General P-Splines for Non-Uniform B-Splines

Zheyuan Li[1] and Jiguo Cao[2*]

[1]School of Mathematics and Statistics, Henan University, Kaifeng, Henan, China.
[2]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada.

*Corresponding author(s). E-mail(s): jiguo_cao@sfu.ca;
Contributing authors: zheyuan.li@vip.henu.edu.cn;

Abstract

We propose a new type of penalized B-splines called general P-splines (GPS) to accommodate non-uniform B-splines on unevenly spaced knots. They feature a novel general difference penalty that accounts for uneven knot spacing, complementing Eilers and Marx's standard P-splines (SPS) tailored for uniform B-splines on equidistant knots. In simulation studies, GPS are as good as SPS in terms of mean squared error (MSE) performance if unevenly spaced knots are not key to uniform approximation of a smooth function, but are much better otherwise. In a practical example with BMC longitudinal data, GPS on a few unevenly spaced knots are more satisfactory than SPS on many equidistant knots, by producing smoother trajectories with no ripples and being faster to compute. GPS are also closely related to O'Sullivan splines or O-splines (OS) through a newly derived sandwich formula that links general difference penalty to derivative penalty. Both penalties are equally powerful for wiggliness control, yet the difference penalty is easier to interpret and compute. In simulation studies, GPS are as good as or even slightly better than OS in terms of MSE performance. In short, GPS are good replacements of OS and more flexible than SPS, hence deserve greater attention. We accompany this paper by two new **R** packages that implement GPS: **gps** and **gps.mgcv**. **R** code is available in supplementary material.

Keywords: derivative penalty, general difference penalty, O-splines, penalized regression splines, sandwich formula, unevenly spaced knots

1 Introduction

P-splines (Eilers and Marx, 1996; Eilers et al, 2015; Eilers and Marx, 2021) are popular for estimating unknown univariate smooth functions. They have been applied in many statistical modeling frameworks, like generalized additive models (Spiegel et al, 2019), single-index models (Wang et al, 2018), generalized partially linear single-index models (Yu et al, 2017), functional mixed-effects models (Chen et al, 2018), survival models (Orbe and Virto, 2021; Bremhorst and Lambert, 2016), models for longitudinal data (Koehler et al, 2017; Andrinopoulou et al, 2018), additive quantile regression models (Muggeo et al, 2021), varying coefficient models (Hendrickx et al, 2018), quantile varying coefficient models (Gijbels et al, 2018), spatial models (Greco et al, 2018; Rodriguez-Alvarez et al, 2018), spatiotemporal models (Minguez et al, 2020; Goicoa et al, 2019) and spatiotemporal quantile/expectile regression models (Franco-Villoria et al, 2019; Spiegel et al, 2020).

P-splines represent a function as a linear combination of B-splines (de Boor, 2001) and penalize their coefficients with a difference penalty to prevent overfitting. Although B-splines can be constructed on arbitrary knot sequences, P-splines are bundled with uniform B-splines (UBS) on equidistant knots. As a result, to handle non-uniform B-splines (NUBS) on unevenly spaced knots, O-splines (O'Sullivan, 1986) that penalize B-splines with a derivative penalty are hitherto the only choice.

Equidistant knots are easy to implement, but in certain cases, unevenly spaced knots are more satisfactory. For example, Maturana-Russel and Meyer (2021) demonstrated that when estimating spectral densities with spikes, O-splines on a few unevenly spaced knots outperform P-splines on a large number of equidistant knots, in terms of both goodness of fit and computation time. As the motivating example of this paper, Figure 1 shows that when smoothing annually measured bone mineral content (BMC, in grams) from 112 subjects, collected in the Pediatric Bone Mineral Accrual Study (Bailey, 1997), trajectories estimated by P-splines on equidistant knots are suspiciously wiggly, whereas trajectories estimated by O-splines on unevenly spaced knots are plausibly smooth (see Section 4.1 for details of BMC data modeling).

The primary aim of this paper is to generalize P-splines so that they can accommodate NUBS. The newly developed P-splines will be termed general P-splines (GPS). By contrast, the original P-splines will hereafter be referred to as standard P-splines (SPS). See Table 1 for a summary of the capability of all relevant penalized B-splines.

This paper has four contributions. Firstly, we propose and justify GPS. This is not a trivial task. Although it has been conjectured (Eilers and Marx, 1996; Wood, 2017a; Perperoglou et al, 2019; Maturana-Russel and Meyer, 2021) that when handling NUBS, a difference penalty should be calculated using weighed differences, possibly divided differences, it is not intuitive at all how to do this because the number of B-splines coefficients does not equal the number of knots. We are the first to formulate the procedure, which is, in fact, not doing divided differences. Secondly, we find a simple interpretation

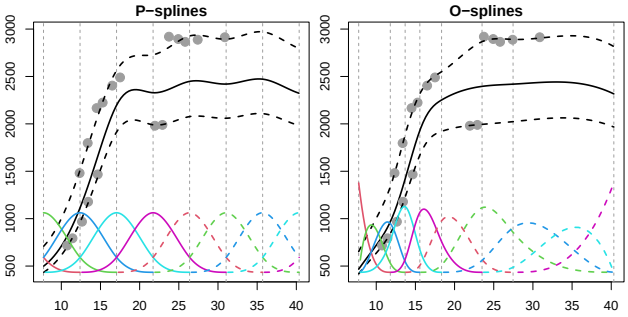

Fig. 1 Estimated population BMC trajectory (solid line) and subject BMC trajectories (dashed lines) for two selected subjects (gray dots). Left: cubic P-splines with 10 UBS (colored lines) on equidistant knots (vertical dashed lines). Right: cubic O-splines with 10 NUBS (colored lines) on unevenly spaced knots (vertical dashed lines).

Table 1 Capability of several penalized B-splines

	derivative penalty	difference penalty
uniform B-splines (UBS) (on equidistant knots)	O-splines (OS)	standard P-splines (SPS) general P-splines (GPS)
non-uniform B-splines (NUBS) (on unevenly spaced knots)	O-splines (OS)	general P-splines (GPS)

for difference penalties. For example, a 2nd order difference penalty is just the sum of squared B-splines coefficients in the 2nd derivative of a spline. This is actually more interpretable than a 2nd order derivative penalty, which stands for the integrated squared 2nd derivative of a spline. Thirdly, we derive a sandwich formula that establishes a one-to-one correspondence between a difference penalty and a derivative penalty. This enhances our understanding of the relation between the two types of penalties, which is very lacking in the literature. Last but not least, we demonstrate through simulation studies that GPS are no inferior to SPS or OS in terms of mean squared error (MSE) performance. On one hand, in all simulated examples, GPS are as good as or even slightly better than OS. On the other hand, when compared with SPS, GPS perform equally well if unevenly spaced knots are not crucial but perform much better otherwise. In summary, GPS are more flexible than SPS, being able to accommodate unevenly spaced knots. They are also good replacements of OS, for the simplicity in the interpretation and computation of difference penalties.

2 The New General P-Splines

A smoothing model for observations $(x_i, y_i)_1^n$ hypothesizes $y_i = f(x_i) + e_i$, where $f(x)$ is a smooth function estimator and e_i is an iid Gaussian error.

A penalized B-splines estimator expresses $f(x) = \sum_{j=1}^{p} B_j(x)\beta_j$ using pre-defined B-splines $(B_j(x))_1^p$. The unknown coefficients $(\beta_j)_1^p$ are then estimated by minimizing a penalized least squares objective:

$$\sum_{i=1}^{n} \left[y_i - \sum_{j=1}^{p} B_j(x_i)\beta_j\right]^2 + \lambda \cdot \text{PEN}(\boldsymbol{\beta}),$$

where the penalty $\text{PEN}(\boldsymbol{\beta})$ is a wiggliness measure for $f(x)$. Several types of penalized B-splines exist, like OS, SPS and GPS listed in Table 1, and they are each characterized by a particular form of $\text{PEN}(\boldsymbol{\beta})$. The non-negative smoothing parameter λ trades off $f(x)$'s closeness to data for $f(x)$'s smoothness and plays a critical role. It is common practice to choose its optimal value that minimizes the leave-one-out generalized cross-validation (GCV) error (Wahba, 1990), although other criteria like maximizing the restricted log-likelihood also exist (Wood, 2017a). Now to motivate GPS, let's first explain why SPS fail for NUBS.

2.1 Standard P-Splines

The penalty for SPS, hereafter called the standard difference penalty, is the sum of squared order-m differences of nearby B-splines coefficients. To be precise, for $m = 1, 2, 3$, we have:

$$\text{PEN}^{(1)}_{\text{sps}}(\boldsymbol{\beta}) = \sum_{j=1}^{p-1}(\beta_{j+1} - \beta_j)^2,\ \text{PEN}^{(2)}_{\text{sps}}(\boldsymbol{\beta}) = \sum_{j=1}^{p-2}(\beta_{j+2} - 2\beta_{j+1} + \beta_j)^2,$$
$$\text{PEN}^{(3)}_{\text{sps}}(\boldsymbol{\beta}) = \sum_{j=1}^{p-3}(\beta_{j+3} - 3\beta_{j+2} + 3\beta_{j+1} - \beta_j)^3.$$

Such penalty is tailored for UBS on equidistant knots, and does not make sense for NUBS on unevenly spaced knots. To confirm this, consider fitting standard cubic P-splines with a 2nd order penalty to 500 noisy observations of a U-shaped curve $y = x^3 - 0.2x^4$, where x-values are simulated from $\text{N}(0, 1)$ distribution. To construct UBS, we place 50 equidistant knots through the range of x-values. To construct NUBS, we place 50 knots at equal quantiles of x-values. Figure 2 shows that the fit under NUBS representation is rather wiggly at the valley, while the fit under UBS representation is as smooth as it should be. The boxplot of MSEs based on 100 simulations reassures that this phenomenon is persistent. In short, standard difference penalty fails to perform roughness control when handling NUBS on unevenly spaced knots.

