Carballo, A., Durbán, M., Kauermann, G. and Lee, D-J. (2020). A general framework for prediction in penalized regression. *Statistical Modelling*.

https://doi.org/10.1177/1471082X19896867

A general framework for prediction in penalized regression

Alba Carballo [1], Maria Durban [1], Göeran Kauermann [2] and Dae-Jin Lee [3]

[1] Departamento de Estadística, Universidad Carlos III de Madrid, Madrid, Spain

[2] Institut fur Statistik, Ludwing-Maximilians-Universitat, Munchen, Germany

[3] BCAM - Basque Center for Applied Mathematics, Bilbao, Basque Country, Spain

---

Address for correspondence: Alba Carballo, Departamento de Estadísitica, Universidad Carlos III de Madrid, Edificio Juan Benet, Av. de la Universidad 30, 28911 Leganés, Madrid, Spain.

E-mail: albcarba@est-econ.uc3m.es.

Phone: (+34) 916 246 201.

Fax: (+34) 916 249 177.

Abstract: There are two main approaches to carrying out prediction in the context of penalized regression: with low-rank basis and penalties or through the smooth mixed models. In this paper, we give further inside in the case of P-splines showing the influence of the penalty on the prediction. In the context of mixed models, we can connect the new predicted values to the observed values through a joint normal distribution, which allows us to compute prediction intervals. In this work, we propose

an alternative approach, called the "*extended mixed model approach*" that allows us to fit and predict data simultaneously. The methodology is illustrated with two real data sets, one of them on aboveground biomass and the other on monthly sulphur dioxide (SO_2) levels in a selection of monitoring sites in Europe.

Key words: Mixed models; penalized regression; prediction; P-splines; smooth models

1 Introduction

There are many situations in which prediction of new observations in the context of regression is needed, in particular, when "out of range" prediction is required, that is beyond the range of observed covariates. This problem extends in the framework of smoothing, i.e. for models where the regression function is a smooth but otherwise unspecified function. Examples are numerous, for instance, hourly temperatures at a weather station or the yearly number of deaths, where the latter have a major impact in areas such as demography (mortality tables). A scatterplot of data often exhibits patterns, such as an upward or downward trend, or a pattern that repeats, seasonal variation, both of which might be used to predict new values.

The generalities of the problem and the particular characteristics of datasets are one of the main reasons that encourages us to work in the prediction field and base our work on the forecasting method proposed in Currie et al. (2004). They have shown how the method of penalized splines (P-splines), introduced by Eilers and Marx (1996) and extensively discussed in Ruppert et al. (2003), can be extended to smooth and predict

two-dimensional mortality tables. In particular, the authors show how to construct the appropriate regression bases and penalty matrices for forecasting.

Most of the existing literature in the area is related to the prediction of new observations in a temporal context, i.e. forecast of new observations. Although we have a more general approach, we start by providing a brief review of the main literature related to forecasting in smoothing models by commenting the main approaches of Ba et al. (2012) and Sacks et al. (1989). We also refer to Hyndman et al. (2008) who give an overview of exponential smoothing methods.

Exponential smoothing or weighted smoothing, respectively, refers to a class of forecasting methods, each of them having the property that forecasts are weighted combinations of past observations, with recent observations given relatively more weight than older observations. Hyndman et al. (2008) provide extensive information about exponential smoothing methods. A summary of the exponential smoothing history shows the great importance of this method. It certainly is the most popular forecasting method used in business and industry since the 1950s.

In this paper we focus on computationally driven methods. A recent strategy is to use penalized splines to fit and forecast time series data (Ba et al. (2012)). In this case, one minimizes the penalized least squares criteria:

$$S = (\boldsymbol{y} - \boldsymbol{B\theta})'\boldsymbol{M}(\boldsymbol{y} - \boldsymbol{B\theta}) + \lambda\boldsymbol{\theta}'\boldsymbol{P\theta},$$

where $\boldsymbol{y}$ is the vector of observed responses, $\boldsymbol{\theta}$ are basis coefficients, $\boldsymbol{B}$ is a spline basis that covers the hole range of the explanatory variable and $\boldsymbol{M}$ is a weight matrix that puts exponentially decreasing weights on the samples, according to the order of their arrival. Matrix $\boldsymbol{P}$ is a penalty matrix controlling the smoothness of the fitted function

and λ is the smoothing parameter. Ba et al. (2012) propose a data driven choice for the exponential weight matrix.

A similar prediction problem is tackled in the framework of global optimization where the interest is to evaluate an unknown function at point x, say. The question is now where to place future values of x to evaluate the function, such that relevant (preferably most) information about the function is achieved. Exemplary we refer to Sacks et al. (1989) and Jones et al. (1998), who fit a stochastic process to data and predict the process at a new point given the already observed data. They treat the observations as if they were generated by a constant and an error component which is modelled as a stochastic process. Their approach is called Bayesian global optimization and the concept is the same as the idea behind the well-known technique in spatial statistics called kriging (Cressie, 1993). Prediction can be addressed in a Bayesian setting using Bayesian P-splines by exploiting the properties of the random walk prior (Besag et al. (1995), Rue and Held (2005)). Computation of the predicted pattern and its credible intervals can be performed via MCMC.

Beyond the aforementioned papers, we have not found methods to predict in penalized regression. In this paper, we propose a general framework for prediction in penalized regression and also delve deeper into our knowledge of the forecasting method proposed in Currie et al. (2004). Although we will use B-spline basis and penalties based on differences, the methodology proposed here can be extended to any basis and quadratic penalty. We have organized the remaining of the paper as follows. Section 2 is dedicated to introducing the fundamentals of the method proposed by Currie et al. (2004) and to show some properties that relate the order of the penalty with the shape of the predictions. In Section 3 we describe how forecasting can be

carried out in the context of smooth mixed models. It can be done through two-stages procedures, following Gilmour et al. (2004) or based on the conditional distribution of the new values given the observed data. We show how the last procedure allow us to compute prediction intervals (see Section 3.1). In addition, we propose an alternative method for prediction smooth mixed models that can be done through a one-stage procedure, by extending the proposal of Currie et al. (2004) to the mixed model framework (see Section 3.2). The equivalence of two-stage and one-stage procedures (in terms of predicted values and choice of smoothing parameter) is shown in the particular case of penalties based on differences between adjacent coefficients (Eilers and Marx, 1996). The proposed methodology is illustrated in Section 4 with the analysis of two real data sets. Finally, concluding remarks are made in Section 5.

2 Prediction with smooth models and quadratic penalties

We start our presentation by giving a brief revision of penalized regression. Consider the case of univariate data with response variable $\boldsymbol{y}$ and regressor $\boldsymbol{x}$. The smooth model is of the form

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{I}), \tag{2.1}$$

where $f(\cdot)$ is an unknown smooth function, subsequently also called trend component. We assume to have the data points (x_i, y_i) drawn from the model, for $i = 1, ..., n$, and ϵ_i being independent and identically distributed errors with variance $\sigma^2_{\boldsymbol{\epsilon}}$. In order to estimate $f(\boldsymbol{x})$ we replace the function by a basis representation, i.e. we assume

$f(\boldsymbol{x}) = \sum_{k=1}^{c} \boldsymbol{B}_k(\boldsymbol{x})\boldsymbol{\theta}_k$ where $\boldsymbol{B}_k(\cdot)$ are appropriately chosen basis function, e.g. B-splines (De Boor, 1972). We can then rewrite model (2.1) in matrix form:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma_{\boldsymbol{\epsilon}}^2 \boldsymbol{I}), \tag{2.2}$$

where $\boldsymbol{B}$, of size $n \times c$, is the regression basis constructed row-wise from $\boldsymbol{B}(x_i)$ for $i = 1, ..., n$ and $\boldsymbol{\theta}$ is the vector of regression coefficients with dimension $c \times 1$. Rather than estimating the coefficients $\boldsymbol{\theta}$ in (2.2) by simple maximum likelihood methods we penalize the coefficients through a quadratic penalty, i.e. the fit is:

$$\hat{\boldsymbol{\mu}} = \boldsymbol{B}\hat{\boldsymbol{\theta}} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{P})^{-1}\boldsymbol{B}'\boldsymbol{y},$$

where $\boldsymbol{P}$ is any quadratic penalty of dimension $c \times c$ that forces the coefficients to vary smoothly, and consequently to obtain a smoothed curve and λ is the smoothing parameter, it can be estimated using a information criteria or a cross-validation criteria method. There are several alternatives for the choice of the regression basis $\boldsymbol{B}$ and the penalty matrix $\boldsymbol{P}$, we choose a B-spline basis and set the penalty to $\boldsymbol{P} = \boldsymbol{D}_q'\boldsymbol{D}_q$, with $\boldsymbol{D}_q$ a difference matrix of order q and dimension $(c-q) \times c$ (P-splines) although the results in the Section 3 are valid for any basis and quadratic penalty.

