Schall (1991), which estimated random effects and dispersion parameters without the need to specify their distribution. The key is that each variance component is connected to an effective dimension. The sum of squares of the corresponding random

coefficients is equal to their variance times their respective effective dimension. This fact can be exploited in an iterative algorithm. After each cycle of smoothing, the sums of squares and effective dimensions are recomputed, which then are used to update the variances for the next round. See Marx (2010) and Rodriguez-Alvarez et al. (2015) for details and for extensions to multidimensional smoothing.

It is important to note a fact about prediction. Although the fitted model is the same regardless of parametrization (i.e. as mixed model or not), the standard errors for the predicted values are not invariant. This results because the variability of the random effects is taken into account in the mixed model case (and not in the other). The confidence interval obtained from the original parametrization is $\hat{f}(\boldsymbol{x}) \pm 2\hat{\sigma}\sqrt{(\boldsymbol{HH})_{ii}}$ (where $\boldsymbol{H}$ is such that $\hat{\boldsymbol{y}} = \boldsymbol{Hy}$). This confidence interval covers $E[\hat{f}(\boldsymbol{x})]$ rather than $f(\boldsymbol{x})$, since $\hat{f}(\boldsymbol{x})$ is not an unbiased estimate of $f(\boldsymbol{x})$. Whereas in the mixed model framework, $\hat{f}(\boldsymbol{x})$ is unbiased due to the random $\boldsymbol{u}$, and the biased adjusted confidence interval is $\hat{f}(\boldsymbol{x}) \pm 2\hat{\sigma}\sqrt{(\boldsymbol{H})_{ii}}$ (Ruppert et al., 2003).

Of course, the interest of the mixed model representation of P-splines has been motivated by the possibility of including smoothing in a larger class of models. In fact, during the last 15 years, there has been an explosion of models: ranging from estimating subject-specific curves in longitudinal data (Durbán et al., 2005), to extending classical models in economics Basile et al. (2014), to playing a key role in the recent advances in functional data analysis (Scheipl et al., 2015; Brockhaus et al., 2015), among others.

10.1. P-splines and correlated errors

Although the mixed model approach has allowed the generalization of many existing models, there is an area in which it has played a key role: data with serial correlation. For many years the main difficulty when fitting a smooth model in the presence of correlation has been the joint estimation of the smoothing and correlation parameters. It is well known that the standard methods for smoothing parameter selection (based on minimization of the mean squared prediction error) generally under-smooth the data in the presence of positive correlation, since a smooth trend plus correlated noise can be seen as a less smooth trend plus white noise.

The solution is to take into account the correlation structure explicitly, i.e. $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ can depend on one or more correlation parameters. Durbán and Currie (2003) presented a strategy to select the smoothing parameter and estimate the correlation based on REML. Krivobokova and Kauermann (2007) showed that maximum likelihood estimation of the smoothing parameter is robust, even under moderately misspecified correlation. This method has allowed the inclusion of temporal non-linear trends and filtering of time series (Kauermann et al., 2011).

Recently, and motivated by the need to improve the speed and stability of forecasting models, Wood et al. (2015) have developed efficient methods for fitting additive models to large data sets with correlated errors.

Correlation also appears in more complex situations, for example in the case of spatial data. Lee and Durbán (2009) combined two-dimensional P-splines and random effects with a CAR (conditional auto-regressive) structure to estimate spatial trends when data are geographically distributed over locations on a map. Other authors have taken different approaches; they combined additive mixed models with spatial effects represented by Markov or Gaussian random fields (Kneib and Fahrmeir, 2006; Fahrmeir et al., 2010).

10.2. Multidimensional P-splines as mixed models

Multidimensional P-splines can be handled as a mixed model too. A first attempt was made by Ruppert et al. (2003) using radial basis functions. Currie et al. (2006) analysed tensor product P-splines as mixed models. Here, the singular value decomposition of the penalty matrix (as in the 1D case) is used to construct the mixed model matrices. This approach works for any sum of quadratic penalties (Wood, 2006a). However, when the penalty is expressed as the sum of Kronecker product of marginal bases (the Kronecker sum of penalties), the representation as a mixed model is based on the reparametrization of the marginal bases. An important by-product of this parametrization is that the transformed penalty matrix (i.e. the covariance matrix of the random effects), and the mixed model matrices lead to an interesting decomposition of the fitted values as the sum of main effects and interactions (Lee and Durbán, 2011):

$$E(\boldsymbol{Y}) = f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) + f_3(\boldsymbol{x}_1, \boldsymbol{x}_2).$$

This decomposition is strongly related to the work proposed by Gu (2002) on smoothing spline analysis of variance.

The model now has multiple smoothing parameters, which makes estimating them less efficient, if numerical optimization were to be used. Several steps have been taken to make computation efficient. Wood (2011) used a Laplace approximation to obtain an approximate REML suitable for efficient direct approximation. Lee et al. (2013) improved computational efficiency by using nested B-spline bases, and modified the penalty so that optimization could be carried out in standard statistical software. Wood and Scheipl (2013) proposed an intermediate low-rank smoother. Recently, Schall's algorithm has been extended (Rodriguez-Alvarez et al., 2015) to the case of multidimensional smoothing. This work also shows the fundamental role of the effective dimensions of the components of the model.

11. Bayesian P-splines

It is a small step to go from a latent distribution in a mixed model to a prior in a Bayesian interpretation. Bayesian P-splines were proposed by Lang and Brezger (2004), and they were made accessible by appropriate software (Brezger et al., 2005). Their approach is based on Markov chain Monte Carlo (MCMC) simulation. As for the mixed model, the penalty leads to a singular distribution. This is solved by simulation using a random walk of the same order as that of the differences.

