UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
17-11
ISSN 2387-0303
Mayo 2017

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91 624-98-48

# A general framework for prediction in penalized regression

Alba Carballo[a], María Durbán[a], Dae-Jin Lee[b]

**Abstract**

We present several methods for prediction of new observations in penalized regression using different methodologies, based on the methods proposed in: i) Currie et al. (2004), ii) Gilmour et al. (2004) and iii) Sacks et al. (1989). We extend the method introduced by Currie et al. (2004) to consider the prediction of new observations in the mixed model framework. In the context of penalties based on differences between adjacent coefficients (Eilers & Marx (1996)), the equivalence of the different methods is shown. We demonstrate several properties of the new coefficients in terms of the order of the penalty. We also introduce the concept memory of a P-spline, this new idea gives us information on how much past information we are using to predict. The methodology and the concept of memory of a P-spline are illustrated with two real data sets, one of them on the yearly mortality rates of Spanish men and other on rental prices.

***Keywords:*** Prediction, Penalized regression, P-splines, Mixed Models

[a] Department of Statistics, Universidad Carlos III de Madrid, Spain.
[b] BCAM - Basque Center for Applied Mathematics, Bilbao, Spain.

Acknowledgements: The first and the second authors acknowledge financial support from the Spanish Ministry of Economy and Competitiveness MTM2014-52184-P. The third author acknowledges financial support from the Basque Government through the BERC 2014-2017 program and by the Spanish Ministry of Economy and Competitiveness MINECO: BCAM Severo Ochoa excellence accreditation SEV-2013-0323.

# A general framework for prediction in penalized regression

Alba Carballo[a], María Durbán[a], Dae-Jin Lee[b]

[a]*Department of Statistics, Universidad Carllos III de Madrid, Spain*
[b]*Basque Center for Applied Mathematics, Bilbao, Spain*

May 2017

**Abstract**

We present several methods for prediction of new observations in penalized regression using different methodologies based on the methods proposed in: i) Currie *et al.* (2004), ii) Gilmour *et al.* (2004) and iii) Sacks *et al.* (1989). We extend the method introduced by Currie *et al.* (2004) to consider the prediction of new observations in the mixed model framework. In the context of penalties based on differences between adjacent coefficients (Eilers & Marx (1996)), the equivalence of the different methods is shown. We demonstrate several properties of the new coefficients in terms of the order of the penalty. We also introduce the concept of memory of a P-spline, this new idea gives us information on how much past information we are using to predict. The methodology and the concept of memory of a P-spline are illustrated with two real data sets, one of them on the yearly mortality rates of Spanish men and other on the rental prices.

*Keywords:* Prediction, Penalized regression, P-splines, Mixed models

## 1 Introduction

There are many situations in which prediction of new observations in the context of smoothing models is needed, for example hourly temperatures at a weather station or yearly number of deaths. This can have a major impact in areas such as demography (mortality tables), epidemiology, particularly in "disease mapping" or environmental sciences. A graph of data often exhibits patterns, such as an upward or downward moment, trend, or a pattern that repeats, seasonal variation, both might be used to predict new values. These are some 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 & Marx (1996), can be extended to smooth and predict two-dimensional mortality tables, showing how to construct the appropriate regression bases and penalty matrices for forecasting.

Most of the existing literature in this area is related to the prediction of new observations in a temporal context, i.e. forecast of new observations. Let us start by doing a brief review of the main literature related to forecasting in smoothing models by commenting the main approaches of Ba *et al.* (2012), Caudel & Frey (2012) and Sacks *et al.* (1989). We also can find in Hyndman *et al.* (2008) an overview of

exponential smoothing methods.

The exponential smoothing family is used to name 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. In Hyndman *et al.* (2008), extensive information about exponential smoothing methods can be found. A brief summary of the exponential smoothing history allows us to know the great importance of this kind of models, they are since 1950 the most popular forecasting methods used in business and industry. They consider a time series as a combination of various components such as the trend (T), cycle (C) (a pattern that repeats with some regularity but with unknown and changing periodicity), seasonal (S), and irregular or error (E) (the unpredictable component of the series) components, and distinguish the models depending on the combination of the components. For instance,

- A *purely additive* model has the form: $\boldsymbol{y} = \mathrm{T} + \mathrm{S} + \mathrm{E}$, the three components are added together to form the observed series.
- A *purely multiplicative* model is written as: $\boldsymbol{y} = \mathrm{T} \times \mathrm{S} \times \mathrm{E}$, the data are formed as the product of the three components.

The trend component is, in turn, a combination of a level term and a growth term. Once a trend component is chosen, they, optionally, introduce a seasonal component, either additively or multiplicatively. Taking into account all possible combinations 15 methods are obtained, and for each of them there are two possible state space models, one corresponding to a model with additive errors and the other to a model with multiplicative errors. If the same parameter values are used, these two models give equivalent point forecasts although different prediction intervals.

Other authors such us Ba *et al.* (2012) use penalized splines to fit and forecast time series data and want to minimize

$$S = (\boldsymbol{y} - \boldsymbol{B\theta})'\boldsymbol{M}(\boldsymbol{y} - \boldsymbol{B\theta}) + \boldsymbol{\theta}'\boldsymbol{P\theta}, \tag{1}$$

where $\boldsymbol{y}$ is the observed response, $\boldsymbol{\theta}$ are the coefficients, $\boldsymbol{B}$ is a spline basis that covers the hole range of the extended explanatory variable and $\boldsymbol{M}$ is a weight matrix that puts different weights on the samples, instead of what Currie *et al.* (2004) do, who give the same weight to each sample. It means, in this case,

$$\boldsymbol{M} = \mathrm{diag}(\omega^{n-1}, \omega^{n-2}, ..., \omega^2, \omega, 1), \tag{2}$$

with $n$ the size of the sample and the *forgetting factor*, $\omega$, in the interval $(0, 1]$. Therefore, $\boldsymbol{M}$ puts exponentially decreasing weights on the samples, according to the order of their arrival. They use the following penalty matrix:

$$\boldsymbol{P} = \mathrm{diag}(\gamma, \gamma, ..., \gamma), \tag{3}$$

with $\gamma > 0$.
To get the fit and forecast they propose two algorithms, both of them start with

initial values for the coefficients and run until the estimated coefficients can be used to compute predictions for new given covariates, following a certain criteria. With one of them the obtained solution is the same as in (1) with $\boldsymbol{M}$ as in (2) for a fixed value of $\omega$ and $\boldsymbol{P}$ as in (3).
In the second proposed algorithm the forgetting factor is not fixed, they use adaptive forgetting factors in order to improve the stability and the tracking behaviour. The basic idea is to choose large forgetting factor during stationary regimes, and small forgetting factors during transient phases. To do it they update the forgetting factor according to a particular formula with the objective of minimizing the expected value of the a priori errors.

In Caudel & Frey (2012) P-splines are also used to fit and forecast time series, their main approach is to update the P-spline as new time series observations arrive, i.e., once a new observation is predicted the P-spline is updated; they have shown that large change in the fitted values of the P-spline caused by new time series data rarely occurs past the fourth knot going backward in time. Because of this, they introduced the concept of *stitching*, i.e., several P-splines are constructed and then combined to make one long continuous P-spline, to stitch the P-splines they force the fitted response values to be equal in the points of window breaks. The main benefit of the construction of windowed P-splines is the use of multiple smoothing parameters, and therefore, take into account each state of the time series. Stitching, they obtain a fit that closely resembles the time series data at each moment and that is unaffected by time historical time series data that occurred in the distant past. It is notable that stitching is also a substantial computational benefit over complete P-spline estimation method.

In the framework of global optimization several authors, Sacks *et al.* (1989) and Jones *et al.* (1998), fit a stochastic process to data and predict at a new point computing the function value that is most consistent with the estimated typical behavior. They treat the observations as if they were generated by a constant and an error that is a stochastic process. Considering the error as a stochastic process and relating the correlation between errors to the distance between the corresponding points allow them to obtain a smooth function with a simple constant as the regression terms. Their approach is called Bayesian global optimization and the concept is the same as the idea behind the well-know technique in spatial statistics called kriging (Cressie (1993)).