Currie et al. (2004) proposed a method to fit and predict simultaneously in penalized regression models. We call their proposal "*the missing value approach*" subsequently, and give a brief summary of their methodology.

In the framework of model (2.1), given a vector $\boldsymbol{y}$ of n observations of the response variable, suppose that we want to predict n_p new values $\boldsymbol{y}_p$ at $\boldsymbol{x}_p$, where $\boldsymbol{x}_p$ may be within or, more interestingly, outside of the range of observed values x_i. In the following we focus on the case when $\boldsymbol{x}_p$ is not in the convex hull of x_i, $i = 1, ..., n$ and it is to the right of $\boldsymbol{x}$ (forward prediction), but $\boldsymbol{x}_p$ could also be to the left of $\boldsymbol{x}$

(backward prediction). We define the new vector of observations as

$$\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}_p')', \tag{2.3}$$

which contains the observed response $\boldsymbol{y}$ and the unknown values $\boldsymbol{y}_p$ to be predicted. A new extended B-spline basis, $\boldsymbol{B}_+$, is built from a new set of knots that consists of the original knots covering x_i, $i = 1, ..., n$, and extended to the range of the n_p values of $\boldsymbol{x}_{p_j}$, $j = 1, ..., n_p$. This leads to the basis $\boldsymbol{B}_+ = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{O} \\ \boldsymbol{B}_1 & \boldsymbol{B}_2 \end{bmatrix}$ of size $n_+ \times c_+$, where $\boldsymbol{B}$ is the $n \times c$ basis used for fitting the trend component, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are auxiliary bases for prediction up to $n_+ = n + n_p$ values, which are of sizes $n_p \times c$ and $n_p \times c_p$, respectively, and $c_+ = c + c_p$. Figure 1 represents an extended splines basis. We show the original basis $\boldsymbol{B}$ in black, the $\boldsymbol{B}_1$ component in grey and the $\boldsymbol{B}_2$ part with dashed line.

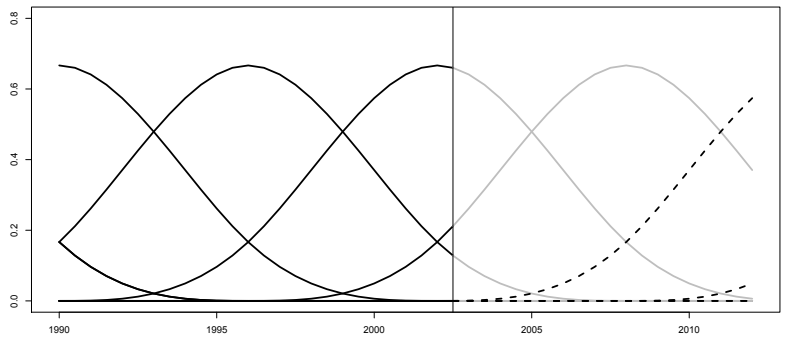


Figure 1: Example of an extended basis to the right of the data (forward).

Associated to the new basis $\boldsymbol{B}_+$, a new vector of coefficients, $\boldsymbol{\theta}_+ = (\boldsymbol{\theta}', \boldsymbol{\theta}_p')'$, is defined, with length $c_+ \times 1$. A new quadratic penalty associated with the new set of coefficients

needs to be introduced, let say $\boldsymbol{P}_+$. Similar to $\boldsymbol{B}_+$, we can also decompose $\boldsymbol{P}_+$ to

$$\boldsymbol{P}_+ = \begin{bmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \\ \boldsymbol{P}_2' & \boldsymbol{P}_3 \end{bmatrix}. \tag{2.4}$$

In the case that $\boldsymbol{P}_+$ is built from q-th order difference matrices:

$$\boldsymbol{P}_+ = \boldsymbol{D}'_+\boldsymbol{D}_+ = \begin{bmatrix} \boldsymbol{D}'\boldsymbol{D} + \boldsymbol{D}_1'\boldsymbol{D}_1 & \boldsymbol{D}_1'\boldsymbol{D}_2 \\ \boldsymbol{D}_2'\boldsymbol{D}_1 & \boldsymbol{D}_2'\boldsymbol{D}_2 \end{bmatrix}, \quad \text{with } \boldsymbol{D}_+ = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_1 & \boldsymbol{D}_2 \end{bmatrix}, \tag{2.5}$$

i.e. $\boldsymbol{P}_1 = \boldsymbol{D}'\boldsymbol{D} + \boldsymbol{D}_1'\boldsymbol{D}_1$, $\boldsymbol{P}_2 = \boldsymbol{D}_1'\boldsymbol{D}_2$ and $\boldsymbol{P}_3 = \boldsymbol{D}_2'\boldsymbol{D}_2$. Notice that $\boldsymbol{D}_+$ has size $(c_+ - q) \times c_+$ and that $\boldsymbol{D}$ is the difference matrix used to build the penalty matrix for the observed data. Moreover, for a second order penalty, $\boldsymbol{D}_+$ is a banded matrix with three non-zero elements per row. Here the subscripts do not indicate the order of the penalty but the blocks of the extended differences matrix.

The model can be fitted and predicted simultaneously by minimizing the following penalized least squares criterion for $\boldsymbol{\theta}_+$:

$$S \;=\; (\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+)'\boldsymbol{W}(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + \lambda\boldsymbol{\theta}'_+\boldsymbol{P}_+\boldsymbol{\theta}_+, \tag{2.6}$$

where the unknown $\boldsymbol{y}_p$ values of $\boldsymbol{y}_+$ are arbitrary and $\boldsymbol{W}$ is a diagonal matrix of dimension $n_+ \times n_+$ with 0 entries if the data is missing, that is for $\boldsymbol{y}_p$, and 1 if the data is observed, that is for $\boldsymbol{y}$. Differentiating with respect to $\boldsymbol{\theta}_+$ leads to

$$\hat{\boldsymbol{\theta}}_+ = (\boldsymbol{B}'_+\boldsymbol{W}\boldsymbol{B}_+ + \lambda\boldsymbol{P}_+)^{-1}\boldsymbol{B}'_+\boldsymbol{W}\boldsymbol{y}_+, \tag{2.7}$$

and $\hat{\boldsymbol{\mu}}_+ = \boldsymbol{H}_+\boldsymbol{y}_+$ with $\boldsymbol{H}_+ = \boldsymbol{B}_+(\boldsymbol{B}'_+\boldsymbol{W}\boldsymbol{B}_+ + \lambda\boldsymbol{P}_+)^{-1}\boldsymbol{B}'_+\boldsymbol{W}$. Note that $\lambda > 0$ is required so that the matrix inversion in (2.7) exists.