It is also possible to start from a mixed model representation. Crainiceanu et al. (2007) did this in one dimension, using truncated power functions. They avoid tensor products of TPFs and switch to radial basis functions for spatial smoothing. These authors also allowed for varying (although isotropic) smoothness and for heteroscedasticity. Jullion and Lambert (2007) proposed a Bayesian model for adaptive smoothing.

As an alternative to MCMC, integrated nested Laplace approximation (INLA) is powerful and fast, and it is gaining in popularity (Rue et al., 2009). INLA avoids stochastic simulation for precision parameters and uses numerical integration instead. Basically INLA uses a parameter for each observation so a (B-spline) regression basis has to be implemented in an indirect way, as a matrix of constraints (Fraaije et al., 2015).

INLA is an attractive choice for anisotropic smoothing. By working with a sum of precision matrices it can handle the equivalent of a mixed model with overlapping penalties (Rodriguez-Alvarez et al., 2015).

12. Varia

In this section we discuss some subjects that do not find a natural home in one of the preceding sections. We take a look at asymptotic properties of P-splines and at boosting.

Several authors have studied the asymptotic behaviour of P-splines. See Li and Ruppert (2008); Claeskens et al. (2009); Kauermann et al. (2009); Wang et al. (2011). Although we admire the technical level of these contributions, we do not fully see their practical relevance. The problem is their very limited interpretation of increasing the number of observations: it is all about more observations on the same domain. In that case it is found that the number of knots should grow as a small power of the number of observations. Yet, the whole idea of P-splines is to use far too many knots and let the penalty do the work. Trying to optimize the number of knots, as Ruppert (2002) did, is not worthwhile. He reports some cases where more knots increase the estimation error, but the numbers are not dramatic. His analysis was based on truncated power functions, which he confusingly calls P-splines, with knots at quantiles of the observed x. It is not clear how this design influences the results. For a proper analysis equally-spaced knots have to be used.

Most asymptotic analyses of penalized splines use the framework of truncated power functions (TPF). There is an unpenalized part, a polynomial in x, of the same degree

as that of the TPF. The penalty on the TPF is on the size of the coefficients, not on differences thereof. This makes analytical work easier. In Section 10, we have presented two alternative representations of P-splines that have a unpenalized polynomial part and a size penalty on the other basis functions. We believe that such are more suitable than TPF, if only because of the decoupling of the degree of the splines and the order of the penalty. Boer (2015) recently presented a variant that keeps the basis sparse.

What is neglected in papers on asymptotic theory is that often we have to deal with observation in time or space, where more observations bring about a proportional increase in the size of the domain.

In Section 4, we have shown that there is no danger in using many splines even when fitting only a few data points. Hence one is always free to use many splines and not worry about optimization of their number. We therefore advise to use 100 B-splines, a safe choice.

Boosting for smoothing basically works as follows (Tutz and Binder, 2006; Schmid and Hothorn, 2008): (1) smooth with a very strong penalty and save the result, (2) smooth the residuals and add (a fraction of) this result to the previous result, (3) repeat step (2) many times. The result gets more flexible with each iteration. So one has to stop at some point, using AIC or another criterion. Boosting has many enthusiastic proponents, and its use has been extended to non-normal data and additive models and other smooth structures (Mayr et al., 2012). We find it difficult to see its advantages, especially when we compare it to Schall's algorithm for tuning multiple smoothing parameters, which we presented in Section 10. On the other hand, boosting allows to select relevant variables in a model and the use of non-standard objective functions.

13. Computation and software

For many applications standard P-splines do not pose computational challenges. The size of the B-spline basis will be moderate and many thousands of observations can be handled with ease. If the data are observed on an equidistant grid and only smoothed values on that grid are wanted, one can just as well use the identity matrix as a basis. This leads to the Whittaker smoother (Whittaker, 1923; Eilers, 2003). The number of coefficients will be equal to the number of observation, but in combination with sparse matrix algorithms a very fast smoother is obtained.

Sparse matrices also are attractive when long data series have to be smoothed with a large B-spline basis (de Rooi et al., 2014). Even though the basis matrix is sparse, one has to take care to avoid computing dense matrices along the way, as is the case when using truncated power functions. The key is to recognize that $\boldsymbol{B}_{j+k}(\boldsymbol{x}) = \boldsymbol{B}_j(\boldsymbol{x} - s\boldsymbol{k})$, where s is the distance between the knots. Each $\boldsymbol{x}_i$ is shifted by $-s\boldsymbol{k}_i$ to a chosen subdomain , and a basis of only four (cubic) B-splines is computed on that domain. In a last step a sparse matrix is constructed with the columns of row i shifted to the right by $\boldsymbol{k}_i$. An added advantage is that numerical roundoff is minimized (Eilers and Marx, 2010).

When the penalty parameter is large, forming $\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D}$ explicitly to solve the penalized normal equations is not optimal and rounding problems can occur. It is better to use an augmented version of $\boldsymbol{B}$ yielding $\bar{\boldsymbol{B}}$, where $\bar{\boldsymbol{B}} = [\boldsymbol{B}^{\mathsf{T}}\ \sqrt{\lambda}\boldsymbol{D}^{\mathsf{T}}]^{\mathsf{T}}$, and an augmented $\boldsymbol{y}$ as $\bar{\boldsymbol{y}} = [\boldsymbol{y}^{\mathsf{T}}\ \boldsymbol{0}^{\mathsf{T}}]^{\mathsf{T}}$ and perform linear regression of $\bar{\boldsymbol{y}}$ on $\bar{\boldsymbol{B}}$ using the QR decomposition. Here $\boldsymbol{0}$ stands for a vector of zeros with length equal to the number of rows of $\boldsymbol{D}$. See Wood (2006a) for advice on stable computation.

The demands of additive models on computer memory and computation time often are modest. However, very large data sets need special treatment when they do not fit in the working memory. Wood et al. (2015) described such an application, forecasting electricity consumption in France. They developed a specialized algorithm, which is part of the the R package (`mgcv`) as the function `bam`.

