Computational Statistics and Data Analysis 61 (2013) 22–37

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Efficient two-dimensional smoothing with *P*-spline ANOVA mixed models and nested bases

Dae-Jin Lee [a,*], María Durbán [b], Paul Eilers [c]

[a] *CSIRO Mathematics Informatics and Statistics, Private Bag 33, Clayton, VIC 3169, Australia*

[b] *Department of Statistics, Universidad Carlos III de Madrid, Escuela Politécnica Superior, Leganés 28911 Madrid, Spain*

[c] *Department of Biostatistics, Erasmus Medical Centre, Rotterdam, The Netherlands*

ARTICLE INFO

Article history:
Received 30 March 2012
Received in revised form 16 November 2012
Accepted 16 November 2012
Available online 20 December 2012

Keywords:
Mixed models
Penalized splines
Schall's algorithm
Smooth-ANOVA decomposition

ABSTRACT

Low-rank smoothing techniques have gained much popularity in non-standard regression modeling. In particular, penalized splines and tensor product smooths are used as flexible tools to study non-parametric relationships among several covariates. The use of standard statistical software facilitates their use for several types of problems and applications. However, when interaction terms are considered in the modeling, and multiple smoothing parameters need to be estimated standard software does not work well when datasets are large or higher-order interactions are included or need to be tested. In this paper, a general approach for constructing and estimating bivariate smooth models for additive and interaction terms using penalized splines is proposed. The formulation is based on the mixed model representation of the smooth-ANOVA model by Lee and Durbán (in press), and several nested models in terms of random effects components are proposed. Each component has a clear interpretation in terms of function shape and model identifiability constraints. The term *PS*-ANOVA is coined for this type of models. The estimation method is relatively straightforward based on the algorithm by Schall (1991) for generalized linear mixed models. Further, a simplification of the smooth interaction term is used by constructing lower-rank basis (nested basis). Finally, some simulation studies and real data examples are presented to evaluate the new model and the estimation method.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

One of the main problems encountered when multidimensional smoothing is used, is the computational time needed when the dataset is large and the number of covariates increases. Many attempts (with different degrees of success) have been made over years to deal with this problem. Buja et al. (1989) and Hastie and Tibshirani (1990) introduced the idea of additive models, where functions of several covariates are expressed as the sum of unidimensional smooth terms. These models use the back-fitting algorithm for the estimation of the smooth functions. This is a very fast method (even for large datasets); however, the models are too simplistic in some situations, and they might not reflect important features (interactions) in the data. A more versatile family of smoothing methods was proposed by Wahba (1990) and Chen (1993), where additive models were extended to include interaction terms of covariates, subject to certain constraints. This was an extension of ANOVA models to smooth functional spaces, but once again, the extension to interactions of order 3 or more is computationally very demanding (see Gu, 2002 for a detailed overview in the context of smoothing splines and reproducing kernel Hilbert spaces).

* Corresponding author. Tel.: +61 3 9545 8071; fax: +61 3 9545 8080.
E-mail addresses: dae-jin.lee@csiro.au (D.-J. Lee), mdurban@est-econ.uc3m.es (M. Durbán), p.eilers@eramusmc.nl (P. Eilers).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.11.013

More recently, low-rank smoothers, and in particular, penalized splines (or simply *P*-splines, Eilers and Marx, 1996) have been used to fit smooth terms of two or more dimensions (see for example, Lang and Brezger, 2004, Currie et al., 2006, Wood, 2006b and Lee and Durbán, in press). The reduction of the number of parameters in the basis used for regression (based on tensor product of *B*-splines) has improved the computational burden. However, these models use penalties for the interaction terms that depend on more than one smoothing parameter. Their estimation can be complicated, in particular when there are non-independent smoothing parameters in the penalty term. This makes the penalty term (that can be viewed as a precision matrix) non-standard, so the estimation of the amount of smoothing needs to be done using numerical optimization methods. Hence, the models are far from useful in practice if the size of the data is not moderately small. A possible solution (taken by most users of the usual statistical packages) is to reduce the size of the basis used in the model, but this may result in over-smoothing and, consequently, ignoring important features in the data.

Our proposal is a generalization of the penalized splines ANOVA models introduced by Lee and Durbán (in press). We put together several strategies to obtain fast and efficient models: (i) the multidimensional model is built as a sum of several nested components; (ii) the smoothness of each term is controlled by a single smoothing parameter; (iii) the parameterization of the model as mixed model allows to impose identifiability constraints in an easy and intuitive way, (iv) Schall's algorithm (Schall, 1991) is used for fast estimation of variance components and calculation of the effective dimension of each smooth term in the model, (v) further computational efficiency is obtained by the use of nested *B*-spline basis for interactions.

The paper is organized as follows: in Section 2 we introduce a bivariate *P*-spline ANOVA mixed model and its parameterization as a mixed model, and address the problem of efficient estimation of anisotropic smoothing structures. Section 3 extends the model to consider a new parameterization with separate and independent smoothing parameter for each term. We called this model *PS − ANOVA*. Based on this new formulation we implement the algorithm by Schall (1991). In Section 4 a computationally efficient construction tensor product of *B*-spline bases is proposed. In Section 5 we conduct several simulation studies to check new model performance and computational efficiency, and in Section 6 we analyze some real data. We close with a discussion in Section 7.

2. Low-rank *P*-splines smooth-ANOVA models

To present the background needed for this paper, we introduce the approach in Lee (2010) and Lee and Durbán (in press), and unify the notation for the rest of the Sections.

2.1. Model basis, penalty and identifiability problems

Consider a model two-dimensional smooth model

$$\boldsymbol{y} = \gamma + f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) + f_{1,2}(\boldsymbol{x}_1, \boldsymbol{x}_2) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{O}, \sigma^2 \boldsymbol{I}), \tag{2.1}$$

where γ is the intercept term, with main effects (f_1 and f_2), and two-way interaction effects ($f_{1,2}$), and where the univariate smooth terms are represented by

$$f_d(\boldsymbol{x}_d) = \sum_{j=1}^{c_d} B_j(\boldsymbol{x}_d)\theta_{dj}, \quad \text{for } d = 1, 2, \text{and } j = 1, \ldots, c_d,$$

B_j is a *B*-spline basis function, and θ_{dj} a vector of regression coefficients, of length c_{dj}. The interaction term is

$$f_{1,2}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{k=1}^{c_1}\sum_{l=1}^{c_2} B_k(\boldsymbol{x}_1)B_l(\boldsymbol{x}_2)\theta_{kl}, \quad \text{with } k = 1, \ldots, c_1 \text{ and } l = 1, \ldots, c_2, \tag{2.2}$$

where $B_k(\boldsymbol{x}_1)B_l(\boldsymbol{x}_2)$ is the tensor product of two marginal *B*-splines bases, and θ_{kl} is a vector of coefficients of length $c_1c_2 \times 1$.

In matrix notation, model (2.1) can be written as

$$\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}_1, \boldsymbol{x}_2] = \boldsymbol{B}\boldsymbol{\theta}, \tag{2.3}$$

where $\boldsymbol{B}$ is the full regression matrix, and $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_{[1,2]})'$ is a vector of regression coefficients. The regression matrix for model is defined by blocks as:

$$\boldsymbol{B} = [\boldsymbol{1}_n|\boldsymbol{B}_1|\boldsymbol{B}_2|\boldsymbol{B}_{[1,2]}], \tag{2.4}$$

with marginal bases of the covariates $\boldsymbol{B}_1 = \boldsymbol{B}_1(\boldsymbol{x}_1)$ and $\boldsymbol{B}_2 = \boldsymbol{B}_2(\boldsymbol{x}_2)$ of dimensions $n \times c_1$ and $n \times c_2$ respectively. For the interaction basis $\boldsymbol{B}_{[1,2]}$, we use the row-tensor product or *box-product* of the two marginal *B*-spline bases (see Eilers et al., 2006), denoted by symbol □, and defined as:

$$\boldsymbol{B}_{[1,2]} = \boldsymbol{B}_1 \square \boldsymbol{B}_2 = (\boldsymbol{B}_1 \otimes \boldsymbol{1}_n') * (\boldsymbol{1}_n' \otimes \boldsymbol{B}_2), \quad \text{of dimension } n \times c_1c_2 \tag{2.5}$$

where symbols $\otimes$, and $*$ are respectively the Kronecker and element-wise matrix products (Eilers et al., 2006). Smoothness is achieved by a penalty term applied on the regression coefficients $\boldsymbol{\theta}'\boldsymbol{P}\boldsymbol{\theta}$, where $\boldsymbol{P}$ has a block-diagonal form:

$$\boldsymbol{P} = \text{blockdiag}(0, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_{[1,2]}), \tag{2.6}$$

$\boldsymbol{P}_d$, for $d = 1, 2$ are penalty matrices for the main effects coefficients, given by $\boldsymbol{P}_d = \lambda_d \boldsymbol{D}_d' \boldsymbol{D}_d$ where λ_d is a smoothing parameter that controls the amount of smoothing along each covariate $\boldsymbol{x}_d$, and $\boldsymbol{D}_d$ is a difference matrix of order *pord* (usually second order). $\boldsymbol{P}_{[1,2]}$ is a penalty matrix for the interaction effect coefficients (where smoothness is controlled by λ_3 and λ_4), that can be written as a Kronecker sum:

$$\boldsymbol{P}_{[1,2]} = \lambda_3 \boldsymbol{D}_1' \boldsymbol{D}_1 \otimes \boldsymbol{I}_{c_2} + \lambda_4 \boldsymbol{I}_{c_1} \otimes \boldsymbol{D}_2' \boldsymbol{D}_2. \tag{2.7}$$

However, model (2.1) with basis in (2.4) has $\text{rank}(\boldsymbol{B}) = c_1 c_2$, so there exists an identifiability problem (this is well known in the context of additive models, but not in the case of multidimensional smoothing). In fact, there are $1 + c_1 + c_2$ linearly dependent columns in (2.4). The reason is that the space spanned by the tensor product $\boldsymbol{B}_{[1,2]}$, contains the space spanned by the marginal bases $\boldsymbol{B}_1$, and $\boldsymbol{B}_2$. Similarly, penalty matrix (2.6) must be carefully considered. For a tensor product B-spline basis $\boldsymbol{B}$ of rank $c_1 c_2$, the penalty matrix (2.6) has to be of rank $\leq c_1 c_2$. Lee and Durbán (in press) showed how to construct models of the form of (2.1) by reparameterizing the model basis and penalty.

2.2. Mixed model representation

Although a mixed model approach is not necessary to fit the models proposed in this paper (a reparameterization is enough), we use the mixed model representation of P-splines since it will allow an automatic estimation of the smoothing parameter using standard mixed model software (Ngo and Wand, 2004) based on restricted or residual maximum likelihood (REML). Reiss and Ogden (2009) showed how REML outperforms other selection criteria for the smoothing parameters such as Akaike's information criteria or (generalized) cross-validation. This representation consists of transforming the basis and coefficients of (2.1) into:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}) \text{ and } \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{G}), \tag{2.8}$$

where $\boldsymbol{X}$ is a model matrix for the fixed effects $\boldsymbol{\beta}$, $\boldsymbol{Z}$ is the model matrix for the random effects $\boldsymbol{\alpha}$ with covariance $\boldsymbol{G}$, and $\boldsymbol{\epsilon}$, is a vector of uncorrelated errors with zero mean and σ^2 variance. We basically have to find a reparameterization of the basis and coefficients, such that:

$$\boldsymbol{B} \rightarrow [\boldsymbol{X}|\boldsymbol{Z}], \quad \text{and} \quad \boldsymbol{\theta} \rightarrow (\boldsymbol{\beta}, \boldsymbol{\alpha})'.$$

Now, the smoothing parameters becomes the ratio $\lambda_d = \sigma^2/\tau_d^2$, for $d = 1, 2$, where τ_d^2 is the variance of the dth random effect. Hence, the covariance matrix becomes $\boldsymbol{G} = \sigma^2 \boldsymbol{F}^{-1}$, for some definite positive matrix $\boldsymbol{F}$ (Searle et al., 1992).

