Statistics and Computing
https://doi.org/10.1007/s11222-018-9818-2

On the estimation of variance parameters in non-standard generalised linear mixed models: application to penalised smoothing

María Xosé Rodríguez-Álvarez[1,2] · Maria Durban[3] · Dae-Jin Lee[1] · Paul H. C. Eilers[4]

Received: 16 March 2018 / Accepted: 5 June 2018

Abstract
We present a novel method for the estimation of variance parameters in generalised linear mixed models. The method has its roots in Harville (J Am Stat Assoc 72(358):320–338, 1977)'s work, but it is able to deal with models that have a precision matrix for the random effect vector that is linear in the inverse of the variance parameters (i.e., the precision parameters). We call the method SOP (separation of overlapping precision matrices). SOP is based on applying the method of successive approximations to easy-to-compute estimate updates of the variance parameters. These estimate updates have an appealing form: they are the ratio of a (weighted) sum of squares to a quantity related to effective degrees of freedom. We provide the sufficient and necessary conditions for these estimates to be strictly positive. An important application field of SOP is penalised regression estimation of models where multiple quadratic penalties act on the same regression coefficients. We discuss in detail two of those models: penalised splines for locally adaptive smoothness and for hierarchical curve data. Several data examples in these settings are presented.

Keywords Generalised linear mixed models · Generalised additive models · Variance parameters · Smoothing parameters · REML · Effective degrees of freedom

1 Introduction

The estimation of variance parameters is a statistical problem that has received extensive attention for more than 50 years. It originated with the ANOVA methodology proposed by Fisher in the 1920's, where estimates where obtained equating mean squared error to its expected value. However, the results yielded by this method were not optimal in some situations, for example, in the case of unbalanced data. Later on, Crump (1951) applied maximum likelihood (ML) under the assumption of normally distributed errors and random effects. But it was not until the 1970's when the estimation of variance parameters based on ML methods gained interest. The method of *Restricted Maximum Likelihood* (REML) (Patterson and Thompson 1971) gave a solution to the problem of biased estimators of the variance parameters. However, one of the main obstacles to the use of this technique, at the time, was the fact that the calculation of ML/REML estimates requires the numerical solution of a nonlinear problem. Patterson and Thompson (1971) proposed an iterative solution using the Fisher Scoring algorithm, but it was Harville (1977) who proposed the first numerical algorithm to compute REML estimates of the variance parameters. His proposal is the inspiration of our work.

Along the years, several computational approaches have appeared with the aim of improving the computational burden of solving the score equations for the variance parameters: Smith (1990) proposed the use of the EM algorithm, Graser et al. (1987) suggested the use of the simplex algorithm to obtain the estimates directly from the likelihood, and Gilmour et al. (1995) developed a method based on the use of an average information matrix.

In the context of generalised linear mixed models (GLMMs), estimation based on iterative re-weighted REML has been proposed independently by a number of authors

✉ María Xosé Rodríguez-Álvarez
mxrodriguez@bcamath.org

1 BCAM - Basque Center for Applied Mathematics, Alameda de Mazarredo, 14, 48009 Bilbao, Basque Country, Spain

2 IKERBASQUE, Basque Foundation for Science, Bilbao, Spain

3 Department of Statistics and Econometrics, Universidad Carlos III de Madrid, Leganés, Spain

4 Erasmus University Medical Centre, Rotterdam, The Netherlands

Published online: 11 June 2018

(e.g.,Schall 1991; Engel and Keen 1994), as an extension of the iterative re-weighted least squares algorithm for Generalised Linear Models (GLM, McCullagh and Nelder 1989). Breslow and Clayton (1993) proposed a general method based on penalised quasi-likelihood (PQL) for the estimation of the fixed and random effects, and pseudo-likelihood for the variance parameters. As noted by Engel and Buist (1996), the estimation procedures discussed in all these papers are equivalent, although motivated from quite different starting points.

The majority of the methods mentioned above impose a strong restriction on the vector of random effects: its variance-covariance matrix has to be linear in the variance parameters. The results we present in this paper relax that assumption to the case in which the linearity in the parameters is necessary on the precision matrix and not on the variance-covariance matrix. Our contribution is motivated by the need to estimate smoothing parameters in the context of penalised regression models with non-standard quadratic penalties.

Penalised spline regression (P-splines, Eilers and Marx 1996) has become a popular method for estimating models in which the mean response (or linear predictor in the non-Gaussian case) is a smooth unknown function of one or more covariates. The method is based on the representation of the smooth component in terms of basis functions and the estimation of the parameters by modifying the likelihood with a quadratic penalty on the coefficients. The size of the penalty is controlled by the so-called *smoothing parameter*. The connection between penalised smoothing and linear mixed models was first established a long time ago (Green 1987), and it has become of common use in the last 15 years (Currie and Durban 2002; Currie et al. 2006; Lee 2010; Wand 2003). The key point of the equivalence is that the smoothing parameter becomes the ratio between two variance parameters. Therefore, the methods mentioned above can be used to estimate directly the amount of smoothing needed in the model, instead of using methods based on the minimisation of some prediction error such as the Akaike information criterion (AIC), generalised cross-validation (GCV) or Mallows' C_p (see, e.g., Eilers and Marx 1996; Wood 2008). In Krivobokova (2009), the asymptotic properties of REML and Mallows' C_p smoothing parameter estimators are considered. The paper shows that, in the frequentist framework, the REML oracle smoothing parameter is asymptotically suboptimal. However, the variance of the REML-based estimator is found to be smaller than that of the Mallows' C_p estimator. Also, Reiss and Ogden (2009) show that at finite sample sizes AIC and GCV are prone to undersmoothing and are more likely to develop multiple minima than REML (see also Wood 2017, Chapter 6). An additional argument in favour of REML is that the inclusion of random effects into the penalised regression model is straightforward.

In the smoothing context, standard methods based on REML can be applied when simple penalties are used, i.e., each regression coefficient is affected by a single penalty (by a single smoothing parameter). However, in some circumstances, the penalties present an overlapping structure, with the same coefficients being penalised simultaneously by several smoothing parameters. This includes important cases such as multidimensional penalised splines with anisotropic penalties or adaptive penalised splines. Estimation methods that can deal with this situation have been proposed in the smoothing literature (e.g.,Wood 2011; Wood et al. 2016), but they have the drawback of being very computationally demanding, especially when the number of smoothing parameters is large.

This work addresses this problem and presents a fast method for estimating the variance parameters/smoothing parameters in generalised linear mixed models/generalised additive models. The method can be used whenever the precision matrix of the random component (or the penalty matrix of the P-spline model) is a linear combination defined over the inverse of the variance parameters (smoothing parameters). We obtain simple expressions for the estimates of the variance parameters that are ratios between a sum of squares and a quantity related to the notion of effective degrees of freedom in the smoothing context (Hastie and Tibshirani 1990). We show the sufficient and necessary conditions that guarantee the positiveness of these estimates and discuss several situations where these conditions can be easily verified. Particular cases of the method presented here have been introduced in Rodríguez-Álvarez et al. (2015b), which solved the problem in the case of anisotropic multidimensional smoothing, and in Rodríguez-Álvarez et al. (2015a), where results for adaptive P-splines were first discussed. More recently, Wood and Fasiolo (2017) extended the above-mentioned works to more general penalised spline models. The proposal discussed here presents two main advantages with respect to Wood and Fasiolo (2017)'s approach. First, the smoothing/variance parameter estimates described in Wood and Fasiolo (2017) rely on Moore–Penrose pseudoinverses of the penalty matrices, which, in our experience, may present numerical instabilities. Second, our proposal establishes an explicit connection between variance component estimates and effective degrees of freedom, which lacks in Wood and Fasiolo (2017). Effective degrees of freedom are key components in smoothing models. They help in summarising a model, as partial effective degrees of freedom are measures of model components' complexity with strong intuitive appeal (see, e.g., Rodríguez-Álvarez et al. 2018 for an example in the agricultural field).

The rest of this paper is organised as follows: Sect. 2 introduces the work by Harville (1977), which constitutes the foundation of the work presented here. Section 3 is the core of the paper: the new method, called SOP (separation of overlapping precision matrices), is presented; and the connection between SOP and the notion of effective degrees of

freedom is discussed. Section 4 describes several P-splines models whose estimation can be approached using SOP. We focus in this paper on adaptive P-splines and P-splines for hierarchical curve data. Illustrations with data examples are provided in Sect. 5. A discussion closes the paper. Some technical details have been added as appendices. The estimating algorithm is detailed there.

2 Estimation of variance parameters in generalised linear mixed models: Harville (1977)'s work and extensions

This section is our little tribute to Harville (1977)'s paper, which was the inspiration for this work. Harville (1977)'s paper deals with ML/REML approaches to variance parameters estimation in linear mixed models (LMM) for Gaussian data. Nonetheless, estimation of GLMM can be done by repeated use of LMM methodology on a *working*-dependent variable (see, e.g., Schall 1991; Engel and Keen 1994 where use is made of the results by Harville 1977). This is the approach we follow in this paper.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ be a vector of n observations. A GLMM can be written as

$$g(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{Z\alpha}, \quad \text{with } \boldsymbol{\alpha} \sim N(\boldsymbol{0}, \boldsymbol{G}), \tag{1}$$

where $\mu_i = \mathbb{E}(y_i \mid \boldsymbol{\alpha})$ and $g(\cdot)$ is the link function. The model assumes that, conditional on the random effects, the observations y_i are independently distributed with mean μ_i and variance $\mathbb{V}\text{ar}(y_i|\boldsymbol{\alpha}) = \phi v(\mu_i)$. Here, $v(\cdot)$ is a specified variance function, and ϕ is the dispersion parameter that may be known or unknown. In model (1), $\boldsymbol{X}$ and $\boldsymbol{Z}$ represent column-partitioned matrices, associated, respectively, with the fixed and random effects. We assume that $\boldsymbol{X}$ has full rank, $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_c]$, and $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_c^\top\right)^\top$. Each $\boldsymbol{Z}_k$ corresponds to the design matrix of the k-th random component $\boldsymbol{\alpha}_k$, with $\boldsymbol{\alpha}_k$ being a $(q_k \times 1)$ vector $(k = 1, \ldots, c)$. We assume further that $\boldsymbol{\alpha}_k \sim N(\boldsymbol{0}, \boldsymbol{G}_k)$ and that

$$\boldsymbol{G} = \bigoplus_{k=1}^{c} \boldsymbol{G}_k = \bigoplus_{k=1}^{c} \sigma_k^2 \boldsymbol{I}_{q_k} = \text{diag}\left(\sigma_1^2 \boldsymbol{I}_{q_1}, \ldots, \sigma_c^2 \boldsymbol{I}_{q_c}\right),$$

where $\boldsymbol{I}_m$ is an identity matrix of order $m \times m$ and $\bigoplus$ denotes the direct sum of matrices. Note that the variance–covariance matrix $\boldsymbol{G}$ is linear in the variance parameters σ_k^2.

As noted before, estimation of model (1) can be approached by iterative fitting of a LMM that involves a *working* dependent variable $\boldsymbol{z}$ and a weight matrix $\boldsymbol{W}$ (updated at each iteration). The specific form of $\boldsymbol{z}$ and $\boldsymbol{W}$ is given in "Appendix C". If ϕ and σ_k^2 $(k = 1, \ldots, c)$ are known, at each iteration, the updates for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ follow from the so-called Henderson equations (Henderson 1963)

$$\underbrace{\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{R}^{-1} \boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix}}_{\boldsymbol{C}} \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{z} \\ \boldsymbol{Z}^\top \boldsymbol{R}^{-1} \boldsymbol{z} \end{bmatrix}, \tag{2}$$

which give rise to closed-form expressions

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= \left(\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{z}, \\ \widehat{\boldsymbol{\alpha}}_k &= \boldsymbol{G}_k \boldsymbol{Z}_k^\top \boldsymbol{P} \boldsymbol{z} \qquad (k = 1, \ldots, c), \end{aligned} \tag{3}$$

where $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{V}^{-1}$ with $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top$ and $\boldsymbol{R} = \phi \boldsymbol{W}^{-1}$. The Henderson equations are of little use if the variances parameters ϕ and σ_k^2 $(k = 1, \ldots, c)$ are unknown. In his 1977 paper, Harville shows how to estimate them by REML by an elegant iterative algorithm. Let's first define

$$\boldsymbol{T} = \left(\boldsymbol{I} + \boldsymbol{Z}^\top \boldsymbol{S} \boldsymbol{Z} \boldsymbol{G}\right)^{-1},$$

where $\boldsymbol{S} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{R}^{-1}\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{R}^{-1}$. We note that $\boldsymbol{T}$ can be partitioned as follows

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_{11} & \boldsymbol{T}_{12} & \cdots & \boldsymbol{T}_{1c} \\ \boldsymbol{T}_{21} & \boldsymbol{T}_{22} & \cdots & \boldsymbol{T}_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{T}_{c1} & \boldsymbol{T}_{c2} & \cdots & \boldsymbol{T}_{cc} \end{bmatrix},$$

where $\boldsymbol{T}_{ij}$ are matrices of order $q_i \times q_j$. In Harville (1977), the updated estimate of σ_k^2 $(k = 1, \ldots, c)$ is

$$\widehat{\sigma}_k^2 = \frac{\widehat{\boldsymbol{\alpha}}_k^{\top[t]} \widehat{\boldsymbol{\alpha}}_k^{[t]}}{\text{ED}_k^{[t]}}, \tag{4}$$

where

$$\text{ED}_k^{[t]} = q_k - \text{trace}\left(\boldsymbol{T}_{kk}^{[t]}\right), \tag{5}$$

and the superscript $[t]$ denotes quantities evaluated at current estimates of the variance parameters. From the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ follow an estimate for $\boldsymbol{z}$: $\widehat{\boldsymbol{z}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{Z}\widehat{\boldsymbol{\alpha}}$. The residuals are $\boldsymbol{z} - \hat{\boldsymbol{z}}$. Harville uses

$$\hat{\phi} = \frac{\boldsymbol{z}^\top \boldsymbol{W}\left(\boldsymbol{z} - \widehat{\boldsymbol{z}}^{[t]}\right)}{n - \text{rank}(\boldsymbol{X})}, \tag{6}$$

to estimate the dispersion parameter (not always needed in GLMM). An alternative expression is (see, e.g., Engel 1990; Rodríguez-Álvarez et al. 2015b)

$$\hat{\phi} = \frac{\left(\boldsymbol{z} - \widehat{\boldsymbol{z}}^{[t]}\right)^\top \boldsymbol{W}\left(\boldsymbol{z} - \widehat{\boldsymbol{z}}^{[t]}\right)}{n - \text{rank}(\boldsymbol{X}) - \sum_{k=1}^{c} \text{ED}_k^{[t]}}. \tag{7}$$

Here $\text{rank}(\boldsymbol{X}) + \sum_{k=1}^{c} \text{ED}_k^{[t]}$ can be interpreted as the effective model dimension. At convergence, Eqs. (6) and (7) give identical numerical values.

