Article

Out-of-Sample Prediction in Multidimensional P-Spline Models

Alba Carballo ^{1,†}, María Durbán ^{1,†} and Dae-Jin Lee ^{2,*,†}

¹ Department of Statistics, Universidad Carlos III de Madrid, 28903 Getafe, Spain; alba22gc@gmail.com (A.C.); mdurban@est-econ.uc3m.es (M.D.)

² BCAM—Basque Center for Applied Mathematics, 48009 Bilbao, Spain

* Correspondence: dlee@bcamath.org

† These authors contributed equally to this work.

Abstract: The prediction of out-of-sample values is an interesting problem in any regression model. In the context of penalized smoothing using a mixed-model reparameterization, a general framework has been proposed for predicting in additive models but without interaction terms. The aim of this paper is to generalize this work, extending the methodology proposed in the multidimensional case, to models that include interaction terms, i.e., when prediction is carried out in a multidimensional setting. Our method fits the data, predicts new observations at the same time, and uses constraints to ensure a consistent fit or impose further restrictions on predictions. We have also developed this method for the so-called smooth-ANOVA model, which allows us to include interaction terms that can be decomposed into the sum of several smooth functions. We also develop this methodology for the so-called smooth-ANOVA models, which allow us to include interaction terms that can be decomposed as a sum of several smooth functions. To illustrate the method, two real data sets were used, one for predicting the mortality of the U.S. population in a logarithmic scale, and the other for predicting the aboveground biomass of Populus trees as a smooth function of height and diameter. We examine the performance of interaction and the smooth-ANOVA model through simulation studies.

Citation: Carballo, A.; Durbán, M.; Lee, D.-J. Out-of-Sample Prediction in Multidimensional P-Spline Models. *Mathematics* **2021**, *9*, 1761. <https://doi.org/10.3390/math9151761>

Academic Editor: Amir Mosavi

Received: 31 May 2021

Accepted: 15 July 2021

Published: 26 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: prediction; penalized regression; P-splines; mixed models

1. Introduction

For more than twenty years, use of penalized splines with B-splines bases (or simply P-splines) [1] has become one of the most popular smoothing techniques in many application fields, see [2] for a detailed review, and also from a theoretical perspective [3,4]. The possibility to reformulate a P-spline as a maximum likelihood estimator and best linear unbiased predictors (BLUPs) in a mixed model framework have led to a wider generalization of the method to longitudinal data, spatial smoothing, correlated errors and Bayesian approaches [5,6]. In a multidimensional setting, P-splines have also been used to model mortality life tables or spatio-temporal data [7–9]. This method is not only used to fit data, but also to obtain out-of-sample predictions. For instance, ref. [10] used P-splines to forecast mortality life tables and [11,12] also used them in a three-dimensional spatio-temporal model for predicting the risk of cancer death in the next few years. However, there are some important issues that still need to be addressed, for example, simultaneous prediction in more than one dimension, or in models that include smooth main effects and interactions terms.

One of the main contributions of this paper is an in depth study of the method proposed in [10], presenting the relationship between the regression coefficients to determine the fit and the coefficients to determine the prediction and extending the method to the mixed model framework. For the one-dimensional case, this research has been completed in [13]. In this work, we study the properties of the method proposed by [10] and show that the fit changes when the fit and the prediction are obtained simultaneously. The difference

between the fits can sometimes be small, but as we shown in the examples, the impact on the prediction can be very important.

We also show that for specific cases where only one of the two covariates is extended, the extended penalty matrix can be modified to maintain the fit. However, when the two covariates are extended, the penalty matrix cannot be modified because the matrix involved in obtaining the estimated parameters is singular.

As a general solution to maintain the fit, we propose imposing constraints on the P-splines regression coefficients. We achieve this by using the standard Lagrangian formulation of the least squares minimization problem [14]. In this article, we mainly use restrictions to ensure that the fit must be maintained, but our approach can be used for other purposes—for example, combining known information or expert opinions (or upper/lower bounds) about the value of the out-of-sample predicted observations.

In other contexts, such as spatial or spatio-temporal modelling, we might be interested in prediction at new locations, this would involve out-of-sample prediction for two covariates (latitude and longitude), and so, we extend the proposal of [10] to predict when the two covariates are extended and to predict with smooth-ANOVA models. These models have been formulated as mixed models by [8,15] and they allow us to decompose multidimensional smooth functions in terms of additive terms and interactions.

The rest of the paper is organized as follows: in Section 2, we give a brief introduction to multidimensional P-splines, present a general approach for out-of-sample prediction with and without constraints and show the properties satisfied, under certain conditions, by the coefficients that determine the prediction. We apply the methodology proposed to the analysis and forecast of US mortality data. In Section 3, we extend the results to the case of smooth-ANOVA models and, in particular, to the case in which P-splines are reformulated as mixed models. We close the section with the analysis and prediction of aboveground biomass data. Section 4 is dedicated to a simulation study to compare the methods proposed under two different scenarios. In Section 5, we present some conclusions of this work.

2. Multidimensional P-Splines

Consider a multidimensional smoothing problem of a response variable y with two covariates, x and z .

$$y = f(z, x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, R) \quad (1)$$

where ϵ is a Gaussian i.i.d error term with zero mean and variance covariance matrix $R = \sigma^2 W$, where W is a diagonal matrix when data are not correlated. The function $f(\cdot)$ is a 2-multidimensional smooth function that depends on the two explanatory variables $z = (z_1, \dots, z_{n_z})'$ and $x = (x_1, \dots, x_{n_x})'$, and each of them have lengths n_z and n_x , respectively. Although, here, we assume i.i.d. for simplicity, the results can be easily extended to a more general variance–covariance matrix R . The model in (1), can be represented in terms of basis functions with

$$f(z, x) = B\theta, \quad (2)$$

with B , a B -spline regression basis, and θ , the vector of coefficients. If we consider data in a regular array structure, a smooth multidimensional surface is constructed via the Kronecker product of the marginal B -spline basis of each covariate, the basis in model (2) is,

$$B = B_x \otimes B_z, \quad (3)$$

where $\otimes$ is the Kronecker product of two matrices, and $B_x = B(x)$ and $B_z = B(z)$, of dimensions $n_x \times c_x$ and $n_z \times c_z$, are the marginal B -spline bases for x and z , respectively. Then, the dimension of (3) is $n_x n_z \times c_x c_z$. On the other hand, if we consider scattered data,

the basis is constructed by the tensor product of the marginal B-spline basis defined in [7] as the so-called box-product, denoted by symbol $\square$:

$$B = B_x \square B_z = (B_x \otimes \mathbf{1}'_{c_z}) \odot (\mathbf{1}'_{c_x} \otimes B_z),$$

where the operator $\odot$ is the element-wise matrix product and $\mathbf{1}_{c_z}$ and $\mathbf{1}_{c_x}$ are column vectors of ones of lengths c_z and c_x .

In both cases (array structured or scattered data), the vector of coefficients θ can be arranged into a $c_z \times c_x$ matrix Θ , that is

$$\Theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1c_x} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2c_x} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{c_z1} & \theta_{c_z2} & \cdots & \theta_{c_zc_x} \end{bmatrix}, \quad (4)$$

and so the two-dimensional P-spline model can be written as

$$f(z, x) = (B_x \otimes B_z)\theta = \text{vec}(B_z \Theta B'_x), \quad (5)$$

where $\text{vec}(\cdot)$ denotes the vectorization operator.

In the two-dimensional case, a matrix P penalizes the difference between adjacent coefficients of the rows and columns of the coefficient matrix Θ (4). Therefore, the penalty matrix P has the following expression:

$$P = \underbrace{\lambda_z I_{c_x} \otimes D'_z D_z}_{P^z} + \underbrace{\lambda_x D'_x D_x \otimes I_{c_z}}_{P^x}, \quad (6)$$

where D_z and D_x are the difference matrices acting on the rows and columns of Θ , respectively. λ_z and λ_x are the smoothing parameters for each dimension of the model. Since λ_z and λ_x are not necessary equal, penalty (6) allows for anisotropic smoothing. To estimate the regression coefficients, we minimize the following function for the penalized sum of squares:

$$S(\theta) = (y - B\theta)'(y - B\theta) + \theta' P \theta. \quad (7)$$

Therefore, for given values of the smoothing parameters λ_z and λ_x , the minimizer of the penalized sum of squares (7) is:

$$\hat{\theta} = (B'B + P)^{-1} B'y. \quad (8)$$

Optimal λ_z and λ_x can be estimated using a information criteria (such as Akaike or Bayesian criteria) or via cross-validation (see [1,7] for details).

Once we have given this brief introduction to multidimensional P-splines, we generalize the approach given in [10] to obtain out-of-sample predictions not only in one, $f(x_+, z)$, but in both independent variables, $f(x_+, z_+)$ (the subscript '+' denotes the dimension for which we are interested in for out-of-sample prediction). For the particular case in which a single covariate is extended, we will state several properties of the prediction method. Since the natural extension of the penalty matrix provides a change in the fit, in order to overcome this problem, we will suggest using constraints on the model coefficients.

2.1. Out-of-Sample Prediction of a Single Covariate

For the multidimensional P-spline model (1), we will consider first the case where the interest is on predicting outside the range of one of the covariates. This was already discussed in [10] for mortality life tables, where death counts were collected at different ages and years, and the aim was to forecast future mortality across ages.