Writing the fit and the forecast as a function of the extended penalty matrix (2.4)

and applying Theorem 9.6.1 given in Harville (2000), we get the fit and the prediction

$$\hat{\boldsymbol{\mu}}_+ = \boldsymbol{B}_+ \begin{bmatrix} \boldsymbol{I} \\ -\boldsymbol{P}_3^- \boldsymbol{P}_2' \end{bmatrix} (\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{P}_1 - \lambda \boldsymbol{P}_2 \boldsymbol{P}_3^- \boldsymbol{P}_2')^{-1} \boldsymbol{B}' \boldsymbol{y}, \tag{2.8}$$

where the superscript $(^-)$ denotes the generalized inverse. Note that $\hat{\boldsymbol{\mu}}_+ = (\hat{\boldsymbol{\mu}}', \hat{\boldsymbol{\mu}}_p')'$ is a fitted mean value, i.e. $\hat{\boldsymbol{\mu}}_+ = \widehat{\mathbb{E}[\boldsymbol{y}_+]}$. Hence, in particular the new values $\boldsymbol{y}_p$ are predicted by their fitted mean, where the fit is based on the observed values $\boldsymbol{y}$. Taking formula (2.8) we can derive the expectation of $\hat{\boldsymbol{\mu}}_+$ which for the new values, result to $\mathbb{E}[\hat{\boldsymbol{\mu}}_p] = (\boldsymbol{B}_1 - \boldsymbol{B}_2 \boldsymbol{P}_3^- \boldsymbol{P}_2')(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{P}_1 - \lambda \boldsymbol{P}_2 \boldsymbol{P}_3^- \boldsymbol{P}_2)^{-1} \boldsymbol{B}'\boldsymbol{B}\boldsymbol{\theta}$. This does not simplify to $[\boldsymbol{B}_1 \mid \boldsymbol{B}_2]\boldsymbol{\theta}_+$. Consequently, a bias is induced in the forecast. This bias cannot be determined in size since $\boldsymbol{\theta}_p$ is unknown. Hence, we can not say anything about the quality of the estimate $\hat{\boldsymbol{\mu}}_p$, nor can we derive any confidence interval statements. This is an unsatisfactory result. But we can change the perspective if we link penalized estimation to mixed models. This is done in Section 3.

When the aim of prediction is to forecast future values, the independent error assumption in (2.2) might not always be appropriate. If this is the case, the estimation of the smoothing parameter will be affected by the correlation that is not being accounted for. The methods presented here are easily extended to the case of non i.i.d. errors where smoothing and correlation parameters can be estimated simultaneously (Durbán and Currie (2003)). See also Krivobokoa and Kauermann (2007) for related results in mixed models.

2.1 Properties of the predictions in the case of P-splines with penalties based on differences

When penalties are based on differences between adjacent coefficients, for interpolating, there is no need for new coefficients and the B-spline coefficients form a polynomial sequence of degree $2q-1$ (Eilers and Marx (2010)), for instance when $q=2$, we get cubic interpolation. If we extrapolate instead of interpolate the method above satisfies certain important properties. The subsequent results are based on a basis constructed from equally spaced knots, however, the results extend also to the non-equal spaced knots case, if we define the appropriately scaled penalty matrices. The central results are the following:

i) The fit remains the same regardless of the forecast horizon (the coefficients that yield the fit do not change).

ii) The shape of the forecast is determined by the order of the penalty.

These properties are an immediate consequence of the following theorems.

Theorem 1. *The coefficients from minimizing (2.6) with extended penalty matrix (2.5) satisfy the following properties:*

I. The first c coefficients of $\hat{\boldsymbol{\theta}}_+$*, are those obtained from the fit of* $\boldsymbol{y}$*:* $\hat{\boldsymbol{\theta}}_{+_{1,\ldots,c}} = \hat{\boldsymbol{\theta}}$.

II. The coefficients for the n_p *predicted values are* $\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}}$.

Proof. Substituting the blocks of $\boldsymbol{P}_+$ by their specific values in (2.8) we have that:

$$\hat{\boldsymbol{\mu}}_+ = \boldsymbol{B}_+ \begin{bmatrix} \boldsymbol{I} \\ -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1 \end{bmatrix} (\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{P})^{-1}\boldsymbol{B}'\boldsymbol{y}$$

i.e., the first c coefficients of $\hat{\boldsymbol{\theta}}_+$ are $(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y}$, the same as the ones that give the fit, and the additional coefficients are:

$$\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y} = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}}. \tag{2.9}$$

□

If the knots are not equally-spaced the expression above would be modified since the penalty would have to account for the difference between the knots (Eilers and Marx (2010)).

Corollary 2 (Theorem 1)**.** *Given penalties of order q, the new coefficients are combinations of order $q-1$ of the last q fitted coefficients.*

As the most popular penalties are of second or third order, the proof of the previous corollary for such cases and for penalties of order 1 is given in the complementary material.

In many situations, the fit is not greatly affected by the order of the penalty. However, there is an immediate connection between the penalty (or prior distribution) on the coefficients on the shape of the out-of-sample prediction shown in the above corollary. This is known in the framework of Bayesian P-splines, where the difference penalty corresponds to assuming a random walk prior on the coefficients (see Lang and Brezger (2004)), however it is not common knowledge in a non-Bayesian context. We believe this is an important result that needs to be addressed when using this methodology.

3 Prediction with mixed-effects smooth models

The connection between penalized smoothing and mixed models was established more than thirty years ago in Green (1987) (see also Currie and Durbán (2002) and Wand (2003)). The key point of this equivalence is the fact that the smoothing parameter becomes a ratio of variances, $\lambda = \frac{\sigma^2_\epsilon}{\sigma^2_{\boldsymbol{\alpha}}}$, and both variance components can be estimated through restricted maximum likelihood procedure (REML) (see Patterson and Thompson (1971) or Kauermann (2005)). The idea is extensively discussed in Ruppert et al. (2003). We here exploit the link to mixed models to extend the results of the previous section and derive properties of the predicted value in (2.8).

In order to reparameterize a penalized smooth model it is necessary to find a new basis that allows the representation of model (2.1) as a mixed model. Like before we replace the smooth function by a basis representation which is now written as $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{\epsilon}$. Coefficients $\boldsymbol{\alpha}$ are penalized to achieve smoothness, but the penalty is rephrased as a normal prior on $\boldsymbol{\alpha}$. This leads to a mixed model of the form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha} + \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{G}) \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}), \tag{3.1}$$

where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are the model matrices and $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are the fixed and random effects coefficients respectively. The random effects have the covariance matrix $\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{G}$, which depends on the variance of the random effects $\sigma^2_{\boldsymbol{\alpha}}$, as stated above $\lambda = \frac{\sigma^2_\epsilon}{\sigma^2_{\boldsymbol{\alpha}}}$.

There are different alternatives for the reparameterization of the original smooth model described before which was based on quadratic penalties as a mixed model in (3.1). The idea is to find a transformation $\boldsymbol{\Omega}$ such that:

$$\boldsymbol{B\Omega} = [\boldsymbol{X} \mid \boldsymbol{Z}] \text{ and } \boldsymbol{\Omega}'\boldsymbol{\theta} = [\boldsymbol{\beta}' \mid \boldsymbol{\alpha}']' \text{ to have } \boldsymbol{B\theta} = \boldsymbol{X\beta} + \boldsymbol{Z\alpha},$$

where $\boldsymbol{\Omega}$ is an orthogonal matrix. We split the matrix $\boldsymbol{\Omega}$ into two submatrices (for the fixed and the random components respectively), i.e., $\boldsymbol{\Omega} = [\boldsymbol{\Omega}_f \mid \boldsymbol{\Omega}_r]$, and achieve that $\boldsymbol{X} = \boldsymbol{B}\boldsymbol{\Omega}_f$ and $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{\Omega}_r$. Since the fixed effects are unpenalized, the matrix $\boldsymbol{X}$, may be replaced by any submatrix such that the composed matrix $[\boldsymbol{X} \mid \boldsymbol{Z}]$ has full rank (this also implies that both $\boldsymbol{X}$ and $\boldsymbol{Z}$ have full column rank). For the submatrix $\boldsymbol{\Omega}_r$ there are different alternatives, following the approach of Currie and Durbán (2002), we use the eigenvalue decomposition of the penalty matrix. We here decompose $\boldsymbol{P} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}'$, where $\tilde{\boldsymbol{\Sigma}}$ is a diagonal matrix that contains the eigenvalues of $\boldsymbol{P}$, and $\boldsymbol{U}$ is the corresponding matrix of eigenvectors. This allows to define $\boldsymbol{\Omega}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}^{-1/2}$, where $\boldsymbol{\Sigma}$ contains the positive eigenvalues and $\boldsymbol{U}_r$ contains the span of the decomposition. With this reparametrization, it is straightforward to obtain the relationship between the inverse of the covariance matrix $\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{G}$ of the random effects and the penalty $\boldsymbol{P}$: $\frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{G}^{-1} = \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{\Omega}_r'\boldsymbol{P}\boldsymbol{\Omega}_r$.