2.1 Effective degrees of freedom in Harville's method

As noted by Harville (1977), the iterates derived from expression (4) have an intuitively appealing form. On each iteration, σ_k^2 is estimated by the ratio between the sum of squares of the estimates for $\boldsymbol{\alpha}_k$ and a number between zero and q_k. We now show that the denominator in expression (4) can, in fact, be interpreted as effective degrees of freedom in smoothing sensu, i.e., as the trace of a "hat" matrix (Hastie and Tibshirani 1990).

First note that expression (3) reveals that $\boldsymbol{Z}_k\widehat{\boldsymbol{\alpha}}_k = \boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{z}$. Thus, the "hat" matrix corresponding to the k-th random component $\boldsymbol{\alpha}_k$ is

$$\boldsymbol{H}_k = \boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P},$$

i.e., $\boldsymbol{H}_k\boldsymbol{z} = \boldsymbol{Z}_k\widehat{\boldsymbol{\alpha}}_k$. We now show that $\text{trace}\,(\boldsymbol{H}_k) = \text{ED}_k$. It is easy to verify that

$$\boldsymbol{T} = \left(\boldsymbol{I} + \boldsymbol{Z}^t\boldsymbol{S}\boldsymbol{Z}\boldsymbol{G}\right)^{-1} = \boldsymbol{G}^{-1}\left(\boldsymbol{G}^{-1} + \boldsymbol{Z}^t\boldsymbol{S}\boldsymbol{Z}\right)^{-1}, \tag{8}$$

where $\left(\boldsymbol{G}^{-1} + \boldsymbol{Z}\boldsymbol{S}\boldsymbol{Z}^t\right)^{-1}$ is that partition of the inverse of $\boldsymbol{C}$ in (2) corresponding to the random vector $\boldsymbol{\alpha}$ (Harville 1977; Johnson and Thompson 1995). Exploiting the block structure of $\boldsymbol{Z}$ and $\boldsymbol{G}$, and making use of result (8) and (A4) in Johnson and Thompson (1995), we have that

$$\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k = \boldsymbol{I}_{q_k} - \boldsymbol{G}_k^{-1}\boldsymbol{C}_{kk}^* = \boldsymbol{I}_{q_k} - \boldsymbol{T}_{kk}, \tag{9}$$

where, to ease the notation, $\boldsymbol{C}^*$ denotes the inverse of $\boldsymbol{C}$ and $\boldsymbol{C}_{kk}^*$ denotes that partition of $\boldsymbol{C}^*$ corresponding to the k-th random component $\boldsymbol{\alpha}_k$. Thus,

$$\begin{aligned}\text{trace}\,(\boldsymbol{H}_k) &= \text{trace}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P}\right) = \text{trace}\left(\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\right)\\ &= \text{trace}\left(\boldsymbol{I}_{q_k} - \boldsymbol{T}_{kk}\right) = q_k - \text{trace}\,(\boldsymbol{T}_{kk})\\ &= \text{ED}_k.\end{aligned}$$

3 Separation of overlapping precision matrices: the SOP method

In the previous section, we have discussed an estimating method for generalised linear mixed models where the variance–covariance matrix of the random component is linear in the variance parameters. However, more complex structures of the variance–covariance matrix appear in practice. The present research was motivated by our work on penalised spline regression. In spite of that, the method to be discussed in this section is not confined to this area: it can be seen as a general estimating method for generalised linear mixed models with a precision matrix of a specific structure. As in Sect. 2, we consider the generalised linear mixed model

$$g\,(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{k=1}^{c}\boldsymbol{Z}_k\boldsymbol{\alpha}_k, \tag{10}$$

with $\boldsymbol{\alpha}_k \sim N\,(\boldsymbol{0}, \boldsymbol{G}_k)$, $\boldsymbol{\alpha} \sim N\,(\boldsymbol{0}, \boldsymbol{G})$, and $\boldsymbol{G} = \bigoplus_{k=1}^{c}\boldsymbol{G}_k$. The main difference with respect to Sect. 2 is that we do not assume that $\boldsymbol{G}_k = \sigma_k^2\boldsymbol{I}_{q_k}$, but we consider precision matrices of the form

$$\boldsymbol{G}_k^{-1} = \sum_{l=1}^{p_k}\sigma_{k_l}^{-2}\boldsymbol{\Lambda}_{k_l}, \tag{11}$$

where $\sigma_{k_l}^2$ ($l = 1, \ldots, p_k$ and $k = 1, \ldots, c$) are the variance parameters and $\boldsymbol{\Lambda}_{k_l}$ are known symmetric positive semi-definite matrices of dimension $q_k \times q_k$. Note that we do not require $\boldsymbol{\Lambda}_{k_l}$ to be positive definite. The only requirement we need is that $\boldsymbol{G}_k^{-1}(k = 1, \ldots, c)$ are positive definite and so are $\boldsymbol{G}^{-1}$ and its inverse, the variance–covariance matrix $\boldsymbol{G}$.

Expression (11) deserves some detailed discussion. Firstly, it is worth noting that we do not work with variance–covariance matrices, but with their inverses, the precision matrices. As said, the developments in this work have their origin on penalised spline methods. In Sect. 4, the need to work with precision matrices will become clear, or, in the terminology of penalised splines, with penalty matrices. Secondly, what constitutes the main contribution of this paper is that we assume that each random component $\boldsymbol{\alpha}_k$ ($k = 1, \ldots, c$) in model (10) may be "affected" (shrunk) by several variance parameters. A particular case would be when $p_k = 1\ \forall k$, in which case we are in the situation discussed in Sect. 2.

For the sake of simplicity, in some cases we will rewrite the precision matrix $\boldsymbol{G}^{-1}$ as follows

$$\boldsymbol{G}^{-1} = \sum_{l=1}^{p}\sigma_l^{-2}\widetilde{\boldsymbol{\Lambda}}_l, \tag{12}$$

where $p = \sum_{k=1}^{c} p_k$. By a slight abuse of notation, let $\boldsymbol{\Lambda}_l$ denote the matrices involved in expression (11). The matrix $\widetilde{\boldsymbol{\Lambda}}_l$ is $\boldsymbol{\Lambda}_l$ padded out with zeroes. Some specific examples will be presented in Sect. 4 below. Expression (12) makes it clear that the present work deals with the situation of generalised linear mixed models with a precision matrix for the random component that is linear in the precision parameters σ_l^{-2}.

The next section presents the proposed estimation method that we call SOP.

3.1 The SOP method

Regardless of the structure of $\boldsymbol{G}$, estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, and, when necessary, the dispersion parameter ϕ, does not pose a problem and can be done as discussed in Sect. 2 above (see also "Appendix C" for a detailed description of the estimating algorithm). Recall that estimates for $\boldsymbol{\alpha}_k$ are obtained as $\widehat{\boldsymbol{\alpha}}_k = \boldsymbol{G}_k \boldsymbol{Z}_k^\top \boldsymbol{P} \boldsymbol{z}$. The hat matrix associated with the k-th random component $\boldsymbol{\alpha}_k$ is, once again, $\boldsymbol{H}_k = \boldsymbol{Z}_k \boldsymbol{G}_k \boldsymbol{Z}_k^\top \boldsymbol{P}$, and the effective degrees of freedom of this component are $\text{ED}_k = \text{trace}\,(\boldsymbol{H}_k)$. The variance parameters $\sigma^2_{k_l}$ ($l = 1, \ldots, p_k$ and $k = 1, \ldots, c$), however, cannot be estimated by means of Harville's approach. This is a consequence of $\boldsymbol{G}$ not being linear in the variance parameters.

The key of our approach is to work with $\boldsymbol{G}^{-1}$ instead of with $\boldsymbol{G}$. Given that $\boldsymbol{G}^{-1}$ is linear in the precision parameters $\sigma^{-2}_{k_l}$, the first-order partial derivatives of the (approximate) REML log-likelihood function can be explicitly obtained, as well as the REML-based estimates of the variance parameters. We state the result in the following Theorem, whose proof is given in "Appendix A".

Theorem 1 *Let $\boldsymbol{G} = \bigoplus_{k=1}^{c} \boldsymbol{G}_k$ be a symmetric positive definite matrix, with $\boldsymbol{G}_k^{-1} = \sum_{l=1}^{p_k} \sigma^{-2}_{k_l} \boldsymbol{\Lambda}_{k_l}$ symmetric positive definite and $\boldsymbol{\Lambda}_{k_l}$ known symmetric positive semi-definite. Then, the REML-based estimate updates of the variance parameters $\sigma^2_{k_l}$ ($l = 1, \ldots, p_k$ and $k = 1, \ldots, c$) are given by*

$$\widehat{\sigma}^2_{k_l} = \frac{\widehat{\boldsymbol{\alpha}}_k^{[t]\top} \boldsymbol{\Lambda}_{k_l} \widehat{\boldsymbol{\alpha}}_k^{[t]}}{ED_{k_l}^{[t]}}, \tag{13}$$

where

$$ED_{k_l}^{[t]} = trace\left(\boldsymbol{Z}_k^\top \boldsymbol{P}^{[t]} \boldsymbol{Z}_k \boldsymbol{G}_k^{[t]} \frac{\boldsymbol{\Lambda}_{k_l}}{\widehat{\sigma}^{2[t]}_{k_l}} \boldsymbol{G}_k^{[t]} \right), \tag{14}$$

with $\widehat{\boldsymbol{\alpha}}^{[t]}$, $\boldsymbol{P}^{[t]}$ and $\boldsymbol{G}^{[t]}$ evaluated at the current estimates $\widehat{\sigma}^{2[t]}_{k_l}$ ($l = 1, \ldots, p_k$ and $k = 1, \ldots, c$), and, when necessary, $\widehat{\phi}^{[t]}$.

Note that when $\boldsymbol{G}_k = \sigma^2_k \boldsymbol{I}_{q_k}$, expressions (13) and (14) reduce to those of Harville [expressions (4) and (5), respectively].

An important and desirable property of the updates given in expression (13) is that they are always non-negative, provided that the previous estimates of the variance parameters are non-negative. In addition, under rather weak conditions, these updates are strictly positive (although it is possible to obtain values very close to zero).

Theorem 2 *If $\widehat{\sigma}^{2[t]}_{k_l} > 0$, then the REML-based estimate updates of the variance parameters given in expression (13) are larger or equal than zero, with strict inequality holding if: (i) rank $\left(\boldsymbol{X}, \boldsymbol{Z}_k \boldsymbol{G}_k^{[t]} \boldsymbol{\Lambda}_{k_l}\right) >$ rank $(\boldsymbol{X})$; and (ii) $\boldsymbol{z}$ (the working response vector) is not in the space spanned by the columns of $\boldsymbol{X}$.*

The proof is provided in "Appendix B". We note that condition (i) is needed for both the numerator and denominator of expression (13) to be strictly positive, while (ii) is only needed for the numerator. From an applied point of view, it would undoubtedly important to be able to check whether the conditions are fulfilled before fitting the model. This may not be an easy task, since they depend on $\boldsymbol{G}_k^{[t]}$ and thus may vary from iteration to iteration. There are, however, common situations where condition (i) could be checked in advance:

- If $\boldsymbol{\Lambda}_{k_l}$ is of full rank, then condition (i) simplifies to rank $(\boldsymbol{X}, \boldsymbol{Z}_k) >$ rank $(\boldsymbol{X})$. We note that this condition is the same as that discussed by Harville (1977) in Lemma 1.
- If $\boldsymbol{G}_k^{[t]}$ and $\boldsymbol{\Lambda}_{k_l}$ commute (i.e., $\boldsymbol{G}_k^{[t]} \boldsymbol{\Lambda}_{k_l} = \boldsymbol{\Lambda}_{k_l} \boldsymbol{G}_k^{[t]}$), then condition (i) simplifies to rank $\left(\boldsymbol{X}, \boldsymbol{Z}_k \boldsymbol{\Lambda}_{k_l}\right) >$ rank $(\boldsymbol{X})$. Examples when $\boldsymbol{G}_k^{[t]}$ and $\boldsymbol{\Lambda}_{k_l}$ commute include, for instance, when both are diagonal.

We discuss these situations in more detail in Sect. 4, where some examples of application of the SOP method are presented.