There may be many explanations for such failure, but we focus on a particular flaw of standard difference penalty by investigating the limiting behavior of $f(x)$ at $\lambda = +\infty$. In this case, $\text{PEN}^{(m)}_{\text{sps}}(\boldsymbol{\beta}) = 0$ and the fitted P-splines lie in the penalty's null space. Let's take $m = 2$ as an example and fit SPS set up with non-uniform cubic B-splines to noisy observations from $y = x$, $x \in [0, 1]$. To examine effects of knot locations, we attempt unevenly spaced knots that follow different distributions (primarily Beta distributions with varying shape parameters). Figure 3 shows that the limiting fit at $\lambda = +\infty$ is sensitive to knot locations. Since unevenly spaced knots can be anywhere in general, this

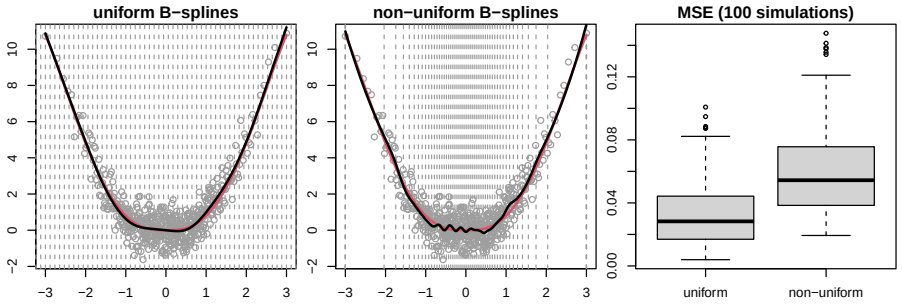

Fig. 2 SPS are tailored for UBS on equidistant knots, and do not make sense for NUBS on unevenly spaced knots.

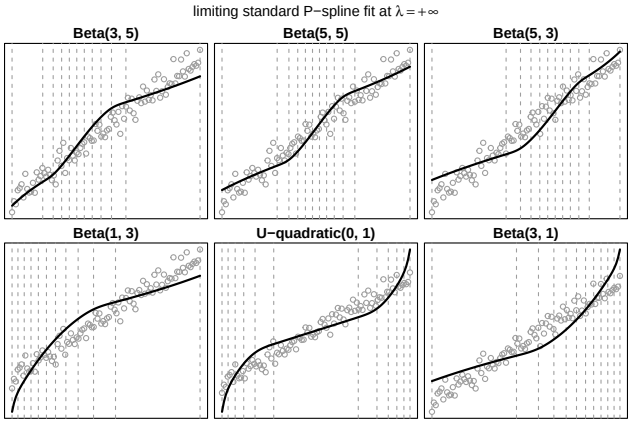

Fig. 3 Standard difference penalty is not a reliable wiggliness measure for splines represented by NUBS, because the limiting SPS fit (solid line) at $\lambda = +\infty$ is inconsistent when knot locations (vertical dashed lines) vary.

limiting fit is utterly unpredictable. Hence, standard difference penalty is not a reliable roughness measure when handling NUBS and will fail to control $f(x)$'s wiggliness as λ varies on $(0, +\infty)$.

To overcome this defect, standard difference penalty must be upgraded to take uneven knot spacing into account. We call the resulting new penalty the general difference penalty, which characterizes our new GPS. To make construction of general difference penalty as transparent as possible, we present Section 2.2 as a tutorial by working through a small example and postpone its mathematical derivation to Section 2.3. As we begin to work on knot sequences, we strongly advise readers to refer to the Appendix from time to time, which clarifies the order (d), domain ($[a, b]$) and knots (k, K) of spline and B-splines.

2.2 Introducing General P-Splines

Let

$$\boldsymbol{\Delta} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}$$

be a fundamental difference matrix whose exact dimension can be inferred from the context, then the 1st, 2nd and 3rd order standard differences of $\boldsymbol{\beta}_0 = \boldsymbol{\beta} = (\beta_j)_1^p$ can be iteratively calculated as $\boldsymbol{\beta}_1 = \boldsymbol{\Delta}\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_2 = \boldsymbol{\Delta}\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_3 = \boldsymbol{\Delta}\boldsymbol{\beta}_2$. Calculating general differences is no more complicated than an extra weighting process at each step. That is, instead of $\boldsymbol{\beta}_m = \boldsymbol{\Delta}\boldsymbol{\beta}_{m-1}$, we compute $\boldsymbol{\beta}_m = \boldsymbol{W}_m^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_{m-1}$ for some diagonal weight matrix $\boldsymbol{W}_m$. The key issue is just how to obtain $\boldsymbol{W}_m$ and we will make this crystal clear through the following example.

Consider non-uniform cubic B-splines ($d = 4$) on domain $[a, b] = [0, 4]$ with $k = 2$ interior knots $s_1 = 1$, $s_2 = 3$ and clamped boundary knots. In this case, there will be $p = k + d = 6$ B-splines and $K = p + d = 10$ knots. Coefficient vector $\boldsymbol{\beta}$ (or $\boldsymbol{\beta}_0$) has 6 elements and the complete knot sequence is:

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
0	0	0	0	1	3	4	4	4	4

Now step by step, we show how to calculate $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$. Specifically, at each step, we identify a set of "active" knots and compute their differences at a correct lag.

(1) Lag-3 differences of the following "active" knots, divided by the lag, give the diagonal elements of $\boldsymbol{W}_1$:

~~t_1~~	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	~~t_{10}~~
~~0~~	0	0	0	1	3	4	4	4	~~4~~

$$\boldsymbol{W}_1 = \frac{1}{3}\begin{bmatrix} t_5 - t_2 & & & & & & \\ & t_6 - t_3 & & & & & \\ & & t_7 - t_4 & & & & \\ & & & t_8 - t_5 & & & \\ & & & & t_9 - t_6 & & \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & & & & & \\ & \frac{3}{3} & & & & \\ & & \frac{4}{3} & & & \\ & & & \frac{3}{3} & & \\ & & & & \frac{1}{3}. & \end{bmatrix}.$$

(2) Lag-2 differences of the following "active" knots, divided by the lag, give the diagonal elements of $\boldsymbol{W}_2$:

~~t_1~~	~~t_2~~	t_3	t_4	t_5	t_6	t_7	t_8	~~t_9~~	~~t_{10}~~
~~0~~	~~0~~	0	0	1	3	4	4	~~4~~	~~4~~

$$\boldsymbol{W}_2 = \frac{1}{2}\begin{bmatrix} t_5 - t_3 & & & & \\ & t_6 - t_4 & & & \\ & & t_7 - t_5 & & \\ & & & t_8 - t_6 & \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & & & \\ & \frac{3}{2} & & \\ & & \frac{3}{2} & \\ & & & \frac{1}{2} \end{bmatrix}.$$

(3) Lag-1 differences of the following "active" knots, divided by the lag, give the diagonal elements of $\boldsymbol{W}_3$:

~~t_1~~	~~t_2~~	~~t_3~~	t_4	t_5	t_6	t_7	~~t_8~~	~~t_9~~	~~t_{10}~~
~~0~~	~~0~~	~~0~~	0	1	3	4	~~4~~	~~4~~	~~4~~

$$\boldsymbol{W}_2 = \frac{1}{1}\begin{bmatrix} t_5 - t_4 & & \\ & t_6 - t_5 & \\ & & t_7 - t_6 \end{bmatrix} = \begin{bmatrix} 1 & & \\ & 2 & \\ & & 1 \end{bmatrix}.$$

We now reach the end of the iteration because there are no lag-0 differences for calculating $\boldsymbol{W}_4$. As a result, $\boldsymbol{\beta}_1 = \boldsymbol{W}_1^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_2 = \boldsymbol{W}_2^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_3 = \boldsymbol{W}_3^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_2$. By expanding the recursion, we observe:

$$\boldsymbol{\beta}_1 = \overbrace{\boldsymbol{W}_1^{-1}\boldsymbol{\Delta}}^{\boldsymbol{D}_1}\boldsymbol{\beta}_0,\ \boldsymbol{\beta}_2 = \boldsymbol{W}_2^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_1 = \overbrace{\boldsymbol{W}_2^{-1}\boldsymbol{\Delta}\boldsymbol{W}_1^{-1}\boldsymbol{\Delta}}^{\boldsymbol{D}_2}\boldsymbol{\beta}_0,$$
$$\boldsymbol{\beta}_3 = \boldsymbol{W}_3^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_2 = \overbrace{\boldsymbol{W}_3^{-1}\boldsymbol{\Delta}\boldsymbol{W}_2^{-1}\boldsymbol{\Delta}\boldsymbol{W}_1^{-1}\boldsymbol{\Delta}}^{\boldsymbol{D}_3}\boldsymbol{\beta}_0.$$