Taking into account the previous ideas, and in order to give a general framework for prediction in penalized regression and to delve deeper into our knowledge of the forecasting method proposed in Currie *et al.* (2004), we have organized the remaining of the paper as follows. Section 2 is dedicated to introduce the fundamentals of three different methods that we can use to predict with smooth models, they have been proposed in Currie *et al.* (2004), Gilmour *et al.* (2004) and Sacks *et al.* (1989). We also extend the proposal of Currie *et al.* (2004) to the mixed model framework and show the equivalence of all methods in the particular case of penalties based on differences between adjacent coefficients (Eilers & Marx (1996)).

In section 3 we follow Currie *et al.* (2004) and show that the order of the penalty determines the shape of the predictions. The proposed methodology is illustrated in Section 4 with the analysis of two real data sets. In the fifth section, we introduce the concept of memory of a P-spline as a tool to know the weight of each observation on the prediction when we use the method proposed in Currie *et al.* (2004). Finally, concluding remarks are made in Section 6.

## 2 Prediction in penalized regression

In this section we present three different methodologies (penalized regression with a quadratic penalty, linear mixed models and stochastic processes) that allow us to estimate, nonparametrically, a smooth curve and to predict new observations. To begin with we are going to give a brief revision of the penalized regression.

Consider the case of a univariate Gaussian data, with response variable $\boldsymbol{y}$ and regressor $\boldsymbol{x}$, the smooth model is of the form:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{I}), \tag{4}$$

where $f(\cdot)$ is an unknown smooth function, it is estimated from the data points $(\boldsymbol{x}_i, \boldsymbol{y}_i)$, for $i = 1, ..., n$, and $\boldsymbol{\epsilon}$ are independent and identically distributed errors with variance $\sigma^2_{\boldsymbol{\epsilon}}$. Let us see how the different methodologies that allow us to obtain $f(\boldsymbol{x})$ and to predict new observations.

The model (4) can be written in matrix form:

$$\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\epsilon}} \boldsymbol{I}), \tag{5}$$

where $\boldsymbol{B}$ is a regression basis constructed from the covariate $\boldsymbol{x}$, such that, $\boldsymbol{B} = \boldsymbol{B}(\boldsymbol{x})$, and $\boldsymbol{\theta}$ is the vector of regression coefficients. Rather than estimating the coefficients $\boldsymbol{\theta}$ in (5) by simple maximum likelihood methods we penalize the coefficients through a quadratic penalty, i.e., the fit is:

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\theta}} = \boldsymbol{B}(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{P})^{-1}\boldsymbol{B}'\boldsymbol{y}, \tag{6}$$

where $\boldsymbol{P}$ is any quadratic penalty that forces the coefficients to vary smoothly, and consequently to obtain a smoothed curve and $\lambda$ a smoothing parameter. There are several alternatives for the choice of the regression basis $\boldsymbol{B}$ and the penalty matrix $\boldsymbol{P}$, like Eilers & Marx (1996), we use a B-spline basis and $\boldsymbol{P} = \boldsymbol{D}'_q \boldsymbol{D}_q$, with $\boldsymbol{D}_q$ a difference matrix of order $q$ (P-splines).

### 2.1 Prediction with smooth models and quadratic penalties

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

In the framework of model (4), given a vector of $n$ observations $\boldsymbol{y}$ of the response variable, suppose that we want to predict $n_p$ new values $\boldsymbol{y}_p$ at $\boldsymbol{x}_p$. If the prediction is within sample everything is straightforward, the coefficients remain the same and we just have to extend the basis accordingly. The fit and the prediction are,

$$\boldsymbol{y}_+ = \tilde{\boldsymbol{B}}\hat{\boldsymbol{\theta}},$$

with $\tilde{\boldsymbol{B}}$ a reconstruction of the basis $\boldsymbol{B}$, keeping the same knots but with new rows, and $\hat{\boldsymbol{\theta}}$ as in (27).

In the case of prediction out-of-sample, we obviously have to extend the basis but also the penalty to penalize the new coefficients. Currie *et al.* (2004) define the new vector of observations

$$\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}_p')', \tag{7}$$

that contains the observed response $\boldsymbol{y}$ and the unknown values to be predicted.

A new extended basis, $\boldsymbol{B}_+$, is built from a new set of knots that extends the original knots used to fit the observed data to cover the range of the $n_p$ observations to forecast:

$$\boldsymbol{B}_+ = \begin{bmatrix} \boldsymbol{B} & \boldsymbol{O} \\ \boldsymbol{B}_1 & \boldsymbol{B}_2 \end{bmatrix}, \text{ of size } n_+ \times c_+, \tag{8}$$

where $\boldsymbol{B}$ is the $n \times c$ basis used for fitting the trend component, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are auxiliary basis for prediction up to $n_+ = n + n_p$ values, of sizes $n_p \times c$ and $n_p \times c_p$ respectively, and $c_+ = c + c_p$. The following figure represents a extended splines basis.

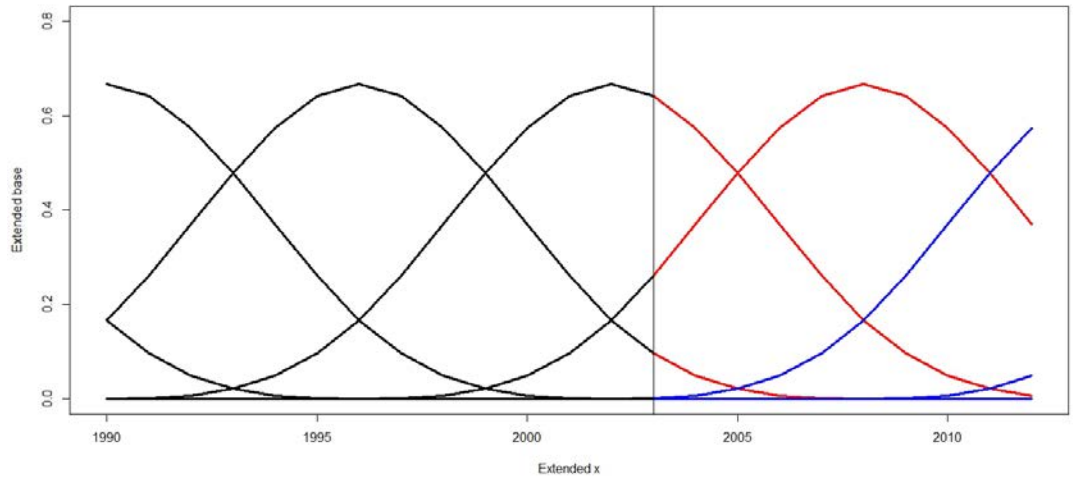


Figure 1: Extended basis example.

In Figure 1 we show the original basis $\boldsymbol{B}$ in black, the $\boldsymbol{B}_1$ component in red and the $\boldsymbol{B}_2$ part in blue. Associated to the new basis $\boldsymbol{B}_+$, a new vector of coefficients, $\boldsymbol{\theta}_+$,

is defined, with length $c_+ \times 1$. A new quadratic penalty associated with the new set of coefficients needs to be introduced, let say $\boldsymbol{P}_+$. As $\boldsymbol{B}_+$, $\boldsymbol{P}_+$ can be splited as:

$$\boldsymbol{P}_+ = \begin{bmatrix} \boldsymbol{P}_1 & \boldsymbol{P}_2 \\ \boldsymbol{P}_2' & \boldsymbol{P}_3 \end{bmatrix}. \tag{9}$$

In the case of P-splines $\boldsymbol{P}_+ = \boldsymbol{D}'_+\boldsymbol{D}_+$, built from a difference matrix,

$$\boldsymbol{D}_+ = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_1 & \boldsymbol{D}_2 \end{bmatrix}, \tag{10}$$

of order $q$ and size $(c_+ - q) \times c_+$, with $\boldsymbol{D}$ the difference matrix used to build the penalty matrix for the fit.