3.2 Effective degrees of freedom in the SOP method

In line with the Harville method discussed in Sect. 2, the denominator of expression (13) has been denoted as $\text{ED}_{\{\cdot\}}$, from effective degrees of freedom. Result (14) makes it easy to show that the sum of the ED_{k_l} over the p_k variance parameters involved in $\boldsymbol{G}_k^{-1}$ [see Eq. (11)] corresponds to ED_k (the effective degrees of freedom of $\boldsymbol{\alpha}_k$)

$$\begin{aligned} \sum_{l=1}^{p_k} \text{ED}_{k_l} &= \sum_{l=1}^{p_k} \text{trace}\left(\boldsymbol{Z}_k^\top \boldsymbol{P} \boldsymbol{Z}_k \boldsymbol{G}_k \frac{\boldsymbol{\Lambda}_{k_l}}{\sigma^2_{k_l}} \boldsymbol{G}_k \right) \\ &= \text{trace}\left(\boldsymbol{Z}_k^\top \boldsymbol{P} \boldsymbol{Z}_k \boldsymbol{G}_k \right) = \text{trace}\,(\boldsymbol{H}_k) = \text{ED}_k. \end{aligned}$$

As a consequence, at convergence, the estimated effective degrees of freedom associated with each random component in model (10) are obtained as a by-product of the SOP method.

To finish this part, we would like to point out an interesting link between the upper bound for ED_{k_l} (denoted with $\text{ED}^{ub}_{k_l}$) and condition (i) in Theorem 2. It can be shown that

$$\begin{aligned}\text{ED}_{k_l} &\leq \text{rank}\left(\boldsymbol{X}, \boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\right) - \text{rank}\left(\boldsymbol{X}\right)\\ &= \text{rank}\left((I_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\right) = \text{ED}_{k_l}^{ub},\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{X}} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top$. Thus, if condition (i) in Theorem 2 is not verified, ED_{k_l} would be exactly zero. We omit the proof of the previous result. It can be obtained in a similar fashion as in the paper by Cui et al. (2010) (see Web Appendix (f) in that paper), by noting that

$$\begin{aligned}\text{ED}_{k_l} &= \text{trace}\left(\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\frac{\boldsymbol{\Lambda}_{k_l}}{\sigma^2_{k_l}}\boldsymbol{G}_k\right)\\ &= \text{trace}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\frac{\boldsymbol{\Lambda}_{k_l}}{\sigma^2_{k_l}}\boldsymbol{G}_k\boldsymbol{Z}_k^\top[(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})\,\boldsymbol{V}\,(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}})]^+\right),\end{aligned}$$

where $\boldsymbol{\Gamma}^+$ denotes the Moore–Penrose pseudoinverse of $\boldsymbol{\Gamma}$. This equivalence has been proved, in a less general situation, by Rodríguez-Álvarez et al. (2018) (see Web Appendix D in that paper). Following a similar reasoning, we obtain the upper bound for the effective degrees of freedom of the k-th random component

$$\begin{aligned}\text{ED}_k &\leq \text{rank}\left(\boldsymbol{X}, \boldsymbol{Z}_k\right) - \text{rank}\left(\boldsymbol{X}\right)\\ &= \text{rank}\left((I_n - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{Z}_k\right) = \text{ED}_k^{ub}.\end{aligned}$$

Using the well-known result that the rank of a matrix sum cannot exceed the sum of the ranks of the summand matrices, we have that $\text{ED}_k^{ub} \leq \sum_{l=1}^{p_k}\text{ED}_{k_l}^{ub}$. In general, however, this is a strict inequality, since most of the cases

$$\mathcal{C}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_u}\right) \cap \mathcal{C}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_v}\right) \neq \{\boldsymbol{0}\} \quad (u \neq v),$$

where $\mathcal{C}\left(\boldsymbol{A}\right)$ denotes the linear space spanned by the columns of $\boldsymbol{A}$ (see, e.g., Theorem 18.5.7 in Harville 1997). Intuitively, we can interpret this as a sort of competition among the p_k "elements" associated with the k-th random component. The ED_{k_l} cannot vary "free" between 0 and $\text{ED}_{k_l}^{ub}$, but they have to fulfil that their sum does not excess ED_k^{ub}.

3.3 Computational aspects

From a computational point of view, the evaluation of the expression given in (14) can be very costly. However, this computation can be relaxed by using the result given in (9). For our purpose, it is easy to show that

$$\begin{aligned}\text{trace}\left(\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\boldsymbol{G}_k\right) &= \text{trace}\left(\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\right)\\ &= \text{trace}\left(\left(\boldsymbol{G}_k - \boldsymbol{C}^*_{kk}\right)\boldsymbol{\Lambda}_{k_l}\right).\end{aligned}$$

This result is exploited in the estimating algorithm described in "Appendix C". We note that $\boldsymbol{C}^*$ [i.e., the inverse of $\boldsymbol{C}$ in (2)] is computed in order to estimate $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$. Thus, the computation of the traces needed to implement the algorithm reduces to the computation of the diagonal elements of $\left(\boldsymbol{G}_k - \boldsymbol{C}^*_{kk}\right)\boldsymbol{\Lambda}_{k_l}$. In addition, in those cases where $\boldsymbol{\Lambda}_{k_l}$ is diagonal, only the element-wise product of the diagonals of $\left(\boldsymbol{G}_k - \boldsymbol{C}^*_{kk}\right)$ and $\boldsymbol{\Lambda}_{k_l}$ is needed. This will considerably reduce the number of operations required and therefore the computing time.

4 Penalised smoothing and the SOP method

This section discusses several situations in the P-spline framework where estimation can be approached using the SOP method. As it will be seen, the method can be used whenever there are multiple penalties acting on the same coefficients. Anisotropic tensor-product P-splines is an example of overlapping penalties, and it has been extensively discussed in the paper by Rodríguez-Álvarez et al. (2015b). However, multiple penalties arise in a broader class of situations. We describe here two of those: spatially adaptive P-splines and P-splines for hierarchical curve data.

4.1 Spatially adaptive P-splines

Consider a generalised regression problem

$$g\left(\mu_i\right) = f\left(x_i\right), \qquad i = 1, \ldots n, \tag{15}$$

where $\mu_i = \mathbb{E}\left(y_i\right)$, $g(\cdot)$ is the link function and $f(\cdot)$ is a smooth and unknown function. We assume further that $\mathbb{V}\text{ar}\left(y_i\right) = \phi v\left(\mu_i\right)$. In the P-spline framework (Eilers and Marx 1996), the unknown function $f(x)$ is approximated by a linear combination of d B-splines basis functions, i.e., $f(x) = \sum_{j=1}^{d}\theta_j B_j\left(x\right)$. In matrix notation, model (15) is thus expressed as

$$g\left(\boldsymbol{\mu}\right) = \boldsymbol{B}\boldsymbol{\theta}, \tag{16}$$

where $\boldsymbol{\mu} = \left(\mu_1, \mu_2, \ldots, \mu_n\right)^\top, \boldsymbol{\theta} = \left(\theta_1, \theta_2, \ldots, \theta_d\right)^\top$ and $\boldsymbol{B}$ is a B-spline regression matrix of dimension $n \times d$, i.e., $b_{ij} = B_j\left(x_i\right)$ is the j-th B-spline evaluated at x_i. Smoothness is achieved by imposing a penalty on the regression coefficients $\boldsymbol{\theta}$ in the form

$$\lambda\sum_{k=q+1}^{d}\left(\Delta^q\theta_k\right)^2 = \lambda\boldsymbol{\theta}^\top\boldsymbol{D}_q^\top\boldsymbol{D}_q\boldsymbol{\theta}, \tag{17}$$

where λ is the smoothing parameter and Δ^q forms differences of order q on adjacent coefficients, i.e., $\Delta\theta_k = \theta_k - \theta_{k-1}$, $\Delta^2\theta_k = \Delta\left(\Delta\theta_k\right) = \theta_k - \theta_{k-1} - \left(\theta_{k-1} - \theta_{k-2}\right) = \theta_k - 2\theta_{k-1} + \theta_{k-2}$, and so on for higher q. Finally, $\boldsymbol{D}_q$ is simply the matrix representation of Δ^q.

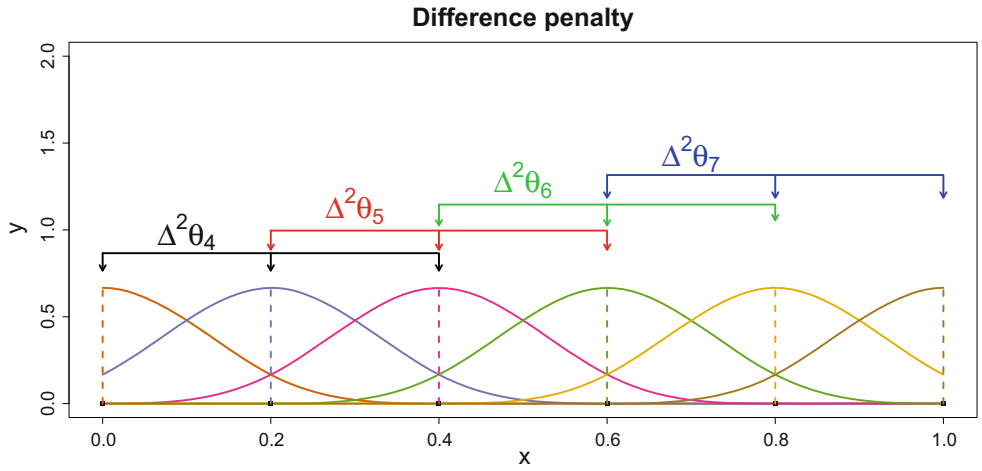

Fig. 1 Graphical representation of differences of order 2 on adjacent coefficients of cubic B-splines basis functions. Note the local and ordered nature of these differences

As can be observed in Eqn. (17), the same smoothing parameter λ applies to all coefficient differences, irrespective of their location (see Fig. 1). Thus, the model assumes that the same amount of smoothing is needed across the whole domain of the covariate. Adaptive P-splines (see, e.g., Krivobokova et al. 2008; Ruppert and Carroll 2000 among others) relax this assumption. The idea is simple, to replace the global smoothing parameter by smoothing parameters that vary locally according to the covariate value. This can be accomplished by specifying a different smoothing parameter for each coefficient difference (Ruppert and Carroll 2000; Wood 2011)

$$\sum_{k=q+1}^{d} \lambda_{k-q} \left(\Delta^q \theta_k\right)^2 = \boldsymbol{\theta}^\top \boldsymbol{D}_q^\top \text{diag}(\boldsymbol{\lambda}) \boldsymbol{D}_q \boldsymbol{\theta}, \tag{18}$$

where $\boldsymbol{\lambda} = \left(\lambda_1, \ldots, \lambda_{d-q}\right)^\top$. Note that this approach would imply as many smoothing parameters as coefficient differences (i.e., $d-q$), which could lead to undersmoothing and unstable computations. Given the local and ordered nature of the coefficient differences (see Fig. 1), we may model the smoothing parameters λ_k as a smooth function of k (its position) and use B-splines for this purpose (here no penalty is assumed)

$$\boldsymbol{\lambda} = \boldsymbol{\Psi} \boldsymbol{\xi}, \tag{19}$$

where $\boldsymbol{\Psi}$ is a B-spline regression matrix of dimension $(d-q) \times p$ with $p < (d-q)$ and $\boldsymbol{\xi} = \left(\xi_1, \ldots, \xi_p\right)^\top$ is the new vector of smoothing parameters. Performing some simple algebraic operations, it can be shown that the *adaptive* penalty (18) is

$$\boldsymbol{\theta}^\top \left(\sum_{l=1}^{p} \xi_l \boldsymbol{D}_q^\top \text{diag}\left(\boldsymbol{\psi}_l\right) \boldsymbol{D}_q \right) \boldsymbol{\theta}, \tag{20}$$

where $\boldsymbol{\psi}_l$ denotes the column l of $\boldsymbol{\Psi}$. Note that under this adaptive penalty, all coefficients are penalised by multiple smoothing parameters, i.e., there are overlapping penalties.

4.1.1 Mixed model reparametrisation

Estimation of the P-spline model (16) subject to the adaptive penalty defined in (20) can be carried out based on the connection between P-splines and mixed models (e.g., Currie and Durban 2002; Wand 2003). It is easy to show that the null space (i.e., the unpenalised function space) of the adaptive penalty matrix $\boldsymbol{P}_{Ad} = \sum_{l=1}^{p} \xi_l \boldsymbol{D}_q^\top \text{diag}\left(\boldsymbol{\psi}_l\right) \boldsymbol{D}_q$ is independent of $\boldsymbol{\xi}$. In addition, note that when $\xi_u = \xi_v = \lambda \; (\forall \; u, v)$ then

$$\begin{aligned} \boldsymbol{P}_{Ad} &= \sum_{l=1}^{p} \xi_l \boldsymbol{D}_q^\top \text{diag}\left(\boldsymbol{\psi}_l\right) \boldsymbol{D}_q = \lambda \boldsymbol{D}_q^\top \left(\sum_{l=1}^{p} \text{diag}\left(\boldsymbol{\psi}_l\right) \right) \boldsymbol{D}_q \\ &= \lambda \boldsymbol{D}_q^\top \boldsymbol{D}_q = \boldsymbol{P}. \end{aligned}$$

This is a consequence of the rows of a B-spline regression matrix adding up to 1. Thus, the null space of $\boldsymbol{P}_{Ad}$ is the same as that of $\boldsymbol{P}$. Different reparametrisations of P-spline models have been suggested in the literature (see, e.g., Currie and Durban 2002; Eilers 1999), all aiming to decompose the model into the unpenalised and the penalised part. The consequence of this decomposition is that the penalty matrix of the reparametrised P-spline model is of full rank and so is the precision matrix of the corresponding mixed model. For our application, we use the proposal given in Eilers (1999). As will be seen, this approach gives rise to diagonal precision matrices. As discussed in Sect. 3.3, this is very convenient

from a computational point of view. Using Eilers (1999)'s transformation, model (16) is re-expressed as

$$g(\boldsymbol{\mu}) = \boldsymbol{B}\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha},$$

where $\boldsymbol{X} = \left[\mathbf{1}_n | \boldsymbol{x} | \ldots | \boldsymbol{x}^{(q-1)}\right]$ and $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{D}_q^\top \left(\boldsymbol{D}_q \boldsymbol{D}_q^\top\right)^{-1}$, and the precision (penalty) matrix of the vector of random (penalised) coefficients $\boldsymbol{\alpha}$ becomes

$$\boldsymbol{G}^{-1} = \frac{1}{\phi}\boldsymbol{F}^\top \boldsymbol{P}_{Ad} \boldsymbol{F} = \sum_{l=1}^{p} \sigma_l^{-2} \text{diag}\left(\boldsymbol{\psi}_l\right) = \sum_{l=1}^{p} \sigma_l^{-2} \widetilde{\boldsymbol{\Lambda}}_l, \tag{21}$$

where $\boldsymbol{F} = \boldsymbol{D}_q^\top \left(\boldsymbol{D}_q \boldsymbol{D}_q^\top\right)^{-1}$, $\widetilde{\boldsymbol{\Lambda}}_l = \text{diag}\left(\boldsymbol{\psi}_l\right)$, and $\sigma_l^2 = \phi/\xi_l$. Thus, the precision matrix is linear in the precision parameters ${\sigma_l}^{-2}$, and the SOP method can therefore be used.