Given a vector of $n_z \times n_x$ observations y of the response variable, suppose we want to predict $n_p = n_z \times n_{x_p}$ new values y_p at z and x_p , where subscript p denotes the 'pre-

dicted' range of the $\mathbf{x}_p = (x_1, x_2, \dots, x_n, x_{n+1}, \dots, x_{n+p})'$ covariate and the response vector is $\mathbf{y}_+ = (\mathbf{y}, \mathbf{y}_p)'$, where $\mathbf{y}$ is the vector of observed values and $\mathbf{y}_p$ is the new unknown observations to be predicted that can take arbitrary values—if we arrange the vector of observations into a matrix $\mathbf{Y}$ of dimension $n_z \times n_x$, such that $\text{vec}(\mathbf{Y}) = \mathbf{y}$ of length $n_z n_x \times 1$ the new matrix of observations can be written as:

$$\mathbf{Y}_+ = [\mathbf{Y} \mid \mathbf{Y}_{zx_p}], \quad (9)$$

where $\mathbf{Y}_{zx_p}$ is a matrix of dimensions $n_z \times (n_x + n_{x_p})$. Hence, matrix $\mathbf{Y}_+$ is, in this case, of dimensions $n_z \times n_{x_p}$. In order to fit and predict $\mathbf{y}_+$ we consider the extended model:

$$\mathbf{y}_+ = \mathbf{B}_+ \boldsymbol{\theta}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{R}_+), \quad (10)$$

where $\mathbf{R}_+$ is the extended error covariance-matrix defined as $\mathbf{R}_+ = \sigma^2 \mathbf{W}_+$, with $\mathbf{W}_+ = \mathbf{W}_{x_+} \otimes \mathbf{W}_z$, and $\mathbf{W}_{x_+}$ is a diagonal matrix with 1's for the first n_x entries (when the data are observed) and infinite values for the last n_{x_p} (when the data are missing). For the z covariate, as we are not predicting new observation, $\mathbf{W}_z$ of a diagonal matrix of one's of dimension $n_z \times n_z$.

Following [10], the extended basis is defined as:

$$\mathbf{B}_+ = \mathbf{B}_{x_+} \otimes \mathbf{B}_z = \begin{bmatrix} \mathbf{B}_x & \mathbf{O} \\ \mathbf{B}_{x_{(1)}} & \mathbf{B}_{x_{(2)}} \end{bmatrix} \otimes \mathbf{B}_z = \begin{bmatrix} \mathbf{B}_x \otimes \mathbf{B}_z & \mathbf{O} \\ \mathbf{B}_{x_{(1)}} \otimes \mathbf{B}_z & \mathbf{B}_{x_{(2)}} \otimes \mathbf{B}_z \end{bmatrix}, \quad (11)$$

where $\mathbf{B}_{x_+} = \mathbf{B}(x_+)$ and $\mathbf{B}_z = \mathbf{B}(z)$ are the regression bases with x_+ and z , the two regressors. The new extended marginal B-spline basis for the x -covariate, $\mathbf{B}_{x_+}$, is constructed from a new set of knots that consists of the original knots covering x_i , $i = 1, \dots, n_x$, and extended to the new range of the n_{x_p} values of x_{p_j} , $j = 1, \dots, n_{x_p}$. In other words, $\mathbf{B}_{x_+}$ is a direct extension of $\mathbf{B}_x$. To estimate the new regression coefficients $\boldsymbol{\theta}_+$, we minimize the extended penalized sum of squares:

$$\mathcal{L}(\boldsymbol{\theta}_+) = (\mathbf{y}_+ - \mathbf{B}_+ \boldsymbol{\theta}_+)' \mathbf{W}_+^{-1} (\mathbf{y}_+ - \mathbf{B}_+ \boldsymbol{\theta}_+) + \boldsymbol{\theta}_+' \mathbf{P}_+ \boldsymbol{\theta}_+, \quad (12)$$

where the extended penalty matrix is given by:

$$\mathbf{P}_+ = \lambda_z \mathbf{P}_+^z + \lambda_x \mathbf{P}_+^{x_+}. \quad (13)$$

since only the covariate x is extended, $\mathbf{P}_+^z$ and $\mathbf{P}_+^{x_+}$ are:

$$\begin{aligned} \mathbf{P}_+^z &= \mathbf{I}_{c_{x_+}} \otimes \mathbf{D}_z' \mathbf{D}_z = \begin{bmatrix} \mathbf{I}_{c_x} & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{c_{x_p}} \end{bmatrix} \otimes \mathbf{D}_z' \mathbf{D}_z = \begin{bmatrix} \mathbf{I}_{c_x} \otimes \mathbf{D}_z' \mathbf{D}_z & \mathbf{O} \\ \mathbf{O} & \mathbf{I}_{c_{x_p}} \otimes \mathbf{D}_z' \mathbf{D}_z \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}_{+11}^z & \mathbf{O} \\ \mathbf{O} & \mathbf{P}_{+22}^z \end{bmatrix}, \end{aligned} \quad (14)$$

and

$$\begin{aligned} \mathbf{P}_+^{x_+} &= \mathbf{D}_{x_+}' \mathbf{D}_{x_+} \otimes \mathbf{I}_{c_z} = \begin{bmatrix} (\mathbf{D}_x' \mathbf{D}_x + \mathbf{D}_{x(1)}' \mathbf{D}_{x(1)}) \otimes \mathbf{I}_{c_z} & \mathbf{D}_{x(1)}' \mathbf{D}_{x(2)} \otimes \mathbf{I}_{c_z} \\ \mathbf{D}_{x(2)}' \mathbf{D}_{x(1)} \otimes \mathbf{I}_{c_z} & \mathbf{D}_{x(2)}' \mathbf{D}_{x(2)} \otimes \mathbf{I}_{c_z} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}_{+11}^{x_+} & \mathbf{P}_{+12}^{x_+} \\ \mathbf{P}_{+21}^{x_+} & \mathbf{P}_{+22}^{x_+} \end{bmatrix}, \end{aligned} \quad (15)$$

where D_{x+} and D_z are the difference matrices acting on the columns and rows of the matrix formed by the extended vector of coefficients:

$$\Theta_+ = \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1\ c_x-1} & \theta_{1c_x} & \theta_{1\ c_x+1} & \theta_{1\ c_x+2} & \cdots \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2\ c_x-1} & \theta_{2c_x} & \theta_{2\ c_x+1} & \theta_{2\ c_x+2} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_{c_z1} & \theta_{c_z2} & \cdots & \theta_{c_z\ c_x-1} & \theta_{c_zc_x} & \theta_{c_z\ c_x+1} & \theta_{c_z\ c_x+2} & \cdots \end{bmatrix}. \quad (16)$$

The minimizer of (12) is given by:

$$\theta_+ = (B_+'W_+^{-1}B_+ + P_+)^{-1}B_+'W_+^{-1}y_+. \quad (17)$$

The fact that we extend just one of the two covariates, and the particular form of the penalties based on differences between adjacent coefficients, yields certain important properties satisfied by the model. These properties are the direct result of the following theorem:

Theorem 1. *The coefficients obtained from minimization of (12) with the extended basis (11), extended error covariance matrix (10) and extended penalty matrix (13) meet the following properties:*

I. *The first c , $c = c_z \times c_x$, coefficients of $\hat{\theta}_+$, are:*

$$\hat{\theta}_{+1,\dots,c} = \left(B'B + \lambda_x P_{+11}^{x+} + \lambda_z P_{+11}^z - \lambda_x^2 P_{+12}^{x+} (\lambda_x P_{+22}^{x+} + \lambda_z P_{+22}^z)^{-1} P_{+21}^{x+} \right)^{-1} B'y, \quad (18)$$

where P_{+11}^z and P_{+22}^z are defined in (14) and P_{+11}^{x+} , P_{+12}^{x+} , P_{+21}^{x+} and P_{+22}^{x+} defined in (15).

II. *The coefficients for the $n_p = n_z \times n_{x_p}$ predicted values are*

$$\hat{\theta}_p = - \left(\frac{\lambda_z}{\lambda_x} P_{+22}^z + P_{+22}^{x+} \right)^{-1} P_{+21}^{x+} \hat{\theta}_{+1,\dots,c}, \quad (19)$$

where P_{+22}^z is defined in (14) and P_{+22}^{x+} , P_{+21}^{x+} is defined in (15).

The proof of Theorem 1 is given in the supplementary materials.

Therefore, according to the previous theorem, when we predict in two dimensions extending one covariate, the predicted values, y_p , obtained by using the new coefficients, θ_p ; depend on the ratio λ_z/λ_x , unless $\lambda_x = \lambda_z$ (i.e., an isotropic smoothing). Therefore, while in the case of one-dimensional smooth models, the smoothing parameter does not play any role in the prediction (see [13] for details), we have found that in two dimensions the smoothing parameters in both directions, λ_x and λ_z , determine the prediction.

Notice that we have proved that the coefficients that give the fit when the fit and the prediction are obtained at the same time (18) are different from the solution we obtain only fitting the data (8) (this property is verified in the case of one-dimensional smooth models [13]). Although, in the one-dimensional case, the extended penalty is not a direct extension of the penalty used to fit the data, the blocks of the extended penalty are simplified and the fit remains the same. This does not happen in the case of two dimensions, unless the block P_{+22}^z in (14) is equal to zero, as shown by the following corollary.

Corollary 1 (Theorem 1). *If $P_{+22}^z = O$ (or equivalently, if $I_{c_{x_p}} = O$) in (14), the solution of (12) verifies the following properties:*

1. *When performing out-of-sample predictions, the fit remains invariant.*
2. *Consider the coefficient matrix obtained from the model fit, $\hat{\Theta}$, and the c matrix of predicted coefficients, $\hat{\Theta}_p$, each row $j = 1, \dots, c_z$, of the additional matrix of coefficients is a linear combination of the last q_z old coefficients of that row (q_z is the order of the penalty acting on rows and c_z is the number of rows of B_z).*

For the particular case of second and third order penalties, i.e., $q_x = q_z = 2$ and $q_x = q_z = 3$, the proof is given in the supplementary materials.

As we have showed in the previous corollary, setting $I_{c_{x_p}}$ to zero, the fit is preserved and everything is analogous to the one-dimensional case. In the literature, there are some works in which $I_{c_{x_p}}$ is considered equal to zero, e.g., [12]. However, in practice, we do not set $I_{c_{x_p}}$ to zero since we would not be imposing the penalty correctly. In addition, we cannot extend it to situations where we want to make out-of-sample predictions in two dimensions. We cannot set $I_{c_{z_p}}$ and $I_{c_{x_p}}$ to zero in (14) and (15), since the matrix $B'_+ W_+^{-1} B_+ + P_+$ would be singular.

It is important to note that, regardless of the value of P_{+22}^{x+} , by (19) the new coefficients are determined by the last q_z columns of the coefficients that give the fit. Therefore, we do not need to set P_{+22}^{x+} equal to zero to know which coefficients determine the prediction.

2.2. Out-of-Sample Prediction in Two Dimensions

In this section, we address the case in which we are interested in making predictions outside the range of the observed values of both covariates x and z .