We call $\boldsymbol{D}_1$, $\boldsymbol{D}_2$ and $\boldsymbol{D}_3$ the 1st, 2nd and 3rd order general difference matrices, and we can work them out:

$$\begin{bmatrix} -3 & 3 & & & & \\ & -1 & 1 & & & \\ & & -\frac{3}{4} & \frac{3}{4} & & \\ & & & -1 & 1 & \\ & & & & -3 & 3 \end{bmatrix}, \begin{bmatrix} 6 & -8 & 2 & & & \\ & \frac{2}{3} & -\frac{7}{6} & \frac{1}{2} & & \\ & & \frac{1}{2} & -\frac{7}{6} & \frac{2}{3} & \\ & & & 2 & -8 & 6 \end{bmatrix}, \begin{bmatrix} -6 & \frac{26}{3} & -\frac{19}{6} & \frac{1}{2} & & \\ & -\frac{1}{3} & \frac{5}{6} & -\frac{5}{6} & \frac{1}{3} & \\ & & -\frac{1}{2} & \frac{19}{6} & -\frac{26}{3} & 6 \end{bmatrix}.$$

Meanwhile, deliberately setting $\boldsymbol{W}_1$, $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ to identity matrices yields standard difference matrices of order 1 to 3:

$$\begin{bmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & -1 & 1 & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{bmatrix}, \begin{bmatrix} 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & 1 & -2 & 1 & \\ & & & 1 & -2 & 1 \end{bmatrix}, \begin{bmatrix} -1 & 3 & -3 & 1 & & \\ & -1 & 3 & -3 & 1 & \\ & & -1 & 3 & -3 & 1 \end{bmatrix},$$

and we denote them by $\boldsymbol{\Delta}_1$, $\boldsymbol{\Delta}_2$ and $\boldsymbol{\Delta}_3$.

It is easy to verify that standard difference penalty can be concisely expressed as:

$$\mathrm{PEN}_{\mathrm{sps}}^{(m)}(\boldsymbol{\beta}) = \|\boldsymbol{\Delta}_m\boldsymbol{\beta}\|^2.$$

Thus, SPS minimize:

$$\textstyle\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{p} B_j(x)\beta_j\right]^2 + \lambda\|\boldsymbol{\Delta}_m\boldsymbol{\beta}\|^2.$$

In an analogy, we define the following general difference penalty:

$$\mathrm{PEN}_{\mathrm{gps}}^{(m)}(\boldsymbol{\beta}) = \|\boldsymbol{D}_m\boldsymbol{\beta}\|^2.$$

and consequently, GPS minimize:

$$\textstyle\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{p} B_j(x)\beta_j\right]^2 + \lambda\|\boldsymbol{D}_m\boldsymbol{\beta}\|^2.$$

The two objectives differ apparently only in the use of difference matrix, but SPS are tailored for UBS on equidistant knots, whereas GPS are well defined

for NUBS on unevenly spaced knots. In addition, $\boldsymbol{D}_m$ and $\boldsymbol{\Delta}_m$ have the same band sparsity, but the latter has constant diagonals, while the former does not due to the weighting procedure. Had knots been equidistant, all weight matrices would be proportional to identity matrices, so that $\boldsymbol{D}_m$ would be proportional to $\boldsymbol{\Delta}_m$, establishing the equivalence between two estimators. In short, GPS comprise SPS as special cases, and can handle B-splines on any knot sequences.

2.3 Justifying General P-Splines

We now justify the construction of general difference penalty exemplified in the previous section and interpret this new penalty. In general, on a full knot sequence $(t_j)_1^K$, we begin with $(K-1)$ order-1 B-splines:

$$B_{j,1}(x) = \begin{cases} 1 & x \in [t_j, t_{j+1}] \\ 0 & \text{otherwise,} \end{cases}$$

and iteratively construct higher-order B-splines using:

$$B_{j,d}(x) = (x - t_j)\frac{B_{j,d-1}(x)}{t_{j+d-1}-t_j} + (t_{j+d} - x)\frac{B_{j+1,d-1}(x)}{t_{j+d}-t_{j+1}}.$$

Any order-d spline $f(x)$ can be expressed as a linear combination of B-splines of the same order:

$$f(x) = \sum_{j=1}^{p} B_{j,d}(x)\beta_{j,0},$$

so its 1st derivative is:

$$f'(x) = \sum_{j=1}^{p} B'_{j,d}(x)\beta_{j,0}.$$

According to de Boor (2001), it holds that:

$$B'_{j,d}(x) = (d-1)\left(\frac{B_{j,d-1}(x)}{t_{j+d-1}-t_j} - \frac{B_{j+1,d-1}(x)}{t_{j+d}-t_{j+1}}\right). \tag{1}$$

Therefore, the 1st derivative becomes:

$$\frac{f'(x)}{d-1} = \sum_{j=1}^{p} \frac{B_{j,d-1}(x)}{t_{j+d-1}-t_j}\beta_{j,0} - \sum_{j=1}^{p} \frac{B_{j+1,d-1}(x)}{t_{j+d}-t_{j+1}}\beta_{j,0}.$$

The first term $B_{1,d-1}(x)$ in the first summation and the last term $B_{p+1,d-1}(x)$ in the second summation turn out to be zeros on the domain, so we can show:

$$\frac{f'(x)}{d-1} = \sum_{j=1}^{p-1} B_{j+1,d-1}(x)\frac{\beta_{j+1,0}-\beta_{j,0}}{t_{j+d}-t_{j+1}},$$

which can also be expressed as:

$$\beta_{j+1,1} = \frac{\beta_{j+1,0}-\beta_{j,0}}{(t_{j+d}-t_{j+1})/(d-1)},\ f'(x) = \sum_{j=1}^{p-1} B_{j+1,d-1}(x)\beta_{j+1,1}. \tag{2}$$

To find the 2nd derivative, we further differentiate $f'(x)$ in (2):

$$f''(x) = \sum_{j=1}^{p-1} B'_{j+1,d-1}(x)\beta_{j+1,1}.$$

Substituting j with $j+1$ and d with $d-1$ in (1) gives:

$$B'_{j+1,d-1}(x) = (d-2)\left(\frac{B_{j+1,d-2}(x)}{t_{j+d-1}-t_{j+1}} - \frac{B_{j+2,d-2}(x)}{t_{j+d}-t_{j+2}}\right).$$

Then, we can show:

$$\frac{f''(x)}{d-2} = \sum_{j=1}^{p-2} B_{j+2,d-2}(x)\frac{\beta_{j+2,1}-\beta_{j+1,1}}{t_{j+d}-t_{j+2}},$$

which can be rearranged as:

$$\beta_{j+2,2} = \frac{\beta_{j+2,1}-\beta_{j+1,1}}{(t_{j+d}-t_{j+2})/(d-2)},\ f''(x) = \sum_{j=1}^{p-2} B_{j+2,d-2}(x)\beta_{j+2,2}. \tag{3}$$

Finally, applying induction to (2)(3), we obtain the result for $f^{(m)}(x)$ $(1 \le m \le d-1)$:

$$\beta_{j+m,m} = \frac{\beta_{j+m,m-1}-\beta_{j+m-1,m-1}}{(t_{j+d}-t_{j+m})/(d-m)},\ f^{(m)}(x) = \sum_{j=1}^{p-m} B_{j+m,d-m}(x)\beta_{j+m,m}. \tag{4}$$

To get rid of distracting subscripts in these equations, let's define:

$$\boldsymbol{B}_{d-m}(x) = \begin{bmatrix} B_{1+m,d-m}(x) \\ B_{2+m,d-m}(x) \\ \vdots \\ B_{p,d-m}(x) \end{bmatrix},\ \boldsymbol{\beta}_m = \begin{bmatrix} \beta_{1+m,m} \\ \beta_{2+m,m} \\ \vdots \\ \beta_{p,m} \end{bmatrix},$$

$$\boldsymbol{W}_m = \frac{1}{d-m}\begin{bmatrix} t_{d+1}-t_{1+m} & & & \\ & t_{d+2}-t_{2+m} & & \\ & & \ddots & \\ & & & t_{p+d-m}-t_p \end{bmatrix},$$

then (4) can be written in a neat matrix-vector form:

$$\boldsymbol{\beta}_m = \boldsymbol{W}_m^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_{m-1},\ f^{(m)}(x) = \boldsymbol{B}_{d-m}(x)^{\mathrm{T}}\boldsymbol{\beta}_m. \tag{5}$$

By expanding the recursion for $\boldsymbol{\beta}_m$ in (5):

$$\boldsymbol{\beta}_m = \boldsymbol{W}_m^{-1}\boldsymbol{\Delta}\boldsymbol{\beta}_{m-1} = \cdots = \underbrace{\boldsymbol{W}_m^{-1}\boldsymbol{\Delta}\boldsymbol{W}_{m-1}^{-1}\boldsymbol{\Delta}\cdots\boldsymbol{W}_1^{-1}\boldsymbol{\Delta}}_{\boldsymbol{D}_m}\boldsymbol{\beta}_0,$$

we can work out the order-m general difference matrix $\boldsymbol{D}_m$, so that:

$$\boldsymbol{\beta}_m = \boldsymbol{D}_m\boldsymbol{\beta}_0 = \boldsymbol{D}_m\boldsymbol{\beta},\ \mathrm{PEN}_{\mathrm{gps}}^{(m)}(\boldsymbol{\beta}) = \|\boldsymbol{D}_m\boldsymbol{\beta}\|^2 = \|\boldsymbol{\beta}_m\|^2.$$

What we do in the previous section is exactly as same as above. In general, "active" knots are $(t_j)_{1+m}^{K-m}$. The lag-$(d-m)$ differences of "active" knots, divided by $d-m$, give the diagonal elements of the weight matrix $\boldsymbol{W}_m$. Order-$(d-m)$ B-splines defined on "active" knots are collected in $\boldsymbol{B}_{d-m}(x)$. In particular, their linear combination with coefficients $\boldsymbol{\beta}_m$ gives $f^{(m)}(x)$. Therefore, general difference penalty is the squared L_2 norm of $f^{(m)}(x)$'s B-splines coefficients. When $\lambda = +\infty$, we have:

$$\mathrm{PEN}_{\mathrm{gps}}^{(m)}(\boldsymbol{\beta}) = 0 \;\Leftrightarrow\; \boldsymbol{\beta}_m = \mathbf{0} \;\Leftrightarrow\; f^{(m)}(x) = 0,$$

which holds regardless of knot locations. Thus, the limiting GPS fit is always an order-m polynomial.