We note that the model has a single random component ($c = 1$). The reparametrisation ensures that $\text{rank}(\boldsymbol{X}, \boldsymbol{Z}) = \text{rank}(\boldsymbol{X}) + \text{rank}(\boldsymbol{Z})$. Thus, $\text{rank}(\boldsymbol{X}, \boldsymbol{Z}\widetilde{\boldsymbol{\Lambda}}_l) = \text{rank}(\boldsymbol{X}) + \text{rank}(\boldsymbol{Z}\widetilde{\boldsymbol{\Lambda}}_l) > \text{rank}(\boldsymbol{X})$. Exploiting the fact that $\boldsymbol{G}$ and $\widetilde{\boldsymbol{\Lambda}}_l$ are diagonal, and thus commute, condition (i) of Theorem 2 is satisfied.

4.2 P-splines for hierarchical curve data

For simplicity, let's assume balanced hierarchical curve data. Our data consist on m individuals each with s different measurements at times $\boldsymbol{t} = (t_1, t_2, \ldots, t_s)$. Our interest is focused on

$$g\left(\mu_{ij}\right) = f\left(t_i\right) + g_j\left(t_i\right), \quad 1 \le i \le s, \ 1 \le j \le m, \tag{22}$$

where $\mu_{ij} \equiv \mathbb{E}\left(y_{ij}\right)$, y_{ij} is the response variable on the j-th subject at time t_i, $f(\cdot)$ is a function describing the population effect, and $g_j(\cdot)$ are random functions measuring the deviation of the j-th subject from the population effect. As before, $\mathbb{V}\text{ar}\left(y_{ij}\right) \equiv \phi v\left(\mu_{ij}\right)$. A simple model would consist in a parametric specification for $f(\cdot)$ and $g_j(\cdot)$, e.g., $f(t) = \beta_0 + \beta_1 t$ and $g_i(t) = \alpha_{0j} + \alpha_{1j} t$, with $\boldsymbol{\alpha}_0 = (\alpha_{01}, \ldots, \alpha_{0m})^\top \sim N\left(\mathbf{0}, \sigma_0^2 \boldsymbol{I}_m\right)$ and $\boldsymbol{\alpha}_1 = (\alpha_{11}, \ldots, \alpha_{1m})^\top \sim N\left(\mathbf{0}, \sigma_1^2 \boldsymbol{I}_m\right)$.

A more flexible approach consists in assuming $f(\cdot)$ and $g_j(\cdot)$ smooth and unknown. Important contributions in the P-spline framework can be found in Durban et al. (2005) and Ruppert et al. (2003). Both approaches are based on modelling $f(\cdot)$ and $g_j(\cdot)$ using truncated line basis, and estimation is based on penalised methods and linear mixed model techniques. More recently, Djeundje and Currie (2010) extend those models and propose the inclusion of an extra penalty for the individual curve coefficients. The authors argue that this extra penalty is needed to address identifiability issues when estimating the population effect. Under the so-called `M0` model in Djeundje and Currie (2010)'s paper, model (22) is expressed in matrix notation as

$$g\left(\boldsymbol{\mu}_{\cdot j}\right) = \underbrace{\boldsymbol{B}\boldsymbol{\theta}}_{f(\boldsymbol{t})} + \underbrace{\breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}_j}_{g_j(\boldsymbol{t})},$$

where $\boldsymbol{\mu}_{\cdot j} = \left(\mu_{1j}, \ldots, \mu_{sj}\right)^\top$ is the mean response vector for the j-th individual and $\boldsymbol{B}$ and $\breve{\boldsymbol{B}}$ are B-spline regression matrices of possibly different size for, respectively, the population curve and the individual deviation. The vector $\boldsymbol{\theta}$ is assumed fixed, but subject to a q-th order penalty of the form $\boldsymbol{P} = \lambda_1 \boldsymbol{D}_q^\top \boldsymbol{D}_q$, and $\breve{\boldsymbol{\theta}}_j$ is a random vector with distribution $N\left(\mathbf{0}, \phi \breve{\boldsymbol{P}}^{-1}\right)$ where

$$\breve{\boldsymbol{P}} = \lambda_2 \breve{\boldsymbol{D}}_{\breve{q}}^\top \breve{\boldsymbol{D}}_{\breve{q}} + \lambda_3 \boldsymbol{I}_{\breve{d}}. \tag{23}$$

The first term $\lambda_2 \breve{\boldsymbol{D}}_{\breve{q}}^\top \breve{\boldsymbol{D}}_{\breve{q}}$ is responsible for the smoothness of the individuals curves, whereas $\lambda_3 \boldsymbol{I}_{\breve{d}}$ addresses the identifiability aspect (see Djeundje and Currie 2010 for more details). Note that each random effect is shrunk (penalised) by both smoothing parameters λ_2 and λ_3, and thus the precision matrix $\breve{\boldsymbol{G}}^{-1} = 1/\phi \breve{\boldsymbol{P}}$ is linear in the precision parameters

$$\breve{\boldsymbol{G}}^{-1} = \sum_{l=2}^{3} \sigma_l^{-2} \breve{\boldsymbol{\Lambda}}_l, \tag{24}$$

where $\sigma_2^2 = \phi/\lambda_2$ and $\sigma_3^2 = \phi/\lambda_3$ and $\breve{\boldsymbol{\Lambda}}_2 = \breve{\boldsymbol{D}}_{\breve{q}}^\top \breve{\boldsymbol{D}}_{\breve{q}}$ and $\breve{\boldsymbol{\Lambda}}_3 = \boldsymbol{I}_{\breve{d}}$.

In more compact way, we express the model for the whole sample as

$$g(\boldsymbol{\mu}) = [\mathbf{1}_m \otimes \boldsymbol{B}]\boldsymbol{\theta} + [\boldsymbol{I}_m \otimes \breve{\boldsymbol{B}}]\breve{\boldsymbol{\theta}}, \tag{25}$$

where $\otimes$ denotes the Kronecker product, $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_{\cdot 1}^\top, \ldots, \boldsymbol{\mu}_{\cdot m}^\top\right)^\top$, $\breve{\boldsymbol{\theta}} = \left(\breve{\boldsymbol{\theta}}_1^\top, \ldots, \breve{\boldsymbol{\theta}}_m^\top\right)^\top$ and

$$\breve{\boldsymbol{\theta}} \sim N\left(\mathbf{0}, \boldsymbol{I}_m \otimes \breve{\boldsymbol{G}}\right).$$

One last step is needed in order to apply the SOP method for the estimation of model (25): the decomposition of the population effect into the unpenalised and the penalised part. We use here the approach based on the eigenvalue decomposition (EVD) of the penalty. Let $\boldsymbol{D}_q^\top \boldsymbol{D}_q = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ be the EVD of $\boldsymbol{D}_q^\top \boldsymbol{D}_q$. Here $\boldsymbol{U}$ denotes the matrix of eigenvectors and $\boldsymbol{\Sigma}$ the diagonal matrix of eigenvalues. Let us also denote by $\boldsymbol{U}_+(\boldsymbol{\Sigma}_+)$ and $\boldsymbol{U}_0(\boldsymbol{\Sigma}_0)$ the sub-matrices corresponding to the nonzero and zero eigenvalues, respectively. In this case, model (25) is re-expressed as

$$g(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha},$$

where $\boldsymbol{\beta} = \boldsymbol{U}_+^\top \boldsymbol{\theta}$, $\boldsymbol{\alpha} = \left(\left(\boldsymbol{U}_0^\top \boldsymbol{\theta} \right)^\top, \breve{\boldsymbol{\theta}}^\top \right)^\top$, $\boldsymbol{X} = [\mathbf{1}_m \otimes \boldsymbol{B}\boldsymbol{U}_0]$ and $\boldsymbol{Z} = [\mathbf{1}_m \otimes \boldsymbol{B}\boldsymbol{U}_+ : \boldsymbol{I}_m \otimes \breve{\boldsymbol{B}}]$. Finally, $\boldsymbol{\alpha} \sim N(\mathbf{0}, \boldsymbol{G})$, where

$$\boldsymbol{G}^{-1} = \begin{pmatrix} \frac{1}{\sigma_1^2}\boldsymbol{\Sigma}_+ & \mathbf{0}_{d \times \left(m\breve{d}\right)} \\ \mathbf{0}_{\left(m\breve{d}\right) \times c} & \boldsymbol{I}_m \otimes \breve{\boldsymbol{G}}^{-1} \end{pmatrix} = \sum_{l=1}^{3} \sigma_l^{-2} \widetilde{\boldsymbol{\Lambda}}_l,$$

with $\sigma_l^2 = \frac{\phi}{\lambda_l}$ and

$$\widetilde{\boldsymbol{\Lambda}}_1 = \begin{pmatrix} \boldsymbol{\Sigma}_+ & \mathbf{0}_{d \times \left(m\breve{d}\right)} \\ \mathbf{0}_{\left(m\breve{d}\right) \times d} & \mathbf{0}_{\left(m\breve{d}\right) \times \left(m\breve{d}\right)} \end{pmatrix}$$

$$\widetilde{\boldsymbol{\Lambda}}_2 = \begin{pmatrix} \mathbf{0}_{d \times d} & \mathbf{0}_{d \times \left(m\breve{d}\right)} \\ \mathbf{0}_{\left(m\breve{d}\right) \times d} & \boldsymbol{I}_m \otimes \breve{\boldsymbol{\Lambda}}_2 \end{pmatrix}$$

$$\widetilde{\boldsymbol{\Lambda}}_3 = \begin{pmatrix} \mathbf{0}_{d \times d} & \mathbf{0}_{d \times \left(m\breve{d}\right)} \\ \mathbf{0}_{\left(m\breve{d}\right) \times d} & \boldsymbol{I}_m \otimes \breve{\boldsymbol{\Lambda}}_3 \end{pmatrix}.$$

The precision matrix is linear in the precision parameters and thus appropriate for the SOP method. There are, in this case, two random components, modelling, respectively, the population curve and the individual deviations, with

$$\boldsymbol{Z}_1 = \mathbf{1}_m \otimes \boldsymbol{B}\boldsymbol{U}_+ \quad \text{and} \quad \boldsymbol{G}_1^{-1} = \sigma_1^{-2}\boldsymbol{\Sigma}_+,$$
$$\boldsymbol{Z}_2 = \boldsymbol{I}_m \otimes \breve{\boldsymbol{B}} \quad \text{and}$$
$$\begin{aligned} \boldsymbol{G}_2^{-1} &= \sigma_2^{-2}\boldsymbol{I}_m \otimes \breve{\boldsymbol{\Lambda}}_2 + \sigma_3^{-2}\boldsymbol{I}_m \otimes \breve{\boldsymbol{\Lambda}}_3 \\ &= \sigma_2^{-2}\boldsymbol{I}_m \otimes \breve{\boldsymbol{D}}_{\breve{q}}^\top \breve{\boldsymbol{D}}_{\breve{q}} + \sigma_3^{-2}\boldsymbol{I}_m \otimes \boldsymbol{I}_{\breve{d}}\,. \end{aligned}$$

Recall that $\boldsymbol{X} = \mathbf{1}_m \otimes \boldsymbol{B}\boldsymbol{U}_0$, and note that $\operatorname{rank}(\mathbf{1}_m \otimes \boldsymbol{B}\boldsymbol{U}_0) = \operatorname{rank}(\boldsymbol{B}\boldsymbol{U}_0)$. We now show that condition (i) of Theorem 2 is satisfied for all variance parameters:

σ_1^2: By construction, $\boldsymbol{\Sigma}_+$ is positive definite and of full rank, and $\operatorname{rank}(\boldsymbol{X}, \boldsymbol{Z}_1) = \operatorname{rank}(\boldsymbol{X}) + \operatorname{rank}(\boldsymbol{Z}_1)$. Noting that $\operatorname{rank}(\boldsymbol{X}, \boldsymbol{Z}_1\boldsymbol{G}_1\boldsymbol{\Sigma}_+) = \operatorname{rank}(\boldsymbol{X}, \boldsymbol{Z}_1) > \operatorname{rank}(\boldsymbol{X})$, the condition is verified.