Given a vector of $n_z n_x$ observations y of the response variable, now we want to predict $n_p = n_z n_{x_p} + n_{z_p} n_x + n_{z_p} n_{x_p}$ new values at (z, x_p) , (z_p, x) and (z_p, x_p) , i.e., if we arrange the observation vector into a matrix Y of dimension $n_z \times n_x$, the observed and predicted values can be arranged into a matrix of dimension $n_{z+} \times n_{x+}$ ($n_{z+} = n_z + n_{z_p}$, $n_{x+} = n_x + n_{x_p}$), as:

$$Y_+ = \begin{bmatrix} Y & Y_{zx_p} \\ Y_{z_p x} & Y_{z_p x_p} \end{bmatrix}. \quad (20)$$

Note that the dimensions of Y_{zx_p} , $Y_{z_p x}$ and $Y_{z_p x_p}$ are $n_z \times n_{x_p}$, $n_{z_p} \times n_x$ and $n_{z_p} \times n_{x_p}$, respectively.

We again use the model in (10), but now the full extended basis is the Kronecker product of the two extended marginal B-spline basis, $B_{x+} = B(x_+)$ and $B_{z+} = B(z_+)$, of dimensions $n_{x+} \times c_{x+}$ and $n_{z+} \times c_{z+}$, respectively:

$$B_+ = B_{x+} \otimes B_{z+} = \begin{bmatrix} B_x & O \\ B_{x(1)} & B_{x(2)} \end{bmatrix} \otimes \begin{bmatrix} B_z & O \\ B_{z(1)} & B_{z(2)} \end{bmatrix}, \quad (21)$$

where the extended bases B_{x+} and B_{z+} are constructed from a new set of knots that consists of the original knots and extended to cover the full range of x_+ and z_+ , respectively. The extended covariance matrix $R_+ = \sigma_\epsilon^2 W_+$ with $W_+ = W_{x+} \otimes W_{z+}$, where W_{x+} and W_{z+} are diagonal matrices of dimensions $n_{x+} \times n_{x+}$ ($n_{x+} = n_x + n_{x_p}$) and $n_{z+} \times n_{z+}$ ($n_{z+} = n_z + n_{z_p}$), respectively, with infinity entry if the data are missing and 1 if the data are observed. The infinity quantities express the uncertainty since we do not have any information about the data y_p . To estimate the extended coefficients, we minimize the penalized sum of squares in (12) with an extended penalty matrix

$$P_+ = \lambda_z P_+^{z+} + \lambda_x P_+^{x+}, \quad (22)$$

with

$$\begin{aligned} P_+^{z+} &= I_{c_{x+}} \otimes D'_{z+} D_{z+} = \begin{bmatrix} I_{c_x} \otimes D'_{z+} D_{z+} & O \\ O & I_{c_{x_p}} \otimes D'_{z+} D_{z+} \end{bmatrix} \\ &= \begin{bmatrix} P_{+11}^{z+} & O \\ O & P_{+22}^{z+} \end{bmatrix}, \end{aligned} \quad (23)$$

and P_+^{x+} as in (15). Note that D_{z+} and D_{x+} are direct extensions of D_z and D_x but P_+^{z+} and P_+^{x+} are not direct extensions of P^z and P^x . Moreover, if θ is given in (8) and we are extending the two covariates, $\theta_+ = \text{vec}(\Theta_+)$, in matrix form Θ_+ is:

$$\Theta_+ = \left[\begin{array}{ccccc|ccccc} \theta_{11} & \theta_{12} & \cdots & \theta_{1\ c_x-1} & \theta_{1c_x} & \theta_{1\ c_x+1} & \theta_{1\ c_x+2} & \cdots & \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2\ c_x-1} & \theta_{2c_x} & \theta_{2\ c_x+1} & \theta_{2\ c_x+2} & \cdots & \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ \theta_{c_z1} & \theta_{c_z2} & \cdots & \theta_{c_z\ c_x-1} & \theta_{c_zc_x} & \theta_{c_z\ c_x+1} & \theta_{c_z\ c_x+2} & \cdots & \\ \hline \theta_{c_z+1\ 1} & \theta_{c_z+1\ 2} & \cdots & \theta_{c_z+1\ c_x-1} & \theta_{c_z+1\ c_x} & \theta_{c_z+1\ c_x+1} & \theta_{c_z+1\ c_x+2} & \cdots & \\ \theta_{c_z+2\ 1} & \theta_{c_z+2\ 2} & \cdots & \theta_{c_z+2\ c_x-1} & \theta_{c_z+2\ c_x} & \theta_{c_z+2\ c_x+1} & \theta_{c_z+2\ c_x+2} & \cdots & \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \cdots & \end{array} \right].$$

Note that in the case of fitting, the extension of the predicted new values depends on the data structure. If we consider scattered data, we set n_p out-of-sample points (z_{+i}, x_{+i}) at which we want to predict new y_{p_i} values for $i = 1, \dots, n_p$ and R_+ is a diagonal matrix with the first n values equal to 1 and the last n_p values equal to infinity. Everything else is independent of the data structure.

As in the case of only one-dimensional out-of-sample prediction, when using model (1) or model (10), the fit to the observed data is not the same, as one might expect. In the next section, we recommend using constraints to maintain the fit when fitting and prediction are obtained at the same time.

2.3. Out-of-Sample Prediction with Constraints

As we showed in the previous section, the natural expansion of the penalty matrix provides a change in fit. To overcome this problem, and as a possible way to incorporate known information about the prediction, we recommend the use of constrained P-splines. This method will allow us to impose constant and fixed constraints, as well as constraints that depend on observed data.

Our proposal is to derive the solution of the extended model (10) subject to a set of l linear constraints given by $C\theta_+ = r$, where C is a known matrix of dimension $l \times c_+$ acting on all coefficients, and r is the restrictions vector of dimension $l \times 1$. Therefore, we have the following extended regression model with constraints:

$$y_+ = B_+\theta_+^* + \epsilon_+, \quad \epsilon_+ \sim \mathcal{N}(0, R_+) \quad (24)$$

subject to

$$C\theta_+^* = r, \quad (25)$$

Depending on whether we are predicting out-of-sample in one or two dimensions, we define y_+ , B_+ and R_+ as shown in Sections 2.1 and 2.2. The superscript (*) refers to the use of constraints.

We extend the results of [14] to the case of penalized least squares and obtain the Lagrange formulation of the penalized least squares minimization problem:

$$\mathcal{L}(\theta_+^*, \omega) = (y_+ - B_+\theta_+^*)'W_+^{-1}(y_+ - B_+\theta_+^*) + \theta_+^{*'}P_+\theta_+^* + 2\omega'(C\theta_+^* - r), \quad (26)$$

where W_+^{-1} is defined depending on whether we extend one or two covariates, P_+ is the extended penalty matrix (13) or (22), θ_+^* denotes the restricted penalized least squares (RPLS) estimator and ω is a $l \times 1$ vector of Lagrange multipliers.

Differentiating the expression in (26), we obtain:

$$\frac{\partial \mathcal{L}}{\partial \theta_+^*} = -2B_+'W_+^{-1}(y_+ - B_+\theta_+^*) + 2P_+\theta_+^* + 2C'\omega, \quad (27)$$

$$\frac{\partial \mathcal{L}}{\partial \omega} = C\theta_+^* - r. \quad (28)$$

The system of equations can be written as:

$$\begin{bmatrix} B'_+ W_+^{-1} B_+ + P_+ & C' \\ C & O \end{bmatrix} \begin{bmatrix} \hat{\theta}_+^* \\ \hat{\omega} \end{bmatrix} = \begin{bmatrix} B'_+ W_+^{-1} y_+ \\ r \end{bmatrix}.$$

yielding the following solutions:

$$\hat{\theta}_+^* = \hat{\theta}_+ - (B'_+ W_+^{-1} B_+ + P_+)^{-1} C' \hat{\omega}, \quad (29)$$

$$\hat{\omega} = [C(B'_+ W_+^{-1} B_+ + P_+)^{-1} C']^{-1} (C \hat{\theta}_+ - r). \quad (30)$$

where $\hat{\theta}_+$ is the unconstrained penalized least squares estimator as in (17).

Therefore, the constrained vector of coefficients $\hat{\theta}_+^*$, are obtained by computing the vector of Lagrange multipliers (30) and substituting in (29), i.e., $\hat{\theta}_+^*$ is the unconstrained solution, $\hat{\theta}_+$, plus a multiple of the discrepancy vector.

Using these results, we can obtain the constrained fitted and predicted values as

$$\hat{y}_+^* = B_+ \hat{\theta}_+^*,$$

defining the matrices

$$\begin{aligned} A_1 &= (B'_+ W_+^{-1} B_+ + P_+)^{-1} B'_+ \\ A_2 &= (B'_+ W_+^{-1} B_+ + P_+)^{-1} C' [C(B'_+ W_+^{-1} B_+ + P_+)^{-1} C']^{-1}, \end{aligned}$$

the response vector of new values $\hat{y}_+^*$ can be written as:

$$\hat{y}_+^* = B_+ (A_1 W_+^{-1} y_+ - A_2 C A_1 W_+^{-1} y_+ + A_2 r).$$

The variance of y_+^* depends on the following conditions:

- (a) If the constraint r is a constant, i.e., it does not depend on the data, the variance is:

$$\text{Var}[\hat{y}_+^*] = \sigma_\epsilon^2 B_+ A_3 W_+^{-1} A_3' B_+',$$

with $A_3 = A_1 - A_2 C A_1$.

- (b) If the constraint depends on the data, we must consider the variability of r . For instance, if the constraint implies that the fit has to be maintained, i.e., $r = \hat{\theta}$, the variance is:

$$\text{Var}[\hat{y}_+^*] = \sigma_\epsilon^2 B_+ A_4 W_+^{-1} A_4' B_+',$$

with

$$A_4 = A_1 - A_2 C A_1 + A_2 \begin{bmatrix} (B' B + P)^{-1} \\ O \end{bmatrix},$$

where O is a null matrix of dimension, $c_p \times n_p$, $c_p = c_z c_{x_p} + c_{z_p} c_x + c_{z_p} c_{x_p}$ is the number of new coefficients, and n_p is the number of new observations.