In the low-rank smooth-ANOVA context, Lee and Durbán (in press) showed that identifiability problems can be easily removed by finding transformation matrix $\boldsymbol{T}$ in such a way that

$$\boldsymbol{B}\boldsymbol{T} = [\boldsymbol{X}|\boldsymbol{Z}] \quad \text{and} \quad \boldsymbol{B}\boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha}.$$

Lee (2010) described this transformation matrix $\boldsymbol{T}$ which is based on the singular value decomposition of the penalty. Let $\boldsymbol{D}_d' \boldsymbol{D}_d = \boldsymbol{U}_d \boldsymbol{\Sigma}_d \boldsymbol{U}_d'$, for $d = 1, 2$, where $\boldsymbol{U}_d$ is the matrix of the eigenvectors, $\boldsymbol{\Sigma}_d$ is the diagonal matrix of eigenvalues, and $\boldsymbol{\Sigma}_{ds}$ is the sub-matrix of non-zero eigenvalues (the number of zero eigenvalues corresponds to the order of the penalty, in this case, 2). Then, the mixed model matrices for model (2.1) are:

$$\boldsymbol{X} = [\boldsymbol{1}_n | \boldsymbol{x}_1 \square \boldsymbol{1}_n | \boldsymbol{1}_n \square \boldsymbol{x}_2 | \boldsymbol{x}_1 \square \boldsymbol{x}_2], \quad \text{and} \tag{2.9}$$

$$\boldsymbol{Z} = [\boldsymbol{Z}_1 \square \boldsymbol{1}_n | \boldsymbol{1}_n \square \boldsymbol{Z}_2 | \boldsymbol{Z}_1 \square \boldsymbol{x}_2 | \boldsymbol{x}_1 \square \boldsymbol{Z}_2 | \boldsymbol{Z}_1 \square \boldsymbol{Z}_2] \tag{2.10}$$

where the random effects matrix as $\boldsymbol{Z}_d = \boldsymbol{B}_d \boldsymbol{U}_{ds}$, for $d = 1, 2$ and $\boldsymbol{U}_{ds}$ are the eigenvectors corresponding to the positive eigenvalues of the singular value decomposition of $\boldsymbol{D}_d' \boldsymbol{D}_d$. Lee and Durbán (in press) proved that this procedure of eliminating linear dependency through the mixed model reparameterization is exactly equivalent to imposing constraints on the regression coefficients $\boldsymbol{\theta}$ of the original model, and yields the usual sum-to-zero constraints in factorial designs (see Lee, 2010, Chapter 4, for details). Their approach is equivalent to define a new penalty matrix $\breve{\boldsymbol{P}}$, such that

$$\boldsymbol{\theta}' \breve{\boldsymbol{P}} \boldsymbol{\theta} = \boldsymbol{\theta}' (\boldsymbol{K}\boldsymbol{P}\boldsymbol{K}) \boldsymbol{\theta}, \tag{2.11}$$

where $\boldsymbol{K}$ is a orthogonal matrix, such that, $\boldsymbol{K} = \text{blockdiag}(1, \boldsymbol{K}_1 \boldsymbol{K}_2, \boldsymbol{K}_1 \otimes \boldsymbol{K}_2)$, where $\boldsymbol{K}_d = \boldsymbol{I}_d - \boldsymbol{1}_d' \boldsymbol{1}_d / c_d$ a centering matrix of order c_d, for $d = 1, 2$. Then, (2.11) is equivalent to $\breve{\boldsymbol{\theta}}' \boldsymbol{P} \breve{\boldsymbol{\theta}}$, where $\breve{\boldsymbol{\theta}}$ are the centered coefficients (more details can be found on Appendix A).

With the new reparameterization, the matrix $\boldsymbol{F}$ such that $\boldsymbol{G} = \sigma^2 \boldsymbol{F}^{-1}$, becomes a block-diagonal matrix:

$$\boldsymbol{F} = \text{blockdiag}(\boldsymbol{F}_1, \boldsymbol{F}_2, \boldsymbol{F}_{[1,2]}), \tag{2.12}$$

where $\boldsymbol{F}_1 = \lambda_1 \boldsymbol{\Sigma}_{1s}$, $\boldsymbol{F}_2 = \lambda_2 \boldsymbol{\Sigma}_{2s}$, correspond to the random effects coefficients for the main effects and $\boldsymbol{F}_{[1,2]}$ corresponds to the interaction random effect coefficients

$$\boldsymbol{F}_{[1,2]} = \begin{pmatrix} \lambda_3 \boldsymbol{\Sigma}_{1s} & & \\ & \lambda_4 \boldsymbol{\Sigma}_{2s} & \\ & & \lambda_3 \boldsymbol{\Sigma}_{1s} \otimes \boldsymbol{I}_{c_2-2} + \lambda_4 \boldsymbol{\Sigma}_{2s} \otimes \boldsymbol{I}_{c_2-2} \end{pmatrix}. \tag{2.13}$$

The estimation of the variance components can be done by *restricted or residual maximum likelihood* (REML) criteria, $\mathcal{L}_R(\tau_1^2, \tau_2^2, \sigma^2)$:

$$-\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}'(\boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}^{-1})\boldsymbol{y}, \tag{2.14}$$

where $\boldsymbol{V} = \sigma^2\boldsymbol{I} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$. Standard mixed models software (functions `lme` in `nlme` R-package Pinheiro and Bates, 2000, Ngo and Wand, 2004 or `PROC MIXED` procedure in SAS Ruppert et al., 2003) can be used for this type of S-ANOVA model only if a single smoothing parameter is used for the interaction term (i.e. $\tau_3 = \tau_4$). However, using the same amount of smoothing for the interaction may not be flexible enough in many situations, especially when covariates are measured on different scales or when heavy penalty is only applied on one covariate (i.e. there is a linear effect on one of the covariates, and hence $\lambda_d \to \infty$). If a more flexible model is preferred, the problem arises from the crossed covariance in the last block of (2.13), that makes the inverse of $\boldsymbol{F}$ non-standard. The function `gamm` in `mgcv` R-package uses a special type of `pdMat` class to consider covariance structures as in (2.13) (Wood, 2006b). Some efficient implementations have been proposed, as in Wood (2011), using Newton's method in several dimensions, but in general, the development of efficient algorithms to deal with such penalty structures are challenging, and in particular for more than two covariates. In the following section we propose a new model that is as flexible as the one introduced in this section, but with the advantage of having a single smoothing parameter controlling each term in the model.

3. Nested penalized splines smooth-ANOVA models

We propose a generalization of model (2.1) that is flexible, computationally efficient, and has (2.1) as a nested model, we use the term *PS*-ANOVA. The reparameterization of model basis and penalties given in the previous section shows that, naturally, the interaction term $f_{1,2}(\boldsymbol{x}_1, \boldsymbol{x}_2)$ in (2.1) could be expressed as:

$$f_{1,2}(\boldsymbol{x}_1, \boldsymbol{x}_2) \equiv g_1(\boldsymbol{x}_1)\boldsymbol{x}_2 + \boldsymbol{x}_1 g_2(\boldsymbol{x}_2) + h(\boldsymbol{x}_1, \boldsymbol{x}_2)$$

where we explicitly decompose $f_{1,2}$ into linear-by-smooth interactions (i.e. $g_1(\boldsymbol{x}_1)\boldsymbol{x}_2$ and $\boldsymbol{x}_1 g_2(\boldsymbol{x}_2)$) and smooth-by-smooth interactions ($h(\boldsymbol{x}_1, \boldsymbol{x}_2)$) of the covariates. The new model is:

$$\boldsymbol{y} = \gamma + f_1(\boldsymbol{x}_1) + f_2(\boldsymbol{x}_2) + g_1(\boldsymbol{x}_1)\boldsymbol{x}_2 + \boldsymbol{x}_1 g_2(\boldsymbol{x}_2) + h(\boldsymbol{x}_1, \boldsymbol{x}_2) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I}). \tag{3.1}$$

To construct model (3.1) from a *P*-spline approach, the full *B*-spline regression matrix has a column vector of ones for the constant and five blocks for each term, i.e.:

$$\boldsymbol{B} = [\boldsymbol{1}_n|\boldsymbol{B}_1|\boldsymbol{B}_2|\boldsymbol{B}_3|\boldsymbol{B}_4|\boldsymbol{B}_5], \tag{3.2}$$

with vector of regression coefficients $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4, \boldsymbol{\theta}_5)'$. The bases for the main effects $\boldsymbol{B}_1$ and $\boldsymbol{b}_2$ were already defined, and for the interaction effects ($\boldsymbol{B}_3$, $\boldsymbol{B}_4$ and $\boldsymbol{B}_5$) we have the tensor products:

$$\boldsymbol{B}_3 = [\boldsymbol{B}_1 \square \boldsymbol{x}_2], \qquad \boldsymbol{B}_4 = [\boldsymbol{x}_1 \square \boldsymbol{B}_2], \quad \text{and} \quad \boldsymbol{B}_5 = [\boldsymbol{B}_1 \square \boldsymbol{B}_2].$$

Observe that, $\boldsymbol{B}_3$ and $\boldsymbol{B}_4$ are the bases for a varying coefficient model (Hastie and Tibshirani, 1993) that allows for the smooth functions g_1 and g_2 to vary smoothly along $\boldsymbol{x}_2$ and $\boldsymbol{x}_1$ respectively (Eilers and Marx, 2002).

The penalty matrix for model (3.1) has a block-diagonal structure:

$$\boldsymbol{P} = \text{blockdiag}(0, \boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3, \boldsymbol{P}_4, \boldsymbol{P}_5). \tag{3.3}$$

We simplify the estimation procedure, by imposing separate and unique smoothing parameters for each block. Note that, we do not loose flexibility, since the interaction is controlled by three terms. The full matrix (3.2) has again rank c_1c_2, so now there are $1 + 2c_1 + 2c_2$ linearly dependent columns. Given the mixed model reparameterization in Section 2.2, it is straightforward to identify which components of the new basis correspond to each component of the nested model (3.1). After removing the linearly dependent columns we would obtain the same matrices given in (2.9) and (2.10), but we modify the random effects structure. We extend the smooth-ANOVA decomposition in Lee and Durbán (in press) by considering five random effects $\boldsymbol{\alpha}_k \sim \mathcal{N}(0, \boldsymbol{G}_k)$, for $k = 1, \ldots, 5$, that is, each group of coefficients is penalized by a single smoothing parameter. The matrices $\boldsymbol{Z}_k$, correspond to the same blocks described in (2.10). Now, since each component of $\boldsymbol{F}$ depends on a single smoothing parameter, we can simplify the model by defining $\boldsymbol{Z}_k^* = \boldsymbol{Z}_k\boldsymbol{F}_k^{-1}$. Then, the covariance matrices of the main random effects are: $\boldsymbol{G}_1 = \tau_1^2\boldsymbol{I}_{c_1-2}$, and $\boldsymbol{G}_2 = \tau_2^2\boldsymbol{I}_{c_2-2}$, and for the interactions are:

$$\boldsymbol{G}_3 = \tau_3^2\boldsymbol{I}_{c_1-2}, \qquad \boldsymbol{G}_4 = \tau_4^2\boldsymbol{I}_{c_2-2}, \quad \text{and} \quad \boldsymbol{G}_5 = \tau_5^2\boldsymbol{I}_{(c_1-2)(c_2-2)}.$$

Now, $\boldsymbol{G}$ has a simple structure that can be easily implemented in standard statistical software.