σ_2^2: By, e.g., Corollary 18.2.2 in Harville (1997), $\boldsymbol{G}_2$ and $\boldsymbol{I}_m \otimes \breve{\boldsymbol{D}}_{\breve{q}}^\top \breve{\boldsymbol{D}}_{\breve{q}}$ commute. Thus, as long as $m > 1$, it is easy to show that

$$\begin{aligned} &\operatorname{rank}\left(\boldsymbol{X}, \boldsymbol{Z}_2\boldsymbol{I}_m \otimes \breve{\boldsymbol{D}}_{\breve{q}}^\top \breve{\boldsymbol{D}}_{\breve{q}}\right) \\ &\quad = \operatorname{rank}\left(\mathbf{1}_m \otimes \boldsymbol{B}\boldsymbol{U}_0, \boldsymbol{I}_m \otimes \breve{\boldsymbol{B}}\breve{\boldsymbol{D}}_{\breve{q}}^\top \breve{\boldsymbol{D}}_{\breve{q}}\right) > \operatorname{rank}(\boldsymbol{B}\boldsymbol{U}_0)\,. \end{aligned}$$

σ_3^2: Note that $\boldsymbol{G}_2$ and $\boldsymbol{I}_m \otimes \boldsymbol{I}_{\breve{d}}$ commute. As before, as long as $m > 1$, it is easy to show that

$$\begin{aligned} \operatorname{rank}\left(\boldsymbol{X}, \boldsymbol{Z}_2\boldsymbol{I}_m \otimes \boldsymbol{I}_{\breve{d}}\right) &= \operatorname{rank}\left(\mathbf{1}_m \otimes \boldsymbol{B}\boldsymbol{U}_0, \boldsymbol{I}_m \otimes \breve{\boldsymbol{B}}\right) \\ &> \operatorname{rank}(\boldsymbol{B}\boldsymbol{U}_0)\,. \end{aligned}$$

5 Examples

This section presents several data examples where the SOP method represents a powerful alternative to existing estimation procedures. We discuss three different analyses: the first two examples are concerned with spatially adaptive P-splines, but each of them deals with a different situation regarding complexity and aim; the last example is devoted to illustrating our method for the analysis of hierarchical curve data. All computations were performed in (64-bit) `R` 3.4.4 (Core Team 2018), and a 2.30GHz × 4 Intel® Core™i5 processor computer with 15.6GB of RAM and Ubuntu 16.04 LTS operating system.

5.1 Doppler function

For our first example, we consider the Doppler function. This is a common example in the adaptive smoothing literature and has been discussed by Ruppert and Carroll (2000), Krivobokova et al. (2008) and Tibshirani (2014), among others. Data are generated according to

$$y_i = \sin(4/x_i) + 1.5 + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $x_i \sim U[0, 1]$, $\varepsilon_i \sim N\left(0, 0.2^2\right)$, and $n = 1000$. For fitting the data, we assume the spatially adaptive P-spline model discussed in Sect. 4.1. We compare the performance of the SOP method with that implemented in the `R`-package `mgcv`, version `1.8-23`, and described in Wood (2011). It is worth noticing that both approaches implement in essence the same adaptive P-spline model; the only difference is the estimation procedure (and, possibly, the reparametrisation). In addition, we also fit the model without assuming an adaptive penalty. In this case, the SOP method reduces to Harville's approach (see Sect. 2). In all cases, we use 200 cubic B-splines to represent the smooth function, jointly with second-order differences. For the adaptive approaches, 15 equally spaced cubic B-splines are used for the smoothing parameters [see Eq. (19)]. These values are chosen to provide enough flexibility to the model. Under this configuration, there are a total of 15 variance parameters. Figure 2a shows the true simulated Doppler function. Figure 2b, c shows, respectively, the estimated curves based on the SOP method without and with an adaptive penalty. Results using `mgcv` are depicted in Fig. 2d. As expected, both adaptive approaches perform similarly. With the specified configuration, they are able to capture 7 cycles of the Doppler function. On the other hand, the non-adaptive approach is able to capture only 4 cycles and presents very wiggly estimates, especially on the right-hand side of the covariate domain. In terms of EDs, for the SOP model without adaptive penalty, we obtain a total ED of 95.8 (out of 200). For models with adaptive smoothing, we obtained identical results, i.e., 50.2 (with SOP) and 50.0 (with

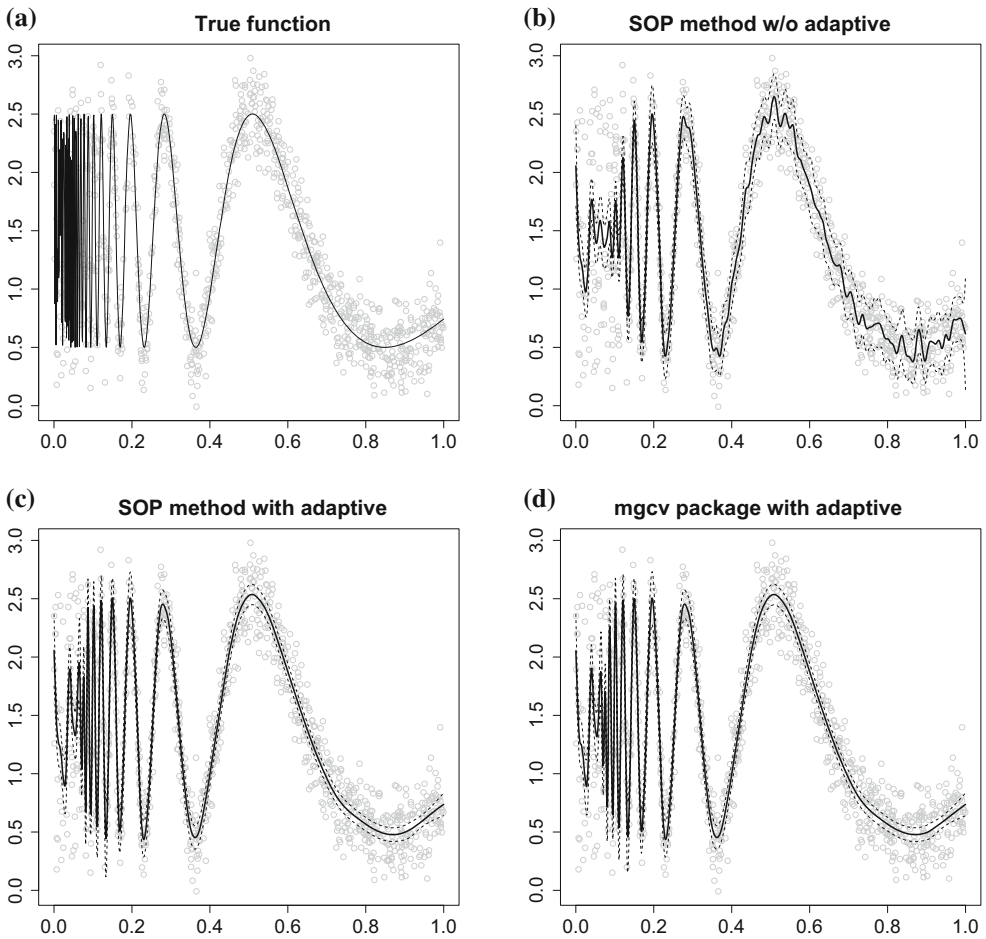

Fig. 2 For the Doppler function: **a** true function (solid line) and simulated data points (grey points), **b** estimated curve using the SOP method without adaptive penalty (solid line) jointly with 95% approximate confidence intervals (dotted lines), **c** estimated curve using the SOP method and adaptive penalty (solid line) jointly with 95% approximate confidence intervals (dotted lines); and **d** estimated curve using the `mgcv` package (solid line) jointly with 95% approximate confidence intervals (dotted lines)

`mgcv`). It is worth remembering that using the SOP method the total ED is obtained by adding up the EDs associated with each variance parameter in the model (plus the dimension of the fixed part). These EDs are the denominator of the estimate update expressions of the variance parameters. The gain of the SOP method is clear when we compare the computing times: 1.0 second with our approach (0.4 if we do not consider adaptive), and 45 seconds using `mgcv`. To gain more insights into the performance of both adaptive approaches beyond the visual similarity shown in Fig. 2 for a single dataset, we run a small simulation study. We perform a total of 100 replicates with the same specification described above and make comparisons regarding root-mean-square error (RMSE) and computing time. In terms of RMSE, the mean (standard deviation) over the 100 replicates is 0.265 (0.0113) for SOP and 0.267 (0.0121) using `mgcv`. Concerning computing times, the mean (standard deviation)—in seconds—is 0.84 (0.22) and 52.79 (30.77) for, respectively, SOP and `mgcv`.

5.2 X-ray diffraction data

For this example, we use data from a X-ray crystallography radiation scan of a thin layer of indium tin oxide. X-ray crystallography allows the exploration of the molecular and atomic structure of crystals. Crystallographers precisely rotate the crystal by entering its desired orientation, while it is illuminated by a beam of X-rays. Depending on the angle, the number of diffracted photons varies and they are detected and counted by an electronic detector. The dataset was analysed by Davies et al. (2008, 2013) and can be found in the R-package `diffractometry` as `indiumoxide`. Figure 3 shows such an X-ray diffraction scan (grey lines). The aim of X-ray diffraction analysis is to determine (a) the signal baseline (and remove it); and (b) the number of peaks (and isolate them to further analysis of their position, height, symmetry, and so forth). This example is solely included to illustrate the potentiality of the method presented in this paper for the anal-

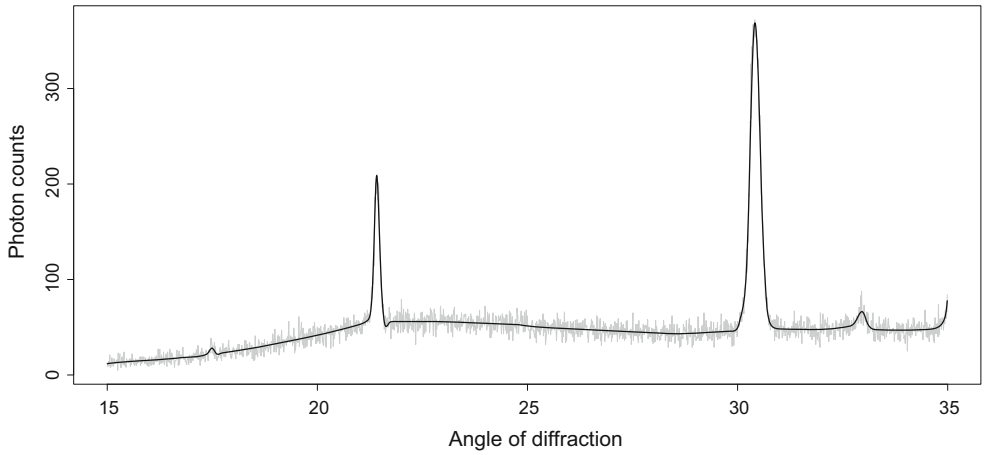

Fig. 3 For the X-ray radiation example: estimated smooth effect of the angle of diffraction on the X-ray radiation using the SOP method (solid black line). The grey lines represent the raw data

ysis of very complex data. For a different modelling approach applied to this dataset, see Camarda et al. (2016). Given that the outcome variable represents count data, a Poisson model is adopted

$$\log\left(\mathbb{E}\left[y_i\right]\right) = \log\left(\mu_i\right) = f\left(x_i\right),$$

where y_i and $x_i (i = 1, \ldots, 2000)$ denote, respectively, the photon counts and the angle of diffraction. To provide enough flexibility to the model in order to make it able to capture the peaks (see Fig. 3), we use 200 cubic B-splines and second-order differences for the function, and 80 cubic B-splines for the adaptive penalty. Results are shown in Fig. 3. The results using the `mgcv` package are almost identical to our proposal and are not depicted. In this case, our method takes less than 3 s, whereas `mgcv` is around 750 times slower (33 min). Regarding the EDs, we obtain a total of 29.5 and 24.9 using SOP and `mgcv`, respectively.

5.3 Diffusion tensor imaging scan data

Our last example deals with hierarchical curve data. We analyse the `DTI` dataset that can be found in the `R`-package `refund` (Goldsmith et al. 2016). A detailed description of the study and data can be found in Goldsmith et al. (2011), Goldsmith et al. (2012) and Greven and Scheipl (2017). In brief, the study aimed at comparing the white matter tracts in patients affected by multiple sclerosis (MS) and healthy individuals. Multiple sclerosis is a disease of the central nervous system that causes lesions in white matter tracts, thus interrupting the travel of nerve impulses to and from the brain and spinal cord. Diffusion tensor imaging (DTI) is a magnetic resonance imaging technique which makes it possible to study the white matter tracts by tracing the diffusion of water on the brain. From DTI scans, fractional anisotropy (FA) measurements can be extracted. FA is related to water diffusion and thus to MS diagnosis and progression. The `DTI` dataset contains FA measurements of healthy and diseased individuals, recorded at several locations along the callosal fibre tract (CCA) and the right corticospinal tract in the brain. In Fig. 4a, the observed FA measurements at different tract locations in the CAA are shown, separately for cases and controls, i.e., individuals affected and non-affected with MS. Each line in these plots represents an individual, and only the first visit is considered. Note that, in general, MS patients present lower FA measurements than healthy individuals.