Prediction in the context of mixed models has always been done as a two-stage procedure: First fit and then predict. We show how to apply the existing results to the context of smooth mixed models, and then, we will propose an alternative one-stage approach. As we will see the variance-covariance matrix for the random effects in the two-stage approach is a direct extension of the variance-covariance matrix of the random effects in the fit. This implies that the extended mixed model transformation used has to accomplished that. One of the advantages of the one-stage approach is that we can use any transformation.

3.1 Two-stage approaches

3.1.1 Standard methodology for prediction

Once we have established the connection between mixed models and P-splines, we can use the results given in Gilmour et al. (2004) to predict new observations. In this case the prediction is a linear function of the best linear unbiased predictor (BLUP) of random effects and the best linear unbiased estimator (BLUE) of the fixed effects in the model. The results are based on the following augmented mixed model,

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\boldsymbol{\beta} + \boldsymbol{Z}_+\boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \tag{3.2}$$

i.e.:

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}_p \end{bmatrix} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{X}_p \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \boldsymbol{Z} & \boldsymbol{O} \\ \boldsymbol{Z}_1 & \boldsymbol{Z}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\alpha}_p \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon} \\ \boldsymbol{\epsilon}_p \end{bmatrix},$$

with $(\boldsymbol{\epsilon}', \boldsymbol{\epsilon}_p')' \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}_+)$, $\boldsymbol{\beta}$ the fixed effects (the same as the ones that give the fit, the fixed part is linear and therefore no new parameters for the linear part are needed), and $\boldsymbol{\alpha}_+ = (\boldsymbol{\alpha}', \boldsymbol{\alpha}_p')'$ the augmented random effects with covariance matrix

$$\text{Var}[\boldsymbol{\alpha}_+] = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{G}_+ = \sigma^2_{\boldsymbol{\alpha}} \begin{bmatrix} \boldsymbol{G} & \boldsymbol{G}_{op} \\ \boldsymbol{G}_{po} & \boldsymbol{G}_{pp} \end{bmatrix},$$

where $\boldsymbol{G}$ is the covariance matrix of the random effects in the model for the observed data,$\boldsymbol{G}_{op}$ is the covariance matrix between the random effects for the observed data and for the unobserved data and $\boldsymbol{G}_{pp}$ is the covariance matrix of the random effects for the unobserved data. The variance components, $\sigma^2_{\boldsymbol{\epsilon}}$ and $\sigma^2_{\boldsymbol{\alpha}}$, are the ones estimated in the fit through restricted maximum likelihood procedure (REML) (Patterson and Thompson (1971)).

Now we need to formulate the extended P-spline model into the extended mixed model (3.2). For that, we define the extended transformation matrix $\boldsymbol{\Omega}_{r_+} = \begin{bmatrix} \boldsymbol{\Omega}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Omega}_{p_r} \end{bmatrix}$,where $\boldsymbol{\Omega}_r$ is the transformation matrix used for the observed data, and $\boldsymbol{\Omega}_{p_r}$ the one for the predicted values. Hence we have $\boldsymbol{Z}_1 = \boldsymbol{B}_1\boldsymbol{\Omega}_r$ and $\boldsymbol{Z}_2 = \boldsymbol{B}_2\boldsymbol{\Omega}_{p_r}$. There are many ways in which $\boldsymbol{\Omega}_{r_+}$ may be chosen. In the context of penalties based on differences, for simplicity, we chose $\boldsymbol{\Omega}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}^{-1/2}$, based on the eigenvalue decomposition of $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}'$, and $\boldsymbol{\Omega}_{p_r} = \boldsymbol{D}_2^{-1}$, with $\boldsymbol{D}$ and $\boldsymbol{D}_2$ blocks of the extended difference matrix $\boldsymbol{D}_+$, see Equation (2.5). We choose this extended transformation matrix to obtain an extended variance-covariance matrix of random effects that is a direct extension of $\boldsymbol{G}$, the variance-covariance matrix of the random effects in the fit.

Then, the new predicted values are

$$\hat{\boldsymbol{\mu}}_p = \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_{(p)}\hat{\boldsymbol{\alpha}}, \tag{3.3}$$

with $\boldsymbol{Z}_{(p)} = \boldsymbol{Z}_1 + \boldsymbol{Z}_2\boldsymbol{G}_{po}\boldsymbol{G}^{-1}$ and $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\alpha}}$ the BLUE and BLUP, respectively, estimated from the observed data, i.e.

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} = \boldsymbol{Q}^{-1} \begin{bmatrix} \frac{1}{\sigma^2_\epsilon}\boldsymbol{X}' \\ \frac{1}{\sigma^2_\epsilon}\boldsymbol{Z}' \end{bmatrix} \boldsymbol{y}, \tag{3.4}$$

where $\boldsymbol{Q} = \begin{bmatrix} \frac{1}{\sigma^2_\epsilon}\boldsymbol{X}'\boldsymbol{X} & \frac{1}{\sigma^2_\epsilon}\boldsymbol{X}'\boldsymbol{Z} \\ \frac{1}{\sigma^2_\epsilon}\boldsymbol{Z}'\boldsymbol{X} & \frac{1}{\sigma^2_\epsilon}\boldsymbol{Z}'\boldsymbol{Z} + \frac{1}{\sigma^2_\alpha}\boldsymbol{G}^{-1} \end{bmatrix}$.

It follows that the predicted random effects vector for $\boldsymbol{\alpha}_p$ is $\hat{\boldsymbol{\alpha}}_p = \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}}$.

Therefore, the 95% **confidence interval** is

$$\hat{\boldsymbol{\mu}}_p \pm 1.96\sqrt{\text{diag}\left(\begin{bmatrix} \boldsymbol{X}_p \mid \boldsymbol{Z}_{(p)} \end{bmatrix} \boldsymbol{Q}^{-1} \begin{bmatrix} \boldsymbol{X}'_p \\ \boldsymbol{Z}'_{(p)} \end{bmatrix}\right)}, \tag{3.5}$$

with $\boldsymbol{Q}$ defined above and the variance components estimated through restricted maximum likelihood procedure (REML) (see Patterson and Thompson (1971)). We denote this method as "*mixed model approach*".

3.1.2 Prediction based on the conditional distribution of $\boldsymbol{y}_p|\boldsymbol{y}$

The previous method can be seen from the point of view of conditional distributions, which allow us to compute prediction intervals. We rewrite (3.2) as

$$\boldsymbol{y}_+ = \left[\boldsymbol{y}' \mid \boldsymbol{y}_p'\right]' \sim \mathcal{N}(\boldsymbol{X}_+\boldsymbol{\beta}, \boldsymbol{V}_+), \tag{3.6}$$

with $\text{Var}[\boldsymbol{y}_+] = \boldsymbol{V}_+ = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{Z}_+\boldsymbol{G}_+\boldsymbol{Z}'_+ + \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}_+$, the mixed model formulation connects the new values $\boldsymbol{y}_p$ to the observed vector $\boldsymbol{y}$ through a joint normal distribution. It appears therefore natural to predict $\boldsymbol{y}_p$ given $\boldsymbol{y}$ based on the conditional model resulting from (3.6), that is

$$\boldsymbol{y}_p|\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}_p\boldsymbol{\beta} + \boldsymbol{V}_{po}\boldsymbol{V}_{oo}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{V}_{pp} - \boldsymbol{V}_{po}\boldsymbol{V}_{oo}^{-1}\boldsymbol{V}_{op}),$$

where $\boldsymbol{V}_{oo}$, $\boldsymbol{V}_{op}$ and $\boldsymbol{V}_{pp}$ are the submatrices of matrix $\boldsymbol{V}_+$ matching to $\boldsymbol{y}$ and $\boldsymbol{y}_p$. The mean value results through

$$\mathbb{E}[\boldsymbol{y}_p|\boldsymbol{y}] = \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \boldsymbol{V}_{po}\boldsymbol{V}_{oo}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \boldsymbol{Z}_1\hat{\boldsymbol{\alpha}} + \boldsymbol{Z}_2\boldsymbol{G}_{po}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}},$$

which equals (3.3). Note that the first term in the equation above is the result of plugging $\boldsymbol{X}_p$ into the regression equation, and represents the adjustment to this prediction based on the covariance between $\boldsymbol{y}$ and $\boldsymbol{y}_p$. The conditional variance of $\boldsymbol{y}_p|\boldsymbol{y}$ gives the prediction error. This follows since