Gu (2002) and Wahba (1990) already addressed this kind of ANOVA-type models, by expansions of the corresponding reproducing kernel Hilbert function spaces, and in general, any parametric-by-smooth interaction can be included in the model. Note that, in our context, depending on the penalty order, we will also consider polynomial interactions. For sake of simplicity, we only consider second order penalties in this paper ($pord = 2$), but other penalty orders can be used. For example, with 3rd order penalties, we would have additional quadratic-by-smooth interactions terms of the form $g_3(\boldsymbol{x}_1)\boldsymbol{x}_2^2$, and $\boldsymbol{x}_1^2 g_4(\boldsymbol{x}_2)$, with the two additional random effects components and bases, i.e. $\boldsymbol{Z}_6 = [\boldsymbol{Z}_1 \square \boldsymbol{x}_2^2]$ and $\boldsymbol{Z}_7 = [\boldsymbol{x}_1^2 \square \boldsymbol{Z}_2]$, and $\boldsymbol{G}_6 = \tau_6^2 \boldsymbol{I}_{c_1-2}$, $\boldsymbol{G}_7 = \tau_7^2 \boldsymbol{I}_{c_2-2}$, respectively. Recently, a similar decomposition has been proposed by Wood et al. (in press), but the penalty term used does not correspond to the usual Kronecker sum of the marginal penalties.

Finally, as shown in Section 2.2, we can obtain a formulation of the model in terms of the original parameterization. In this case, we have to impose constraints on the interaction penalty blocks of matrix (3.3) (see details on Appendix B).

3.1. Estimation for nested P-splines smooth-ANOVA models

The estimation of the variance components in standard mixed models software, can be computationally intensive when it is used for multidimensional models. Here, we propose an alternative when the nested *PS*-ANOVA model formulation in Section 3 is used. The estimation approach is based on the work presented in Schall (1991) for the estimation of fixed and random effects coefficients ($\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$), and variance components ($\tau_1^2, \ldots, \tau_K^2$ and σ^2). The algorithm can be implemented for any number of covariates, but this paper we consider the bivariate case (where up to five random effects components are estimated). This algorithm yields maximum likelihood estimates in the case of Gaussian data and quasi-maximum likelihood estimates in the case of a Generalized Linear Mixed Model, and it is also equivalent to optimizing the adjusted profile likelihood used in Lee et al. (2006). To simplify the notation, we write model (3.1) as

$$\boldsymbol{y} = \gamma + \sum_{j=1}^{5} f_j + \boldsymbol{\epsilon}. \tag{3.4}$$

For two covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, the new model estimates a total of 5 random effects components. The log-likelihood in (2.14) can be written as:

$$-2\mathcal{L}(\tau_1, \ldots, \tau_5) = n \log \sigma + \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\alpha})}{\sigma^2} + \sum_{j=1}^{5} \left(k_j \log \tau_j + \log \sigma + \frac{\boldsymbol{\alpha}_j' \boldsymbol{\alpha}_j}{\tau_j^2} \right), \tag{3.5}$$

where k_j is the length of the jth random effect $\boldsymbol{\alpha}_j$. Given (3.5), it is easy to show that the estimates of the variance components are given by:

$$\widehat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{\alpha}})'(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\widehat{\boldsymbol{\alpha}})}{n - \mathrm{ED}}, \quad \text{and} \quad \widehat{\tau}_j^2 = \frac{\widehat{\boldsymbol{\alpha}}_j' \widehat{\boldsymbol{\alpha}}_j}{\mathrm{ED}_j}, \tag{3.6}$$

where "ED" is the effective dimension of the model, computed as the trace of the so-called hat-matrix, $\boldsymbol{H}$ (Hastie and Tibshirani, 1990), such that $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, where:

$$\boldsymbol{H} = \boldsymbol{C} \left(\boldsymbol{C}'\boldsymbol{C} + \boldsymbol{\Omega} \right)^{-1} \boldsymbol{C}', \tag{3.7}$$

with $\boldsymbol{C} = [\boldsymbol{X}|\boldsymbol{Z}]$, $\boldsymbol{\Omega} = \mathrm{blockdiag}(\boldsymbol{O}_p, \boldsymbol{G}^{-1})$, and p is the number of fixed effects (or columns of $\boldsymbol{X}$). The trace of $\boldsymbol{H}$ can be efficiently computed by:

$$\mathrm{ED} = \mathrm{trace}\left[\left(\boldsymbol{C}'\boldsymbol{C} + \boldsymbol{\Omega} \right)^{-1} \boldsymbol{C}'\boldsymbol{C} \right]. \tag{3.8}$$

Similarly, the effective dimensions associated with the jth smooth terms, ed_j, are computed as the trace of the hat matrix corresponding to each smooth term, i.e.

$$\boldsymbol{H}_j = \boldsymbol{C}_j \left\{ \left(\boldsymbol{C}'\boldsymbol{C} + \boldsymbol{\Omega} \right)^{-1} \right\}_j \boldsymbol{C}_j'. \tag{3.9}$$

Then we compute:

$$\mathrm{ED}_j = \mathrm{trace}\left[\left\{ (\boldsymbol{C}'\boldsymbol{C} + \boldsymbol{\Omega})^{-1} \right\}_j \boldsymbol{C}_j'\boldsymbol{C}_j \right], \tag{3.10}$$

where $\{\cdot\}_j$ denotes the diagonal block associated with the jth component, and $\boldsymbol{C}_j$ is the compound matrix of the fixed and random effects matrices of each jth component. Then:

$$\hat{\boldsymbol{y}} = \sum_{j=1}^{5} \boldsymbol{H}_j \boldsymbol{y}, \tag{3.11}$$

and $\mathrm{ed} = p + \sum_j^J \mathrm{ed}_j$, where p is the number of fixed effects (or columns of $\boldsymbol{X}$).

Given some starting values for the variance components, we compute ED_j as part of the estimation process, and obtain estimates of $\widehat{\sigma}^2$ and $\widehat{\tau}_j^2$ from the fit, and iterate until convergence. Notice that Schall's algorithm cannot be used unless we assume a single variance for the interaction random effect, otherwise it is not possible to obtain the effective dimensions associated with the variance components from a weighted sum of variances. In Section 5 we will use a simulation study to show the efficiency of this method of estimation compared with existing alternatives and software.

3.1.1. Standard errors and confidence bands

Standard errors and confidence intervals for each smooth term $\hat{f}_j$ can be easily obtained. Taking into account the randomness in the random effects $\boldsymbol{\alpha}$, the variance should be calculated with respect to the conditional distribution (Ruppert et al., 2003). The covariance matrix associated with the jth smooth component is:

$$\mathbb{V}\mathrm{ar}(\hat{f}_j|\boldsymbol{\alpha}) = \hat{\sigma}^2\boldsymbol{C}_j\left\{(\boldsymbol{C}'\boldsymbol{C}+\boldsymbol{\Omega})^{-1}\boldsymbol{C}'\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{C}+\boldsymbol{\Omega})^{-1}\right\}_j\boldsymbol{C}_j'. \tag{3.12}$$

The square root of the diagonal elements of (3.12) are used for confidence bands.

3.1.2. Non-Gaussian data

We can extend the algorithm to non-Gaussian responses in the context of GLMM's, such as Poisson or Binomial distributions. Consider the Poisson case with log link function. The linear predictor is $\boldsymbol{\eta} = \exp(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\alpha})$. Estimation is done by the iterative re-weighted least squares (IRLS) method, where the scoring algorithm becomes:

$$\begin{pmatrix}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{W}\boldsymbol{Z}+\boldsymbol{G}^{-1}\end{pmatrix}\begin{pmatrix}\widetilde{\boldsymbol{\beta}}\\ \widetilde{\boldsymbol{\alpha}}\end{pmatrix} = \begin{pmatrix}\boldsymbol{X}'\boldsymbol{W}\\ \boldsymbol{Z}'\boldsymbol{W}\end{pmatrix}\boldsymbol{z}, \tag{3.13}$$

where $\boldsymbol{z} = \boldsymbol{\eta}+\boldsymbol{W}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$ is the working vector, and $\boldsymbol{W}$ is a diagonal matrix of weights equal to $\boldsymbol{W} = \mathrm{diag}(\boldsymbol{\mu})$ with $\boldsymbol{\mu} = \exp(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\alpha})$. Now, the variance $\boldsymbol{V}$ is given by $\boldsymbol{V} = \boldsymbol{W}^{-1}+\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$. The hat-matrix is

$$\boldsymbol{H} = \boldsymbol{C}(\boldsymbol{C}'\hat{\boldsymbol{W}}\boldsymbol{C}+\boldsymbol{\Omega})^{-1}\boldsymbol{C}'\hat{\boldsymbol{W}}, \tag{3.14}$$

where $\hat{\boldsymbol{W}}$ is the weight matrix of the last iteration at convergence. The effective dimensions are now computed as:

$$\mathrm{ED} = \mathrm{trace}\left[(\boldsymbol{C}'\hat{\boldsymbol{W}}\boldsymbol{C}+\boldsymbol{\Omega})^{-1}\boldsymbol{C}'\hat{\boldsymbol{W}}\boldsymbol{C}\right], \quad \text{and} \tag{3.15}$$

$$\mathrm{ED}_j = \mathrm{trace}\left[\left\{(\boldsymbol{C}'\hat{\boldsymbol{W}}\boldsymbol{C}+\boldsymbol{\Omega})^{-1}\right\}_j\boldsymbol{C}_j'\hat{\boldsymbol{W}}_j\boldsymbol{C}_j\right]. \tag{3.16}$$

For Poisson data, $\sigma^2 = 1$, and we simply use $\lambda_j^{-1} = \boldsymbol{\alpha}_j'\boldsymbol{\alpha}_j/\mathrm{ED}_j$. As shown in the Gaussian case, standard errors and confidence intervals for $\hat{f}_j$ can be obtained including the diagonal matrix of weights $\boldsymbol{W}_j$ in (3.12).

Schall's algorithm provides an automatic procedure to check the presence of over or underdispersion, a common problem in Poisson and Binomial models, if $\mathbb{V}\mathrm{ar}[\boldsymbol{y}] = \sigma^2\boldsymbol{\mu}$. At convergence, we compute $\widehat{\sigma}^2$ as:

$$\widehat{\sigma}^2 = \frac{(\boldsymbol{y}-\widehat{\boldsymbol{\mu}})'\hat{\boldsymbol{W}}^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\mu}})}{n-\mathrm{ED}}. \tag{3.17}$$

Then, for values of $\hat{\sigma}^2$ significantly larger or smaller than 1 would indicate over or underdispersion respectively. In general, it is preferable to incorporate over/underdispersion directly into the estimation process and this can be easily done in mixed models with a little modification of Schall's algorithm and estimate as (3.17).