For illustration purposes, we present two different analyses and comparisons. We first focus our interest on the subgroup of individuals affected with MS. In this group, there are a total of $m = 99$ individuals, each with $s = 93$ FA measurements at different CAA tract locations. The SOP method is used to estimate the model described in Sect. 4.2 (for Gaussian homoskedastic errors) and presented in Djeundje and Currie (2010). To compare results and computing times, the code associated with the paper by Djeundje and Currie (2010) is also tested. For this example, both implementations take the advantage of the array structure of the data: generalised linear array models (GLAM,Currie et al. 2006) are used to efficiently compute the inner products for the Henderson equations [Eq. (27)]. Here, 43 cubic B-spline basis are used for the population curve, and 23 for the individual curves. This configuration gives rise to a model with 2320 $(= 43 + 23 \times 99)$ coefficients (both random and fixed). SOP method needs about 150 seconds to fit the model, and Djeundje and Currie (2010)'s code is 14 times slower. We note that the computational time can be further improved if the sparse structure of the matrices involved in the model is exploited. Using the `R`-package `Matrix`, we are able to reduce the computing time using SOP to 35 seconds. Figure 4b shows the estimated population effect

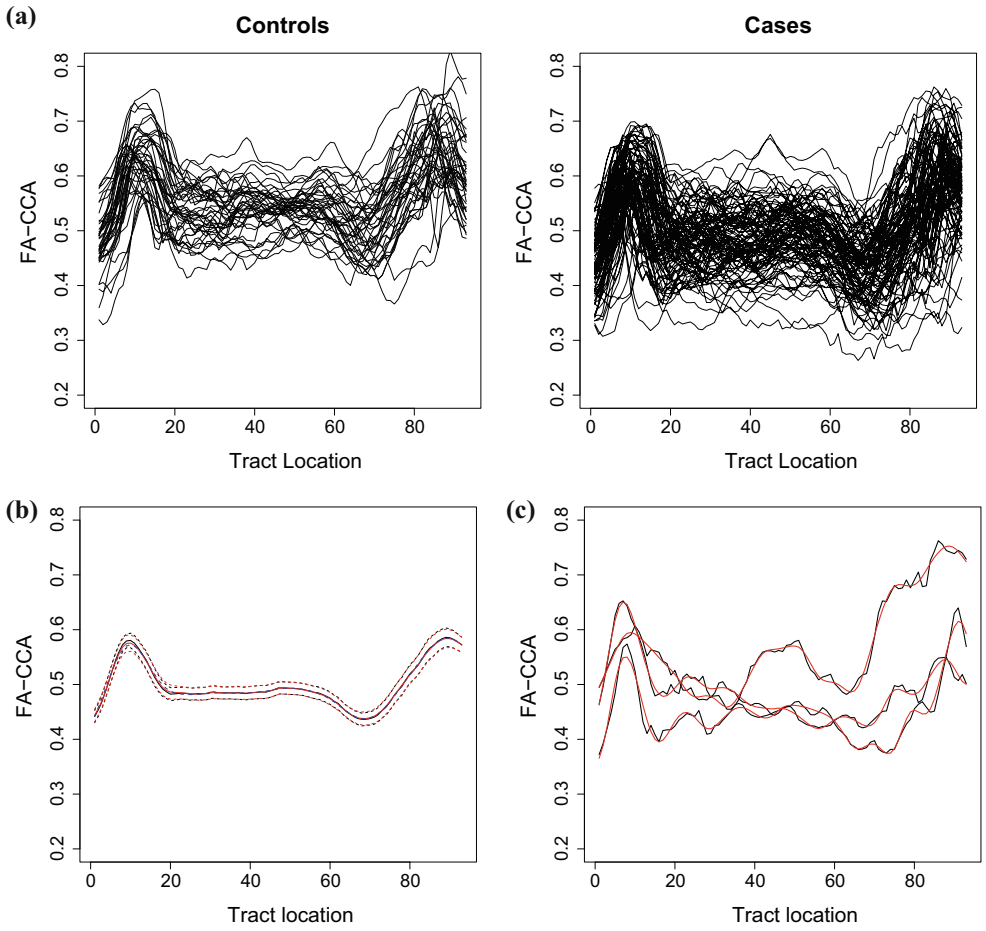

Fig. 4 For the diffusion tensor imaging scan data: **a** observed FA values along the CCA tract. Left: healthy controls. Right: MS patients; **b** estimated population (group) FA profile for individuals affected by MS. Solid red line: SOP method. Dotted blue line: Djeundje and Currie (2010)'s code. The dotted red lines are the pointwise 95% confidence intervals based on the full-sandwich standard errors proposed by Heckman et al. (2013). The back line is the observed mean, and the dotted black lines are the empirical confidence intervals; and **c** estimated (red lines) and observed (black lines) individual FA profiles for 3 selected individuals. (Color figure online)

using both approaches, which provide very similar results. 95% pointwise confidence intervals are calculated using the full-sandwich standard errors proposed by Heckman et al. (2013) but adapted to our case. Figure 4c shows, for several MS patients, the estimated (and observed) FA profiles. In terms of EDs, we obtain 35.03 (out of 43) for the population curve (including the unpenalised or fixed part), and a total of 2025.78 (out of 2275 ($= 23 \times 99 - 2$)) for all individual curves. We note that this total corresponds to the sum of the EDs associated with the variances σ_2^2 and σ_3^2 involved in the modelling of these individual curves [see expressions (23) and (24) for details]. More precisely, using the SOP method we obtain an ED of 870.44 for σ_2^2 and of 1155.34 for σ_3^2.

Our second analysis considers all individuals, cases and controls. The interest here is to compare the FA profiles at the first visit between these two groups. To that aim, the following factor-by-curve interaction model is considered.

$$y_{ij} = f_{z_j}(t_i) + g_j(t_i) + \varepsilon_{ij} \quad 1 \le i \le s,\ 1 \le j \le m,$$

where $z_j = 1$ if the j-th individual is affected by MS (case) and $z_j = 0$ otherwise (control), and $\varepsilon_{ij} \sim N(0, \phi)$. Note that this model assumes a different FA profile for each group. For this analysis, there are a total of $m_0 = 42$ controls and $m_1 = 99$ MS patients ($m = m_0 + m_1 = 141$), and $s = 93$ different tract locations. A detailed description of the model can be found in "Appendix D". As for the first analysis 43 cubic B-spline basis are used for the population curves (FA profiles), and 23 for the individual curves, yielding a total of 3329 ($= 43 \times 2 + 23 \times 141$) coefficients. Using GLAM and sparse matrix techniques (`R`-package `Matrix`), the fit takes 65 seconds. Figure 5a shows the estimated FA profiles for both cases and controls, jointly with 95% pointwise confidence intervals (Heckman et al. 2013). The ED for the FA profile in controls is 32.21 and in MS patients is 35.55. In both cases, we include the fixed part. Regarding the indi-

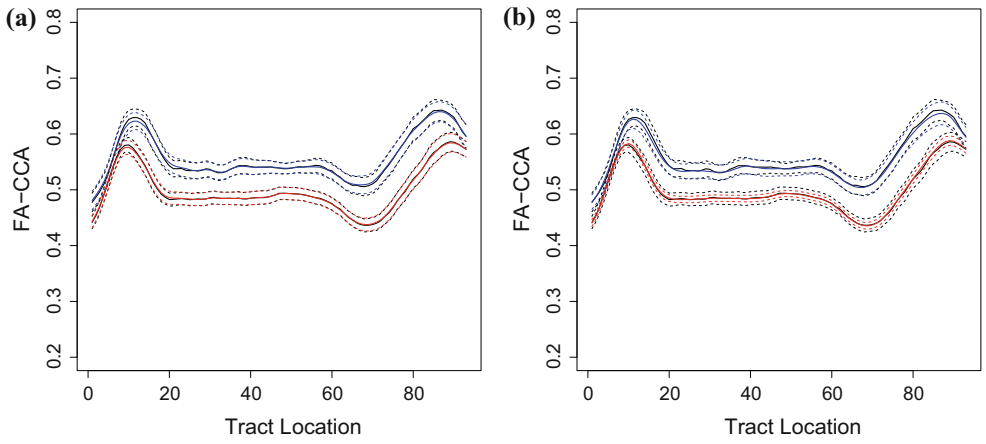

Fig. 5 For the diffusion tensor imaging scan data: estimated population (group) FA profiles. From left to right: results using the SOP method and the functional regression approach by Greven and Scheipl (2017). Solid red line: MS patients. Solid blue line: controls. The dotted blue and red lines are the pointwise 95% confidence intervals. For the SOP method confidence intervals are based on the full-sandwich standard errors proposed by Heckman et al. (2013). The solid back line is the observed mean, and the dotted black lines are the empirical confidence intervals. (Color figure online)

vidual curves, we obtain a total ED of 2863.46, the sum of 1263.26 and 1600.20.

We compare the results with the functional regression model presented in the paper by Greven and Scheipl (2017) [see model (1.1) and Fig. 2 in that paper]. We would like to note that the P-spline model used in this paper was discussed in Durban and Aguilera-Morillo (2017) as a competitive alternative to the functional approach by Greven and Scheipl (2017). For the functional approach, we consider Gaussian homoskedastic errors, and, as suggested by the authors, 25 cubic B-spline basis and a first-order difference penalty, as well as 8 functional principal components (FPC) functions. The code for fitting the model was kindly provided by the authors. Results are depicted in Fig. 5b. As can be observed, both approaches provide very similar results. However, note that the pointwise 95% confidence interval for the estimated FA profile in MS patients is narrower than the empirical confidence interval. This results can be explained by the (possibly wrong) assumption of Gaussian homoskedastic errors. In terms of computing times, the functional regression model needs 895 s to be fitted (in contrast to 65 s using SOP). We are aware that the computing times of both approaches (the SOP method and functional approach) are not fully comparable since they assume different model specifications.

6 Discussion

This paper presents a new estimation method, called SOP, for (generalised) linear mixed models. The method is an extension of a previous proposal by Harville (1977), and generalised by Schall (1991) among others. In contrast to those previous approaches, the SOP method is suitable for models where the precision matrix is linear in the precision parameters. These precision matrices are common when penalised smooth models are reformulated as (generalised) linear mixed models. They appear when there are multiple quadratic penalties acting on the same regression coefficients. One special case is anisotropic tensor-product P-splines. This situation was discussed in the paper by Rodríguez-Álvarez et al. (2015b) where the SAP algorithm was proposed. The present paper goes one step further and generalises SAP to a more general case. SOP is, as far as we know, the only method in this context where the variance parameters estimates involve "partial" effective degrees of freedom. As a consequence, the method provides, at convergence, the estimated effective partial degrees of freedom associated with each smooth/random component in the model. This is especially relevant when working with smoothing models, where the effective degrees of freedom of each smooth term give important insights on the complexity of the fitted function. Furthermore, we show in the paper that the SOP method ensures non-negative estimates of the variance parameters and discuss the conditions under which these estimates are strictly positive.

We present in the paper several situations in the context of penalised spline regression in which SOP represents a powerful alternative to existing estimation methods. In particular, we show the use of SOP in the case of spatially adaptive P-splines and for the estimation of subject-specific curves in longitudinal data. We discuss several real data analyses dealing with these situations and show the outperformance of SOP in terms of computing times. We use simple modelling situations with the aim of describing the method, models and examples in a detailed way, avoiding generalisations

that could complicate the reading of the paper and obscure the simplicity of the proposal. However, there are several other fields of application of SOP. For instance, overlapping penalties appear in brain imaging research (Karas et al. 2017) and derivative curve estimation (Simpkin and Newell 2013). Also, the method can be used for more complex models including linear effects (multidimensional) smooth functions, random Gaussian effects, etc. The use of other basis functions beyond B-splines is also possible as long as quadratic penalties are combined.

The proposal and examples presented in this paper pave the way for further research efforts. For instance, the approach discussed for adaptive P-splines is based on smoothing the locally varying smoothing parameter by means of B-splines. This implies that smoothness is solely controlled by the number of B-spline basis. The selection of the appropriate basis dimension may not be an easy task, with the undesirable consequence that if a large basis dimension is chosen (larger than needed), we may end up with a local linear fit. To reduce the impact of the basis dimension, we will explore the inclusion of a penalty on the coefficients (variance parameters) associated with the B-spline basis. This can be accomplished by means of a hierarchical structure for the random effects (see, e.g., Krivobokova et al. 2008). Another challenging field is the study of suitable penalties and efficient estimation methods for adaptive P-splines in more than one dimension. Whereas some attempts have been done in two dimensions (see, e.g., Crainiceanu et al. 2007; Krivobokova et al. 2008), to the best of our knowledge the literature is lacking in three dimensional approaches (e.g., space and time). Some preliminary results using SOP are available at Rodríguez-Álvarez et al. (2016), but further work still needs to be done. Also, the use and extension of the SOP method to the analysis of multilevel and longitudinal functional data (see, e.g., Greven and Scheipl 2017) represents an exciting field of research. Finally, for variable selection there exists some works that propose sparse regression models using local quadratic approximations to the L1-norm adopting a penalised likelihood approach (see Fan and Li 2001; Hunter and Li 2005; Zou and Li 2008). More recently, (generalised) linear mixed effects approaches have also been proposed (see Taylor et al. 2012; Groll and Tutz 2014) allowing for the penalty to be estimated simultaneously with the variance parameters using REML. We intend to extend the SOP method in this direction. In conclusion, this paper opens up a pathway for a general estimating method allowing for both smoothing and variable selection in reasonable computing times.

The `R`-code used for the real data examples presented in Sect. 5 as well as an `R`-package implementing the SOP method for generic generalised linear mixed models, spatially adaptive P-spline models and P-spline models for hierarchical curve data can be downloaded from https://bitbucket.org/mxrodriguez/sop.

Acknowledgements This research was supported by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through projects MTM2017-82379-R funded by (AEI/FEDER, UE) and acronym "AFTERAM", MTM2014-52184-P and MTM2014-55966-P. The MRI/DTI data were collected at Johns Hopkins University and the Kennedy-Krieger Institute. We are grateful to Pedro Caro and Iain Currie for useful discussions, to Martin Boer and Cajo ter Braak for the detailed reading of the paper and their many suggestions, and to Bas Engel for sharing with us his knowledge. We are also grateful to the two peer referees for their constructive comments of the paper.

A Proof of Theorem 1

Proof We first note that the first-order partial derivatives of the (approximate) REML log-likelihood function can be expressed as (see, e.g., Rodríguez-Álvarez et al. 2015b)

$$\frac{\partial l}{\partial \sigma^2_{k_l}} = -\frac{1}{2}\text{trace}\left(\boldsymbol{Z}^\top \boldsymbol{P}\boldsymbol{Z}\frac{\partial \boldsymbol{G}}{\partial \sigma^2_{k_l}}\right) + \frac{1}{2}\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{G}^{-1}\frac{\partial \boldsymbol{G}}{\partial \sigma^2_{k_l}}\boldsymbol{G}^{-1}\widehat{\boldsymbol{\alpha}}.$$

Given that $\boldsymbol{G}$ is a positive definite matrix, we have the identity

$$\frac{\partial \boldsymbol{G}}{\partial \sigma^2_{k_l}} = -\boldsymbol{G}\frac{\partial \boldsymbol{G}^{-1}}{\partial \sigma^2_{k_l}}\boldsymbol{G},$$

and thus

$$\frac{\partial \boldsymbol{G}}{\partial \sigma^2_{k_l}} = \frac{1}{\sigma^4_{k_l}}\text{diag}\left(\mathbf{0}^{(1)}, \boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\boldsymbol{G}_k, \mathbf{0}^{(2)}\right),$$

where $\mathbf{0}^{(1)}$ and $\mathbf{0}^{(2)}$ are zero square matrices of appropriate dimensions.