$$\mathbb{E}_{\boldsymbol{y},\boldsymbol{y}_p}[(\hat{\boldsymbol{\mu}}_p - \boldsymbol{y}_p)^2] = \mathbb{E}_{\boldsymbol{y}}[\mathbb{E}_{\boldsymbol{y}_p}[(\hat{\boldsymbol{\mu}}_p - \boldsymbol{y}_p)^2|\boldsymbol{y}]] = \mathbb{E}_{\boldsymbol{y}}[\text{Var}[\boldsymbol{y}_p|\boldsymbol{y}]] = \text{Var}[\boldsymbol{y}_p|\boldsymbol{y}],$$

where the latter equality holds since in the Normal model the conditional variance does not depend on the value of the variable we condition on. With these results we can construct 95% **prediction interval**:

$$\hat{\boldsymbol{\mu}}_p \pm 1.96\sqrt{\text{Var}[\boldsymbol{y}_p|\boldsymbol{y}]},$$

where observed values are considered as fixed. Notice that, as we have mentioned before, this approach allow us to work out the posterior predictive distribution $\boldsymbol{y}_p|\boldsymbol{y}$ as a Gaussian distribution and therefore compute the prediction intervals. While we can not compute prediction intervals with the standard methodology described in 3.1.1 unless we link it with a Gaussian process.

3.2 One-stage approach

The previous method was a two-stage procedure. First, the model is fitted to the available data and second, based on the fitted model we predict the new observations. As mentioned previously, this approach imposes constraints on the reparametrization used to obtain the smooth mixed model, since the variance-covariance matrix of the extended model (the model that includes out-of-sample prediction) needs to be an extension of the variance-covariance matrix of the original model. Now we propose an alternative approach which can be used with any reparametrization and yields the same results as the two-step approach for appropriate transformations. This approach relates the above results to the method presented in Section 2. We will call it "*extended mixed model approach*", since we include $\boldsymbol{y}_p$ in the model but with infinite variance (zero weight). In this case we consider the model

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\boldsymbol{\beta} + \boldsymbol{Z}_+\boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \quad \boldsymbol{\alpha}_+ \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{G}_+), \quad \boldsymbol{\epsilon}_+ \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{R}_+), \tag{3.7}$$

where $\boldsymbol{R}_+$ is a diagonal weight matrix of dimension $n_+ \times n_+$, with 1 entries if the data is observed, i.e. for $\boldsymbol{y}$, and infinity if the data is considered to be forecasted, i.e. for $\boldsymbol{y}_p$. The quantity infinity expresses that we do not have any information about the data $\boldsymbol{y}_p$. Its estimation is done using the extended mixed model equations of Henderson (1975):

$$\begin{aligned} \hat{\hat{\boldsymbol{\beta}}} &= (\boldsymbol{X}'_+ \tilde{\boldsymbol{V}}^-_+ \boldsymbol{X}_+)^{-1} \boldsymbol{X}'_+ \tilde{\boldsymbol{V}}^-_+ \boldsymbol{y}_+, \\ \hat{\hat{\boldsymbol{\alpha}}}_+ &= \sigma^2_{\boldsymbol{\alpha}} \boldsymbol{G}_+ \boldsymbol{Z}'_+ \tilde{\boldsymbol{V}}^-_+ (\boldsymbol{y}_+ - \boldsymbol{X}_+ \hat{\hat{\boldsymbol{\beta}}}), \end{aligned} \tag{3.8}$$

where $\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}'_p)'$ as in (2.3), $\tilde{\boldsymbol{V}}_+ = \sigma^2_{\boldsymbol{\alpha}} \boldsymbol{Z}_+ \boldsymbol{G}_+ \boldsymbol{Z}'_+ + \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{R}_+$ and by Theorem 18.2.8 given in Harville (2000) $\tilde{\boldsymbol{V}}^-_+ = \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}} \boldsymbol{R}^{-1}_+ - \frac{1}{\sigma^4_{\boldsymbol{\alpha}}} \boldsymbol{R}^{-1}_+ \boldsymbol{Z}_+ (\frac{1}{\sigma^2_{\boldsymbol{\alpha}}} \boldsymbol{G}^{-1}_+ + \boldsymbol{Z}'_+ \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}} \boldsymbol{R}^{-1}_+ \boldsymbol{Z}_+)^{-1} \boldsymbol{Z}'_+ \boldsymbol{R}^{-1}_+$. Finally $\boldsymbol{Z}_+ = \boldsymbol{B}_+ \boldsymbol{\Omega}_{+_r}$, with $\boldsymbol{\Omega}_{+_r}$ any orthogonal transformation such that

$$\boldsymbol{B}_+ \left[\boldsymbol{\Omega}_{+_f} \mid \boldsymbol{\Omega}_{+_r} \right] = [\boldsymbol{X}_+ \mid \boldsymbol{Z}_+] \text{ and } \left[\boldsymbol{\Omega}_{+_f} \mid \boldsymbol{\Omega}_{+_r} \right]' \boldsymbol{\theta}_+ = \left[\boldsymbol{\beta}' \mid \boldsymbol{\alpha}'_+ \right]',$$

for the particular case of penalties based on differences we choose $\boldsymbol{\Omega}_{+_r}$ based on the eigenvalue decomposition of $\boldsymbol{D}'_+ \boldsymbol{D}_+$. This yields $\boldsymbol{B}_+ \boldsymbol{\theta}_+ = \boldsymbol{X}_+ \boldsymbol{\beta} + \boldsymbol{Z}_+ \boldsymbol{\alpha}_+$.
We use the double hat symbol ($\hat{\hat{}}$) here to remark that in this case the estimation is based on the extended version assuming that the unknown values $\boldsymbol{y}_p$ have infinity variance. Notice that to compute the fixed and random effects we do not need $\boldsymbol{R}_+$, we need its inverse, $\boldsymbol{R}^{-1}_+$, with 1 entries if the data is observed and 0 if the data is considered to be predicted.

Like above we have $\widehat{\widehat{\boldsymbol{\alpha}}}_+ = \widehat{\widehat{\mathbb{E}[\boldsymbol{\alpha}_+ | \boldsymbol{y}]}}$ with the double hat on the expectation denoting that $\boldsymbol{\beta}$ has been estimated. It is easy to see that taking expectation over $\boldsymbol{y}$ (or $\boldsymbol{y}_+$) we get $\mathbb{E}[\hat{\hat{\boldsymbol{\alpha}}}_+] = \boldsymbol{0}$ so that the estimate is unbiased in the mixed model.

As we mentioned earlier, the above extension of the missing value approach to the mixed model framework fits and predicts simultaneously while the approach of

Gilmour et al. (2004) is a two-stage method. In order to know the relationship between the two methods, we need to know the relationship between the covariance matrix of the random effects that gives the fit and the extended covariance matrix. This is shown in the following theorem.

Theorem 3. *Given model in (2.1) with penalty based on differences between adjacent coefficients, the fit and the prediction of new observations given by extended mixed model approach and mixed model approach are the same if the transformation matrix used to obtain the model components in (3.8) is the direct extension of the original transformation,*

$$\boldsymbol{\Omega}_{+_r} = \begin{bmatrix} \boldsymbol{\Omega}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Omega}_{p_r} \end{bmatrix}, \tag{3.9}$$

Under the previous hypothesis the variance components ($\sigma^2_{\boldsymbol{\alpha}}$, $\sigma^2_{\boldsymbol{\epsilon}}$) that maximize the REML criteria, l, and the REML criteria corresponding to the extended weighted model, l_+, are equal,

$$l(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}}) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)'\boldsymbol{V}^{-1}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right), \tag{3.10}$$

$$l_+(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}}) = -\frac{1}{2}\log|\tilde{\boldsymbol{V}}_+| - \frac{1}{2}\log|\boldsymbol{X}'_+\tilde{\boldsymbol{V}}^-_+\boldsymbol{X}_+| - \frac{1}{2}\left(\boldsymbol{y}_+ - \boldsymbol{X}_+\boldsymbol{\beta}\right)'\tilde{\boldsymbol{V}}^-_+\left(\boldsymbol{y}_+ - \boldsymbol{X}_+\boldsymbol{\beta}\right). \tag{3.11}$$

Equivalent results would be obtained using maximum likelihood (ML).