In two-dimensional smoothing of large grids, using tensor products, one can run into yet another problem. The data and the one-dimensional bases may easily fit into memory, but the (inner products of) Kronecker products cannot be handled. The two-dimensional basis, $\breve{\boldsymbol{B}} \otimes \boldsymbol{B}$, has $m_1 m_2$ rows and $n_1 n_2$ columns. When smoothing a large 1000 by 1000 image using $n_1 n_2 \approx 1000$, the basis has one billion elements, taking 8 bytes each, and so will not fit into 8 Gb of main memory. Even if it would, the computation of the inner products will be extremely taxing. Note that in this case we have around 1000 coefficients, so it is not the size of the final system of penalized normal equations that is the problem. Fortunately, by rearranging the calculations, one can avoid the explicit Kronecker products and gain orders of magnitudes in computation speed and memory use (Currie et al., 2006; Eilers et al., 2006). This array algorithm, so-called GLAM, allows arbitrary weights and so is suitable for generalized linear smoothing.

When no weights are involved, even larger improvements are possible, by using the "sandwich smoother" (Xiao et al., 2013). The basic idea is that one can first apply one-dimensional smoothing to the rows of a matrix and then to the columns (or the other way around). The order is immaterial, as is easy to see from the explicit equation for $\boldsymbol{A} = (\boldsymbol{B}^{\mathsf{T}}\boldsymbol{B} + \lambda \boldsymbol{D}^{\mathsf{T}}\boldsymbol{D})^{-1}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{Y}\breve{\boldsymbol{B}}(\breve{\boldsymbol{B}}^{\mathsf{T}}\breve{\boldsymbol{B}} + \breve{\lambda}\breve{\boldsymbol{D}}^{\mathsf{T}}\breve{\boldsymbol{D}})^{-1}$, the matrix of coefficients. A similar approach was followed by Eilers and Goeman (2004), using a modified Whittaker smoother.

Nowadays it is quite common to publish computer code on a website, or as supplementary material, to accompany statistical papers. This is certainly true for the literature on P-splines. We do not try to describe these individual efforts. Instead we point to some packages for R (R Core Team, 2015) with a rather wide scope.

Originally designed for fitting generalized additive model, and accompanying Wood (2006b), `mgcv` has grown into the Swiss army knife of smoothing. It offers a diversity of basis functions and their tensor products for multidimensional smoothing. Furthermore, it can fit varying-coefficient models and signal regression, and it can mix and match components in an additive way. It offers a diversity of distributions to handle (over-dispersed) non-normal data.

We described the GAMLSS model in Section 5. A very extensive package is available. Its core is aptly called `gamlss`; it can be extended with a suite of add-ons for censored data, mixed models, and a variety of continuous and discrete distributions.

The package `MortalitySmooth` focuses on smoothing of counts in one and two dimensions (Camarda, 2012). It also is a nice source for mortality data from several countries.

BayesX (Brezger et al., 2005) is a stand-alone program for Windows and Linux. It covers all the models that fit in the generalized linear additive smooth structure (or structured additive regression) framework. The Bayesian algorithms are based on Markov chain Monte Carlo. It also offers mixed model based algorithms. There are R packages to install BayesX and to communicate with it.

It is also possible to use the R-INLA (Rue et al., 2009) package for fitting additive models with P-splines. See Fraaije et al. (2015) and the accompanying software.

The package `mboost` offers boosting for a variety of models, including P-splines and generalized additive models (Hofner et al., 2014). With the extension `gamboostLSS`, one can apply boosting to models for location, scale and shape, similar to GAMLSS.

To estimate smooth expectile curves or surfaces, the package `expectreg` is available.

14. Discussion

The paper by Eilers and Marx (1996) that started it all contained a "consumer score card", comparing various smoothing algorithms. P-splines received the best marks and their inventors concluded that they should be the smoother of choice. Two decades later, it is gratifying to see that this opinion is being shared by many statisticians and other scientists. Once prominent tools like kernel smoothers and local likelihood are gradually fading into obscurity.

In twenty years, P-spline methodology has been extended in many directions. The analogy with mixed models is being exploited to the fullest, as is the Bayesian approach, leading to new interpretations of penalties and powerful recipes for optimizing the amount of smoothing. Multidimensional smoothing with tensor products has become practical and fast, thanks to array algorithms. Regression on (multidimensional) signals has also become practical. Smooth additive structures allow the combination of various models. The key is the combination of rich B-spline regression and a simple roughness penalty. Actually the penalties are the core and many variations have been developed, while the B-spline basis did not change. We expect to see exciting developments in the near future. For a start, we sketch some aspects that we hope will get much attention.

We wrote that the penalty forms the skeleton and that the B-splines put flesh on the bones. That means that new ideas for penalties have to be developed. One promising avenue is the application to differential equations. One can write the solution as a sum of B-splines (the collocation method) and use the differential equation (DE) as the penalty (Ramsay et al., 2007). In this light the usual penalty for smoothing splines is equivalent to a differential equation that says that the second derivative of the solution is zero everywhere. O'Sullivan (1986) took the step from a continuous penalty to a discrete one. This can also be done with a DE-based penalty. However, if the coefficients of the DE are

not fixed (e.g. estimated from the data), then this generates a significant computational load. It will be useful to study (almost) equivalent discrete penalties, based on difference equations.