Then, the model can be fitted and predicted simultaneously by minimizing the following function of $\boldsymbol{\theta}_+$:

$$S = (\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+)'\boldsymbol{W}(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + \lambda\boldsymbol{\theta}'_+\boldsymbol{P}_+\boldsymbol{\theta}_+, \tag{11}$$

where the $\boldsymbol{y}_p$ 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 and 1 if the data is observed.
Deriving with respect to $\boldsymbol{\theta}_+$:

$$\frac{\partial S}{\partial \boldsymbol{\theta}_+} = -2\boldsymbol{B}'_+\boldsymbol{W}(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_+) + 2\lambda\boldsymbol{P}_+\boldsymbol{\theta}_+. \tag{12}$$

Therefore, the penalized least square solution is given by:

$$\hat{\boldsymbol{\theta}}_+ = (\boldsymbol{B}'_+\boldsymbol{W}\boldsymbol{B}_+ + \lambda\boldsymbol{P}_+)^{-1}\boldsymbol{B}'_+\boldsymbol{W}\boldsymbol{y}_+, \tag{13}$$

and $\hat{\boldsymbol{y}} = \boldsymbol{H}_+\boldsymbol{y}_+$ with $\boldsymbol{H}_+ = \boldsymbol{B}_+(\boldsymbol{B}'_+\boldsymbol{W}\boldsymbol{B}_+ + \lambda\boldsymbol{P}_+)^{-1}\boldsymbol{B}'_+\boldsymbol{W}$.

Moreover, writing the fit and the forecast in function of the extended penalty matrix, (9), and applying Theorem 9.6.1 given in Harville (2000), we have that:

$$\hat{\boldsymbol{y}}_+ = \boldsymbol{B}_+ \begin{bmatrix} \boldsymbol{I} \\ -\boldsymbol{P}_3^- \boldsymbol{P}_2' \end{bmatrix} (\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{P}_1 - \lambda\boldsymbol{P}_2\boldsymbol{P}_3^-\boldsymbol{P}_2')^{-1}\boldsymbol{B}'\boldsymbol{y}. \tag{14}$$

The 95% confidence interval is $\hat{\boldsymbol{y}} \pm 1.96\sqrt{\sigma^2_{\boldsymbol{\epsilon}}\text{diag}(\boldsymbol{H}_+\boldsymbol{H}'_+)}$.

## 2.2 Prediction with mixed effects smooth models

The connection between penalized smoothing and mixed models was established thirty years ago in Green (1987). The key point of this equivalence is the fact that the smoothing parameter becomes a variance components ratio. Both variance components can be estimated through restricted maximum likelihood procedure (REML), see Patterson & Thompson (1971), and, therefore, it is not longer necessary to estimate $\lambda$ via a cross-validation method or an information criterion.

To represent a penalized smooth model as a mixed model it is necessary to find a new basis that allows the representation of model (5) with its associated penalty as a mixed model of the form:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad \text{with} \quad \boldsymbol{\alpha} \sim \mathcal{N}(0, \boldsymbol{G}) \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}), \tag{15}$$

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 covariance matrix $\boldsymbol{G}$, which depends on the variance of the random effects $\sigma^2_{\boldsymbol{\alpha}}$.

There are different alternatives for the reparameterization of the original smooth model based on quadratic penalties into a mixed model depending on the basis and the penalty used. The idea is to find a transformation $\boldsymbol{\Omega}$ such that:

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

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., $\boldsymbol{\Omega} = [\boldsymbol{\Omega}_f : \boldsymbol{\Omega}_r]$, and such 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 sub-matrix such that: (i) the composed matrix $[\boldsymbol{X} : \boldsymbol{Z}]$ has full rank (this also implies that both $\boldsymbol{X}$ and $\boldsymbol{Z}$ have full column rank) and (ii) $\boldsymbol{X}$ and $\boldsymbol{Z}$ are orthogonal, i.e., $\boldsymbol{X}'\boldsymbol{Z} = 0$. For the sub-matrix $\boldsymbol{\Omega}_r$ there are different alternatives, following the approach of Currie & Durbán (2002), we use the singular value decomposition of the penalty matrix, $\boldsymbol{P} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}'$, where $\tilde{\boldsymbol{\Sigma}}$ is a diagonal matrix that contains the eigenvalues of $\boldsymbol{P}$, and $\boldsymbol{U}$ is the corresponding matrix of eigenvectors, and we define

$$\boldsymbol{\Omega}_r = \boldsymbol{U}_r\boldsymbol{\Sigma}^{-1/2},$$

$\boldsymbol{\Sigma}$ containing the positive eigenvalues and $\boldsymbol{U}_r$ containing the span of the decomposition. With this reparametrization, it is straightforward to obtain the relationship between the inverse of the covariance matrix $\boldsymbol{G}$ of the random effects and the penalty $\boldsymbol{P}$:

$$\boldsymbol{G}^{-1} = \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{\Omega}_r'\boldsymbol{P}\boldsymbol{\Omega}_r. \tag{16}$$

Once we have stablished the connection between mixed models and P-splines, we can use the results given in Gilmour *et al.* (2004) to predict new observations. In the paper the authors defined the prediction to be a linear function of the best linear unbiased predictor (BLUP) of random effects with the best linear unbiased estimator (BLUE) of the fixed effects in the model.

They consider the augmented mixed model,

$$\boldsymbol{y}_+ = \boldsymbol{X}_+\boldsymbol{\beta}_+ + \boldsymbol{Z}_+\boldsymbol{\alpha}_+ + \boldsymbol{\epsilon}_+, \tag{17}$$

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 $\boldsymbol{\epsilon}_+ \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})$, $\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{X}_p\end{bmatrix}$ the extended fixed part, $\boldsymbol{\beta}_+$ the fixed effects (the same as the ones that give the fit), $\begin{bmatrix}\boldsymbol{Z} & \boldsymbol{O}\\ \boldsymbol{Z}_1 & \boldsymbol{Z}_2\end{bmatrix}$ the extended random part (with $\boldsymbol{Z}$ the random part that gives the fit) and $\boldsymbol{\alpha}_+ = [\boldsymbol{\alpha}', \boldsymbol{\alpha}'_p]'$ the augmented random effects with covariance matrix

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

where $\boldsymbol{G}$ is the covariance matrix of the random effects in the model for the observed data.
They showed that the new predicted values are

$$\hat{\boldsymbol{y}}_p = \boldsymbol{X}_p\hat{\boldsymbol{\beta}}_+ + \boldsymbol{Z}_{(p)}\hat{\boldsymbol{\alpha}}, \tag{18}$$

with

$$\boldsymbol{Z}_{(p)} = \boldsymbol{Z}_1 + \boldsymbol{Z}_2\boldsymbol{G}_{po}\boldsymbol{G}^{-1} \tag{19}$$

Therefore, the random effects vector of the predicted values is $\hat{\boldsymbol{\alpha}}_p = \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\hat{\boldsymbol{\alpha}}$. We denote this method as "MM".