The proof can be found in the complementary material. Since the fitted and predicted values in both approaches do not depend on the transformations used, the previous theorem is stating a stronger result: the approaches always give the same solution, regardless of the used transformations. We have stated the theorem for a particular transformation to established the relationship between both methods.

The last statement of the previous theorem means that the variance parameters used to predict are the same as the ones used for estimating the original fit. In other words prediction in- and out-of-sample can not only be done simultaneously, but also the optimal smoothing parameter is the same.

4 Applications

In this section, we apply the proposed methods to two real data sets. The first one, allows us to show an example of predicting within the framework of additive models and in the case when out of sample prediction is needed to the left and right of the interval where the covariate is observed. The second dataset illustrates a classic example were forecasting is needed, when data are collected over time. Although, as we have shown all the methods give us the same result, in order to obtain the prediction intervals we use the two-stage approach.

4.1 Prediction of aboveground biomass

All the results presented in the previous sections are obtained in the case of smooth models with a single covariate. However, it is immediate to extend these results to the case of semiparametric or additive models. In this section, we apply the proposed methodology to such data set. The data set corresponds to an agricultural trial carried out in Spain (Rivas-Martínez et al. (2002)) with the aim of evaluating the economic viability of *Populus* trees prior to harvesting. In this context, it is essential to estimate aboveground biomass and obtain accurate predictions using only a minimum

set of easily obtainable information i.e., diameter and height. Sánchez-González et al. (2016) proposed the use of smooth additive mixed models for predicting aboveground biomass. Here we analyze data for a single clone of the nine included in the trials. The aim is to estimate the productions (measured as aboveground dry biomass) as a function of diameter and height and give out-of-sample predictions. The observed data consists of 315 observations, for diameter values measured at 1.30 m breast height. This data is illustrated in Figure 2.

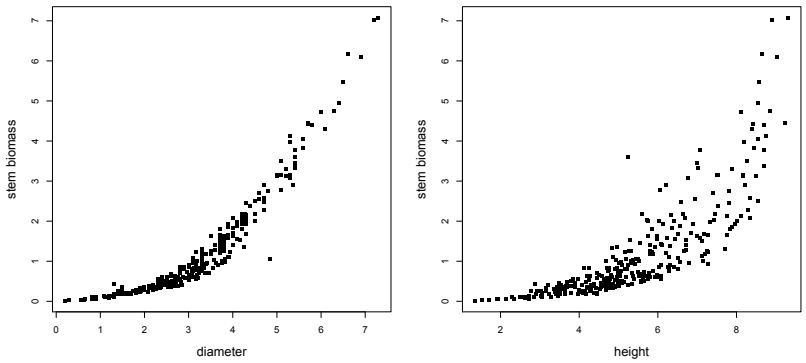


Figure 2: Plot of weight versus diameter (left panel) and plot of weight versus height (right panel).

From the plot it is immediate to see that the weight is a non-linear function of diameter and height. Therefore we fit the following model:

$$\mathbb{E}[y_i|x_i, y_i] = f(x_i) + f(z_i), \tag{4.1}$$

where $\boldsymbol{x}$ is the diameter and $\boldsymbol{z}$ is the height, i.e. $f(\boldsymbol{x})$ and $f(\boldsymbol{z})$ are the functions that represent the main effects of the diameter and of the height, respectively. The regression matrix is then defined by blocks as $\boldsymbol{B} = [\boldsymbol{B}_{\boldsymbol{x}} \mid \boldsymbol{B}_{\boldsymbol{z}}]$ with marginal B-spline bases of degree three of the covariates diameter and height, $\boldsymbol{B}_{\boldsymbol{x}}$ and $\boldsymbol{B}_{\boldsymbol{z}}$, respectively. The penalty matrix associated with model (4.1) has a block-diagonal form: $\boldsymbol{P} =$

$blockdiag(\lambda_{\boldsymbol{x}}\boldsymbol{P}_{\boldsymbol{x}}, \lambda_{\boldsymbol{z}}\boldsymbol{P}_{\boldsymbol{z}})$, where $\boldsymbol{P}_{\boldsymbol{x}}$ and $\boldsymbol{P}_{\boldsymbol{z}}$ are the marginal second-order difference penalties for diameter and height.

We predict weight for 6 new out-of-sample values for diameter and 15 new values for height where 5 are to the left and 10 to the right of the range of the observed height values. Applying the previous methodology we extend the basis and the penalty: $\boldsymbol{B}_+ = \left[\boldsymbol{B}_{\boldsymbol{x}_+} \mid \boldsymbol{B}_{\boldsymbol{z}_+}\right]$, $\boldsymbol{P}_+ = blockdiag(\lambda_{\boldsymbol{x}}\boldsymbol{P}_{\boldsymbol{x}_+}, \lambda_{\boldsymbol{z}}\boldsymbol{P}_{\boldsymbol{z}_+})$, with $\boldsymbol{B}_{\boldsymbol{x}_+} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{x}} & \boldsymbol{O} \\ \boldsymbol{B}_{\boldsymbol{x}_{(1)}} & \boldsymbol{B}_{\boldsymbol{x}_{(2)}} \end{bmatrix}$ and $\boldsymbol{B}_{\boldsymbol{z}_+} = \begin{bmatrix} \boldsymbol{B}_{\boldsymbol{z}_{(0)}} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B}_{\boldsymbol{z}} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{B}_{\boldsymbol{z}_{(1)}} & \boldsymbol{B}_{\boldsymbol{z}_{(2)}} \end{bmatrix}$.

Once we have extended the basis and the penalty, it is straightforward to obtain the fit and the prediction applying Equation (2.7), but in order to obtain confidence and prediction intervals and to avoid identifiability problems, since the column of $\mathbf{1}$'s is contained in the space spanned by the columns of $\boldsymbol{B}_{\boldsymbol{x}_+}$ and $\boldsymbol{B}_{\boldsymbol{z}_+}$, we reparameterize the model using the representation of a penalized spline model as a mixed model. Figure 3 shows the smooth fitted and predicted trend for diameter and height, the 95% confidence interval (grey line) and the 95% prediction interval (dashed lines). Notice that the prediction is done backward and forward, the proposed methodology allows us to obtain the prediction for any range of the independent variable. This could not be done by using methods developed in the series temporal framework.

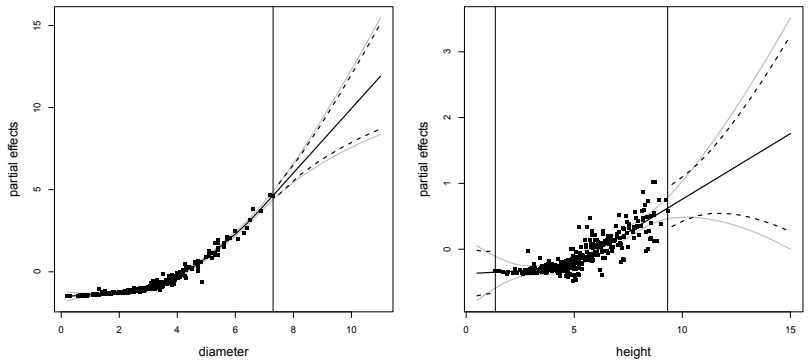


Figure 3: Fit, forecast, 95% confidence interval (grey lines) and 95% prediction interval (dashed lines) of the additive smooth term for diameter (left panel) and for height (right panel), result of applying the methodology of a data set on the stem biomass.

The estimation of the covariance parameters was carried out by the REML through the SOP algorithm, proposed in Rodríguez-Álvarez et al. (2018).

4.2 Forecasting SO_2 concentration levels

Now, we analyze data where out-of-sample prediction is needed and missing observations are present in the data. In this situation the one-stage approach offers an elegant solution to solve simultaneously both problems. We consider measurements on sulphur dioxide (SO_2) concentration levels (in $\mu g/m^3$) over station AT02 from January 1990 to December 2001, Figure 4 shows the data set, in which we can see that there are some missing observations between October 1995 and March 1999.