It is remarkable that in one-dimensional smoothing, kriging is almost absent. Altman (2000) compared splines and kriging and found that serial correlation is a key issue. If it is present and ignored, splines do not perform well. There are ways to handle correlation, as discussed in Section 10.. In spatial data analysis, kriging is still dominant. We believe that for many applications, tensor product P-splines would be a much better choice, especially if one is more interested in estimating a trend rather than doing spatial interpolation. It may appear that attempting to estimate a covariance structure from the data is a worthwhile effort, but in practice it often leads to unstable procedures. Handling non-normal data with kriging is cumbersome. In contrast, P-splines impose a relatively simple covariance structure, and in practice do the job in a very stable way, as our experiences with the analysis of agricultural field trials has shown. Smoothing of data on large grids is problematic for kriging, but P-splines and array algorithms handle such data with ease. In some cases it might even be attractive to summarize the data (as counts and sums) on a grid before analysis. Combined with the PS-ANOVA approach (Lee et al., 2013), which avoids detailed modelling of higher-order interactions, powerful tools for large data sets can be developed.

In some applications extrapolation is very important. Mortality data are a prime example. The order of the differences in the penalty determines the result: for first order differences it is a constant, for second order a straight line and a weighted combination of both gives an exponential curve. A challenge is to determine which penalty to use and to set its tuning parameter(s) for optimal extrapolation. In one dimension extrapolation does not influence the fit to the data. This is not true in two dimensions, for example with life tables. The penalties for the age and the time dimension interact and the size of the extrapolation region also has an influence. More research is needed to better understand these issues.

In several places in this paper, we have encountered the effective dimension of (components) of a model. It is an important parameter when optimizing penalties. Yet it deserves more attention on its own right. The definition, by Ye (1998), in (10) is very powerful. The contribution to $\mathrm{ED} = \sum_i \partial \hat{y}_i / \partial y_i$ by a component of an additive model can be determined clearly by following a change in y_i down the model to the coefficients, and from there back up again to the corresponding change in $\hat{y}_i$. Partial effective dimensions can be calculated this way; they are important summaries of the contributions of the model components.

In this paper, we have tried to give a glimpse of the many landmarks created in the last 20 years. It has been a collective achievement, the result of the work of many researchers who believed in the power of P-splines. We see a great and exciting future ahead, as there are many problems to solve, new complex data to model, and especially a new generation of bright statisticians who are already showing that P-splines have much more to contribute to this century, the century of data.

Appendix A. O-splines or P-splines?

Wand and Ormerod (2008) introduced O'Sullivan splines, or O-splines for short. They were not entirely pleased with the pure discrete penalty of P-splines and returned to the integral of the squared second (or higher) derivative of the fitted function. This can be attractive, especially when the knots of the B-spline basis are not evenly spaced. There are cases when this can be very valuable. As an example, Whitehorn et al. (2013) presented an example of high-dimensional smoothing with tensor products in high-energy physics to model the response of a detector. In this case, more detail was needed in the centre than near the boundaries. However, this was not the motivation of Wand and Ormerod. They rather favour the use of quantiles of x for the knots.

The paper claims that P-splines do not extrapolate well, when compared to the smoothing spline. Hence O-splines should be preferred. This claim was repeated by Ruppert et al. (2009). The paper by Wand and Ormerod (2008) has been cited more than 50 times, so apparently the message did not get lost.

We were concerned about this analysis because a basis with multiple knots at the domain boundaries had been used for the O-splines. If multiple knots had also been used for the P-splines (similar to the one in the bottom panel of Figure 3), artifacts could have occurred. So we decided to take the example data from their paper (dataset `fossil` in the R package `SemiPar`) and re-analyse them. We downloaded the file `WandOrmerod08.Rs` from Matt Wand's personal web page. For fitting, we extracted the section that invokes the function `lme` in the package `nlme` for estimating an equivalent mixed model. This program was adapted for P-splines by changing the basis and the penalty matrix. For comparison we use the function `smooth.spline` that is a standard part of R. It tunes the amount of smoothing automatically to the data, using cross-validation.

Figure 9 shows the fits of P-splines and a smoothing spline. We used 40 B-splines on the domain from 85 to 130. There is strong correspondence between the two splines. This is also true for the estimated derivative, which was approximated by taking differences.

Surprisingly, the O-splines do not work as well as P-splines, as Figure 10 shows. This especially can be seen in the derivatives. It appears that the O-spline fit struggles near $x = 100$. The reason is that the knot density is low there, because the low local data density. What is more worrying is that derivatives of the extrapolated part is not constant, as they should be for a linear result.

We believe that the anomalous behaviour of the O-splines is caused by the choice of basis. Multiple knots do not go together with a discrete penalty on the spline coefficients. The root of all evil is the choice to use quantiles of x for the knots; there is absolutely no need for it.