2.4 Connection with O-Splines

OS is characterized by a derivative penalty:

$$\mathrm{PEN}_{\mathrm{os}}^{(m)}(\boldsymbol{\beta}) = \int_a^b f^{(m)}(x)^2 \mathrm{d}x = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{S}_m \boldsymbol{\beta},$$

where $\boldsymbol{S}_m$ is a symmetric matrix with element $\int_a^b B_u^{(m)}(x) B_v^{(m)}(x) \mathrm{d}x$ at row u and column v. It used to be increasingly difficult to compute $\boldsymbol{S}_m$ as m grows, so SPS with easy-to-compute standard difference matrix $\boldsymbol{\Delta}_m$ were proposed as a simple alternative. However, SPS are not full replacements of OS due to their incapability to handle NUBS. The new GPS has the potential to fully replace OS, but we need to establish the equivalence between the general difference penalty and the derivative penalty to be confident about that.

To explore the relation between two penalties, we start with a new way to compute $\boldsymbol{S}_m$. Plugging $\boldsymbol{\beta}_m = \boldsymbol{D}_m \boldsymbol{\beta}$ into (5), we get:

$$f^{(m)}(x) = \boldsymbol{B}_{d-m}(x)^{\mathrm{T}} \boldsymbol{D}_m \boldsymbol{\beta}.$$

It follows that:

$$\mathrm{PEN}_{\mathrm{os}}^{(m)}(\boldsymbol{\beta}) = \int_a^b f^{(m)}(x)^{\mathrm{T}} f^{(m)}(x) \mathrm{d}x = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{D}_m^{\mathrm{T}} \bar{\boldsymbol{S}}_m \boldsymbol{D}_m \boldsymbol{\beta},$$

where $\bar{\boldsymbol{S}}_m$ is a positive definite matrix with element $\int_a^b B_{u,d-m}(x) B_{v,d-m}(x) \mathrm{d}x$ at row u and column v, whose computation can employ the algorithm of Wood (2017b). The general difference matrix $\boldsymbol{D}_m$ is also routine to compute (see Section 2.2), thus, we can build $\boldsymbol{S}_m$ from a sandwich formula:

$$\boldsymbol{S}_m = \boldsymbol{D}_m^{\mathrm{T}} \bar{\boldsymbol{S}}_m \boldsymbol{D}_m.$$

This formula explicitly links the two penalties together. Expressing the general difference penalty as:

$$\mathrm{PEN}_{\mathrm{gps}}^{(m)}(\boldsymbol{\beta}) = \|\boldsymbol{D}_m \boldsymbol{\beta}\|^2 = \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{D}_m^{\mathrm{T}} \boldsymbol{D}_m \boldsymbol{\beta},$$

we observe that it would equal the derivative penalty if $\bar{\boldsymbol{S}}_m$ were an identity matrix. So how far is $\bar{\boldsymbol{S}}_m$ from being exactly or proportional to identity? Consider the example knot sequence used in Section 2.2 again. For cubic B-splines ($d = 4$), it can be computed that:

$$\bar{\boldsymbol{S}}_1 = \begin{bmatrix} \frac{1}{5} & \frac{11}{90} & \frac{2}{15} & & \\ \frac{11}{90} & \frac{2}{5} & \frac{1}{5} & \frac{1}{135} & \\ \frac{2}{15} & \frac{1}{5} & \frac{22}{27} & \frac{1}{5} & \frac{2}{15} \\ & \frac{1}{135} & \frac{1}{5} & \frac{2}{5} & \frac{11}{90} \\ & & \frac{2}{15} & \frac{11}{90} & \frac{1}{5} \end{bmatrix}, \quad \bar{\boldsymbol{S}}_2 = \begin{bmatrix} \frac{1}{3} & \frac{1}{6} & & \\ \frac{1}{6} & 1 & \frac{1}{3} & \\ & \frac{1}{3} & 1 & \frac{1}{6} \\ & & \frac{1}{6} & \frac{1}{3} \end{bmatrix}, \quad \bar{\boldsymbol{S}}_3 = \begin{bmatrix} 1 & & \\ & 2 & \\ & & 1 \end{bmatrix}.$$

In general, the bigger m is, the closer $\bar{\boldsymbol{S}}_m$ is to diagonal. It is diagonal when $m = d - 1$, but not proportional to identity unless all knots are equidistant. Therefore, the general difference penalty, taking $\bar{\boldsymbol{S}}_m$ to be identity, is always a simplification of the derivative penalty.

The distinction between the two penalties actually does not matter. Defining $\boldsymbol{K}_m = \boldsymbol{U}_m \boldsymbol{D}_m$, where $\boldsymbol{U}_m$ is $\bar{\boldsymbol{S}}_m$'s upper triangular Cholesky factor such that $\bar{\boldsymbol{S}}_m = \boldsymbol{U}_m^{\mathrm{T}} \boldsymbol{U}_m$, we can express the derivative penalty as:

$$\mathrm{PEN}_{\mathrm{os}}^{(m)}(\boldsymbol{\beta}) = \|\boldsymbol{U}_m \boldsymbol{D}_m \boldsymbol{\beta}\|^2 = \|\boldsymbol{K}_m \boldsymbol{\beta}\|^2.$$

Since $\boldsymbol{U}_m$ has full rank, $\boldsymbol{D}_m$ and $\boldsymbol{K}_m$ have the same null space. Decomposing $\boldsymbol{\beta} = \boldsymbol{\xi} + \boldsymbol{\theta}$, where $\boldsymbol{\xi}$ is $\boldsymbol{\beta}$'s projection on this null space and $\boldsymbol{\theta}$ is orthogonal to $\boldsymbol{\xi}$, we have $\boldsymbol{D}_m \boldsymbol{\beta} = \boldsymbol{D}_m \boldsymbol{\theta}$, $\boldsymbol{K}_m \boldsymbol{\beta} = \boldsymbol{K}_m \boldsymbol{\theta}$ and $\boldsymbol{D}_m \boldsymbol{\xi} = \boldsymbol{K}_m \boldsymbol{\xi} = \boldsymbol{0}$. In other words, both penalties act on $\boldsymbol{\theta}$ only, so we may instead write:

$$\mathrm{PEN}_{\mathrm{gps}}^{(m)}(\boldsymbol{\theta}) = \|\boldsymbol{D}_m \boldsymbol{\theta}\|^2, \ \mathrm{PEN}_{\mathrm{os}}^{(m)}(\boldsymbol{\theta}) = \|\boldsymbol{K}_m \boldsymbol{\theta}\|^2.$$

Since $\boldsymbol{U}_m$ has full rank, for any $\boldsymbol{\theta}_{\mathrm{gps}}$, there is a unique $\boldsymbol{\theta}_{\mathrm{os}}$ such that $\boldsymbol{D}_m \boldsymbol{\theta}_{\mathrm{gps}} = \boldsymbol{K}_m \boldsymbol{\theta}_{\mathrm{os}}$ and $\mathrm{PEN}_{\mathrm{gps}}^{(m)}(\boldsymbol{\theta}_{\mathrm{gps}}) = \mathrm{PEN}_{\mathrm{os}}^{(m)}(\boldsymbol{\theta}_{\mathrm{os}})$, establishing a one-to-one correspondence between two functions. Therefore, while the two penalties do not equal each other for a common $\boldsymbol{\theta}$, they are equivalent in wiggliness control. As $\lambda \to +\infty$, both OS and GPS tend to the same least squares order-m polynomial. As $\lambda \to 0$, both estimators tend to the same least squares regression spline.