4. Nested *B*-splines bases

Tensor product B-spline models may be constrained to the total number of parameters to be estimated. In the construction of the model bases, the number of B-spline basis functions c_d depends on the selection of some parameters: (i) the number of equally spaced knots ndx_d, (strictly $ndx_d - 1$ is the number of internal knots in the domain of the covariate); (ii) the degree of the B-spline basis, *bdeg*, usually a cubic spline, and (iii) the order of the penalty, *pord* (usually of second order). The total number of knots for the construction of a marginal B-spline basis, $\boldsymbol{B}_d$, are $ndx_d + 2bdeg + 1$, and the number of B-splines coefficients are $c_d = ndx_d + bdeg$. Then, for the smooth-ANOVA model in (2.1), the total number of coefficients to estimate are $1 + c_1 + c_2 + c_1c_2$. The usual choice is a moderate number of equally-spaced knots, in general a number between 20 and 40 knots (Ruppert et al., 2003).

Using a moderate number of knots, the interaction basis $\boldsymbol{B}_{[1,2]}$ in (2.5) is not too large, and computation may not be in general very intensive. However, in some situations, we need to provide more flexibility increasing the number of knots for $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$, this may lead to a very large full regression matrix $\boldsymbol{B}$ and the computational burden might be prohibitive. Model bases in (2.4) and (3.2) are constructed using the same marginal basis to ensure that both additive and ANOVA models are strictly nested. To avoid computational limitations, a simple solution is to reduce the number of parameters

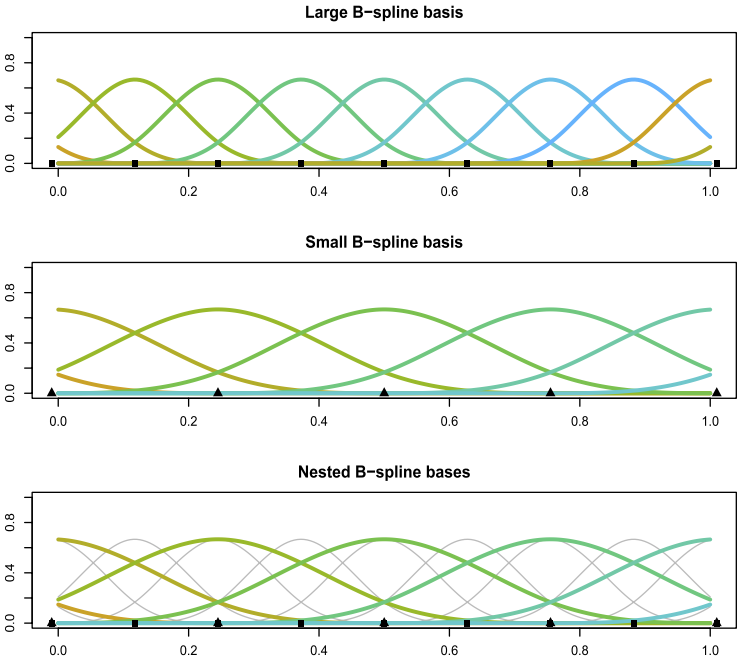


Fig. 4.1. Illustrative visualization of nested *B*-spline bases.

for the interaction term, by reducing the number of knots for the marginal bases for the tensor product. This idea can be explained by analogy to classical ANOVA models, where in general the main effects are more significant than interactions. In the smoothing context, additive terms would explain most of the structure, and we expect less complex interaction effects. Then, a lower dimension interaction basis would be the tensor product of two marginal lower rank *B*-spline bases: $\boldsymbol{B}_{[1,2]} = \breve{\boldsymbol{B}}_1 \square \breve{\boldsymbol{B}}_2$, of size $n \times \breve{c}_1\breve{c}_2$, such that:

$$\mathsf{rank}(\breve{\mathbf{B}}_d) < \mathsf{rank}(\boldsymbol{B}_d) \rightarrow \breve{c}_d < c_d, \quad \text{for } d = 1, 2.$$

Reducing the rank of the *B*-spline basis for the interaction, we reduce the number of parameters to estimate for the interaction to $\breve{c}_1\breve{c}_2 < c_1c_2$. However, taking a reduced basis of arbitrary size will yield a model that will not be nested to the additive model, hence the comparison between additive and ANOVA models will not be straightforward.

We propose the use of *nested B-spline bases* for the interaction term: a *B*-spline basis such that the space spanned by $\breve{\boldsymbol{B}}_d$, is a subset of the space spanned by $\boldsymbol{B}_d$, such that the hierarchical nature of the models is preserved. Now, with the new nested *PS*-ANOVA decomposition in (3.1).

In the ANOVA context, the main effects are more important than the interactions, so in most situations this would be reasonable. Using these nested *B*-spline bases, the identifiability constraints and model formulation remain the same, and the total number of parameters and size of the full regression basis is dramatically reduced. The way to ensure that the new basis is nested relative to the original basis is choose the number of knots (`ndx`*) as a divisor of the number of knots used in the original basis (`ndx`), i.e.:

$$ndx^* \text{ of } \breve{\boldsymbol{B}}_d = \frac{ndx \text{ of } \boldsymbol{B}_d}{\text{div}} \Rightarrow \mathsf{span}(\breve{\boldsymbol{B}}_d) \subset \mathsf{span}(\boldsymbol{B}_d),$$

and div is any divisor of the number of knots used to construct $\boldsymbol{B}_d$. Given this reduction, we construct the marginal penalty for the interaction coefficients as the Kronecker sum (2.7) with $\breve{\boldsymbol{D}}_d'\breve{\boldsymbol{D}}_d$, of dimension $\breve{c}_d \times \breve{c}_d$.

Fig. 4.1 shows two *B*-spline bases with $ndx = 8$ and $ndx^* = 4$ knots (div $= 2$), for $bdeg = 3$, and $pord = 2$. Then, we have large *B*-spline basis $\boldsymbol{B}$ of size $n \times c$, with $c = 11$, and a smaller $\breve{\boldsymbol{B}}$ of size $n \times \breve{c}$, with $\breve{c} = 8$. The plot at the bottom shows both bases overlapped. The inner knots of $\breve{\boldsymbol{B}}$ are placed between the inner knots of $\boldsymbol{B}$, and consequently the space spanned by one *B*-spline in the small basis is spanned by 3 *B*-splines of the large basis.

5. Simulation studies

In this section, we conduct three simulation studies to compare the performance of the different smooth models and fitting algorithms discussed in the previous sections. The aims of these studies are: (i) check model and estimation algorithms performance; (ii) compare computing time required of Schall's algorithm and standard mixed models software, and (iii) check computational efficiency and model performance of using nested *B*-spline bases for interactions.

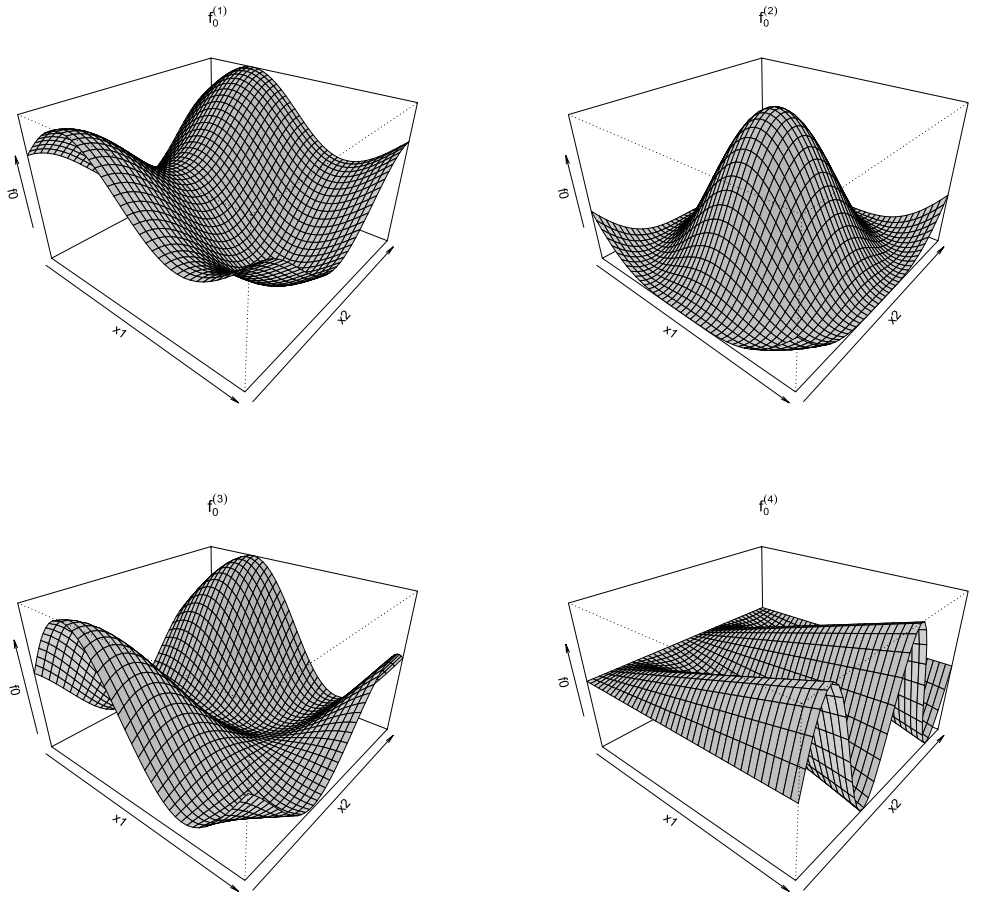


Fig. 5.1. Perspective plots of surfaces used in the simulation study.