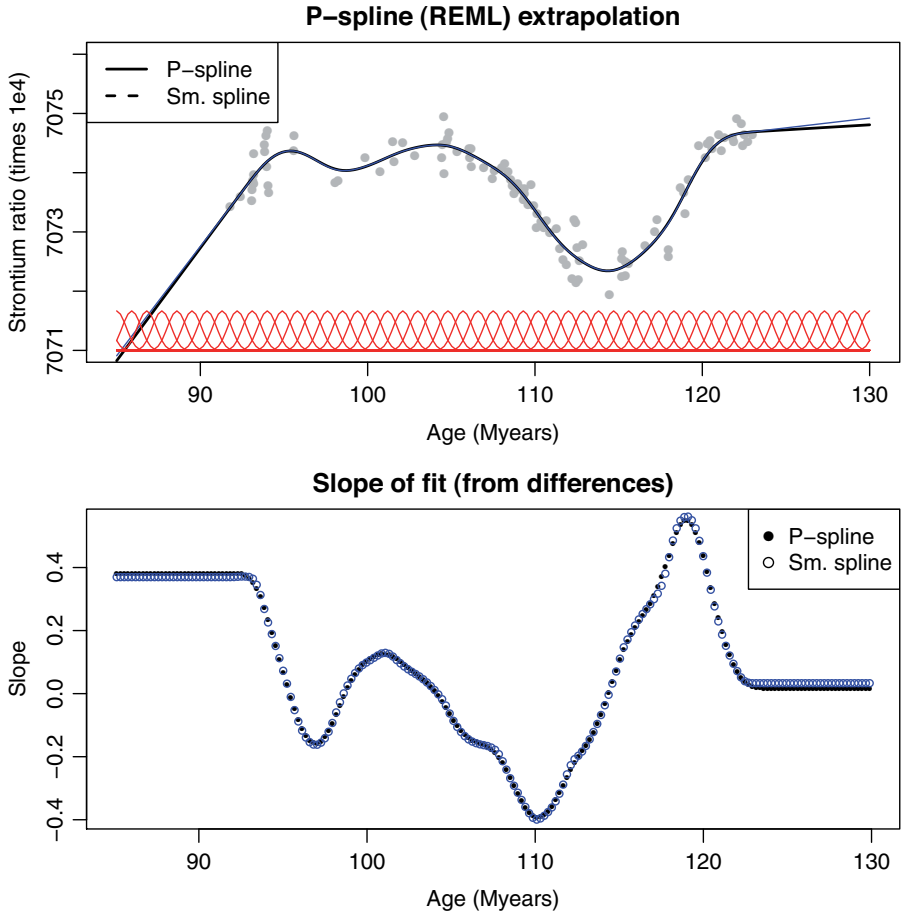

Figure 9: *Upper panel: P-spline and smoothing spline fit to the fossil data. On both sides the fit is extrapolated automatically. Lower panel: derivatives of both splines, as computed from first differences.*

Appendix B. Proof that the effective dimension is smaller than m

A formal proof starts from a simplified case, with $d = 0$ and a general basis $\boldsymbol{Z}$, where the system of equations to solve is

$$(\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \lambda \boldsymbol{I})\boldsymbol{u} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}. \tag{18}$$

The singular value decomposition of $\boldsymbol{Z}$ gives: $Z = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$, with $\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U} = \boldsymbol{I}_m$ and $\boldsymbol{V}^{\mathsf{T}}\boldsymbol{V} = \boldsymbol{I}_n$. Through substitution:

$$(\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}} + \lambda \boldsymbol{I})\boldsymbol{u} = (\boldsymbol{V}\boldsymbol{S}^2\boldsymbol{V}^{\mathsf{T}} + \lambda \boldsymbol{I})\hat{\boldsymbol{a}} = \boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y} \tag{19}$$

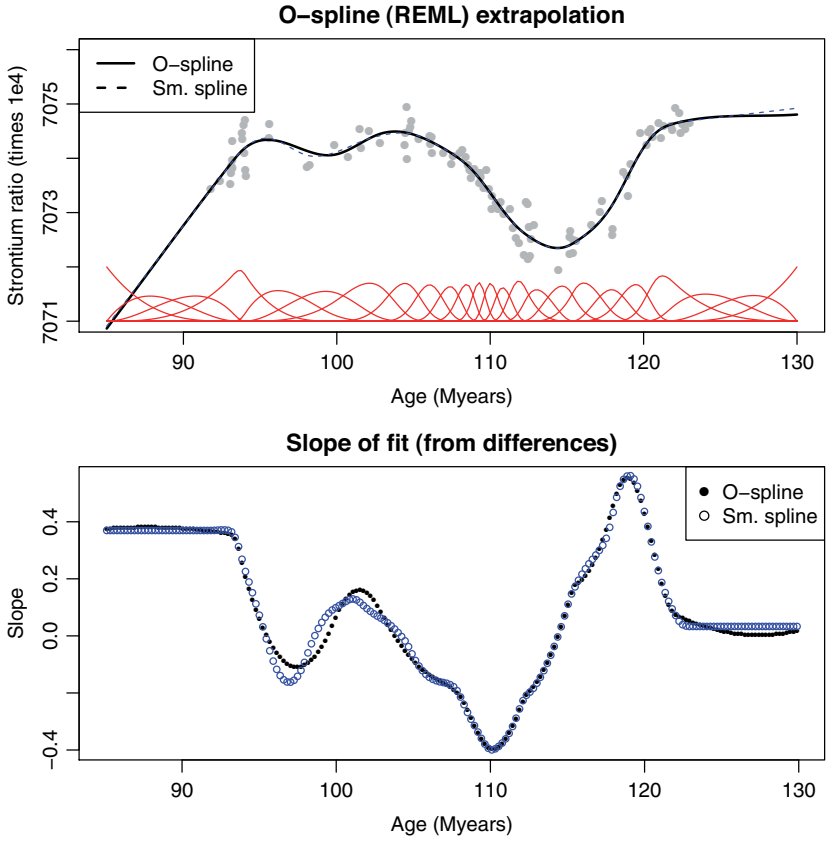

Figure 10: *Upper panel: O-spline and smoothing spline fit to the fossil data. On both sides the fit is extrapolated automatically. Lower panel: derivatives of both splines, as computed from first differences.*

Now assume that $\boldsymbol{u} = \boldsymbol{V}\boldsymbol{\gamma}$. Fill in and multiply by $\boldsymbol{V}^{\mathsf{T}}$:

$$\boldsymbol{V}^{\mathsf{T}}(\boldsymbol{V}\boldsymbol{S}^2\boldsymbol{V}^{\mathsf{T}} + \lambda \boldsymbol{I})\boldsymbol{V}\boldsymbol{\gamma} = \boldsymbol{V}^{\mathsf{T}}\boldsymbol{V}\boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}. \tag{20}$$

Hence

$$(\boldsymbol{S}^2 + \lambda \boldsymbol{I}_m)\boldsymbol{\gamma} = \boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}\boldsymbol{y}. \tag{21}$$

This is a system with m equations in m unknowns. The system matrix is diagonal and non-singular.