Note that it is not our intention to approximate the derivative penalty as best as we can. Technically, the general difference penalty can be modified to better approximate the derivative penalty, by an extra weighting: $\tilde{\boldsymbol{D}}_m = \boldsymbol{B}_m^{1/2} \boldsymbol{D}_m$, where $\boldsymbol{B}_m$ is a diagonal matrix containing $\bar{\boldsymbol{S}}_m$'s diagonal elements, i.e., the j^{th} diagonal entry in $\boldsymbol{B}_m$ is $\int_a^b B_{j,d-m}(x)^2 \mathrm{d}x$. In this way, $\tilde{\boldsymbol{D}}_m$ also accounts for NUBS' inhomogeneous shapes. However, we do not appreciate this idea for three reasons. Firstly, it obscures the simple interpretation of the difference penalty. Without this extra weighting, the general difference penalty stands for the squared L_2 norm of $f^{(m)}(x)$'s B-splines coefficients. With this extra weighting, it is the squared L_2 norm of $f^{(m)}(x)$'s scaled B-splines coefficients, where the j^{th} coefficient is multiplied by the square root of the

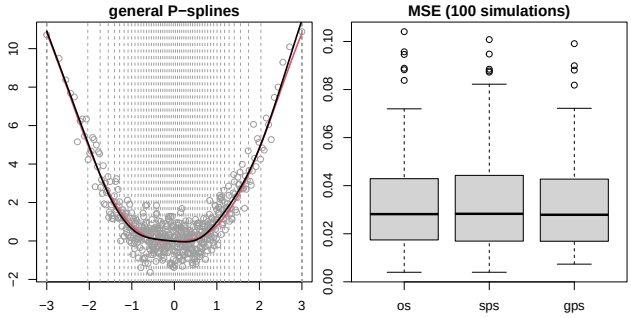

Fig. 4 Left: GPS (black) on quantile knots (vertical dashed lines) fitted to noisy observations (gray dots) from a U-shaped $g(x)$ (red). Right: MSE performance of OS, SPS and GPS.

area under the j^{th} B-splines curve, which is rather cumbersome. Secondly, with this extra weighting, the general difference penalty loses its computational simplicity as it requires numerical integration to work out $\boldsymbol{B}_m$. Last but not least, as our simulation studies show, the general difference penalty works very well as a wiggliness penalty without this extra weighting. To conclude, compared with the derivative penalty, the general different penalty is simpler in both interpretation and computation, yet equally powerful and successful for wiggliness control.

3 Simulation Studies

In this section, we fit cubic OS, SPS and GPS to simulated $(x_i, y_i)_1^n$ from $y_i = g(x_i) + \varepsilon_i$ and compare their MSE performance. Here, $g(x)$ is a true signal function, ε_i is a Gaussian white noise and we will give four examples of $g(x)$. For each example, we choose the same number of interior knots k (or the same number of B-splines p) and the same penalty order m for all spline estimators; only knot locations differ. For SPS, we construct UBS on equidistant knots through the range of $(x_i)_1^n$. For OS and GPS, we construct NUBS on unevenly spaced knots positioned at equal quantiles of $(x_i)_1^n$. To make two knot sequences substantially different, we will generate unevenly spaced $(x_i)_1^n$ in the first place.

U-shaped Curve We simulated $n = 500$ noisy observations from $g(x) = \mathrm{x}^3 - 0.2x^4$, where $(x_i)_1^n$ were drawn from $\mathrm{N}(0, 1)$ distribution, then penalized cubic B-splines on $k = 50$ interior knots by a 2nd order penalty ($m = 2$) for smoothing. The fitted OS, SPS and GPS were indistinguishable. See Figure 4 (left) for the fitted GPS. In Figure 4 (right), the boxplot of MSEs based on 100 simulations suggests that OS, SPS and GPS perform equally well on average.

Random Curve We can construct a more complex $g(x)$ on $[0, 1]$, by setting it up as a linear combination of 8 uniform cubic B-splines. To make it smooth, we may take B-splines coefficients to be a realization of an AR(1) process with auto-correlation coefficient $-1/3$. See Figure 5 (left) for an example random curve created this way. For this example, we sampled $n = 500$ x-values

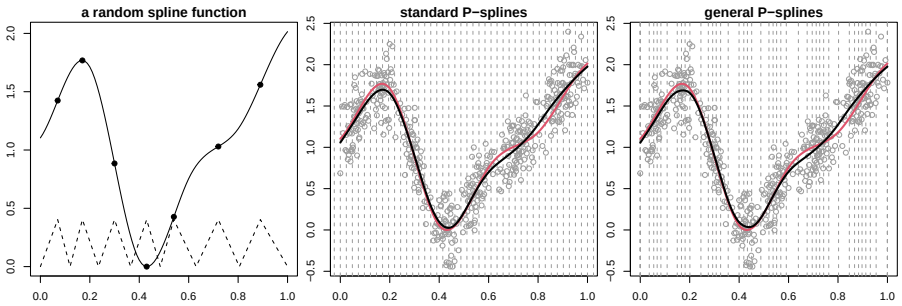

Fig. 5 Left: an example random $g(x)$ (solid). Unevenly spaced $(x_i)_1^n$ are to be sampled from a "tent" distribution whose density (dashed) peaks at local extrema (dots) of $g(x)$ and $g'(x)$. Middle: SPS (black) on equidistant knots (vertical dashed lines) fitted to noisy observations (gray dots) from $g(x)$ (red). Right: the fitted GPS on quantile knots.

from a "tent" distribution whose density peaks at local extrema points of $g(x)$ and $g'(x)$. As a result, $(x_i)_1^n$ were more concentrated around those critical locations that cut $g(x)$ into monotonic/convex/concave pieces. We then simulated noisy observations from $g(x)$ and penalized cubic B-splines on $k = 40$ interior knots by a 2nd order penalty ($m = 2$) for smoothing. The fitted OS, SPS and GPS were indistinguishable. See Figure 5 (middle and right) for the fitted SPS and GPS.

We also tried 1st order and 3rd order penalties. In each case, the fitted OS, SPS and GPS were highly alike, but the 1st order penalty led to a rougher fit and inferior MSE performance. To better illustrate the quality of the fit, we subtracted $g(x)$ from the fitted spline and sketched this difference in Figure 6 (first row). Clearly, the deviation curve is more wiggly when $m = 1$, while $m = 2$ and $m = 3$ almost give identically smooth result. This explains why $m = 2$ is widely used in practice. The boxplots of MSEs in Figure 6 (second row) confirms that $m = 1$ yields worse MSE performance. In addition, OS, SPS and GPS perform equally well when $m = 2$ or 3.

Many Random Curves We can take a step further and base our simulations on N random curves instead of any specific one. To be precise, we repeat the following procedure for $l = 1, 2, \ldots, N$:

1. generate a random curve $g^{[l]}(x)$ on $[0, 1]$;
2. draw $(x_i^{[l]})_1^n$ from a "tent" distribution that has higher density around local extrema points of $g^{[l]}(x)$ and $g^{[l]\prime}(x)$;
3. evaluate $g_i^{[l]} = g^{[l]}(x_i^{[l]})$ and simulate noisy observations $y_i^{[l]} = g_i^{[l]} + \varepsilon_i^{[l]}$, with $\varepsilon_i^{[l]} \overset{\text{iid}}{\sim} \mathrm{N}(g_i^{[l]}, \sigma_l^2)$;
4. estimate $\hat{f}_i^{[l]} = \hat{f}^{[l]}(x_i^{[l]})$ from $(x_i^{[l]}, y_i^{[l]})$ using OS, SPS and GPS;
5. compute relative MSE: $\delta_l = \frac{1}{n}\sum_{i=1}^n (\hat{f}_i^{[l]} - g_i^{[l]})^2/\sigma_l^2$, for each fitted spline.

Steps 1-2 are well described in the previous example. At step 3, we specify $\sigma_l = \gamma \times \text{s.e.}\{(g_i^{[l]})_1^n\}$, where $\text{s.e.}\{(g_i^{[l]})_1^n\}$ is the standard error of the signal values and γ is a *noise-to-signal ratio*. Note that although we hold γ fixed through $l = 1, 2, \ldots, N$, σ_l will vary from signal to signal. This explains why in

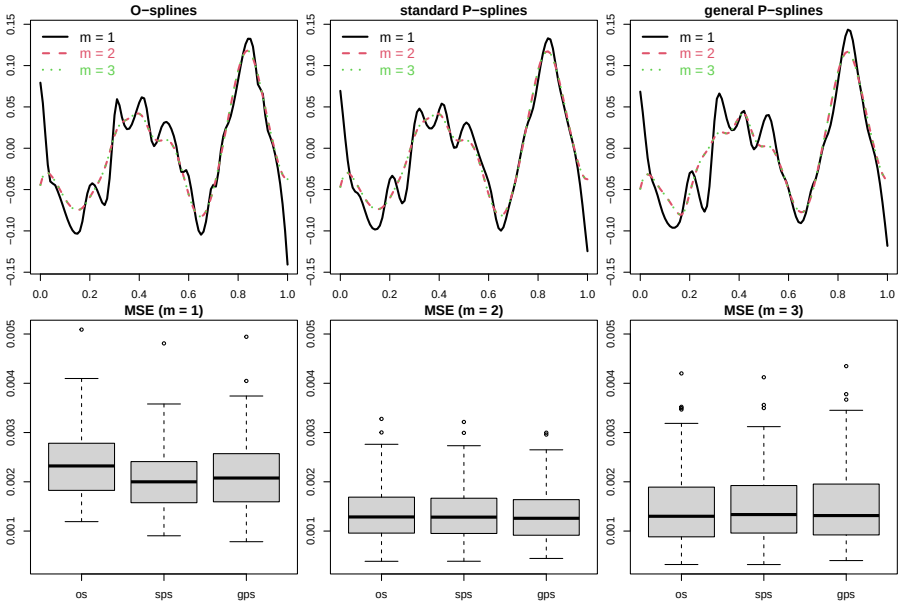

Fig. 6 This figure continues with Figure 5. First Row: the difference between fitted splines and $g(x)$ for different penalty order m. Second Row: MSE performance of OS, SPS and GPS when m varies.

step 5 we need to divide raw MSE by σ_ι^2. At step 4, we represent $f^{[\iota]}(x)$ using 40 B-splines. Such basis dimension is already more than adequate, because the random curve in step 1 is constructed with 8 B-splines only.

To start the simulation we need to specify the number of simulations N, the number of observations n and the noise-to-signal ratio γ. For penalty order, we attempt all $m = 1, 2, 3$. Figure 7 shows the boxplots of MSEs for $N = 100$ and four different (n, γ) pairs: (1000, 0.1), (1000, 0.5), (5000, 0.1) and (5000, 0.5). In all cases, GPS are no inferior to OS or SPS.