To help readers understand the process, we now explain how to impose constraints on the fit in practice. Suppose we just make out-of-sample predictions in one of the two covariates and, for example, the estimated coefficients matrix obtained when only fitting the observed data has dimensions of 4×3 , and the one obtained when fitting and forecasting simultaneously has dimension 4×5 , then

$$\hat{\Theta} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_5 & \hat{\theta}_9 \\ \hat{\theta}_2 & \hat{\theta}_6 & \hat{\theta}_{10} \\ \hat{\theta}_3 & \hat{\theta}_7 & \hat{\theta}_{11} \\ \hat{\theta}_4 & \hat{\theta}_8 & \hat{\theta}_{12} \end{bmatrix}, \quad \Theta_+^* = \begin{bmatrix} \bullet & \bullet & \bullet & \circ & \circ \\ \theta_1 & \theta_5 & \theta_9 & \theta_{13} & \theta_{17} \\ \bullet & \bullet & \bullet & \circ & \circ \\ \theta_2 & \theta_6 & \theta_{10} & \theta_{14} & \theta_{18} \\ \bullet & \bullet & \bullet & \circ & \circ \\ \theta_3 & \theta_7 & \theta_{11} & \theta_{15} & \theta_{19} \\ \bullet & \bullet & \bullet & \circ & \circ \\ \theta_4 & \theta_8 & \theta_{12} & \theta_{16} & \theta_{20} \end{bmatrix},$$

where in Θ_+^* the coefficients that determine the fit have superscript $\bullet$ and the coefficients that determine the forecast have superscript symbol $\circ$. If we impose the constraint that the fit has to be maintained, we use the constraint equation in (25) with $\theta_+^* = \text{Vec}(\Theta_+^*)$, $C = [I_{12 \times 12} \mid O_{12 \times 8}]$ ($I_{12 \times 12}$ is an identity matrix of dimension 12 and $O_{12 \times 8}$ a zero matrix of dimension 12×8) and $r = \hat{\theta} = \text{vec}(\hat{\Theta})$.

On the other hand, if we extend the two covariates, the coefficient matrices obtained when only fitting and when fit and prediction is carried out simultaneously are:

$$\hat{\Theta} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_5 & \hat{\theta}_9 \\ \hat{\theta}_2 & \hat{\theta}_6 & \hat{\theta}_{10} \\ \hat{\theta}_3 & \hat{\theta}_7 & \hat{\theta}_{11} \\ \hat{\theta}_4 & \hat{\theta}_8 & \hat{\theta}_{12} \end{bmatrix}, \quad \Theta_+^* = \begin{bmatrix} \bullet & \bullet & \bullet & \circ & \circ \\ \theta_1 & \theta_7 & \theta_{13} & \theta_{19} & \theta_{25} \\ \bullet & \bullet & \bullet & \circ & \circ \\ \theta_2 & \theta_8 & \theta_{14} & \theta_{20} & \theta_{26} \\ \bullet & \bullet & \bullet & \circ & \circ \\ \theta_3 & \theta_9 & \theta_{15} & \theta_{21} & \theta_{27} \\ \bullet & \bullet & \bullet & \circ & \circ \\ \theta_4 & \theta_{10} & \theta_{16} & \theta_{22} & \theta_{28} \\ \circ & \circ & \circ & \circ & \circ \\ \theta_5 & \theta_{11} & \theta_{17} & \theta_{23} & \theta_{29} \\ \circ & \circ & \circ & \circ & \circ \\ \theta_6 & \theta_{12} & \theta_{18} & \theta_{24} & \theta_{30} \end{bmatrix},$$

i.e., $c_z = 4$, $c_{z_p} = 2$, $c_x = 3$ and $c_{x_p} = 2$. To impose the restriction the fit has to be maintained, we set $\theta_+^* = \text{Vec}(\Theta_+^*)$, $C = \text{blockdiag}(I_{4 \times 4}, [O_{2 \times 2} \mid I_{4 \times 4}], [O_{2 \times 2} \mid I_{4 \times 4}])$ ($I_{4 \times 4}$ is an identity matrix of dimension 4 and $O_{2 \times 2}$ is a zero matrix of dimension 2) and $r = \hat{\theta} = \text{vec}(\hat{\Theta})$ in Equation (25).

In general, if the constraint implies that the fit has to be maintained and $c_{x_p} \geq c_{z_p}$, C is a block diagonal matrix with the first block an identity matrix of dimensions $c_z \times c_z$ and c_{x_p} blocks equal to $[O_{c_{z_p} \times c_{z_p}} \mid I_{c_z \times c_z}]$, i.e.,

$$C = \begin{bmatrix} I_{c_z \times c_z} & & & \\ & [O_{c_{z_p} \times c_{z_p}} \mid I_{c_z \times c_z}] & & \\ & & [O_{c_{z_p} \times c_{z_p}} \mid I_{c_z \times c_z}] & \\ & & & [O_{c_{z_p} \times c_{z_p}} \mid I_{c_z \times c_z}] \\ & & & & \ddots \end{bmatrix},$$

and $r = \text{vec}(\hat{\Theta})$.

2.4. Example 1: Prediction of Mortality Data

We illustrate our constrained out-of-sample prediction method with the analysis of male mortality data in the US. Although mortality data are usually analyzed by Poisson distribution, for the simple purpose of illustrating the proposed method, we will assume that the logarithm of the mortality rates are normally distributed. The data come from the Human Mortality Database (Link: <https://www.mortality.org/> (last accessed date: 1 July 2021)). We observed the log of mortality rates from ages 0 to 110+ over the period 1960–2014, and forecasted up to 2050, i.e., we obtained out-of-sample predictions in one of the two covariates, in this case, year.

The Lee–Carter method [16] is one of the most commonly used methods for estimating and forecasting mortality data; however, this method may produce an unwanted cross-over of forecasted mortality (higher mortality rates for younger ages than for older ages). The Lee–Carter model is defined as:

$$\log(m_{x,y}) = \alpha_x + \beta_x \kappa_y + \epsilon_y,$$

where $m_{x,y}$ is the central rate of mortality at age x and year y , α_x , β_x and κ_y are parameters to be estimated, and ϵ_y is the error term with mean zero and variance σ^2 . This model is fitted to historical data and the resulting estimated κ_y s are then modelled and projected as a stochastic time series using standard Box–Jenkins methods. Ref. [17] proposed a modification to the Lee–Carter model based on the estimation of smooth versions of β_x to

avoid the cross-over of projected mortality rates. In order to compare our method with the solution given by [17], we obtained fits and predictions through four different models:

(a) Model 1: an unconstrained model:

$$y_+ = f(\text{age}, \text{year}_+) + \epsilon_+, \quad \epsilon_+ \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_+)$$

where y_+ is the log mortality rate and $\mathbf{R}_+$ is defined as in (10).

(b) Model 2: The model defined in (a) subject to the constraint that the fit is maintained when fitting and forecasting at the same time.

(c) Model 3: The model defined in (a) is subject to two constraints:

- The fit is maintained;
- The inter-age structure is preserved. We impose this constraint to avoid age crossing, that is, to prevent the death rate of younger ages from being higher than that of older ages. To achieve this, we use the pattern of the estimated coefficients in the past few years and forecast it. In particular, we ensure that the difference between the coefficients of every two consecutive projections is constant and equal to the difference between the corresponding last coefficient in the fitting. Let us use an example to explain how to impose these two constraints at the same time. Suppose that the dimension of the coefficient matrix from the fitting is 4×3 , and the dimensions of the coefficient matrix given the fitting and prediction is 4×5 , that is,

$$\hat{\Theta} = \begin{bmatrix} \hat{\theta}_1 & \hat{\theta}_5 & \hat{\theta}_9 \\ \hat{\theta}_2 & \hat{\theta}_6 & \hat{\theta}_{10} \\ \hat{\theta}_3 & \hat{\theta}_7 & \hat{\theta}_{11} \\ \hat{\theta}_4 & \hat{\theta}_8 & \hat{\theta}_{12} \end{bmatrix}, \quad \Theta_+^* = \begin{bmatrix} \bullet \hat{\theta}_1 & \bullet \hat{\theta}_5 & \bullet \hat{\theta}_9 & \circ \hat{\theta}_{13} & \circ \hat{\theta}_{17} \\ \bullet \hat{\theta}_2 & \bullet \hat{\theta}_6 & \bullet \hat{\theta}_{10} & \circ \hat{\theta}_{14} & \circ \hat{\theta}_{18} \\ \bullet \hat{\theta}_3 & \bullet \hat{\theta}_7 & \bullet \hat{\theta}_{11} & \circ \hat{\theta}_{15} & \circ \hat{\theta}_{19} \\ \bullet \hat{\theta}_4 & \bullet \hat{\theta}_8 & \bullet \hat{\theta}_{12} & \circ \hat{\theta}_{16} & \circ \hat{\theta}_{20} \end{bmatrix},$$

in Θ_+^* , the coefficients that determine the fit have the superscript $\bullet$ and the coefficients that determine the forecast have the superscript $\circ$. Now, in the restriction Equation (25), $\theta_+^* = \text{Vec}(\Theta_+^*)$, and $\mathbf{C} = \text{blockdiag}(\mathbf{I}_{12 \times 12}, \mathbf{U}, \mathbf{U})$ with

$$\mathbf{U} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{r} = \begin{bmatrix} \hat{\theta} \\ \hat{\theta}_9 - \hat{\theta}_{10} \\ \hat{\theta}_{10} - \hat{\theta}_{11} \\ \hat{\theta}_{11} - \hat{\theta}_{12} \\ \hat{\theta}_9 - \hat{\theta}_{10} \\ \hat{\theta}_{10} - \hat{\theta}_{11} \\ \hat{\theta}_{11} - \hat{\theta}_{12} \end{bmatrix}$$

and $\hat{\theta} = \text{Vec}(\hat{\Theta})$, i.e., we are imposing that the coefficients that determine the fit (when fitting and forecasting simultaneously) have to be the ones we obtain when only fitting the data, and that the difference between the coefficients that determine the forecast of two consecutive rows has to be equal to the difference between the last coefficients from the fit of those rows.

(d) Model 4: is the Lee–Carter model proposed in [16], but modified to get regular projected life tables [17].