Now we propose an alternative approach based on the reparameterization of the method presented in Section 2.1 as a mixed model, in this case the estimation of model (17) is done using the extended mixed model equations of Henderson:

$$\begin{bmatrix}\hat{\boldsymbol{\beta}}_+\\ \hat{\boldsymbol{\alpha}}_+\end{bmatrix} = \boldsymbol{C}^{-1}\begin{bmatrix}\boldsymbol{X}'_+\boldsymbol{M}\\ \boldsymbol{Z}'_+\boldsymbol{M}\end{bmatrix}\boldsymbol{y}_+, \tag{20}$$

where $\boldsymbol{C} = \begin{bmatrix}\boldsymbol{X}_+{}'\boldsymbol{M}\boldsymbol{X}_+ & \boldsymbol{X}_+{}'\boldsymbol{M}\boldsymbol{Z}_+\\ \boldsymbol{Z}_+{}'\boldsymbol{M}\boldsymbol{X}_+ & \boldsymbol{Z}_+{}'\boldsymbol{M}\boldsymbol{Z}_+ + \boldsymbol{G}_+^{-1}\end{bmatrix}$ and $\boldsymbol{Z}_+ = \boldsymbol{B}_+\boldsymbol{\Omega}_{+_r}$, with $\boldsymbol{\Omega}_{+_r}$ any orthogonal transformation such that

$\boldsymbol{B}_+[\boldsymbol{\Omega}_{+_f} : \boldsymbol{\Omega}_{+_r}] = [\boldsymbol{X}_+ : \boldsymbol{Z}_+]$ and $\begin{bmatrix}\boldsymbol{\Omega}'_{+_f}\\ \boldsymbol{\Omega}'_{+_r}\end{bmatrix}\boldsymbol{\theta}_+ = \begin{bmatrix}\boldsymbol{\beta}_+\\ \boldsymbol{\alpha}_+\end{bmatrix}$ to have $\boldsymbol{B}_+\boldsymbol{\theta}_+ = \boldsymbol{X}_+\boldsymbol{\beta}_+ + \boldsymbol{Z}_+\boldsymbol{\alpha}_+$,

$\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}'_p)'$ as in (7) , and $\boldsymbol{M} = \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{W}$, with $\boldsymbol{W}$ a diagonal matrix of dimension $n_+ \times n_+$ with 0 entries if the data is forecasted and 1 if the data is observed. The solutions are

$$\boldsymbol{\beta}_+ = (\boldsymbol{X}'_+\boldsymbol{V}_+^{-1}\boldsymbol{X}_+)^{-1}\boldsymbol{X}'_+\boldsymbol{V}_+^{-1}\boldsymbol{y}_+, \tag{21}$$

$$\boldsymbol{\alpha}_+ = \boldsymbol{G}_+\boldsymbol{Z}'_+\boldsymbol{V}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{X}_+\boldsymbol{\beta}_+), \tag{22}$$

where $\boldsymbol{V}_+ = \boldsymbol{Z}_+\boldsymbol{G}_+\boldsymbol{Z}'_+ + \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{W}$ and $\boldsymbol{V}_+^{-1} = \boldsymbol{M} - \boldsymbol{M}\boldsymbol{Z}_+(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}'_+\boldsymbol{M}\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}'_+\boldsymbol{M}$.

Note that $\boldsymbol{V}_+$ includes the variance components $\sigma^2_{\boldsymbol{\epsilon}}$ and $\sigma^2_{\boldsymbol{\alpha}}$, through the covariance matrices $\sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I}$ and $\boldsymbol{G}_+$, respectively. Let us denote this last method as "MMM".

The 95% confidence interval is $\hat{\boldsymbol{y}} \pm \sqrt{\sigma^2_{\boldsymbol{\epsilon}} \text{diag}\left(\begin{bmatrix}\boldsymbol{X}'_+ \\ \boldsymbol{Z}'_+\end{bmatrix} \boldsymbol{C}^{-1} \left[\boldsymbol{X}_+ | \boldsymbol{Z}_+\right]\right)}$, with $\boldsymbol{C}$ as in (20).

The method proposed in Gilmour *et al.* (2004) is a two-stage procedure (first fit the actual data, and then predict new values) and our extension of the missing value approach to the mixed model framework fits and predicts simultaneously. In order to know the relationship between the two methods, MM and MMM, we need to know the relationship between the covariance matrix of the random effects that gives the fit and the extended covariance matrix.

The following theorem shows the relation between the MMM method and the method proposed in Gilmour *et al.* (2004) in the context of penalties based on differences.

**Theorem 1.** *Given the model in (6) with penalty based on differences between adjacent coefficients. The fit and the prediction of new observations given by MMM and MM are the same if the transformation matrix in (20) is the direct extension of the original transformation,*

$$\boldsymbol{\Omega}_{+_r} = \begin{bmatrix}\boldsymbol{\Omega}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{\Omega}_{p_r}\end{bmatrix}, \tag{23}$$

*where* $\boldsymbol{\Omega}_r = \boldsymbol{U}_r \boldsymbol{\Sigma}^{-1/2}$, *based on the SVD of* $\boldsymbol{D}'\boldsymbol{D} = \boldsymbol{U}\tilde{\boldsymbol{\Sigma}}\boldsymbol{U}'$, *is the transformation matrix for the random component used for the observed data and* $\boldsymbol{\Omega}_{p_r} = \boldsymbol{D}_2^{-1}$ *is the transformation matrix for the random component of the predicted values, with* $\boldsymbol{D}$ *and* $\boldsymbol{D}_2$ *blocks of the extended difference matrix* $\boldsymbol{D}_+$, *(10).*

The proof can be found in Appendix A.

We also prove in Appendix B that the variance components $(\sigma^2_{\boldsymbol{\alpha}}, \sigma^2)$ that maximize the REML, $l$, and the REML corresponding to (17), $l_+$, are equal,

$$l(\sigma^2, \sigma^2_{\boldsymbol{\alpha}}, \rho) = \underbrace{-\frac{1}{2}\log|\boldsymbol{V}|}_{\text{Part I}} \underbrace{-\frac{1}{2}\log|\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}|}_{\text{Part II}} \underbrace{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}_{\text{Part III}}, \tag{24}$$

$$l_+(\sigma^2, \sigma^2_{\boldsymbol{\alpha}}, \rho) = \underbrace{-\frac{1}{2}\log|\boldsymbol{V}_+|}_{\text{Part I}} \underbrace{-\frac{1}{2}\log|\boldsymbol{X}'_+\boldsymbol{V}_+^{-1}\boldsymbol{X}_+|}_{\text{Part II}} \underbrace{-\frac{1}{2}\left(\boldsymbol{y}_+-\boldsymbol{X}_+\boldsymbol{\beta}_+\right)'\boldsymbol{V}_+^{-1}\left(\boldsymbol{y}_+-\boldsymbol{X}_+\boldsymbol{\beta}_+\right)}_{\text{Part III}}. \tag{25}$$

Notice the importance of the previous statement, it means that the variance parameters used to predict are the same as the ones that estimating the original fit.

## 2.3 Prediction with penalized Gaussian process regression

As it is known, one of the attractive features of penalized regression is its link to stochastic processes. The title of this section is inspired in Yi *et al.* (2011), they use the penalized Gaussian process regression to provide an alternative solution to the

Gaussian process regression variable selection problem, since when dimension of the data is high, it suffers from large variance of parameter estimation and high predictive errors. They apply several penalized methods to a Gaussian process model, including Ridge, LASSO, Bridge, SCAD and adaptive LASSO penalties.

We also use Gaussian processes but not on the curve, we make a representation of the curve in terms of bases and coefficients, and we have a Gaussian process on the coefficients. Our proposal to predict new values, in the context of Gaussian process smoothing, is to use a model based on Gaussian process prior and a P-spline covariance matrix to fit non linear data. That is, we will harness the flexibility of the Gaussian process and the choice of a suitable covariance matrix to model any nonlinear model nonparametrically. In addition, the prediction is quite straightforward due to the properties of Gaussian processes.

Prediction with Gaussian processes has a long history in at least three literatures: mathematical geology (where the approach is called 'kriging', see Cressie (1993)), neural networks (Poggio & Girosi (1990) and Girosi *et al.* (1995)) and global optimization in the analysis of computer experiments (e.g Sacks *et al.* (1989)).