5.1. Simulation study 1: model performance

The aim of the first simulation study is to check the performance of alternative models in terms of recovering the true simulated surface. We simulated data $\boldsymbol{y}$ as a function of covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We considered the following true functions (shown in Fig. 5.1):

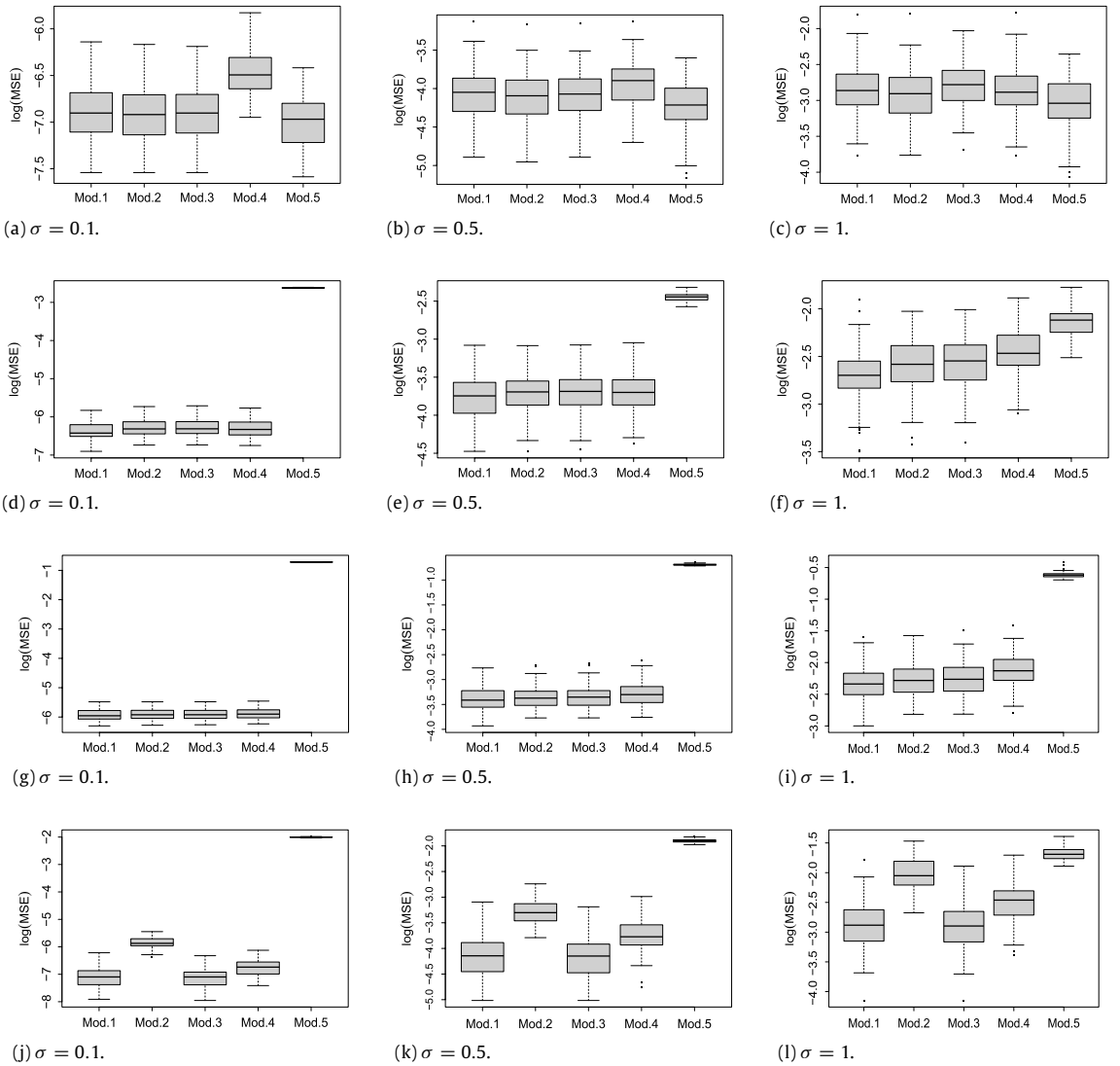
$$f_0^{(1)} = \sin(2\pi\, \boldsymbol{x}_1) + \cos(2\pi\, \boldsymbol{x}_2), \quad \text{(Additive surface)}$$
$$f_0^{(2)} = \cos(2\pi\, \sqrt{(\boldsymbol{x}_1 - 0.5)^2 + (\boldsymbol{x}_2 - 0.5)^2}), \quad \text{(Non-linear interaction surface)}$$
$$f_0^{(3)} = f_0^{(1)} + \sin(2\pi\, (\boldsymbol{x}_2 - \boldsymbol{x}_1)), \quad \text{(Additive plus interaction surface)}$$
$$f_0^{(4)} = 2\boldsymbol{x}_1 \sin(4\pi \boldsymbol{x}_2), \quad \text{(Linear by non-linear interaction surface).}$$

For each function $f_0^{(1-4)}$, we fitted five models: `Mod.1`–`Mod.5` (see Table 5.1). `Mod.1`–`Mod.3` are smooth-ANOVA models with different penalties for the interaction term. We calculated the Mean Square Error (MSE) as a measure of fit. In particular, `Mod.1` is the *PS*-ANOVA in (3.1) which penalizes the interaction term at three levels (*linear-by-smooth*, *smooth-by-linear*, and *smooth-by-smooth*) and estimates a total of five smoothing parameters, `Mod.2` does not estimate explicitly the *smooth-by-linear* interaction, and penalizes the interaction with a single smoothing parameter (so that Schall's algorithm can be used); and `Mod.3` is similar to `Mod.2`, but it controls the smoothness of the whole interaction with two smoothing parameters (model (2.1)). Notice that, the smoothing parameters of `Mod.3` and `Mod.4` are estimated by direct optimization of the REML function using the `optim` function in R. Finally, we fitted and additive model to use it as a baseline for comparison (`Mod.5`). Simulations were performed using R version 2.13.0, with a 2.66 GHz Intel Core-Duo Pentium processor and 2 GB of RAM.

We considered a relatively small sample size of $n = 200$ scattered data points, chosen from two covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ from a uniform distribution on the interval [0, 1], with different levels of noise: $\sigma = \{0.1; 0.5; 1\}$ and $R = 100$ replicates. Additionally, we chose the parameters: $bdeg = 3$ and $pord = 2$, and for the number of knots, $ndx_1 = ndx_2 = 12$. Fig. 5.2 show the boxplots of the log(MSE) for the fitted models.

Table 5.1
Summary of fitted models, number of smoothing parameters in parenthesis, and the estimation method used.

	Smooth model formulation	Estimation
`Mod.1(5)`	$f_1(\boldsymbol{x}_1)+f_2(\boldsymbol{x}_2)+g_1(\boldsymbol{x}_1)\boldsymbol{x}_2+\boldsymbol{x}_1 g_2(\boldsymbol{x}_2)+h(\boldsymbol{x}_1,\boldsymbol{x}_2)$	Schall
`Mod.2(3)`	$f_1(\boldsymbol{x}_1)+f_2(\boldsymbol{x}_2)+f_{1,2}(\boldsymbol{x}_1,\boldsymbol{x}_2)$	Schall
`Mod.3(4)`	$f_1(\boldsymbol{x}_1)+f_2(\boldsymbol{x}_2)+f_{1,2}(\boldsymbol{x}_1,\boldsymbol{x}_2)$	REML
`Mod.4(2)`	$f_{1,2}(\boldsymbol{x}_1,\boldsymbol{x}_2)$	REML
`Mod.5(2)`	$f_1(\boldsymbol{x}_1)+f_2(\boldsymbol{x}_2)$	Schall



(a) $\sigma = 0.1$. (b) $\sigma = 0.5$. (c) $\sigma = 1$.

(d) $\sigma = 0.1$. (e) $\sigma = 0.5$. (f) $\sigma = 1$.

(g) $\sigma = 0.1$. (h) $\sigma = 0.5$. (i) $\sigma = 1$.

(j) $\sigma = 0.1$. (k) $\sigma = 0.5$. (l) $\sigma = 1$.

Fig. 5.2. From top to bottom: log(MSE) of fitted smooth models with $f_0^{(1)} - f_0^{(4)}$ true functions.

The results of the simulation study for each scenario are summarized as follows: `Mod.5` has the worst performance in all situations except when $f_0^{(1)}$ was used to simulate the data (this could be expected since $f_0^{(1)}$ is the sum of two univariate smooth functions). `Mod.4` fails when $f_0^{(1)}$ and $f_0^{(4)}$ are used (main effects or linear by non-linear interactions are present), since the model uses the same amount of smoothness for main effects and interactions. `Mod.3` has a poor performance in the last scenario because the interaction is estimated using a single smoothing parameter. Finally, `Mod.1` and `Mod.3` estimate the have the smallest MSE in all situations, indicating that the interaction should be controlled by more than one smoothing parameters (3 in the case of `Mod.1`, and 2 for `Mod.3`), and additive terms should also be in the model to guarantee flexibility. However, as we will see in the next simulation study, `Mod.1` will outperform `Mod.3` in terms of computational efficiency.

Table 5.2
Comparison of CPU times (in seconds) of Schall's algorithm with respect to standard mixed models software in R. Symbol '–' denotes that models could not be fitted due to memory allocation problems or convergence issues.

Sample size	(ndx_1, ndx_2)	Schall	`lme`	`gamm`
1000	(20,20)	6.77	813.69	37.84
1500	(30,30)	57.78	–	424.73
3000	(40,40)	330.74	–	–

Table 5.3
Number of knots for marginal *B*-spline bases (ndx_1, and ndx_2) and reduced nested bases (ndx_1^* and ndx_2^*), residual sum of squares, CPU time in seconds, and effective dimensions.

Model	ndx_1	ndx_2	ndx_1^*	ndx_2^*	RSS	CPU time	ED
A	30	30	30	30	243.23	72	83.28
B	30	30	15	15	246.66	3.27	75.12
C	30	30	10	10	249.78	0.97	68.04

5.2. Simulation study 2: CPU time and standard mixed models software

In this simulation study, we compare the performance of the *PS*-ANOVA model (`Mod.1`) in terms of central processing unit (CPU) times for larger samples sizes with respect to available standard mixed model software. We made comparisons using the `lme` function in `nlme` R-package (version 3.1–101, Pinheiro and Bates, 2000), and `gamm` in R-packages `mgcv` (version 1.7–6), available from `cran.r-project.org`. Recent versions of the `mgcv` package incorporate a function `t2` to estimate tensor product smooths as shown in Section 3, where the interaction terms are decomposed as a sum of random effects.

We simulated data from the main effects with interaction true surface $f_0^{(3)}$ and $\sigma = 0.5$, and compared the performance for a single replicate (the aim of this simulation exercise is to check the estimation timings, so considering different values of σ and increasing the number of replicates does not affect the conclusions of this exercise). Table 5.2 shows the sample size, number of knots (ndx_1, and ndx_2) for the construction of the marginal *B*-spline bases, and CPU time in seconds for Schall and the alternative methods. As the sample size and number of knots increases, standard software computation becomes more intensive. For sample sizes of 1500 and 3000 data points, and 30 and 40 knots respectively, the standard mixed model functions `lme` and `gamm` were not able to fit the data.

5.3. Simulation study 3: computational efficiency and nested B-spline bases

In this simulation study we simulated data from an additive plus interaction true function given by:

$$f_0^{(5)} = 1.5\sin(12\pi \boldsymbol{x}_1^2) + \cos(2\pi(\boldsymbol{x}_2)) + \sin(2\pi(\boldsymbol{x}_2 - \boldsymbol{x}_1)).$$

The true surface, main effects and interaction surfaces are shown in Fig. 5.3. We simulated a main effect for $\boldsymbol{x}_1$ with a high frequency sine, to illustrate the need for constructing a large basis $\boldsymbol{B}_1$ to smooth along $\boldsymbol{x}_1$. As in Section 5.2, we simulated a single replicate with $\sigma = 0.5$ and $n = 3000$.

Table 5.3 shows the number knots chosen to construct the marginal *B*-spline bases. Model `A` corresponds to a *PS*-ANOVA model with no reduction on the interaction terms bases, and models `B` and `C` correspond to reductions of the bases by divisors div $= \{2, 3\}$ from the original basis. Additionally, we show the residual sum of squares and CPU time. The table shows how the use of nested *B*-spline bases, has no significant loss in terms of the fit to the data, and remarkable results in computing time. Differences among effective dimension correspond to the reduced bases considered in the interaction terms, and reflects that a parsimonious model (with much less coefficients) is able to describe the simulated interaction. For the sake of comparison, we have simulated data where a model with no reduction can be fitted (using the Schall's algorithm, since using `lme` or `gamm` functions could not cope with the size of the basis); however, nested *B*-splines basis are particularly suitable when the dataset is so large that interactions cannot be estimated (due to computational issues) unless the number of parameters used is small.