Figure 1 shows the fit and the forecast obtained with all four models. In order to highlight the difference between the results, we have selected four ages: 20, 40, 60, and 80. Figure 2 shows the fit and forecast for these ages obtained by model 1 (red line), model 2 (green line), model 3 (blue line) and model 4 (orange line). The first three models (models 1, 2 and 3) provide almost the same fit. However, the fit given by model 4 is quite different and worse than the fits of other models, at 20 and 40 years. This is to be expected because

the Lee–Carter model is not flexible enough to capture the increase in mortality in the early 1980s.

Figure 1. Fit and forecast of a data set on the log mortality rates of US males aged 0–110+ over the period 1960–2014, through model 1 (top left), model 2 (top right), model 3 (bottom left) and model 4 (bottom right). The horizontal line indicates the year from which we predict (2014).

Figure 2. Fit and forecast of selected ages: 20, 40, 60 and 80 obtained using model 1 (red line), model 2 (green line), model 3 (blue line) and model 4 (orange line). The vertical line indicates the year from which we predict (2014).

The predictions with models 1 and 2 are almost identical. Although it gave very similar results in the fitting, model 3 provided completely different results in the predictions, as expected. For 60 years of age, model 1 and model 2 predict that the logarithmic death rate will increase during 2020–2050, because they predict an upward trend in death rates

from 2010 to 2014. This is inconsistent with the expected result. In the case of model 4, forecasts are close to model 3 for ages 40 and 80, but in the case of age 20 it seems to clearly overestimate the mortality. In model 3, the increasing trend is corrected after the initial years to maintain the structure of the adjacent ages, thus avoiding irregular predictions (as happens at age 60).

As we have seen, when we impose the two constraints, we obtain the most realistic results: maintain the fit and the cross-age structure, that is, those given by model 3. If we do not maintain the coefficient pattern, age crossover may occur. We further illustrate this fact in Figure 3, we plot the projections obtained with model 2 (green line) and model 3 (blue line) for ages 66 and 67. We can see that the fit of the two models is the same, and within the known data range, the log mortality of age 66 is lower than that of age 67. However, in the forecast, model 2 does not avoid the cross-over for ages 66 and 67 and so, in 2050, the log mortality rate is larger for age 66 than for age 67, this cross-over is not biologically acceptable since it is expected that mortality rates of older people are higher. This does not occur with model 3 where the restrictions imposed preserve the cross-age structure among ages.

Figure 3. Fit and forecast for ages 66 and 67 obtained through model 2 (green line) and model 3 (blue line). Model 3 prevent cross-over for ages. The vertical line indicates the year from which we predict (2014).

3. Component-Wise Prediction with P-Spline Smooth-ANOVA Models

In some cases, the multidimensional model in (1) is not flexible enough, and it may result in unnecessary complex models (that is, use too many degrees of freedom to fit the data). In order to add more flexibility and avoid unnecessary and irrelevant terms, the authors of [8] propose the use of P-spline smooth-ANOVA models (or S-ANOVA). This method decomposes the sum of smooth functions in a similar way to the analysis of variance decomposition, i.e., main and interaction effects. In this section, we will show how to obtain out-of-sample predictions in this setting. First, we briefly review how to fit the S-ANOVA model and show how the P-spline mixed model representation can avoid the identifiability problem that occurs when multiple smoothing terms are used in the same model.

Suppose we have data laid out as a rectangular array, and $\mathbf{y}$ is the response vector of length $n \times 1$, where $n = n_z n_x$, and $\mathbf{z} = (z_1, z_2, \dots, z_{n_z})'$ and $\mathbf{x} = (x_1, x_2, \dots, x_{n_x})'$ are the covariates. Let us consider the following S-ANOVA model:

$$\mathbf{y} = \gamma + f_1(\mathbf{z}) + f_2(\mathbf{x}) + f_{1,2}(\mathbf{z}, \mathbf{x}) + \epsilon = \mathbf{B}\boldsymbol{\theta} + \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{R}) \quad (31)$$

where $\mathbf{R} = \sigma_e^2 \mathbf{I}$, i.e., the errors are independent and identically distributed and the B-spline basis $\mathbf{B}$ is defined as:

$$\mathbf{B} = [\mathbf{1}_n \mid \mathbf{1}_{n_x} \otimes \mathbf{B}_z \mid \mathbf{B}_x \otimes \mathbf{1}_{n_z} \mid \mathbf{B}_x \otimes \mathbf{B}_z], \quad (32)$$

of dimension $n \times (1 + c_z + c_x + c_z c_x)$, and where $\mathbf{1}_{n_z}$ and $\mathbf{1}_{n_x}$ are column vectors of ones of length n_z and n_x respectively, and the vector of regression coefficients are $\boldsymbol{\theta} = (\gamma, \boldsymbol{\theta}'_z, \boldsymbol{\theta}'_x, \boldsymbol{\theta}'_s)'$, where $\boldsymbol{\theta}_z$ and $\boldsymbol{\theta}_x$ are the vectors of coefficients for the main effects, of dimension $c_z \times 1$ and $c_x \times 1$, respectively, and $\boldsymbol{\theta}_s$ is the vector of coefficients for the interaction, of dimension $c_z c_x \times 1$. Therefore, model (31) is written as:

$$\mathbf{y} = \gamma \mathbf{1}_n + (\mathbf{1}_{n_x} \otimes \mathbf{B}_z) \boldsymbol{\theta}_z + (\mathbf{B}_x \otimes \mathbf{1}_{n_z}) \boldsymbol{\theta}_x + (\mathbf{B}_x \otimes \mathbf{B}_z) \boldsymbol{\theta}_s, \quad (33)$$

with penalty equal to the following block-diagonal matrix:

$$\mathbf{P} = \begin{bmatrix} 0 & & & \\ & \lambda_z \mathbf{D}'_z \mathbf{D}_z & & \\ & & \lambda_x \mathbf{D}'_x \mathbf{D}_x & \\ & & & \tau_z \mathbf{I}_{c_x} \otimes \mathbf{D}'_z \mathbf{D}_z + \tau_x \mathbf{D}'_x \mathbf{D}_x \otimes \mathbf{I}_{c_z} \end{bmatrix}, \quad (34)$$

of dimension $(1 + c_z + c_x + c_z c_x) \times (1 + c_z + c_x + c_z c_x)$, where each block corresponds to the penalty over each set of coefficients in the model. Ref. [15] pointed out that model (33) and penalty (34) should be modified in order to preserve identifiability. Their proposal is to construct identifiable bases and penalties by reparameterizing the model as a mixed model instead of imposing numerical constraints as suggested by [18]. Ref. [15] defines the following transformation matrix $\boldsymbol{\Omega} = [\boldsymbol{\Omega}_f \mid \boldsymbol{\Omega}_r]$ with dimension $(1 + c_z + c_x + c_z c_x)$, where:

$$\boldsymbol{\Omega}_f = \begin{bmatrix} 1 & & & \\ & 1 \otimes \mathbf{u}_{zf}^{(2)} & & \\ & & \mathbf{u}_{xf}^{(2)} \otimes 1 & \\ & & & \mathbf{u}_{xf}^{(2)} \otimes \mathbf{u}_{zf}^{(2)} \end{bmatrix}, \quad (35)$$

$$\boldsymbol{\Omega}_r = \begin{bmatrix} 1 & & & \\ & 1 \otimes \mathbf{U}_{zr} & & \\ & & \mathbf{U}_{xr} \otimes 1 & \\ & & & \mathbf{u}_{xf}^{(2)} \otimes \mathbf{U}_{zr} \mid \mathbf{U}_{xr} \otimes \mathbf{u}_{zf}^{(2)} \mid \mathbf{U}_{xr} \otimes \mathbf{U}_{zr} \end{bmatrix}, \quad (36)$$

and $\mathbf{U}_{zf}$ and $\mathbf{U}_{xf}$ are the eigenvectors corresponding to the positive values of the SVD of the marginal penalties, $\mathbf{D}'_z \mathbf{D}_z$ and $\mathbf{D}'_x \mathbf{D}_x$, and $\mathbf{u}_{zf}^{(2)}$ and $\mathbf{u}_{xf}^{(2)}$ are the second columns of $\mathbf{U}_{zr}$ and $\mathbf{U}_{xr}$, respectively.

Given the previous transformation matrix, ref. [15] showed that the fixed and random effects matrices $\mathbf{X}$ and $\mathbf{Z}$ are:

$$\mathbf{X} = [\mathbf{1}_n \mid \mathbf{1}_{n_x} \otimes \tilde{\mathbf{z}} \mid \tilde{\mathbf{x}} \otimes \mathbf{1}_{n_z} \mid \tilde{\mathbf{x}} \otimes \tilde{\mathbf{z}}] \quad (37)$$

$$\mathbf{Z} = [\mathbf{1}_n \mid \mathbf{1}_{n_x} \otimes \mathbf{Z}_z \mid \mathbf{Z}_x \otimes \mathbf{1}_{n_z} \mid \tilde{\mathbf{x}} \otimes \mathbf{Z}_z \mid \mathbf{Z}_x \otimes \tilde{\mathbf{z}} \mid \mathbf{Z}_x \otimes \mathbf{Z}_z] \quad (38)$$

where $\tilde{\mathbf{z}} = \mathbf{B}_z \mathbf{u}_{zf}^{(2)}$, $\tilde{\mathbf{x}} = \mathbf{B}_x \mathbf{u}_{xf}^{(2)}$, $\mathbf{Z}_z = \mathbf{B}_z \mathbf{U}_{zr}$ and $\mathbf{Z}_x = \mathbf{B}_x \mathbf{U}_{xr}$. Moreover, the mixed model penalty (or equivalently, the precision matrix of random effects) is:

$$\mathbf{F} = \text{blockdiag}(\mathbf{F}_{(1)}, \mathbf{F}_{(2)}, \mathbf{F}_{(1,2)}), \quad (39)$$

where for a second order penalty, it has size $(c_z c_x - 4) \times (c_z c_x - 4)$, and with:

$$\begin{aligned} \mathbf{F}_{(1)} &= \lambda_z \tilde{\Sigma}_z, \\ \mathbf{F}_{(2)} &= \lambda_x \tilde{\Sigma}_x, \\ \mathbf{F}_{(1,2)} &= \text{blockdiag}(\tau_z \tilde{\Sigma}_z, \tau_x \tilde{\Sigma}_x, \tau_z \mathbf{I}_{c_x-2} \otimes \tilde{\Sigma}_z + \tau_x \tilde{\Sigma}_x \otimes \mathbf{I}_{c_z-2}) \end{aligned}$$

where $\tilde{\Sigma}_z$ and $\tilde{\Sigma}_x$ are the nonzero eigenvalues of the SVD of $D'_z D_z$ and $D'_x D_x$, respectively. Then, model (31) becomes:

$$y = X\beta + Z\alpha + \epsilon \quad \alpha \sim N(\mathbf{0}, F^{-1}).$$

With the previous representation of model (31), ref. [15] avoid the identifiability problem removing the column vector of 1s in the random effects matrix (38). For more details, see [15].