Building on the previous approaches, we predict new values by proposing the penalized regression framework to a Gaussian process model. In the context of model (4), we can assume that the stochastic behaviour of the random vector $\boldsymbol{y}$ depends on the observed covariate and a latent process $\boldsymbol{s}$, according to a linear mixed model,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{s} + \boldsymbol{\epsilon},$$

with $\boldsymbol{X}\boldsymbol{\beta}$ the trend and $\boldsymbol{\epsilon}$ an independent Gaussian process with zero mean and variance $\sigma^2_{\boldsymbol{\epsilon}}$, modelling the measurement error, i.e., $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})$. Random effects are assumed to account for variability and represented in terms of basis functions $\boldsymbol{s} = \boldsymbol{B}\boldsymbol{\alpha}$, with $\boldsymbol{B}$ any basis. Imposing a prior structure on $\boldsymbol{\alpha}$ through $\boldsymbol{\alpha} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{P}^-)$, with $\boldsymbol{P}^-$ the covariance matrix of the vector of coefficients, we have the Gaussian process

$$\boldsymbol{s} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{ss}),$$

with $\boldsymbol{\Sigma}_{ss} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B}\boldsymbol{P}^-\boldsymbol{B}'$. Independence of $\boldsymbol{s}$ and $\boldsymbol{\epsilon}$ implies that elements of $\boldsymbol{y}$ are independent and normally distributed conditionally on $\boldsymbol{X}$ and $\boldsymbol{s}$. Then, the marginal distribution of the process $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{yy}}),$$

with $\boldsymbol{\Sigma}_{\boldsymbol{yy}} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I} + \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B}\boldsymbol{P}^-\boldsymbol{B}'$.

Assuming that the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{yy}}$ is known, the maximum likelihood estimator of the trend parameter vector $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{yy}}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}_{\boldsymbol{yy}}\boldsymbol{y}. \tag{26}$$

As is well known from Normal distribution theory, the conditional normal distribution of $\boldsymbol{s}|\boldsymbol{y}$ is $\mathcal{N}(\mathbb{E}[\boldsymbol{s}|\boldsymbol{y}], \boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{y}})$, with:

$$\begin{aligned}\mathbb{E}[\boldsymbol{s}|\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{s}] + \boldsymbol{\Sigma}_{\boldsymbol{sy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}]) \\ &= \mathbf{0} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}'(\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}')^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \boldsymbol{B}(\lambda\boldsymbol{P} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),\end{aligned}$$

since $\boldsymbol{\Sigma}_{\boldsymbol{sy}} = \mathbb{E}[\boldsymbol{ss}] = \boldsymbol{\Sigma}_{\boldsymbol{ss}} = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}'$, and

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{s}|\boldsymbol{y}} &= \boldsymbol{\Sigma}_{\boldsymbol{ss}} - \boldsymbol{\Sigma}_{\boldsymbol{sy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{ys}} \\ &= \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}' - \sigma_{\boldsymbol{\alpha}}^4\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}'(\sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}')^{-1}\boldsymbol{B}\boldsymbol{P}^{-}\boldsymbol{B}' \\ &= \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{B}\left[\lambda\boldsymbol{P} + \boldsymbol{B}'\boldsymbol{B}\right]^{-1}\boldsymbol{B}'.\end{aligned}$$

Therefore, the fit is:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{s}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{B}(\lambda\boldsymbol{P} + \boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}). \tag{27}$$

Let $\boldsymbol{x}_p$ be a vector of $n_p$ unobserved values of the process with

$$\boldsymbol{y}_p = \boldsymbol{X}_p\boldsymbol{\beta} + \boldsymbol{s}_p + \boldsymbol{\epsilon}_p,$$

where $\boldsymbol{s}_p \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{\Sigma}_{\boldsymbol{s}_p\boldsymbol{s}_p})$, $\boldsymbol{\epsilon}_p \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I})$. Therefore, the joint distribution of observed and unobserved values is given by:

$$\begin{bmatrix}\boldsymbol{y} \\ \boldsymbol{y}_p\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}\boldsymbol{\beta} \\ \boldsymbol{X}_p\boldsymbol{\beta}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{yy}} & \boldsymbol{\Sigma}_{\boldsymbol{yy}_p} \\ \boldsymbol{\Sigma}_{\boldsymbol{yy}_p}^{'} & \boldsymbol{\Sigma}_{\boldsymbol{y}_p\boldsymbol{y}_p}\end{bmatrix}\right),$$

where $\boldsymbol{\Sigma}_{\boldsymbol{yy}} = \sigma_{\boldsymbol{\epsilon}}^2\boldsymbol{I} + \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{\Sigma}_{\boldsymbol{ss}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{yy}_p} = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{\Sigma}_{\boldsymbol{ss}_p}$.

Pollice & Bilancia (2001) showed that the minimum variance predictor of $\boldsymbol{y}_p$ conditional on values of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{yy}}$, is given by

$$\mathbb{E}[\boldsymbol{y}_p|\boldsymbol{y}] = \boldsymbol{X}_p\boldsymbol{\beta} + \boldsymbol{\Sigma}'_{\boldsymbol{yy}_p}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Therefore, in order to calculate the predicted values we need to compute $\boldsymbol{\Sigma}'_{\boldsymbol{yy}_p}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}$. Imposing a prior structure on $\boldsymbol{\alpha}_+$ through $\boldsymbol{\alpha}_+ \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{P}_+^{-})$, where $\boldsymbol{P}_+^{-}$ is the covariance of the extended vector of coefficients:

$$\boldsymbol{P}_+^{-} = \begin{bmatrix}\boldsymbol{P}^1 & \boldsymbol{P}^2 \\ \boldsymbol{P}^{2'} & \boldsymbol{P}^3\end{bmatrix},$$

and since the extended basis is $\boldsymbol{B}_+$ is (8), we have:

$$\boldsymbol{B}_+\boldsymbol{P}_+^{-}\boldsymbol{B}_+' = \begin{bmatrix}\boldsymbol{B}\boldsymbol{P}^1\boldsymbol{B}' & \boldsymbol{B}(\boldsymbol{P}^1\boldsymbol{B}_1' + \boldsymbol{P}^2\boldsymbol{B}_2') \\ (\boldsymbol{B}_1\boldsymbol{P}^1 + \boldsymbol{B}_2\boldsymbol{P}^{2'})\boldsymbol{B}' & (\boldsymbol{B}_1\boldsymbol{P}^1 + \boldsymbol{B}_2\boldsymbol{P}^{2'})\boldsymbol{B}_1' + (\boldsymbol{B}_1\boldsymbol{P}^2 + \boldsymbol{B}_2\boldsymbol{P}^3)\boldsymbol{B}_2'\end{bmatrix},$$

i.e., $\boldsymbol{\Sigma}_{\boldsymbol{yy}_p} = \sigma_{\boldsymbol{\alpha}}^2\boldsymbol{B}(\boldsymbol{P}^1\boldsymbol{B}_1' + \boldsymbol{P}^2\boldsymbol{B}_2')$. Then, applying Theorem 18.2.8 and Lemma 18.2.1 given in Harville (2000) we have that:

$$\begin{aligned}\boldsymbol{\Sigma}'_{\boldsymbol{y}\boldsymbol{y}_p}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{y}\boldsymbol{y}} &= \sigma^2_{\boldsymbol{\alpha}}(\boldsymbol{B}_1\boldsymbol{P}^1\boldsymbol{B}' + \boldsymbol{B}_2\boldsymbol{P}^{2'}\boldsymbol{B}')(\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B}\boldsymbol{P}^-\boldsymbol{B}' + \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I})^{-1} \\ &= \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}_1\boldsymbol{P}^1\boldsymbol{P}(\sigma^{-2}_{\boldsymbol{\alpha}}\boldsymbol{P} + \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}' + \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}_2\boldsymbol{P}^{2'}\boldsymbol{P}(\sigma^{-2}_{\boldsymbol{\alpha}}\boldsymbol{P} + \sigma^{-2}_{\boldsymbol{\epsilon}}\boldsymbol{B}'\boldsymbol{B})^{-1}\boldsymbol{B}'.\end{aligned}$$

Therefore, predictions are written as function of the extended penalty matrix:

$$\hat{\boldsymbol{y}}_p = \boldsymbol{X}_p\hat{\boldsymbol{\beta}} + \begin{bmatrix}\boldsymbol{B}_1 & \boldsymbol{B}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{I} \\ -\boldsymbol{P}_3^{-}\boldsymbol{P}'_2\end{bmatrix}(\boldsymbol{P}_1 - \boldsymbol{P}_2\boldsymbol{P}_3^{-1}\boldsymbol{P}'_2)^{-}\boldsymbol{P}(\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{P})^{-1}\boldsymbol{B}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{28}$$

with $\hat{\boldsymbol{\beta}}$ as in (26).