Inhomogeneously Smooth Curve We now demonstrate an example where GPS are much better than SPS. Consider smoothing $n = 1000$ noisy observations of $g(x) = x + \sin(10\pi x^5)$, $x \in [0, 1]$ at unevenly spaced $(x_i)_1^n$ such that $(x_i^5)_1^n$ is equidistant. Figure 8 shows that SPS on 50 equidistant knots end up with an extremely wiggly fit, while GPS or OS on quantile knots successfully yield a smooth estimate, with huge improvement in MSE performance. In general, SPS would never approximate an inhomogeneously smooth function uniformly, no matter how many equidistant knots are positioned. GPS on a set of carefully selected unevenly spaced knots will overcome such flaw.

In summary, GPS are no inferior to SPS or OS in terms of MSE performance. On one hand, in all examples, GPS are as good as or even slightly better than OS. On the other hand, when compared with SPS, GPS perform equally well if unevenly spaced knots are not crucial but performs much better otherwise.

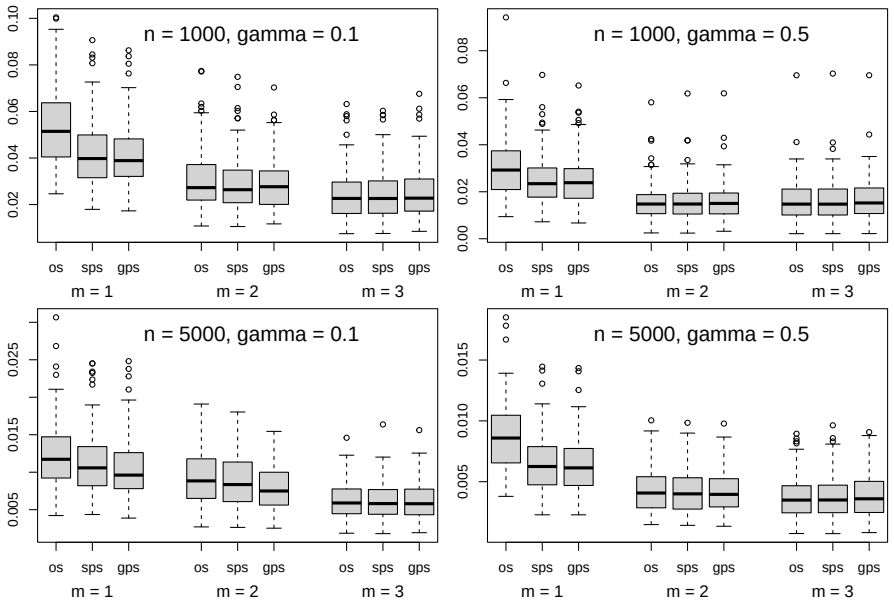

Fig. 7 MSE performance of OS, SPS and GPS for various choices of n (sample size) and γ (noise-to-signal ratio) under $N = 100$ simulations.

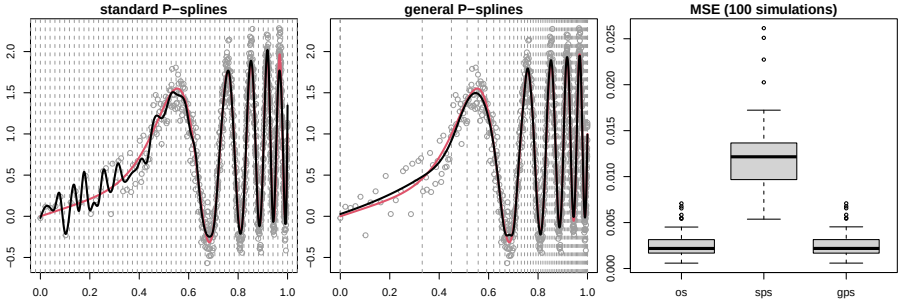

Fig. 8 Smoothing noisy observations (gray dots) from an inhomogeneously smooth $g(x)$ (red). Left: SPS (black) on equidistant knots (vertical dashed lines). Middle: GPS on quantile knots. Right: MSE performance of OS, SPS and GPS.

4 Real Data Examples

In this section, we demonstrate the impact of knot placement on penalized B-splines, by comparing SPS and GPS fitted to real datasets. As before, we use the same number of interior knots k (or the same number of B-splines p) but different knot placement strategies for the two estimators. For SPS we position equidistant knots, while for GPS we position quantile knots (Ruppert et al, 2003) to exemplify unevenly spaced knots. We will see that for small to moderate k, the BMC longitudinal data favor GPS on quantile knots, while the fossil shell data favor SPS on equidistant knots. We do not interpret this

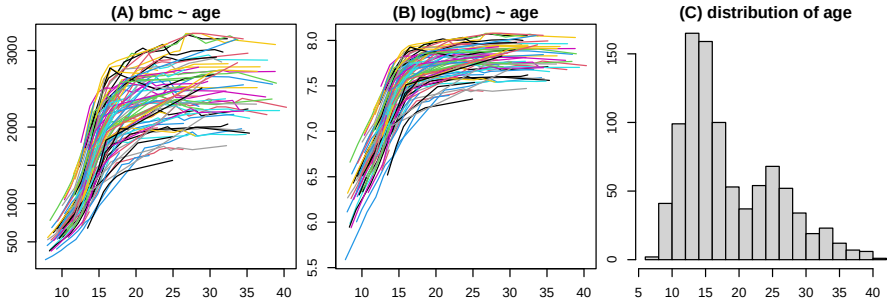

Fig. 9 (A) 932 BMC measurements (in grams) for 112 white ethnic males. (B) log-transformed measurements. (C) distribution of age when measurement was taken.

phenomenon as one type of knots being superior to the other. Rather, we would say that for a particular dataset, one type of knots is more effective than the other when k or p is restricted in size. For both examples, as k or p becomes adequately big, exact knot locations no longer matter.

4.1 BMC Longitudinal Data

The Pediatric Bone Mineral Accrual Study (PBMAS) (Bailey, 1997; Baxter-Jones et al, 2011) was launched in 1991 by the University of Saskatchewan, aiming to investigate the accumulation of bone mineral content (BMC) in growing children. Here, we build an additive mixed model (Zuur et al, 2014; Pedersen et al, 2019) for 112 white ethnic males in this dataset, the same cohort studied by Elhakeem et al (2022) using linear mixed models (Pinheiro and Bates, 2000; Bates et al, 2015). We formulate our model as:

$$\log(\text{bmc})_{j,\text{age}} = f(\text{age}) + f_j(\text{age}) + e_{j,\text{age}},$$

where f (a smooth fixed-effect function) is the population log(BMC) trajectory, f_j (a smooth random-effect function) is the deviation of subject j's log(BMC) trajectory from f and $e_{j,\text{age}} \overset{\text{iid}}{\sim} \text{N}(0, \sigma_e^2)$ is a measurement error. Constructing f and f_j as penalized B-splines, the model can be estimated in the framework of generalized additive models (GAMs) (Wood, 2017a) using **R** package **mgcv** (Wood, 2022).

We represent both f and f_j by cubic B-splines. For penalty order, we choose $m = 2$ for f, but $m = 1$ for f_j. This subtlety comes from the fact that fewer subjects came back for measurements as they grew older. For example, the number of subjects (out of 112 in total) with no observations after the age of 20, 25 and 30 are 36, 50 and 60, respectively. Using $m = 2$ for f_j includes a random slope in the penalty's null space that would lead to linear extrapolation for these subjects. But Figure 9 shows that log(BMC) trajectories are relatively flat after the age of 20. Thus, it is more reasonable to exclude this random slope by choosing $m = 1$, leaving only a random intercept in the penalty's null space.

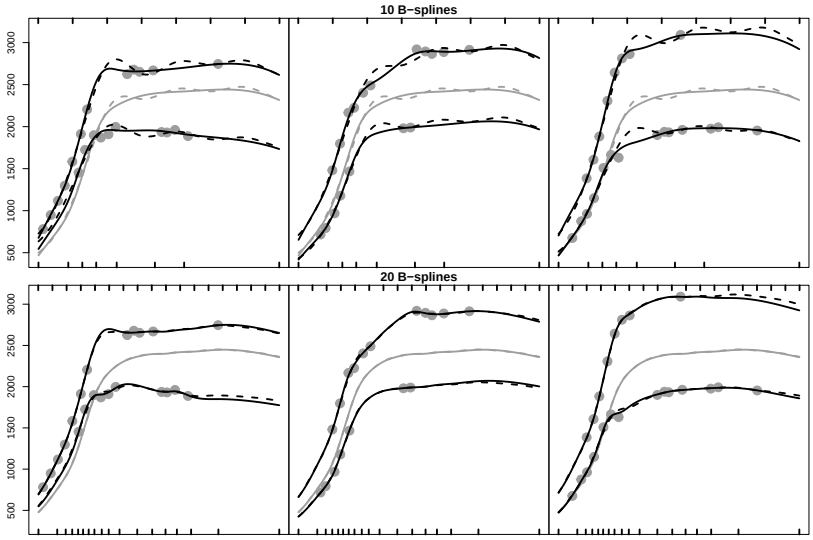

Fig. 10 Estimated BMC trajectories with GPS (solid) and SPS (dashed) when $p = 10$ (top row) and 20 (bottom row). Gray curves: population trajectories; black curves: subject trajectories for six selected subjects; gray dots: data for the selected subjects; rugs on the bottom axis: quantile knots for GPS; rugs on the top axis: equidistant knots for SPS.