Although the out-of-sample prediction will be performed in the context of a mixed model, we introduced the method in the original P-spline formulation because the reparameterization required for out-of-sample prediction will be based on this definition. In particular, we need to know the extended penalty matrix in order to calculate the precision matrix of random effects.

In model (31), given a vector of $n_z n_x$ observations y of the response variable, suppose that we want to predict $n_p = n_z n_{x_p} + n_{z_p} n_x + n_{z_p} n_{x_p}$ new values at (z, x_p) , (z_p, x) and (z_p, x_p) of observed and predicted values can be arranged as in (20). Then, we consider the following extended smooth-ANOVA model:

$$y_+ = \gamma + f_1(z_+) + f_2(x_+) + f_{1,2}(z_+, x_+) + \epsilon_+, \quad \epsilon_+ \sim \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 R_+) \quad (40)$$

where R_+ is defined as in (10) with an extended B-spline basis B_+ defined as:

$$B_+ = [\mathbf{1}_{n_+} \mid \mathbf{1}_{n_{x_+}} \otimes B_{z_+} \mid B_{x_+} \otimes \mathbf{1}_{n_{z_+}} \mid B_{x_+} \otimes B_{z_+}], \quad (41)$$

of dimension $n \times (1 + c_{z_+} + c_{x_+} + c_{z_+} c_{x_+})$, and where $\mathbf{1}_{n_{z_+}}$ and $\mathbf{1}_{n_{x_+}}$ are column vectors of ones of length n_{z_+} and n_{x_+} , respectively.

Therefore, model (40) is written as:

$$y_+ = \gamma \mathbf{1}_{n_+} + (\mathbf{1}_{n_{x_+}} \otimes B_{z_+}) \theta_{z_+} + (B_{x_+} \otimes \mathbf{1}_{n_{z_+}}) \theta_{x_+} + (B_{x_+} \otimes B_{z_+}) \theta_{s_+}, \quad (42)$$

with extended penalty the following block-diagonal matrix:

$$P_+ = \begin{bmatrix} 0 & & & \\ & \lambda_z D'_{z_+} D_{z_+} & & \\ & & \lambda_x D'_{x_+} D_{x_+} & \\ & & & \tau_z I_{c_{x_+}} \otimes D'_{z_+} D_{z_+} + \tau_x D'_{x_+} D_{x_+} \otimes I_{c_{z_+}} \end{bmatrix}, \quad (43)$$

of dimension $(1 + c_{z_+} + c_{x_+} + c_{z_+} c_{x_+}) \times (1 + c_{z_+} + c_{x_+} + c_{z_+} c_{x_+})$, where each block corresponds to the penalty over each of the coefficients of the model.

In order to reformulate the extended S-ANOVA model (40) as a mixed model, we need to define an extended transformation matrix. The obvious choice would be to use the natural extended transformation matrix based on the SVD of the extended difference matrices; however, this method cannot ensure the invariance of the fit when performing out-of-sample predictions. In the following section, we propose the use of an extended transformation matrix, Ω_+^* , that ensures coherent predictions.

3.1. Coherent Prediction with S-ANOVA Model

In order to use the S-ANOVA model for prediction under the constraint that the fit must be maintained, an extended transformation matrix that preserves the model matrix must be used. When the two covariates are extended, we define the following extended transformation matrix $\Omega_+^* = [\Omega_{+f}^* \mid \Omega_{+r}^*]$ with dimension $(1 + c_{z+} + c_{x+} + c_{z+}c_{x+})$, and:

$$\Omega_{+f}^* = \begin{bmatrix} 1 & & & \\ & 1 \otimes \mathbf{u}_{z+f}^{*(2)} & & \\ & & \mathbf{u}_{x+f}^{*(2)} \otimes 1 & \\ & & & \mathbf{u}_{x+f}^{*(2)} \otimes \mathbf{u}_{z+f}^{*(2)} \end{bmatrix}, \quad (44)$$

$$\Omega_{+r}^* = \begin{bmatrix} 1 & & & & \\ & 1 \otimes \mathbf{U}_{z+r}^* & & & \\ & & \mathbf{U}_{x+r}^* \otimes 1 & & \\ & & & \mathbf{u}_{x+f}^{*(2)} \otimes \mathbf{U}_{z+r}^* \mid \mathbf{U}_{x+r}^* \otimes \mathbf{u}_{z+f}^{*(2)} \mid \mathbf{U}_{x+r}^* \otimes \mathbf{U}_{z+r}^* \end{bmatrix}. \quad (45)$$

where $\mathbf{u}_{z+f}^{*(2)}$ and $\mathbf{u}_{x+f}^{*(2)}$ are the second columns of $\mathbf{U}_{z+f}^*$ and $\mathbf{U}_{x+f}^*$, respectively,

$$\mathbf{U}_{z+f}^* = \begin{bmatrix} \mathbf{U}_{zf} \\ -D_{z(2)}^{-1} D_{z(1)} \mathbf{U}_{zf} \end{bmatrix}, \text{ and} \quad (46)$$

$\mathbf{U}_{zf}$ are the eigenvectors corresponding to the positive values of the SVD of $D_z' D_z$ and $D_{z(2)}$ and $D_{z(1)}$ are blocks of (14). $\mathbf{U}_{z+f}^*$, $\mathbf{U}_{x+f}^*$, $\mathbf{U}_{xf}$, $D_{x(2)}$ and $D_{x(1)}$ have analogous definitions. Then, the fixed effects matrix is:

$$\mathbf{X}_+^* = \mathbf{B}_+ \Omega_{+f}^* = [\mathbf{1}_{n_+} \mid \mathbf{1}_{n_{x+}} \otimes \tilde{\mathbf{z}}_+ \mid \tilde{\mathbf{x}}_+ \otimes \mathbf{1}_{n_{z+}} \mid \tilde{\mathbf{x}}_+ \otimes \tilde{\mathbf{z}}_+], \quad (47)$$

where $\tilde{\mathbf{z}}_+$ and $\tilde{\mathbf{x}}_+$ are $\mathbf{B}_{z+} \mathbf{u}_{z+f}^{*(2)}$ and $\mathbf{B}_{x+} \mathbf{u}_{x+f}^{*(2)}$, respectively, and the random effects matrix is:

$$\mathbf{Z}_+^* = \mathbf{B}_+ \Omega_{+r}^* = [\mathbf{1}_{n_+} \mid \mathbf{1}_{n_{x+}} \otimes \mathbf{Z}_{z+}^* \mid \mathbf{Z}_{x+}^* \otimes \mathbf{1}_{n_{z+}} \mid \tilde{\mathbf{x}}_+ \otimes \mathbf{Z}_{z+}^* \mid \mathbf{Z}_{x+}^* \otimes \tilde{\mathbf{z}}_+ \mid \mathbf{Z}_{x+}^* \otimes \mathbf{Z}_{z+}^*], \quad (48)$$

where $\mathbf{Z}_{z+}^*$ and $\mathbf{Z}_{x+}^*$ are $\mathbf{B}_{z+} \mathbf{U}_{z+r}^*$ and $\mathbf{B}_{x+} \mathbf{U}_{x+r}^*$, respectively. The covariance matrix of random effects is $\mathbf{G}_+^* = \sigma_e^2 \mathbf{F}_+^{*-1}$, with $\mathbf{F}_+^*$ given by the following theorem.

Theorem 2. The extended precision matrix of the random effects for S-ANOVA model in (40) with extended transformation matrix given in (45), and extended penalty given in (43), is a block-diagonal matrix given by:

$$\mathbf{F}_+^* = \text{blockdiag}(\mathbf{F}^{(1)}, \mathbf{F}_+^{(2)}, \mathbf{F}_+^{(1,2)}), \quad (49)$$

where:

$$\begin{aligned} \mathbf{F}^{(1)} &= \lambda_z \mathbf{U}_{z+r}^{*'} D_{z+}' D_{z+} \mathbf{U}_{z+r}^*, \\ \mathbf{F}_+^{(2)} &= \lambda_x \mathbf{U}_{x+r}^{*'} D_{x+}' D_{x+} \mathbf{U}_{x+r}^*, \\ \mathbf{F}_+^{(1,2)} &= \begin{bmatrix} \mathbf{F}_{+11}^{(1,2)} & \mathbf{O} & \mathbf{F}_{+13}^{(1,2)} \\ \mathbf{O} & \mathbf{F}_{+22}^{(1,2)} & \mathbf{F}_{+23}^{(1,2)} \\ \mathbf{F}_{+13}^{(1,2)'} & \mathbf{F}_{+23}^{(1,2)'} & \mathbf{F}_{+33}^{(1,2)} \end{bmatrix}, \end{aligned}$$

with

$$\begin{aligned}
 F_{+11}^{(1,2)} &= \tau_z \mathbf{u}_{x+f}^{*(2)'} \mathbf{u}_{x+f}^{*(2)} \otimes \mathbf{U}_{z+r}^{*'} D'_{z+} D_{z+} \mathbf{U}_{z+r}^* \\
 F_{+13}^{(1,2)} &= \tau_z \mathbf{u}_{x+f}^{*(2)'} \mathbf{U}_{x+r}^* \otimes \mathbf{U}_{z+r}^{*'} D'_{z+} D_{z+} \mathbf{U}_{z+r}^* \\
 F_{+22}^{(1,2)} &= \tau_x \mathbf{U}_{x+r}^{*'} D'_{x+} D_{x+} \mathbf{U}_{x+r}^* \otimes \mathbf{u}_{z+r}^{*(2)'} \mathbf{u}_{z+r}^{*(2)}, \\
 F_{+23}^{(1,2)} &= \tau_x \mathbf{U}_{x+r}^{*'} D'_{x+} D_{x+} \mathbf{U}_{x+r}^* \otimes \mathbf{u}_{z+f}^{*(2)'} \mathbf{U}_{z+r}^* \\
 F_{+33}^{(1,2)} &= \tau_z \mathbf{U}_{x+r}^{*'} \mathbf{U}_{x+r}^* \otimes \mathbf{U}_{z+r}^{*'} D'_{z+} D_{z+} \mathbf{U}_{z+r}^* + \tau_x \mathbf{U}_{x+r}^{*'} D'_{x+} D_{x+} \mathbf{U}_{x+r}^* \otimes \mathbf{U}_{z+r}^{*'} \mathbf{U}_{z+r}^*.
 \end{aligned}$$

The proof is given in the supplementary materials.