Since, essentially splines correspond to Gaussian processes with a particular choice of covariance function. Knowing that in the case of penalties based on differences the extended penalty matrix is

$$\boldsymbol{P}_+ = \begin{bmatrix}\boldsymbol{D}'\boldsymbol{D} + \boldsymbol{D}'_1\boldsymbol{D}_1 & \boldsymbol{D}'_1\boldsymbol{D}_2 \\ \boldsymbol{D}'_2\boldsymbol{D}_1 & \boldsymbol{D}'_2\boldsymbol{D}_2\end{bmatrix}, \tag{29}$$

built from the difference matrix $\boldsymbol{D}_+$, (10), it is straigthforward to prove that the solution of the missing value approach (14) and the solution assuming that the response is a realization of a Gaussian process (28) are equal.

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

To study the properties of the proposed methods we focus on the case of penalties based on differences between adjacent coefficients, in which case, the extended penalty matrix in (11) is (29). The following theorem introduces two attractive properties of the coefficients of the model.

**Theorem 2.** *The coefficients from the minimizing of (11) with extended penalty matrix (29) satisfy the following properties:*

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

$$\hat{\boldsymbol{\theta}}_{+_{1,\ldots,c}} = \hat{\boldsymbol{\theta}}.$$

*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 (14) we have that:

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

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:

$$\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1(\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'\boldsymbol{y} = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}}. \tag{30}$$

□

**Corollary 3** (Theorem 2)**.** *The solution of the extended missing value problem, (11), and the solution of the extended problem,*

$$S_+ = (\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_e)'(\boldsymbol{y}_+ - \boldsymbol{B}_+\boldsymbol{\theta}_e) + \lambda\boldsymbol{D}_+'\boldsymbol{D}_+, \tag{31}$$

*where $\boldsymbol{y}_+$, $\boldsymbol{B}_+$, $\boldsymbol{D}_+$ are as in (7), (8) and (10) respectively, are equal.*

*Proof.* Then if $\hat{\boldsymbol{\theta}}_e$ is the vector that minimizes (31), written $\boldsymbol{B}_+$ as in (8) in

$$\boldsymbol{B}_+'\boldsymbol{B}_+\hat{\boldsymbol{\theta}}_e + \lambda\boldsymbol{D}_+'\boldsymbol{D}_+\hat{\boldsymbol{\theta}}_e = \boldsymbol{B}_+'\boldsymbol{y}_+$$

we have:

$$\boldsymbol{B}_o'\boldsymbol{B}_o\hat{\boldsymbol{\theta}}_e + \boldsymbol{B}_p'\boldsymbol{B}_p\hat{\boldsymbol{\theta}}_e + \lambda\boldsymbol{D}_+'\boldsymbol{D}_+\hat{\boldsymbol{\theta}}_e = \boldsymbol{B}_o'\boldsymbol{y} + \boldsymbol{B}_p'\boldsymbol{y}_p. \tag{32}$$

On the other hand, equation (31) can be written as

$$\begin{bmatrix}\boldsymbol{y} - \boldsymbol{B}_o\boldsymbol{\theta}_e \\ \boldsymbol{y}_p - \boldsymbol{B}_p\boldsymbol{\theta}_e\end{bmatrix}' \begin{bmatrix}\boldsymbol{y} - \boldsymbol{B}_o\boldsymbol{\theta}_e \\ \boldsymbol{y}_p - \boldsymbol{B}_p\boldsymbol{\theta}_e\end{bmatrix} + \lambda\boldsymbol{\theta}_e'\boldsymbol{D}_+'\boldsymbol{D}_+\boldsymbol{\theta}_e,$$

so taking derivatives with respect to $\boldsymbol{y}_p$, we get

$$\hat{\boldsymbol{y}}_p = \boldsymbol{B}_p\hat{\boldsymbol{\theta}}_e \tag{33}$$

Now, using (33) to rewrite (32) and simplifying:

$$\boldsymbol{B}_o'\boldsymbol{B}_o\hat{\boldsymbol{\theta}}_e + \lambda\boldsymbol{D}_+'\boldsymbol{D}_+\hat{\boldsymbol{\theta}}_e = \boldsymbol{B}_o'\boldsymbol{y}.$$

Hence $\hat{\boldsymbol{\theta}}_e$ is

$$\hat{\boldsymbol{\theta}}_e = (\boldsymbol{B}_o'\boldsymbol{B}_o + \lambda\boldsymbol{D}_+'\boldsymbol{D}_+)^{-1}\boldsymbol{B}_o'\boldsymbol{y}. \tag{34}$$

I.e., the solution of the extended missing value problem, (13), and the solution of the extended problem, (34), are the equal.

□

**Corollary 4** (Theorem 2)**.** *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 showed in Appendix C.

This is an important result since it shows an inmediate connection between the penalty (or prior distribution) and the coefficients and the shape of the prediction.

# 4 Application

In this section, we apply the proposed methods to two real data sets. One of them on the log mortality rates, since modelling and forecasting mortality data is a challenging task in areas such as demography or insurance industry. The other dataset, rental prices, allows us to show a simple example of predicting within the framework of additive models.

## 4.1 Mortality data

To illustrate the proposed methodology we use a data set on the log mortality rates of Spanish men aged 73 between 1960 and 2009. In order to predict the log mortality rates of Spanish men aged 73 between 2010 and 2019, we apply the missing value approach with B-splines of degree 3 as basis and second-order difference penalties between adjacent coefficients as the penalty. The result and the 95% confidence interval are showed in Figure 2.

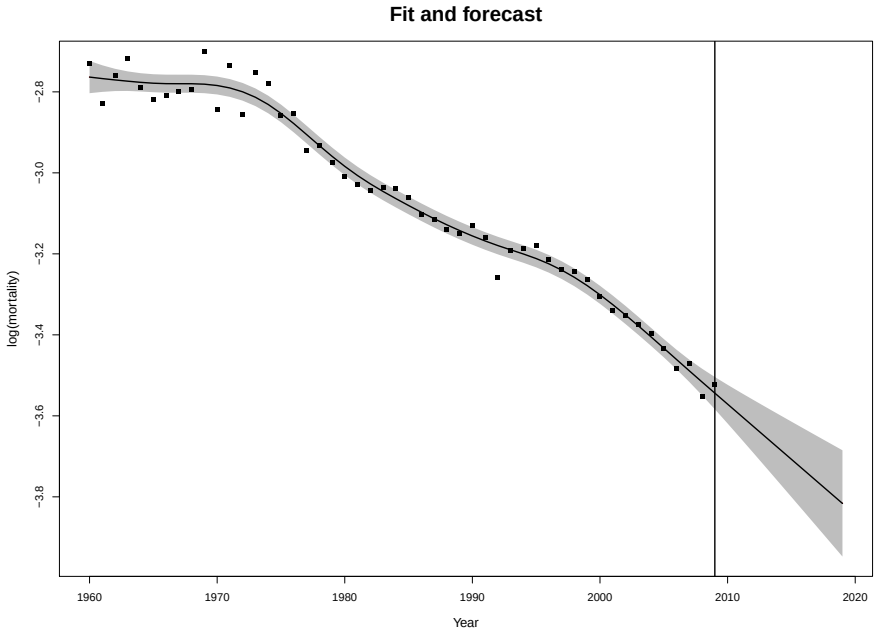


Figure 2: Fit, forecast and 95% confidence interval result of applying the missing value approach of a data set on the log mortality rates of Spanish men aged 73, between 1960 and 2009.

In order to illustrate the result of Corollary 4 we repeat the procedure of the previous example with three different penalty orders (1, 2 and 3). As it can be seen in Figure 3, if the penalty has order 1, the forecast is constant, if the penalty is of order 2 the forecast is a line and if penalty is of third order the forecast is quadratic.