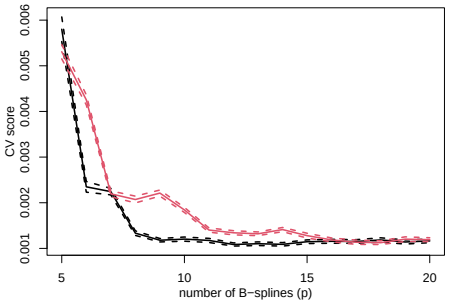

Fig. 11 Cross-validation score (solid) and its 2 standard error bounds (dashed) for GPS (black) and SPS (red) at $p = 5, 6, \ldots, 20$. For BMC data, GPS on quantile knots are more effective than SPS on equidistant knots.

We set up both f and f_j using p cubic B-splines on the same knot sequence, which comprises $k = p - 4$ interior knots and other necessary boundary knots. Note that there are more measurements at early ages (see Figure 9(C)), so quantile knots are right-skewed. Figure 10 shows that when $p = 10$, SPS are suspiciously wiggly, while GPS are plausibly smooth. But the difference between two P-splines are much less noticeable when $p = 20$. The message is that for BMC data, GPS on quantile knots are more effective than SPS on equidistant knots. Let's demonstrate this in a more quantitative way by computing the 10-fold cross-validation score for $p = 5, 6, \ldots, 20$. Figure 11 shows that the score for GPS decreases with p faster. It already achieves a

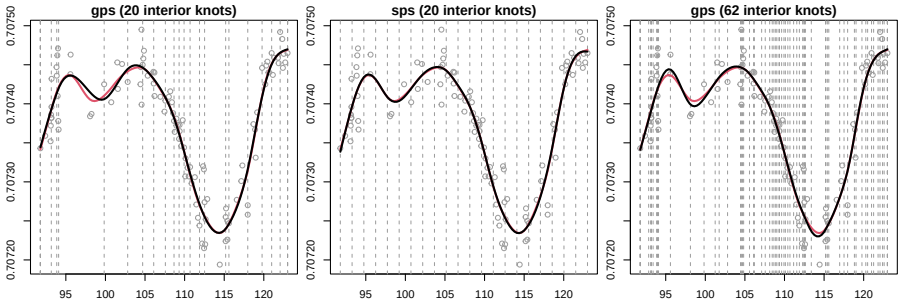

Fig. 12 Fitted SPS and GPS (black) and `smooth.spline` fit (red) for fossil shell data (gray dots). Vertical dashed lines: knot locations for SPS and GPS.

satisfactory fit when $p = 9$, but SPS require $p = 17$ to attain about the same quality.

It is well worth pointing out that keeping p small is crucial to computational efficiency. Estimation of GAMs involves $O(np^2)$ (n being the number of observations) and $O(p^3)$ at different stages, both growing quickly with p. On our computer, fitting BMC data with $p = 10$, 20 and 30 takes 19, 90 and 290 seconds, respectively. Clearly, GPS with $p = 10$ have a great computational advantage over SPS with $p = 20$.

4.2 Fossil Shell Data

The fossil shell data raised a debate (Ruppert et al, 2009; Wand and Ormerod, 2008; Eilers et al, 2015) on the "superiority" of either equidistant knots or quantile knots. We revisit this example, and aligned with these works, consider an estimate by **R** function `smooth.spline` as "truth". Note that previous demonstration with quantile knots was via OS. Here we use GPS.

Eilers et al (2015) claimed that equidistant knots are superior, because Figure 12 (left and middle) shows that GPS on 20 quantile knots fail to capture the local minimum near $x = 98$ correctly, while SPS on 20 equidistant knots have no trouble with this. They criticized that placing quantile knots is "the root of all evil". However, they failed to realize that `smooth.spline` also fits OS on quantile knots, but using more knots. Figure 12 (right) shows that re-estimating GPS on the same 62 quantile knots positioned by `smooth.spline` seems to outperform both OS and SPS, with better approximation to data near $x = 96$, 98 and 114 without sacrificing any smoothness. Since the true function is actually not known, we report residual sum of squares (RSS) rather than MSE for these estimates, which are $\text{RSS}_{\text{sps}} = 5.87 \times 10^{-8}$, $\text{RSS}_{\text{os}} = 5.78 \times 10^{-8}$ and $\text{RSS}_{\text{gps}} = 5.74 \times 10^{-8}$. Clearly, with sufficient number of knots, GPS attains the best fit.

In our view, there is no so-called "superiority" for either type of knots. Which one appears more advantageous is conditional on the number of interior knots k. Figure 13 shows the cross-validation score of SPS and GPS for $k = 10, 11, \ldots, 60$. It is astounding to spot a zigzag in GPS' score. Figure 14 shows

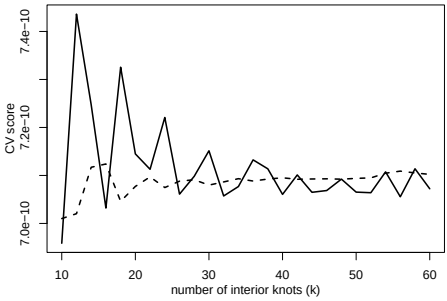

Fig. 13 Leave-one-out generalized cross-validation scores of GPS (solid) and SPS (dashed) for $k = 10, 11, \ldots, 60$. For fossil shell data, SPS on equidistant knots are more effective than GPS on quantile knots.

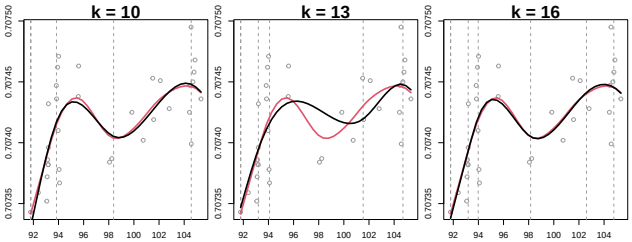

Fig. 14 A zoomed-in display of GPS (black) and `smooth.spline` (red) on $[92, 106]$. A small k may unfortunately miss the critical knot around $x = 98$.

that for fossil shell data, GPS on quantile knots are very sensitive to the existence of a knot around $x = 98$, which a small k may unfortunately miss. Such sensitivity gets reduced as k increases, and after $k = 40$, GPS begin to outperform SPS by a small margin. But before that, SPS almost always gain the upper hand. Hence, for small to moderate k, SPS are more effective for fossil shell data.

5 Conclusion

We propose a new penalized B-splines estimator, the GPS, to accommodate NUBS on unevenly spaced knots. They feature a novel general difference penalty that accounts for uneven knot spacing, complementing Eilers and Marx's SPS tailored for UBS on equidistant knots. Simulation studies show that when unevenly spaced knots are not key to uniform approximation of a smooth function, GPS are as good as SPS in terms of MSE performance. However, for inhomogeneously smooth functions, GPS on carefully positioned unevenly spaced knots perform much better than SPS on blindly positioned equidistant knots. In a practical example with BMC longitudinal data, GPS on a few unevenly spaced knots are more satisfactory than SPS on many equidistant knots, as they are faster to compute and produce smoother trajectories with no ripples.

GPS are also closely related to OS characterized by a derivative penalty. Although such penalty has a longer history, it is less straightforward to compute and harder to interpret. By contrast, the general difference penalty associated with our GPS is both easy to interpret and simple to compute. We derive a sandwich formula to establish a one-to-one correspondence between the two penalties, showing that they are equally powerful for wiggliness control. In simulation studies, GPS are as good as or even slightly better than OS in terms of MSE performance. Therefore, GPS are good replacements of OS and deserve greater attention.

Declarations

Funding. Zheyuan Li was supported by National Natural Science Foundation of China under Young Scientists Fund NSFC-12001166. Jiguo Cao was supported by Natural Sciences and Engineering Research Council of Canada under Discovery Grant 2018-06008.

Conflict of interest. The authors declare that they have no conflict of interest.

Code availability. The **R** code for this paper is available as a vignette at https://github.com/ZheyuanLi/gps-vignettes/blob/main/gps1.pdf.

Data availability. BMC longitudinal data are not open data. For this paper, a high-quality synthetic version is used and provided in **gps.mgcv**. Researchers who are interested in accessing the authentic data from PBMAS should contact Adam Baxter-Jones (baxter.jones@usask.ca).

Appendix A Spline and B-Splines

An order-d spline $f(x)$ defined on domain $[a, b]$ comprises smoothly connected polynomial segments of degree $d-1$, with $d-2$ continuous derivatives at their interior knots (or break points) $a < s_1 < s_2 < \ldots < s_k < b$. Equivalently, it may be expressed as a linear combination of $p = k+d$ order-d B-splines, whose construction requires $d-1$ arbitrarily positioned auxiliary boundary knots on each side of $[a, b]$. In total, there are $K = k + 2d$ knots (including a and b), denoted by $(t_j)_1^K$.