3.2. Example 2: Prediction of Aboveground Biomass

In this section, we apply the proposed 2D interaction P-spline and S-ANOVA models to another real data sets. The data come from an agricultural experiment conducted in Spain [19] to evaluate the economic viability of Populus trees prior to harvesting. In this context, it is important to estimate the aboveground biomass and obtain accurate predictions using only a minimal set of readily available information, i.e., diameter and height (denoted by z and x). Ref. [20] proposed the use of smooth additive mixed models for predicting aboveground biomass. Here, we analyzed the data of one of the nine clones included in the experiment. The observation data consist of 315 observations for the diameter value measured at a breast height of 1.30 m. The data are shown in Figure 4, where it can be seen that the weight is a nonlinear function of diameter and height. We propose predicting out-of-sample aboveground biomass as a smooth function of height and diameter using the following two models:

- 2D interaction P-spline model:

$$\mathbf{y}_+ = f(\mathbf{z}_+, \mathbf{x}_+) + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_+)$$

- Smooth-ANOVA model:

$$\mathbf{y}_+ = f(\mathbf{z}_+) + f(\mathbf{x}_+) + f(\mathbf{z}_+, \mathbf{x}_+) + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_+)$$

Figure 4. Plot of weight versus height (left) and plot of total weight versus diameter (right).

We have predicted weights for 10 new out-of-sample values for diameter and height. In Figure 5, we show the estimated smooth main effects for height (left panel) and diameter (right panel) obtained after fitting and predicting with the S-ANOVA model imposing that the fit is maintained. As shown, the effect of height is weaker than diameter, but both smooth terms are significantly different from zero. Figure 6 shows the fitted and predicted interaction function ($f(\mathbf{z}_+, \mathbf{x}_+)$) for the restricted S-ANOVA model.

Figure 5. Fitted and predicted smooth curves for height (**left**) and for diameter (**right**) using the restricted S-ANOVA model. The vertical line indicates the height and diameter values from which we predict (9.32 and 7.3).

Figure 6. Fitted and predicted interaction function for the the restricted S-ANOVA model. The vertical line indicates the height value from which we predict (9.32) and the horizontal line indicates the diameter value from which we predict (7.3).

Figure 7 shows the predictions when the fit and the forecast are obtained with the S-ANOVA model (top left panel), with the S-ANOVA model subject to the constraint that the fit is maintained (top right panel), with the 2D interaction P-spline model (bottom left panel) and with the 2D interaction P-spline models imposing that the fit is maintained (bottom right panel). As shown in the figure, the results obtained by the S-ANOVA model and the restricted S-ANOVA model are almost equal. However, the solution obtained from the 2D interaction P-spline model is different depending on whether the constraint of maintaining the fit is imposed. However, the solutions obtained from the 2D interaction P-spline models are different depending on whether the restriction the fit is maintained is imposed or not. In particular, if no constraints are imposed, the fit will change significantly.

In the prediction, the most significant difference is that weight values predicted by the 2D interaction P-spline model at the largest diameter and height values are lower than the ones obtained by the S-ANOVA model. It is important to note that, in the fit, the degrees of freedom are 32 for the 2D interaction P-spline model and 12 for the S-ANOVA model, i.e., with the 2D interaction P-spline model, we are adding unnecessary complexity.

Furthermore, comparing the S-ANOVA model and the 2D interaction P-spline model, we conclude that the S-ANOVA model gives the most coherent solution because it provides the largest weight value for the highest values of height and diameter.

Figure 7. Fit and prediction with the S-ANOVA model (**top left**), with the S-ANOVA model, imposing that the fit is maintained (**top right**), with the 2D interaction P-spline model (**bottom left**) and with the 2D interaction P-spline models imposing that the fit is maintained (**bottom right**) at out-of-sample values of diameter $[7.3, 10]$ and height $[9.32, 15]$.

4. Simulation Study

We propose two simulation studies to examine the performance of the interaction 2D model and smooth-ANOVA model compared to their constrained versions (imposing the invariance of the fit when performing out-of-sample predictions). For this propose we have simulated the data in two different scenarios:

(a) Scenario 1. From an interaction model:

$$S^1 = f_{1,2}(z, x).$$

(b) Scenario 2. From a two main effects with interaction model:

$$S^2 = f_1(z) + f_2(x) + f_{1,2}(z, x).$$

where $f_1(z) = \sin(2\pi z)$, $f_2(x) = \cos(3\pi x)$ and $f_{1,2}(z, x) = 3 \sin(2\pi z)(2x - 1)$. To simulate the data, we have generated a grid of 4900 values. Both covariates, z and x , take 70 equidistant values in the interval $[0, 1]$, and the errors are i.i.d., with mean 0 and variance $\sigma_\epsilon^2 = 0.25$. Figure 8 shows $f_1(z)$, $f_2(x)$ and the surfaces proposed in scenarios 1 and 2.

The models compared are:

- 2D interaction P-spline model, i.e.,

$$y_+ = f(z_+, x_+) + \epsilon_+ \quad \epsilon_+ \sim \mathcal{N}(0, \sigma_\epsilon^2 R_+).$$

- Smooth-ANOVA model, i.e.,

$$y_+ = f_1(z_+) + f_2(x_+) + f_{1,2}(z_+, x_+) + \epsilon_+ \quad \epsilon_+ \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

- Constrained 2D interaction P-spline model;
- Constrained smooth-ANOVA model.

Figure 8. Functions (a,b) are the nonlinear main effects of z and x , (c) is the interaction surface and (d) is the sum of the main effects and the interaction surfaces.

The constraint imposed is that the coefficients of the fit remain invariant when out-of-sample-prediction is performed. In all cases, smoothing parameters are estimated using REML.

For each scenario and each model, we repeated the following steps 100 times:

- (1) Start with a data set of observations arranged into a matrix Y_+ of dimension $n_{z_+} \times n_{x_+}$:

$$Y_+ = \begin{bmatrix} Y & Y_{zx_p} \\ Y_{z_px} & Y_{z_px_p} \end{bmatrix},$$

where Y , Y_{zx_p} , Y_{z_px} and $Y_{z_px_p}$ have dimensions $n_z \times n_x$, $n_z \times n_{x_p}$, $n_{z_p} \times n_x$ and $n_{z_p} \times n_{x_p}$, respectively;

- (2) Split our data into two groups: the training data Y and the test data Y_{zx_p} , Y_{z_px} and $Y_{z_px_p}$;
- (3) Use the training data to predict $4900 - n$ new observations ($n = n_z \times n_x$);
- (4) Check the accuracy of each model.

The marginal B-spline bases B_z and B_x are constructed with 15 knots and cubic splines and then extended to cover the whole range of z_+ and x_+ , the penalty orders are two.

As is known, a model that fits the data well does not necessarily yields good predictions, so we compared the fit performance, the forecast performance, and the overall method performance. In order to check the accuracy of the methods proposed, we followed [21] and used the difference between the function values from which we simulated the data and the fit and forecast produced using only the data in the training set:

$$E_+ = S^k - \hat{Y}_+, \quad k = 1, 2$$

i.e., we have the errors matrix $E_+ = \begin{bmatrix} E & E_{zx_p} \\ E_{z_px} & E_{z_px_p} \end{bmatrix}$, and therefore the vectors containing the errors in the fit, in the forecast and in the overall performance:

- Vector with the fit errors: $e^{(f)} = \text{vec}(E)$;
- Vector with the prediction errors: $e^{(p)} = (\text{vec}(E_{zx_p}), \text{vec}(E_{z_px}), \text{vec}(E_{z_px_p}))$;
- Vector with the total errors: $e^{(t)} = \text{vec}(E_+)$.

The errors measure that we use is the mean absolute error because, as it is concluded in [21], it is less sensitive to outliers than the root mean square error: Mean absolute error.

$$\text{MAE} = \frac{\sum_{i=1}^N |e_i^{(l)}|}{N}, \quad N = \text{length}(e^{(l)}) \text{ for } l = f, p, t, \text{ i.e., for the errors in the fit, in the forecast or in total.}$$

4.1. Simulations Results for Scenario 1

Here, we present results from Scenario 1. We use different ranges for the out of sample prediction, $n_{x_p}/n_{z_p} = 0, 5, 10, 15, 20$ and fit and predict simultaneously with the four models described above. Note that for the different values of n_{z_p} and n_{x_p} we consider as the training data an observations matrix of dimension $(70 - n_{z_p}) \times (70 - n_{x_p})$ and we predict $(70 - n_{z_p}) \times n_{x_p} + n_{z_p} \times (70 - n_{x_p}) + n_{z_p} \times n_{x_p}$ new values.

In Figure 9, we report MAE values from the fit, the forecast, and the overall performance. The x-axis label of boxplots refers to the model: unrestricted (2D interaction P-spline model), restricted (2D interaction P-spline model with constraints), ANOVA (S-ANOVA Model) and restricted S-ANOVA (S-ANOVA model with constraints).

Plots for the different values of n_{z_p} and n_{x_p} are available in the supplementary materials, and below we show the particular case of $n_{z_p} = 0$ and $n_{x_p} = 5$.

Figure 9. MAE in the fit (left), in the forecast (middle) and in total (right) of smooth models in scenario 1 and $n_{z_p} = 0$ and $n_{x_p} = 5$.

In most cases, the fit remains relatively invariant in the case of the 2D P-spline models, with and without constraints, and in the restricted S-ANOVA model, while the fit in unrestricted S-ANOVA model is modified in most scenarios.