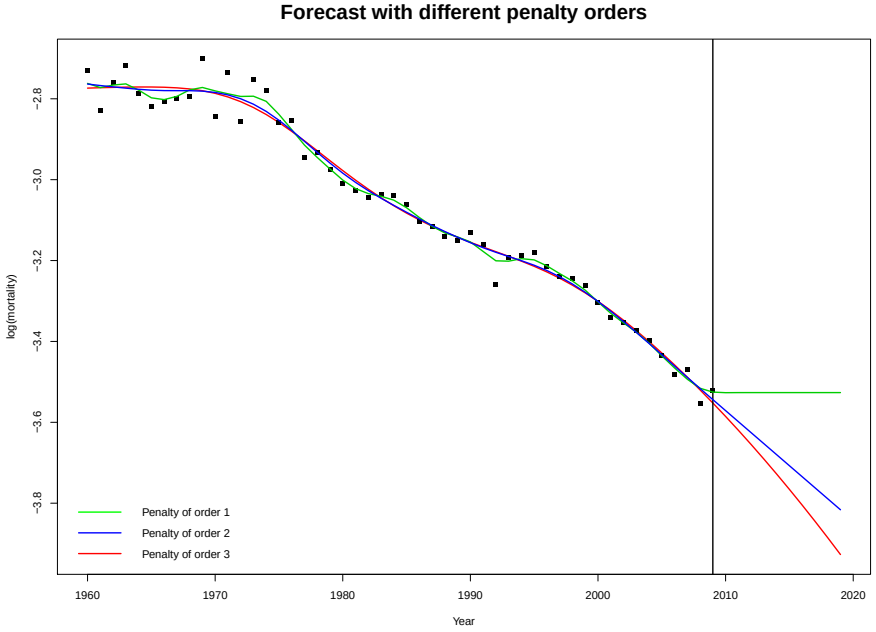


Figure 3: Fit and forecast result of applying the missing value approach with penalty orders 1, 2 and 3 of a data set on the log mortality rates of Spanish men aged 73 between 1960 and 2009.

## 4.2 Rental data

Although all results presented in the previous sections are obtained in the case of smooth models with a single covariate. It is immediately to extend these results to the case of semiparametric or additive models. In this case we analyze data corresponding to rental prices of 3153 houses. The aim of the analysis is to see the relationship between the net rent and the size in square meters and the location of a house. We can fit the following model:

$$\boldsymbol{y}_i = \boldsymbol{\beta}_0 + \boldsymbol{c}_i\boldsymbol{\beta}_1 + f(\boldsymbol{x}_i), \tag{35}$$

where

$$\boldsymbol{c}_i = \begin{cases} 0 & \text{if house } i \text{ is not located in the center} \\ 1 & \text{if house } i \text{ is located in the center} \end{cases}$$

and $f(\boldsymbol{x})$ is the function that represent the main effects of the area (it is measured in square meters and its values are between 20 and 230). The regression matrix is then defined by blocks as

$$\boldsymbol{B} = \left[ \mathbf{1} \mid \boldsymbol{c} \mid \boldsymbol{B}_{\boldsymbol{x}} \right],$$

with marginal B-spline basis of degree three of the covariate area, $\boldsymbol{B}_{\boldsymbol{x}}$. The penalty matrix associated with model (35) has a block-diagonal form:

$$\boldsymbol{P} = blockdiag(0, 0, \lambda_{\boldsymbol{x}}\boldsymbol{P}_{\boldsymbol{x}}),$$

where $\boldsymbol{P}_{\boldsymbol{x}}$ is the marginal second-order difference penalty for area.

Suppose that we want to predict the prices for houses with area between 231 and 280 square meters, applying the missing value approach we extend the basis and the penalty:

$$\boldsymbol{B}_+ = \begin{bmatrix} \mathbf{1} & \boldsymbol{c} & \boldsymbol{B_x} & \boldsymbol{O} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{B}_{\boldsymbol{x}_{(1)}} & \boldsymbol{B}_{\boldsymbol{x}_{(2)}} \end{bmatrix}, \quad \boldsymbol{P}_+ = blockdiag(0, 0, \lambda_{\boldsymbol{x}} \boldsymbol{P}_{\boldsymbol{x}_+}).$$

Once we have extended the basis and the penalty, it is straightforward to obtain the fit and the forecast applying equation (13), but in order to avoid identifiability problems, since $\mathbf{1}$ is contained in the space spanned by the columns of $\boldsymbol{B_x}$, we reparameterize the model using the representation of a penalized spline model as a mixed model, i.e. we apply method MMM. Figure 4 shows the smooth fitted and forecasted trend for area and the 95% confidence interval.

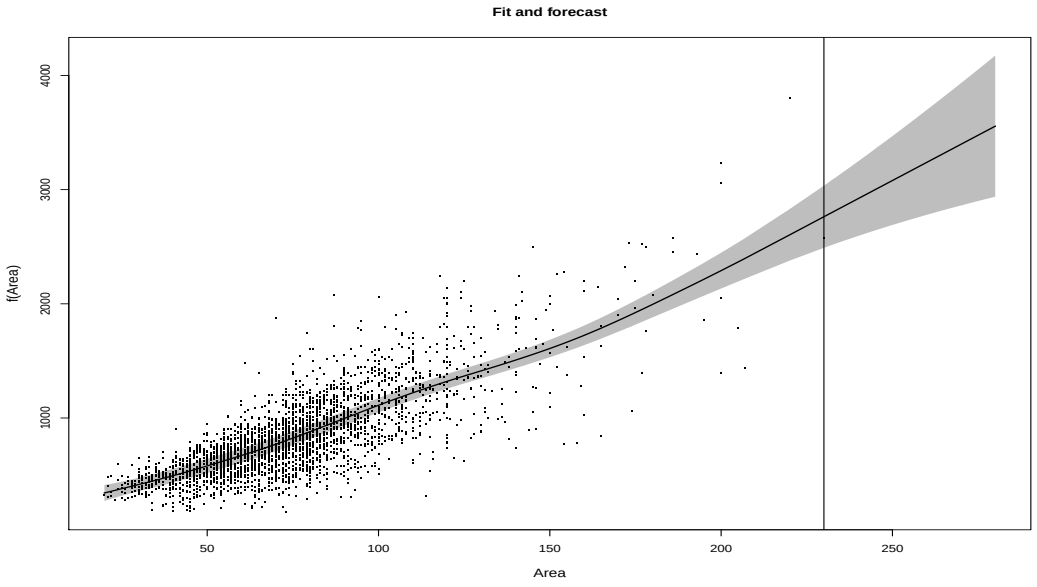


Figure 4: Fit, forecast and 95% confidence interval of the additive smooth term for area, result of applying the MMM approach of a data set on the rental prices of houses with area between 20 and 230 $m^2$.

## 5 Memory of a P-spline

In some occasions, our knowledge of the data can influence our decision on the proportion of the data set that we want to use to predict new observations. I.e., it may be important to know how much of the known information we are using to predict. In this section we introduce the concept of memory of a P-spline as a tool to provide that information and show some of its properties.

It is important to notice that, because the matrix $\boldsymbol{W}$ in (13) is a block diagonal matrix with entries zeros or ones, $\boldsymbol{H}_+$ in (13) has the following form:

$$\boldsymbol{H}_+ = \begin{bmatrix} \boldsymbol{H} & \boldsymbol{O}_1 \\ \boldsymbol{H}_p & \boldsymbol{O}_2 \end{bmatrix}, \tag{36}$$

with $\boldsymbol{H}$ of size $n \times n$, $\boldsymbol{H}_p$ of size $n_p \times n$ and $\boldsymbol{O}_1$ and $\boldsymbol{O}_2$ matrices of zeros of size $n \times n_p$ and $n_p \times n_p$, respectively. I.e., the predicted values given by the missing value approach in the case of penalties based on differences between adjacent coefficients are

$$\hat{\boldsymbol{y}}_p = \boldsymbol{H}_p \boldsymbol{y},$$

where $\boldsymbol{H}_p = \boldsymbol{B}_1(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}' - \boldsymbol{B}_2\boldsymbol{D}_2^{-1}\boldsymbol{D}_1(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\boldsymbol{B}'$, with $\boldsymbol{B}$, $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ as in (8) and $\boldsymbol{D}$, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ as in (10). Therefore, summarizing the rows and columns of $\boldsymbol{H}_p$ we will give us an insight of how the past is affecting the prediciton.