The penalty is a special case here, but in Section 10 it was shown that P-splines can be transformed into a mixed model, specifically with $\boldsymbol{B}\boldsymbol{a} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}$ and with a ridge penalty on $\boldsymbol{u}$. The Henderson equations (17) contain a part for $\boldsymbol{\beta}$. We now have

$$\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} + (\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z} + \lambda \boldsymbol{I})\boldsymbol{u} = \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}, \tag{22}$$

or

$$(\boldsymbol{S}^2 + \lambda \boldsymbol{I}_m)\boldsymbol{u} = \boldsymbol{S}\boldsymbol{U}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}\boldsymbol{\beta}). \tag{23}$$

Note that the value of $\boldsymbol{\beta}$ is immaterial, as it does not change the properties of the system matrix and only the right-hand side of the equation changes.

Acknowledgment

A part of the work of the first author was supported by a Chair of Excellence Grant from Carlos III University in Madrid, Spain.

References

Altman, N. (2000). Theory & methods: Krige, smooth, both or neither? *Australian & New Zealand Journal of Statistics*, 42, 441–461.

Andriyana, Y., Gijbels, I. and Verhasselt, A. (2014). P-splines quantile regression estimation in varying coefficient models. *Test*, 23, 153–194.

Antoniadis, A., Gregoire, G. and McKeague, I. (2004). Bayesian estimation in single-index models. *Statistica Sinica*, 14, 1147–1164.

Azmon, A., Faes, C. and Hens, N. (2014). On the estimation of the reproduction number based on misreported epidemic data. *Statistics In Medicine*, 33, 1176–1192.

Basile, R., Durbán, M., Minguez, R., Montero, J. and Mur, J. (2014). Modeling regional economic dynamics: spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic and Dynamics Control*, 48, 229–245.

Belitz, C. and Lang, S. (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics & Data Analysis*, 53, 61–81.

Bollaerts, K., Eilers, P. H. C. and Aerts, M. (2006). Quantile regression with monotonicity restrictions using P-splines and the l_1-norm. *Statistical Modelling*, 6, 189–207.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, 50, 967–991.

Brezger, A., Kneib, T. and Lang, S. (2005). BayesX: analysing Bayesian structured additive regression models. *Journal of Statistical Software*, 14.

Brockhaus, S., Scheipl, F., Torsten, H. and Greven, S. (2015). The functional linear array model. *Statistical Modelling*, 15, 279–300.

Brumback, B., Ruppert, D. and Wand, M. (1999). Comment on: variable selection and function estimation in additive nonparametric regression using a data-based prior. *Journal of the American Statistical Association*, 94, 794–797.

Bugli, C. and Lambert, P. (2006). Functional ANOVA with random functional effects: an application to event-related potentials modelling for electroencephalograms analysis. *Statistics in Medicine*, 25, 3718–3739.

Camarda, C. G. (2012). MortalitySmooth: an R package for smoothing Poisson counts with P-splines. *Journal of Statistical Software*, 50, 1–24.

Camarda, C. G., Eilers, P. H. C. and Gampe, J. (2008). Modelling general patterns of digit preference. *Statistical Modelling*, 8, 385–401.

Claeskens, G., Krivobokova, T. and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96, 529–544.

Coull, B., Schwartz, J. and Wand, M. (2001). Respiratory health and air pollution: additive mixed model analyses. *Biostatistics*, 2, 337–349.

Crainiceanu, C., Ruppert, D. and Carroll, R. (2007). Spatially adaptive Bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 17, 265–288.

Currie, I., Durbán, M. and Eilers, P. (2004). Smoothing and forecasting mortality rates. *Statistical Modelling*, 4, 279–298.

Currie, I. D. and Durbán, M. (2002). Flexible smoothing with p-splines: a unified approach. *Statistical Modelling*, 2, 333–349.

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

de Rooi, J. J., van der Pers, N. M., Hendrikx, R. W. A., Delhez, R., Bottger, A. J. and Eilers, P. H. C. (2014). Smoothing of X-ray diffraction data and K alpha(2) elimination using penalized likelihood and the composite link model. *Journal of Applied Crystallography*, 47, 852–860.

Dobson, A. and Barnett, A. (2008). *An Introduction to Generalized Linear Models, 3d ed.* CRC Press.

Durbán, M. and Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics*, 18, 251–262.

Durbán, M., Harezlak, J., Wand, M. and Carroll, R. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, 24, 1153–1167.

Eilers, P. (1999). Discussion on: the analysis of designed experiments and longitudinal data by using smoothing splines. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 48, 307–308.

Eilers, P. (2005). Unimodal smoothing. *Journal Of Chemometrics*, 19, 317–328.

Eilers, P. (2009). The smooth complex logarithm and quasi-periodic models. In T. Kneib and G. Tutz, editors, *Statistical Modelling and Regression Structures*. Springer.

Eilers, P. and de Menezes, R. (2005). Quantile smoothing of array CGH data. *Bioinformatics*, 21, 1146–1153.

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

Eilers, P. H. C. (2003). A perfect smoother. *Analytical Chemistry*, 75, 3631–3636.

Eilers, P. H. C. (2007). III-posed problems with counts, the composite link model and penalized likelihood. *Statistical Modelling*, 7, 239–254.

Eilers, P. H. C. and Borgdorff, M. W. (2007). Non-parametric log-concave mixtures. *Computational Statistics & Data Analysis*, 51, 5444–5451.

Eilers, P. H. C. and Goeman, J. J. (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics*, 20, 623–628.

Eilers, P. H. C. and Marx, B. D. (1992). Generalized linear models with P-splines. In *Proceedings of GLIM 92 and the 7th International Workshop on Statistical Modelling*.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11, 89–102.

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

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. (2010). Splines, knots and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 637–653.