In this simple simulation study, we showed how it can lead to good solution for multidimensional problems, and it might work well for more complex situations, where in general the structure of the main effects have more important features than the interactions, as for example in signal regression problems and multidimensional (optical) spectra. The selection of the divisor was also arbitrary, but in general div $= \{2, 3\}$ might be enough to smooth interaction effects, without significant loss on model performance.

6. Examples

In this section, we present the analysis of the nested *PS*-ANOVA model in (3.1) applied to two real datasets.

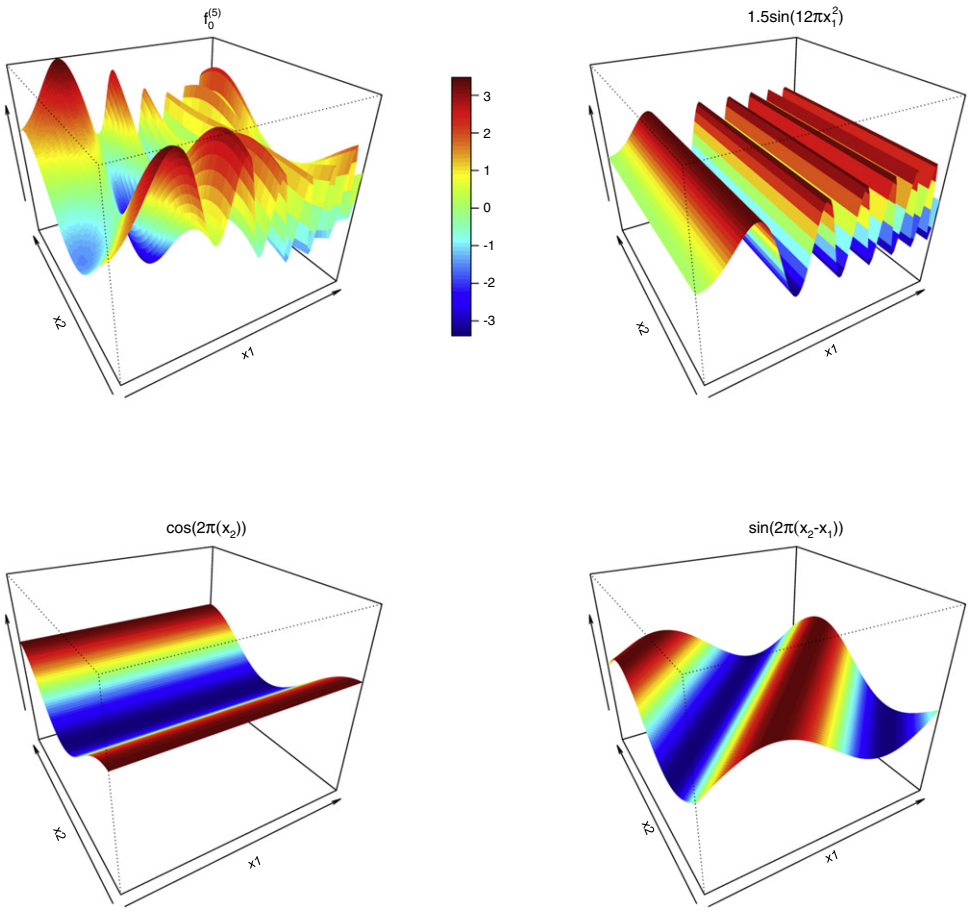


Fig. 5.3. True surfaces for simulation study 3.

6.1. Prestige data

The data consist of prestige scores for occupation, from a social survey conducted in the mid-1960s (Duncan, 1961). We take the average `education` (in years) and average `income` (in logarithm of thousands of dollars) for 102 occupations in Canada as explanatory variables. Fig. 6.1 shows the prestige scores by log income and education. We fitted model (3.4), with 20 knots for each marginal *B*-spline basis. The fitted model shows that, the effect of the log(`income`) is linear; the main effect of `education` has an effective dimension of 3.125, and the only significant interaction is g_1(log(`income`)) : `education` with effective dimension of 3.023. The rest of smooth terms have an estimated effective dimension close to zero, and, therefore, are dropped from the model. If a model that did not include linear by non-linear term was used, a simple additive model would have been selected. Fig. 6.2 shows the smooth terms with effective dimension larger than zero, and the fitted surface. The total effective dimension was 8.148. As expected, the fitted model shows that, higher prestige scores are related to high incomes and more years of education. However, high incomes and few years of education does not result in high prestige scores (this effect would have been lost without the interaction term).

6.2. Mortality data

We consider the male policyholders data from the Continuous Mortality Investigation Bureau in the UK. For each calendar year ($\boldsymbol{x}_a$) from 1947 to 1999 and each age ($\boldsymbol{x}_y$) from 11 to 100 we have the number of policy claims (deaths, the response $\boldsymbol{y}$) and the number of years lived (the exposure, $\boldsymbol{e}$). The aim of this study is to model mortality trends over time and age of death. The mortality of male policyholders has improved rapidly over the last thirty years and this has important financial implications for the insurance industry. Currie et al. (2004) analyzed this dataset and studied mortality trends over time and age and proposed methods for projection of mortality tables. We considered three models shown in Table 5.1: additive mixed model (`Mod.5`), a 2*d* model (`Mod.4`) and *PS*-ANOVA model (`Mod.1`). Model matrices are constructed using the Kronecker product of *B*-spline bases of age and year. Let $\boldsymbol{B}_a = \boldsymbol{B}(\boldsymbol{x}_a)$, $n_a \times c_a$, be the marginal *B*-spline bases of the age covariate,

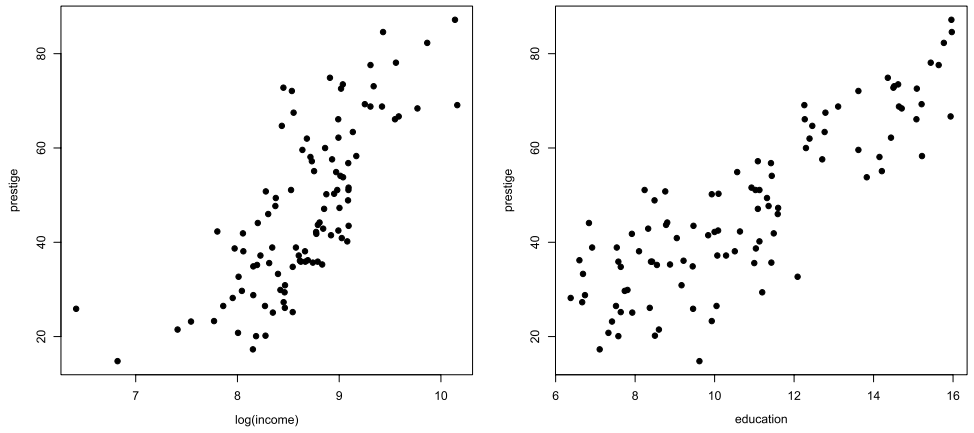


Fig. 6.1. Prestige scores for occupation by education and log income.

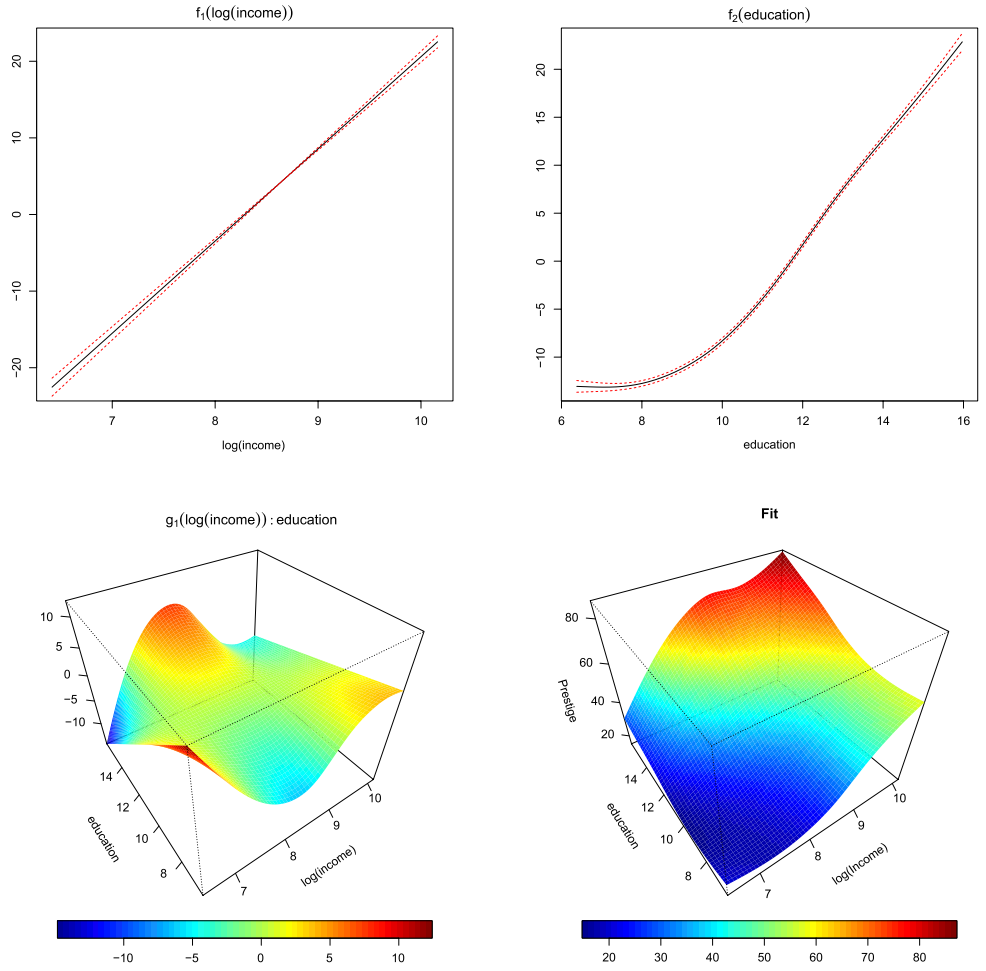


Fig. 6.2. Top: main effects of income and education. Bottom: linear-by-smooth interaction effect of income and education, and fitted surface. The total effective dimension of the fitted model is 8.148.

and similarly $\boldsymbol{B}_y = \boldsymbol{B}(\boldsymbol{x}_y)$, $n_y \times c_y$, for the year covariate $\boldsymbol{x}_y$. We chose 20 and 12 knots respectively, such that $\boldsymbol{B}_a$ and $\boldsymbol{B}_y$ are of sizes, 90×23, and 53×15 respectively. The additive and *PS*-ANOVA models were fitted using Schall's algorithm in Section 3.1 for Poisson data, and the 2*d* model was fitted by PQL (Breslow and Clayton, 1993). Fig. 6.4 shows the fitted

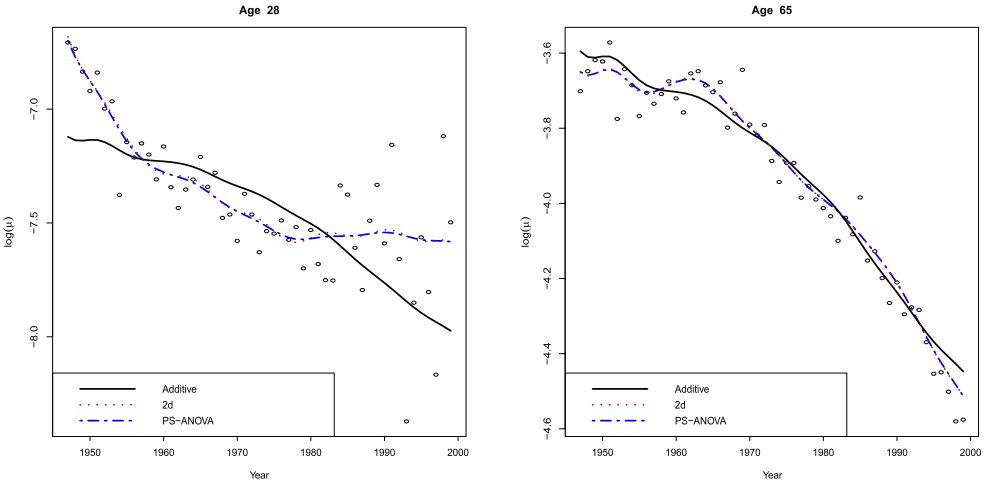


Fig. 6.3. Log mortality against year for ages 28 (left) and 60 (right) of additive, 2*d* and *PS*-ANOVA models.