The first-order partial derivatives of the REML log-likelihood function are then expressed as

$$2\frac{\partial l}{\partial \sigma^2_{k_l}} = -\frac{1}{\sigma^4_{k_l}}\text{trace}\left(\boldsymbol{Z}_k^\top \boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\boldsymbol{G}_k\right) + \frac{1}{\sigma^4_{k_l}}\widehat{\boldsymbol{\alpha}}_k^\top \boldsymbol{\Lambda}_{k_l}\widehat{\boldsymbol{\alpha}}_k.$$

When the REML estimates are positive, they are obtained by equating the former expression to zero (see, e.g., Engel 1990)

$$\frac{\widehat{\boldsymbol{\alpha}}_k^\top \boldsymbol{\Lambda}_{k_l}\widehat{\boldsymbol{\alpha}}_k}{\text{trace}\left(\boldsymbol{Z}_k^\top \boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\boldsymbol{G}_k\right)} = 1.$$

We now multiply both sides with $\sigma^2_{k_l}$ and evaluate the left-hand side for the previous iterates and the right-hand side for the update, obtaining

$$\widehat{\sigma}^2_{k_l} = \frac{\widehat{\boldsymbol{\alpha}}_k^{[t]\top}\boldsymbol{\Lambda}_{k_l}\widehat{\boldsymbol{\alpha}}_k^{[t]}}{\text{trace}\left(\boldsymbol{Z}_k^\top \boldsymbol{P}^{[t]}\boldsymbol{Z}_k\boldsymbol{G}_k^{[t]}\boldsymbol{\Lambda}_{k_l}\boldsymbol{G}_k^{[t]}\right)}\widehat{\sigma}^{2[t]}_{k_l}$$
$$= \frac{\widehat{\boldsymbol{\alpha}}_k^{[t]\top}\boldsymbol{\Lambda}_{k_l}\widehat{\boldsymbol{\alpha}}_k^{[t]}}{\text{trace}\left(\boldsymbol{Z}_k^\top \boldsymbol{P}^{[t]}\boldsymbol{Z}_k\boldsymbol{G}_k^{[t]}\frac{\boldsymbol{\Lambda}_{k_l}}{\widehat{\sigma}^{2[t]}_{k_l}}\boldsymbol{G}_k^{[t]}\right)}.$$

□

B Proof of Theorem 2

Proof First let us recall some notation and introduce some needed results. We denote as $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}\right)\boldsymbol{X}^\top\boldsymbol{V}^{-1}$, where $\boldsymbol{V} = \boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^\top$, $\boldsymbol{R} = \phi\boldsymbol{W}^{-1}$ with $\boldsymbol{W}$ being the diagonal weight matrix involved in the Fisher scoring algorithm.

Denote as $\mathcal{C}(\boldsymbol{A})$ the linear space spanned by the columns of $\boldsymbol{A}$, and let $\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{V}^{-1}} = \boldsymbol{X}\left(\boldsymbol{X}^\top\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^\top\boldsymbol{V}^{-1}$ be the orthogonal projection matrix for $\mathcal{C}(\boldsymbol{X})$ with respect to $\boldsymbol{V}^{-1}$. It is easy to show that

$$\boldsymbol{P} = \boldsymbol{V}^{-1}\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{V}^{-1}}\right)$$
$$= \left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{V}^{-1}}\right)\boldsymbol{V}^{-1}\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{V}^{-1}}\right).$$

By Theorem 14.2.9 in Harville (1997), $\boldsymbol{P}$ is a (symmetric) positive semi-definite matrix. In addition,

$$\boldsymbol{P}\boldsymbol{X} = \boldsymbol{0},$$

and

$$\text{rank}(\boldsymbol{P}) = \text{rank}\left(\boldsymbol{V}^{-1}\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{V}^{-1}}\right)\right)$$
$$= \text{rank}\left(\left(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}\boldsymbol{V}^{-1}}\right)\right)$$
$$= n - \text{rank}\left(\boldsymbol{P}_{\boldsymbol{X}\boldsymbol{V}^{-1}}\right)$$
$$= n - \text{rank}\left(\boldsymbol{X}\right).$$

Thus,

$$ker\left(\boldsymbol{P}\right) = \mathcal{C}\left(\boldsymbol{X}\right), \tag{26}$$

i.e., $\boldsymbol{P}\boldsymbol{x} = \boldsymbol{0}$ if and only if $\boldsymbol{x}$ is in $\mathcal{C}(\boldsymbol{X})$.

Let $\boldsymbol{\Lambda}_{k_l} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$ be the eigenvalue decomposition of $\boldsymbol{\Lambda}_{k_l}$. Note that $\boldsymbol{\Lambda}_{k_l} = \boldsymbol{U}_+\boldsymbol{\Sigma}_+\boldsymbol{U}_+^\top$, where $\boldsymbol{U}_+$ and $\boldsymbol{\Sigma}_+$ are the sub-matrices corresponding to the nonzero eigenvalues. Then

$$\widehat{\boldsymbol{\alpha}}_k^\top\boldsymbol{\Lambda}_{k_l}\widehat{\boldsymbol{\alpha}}_k = \widehat{\boldsymbol{\alpha}}_k^\top\boldsymbol{U}_+\boldsymbol{\Sigma}_+\boldsymbol{U}_+^\top\widehat{\boldsymbol{\alpha}}_k$$
$$= \boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{U}_+\boldsymbol{\Sigma}_+\boldsymbol{U}_+^\top\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{z} \geqslant 0,$$

with equality holding if and only if $\boldsymbol{U}_+^\top\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{z} = \boldsymbol{0}$ (since $\boldsymbol{\Sigma}_+$ is positive definite). Thus, using result (26), the equality holds if $\boldsymbol{z}$ is in $\mathcal{C}(\boldsymbol{X})$ or $\mathcal{C}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{U}_+\right) \subset \mathcal{C}(\boldsymbol{X})$. By Lemma 4.2.2 and Corollary 4.5.2 in Harville (1997), we have

$$\mathcal{C}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{U}_+\right) = \mathcal{C}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\right) \subset \mathcal{C}\left(\boldsymbol{X}\right)$$
$$\iff \text{rank}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}, \boldsymbol{X}\right) = \text{rank}\left(\boldsymbol{X}\right).$$

Regarding the denominator of the REML-based estimates updates, we follow a similar reasoning. Using Corollary 14.7.5 (and Theorem 14.2.9) in Harville (1997), we have

$$\text{trace}\left(\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}\boldsymbol{G}_k\right)$$
$$= \text{trace}\left(\boldsymbol{U}_+^\top\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{U}_+\boldsymbol{\Sigma}_+\right) \geqslant 0,$$

with equality holding if and only if $\boldsymbol{U}_+^\top\boldsymbol{G}_k\boldsymbol{Z}_k^\top\boldsymbol{P}\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{U}_+ = \boldsymbol{0}$. Again, this equality holds if and only if $\mathcal{C}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{U}_+\right) \subset \mathcal{C}(\boldsymbol{X})$ (i.e., $\iff$ $\text{rank}\left(\boldsymbol{Z}_k\boldsymbol{G}_k\boldsymbol{\Lambda}_{k_l}, \boldsymbol{X}\right) = \text{rank}\left(\boldsymbol{X}\right)$). □

C Estimating algorithm

This appendix summarises the steps of the estimating algorithm for model (10) based on the SOP method. Recall that interest lies in estimating model

$$g\left(\boldsymbol{\mu}\right) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{k=1}^{c}\boldsymbol{Z}_k\boldsymbol{\alpha}_k,$$

where $\boldsymbol{Z} = [\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_c]$, $\boldsymbol{\alpha} = \left(\boldsymbol{\alpha}_1^\top, \ldots, \boldsymbol{\alpha}_c^\top\right)^\top$, $\boldsymbol{\alpha}_k \sim N\left(\boldsymbol{0}, \boldsymbol{G}_k\right)$ with $\boldsymbol{G}_k^{-1} = \sum_{l=1}^{p_k}\sigma_{k_l}^{-2}\boldsymbol{\Lambda}_{k_l}$, and $\boldsymbol{\alpha} \sim N\left(\boldsymbol{0}, \boldsymbol{G}\right)$ with $\boldsymbol{G} = \bigoplus_{k=1}^{c}\boldsymbol{G}_k$.

Initialise Set initial values for $\widehat{\boldsymbol{\mu}}^{[0]}$ and the variance parameters $\widehat{\sigma}^{2[0]}_{k_l}$ ($l = 1, \ldots, p_k$ and $k = 1, \ldots, c$). In those situations where ϕ is unknown, establish an initial value for this parameter, $\widehat{\phi}^{[0]}$. Set $t = 0$.

Step 1 Construct the *working* response vector $\boldsymbol{z}$ and the matrix of weights $\boldsymbol{W}$ as follows

$$z_i = g(\widehat{\mu}_i^{[t]}) + (y_i - \widehat{\mu}_i^{[t]})g'(\widehat{\mu}_i^{[t]}),$$
$$w_{ii} = \left\{g'(\widehat{\mu}_i^{[t]})^2 v(\widehat{\mu}_i^{[t]})\right\}^{-1}.$$

Step 1.1 Given the initial *estimates* of variance parameters, estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by solving

$$\underbrace{\begin{bmatrix}\boldsymbol{X}^\top\boldsymbol{R}^{[t]^{-1}}\boldsymbol{X} & \boldsymbol{X}^\top\boldsymbol{R}^{[t]^{-1}}\boldsymbol{Z} \\ \boldsymbol{Z}^\top\boldsymbol{R}^{[t]^{-1}}\boldsymbol{X} & \boldsymbol{Z}^\top\boldsymbol{R}^{[t]^{-1}}\boldsymbol{Z} + \boldsymbol{G}^{[t]^{-1}}\end{bmatrix}}_{\boldsymbol{C}^{[t]}}\begin{bmatrix}\widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}^\top\boldsymbol{R}^{[t]^{-1}}\boldsymbol{z} \\ \boldsymbol{Z}^\top\boldsymbol{R}^{[t]^{-1}}\boldsymbol{z}\end{bmatrix}, \tag{27}$$

where $\boldsymbol{R}^{[t]} = \widehat{\phi}^{[t]}\boldsymbol{W}^{-1}$. Let $\widehat{\boldsymbol{\alpha}}^{[t]}$ and $\widehat{\boldsymbol{\beta}}^{[t]}$ be these estimates.

Step 1.2 Update the variance parameters as follows

$$\widehat{\sigma}^2_{k_l} = \frac{\widehat{\boldsymbol{\alpha}}_k^{[t]\top}\boldsymbol{\Lambda}_{k_l}\widehat{\boldsymbol{\alpha}}_k^{[t]}}{\mathrm{ED}_{k_l}^{[t]}},$$

and, when necessary,

$$\widehat{\phi} = \frac{\left(z - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{[t]} - \boldsymbol{Z}\widehat{\boldsymbol{\alpha}}^{[t]}\right)^\top \boldsymbol{W}\left(z - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{[t]} - \boldsymbol{Z}\widehat{\boldsymbol{\alpha}}^{[t]}\right)}{n - \mathrm{rank}(\boldsymbol{X}) - \sum_{k=1}^{c}\sum_{l=1}^{p_k}\mathrm{ED}_{k_l}^{[t]}},$$

with

$$\mathrm{ED}_{k_l}^{[t]} = \mathrm{trace}\left(\left(\boldsymbol{G}_k^{[t]} - \boldsymbol{C}^{[t]*}_{kk}\right)\frac{\boldsymbol{\Lambda}_{k_l}}{\widehat{\sigma}^{2[t]}_{k_l}}\right).$$

Recall that $\boldsymbol{C}^{[t]*}$ denotes the inverse of $\boldsymbol{C}^{[t]}$ in (27), and $\boldsymbol{C}^{[t]*}_{kk}$ is that partition of $\boldsymbol{C}^{[t]*}$ corresponding to the k-th random component $\boldsymbol{\alpha}_k$.

Step 1.3 Repeat Step 1.1 and Step 1.2, replacing $\widehat{\sigma}^{2[t]}_{k_l}$, and, if updated, $\widehat{\phi}^{[t]}$, by $\widehat{\sigma}^2_{k_l}$ and $\widehat{\phi}$, until convergence. For the examples presented in Sect. 5, we use the REML deviance as the convergence criterion.

Step 2 Repeat Step 1, replacing the variance parameters and the model's fixed and random effects (and thus $\widehat{\boldsymbol{\mu}}^{[t]}$) by those obtained in the last iteration of Steps 1.1–Step 1.3, until convergence.

It is worth noting that for notational convenience, in the examples described in Sect. 4, the precision matrix was rewritten as $\boldsymbol{G}^{-1} = \sum_{l=1}^{p}\sigma_l^{-2}\widetilde{\boldsymbol{\Lambda}}_l$, where $p = \sum_{k=1}^{c}p_k$ is the number of variance parameters, and $\widetilde{\boldsymbol{\Lambda}}_l$ are the matrices $\boldsymbol{\Lambda}_{k_l}$ padded out with zeroes. Here, $\widetilde{\boldsymbol{\Lambda}}_l$ are matrices of dimension $q \times q$, where $q = \sum_{k=1}^{c}q_k$ is the number of random effects coefficients. For this specification, the estimating algorithm discussed above remains essentially the same, but in Step 1.2, $\boldsymbol{\alpha}_k^{[t]}$, $\boldsymbol{G}_k^{[t]}$, and $\boldsymbol{\Lambda}_{k_l}$ are replaced by, respectively, $\boldsymbol{\alpha}^{[t]}$, $\boldsymbol{G}^{[t]}$ and $\widetilde{\boldsymbol{\Lambda}}_l$, and $\boldsymbol{C}^{[t]*}_{kk}$ would be that partition of $\boldsymbol{C}^{[t]*}$ corresponding to the random vector $\boldsymbol{\alpha}$ (and thus the same for all variance parameters).