Eilers, P. H. C., Gampe, J., Marx, B. D. and Rau, R. (2008). Modulation models for seasonal time series and incidence tables. *Statistics In Medicine*, 27, 3430–3441.

Eilers, P. H. C., Li, B. and Marx, B. D. (2009). Multivariate calibration with single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, 96, 196–202.

Fahrmeir, L. and Kneib, T. (2009). Propriety of posteriors in structured additive regression models: Theory and empirical evidence. *Journal of Statistical Planning and Inference*, 139, 843–859.

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

Fahrmeir, L., Kneib, T. and Konrath, S. (2010). Bayesian regularization in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20, 203–219.

Fraaije, R. G. A., ter Braak, C. J. F., Verduyn, B., Breeman, L. B. S., Verhoeven, J. T. A. and Soons, M. B. (2015). Early plant recruitment stages set the template for the development of vegetation patterns along a hydrological gradient. *Functional Ecology*, 29, 971–980.

Frasso and Eilers (2015). L- and V-curves for optimal smoothing. *Statistical Modelling*, 15, 91–111.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *International Statistical Review/Revue Internationale de Statistique*, 245–259.

Greven, S. (2008). *Non-Standard Problems in Inference for Additive and Linear Mixed Models*. Cuvillier Verlag.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer.

Hansen, P. C. (1992). Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Review*, 34, 561–580.

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R. and Christiani, D. C. (2007). Penalized solutions to functional regression problems. *Computational Statistics & Data Analysis*, 51, 4911–4925.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72, 320–338.

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

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

Heim, S., Fahrmeir, L., Eilers, P. H. C. and Marx, B. D. (2007). 3D space-varying coefficient models with application to diffusion tensor imaging. *Computational Statistics & Data Analysis*, 51, 6212–6228.

Henderson, C. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423–447.

Hofner, B., Mayr, A., Robinzonov, N. and Schmid, M. (2014). Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational Statistics*, 29, 3–35.

Jarrow, R., Ruppert, D. and Yu, Y. (2004). Estimating the interest rate term structure of corporate debt with a semiparametric penalized spline model. *Journal of the Americal Statistical Association*, 99, 57–66.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics & Data Analysis*, 51, 2542–2558.

Kammann, E. and Wand, M. (2003). Geoadditive models. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 52, 1–18.

Kauermann, G. (2005a). A note on smoothing parameter selection for penalized spline smoothing. *Journal of statistical planning and inference*, 127, 53–69.

Kauermann, G. (2005b). Penalized spline smoothing in multivariable survival models with varying coefficients. *Computational Statistics & Data Analysis*, 49, 169–186.

Kauermann, G. and Khomski, P. (2006). Additive two-way hazards model with varying coefficients. *Computational Statistics & Data Analysis*, 51, 1944–1956.

Kauermann, G., Krivobokova, T. and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 71, 487–503.

Kauermann, G., Krivobokova, T. and Semmler, W. (2011). Filtering time series with penalized splines. *Studies in Nonlinear Dynamic & Econometrics*, 15, 1–26.

Kauermann, G., Schellhase, C. and Ruppert, D. (2013). Flexible copula density estimation with penalized hierarchical B-splines. *Scandinavian Journal of Statistics*, 40, 685–705.

Kneib, T. and Fahrmeir, L. (2006). Structured additive regression for categorical space-time data: A mixed model approach. *Biometrics*, 62, 109–118.

Kneib, T., Konrath, S. and Fahrmeir, L. (2011). High dimensional structured additive regression models: Bayesian regularization, smoothing and predictive performance. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 60, 51–70.

Kohn, R., Ansley, C. and Tharm, D. (1991). The performance of cross-validation and maximum likelihood estimators of spline smoothing parameters. *Journal of the American Statistical Association*, 86, 1042–1050.

Krivobokova, T. and Kauermann, G. (2007). A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, 102, 1328–1337.

Krivobokova, T., Kauermann, G. and Archontakis, T. (2006). Estimating the term structure of interest rates using penalized splines. *Statistical Papers*, 47, 443–459.

Lambert, P. (2011). Smooth semiparametric and nonparametric Bayesian estimation of bivariate densities from bivariate histogram data. *Computational Statistics & Data Analysis*, 55, 429–445.

Lambert, P. and Eilers, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics & Data Analysis*, 53, 1388–1399.

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

Lee, D.-J. and Durbán, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics & Data Analysis*, 53, 2968–2979.

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

Lee, D.-J., Durbán, M. and Eilers, P. (2013). Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Computational Statistics & Data Analysis*, 61, 22–37.

Leitenstorfer, F. and Tutz, G. (2011). Estimation of single-index models based on boosting techniques. *Statistical Modelling*, 11, 203–217.

Li, B. and Marx, B. D. (2008). Sharpening P-spline signal regression. *Statistical Modelling*, 8, 367–383.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95, 415–436.

Lu, X., Chen, G., Singh, R. and Song, P. (2006). A class of partially linear single-index survival models. *Canadian Journal of Statistics*, 34, 97–112.

Lu, Y., Zhang, R. and Zhu, L. (2008). Penalized spline estimation for varying-coefficient models. *Communications in Statistics-Theory and Methods*, 37, 2249–2261.

Malfait, N. and Ramsay, J. (2003). The historical functional linear model. *Canadian Journal of Statistics*, 31, 185–201.

Marx, B. D. (2010). P-spline varying coefficient models for complex data. In G. Tutz and T. Kneib, editors, *Statistical Modelling and Regression Structures*, 19–43. Springer.

Marx, B. D. (2015). Varying-coefficient single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, 143, 111–121.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28, 193–209.

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.

Marx, B. D. and Eilers, P. H. C. (2005). Multidimensional penalized signal regression. *Technometrics*, 47, 13–22.