Table 6.1
Values of ed, AIC and BIC.

Model	ED	AIC	BIC
Additive	32.44	11 905.94	12 115.85
2*d*	150.83	8093.51	9069.43
PS-ANOVA	128.61	8090.79	8922.91

surface, main effects of age and year, and the interaction effects decomposition of the *PS*-ANOVA model. The model fit also shows a more important smooth interaction effect of age and year (with an effective dimension of 81.03), rather than the linear-by-smooth interaction effects of age and year. Table 6.1 shows the estimated effective dimension, and (Akaike, 1973) and Bayesian information criteria (Schwarz, 1978) for model selection:

$$\mathrm{IC} = \mathrm{Dev} + \delta \times \mathrm{ED}, \tag{6.1}$$

where Dev is the deviance of a Poisson generalized linear model (see Cameron and Trivedi, 1998, for details), ED is the effective dimension, and δ is a parameter that penalizes the dimension of the model, $\delta = 2$ for AIC and $\delta = \log(n)$ for BIC. Best results are obtained by the 2*d* and *PS*-ANOVA models, with less effective dimension and model selection criteria than for the additive model. This is not surprising as the additive model fits a single smooth effect for age and a single smooth effect for year (no interaction), and the fitted log mortality against year for different ages are parallel. Notice that, 2*d* and *PS*-ANOVA models have similar AIC values, but in terms of BIC, *PS*-ANOVA model gives better results, as BIC tends to select parsimonious models, and shows the effectiveness of the *PS*-ANOVA model penalty.

Fig. 6.3 plots the log mortality against year at ages 28 and 60 to show the different pattern of mortality at younger ages from those at older ages. At age 28, the additive model fit is very poor, and the 2*d* and *PS*-ANOVA models have similar fits. At younger ages, the *PS*-ANOVA model fit is slightly smoother than 2*d* showing more flexibility allowed by the extra smoothing parameters for the interaction effects. At older ages, both fits are indistinguishable.

A decomposition of the smooth components of the *PS*-ANOVA model fit is given in Fig. 6.4. Note that, for this data set, the smooth-by-smooth interaction effect has a significant structure (with effective dimensions 9.717 and 5.501).

7. Discussion and further work

We considered the problem of efficient estimation in bivariate smoothing problems. The model formulation proposed in Lee and Durbán (in press) is extended to consider interaction terms as a sum of several nested constructions. This formulation allows us to separate interaction effects into parametric-by-smooth, and smooth-by-smooth interaction terms in the *P*-spline framework. This type of decomposition was addressed by Wahba (1990) and Gu (2002), but the use of *P*-splines gives a simple procedure that shows how the different ways of formulating a S-ANOVA model are equivalent to consider different penalizations on the coefficients' parameter space, and allow an efficient estimation of the model parameters.

Recently other authors have proposed similar low-rank smooth-ANOVA approaches (Wood, 2006a; Belitz and Lang, 2008). All models are basically similar, but differ in how the identifiability constraints are imposed. Wood (2006a) uses a similar decomposition based on a mixed model reparameterization, but removes the linear dependency numerically using a QR algorithm, and Belitz and Lang (2008) consider a penalty matrix as a Kronecker sum of three terms: first two terms

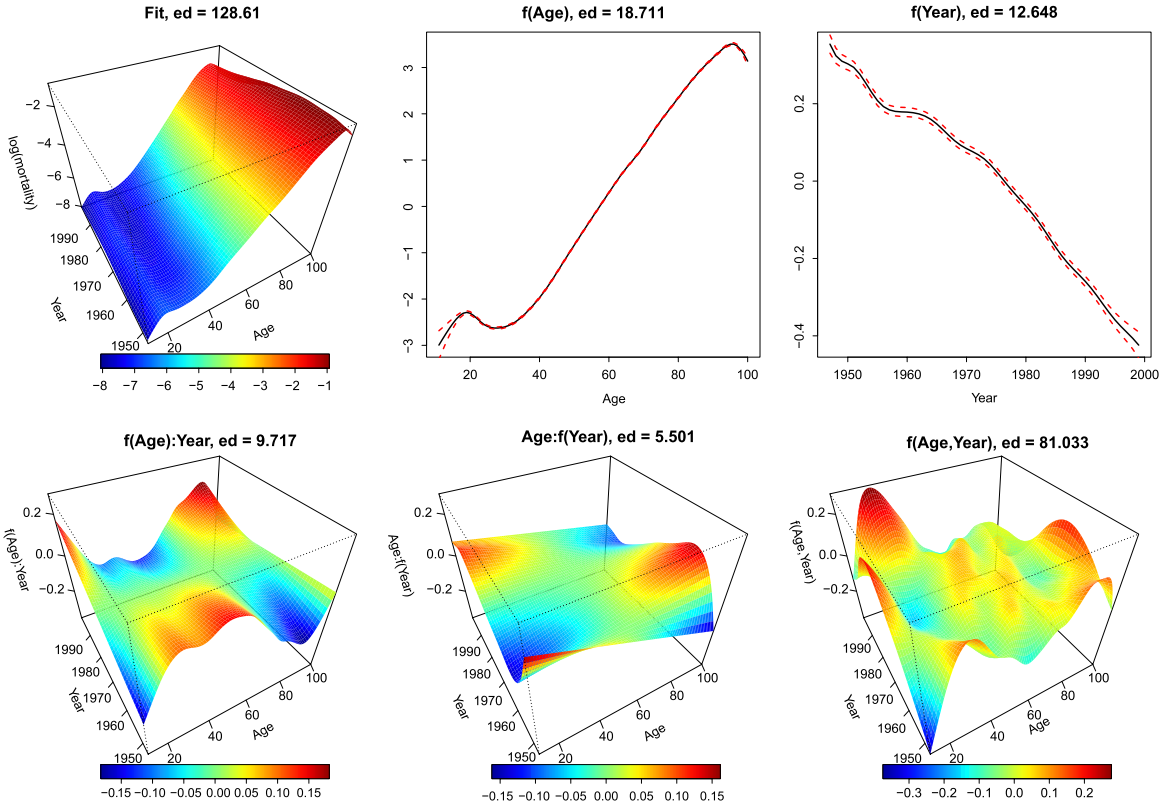


Fig. 6.4. Top: fitted surface, and additive terms for `Age`, and `Year`. Bottom: decomposition of the interaction terms for `Age`, and `Year`.

for the main effects (with some constraints), and a third penalty form as the Kronecker product of two marginal first order penalties. In fact, all approaches yield to smooth-ANOVA type models, and differ only on how penalties are imposed.

The proposed model construction procedure makes it possible to adapt the algorithm introduced by Schall (1991) for variance components estimation. In general, the algorithm converges in less than 100 iterations. The number of iterations until convergence may depend on the starting values for the variance components. In practice a good strategy is to fit the additive model and take the estimates of the additive variance components as starting values. An important strength of the algorithm is that effective dimensions are computed as part of the algorithm at each iteration. This might be important in practice, if after several iterations a smooth terms has an effective dimension close to zero, it can be dropped out of the model, and therefore a model selection procedure is implicitly part of the estimation process. For formal testing procedures, the algorithm provides an automatic method for testing for additive versus ANOVA models based on F-tests and efficient computations based on quadratic forms (Bowman and Azzalini, 1997). However, these tests are just an approximation and might be problematic in combination with mixed models. another approach that can be used in this context is the one proposed by Scheipl et al. (2008) where they present tests for a zero random effect variance in additive and linear mixed models. This method is easily applied in our case, since each smooth term depends only on one single smoothing parameter. This is a current line of research, as the nested *PS*-ANOVA model gives the possibility of testing among alternative models and interaction types. A supplementary R-code can be downloaded from http://www.est.uc3m.es/durban/software/DemoSchall-PS-ANOVA.html, which implements the Schall *PS*-ANOVA model proposed in this paper. An extension of the Schall algorithm to estimate models with multiple variance components for the interaction is a current topic of research. This extension is not straightforward, and requires a new reformulation of the mixed model equations. This extension is working progress and out of the scope of this manuscript. Our idea is to implement all these models in to an R package.

The computational efficiency was illustrated in the simulations studies of Section 5, where the algorithm results are very fast compared to available standard mixed models software tested. We also provide the construction of B-spline bases of lower dimension for the interaction terms. This simple approach is an efficient method for fitting models with interactions. However, there might be situation where a more complex structure is needed for interactions (as high frequency interaction signals), hence other type of strategies should also be considered.

For large multidimensional grids it is possible to use the array arithmetic proposed in Currie et al. (2006) and Eilers et al. (2006), called generalized linear array methods (or GLAM). We are researching the application of the Schall algorithm and nested bases proposed in this paper with array methods. We have some positive results in particular cases and large sample sizes and work continues on a general solution.

Acknowledgments

We are grateful to the associate editor and the anonymous referees for their very thoughtful and helpful comments and suggestions that improved this paper. This work is supported by the Spanish Ministry of Science and Innovation (projects MTM 2008-02901 and MTM2011-28285-C02-02). The research of Dae-Jin Lee was funded by the US National Institutes of Health grant for the Superfund Metal Mixtures, Biomarkers and Neurodevelopment (project 1PA2ES016454-01A2).