D Factor-by-curve hierarchical curve model

This appendix describes in detail the factor-by-curve interaction model discussed in Sect. 5.3, i.e.,

$$y_{ij} = f_{z_j}(t_i) + g_j(t_i) + \varepsilon_{ij} \quad 1 \le i \le s,\ 1 \le j \le m,$$

where $z_j = 1$ if the j-th individual is affected by MS (case) and $z_j = 0$ otherwise (control). Let's order the data with the observations on controls first, followed by observations on MS patients. In matrix notation, the model can be expressed as

$$\boldsymbol{y} = [\boldsymbol{Q} \otimes \boldsymbol{B}]\boldsymbol{\theta} + [\boldsymbol{I}_m \otimes \breve{\boldsymbol{B}}]\breve{\boldsymbol{\theta}} + \boldsymbol{\varepsilon}, \tag{28}$$

with $\boldsymbol{B}, \breve{\boldsymbol{B}}, \breve{\boldsymbol{\theta}}$ as defined in Sect. 4.2, and $\boldsymbol{\varepsilon} = \left(\boldsymbol{\varepsilon}_{\cdot,1}^\top, \ldots, \boldsymbol{\varepsilon}_{\cdot,m}^\top\right)^\top$, with $\boldsymbol{\varepsilon}_{\cdot j} = \left(\varepsilon_{1j}, \ldots, \varepsilon_{sj}\right)^\top$. Matrix $\boldsymbol{Q}$ is any suitable contrast matrix of dimension $m \times 2$, where $m = m_0 + m_1$, with m_0 being the number of controls and m_1 the number of MS patients. For our application, we consider

$$\boldsymbol{Q} = \begin{pmatrix} \mathbf{1}_{m_0} & \mathbf{0}_{m_0} \\ \mathbf{0}_{m_1} & \mathbf{1}_{m_1} \end{pmatrix},$$

and a different amount of smoothing is assumed for f_0 and f_1, i.e., the penalty matrix acting over the vector of coefficients $\boldsymbol{\theta}$ is of the form

$$\boldsymbol{P} = \begin{pmatrix} \lambda_1\boldsymbol{D}_q^\top\boldsymbol{D}_q & \mathbf{0}_{c\times c} \\ \mathbf{0}_{c\times c} & \lambda_2\boldsymbol{D}_q^\top\boldsymbol{D}_q \end{pmatrix}.$$

The reformulation as a mixed model can be done in a similar fashion to that described in Sect. 4.2, with, in this case

$$\boldsymbol{X} = [\boldsymbol{Q} \otimes \boldsymbol{B}\boldsymbol{U}_0],$$
$$\boldsymbol{Z} = [\boldsymbol{Q} \otimes \boldsymbol{B}\boldsymbol{U}_+ : \boldsymbol{I}_m \otimes \breve{\boldsymbol{B}}],$$

and

$$\boldsymbol{G}^{-1} = \begin{pmatrix} \sigma_1^{-2}\boldsymbol{\Sigma}_+ & \mathbf{0}_{d\times d} & \mathbf{0}_{d\times\left(m\breve{d}\right)} \\ \mathbf{0}_{d\times d} & \sigma_2^{-2}\boldsymbol{\Sigma}_+ & \mathbf{0}_{d\times\left(m\breve{d}\right)} \\ \mathbf{0}_{\left(m\breve{d}\right)\times d} & \mathbf{0}_{\left(m\breve{d}\right)\times c} & \boldsymbol{I}_m \otimes \breve{\boldsymbol{G}}^{-1} \end{pmatrix},$$

where

$$\breve{\boldsymbol{G}}^{-1} = \sigma_3^{-2}\breve{\boldsymbol{D}}_{\breve{q}}^\top\breve{\boldsymbol{D}}_{\breve{q}} + \sigma_4^{-2}\boldsymbol{I}_{\breve{d}}.$$

References

Breslow, N.E., Clayton, D.G.: Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. **88**(421), 9–25 (1993)

Camarda, C.G., Eilers, P.H., Gampe, J.: Sums of smooth exponentials to decompose complex series of counts. Stat. Model. **16**(4), 279–296 (2016)

Crainiceanu, C.M., Ruppert, D., Carroll, R.J., Joshi, A., Goodner, B.: Spatially adaptive Bayesian penalized splines with heteroscedastic errors. J. Comput. Graph. Stat. **16**(2), 265–288 (2007)

Crump, S.L.: The present status of variance component analysis. Biometrics **7**(1), 1–16 (1951)

Cui, Y., Hodges, J.S., Kong, X., Carlin, B.P.: Partitioning degrees of freedom in hierarchical and other richly-parameterized models. Technometrics **52**, 124–136 (2010)

Currie, I.D., Durban, M.: Flexible smoothing with P-splines: a unified approach. Stat. Model. **2**(4), 333–349 (2002)

Currie, I.D., Durban, M., Eilers, P.H.C.: Generalized linear array models with applications to multidimensional smoothing. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **68**(2), 259–280 (2006)

Davies, P.L., Gather, U., Meise, M., Mergel, D., Mildenberger, T.: Residual-based localization and quantification of peaks in X-ray diffractograms. Ann. Appl. Stat. **2**(3), 861–886 (2008)

Davies, P.L., Gather, U., Meise, M. Mergel, D., Mildenberger, T., Bernholt, T., Hofmeister, T.: diffractometry: baseline identification and peak decomposition for x-ray diffractograms. R package version 0.1-10 (2018)

Djeundje, V.A., Currie, I.D.: Appropriate covariance-specification via penalties for penalized splines in mixed models for longitudinal data. Electron. J. Stat. **4**, 1202–1224 (2010)

Durban, M., Aguilera-Morillo, M.C.: On the estimation of functional random effects. Stat. Model. **17**(1–2), 50–58 (2017)

Durban, M., Harezlak, J., Wand, M.P., Carroll, R.J.: Simple fitting of subject-specific curves for longitudinal data. Stat. Med. **24**(8), 1153–1167 (2005)

Eilers, P.H.C.: Discussion of Verbyla et al. J. R. Stat. Soc. Ser. C (Appl. Stat.) **48**, 300–311 (1999)

Eilers, P.H.C., Marx, B.D.: Flexible smoothing with B-splines and penalties. Stat. Sci. **11**(2), 89–121 (1996)

Engel, B.: The analysis of unbalanced linear models with variance components. Stat. Neerl. **44**, 195–219 (1990)

Engel, B., Buist, W.: Analysis of a generalized linear mixed model: a case study and simulation results. Biom. J. **38**(1), 61–80 (1996)

Engel, B., Keen, A.: A simple approach for the analysis of generalizea linear mixed models. Stat. Neerl. **48**(1), 1–22 (1994)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)

Gilmour, A.R., Thompson, R., Cullis, B.R.: Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics **51**(4), 1440–1450 (1995)

Goldsmith, J., Bobb, J., Crainiceanu, C.M., Caffo, B., Reich, D.: Penalized functional regression. J. Comput. Graph. Stat. **20**(4), 830–851 (2011)

Goldsmith, J., Crainiceanu, C.M., Caffo, B., Reich, D.: Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. J. R. Stat. Soc. Ser. C (Appl. Stat.) **61**(3), 453–469 (2012)

Goldsmith, J., Scheipl, F., Huang, L., Wrobel, J., Gellar, J., Harezlak, J., McLean, M.W., Swihart, B., Xiao, L., Crainiceanu, C., Reiss, P.T.: refund: Regression with Functional Data. R package version 0.1-16 (2016)

Graser, H.-U., Smith, S.P., Tier, B.: A derivative-free approach for estimating variance components in animal models by restricted maximum likelihood. J. Anim. Sci. **2**(64), 1362–1373 (1987)

Green, P.J.: Penalized likelihood for general semi-parametric regression models. Int. Stat. Rev./Revue Internationale de Statistique **55**(3), 245–259 (1987)

Greven, S., Scheipl, F.: A general framework for functional regression modelling. Stat. Model. **17**(1–2), 1–35 (2017)

Groll, A., Tutz, G.: Variable selection for generalized linear mixed models by L1-penalized estimation. Stat. Comput. **24**(2), 137–154 (2014)

Harville, D.A.: Maximum likelihood approaches to variance component estimation and to related problems. J. Am. Stat. Assoc. **72**(358), 320–338 (1977)

Harville, D.A.: Matrix Algebra from a Statistician's Perspective. Springer, Berlin (1997)

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

Heckman, N., Lockhart, R., Nielsen, J.D.: Penalized regression, mixed effects models and appropriate modelling. Electron. J. Stat. **7**, 1517–1552 (2013)

Henderson, C.R.: Selection index and expected genetic advance. Stat. Genet. Plant Breed. **982**, 141–163 (1963)

Hunter, D.R., Li, R.: Variable selection using MM algorithms. Ann. Stat. **33**(4), 1617–1642 (2005)

Johnson, D.L., Thompson, R.: Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. J. Dairy Sci. **78**, 449–456 (1995)

Karas, M., Brzyski, D., Dzemidzic, M., Goñi, J., Kareken, D.A., Randolph, T.W., Harezlak, J.: Brain connectivity-informed regularization methods for regression. Stat. Biosci. (2017). https://doi.org/10.1007/s12561-017-9208-x

Krivobokova, T.: Smoothing parameter selection in two frameworks for penalized splines. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **75**(4), 725–741. https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12010 (2009)

Krivobokova, T., Crainiceanu, C.M., Kauermann, G.: Fast adaptive penalized splines. J. Comput. Graph. Stat. **17**(1), 1–20 (2008)

Lee, D.-J.: Smoothing mixed model for spatial and spatio-temporal data. PhD thesis. Department of Statistics, Universidad Carlos III de Madrid, Spain (2010)

McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series, 2nd edn. Chapman & Hall, London (1989)

Patterson, H.D., Thompson, R.: Recovery of inter-block information when block sizes are unequal. Biometrika **58**(3), 545–554 (1971)

R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018)

Reiss, P .T., Ogden, R .T.: Smoothing parameter selection for a class of semiparametric linear models. J. R. Stat. Soc. Ser. B Stat. Methodol. **71**(2), 505–523 (2009)

Rodríguez-Álvarez, M.X., Durban, M., Lee, D.-J., Eilers, P.H.C.: Fast estimation of multidimensional adaptive p-spline models. In: Friedl, H., Wagner, H. (eds.) Proceedings of the 30th International Workshop on Statistical Modelling, pp 330 – 335. arXiv:1610.06861 (2015a)

Rodríguez-Álvarez, M.X., Lee, D.-J., Kneib, T., Durban, M., Eilers, P.H.C.: Fast smoothing parameter separation in multidimensional generalized P-splines: the sap algorithm. Stat. Comput. **25**, 941–957 (2015b)

Rodríguez-Álvarez, M.X., Durban, M., Lee, D.-J., Eilers, P.H.C., Gonzalez, F.: Spatio-temporal adaptive penalized splines with application to neuroscience. In: Dupuy, J.-F., Josse, J. (eds.) Proceedings of the 31th International Workshop on Statistical Modelling, pp. 267–272. arXiv:1610.06860 (2016)

Rodríguez-Álvarez, M.X., Boer, M.P., van Eeuwijk, F.A., Eilers, P.H.: Correcting for spatial heterogeneity in plant breeding experiments with p-splines. Spat. Stat. **23**, 52–71 (2018)

Ruppert, D., Carroll, R.J.: Spatially-adaptive penalties for spline fitting. Aust. N. Z. J. Stat. **42**(2), 205–223 (2000)

Ruppert, D., Wand, M.P., Carroll, R.: Semiparametric Regression. Cambridge University Press, Cambridge (2003)

Schall, R.: Estimation in generalized linear models with random effects. Biometrika **78**(4), 719–727 (1991)

Simpkin, A., Newell, J.: An additive penalty p-spline approach to derivative estimation. Comput. Stat. Data Anal. **68**, 30–43 (2013)

Smith S.P.: Estimation of genetic parameters in non-linear models. In: Gianola, D., Hammond, K. (eds.) Advances in Statistical Methods for Genetic Improvement of Livestock. Advanced Series in Agricultural Sciences, vol. 18. Springer, Berlin, Heidelberg (1990)

Taylor, J.D., Verbyla, A.P., Cavanagh, C., Newberry, M.: Variable selection in linear mixed models using an extended class of penalties. Aust. N. Z. J. Stat. **54**(4), 427–449 (2012)

Tibshirani, R.J.: Adaptive piecewise polynomial estimation via trend filtering. Ann. Stat. **42**(1), 285–323 (2014)

Wand, M.P.: Smoothing and mixed models. Comput. Stat. **18**(2), 223–249 (2003)

Wood, S.N.: Fast stable direct fitting and smoothness selection for generalized additive models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **70**(3), 495–518 (2008)

Wood, S .N.: Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **73**(1), 2–36 (2011)

Wood, S.N.: Generalized Additive Models: An Introduction with R, 2nd edn. Chapman & Hall CRC, London (2017)

Wood, S.N., Fasiolo, M.: A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. Biometrics **73**, 1071–1081 (2017)

Wood, S.N., Pya, N., Säfken, B.: Smoothing parameter and model selection for general smooth models. J. Am. Stat. Assoc. **111**(516), 1548–1563 (2016)

Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. Ann. Stat. **36**(4), 1509–1533 (2008)