Marx, B. D., Eilers, P. H. C., Gampe, J. and Rau, R. (2010). Bilinear modulation models for seasonal tables of counts. *Statistics and Computing*, 20, 191–202.

Marx, B. D., Eilers, P. H. C. and Li, B. (2011). Multidimensional single-index signal regression. *Chemometrics and Intelligent Laboratory Systems*, 109, 120–130.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data. A flexible approach based on boosting. *Journal of the Royal Statistical Society Series C*, 61, 403–427.

McLean, M. W., Hooker, G., Staicu, A.-M., Scheipl, F. and Ruppert, D. (2014). Functional Generalized Additive Models. *Journal of Computational and Graphical Statistics*, 23, 249–269.

Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and its Application*, 2, 321–359.

Myers, R. (1989). *Classic and Modern Regression with Applications*. PWS-KENT.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science*, 1, 505–527.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58, 545–554.

Pya, N. and Wood, S. (2015). Shape constrained additive models. *Statistics and Computing*, 25, 543–559.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. and Silverman, B. (2003). *Functional Data Analysis, 2nd ed.* Springer.

Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69, 741–796.

Rigby, R. and Stasinopoulos, D. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Computational Statistics & Applied Statistics*, 54, 507–544.

Rippe, R. C. A., Meulman, J. J. and Eilers, P. H. C. (2012a). Reliable single chip genotyping with semi-parametric log-concave mixtures. *Plos One*, 7.

Rippe, R. C. A., Meulman, J. J. and Eilers, P. H. C. (2012b). Visualization of genomic changes by segmented smoothing using an L_0 penalty. *PLoS ONE*, 7.

Rizzi, S., Gampe, J. and Eilers, P. H. C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*, 182, 138–147.

Rodriguez-Alvarez, M., Lee, D., Kneib, T., Durbán, M. and Eilers, P. (2015). Fast smoothing parameter separation in multidimensional generalized P-splines: the SAP algorithm. *Statistics and Computing*, 25, 941–957.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71, 319–392.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11, 735–757.

Ruppert, D. and Carroll, R. (2000). Spatially-adaptive penalties for spline fitting. *Australian & New Zealand Journal Of Statistics*, 42, 205–223.

Ruppert, D., Wand, M. and Carroll, R. (2003). *Semiparametric Regression*. Cambridge.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2009). Semiparametric regression during 2003-2007. *Electronic Journal of Statistics*, 3, 1193–1256.

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

Scheipl, F., Staicu, A. and Greven, S. (2015). Functional additive mixed models. *Journal of Computational and Graphical Statistics*, 24, 477–501.

Schellhase, C. and Kauermann, G. (2012). Density estimation and comparison with a penalized mixture approach. *Computational Statistics*, 27, 757–777.

Schmid, M. and Hothorn, T. (2008). Boosting additive models using component-wise P-Splines. *Computational Statistics & Data Analysis*, 53, 298–311.

Schnabel, S. K. and Eilers, P. H. C. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53, 4168–4177.

Schnabel, S. K. and Eilers, P. H. C. (2013). Simultaneous estimation of quantile curves using quantile sheets. *ASTA-Advances In Statistical Analysis*, 97, 77–87.

Sobotka, F. and Kneib, T. (2012). Geoadditive expectile regression. *Computational Statistics & Data Analysis*, 56, 755–767.

Sobotka, F., Kauermann, G., Waltrup, L. S. and Kneib, T. (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, 23, 135–148.

Speed, T. (1991). Comment on: that BLUP is a good thing: the estimation of random effects. *Statistical Science*, 6, 15–51.

Thompson, R. and Baker, R. (1981). Composite link functions in generalized linear models. *Applied Statistics*, 30, 125–131.

Tutz, G. and Binder, H. (2006). Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*, 62, 961–971.

Verbyla, A., Cullis, B., Kenward, M. and Welham, S. (1999). The analysis of designed experiments and longitudinal data using smoothing splines. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 48, 269–312.

Wand, M. (2003). Smoothing and mixed models. *Computational Statistics*, 18, 223–249.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal Of Statistics*, 50, 179–198.

Wang, X., Shen, J. and Ruppert, D. (2011). On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, 5, 1–17.

Wang, X.-F., Hu, B., Wang, B. and Fang, K. (2014). Bayesian generalized varying coefficient models for longitudinal proportional data with errors-in-covariates. *Journal of Applied Statistics*, 41, 1342–1357.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance. *Journal of the Royal Statistical Society. Series B-Statistical Methodology*, 60, 159–174.

Whitehorn, N., van Santen, J. and Lafebre, S. (2013). Penalized Splines for smooth representation of high-dimensional Monte Carlo datasets. *Computer Physics Communications*, 184, 2214–2220.

Whittaker, E. (1923). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.

Wood, S. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, 99, 673–686.

Wood, S. (2006a). *Generalized Additive Models: An Introduction with R*. CRC Press.

Wood, S. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semi-parametric generalized linear models. *Journal of the Royal Statistical Society. Series B-Statistical Methodology*, 73, 3–36.

Wood, S., Scheipl, F. and Faraway, J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 23, 341–360.

Wood, S. N. (2006b). On confidence intervals for generalized additive models based on penalized regression splines. *Australian & New Zealand Journal of Statistics*, 48, 445–464.

Wood, S. N., Goude, Y. and Shaw, S. (2015). Generalized additive models for large data sets. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 64, 139–155.

Xiao, L., Li, Y. and Ruppert, D. (2013). Fast bivariate P-splines: the sandwich smoother. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 75, 577–599.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection . *Journal of the American Statistical Association*, 93, 120–131.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the Americal Statistical Association*, 97, 1042–1054.

Zhang, D., Lin, X., Raz, J. and Sowers, M. (1998). Semiparametric stochastic mixed models for longitudinal data. *Journal of the Americal Statistical Association*, 93, 710–719.