For example, the following is a cubic spline ($d = 4$) on $[a, b] = [1, 6]$, with interior knots at $s_1 = 2, s_2 = 3, s_3 = 4, s_4 = 5$:

$$f(x) = \begin{cases} 1.09 + 0.610(x-1) - 0.060(x-1)^2 - \frac{23}{75}(x-1)^3 & x \in [1,2), \\ \frac{4}{3} \quad - 0.430(x-2) - 0.980(x-2)^2 + \frac{59}{75}(x-2)^3 & x \in [2,3), \\ 0.71 - 0.030(x-3) + 1.380(x-3)^2 - \frac{107}{150}(x-3)^3 & x \in [3,4), \\ \frac{101}{75} + 0.590(x-4) - 0.760(x-4)^2 + \frac{7}{24}(x-4)^3 & x \in [4,5), \\ \frac{881}{600} - 0.055(x-5) + 0.115(x-5)^2 + \frac{37}{300}(x-5)^3 & x \in [5,6). \end{cases}$$

To express $f(x)$ as a linear combination of B-splines, we need to arbitrarily place 3 auxiliary boundary knots on each side of $[1, 6]$. For example, with the following full knot sequence:

			a	s_1	s_2	s_3	s_4	b			
t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}
-2	-1	0	1	2	3	4	5	6	7	8	9

we can represent $f(x)$ using 8 UBS (see Figure A1) with coefficients 0.44, 1.11, 1.66, 0.25, 1.60, 1.43, 1.49 and 2.52. Placing auxiliary boundary knots elsewhere only results in different B-splines coefficients. For example, with clamped boundary knots $t_1 = t_2 = t_3 = t_4 = 1$ and $t_9 = t_{10} = t_{11} = t_{12} = 6$, B-splines coefficients are 1.09, $\frac{97}{75}$, 1.66, 0.25, 1.60, 1.43, 1.47 and $\frac{991}{600}$.

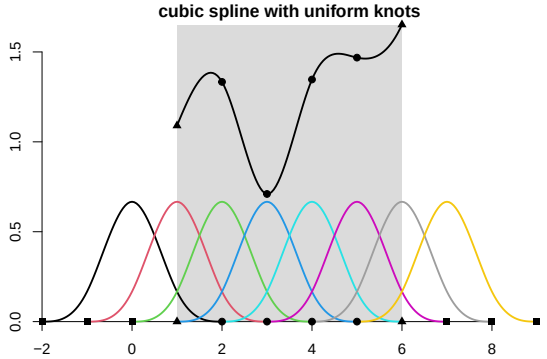

Fig. A1 B-splines (bottom) representation of a cubic spline (top). Shaded area: domain; triangles: domain endpoints; circles: interior knots (break points of polynomial pieces); squares: auxiliary boundary knots for constructing B-splines.

References

Andrinopoulou ER, Eilers PHC, Takkenberg JJM, et al (2018) Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines. Biometrics 74(2):685–693. https://doi.org/10.1111/biom.12814

Bailey DA (1997) The Saskatchewan Pediatric Bone Mineral Accrual Study: Bone mineral acquisition during the growing years. International Journal of Sports Medicine 18:s191–s194. https://doi.org/10.1055/s-2007-972713

Bates D, Mächler M, Bolker B, et al (2015) Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67(1):1–48. https://doi.org/10.18637/jss.v067.i01

Baxter-Jones AD, Faulkner RA, Forwood MR, et al (2011) Bone mineral accrual from 8 to 30 years of age: An estimation of peak bone mass. Journal

of Bone and Mineral Research 26(8):1729–1739. https://doi.org/10.1002/jbmr.412

de Boor C (2001) A Practical Guide to Splines (Revised Edition), Applied Mathematical Sciences, vol 27. Springer New York

Bremhorst V, Lambert P (2016) Flexible estimation in cure survival models using Bayesian P-splines. Computational Statistics & Data Analysis 93(SI):270–284. https://doi.org/10.1016/j.csda.2014.05.009

Chen J, Ohlssen D, Zhou Y (2018) Functional mixed effects model for the analysis of dose-titration studies. Statistics in Biopharmaceutical Research 10(3):176–184. https://doi.org/10.1080/19466315.2018.1458649

Eilers PH, Marx BD (2021) Practical Smoothing: The Joys of P-Splines. Cambridge University Press

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Statistical Science 11(2):89–102. https://doi.org/10.1214/ss/1038425655

Eilers PHC, Marx BD, Durbán M (2015) Twenty years of P-splines. Statistics and Operations Research Transactions 39(2):149–186

Elhakeem A, Hughes R, Tilling K, et al (2022) Using linear and natural cubic splines, SITAR, and latent trajectory models to characterise nonlinear longitudinal growth trajectories in cohort studies. BMC Medical Research Methodology 22(68). https://doi.org/10.1186/s12874-022-01542-8

Franco-Villoria M, Scott M, Hoey T (2019) Spatiotemporal modeling of hydrological return levels: A quantile regression approach. Environmetrics 30(2, SI). https://doi.org/10.1002/env.2522

Gijbels I, Ibrahim MA, Verhasselt A (2018) Testing the heteroscedastic error structure in quantile varying coefficient models. Canadian Journal of Statistics 46(2):246–264. https://doi.org/10.1002/cjs.11346

Goicoa T, Adin A, Etxeberria J, et al (2019) Flexible Bayesian P-splines for smoothing age-specific spatio-temporal mortality patterns. Statistical Methods in Medical Research 28(2):384–403. https://doi.org/10.1177/0962280217726802

Greco F, Ventrucci M, Castelli E (2018) P-spline smoothing for spatial data collected worldwide. Spatial Statistics 27:1–17. https://doi.org/10.1016/j.spasta.2018.08.008

Hendrickx K, Janssen P, Verhasselt A (2018) Penalized spline estimation in varying coefficient models with censored data. TEST 27(4):871–895. https:

//doi.org/10.1007/s11749-017-0574-y

Koehler M, Umlauf N, Beyerlein A, et al (2017) Flexible Bayesian additive joint models with an application to type 1 diabetes research. Biometrical Journal 59(6, SI):1144–1165. https://doi.org/10.1002/bimj.201600224

Maturana-Russel P, Meyer R (2021) Bayesian spectral density estimation using P-splines with quantile-based knot placement. Computational Statistics 36(3):2055–2077. https://doi.org/10.1007/s00180-021-01066-7

Minguez R, Basile R, Durban M (2020) An alternative semiparametric model for spatial panel data. Statistical Methods and Applications 29(4):669–708. https://doi.org/10.1007/s10260-019-00492-8

Muggeo VMR, Torretta F, Eilers PHC, et al (2021) Multiple smoothing parameters selection in additive regression quantiles. Statistical Modelling 21(5):428–448. https://doi.org/10.1177/1471082X20929802

Orbe J, Virto J (2021) Selecting the smoothing parameter and knots for an extension of penalized splines to censored data. Journal of Statistical Computation and Simulation 91(14):2953–2985. https://doi.org/10.1080/00949655.2021.1913737

O'Sullivan F (1986) A statistical perspective on ill-posed inverse problems. Statistical Science 1(4):502–518. https://doi.org/10.1214/ss/1177013525

Pedersen EJ, Miller DL, Simpson GL, et al (2019) Hierarchical generalized additive models in ecology: an introduction with mgcv. PeerJ 7:e6876–e6876. https://doi.org/10.7717/peerj.6876

Perperoglou A, Sauerbrei W, Abrahamowicz M, et al (2019) A review of spline function procedures in R. BMC Medical Research Methodology 19(1):46–46. https://doi.org/10.1186/s12874-019-0666-3

Pinheiro JC, Bates DM (2000) Mixed-Effects Models in S and S-PLUS. Springer

Rodriguez-Alvarez MX, Boer MP, van Eeuwijk FA, et al (2018) Correcting for spatial heterogeneity in plant breeding experiments with P-splines. Spatial Statistics 23:52–71. https://doi.org/10.1016/j.spasta.2017.10.003

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric Regression. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press

Ruppert D, Wand MP, Carroll RJ (2009) Semiparametric regression during 2003-2007. Electronic Journal of Statistics 3:1193–1256. https://doi.org/10.

1214/09-EJS525

Spiegel E, Kneib T, Otto-Sobotka F (2019) Generalized additive models with flexible response functions. Statistics and Computing 29(1):123–138. https://doi.org/10.1007/s11222-017-9799-6

Spiegel E, Kneib T, Otto-Sobotka F (2020) Spatio-temporal expectile regression models. Statistical Modelling 20(4):386–409. https://doi.org/10.1177/1471082X19829945

Wahba G (1990) Spline Models for Observational Data. Society for Industrial and Applied Mathematics

Wand MP, Ormerod JT (2008) On semiparametric regression with O'Sullivan penalized splines. Australian & New Zealand Journal of Statistics 50(2):179–198. https://doi.org/10.1111/j.1467-842X.2008.00507.x

Wang X, Roy V, Zhu Z (2018) A new algorithm to estimate monotone nonparametric link functions and a comparison with parametric approach. Statistics and Computing 28(5):1083–1094. https://doi.org/10.1007/s11222-017-9781-3

Wood S (2022) mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation. URL https://CRAN.R-project.org/package=mgcv, r package version 1.8-40

Wood SN (2017a) Generalized Additive Models: An Introduction with R, 2nd edn. Chapman and Hall

Wood SN (2017b) P-splines with derivative based penalties and tensor product smoothing of unevenly distributed data. Statistics and Computing 27(4):985–989. https://doi.org/10.1007/s11222-016-9666-x

Yu Y, Wu C, Zhang Y (2017) Penalised spline estimation for generalised partially linear single-index models. Statistics and Computing 27(2):571–582. https://doi.org/10.1007/s11222-016-9639-0

Zuur AF, Saveliev AA, Ieno EN (2014) A Beginner's Guide to Generalized Additive Mixed Models with R. Highland Statistics