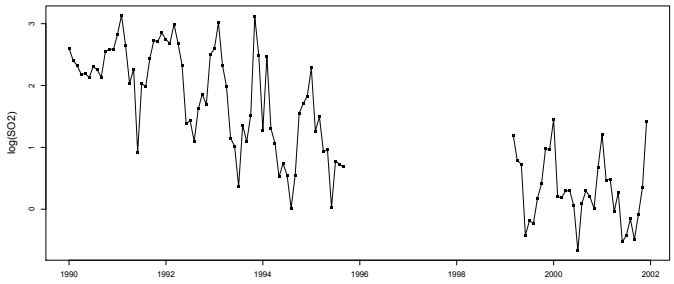


Figure 4: Time series plot of $log(SO_2)$ data for station AT02.

The data were collected through the 'European monitoring and evaluation programme' (EMEP) under the Co-operative Programme for Monitoring and Evaluation of the Long-range Transmission of Air Pollutants in Europe (see further information available at `htpp:\\www.emep.int`). Notice that there is a clear evidence of temporal trends and seasonal effects and that there are some missing observations (mostly due to equipment failure, replacement or calibration), there is a big gap between October 1995 and March 1999.

First of all, let us introduce the smooth modulation model based on P-splines suggested by Eilers et al. (2008):

$$\mathbb{E}\left[y_i|\, x_i\right] = f(x_i) + \sum_{j=1}^{J} \{g_j(x_i)\cos(j\omega x_i) + h_j(x_i)\sin(j\omega x_i)\}, \tag{4.2}$$

where $f(\cdot)$ accounts for the smooth trend, $g(\cdot)$ and $h(\cdot)$ are smooth functions that describe the local amplitudes of cosine and sine waves, and $\omega = 2\pi/p$, with p the period (e.g. $p = 12$ for monthly data). The number of harmonics functions, J, required for the seasonal component is usually taken as 1 or 2 to reduce the number of parameters to be estimated. If $J = 1$ the model (4.2) can be written in matrix

form as:

$$\boldsymbol{y} = \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{R}), \tag{4.3}$$

where $\boldsymbol{R}$ is the covariance matrix of the error, we work with uncorrelated i.i.d. errors, i.e., $\boldsymbol{R} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}$; and $\breve{\boldsymbol{B}}$ is the regression matrix, $\breve{\boldsymbol{B}} = [\boldsymbol{B} \mid \boldsymbol{CB} \mid \boldsymbol{SB}]$, where $\boldsymbol{B}$ is a B-spline basis, $\boldsymbol{C} = \text{diag}\{\cos(\omega \boldsymbol{x})\}$ and $\boldsymbol{S} = \text{diag}\{\sin(\omega \boldsymbol{x})\}$.

If we want to obtain the coefficients $\breve{\boldsymbol{\theta}} = (\boldsymbol{\theta}, \boldsymbol{\theta}_c, \boldsymbol{\theta}_s)$ in (4.3) with the P-splines method we have to minimize the following function of $\breve{\boldsymbol{\theta}}$:

$$S = (\boldsymbol{y} - \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}})'(\boldsymbol{y} - \breve{\boldsymbol{B}}\breve{\boldsymbol{\theta}}) + \breve{\boldsymbol{\theta}}'\breve{\boldsymbol{P}}\breve{\boldsymbol{\theta}}, \tag{4.4}$$

where $\breve{\boldsymbol{P}} = \breve{\boldsymbol{\lambda}} \otimes \breve{\boldsymbol{D}}'\breve{\boldsymbol{D}}$, with $\breve{\boldsymbol{\lambda}} = diag(\lambda, \lambda_c, \lambda_s)$ and $\breve{\boldsymbol{D}} = \text{blockdiag}(\boldsymbol{D}'_q\boldsymbol{D}'_q, \boldsymbol{I}_2 \otimes \boldsymbol{D}'_{q_{cs}}\boldsymbol{D}_{q_{cs}})$, q and q_{cs} are usually 2 and 1, respectively.

To obtain the fit and the forecast simultaneously, we just have to extend the B-spline basis for the trend and the modulation components $\breve{\boldsymbol{B}}_+ = [\boldsymbol{B}_+ \mid \boldsymbol{C}_+\boldsymbol{B}_+ \mid \boldsymbol{S}_+\boldsymbol{B}_+]$, where $\boldsymbol{C}_+ = \text{diag}(\cos(\omega \boldsymbol{x}_+))$ and $\boldsymbol{S}_+ = \text{diag}(\sin(\omega \boldsymbol{x}_+))$, for the additive modulation blocks $\boldsymbol{C}_+\boldsymbol{B}_+$ and $\boldsymbol{S}_+\boldsymbol{B}_+$, and consider the following penalty matrix:

$$\breve{\boldsymbol{P}}_+ = \text{blockdiag}(\lambda \boldsymbol{D}_{+q}'\boldsymbol{D}_{+q}, \lambda_c \boldsymbol{D}_{+q_c}'\boldsymbol{D}_{+q_c}, \lambda_s \boldsymbol{D}_{+q_s}'\boldsymbol{D}_{+q_s}).$$

Once $\breve{\boldsymbol{B}}_+$ and $\breve{\boldsymbol{P}}_+$ are computed, $\breve{\boldsymbol{\theta}}_+$ can be easily computed through the formula in (2.7), by using $\breve{\boldsymbol{B}}_+$ instead of $\boldsymbol{B}_+$ and $\breve{\boldsymbol{P}}_+$ instead of $\lambda \boldsymbol{P}_+$.

Keeping in mind that in a penalized spline modulation model, we have the trend and the seasonality components, it is straightforward to represent it as a mixed model.

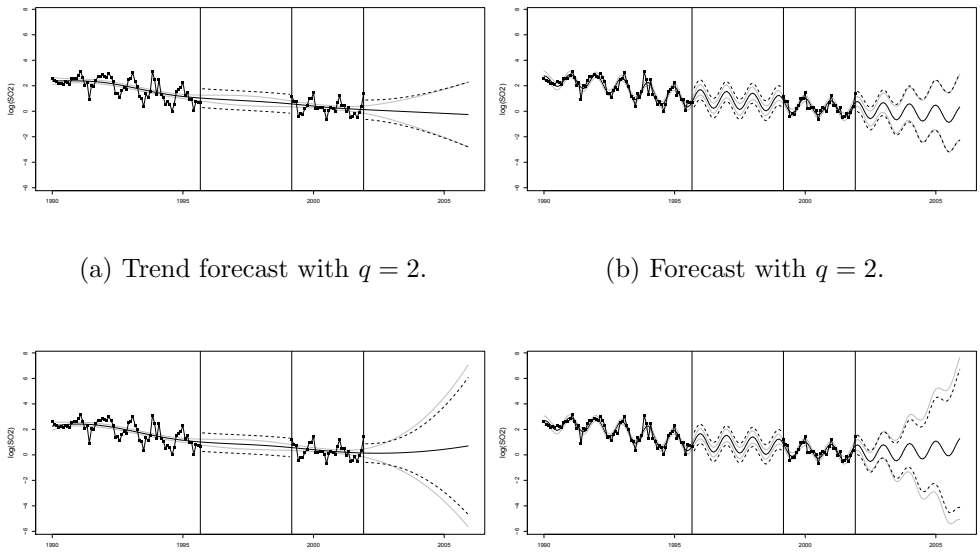
As we have seen the order of the penalty q denotes the $q-1$ grade polynomial of the fixed part when the smoothing parameters are very large (the null model), and

hence, the random part can be considered as smooth deviates from the null model. In model (4.3) it is considered a first order penalty for the modulation terms, then the null terms for the modulation are $\cos(\omega \boldsymbol{x})$ and $\sin(\omega \boldsymbol{x})$. For instance, for $J = 1$ in (4.2) the fixed effect matrix for the smooth modulation model (4.3) is a design matrix of a harmonic regression model: $\breve{\boldsymbol{X}} = [\mathbf{1}_n \mid \boldsymbol{x} \mid \boldsymbol{x}^2 \mid ... \mid \boldsymbol{x}^{q-1} \mid \cos(\omega \boldsymbol{x}) \mid \sin(\omega \boldsymbol{x})]$.