Appendix A. Interpretation of penalty in *PS*-ANOVA models

The formulation in Lee and Durbán (in press) gives a direct interpretation of the identifiability constraints in terms of the penalties on the regression *P*-spline coefficients. To illustrate this interpretation, consider a smooth term with vector of coefficients $\boldsymbol{\theta}$, of length $c \times 1$, and penalty $\boldsymbol{D}'\boldsymbol{D}$, $c \times c$, and the singular value decomposition $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}'$. Suppose a second order penalty is used, then $\boldsymbol{\Sigma}$ has two zero eigenvalues and $\boldsymbol{\Sigma}_s$ are the positive eigenvalues. Let $\boldsymbol{U}_n$, $(c \times 2)$, be the submatrix containing the eigenvectors corresponding to the zero eigenvalues and $\boldsymbol{U}_s$, $c \times (c-2)$, the submatrix containing the eigenvectors corresponding to the non-zero eigenvalues. Notice that $\boldsymbol{U}_n$ forms a basis of the null space of the penalty (with a second order penalty this space corresponds to the constant and linear terms spaces). Similarly, $\boldsymbol{U}_s$ forms a basis of the non-null space of the penalty (the smooth term). Therefore, we can split the penalty on $\boldsymbol{\theta}$ into null and non-null space penalties, as

$$\boldsymbol{\theta}'\boldsymbol{U}_n\boldsymbol{U}_n'\boldsymbol{\theta} \quad \text{and} \quad \boldsymbol{\theta}'\boldsymbol{U}_s\boldsymbol{U}_s'\boldsymbol{\theta},$$

where $\boldsymbol{U}_n\boldsymbol{U}_n'$, $c \times c$, is a penalty matrix on the null space of the penalty, and $\boldsymbol{U}_s\boldsymbol{U}_s'$, $c \times c$, is a penalty matrix on the non-null space of the penalty. Both submatrices are orthogonal, and therefore:

$$\boldsymbol{I}_c = \boldsymbol{U}_s\boldsymbol{U}_s' + \boldsymbol{U}_n\boldsymbol{U}_n', \tag{A.1}$$

and

$$\boldsymbol{K}_s = \boldsymbol{U}_s\boldsymbol{U}_s' + \boldsymbol{u}_n^{(2)}\boldsymbol{u}_n^{(2)\prime}, \tag{A.2}$$

where $\boldsymbol{u}_n^{(2)}$ is the second column of the submatrix $\boldsymbol{U}_n$, with $\mathrm{rank}(\boldsymbol{K}_c) = c - 1$. Then, the penalty matrix for $\boldsymbol{\Theta}$ is:

$$\boldsymbol{P}_{[1,2]} = \lambda_3\boldsymbol{D}_1'\boldsymbol{D}_1 \otimes \boldsymbol{K}_{2s} + \lambda_4\boldsymbol{K}_{1s} \otimes \boldsymbol{D}_2'\boldsymbol{D}_2, \quad \text{of rank } (c_2-1)(c_1-1) - 1, \tag{A.3}$$

and the full block diagonal penalty matrix $\boldsymbol{P}$ in (2.6) has rank $c_1c_2 - 4$. Lee (2010) showed that taking $\boldsymbol{u}_n^{(2)}$ as a vector $(1, \ldots, c)'$ centered and scaled to have unit length, matrix (A.2) is a centering matrix, i.e. $\boldsymbol{K}_s = \boldsymbol{I}_c - \boldsymbol{1}_c\boldsymbol{1}_c'/c$, applied on one of the dimensions (row or columns) of the interaction matrix of coefficients $\boldsymbol{\Theta}$, $c_1 \times c_2$, that yields to the sum-to-zero constraints on the vector of regression coefficients as in a classical ANOVA model (see Lee, 2010, for details).

Appendix B. Penalty in nested S-ANOVA models

We can obtain an equivalence of the mixed model formulation in terms of the original model. In this case, this model is equivalent to imposing constraints on the interaction penalty blocks of matrix (3.3), such that, penalties for coefficients $\boldsymbol{\theta}_3$, $\boldsymbol{\theta}_4$ and $\boldsymbol{\theta}_5$ are:

$$\boldsymbol{P}_3 = \lambda_3\boldsymbol{D}_1'\boldsymbol{D}_1 \otimes \breve{\boldsymbol{K}}_{2n}, \tag{B.1}$$

$$\boldsymbol{P}_4 = \lambda_4\breve{\boldsymbol{K}}_{1n} \otimes \boldsymbol{D}_2'\boldsymbol{D}_2, \tag{B.2}$$

$$\boldsymbol{P}_5 = \lambda_5(\boldsymbol{D}_1'\boldsymbol{D}_1 \otimes \breve{\boldsymbol{K}}_{1s} + \breve{\boldsymbol{K}}_{1s} \otimes \boldsymbol{D}_2'\boldsymbol{D}_2) \tag{B.3}$$

where $\breve{\boldsymbol{K}}_{1n}$ and $\breve{\boldsymbol{K}}_{2n}$ are constraints on the null space of the coefficients vector $\boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_4$ respectively, and $\breve{\boldsymbol{K}}_{1s}$, $\breve{\boldsymbol{K}}_{2s}$ on the non-null space of the interaction vector of coefficients $\boldsymbol{\theta}_5$. In other words, $\breve{\boldsymbol{K}}_{dn}$, for $d = 1, 2$, penalizes the linear component of the linear-by-smooth varying coefficient interaction, shrinking towards the linear component of $\boldsymbol{x}_d$ (more details on penalties interpretation can be found on Appendix B).

We use the same procedure as in Appendix A to derive the equivalent penalty term for the interaction in the nested *PS*-ANOVA mixed model formulation, where we had an additional smoothing parameter λ_5 for the last part of the interaction block. Now, the penalty matrix for the interaction block becomes:

$$\boldsymbol{P}_3 = \lambda_3\boldsymbol{u}_{1n}^{(2)}\boldsymbol{u}_{1n}^{(2)\prime} \otimes \boldsymbol{D}_2'\boldsymbol{D}_2 \tag{B.4}$$

$$\boldsymbol{P}_4 = \lambda_4\boldsymbol{D}_1'\boldsymbol{D}_1 \otimes \boldsymbol{u}_{2n}^{(2)}\boldsymbol{u}_{2n}^{(2)\prime} \tag{B.5}$$

$$\boldsymbol{P}_5 = \lambda_5(\boldsymbol{U}_{1s}\boldsymbol{U}_{1s}' \otimes \boldsymbol{D}_2'\boldsymbol{D}_2 + \boldsymbol{D}_1'\boldsymbol{D}_1 \otimes \boldsymbol{U}_{2s}\boldsymbol{U}_{2s}'), \tag{B.6}$$

with rank $(c_2-1)(c_1-1)-1$, where (B.4) and (B.5) penalizes the null and non-null spaces of the matrix $\boldsymbol{\Theta}$ in both dimensions (with λ_3 and λ_4), and (B.6) penalizes the space of the non-null penalty spaces of $\boldsymbol{\Theta}$ with the smoothing parameter λ_5.

Then, the new penalty terms (B.1)–(B.3) are the penalties for the interaction of the linear-by-smooth, smooth-by-linear, and smooth-by-smooth terms on the matrix $\boldsymbol{\Theta}$, that arises from the nested *PS*-ANOVA mixed model in Section 3, with $\breve{\boldsymbol{K}}_{1n} = \boldsymbol{u}_{1n}^{(2)}\boldsymbol{u}_{1n}^{(2)\prime}$, and $\breve{\boldsymbol{K}}_{1s} = \boldsymbol{U}_{1s}\boldsymbol{U}_{1s}'$, and similar definitions for $\breve{\boldsymbol{K}}_{1n}$ and $\breve{\boldsymbol{K}}_{1s}$.

References

Akaike, H., 1973. Maximum likelihood identification of gaussian autoregressive moving average models. Biometrika 60, 255–265.

Belitz, C., Lang, S., 2008. Simultaneous selection of variables and smoothing parameters in structured additive regression models. Computational Statistics and Data Analysis 53 (1), 61–81.

Bowman, A.W., Azzalini, A., 1997. Applied Smoothing Techniques for Data Analysis: The Kernel Approach with *S*-Plus Illustrations. In: Oxford Statistical Science Series, Oxford University Press, Oxford.

Breslow, N.E., Clayton, D.G., 1993. Aproximated inference in generalised linear mixed models. Journal of the American Statistical Association 88 (421), 9–25.

Buja, A., Hastie, T., Tibshirani, R., 1989. Linear smoothers and additive models (with discussion). The Annals of Statistics 17, 453–555.

Cameron, A.C., Trivedi, P.K., 1998. Regression Analysis of Count Data. In: Econometric Society Monograph, vol. 30. Cambridge University Press.

Chen, Z., 1993. Fitting multivariate regression functions by interaction spline models. Journal of the Royal Statistical Society. Series B 55, 473–491.

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

Currie, I.D., Durbán, M., Eilers, P.H.C., 2006. Generalized linear array models with applications to multidimensional smoothing. Journal of the Royal Statistical Society. Series B 68, 1–22.

Duncan, O., 1961. A Socioeconomic Index for All Occupations. Free Press of Glencoe, 109–138.

Eilers, P.H.C., Currie, I.D., Durbán, M., 2006. Fast and compact smoothing on large multidimensional grids. Computational Statistics and Data Analysis 50 (1), 61–76.

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

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

Gu, C., 2002. Smoothing Spline ANOVA Models. In: Springer Series in Statistics, Springer.

Hastie, T., Tibshirani, R., 1990. Generalized Additive Models. In: Monographs on Statistics and Applied Probability, Chapman and Hall, London.

Hastie, T., Tibshirani, R., 1993. Varying-coefficient models. Journal of the Royal Statistical Society. Series B 55 (4), 757–796.

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

Lee, D.-J., 2010. Smothing mixed model for spatial and spatio-temporal data. Ph.D. Thesis, Department of Statistics, Universidad Carlos III de Madrid, Spain.

Lee, D.-J., Durbán, M., 2011. *P*-spline ANOVA-type interaction models for spatio-temporal smoothing. Statistical Modelling 1, 49–69 (in press).

Lee, Y., Nelder, J., Pawitan, Y., 2006. Generalized Linear Models with Random Effects: Unified Analysis Via *H*-Likelihood. In: Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Ngo, L., Wand, M.P., 2004. Smoothing with mixed model software. Journal of Statistical Software 9 (1).

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in *S* and *S*-PLUS. In: Statistics and Computing, Springer-Verlag.

Reiss, P., Ogden, T., 2009. Smoothing parameter selection for a class of semiparametric linear models. Journal of the Royal Statistical Society: Series B 71 (2), 505–523.

Ruppert, D., Wand, M.P., Carroll, R.J., 2003. Semiparametric Regression. In: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, UK, ISBN: 0521785162.

Schall, R., 1991. Estimation in generalized linear models with random effects. Biometrika 78 (4), 719–721.

Scheipl, F., Greven, S., Küchenhoff, H., 2008. Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. Computational Statistics and Data Analysis 52 (7), 3283–3299.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2), 461–464.

Searle, S., Casella, G., McCulloch, C., 1992. Variance Components. In: Wiley Series in Probability and Mathematical Statistics.

Wahba, G., 1990. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, Philadelphia.

Wood, S.N., 2006a. Generalized Additive Models—An Introduction With R. In: Texts in Statistical Science, Chapman & Hall.

Wood, S.N., 2006b. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. Biometrics 62 (4), 1025–1036.

Wood, S.N., 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. Journal of the Royal Statistical Society. Series B 73, 3–36.

Wood, S.N., Scheipl, F., Faraway, J., 2012. Straightforward intermediate rank tensor product smoothing in mixed models. Statistics and Computing, in press (http://dx.doi.org/10.1007/s11222-012-9314-z).