In the prediction, the three models behave similarly; however, the performance of the restricted S-ANOVA model is a bit worse and its MAE error values increase with an increasing prediction horizon (see Figures S1 and S2 of the supplementary materials).

Based on this simulated scenario, we conclude that simultaneous fit and prediction with S-ANOVA models produces significant changes in the fit and that in the overall performance (including fit and prediction), the 2D P-spline models (with and without restriction) and the restricted S-ANOVA models are the most accurate.

4.2. Simulations Results for Scenario 2

In the supplementary materials (Figure S12), we show the results obtained for Scenario 2, i.e., in the case in which the true surface is constructed from a model with two main effects and an interaction. Figure 10 shows the boxplots for the particular case of $n_{z_p} = 0$ and $n_{x_p} = 5$.

Figure 10. MAE in the fit (left), in the forecast (middle) and in total (right) of smooth models in scenario 2 and $n_{z_p} = 0$ and $n_{x_p} = 5$.

In this case, the conclusion is that there is a significant difference between the results obtained with the interaction models and the S-ANOVA models: two-dimensional P-spline models (restricted and unrestricted) are always less accurate than the S-ANOVA models. This is due to the fact that 2D interaction models try to fit the data without considering the main effects.

When fitting the observed data, the performance of 2D P-spline interaction models, with and without constraints, is quite similar. However, the restricted interaction model is better than the unrestricted model in the out-of-sample prediction, the difference between the two models is apparent for most scenarios.

In the case of the S-ANOVA models, the restricted model is always better than the unrestricted one, in the fit, in the prediction and, therefore, in the overall performance. Although, in the particular case in which we only predict in one dimension, the prediction errors for both models are almost equal (see Supplementary Materials, scenario 2). The conclusion for this scenario is that the restricted S-ANOVA model is clearly the most accurate.

Therefore, considering all test scenarios, our recommendation is that, as a general rule, restricted S-ANOVA models should be used to fit and predict models that require interaction terms.

5. Conclusions

In this article, we propose a general framework for out-of-sample prediction in a smooth additive model with interaction terms. We build our proposal from the method proposed in [10], and then we extend their method to situations where out-of-sample predictions are required in both directions of the interaction term. The method proposed in [10] considers out-of-sample predictions as missing values with a weight of 0, and performs fitting and prediction at the same time. We have shown that this method produces different fitted values, depending on whether we only fit or fit and predict simultaneously.

In order to solve this inconsistency, we recommend maximizing the penalized likelihood under linear constraints to ensure that the coefficients obtained in the data fitting remain unchanged when the fitting and out-of-sample prediction are performed at the same time. For this purpose, we use Lagrange multipliers. This general method can be

used to impose other related constraints, such as mortality forecasting when the cross-age structure needs to be preserved.

The methodology proposed has been extended to the case of smooth mixed models since it will allow us to predict out-of-sample in a wide range of models. We have shown that care should be taken when formulating the models as mixed models, since the matrices of fixed and random effects for out-of-sample prediction need to be direct extensions of the matrices used in the fit. In this context, we have developed a method for out-of-sample prediction for the S-ANOVA model proposed in [8]. This is an important step forward since, until now, there was no general method available for prediction in models that include main effects and interactions terms. A simulation study has been carried out to compare restricted and unrestricted out-of-sample predictions when using 2D interaction models and S-ANOVA models. From the results of the simulation study, we conclude that, in most cases, when the prediction is made on the two dimensions of the interaction term, the restricted S-ANOVA model performs better in fitting and out-of-sample prediction. The results for only one of the covariates depends on the simulation scenario, although in most cases, the S-ANOVA model is better than the full interaction (restricted or unrestricted) model.

The multidimensional prediction P-splines method with constraints proposed has also been used in two real data examples, one in which mortality rates are forecasted over the years, and the importance of imposing constraints on the coefficients to ensure coherent forecast is shown. The other example predicts tree biomass as a function of the tree height and diameter using the S-ANOVA model; this example shows the ability of this model to produce separate out-of-sample predictions for main effects and interaction terms in the model. Although here, due to the nature of the data used, we have only shown examples of forward out-of-sample prediction, our methodology allows for backwards out-of-sample predictions.

Currently, the work presented here is being extended to non-Gaussian responses in the context of generalized additive and mixed models.

Supplementary Materials: The following are available online at <https://www.mdpi.com/article/10.3390/math9151761/s1> with Proof of theorem 1 and Corollary of theorem 1; Proof of theorem 2; Figures S1–S11 contains additional settings for the parameters in scenario 1; Figures S12–S21 contains additional settings for the parameters in scenario 2.

Author Contributions: Conceptualization, A.C., M.D. and D.-J.L.; methodology, A.C., M.D. and D.-J.L.; software, A.C.; validation, A.C.; formal analysis, A.C., M.D. and D.-J.L.; investigation, A.C., M.D. and D.-J.L.; resources, M.D. and D.-J.L.; data curation, M.D.; writing—original draft preparation, A.C.; writing—review and editing, M.D. and D.-J.L.; visualization, D.-J.L.; supervision, M.D. and D.-J.L.; project administration, M.D. and D.-J.L.; funding acquisition, M.D. and D.-J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by Ministerio de Ciencia e Innovación grant numbers PID2019-104901RB-I00. The third author gratefully acknowledges support by the Department of Education, Language Policy and Culture from the Basque Government (BERC 2018-2021 program), the Spanish Ministry of Economy and Competitiveness MINECO and FEDER: PID2020-115882RB-I00/AEI/10.13039/501100011033 funded by Agencia Estatal de Investigación and acronym “S3M1P4R”, and BCAM Severo Ochoa excellence accreditation SEV-2017-0718).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Mortality data are openly available in the Human Mortality Database <https://www.mortality.org> (accessed on 1 July 2021). Biomass data are not available due to confidentiality issues.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eilers, P.H.C.; Marx, B.D. Flexible smoothing with B-splines and penalties. *Stat. Sci.* **1996**, *11*, 89–121. [[CrossRef](#)]
2. Eilers, P.H.C.; Marx, B.D.; Maria, D. Twenty years of P-splines. *Stat. Oper. Res. Trans.* **2015**, *39*, 149–186.
3. Li, Y.; Ruppert, D. On the asymptotics of penalized splines. *Biometrika* **2008**, *95*, 415–436. [[CrossRef](#)]
4. Xiao, L. Asymptotic theory of penalized splines. *Electron. J. Stat.* **2019**, *13*, 747–794. [[CrossRef](#)]
5. Wand, M. Smoothing and mixed models. *Comput. Stat.* **2003**, *18*, 223–249. [[CrossRef](#)]
6. Kauermann, G. Penalized splines, Mixed Models and Bayesian Ideas. In *Statistical Modelling and Regression Structures: Festschrift in Honour of Ludwig Fahrmeir*; Physica-Verlag: Heidelberg, Germany, 2010; pp. 45–58. [[CrossRef](#)]
7. Currie, I.D.; Durbán, M.; Eilers, P.H.C. Generalized linear array models with applications to multidimensional smoothing. *J. R. Stat. Soc. Ser. B* **2006**, *68*, 1–22. [[CrossRef](#)]
8. Lee, D.J.; Durbán, M. P-spline ANOVA-type interaction models for spatio-temporal smoothing. *Stat. Model.* **2011**, *11*, 49–69. [[CrossRef](#)]
9. Lee, D.J.; Durbán, M.; Eilers, P.H.C. Efficient two-dimensional smoothing with P-spline ANOVA mixed models and nested bases. *Comput. Stat. Data Anal.* **2013**, *61*, 22–37. [[CrossRef](#)]
10. Currie, I.D.; Durbán, M.; Eilers, P.H.C. Smoothing and forecasting mortality rates. *Stat. Model.* **2004**, *4*, 279–298. [[CrossRef](#)]
11. Etxeberria, J.; Ugarte, M.; Goicoa, T.; Militino, A.F. On predicting cancer mortality using anova-type P-splines models. *Rev. Stat. J.* **2015**, *13*, 21–40.
12. Ugarte, M.D.; Goicoa, T.; Etxeberria, J.; Militino, A.F. Projections of cancer mortality risks using spatio-temporal P-spline models. *Stat. Methods Med. Res.* **2012**, *21*, 545–560. [[CrossRef](#)] [[PubMed](#)]
13. Carballo, A.; Durban, M.; Kauermann, G.; Lee, D.J. A general framework for prediction in penalized regression. *Stat. Model.* **2020**. [[CrossRef](#)]
14. Greene, W.H.; Seaks, G.T. The Restricted Least Squares Estimator: A Pedagogical Note. *Rev. Econ. Stat.* **1991**, *73*, 563–567. [[CrossRef](#)]
15. Lee, D.J. Smoothing Mixed Models for Spatial and Spatio-Temporal Data. Ph.D. Thesis, Department of Statistics, Universidad Carlos III de Madrid, Madrid, Spain, 2010.
16. Lee, R.; Carter, L. Modelling and forecasting the time series of US mortality. *J. Am. Stat. Assoc.* **1992**, *87*, 659–671.
17. Delwarde, A.; Denuit, M.; Eilers, P.H.C. Smoothing the Lee-Carter and Poisson log-bilinear models for mortality forecasting: A penalized log-likelihood approach. *Stat. Model.* **2007**, *7*, 29–48. [[CrossRef](#)]
18. Wood, S. Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics* **2006**, *62*, 1025–1036. [[CrossRef](#)] [[PubMed](#)]
19. Rivas-Martínez, S.; Díaz, T.E.; Fernández-González, F.; Izco, J.; Loidi, J.; Lousã, M.; Penas, A. Vascular plant communities of Spain and Portugal. Addenda to the syntaxonomical checklist of 2001. *Itinera Geobot.* **2002**, *15*, 5–922.
20. Sánchez-González, M.; Durbán, M.; Lee, D.; Cañellas, I.; Sixto, H. Smooth additive mixed models for predicting aboveground biomass. *J. Agric. Biol. Environ. Stat.* **2016**, *22*, 23–41. [[CrossRef](#)]
21. Hyndman, R.J.; Koehler, A.B. Another look at measures of forecast accuracy. *Int. J. Forecast.* **2006**, *22*, 679–688. [[CrossRef](#)]