The matrix of the random component, $\breve{\boldsymbol{Z}}$, is a block matrix: $\breve{\boldsymbol{Z}} = [\boldsymbol{Z} \mid \boldsymbol{C}\boldsymbol{Z}_{cs} \mid \boldsymbol{S}\boldsymbol{Z}_{cs}]$, where $\boldsymbol{Z} = \boldsymbol{B}\boldsymbol{\Omega}_r$ with $\boldsymbol{\Omega}_r = \boldsymbol{U}_r \tilde{\boldsymbol{\Sigma}}^{-1/2}$ ($\boldsymbol{U}_r$ and $\tilde{\boldsymbol{\Sigma}}$ are obtained from the eigenvalue decomposition of $\boldsymbol{D}_q' \boldsymbol{D}_q$, with $\boldsymbol{D}_q$ a penalty matrix of order usually $q = 2$) and $\boldsymbol{Z}_{cs} = \boldsymbol{B}\breve{\boldsymbol{\Omega}}$, with $\breve{\boldsymbol{\Omega}} = \breve{\boldsymbol{U}}_r \breve{\tilde{\boldsymbol{\Sigma}}}^{-1/2}$ ($\breve{\boldsymbol{U}}_r$ and $\breve{\tilde{\boldsymbol{\Sigma}}}$ are obtained from the eigenvalue decomposition of $\boldsymbol{D}_{q_{cs}}' \boldsymbol{D}_{q_{cs}}$, with $\boldsymbol{D}_{q_{cs}}$ a penalty matrix of order usually $q_{cs} = 1$).

The smoothing parameters λ, λ_c, λ_s in (4.4) become the ratio between the error variance term and the random effect variances σ^2, σ_c^2 and σ_s^2 for the trend and modulation terms respectively. The covariance matrix is a block diagonal matrix, $\breve{\boldsymbol{G}} = \text{blockdiag}(\boldsymbol{G}, \boldsymbol{G}_c, \boldsymbol{G}_s)$, where $\boldsymbol{G} = \sigma^2 \boldsymbol{I}_{c-q}$, $\boldsymbol{G}_c = \sigma_c^2 \boldsymbol{I}_{c-1}$ and $\boldsymbol{G}_s = \sigma_s^2 \boldsymbol{I}_{c-1}$.

The one-stage approach can deal with missing observations within and out of sample just by setting to zero the corresponding diagonal entries of $\boldsymbol{R}_+^{-1}$. Figure 5 shows the forecasted trend and final predictions (including the seasonal projections) for AT02 station with second and third penalty orders for the trend component. Notice that, as it is stated in Corollary 2, for second order penalty ($q = 2$) the trend forecast is linear, and for third order ($q = 3$) the trend forecast is quadratic.



(a) Trend forecast with $q = 2$.

(b) Forecast with $q = 2$.

(c) Trend forecast with $q = 3$.

(d) Forecast with $q = 3$.

Figure 5: Forecast for AT02 station. Top and bottom left figures show the data (points), the fitted and forecasted trend (black line), 95% confidence interval (grey lines) and 95% prediction interval (dashed lines) for second and third order penalties, respectively. Top and bottom right figures show the data (points), the fit and the forecast in the modulation model (black line), 95% confidence interval (grey lines) and 95% prediction interval (dashed lines) for second and third order penalties, respectively.

5 Conclusions

Smoothing techniques have become a very popular tool for describing patterns in noisy data. However, prediction is still an open area of research. In this paper we have proposed a general framework for prediction of new observations in penalized regression, the methodology proposed can be accommodated to the different frame-

works in which smoothing is carried out, regardless of the basis and quadratic penalty used:

i) Extend the basis used for regression and the penalty to control the smoothness in the framework of penalized regression based on quadratic penalties.

ii) Extend the fixed and random components in the context of mixed models and carry out the prediction as a two-stage procedure or using a new approach that allow us to fit and predict simultaneously.

The approach given in Section 3.1.2 is a particular case of prediction in the context of Gaussian process regression, under the assumption that the regression function is a realization of a stochastic process. Then, prediction of new values are based on Gaussian process prior and a P-spline covariance matrix. Hence, we can use the flexibility of the Gaussian process and the choice of suitable covariance matrix to model any non linear model non parametrically. In this context, prediction is straightforward due to the properties of Gaussian processes.

In the context of penalties based on differences between adjacent coefficients, we have proved that all the approaches compared in this work are equivalent, but the one-stage approach can be used regardless of the mixed model reparameterization used. Furthermore, when predicting out-of-sample observations, the horizon of prediction has no effect on the fit or on the selection of the smoothing parameter.

These results are based on basis constructed from equally spaced knots, however, results apply also to the non-equal spaced knots case, defining the appropriate scaled penalty matrices.

To illustrate the methodology and the proved results, we have analyzed two real data sets and showed the performance of the extended mixed model approach with

B-spline basis and penalties based on differences. These examples illustrate how to obtain out-of-sample prediction in the case of additive models. In the first example we show how the method can do backward and forward prediction and in the second example, we show how the one-stage approach can deal with out-of-sample and missing value prediction simultaneously.

The results obtained here have opened new areas of research, in particular, we have already obtained similar results in the case of multidimensional smoothing and preliminary work is being done to extend this work to the case of non-Gaussian data.

6 Supplementary materials

Supplementary materials for this paper [you may specify more details] are available from http://www.statmod.org/smij/archive.html.

Acknowledgements

The first and the second authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness MTM2014-52184-P. The fourth author acknowledges the financial support by the Basque Government through the BERC 2018-2021 program and by Spanish Ministry of Economy and Competitiveness MINECO through BCAM Severo Ochoa excellence accreditation SEV-2013-0323 and through project MTM2017-82379-R funded by (AEI/FEDER, UE) and acronym "AFTERAM".

References

Ba, A., Sinn, M., Goude, Y., and Pompey, P. (2012). Adaptive learning of smoothing functions: Application to electricity load forecasting. In Pereira, F., Burges, C., Bottou, L., and Weinberger., K., editors, *Advances in Neural Information Processing Systems 25*.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic system. *Statistical Science*, **10**(1), 3–66.

Cressie, N. A. C. (1993). *Statistics for Spatial data*. Wiley: New York.

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

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

De Boor, C. (1972). On calculating with b-splines. *Journal of Approximation theory*, **6**(1), 50–62.

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

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

Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots, and penalties. *Computational Statistics*, **2**(6), 637–653.

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

Gilmour, A., Cullis, B., Welham, S., Gogel, B., and Thompson, R. (2004). An efficient computing strategy for prediction in mixed linear models. *Computational Statistics and Data Analysis*, **44**, 571–586.

Green, P. J. (1987). Penalized likelihod for general semi-parametric regression models. *International Statistical Review*, **55**(3), 245–259.

Harville, D. (2000). *Matrix Algebra from a Statistician's Perspective*. Springer.

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

Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing*. Springer Series in Statistics.

Jones, D. R., Schonlau, M., and William, J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimiztion*, **13**, 455–492.

Kauermann, G. (2005). A note on smoothing parameter selection for penalised spline smoothing. *Journal of Statistical Planing and Inference*, **127**, 53–69.

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

Lang, S. and Brezger, A. (2004). Bayesian p-splines. *Journal of computational and graphical statistics*, **13**(1), 183–212.

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

Rivas-Martínez, D., Díaz, T., Fernández-González, F., Izco, J., Loidi, J., Lousã, M., and Penas, A. (2002). Vascular plant communities of spain and portugal. addenda to the syntaxonomical checklist of 2001. *Itinera Geobotanica*, pages 15, 1–2, 5–22.

Rodríguez-Álvarez, M., Lee, D.-J., Kneib, T., Durbán, M., and Eilers, P. (2018). On the estimation of variance parameters in non-standard generalised linear mixed models: application to penalised smoothing. *Statistics and Computing*, **29**, 1–18.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman & Hall.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Series in Statistical and Probabilistic Mathematics. Cambridge University Press, UK.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, **4**(4), 409–435.

Sánchez-González, M., Durbán, M., Lee, D., Cañellas, I., and Sixto, H. (2016). Smooth additive mixed models for predicting aboveground biomass. *Journal of Agricultural, Biological and Environmental Statistics*, **22**, 23–41.

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