To illustrate the concept of *memory of a P-spline* we use the mortality data set of Section 4.1. The data set contains 50 observations, i.e., the size of the hat matrix that give us the fit is $50 \times 50$. If we forecast up to 2019, i.e., we compute 10 new observations, the hat matrix $\boldsymbol{H}_p$ has size $10 \times 50$. Panel (a) of Figure 5 shows the fit and forecast of the log mortality rates until 2019. Panel (b) dislplays the rows of $\boldsymbol{H}_p$, panel (c): columns of $\boldsymbol{H}_+$, the black lines separates the elements of $\boldsymbol{H}$ and $\boldsymbol{H}_p$. Panel (d) shows the image of the $\boldsymbol{H}_+$ matrix.

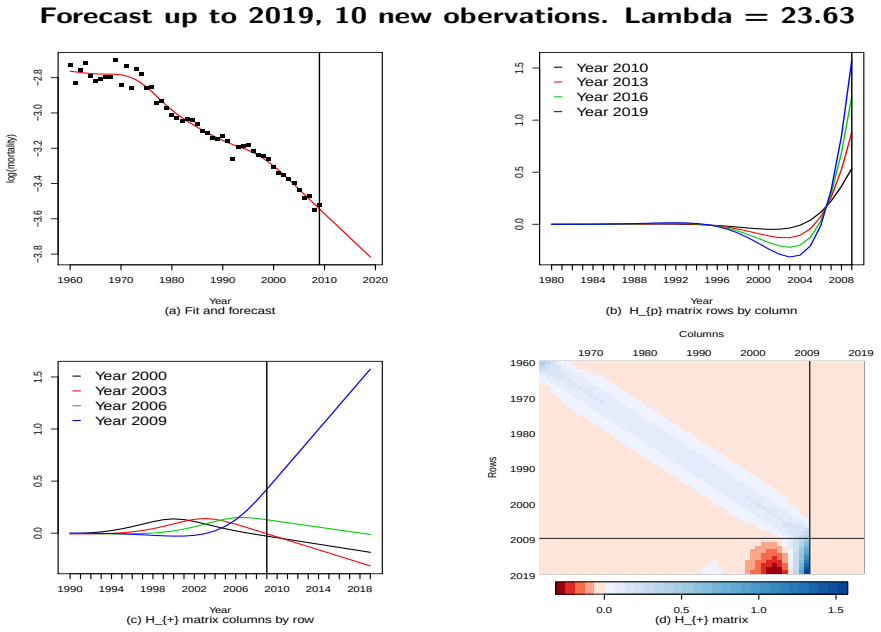


Figure 5: Panel (a): fit and forecast of the log mortality rates until 2019. Panel (b): rows of $\boldsymbol{H}_p$. Panel (c): columns of $\boldsymbol{H}_+$, the black lines separates the elements of $\boldsymbol{H}$ and $\boldsymbol{H}_p$. Panel (d): $\boldsymbol{H}_+$ matrix.

We have noticed that all rows of $\boldsymbol{H}_p$ follow a similar pattern, i.e., if we consider each row as a function, we find that they behave similarly (see panel (b) of figure 5). For instance, if the maximum of the last row is taken at the last column, this also happens in the rest of columns. Moreover, the contribution of each point in the past reduces gradually as we move away from the present.

As each column of $\boldsymbol{H}_p$ gives the contribution of each point of the past in the future values. We have also observed that these contributions are a polynomial function of time of order $q-1$, where $q$ is the order of the penalty (see panel (c) of Figure 5, where $q=2$).

Based on these ideas we have developed the concept *memory of a P-spline*, this new idea will give us information on the overall weight of each observation on the prediction.

We have summarized the columns of $\boldsymbol{H}_p$ as follows: we add them (in absolute value) and standardize them by their sum, this will give us a vector of weights $\boldsymbol{w}$ of the same length of the response variable, $n \times 1$. Considering $T$ as the number of steps backward from the last observation and associating the vector of weights to these values. Then, the ***memory of the P-spline*** is the $99^{\text{th}}$ percentile, $t_0$. That would mean that beyond $t_0$ steps backward no relevant information is affecting the prediction.
Notice that considering the memory as the $99^{\text{th}}$ percentile is just one possible way to summarize the vector of weights. Summary statistics that treat the weights as if they are a discrete distribution (mean, quantiles, expectiles) are other choices.

To calculate the memory of the $P$-spline in the previous example, we compute the vector of weights, $\boldsymbol{w}$, its values are shown in Table 1 (the values of $\boldsymbol{w}_t$ for $t = 25, ..., 50$ are not shown in the table since they are approximately 0) and obtain the $99^{\text{th}}$ percentile. In this case the memory of the $P$-spline is $t_0 = 18$, i.e., what has happened 18 years backward, before 1992, does not influence on the future.

| $t$ | $\boldsymbol{w}_t$ | $t$ | $\boldsymbol{w}_t$ | $t$ | $\boldsymbol{w}_t$ | $t$ | $\boldsymbol{w}_t$ |
|---|---|---|---|---|---|---|---|
| 1 | 0.3315 | 7 | 0.0676 | 13 | 0.0100 | 19 | 0.0029 |
| 2 | 0.1765 | 8 | 0.0617 | 14 | 0.0044 | 20 | 0.0025 |
| 3 | 0.0663 | 9 | 0.0511 | 15 | 0.0006 | 21 | 0.0020 |
| 4 | 0.0150 | 10 | 0.0391 | 16 | 0.0018 | 22 | 0.0014 |
| 5 | 0.0470 | 11 | 0.0276 | 17 | 0.0028 | 23 | 0.0009 |
| 6 | 0.0644 | 12 | 0.0178 | 18 | 0.0031 | 24 | 0.0006 |

Table 1: Normalized weights, $\boldsymbol{w}_t$, for the number of steps backward from the last observed year.

Figure 6 illustrates the result. Left panel shows the vector of weights, the red line corresponds to the year from which we are taking information, 1992. Right panel of Figure 6 shows the fit and the forecast of the log mortality rates until 2019, the data that are between the red and the black lines correspond to the data that contributes to the prediction.

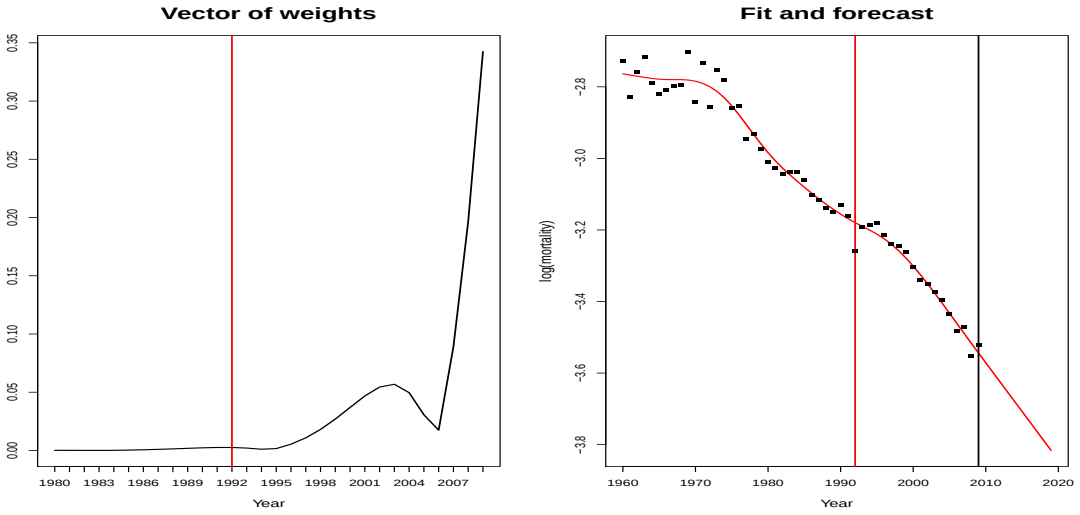


Figure 6: Left panel: vector of weights, the red line corresponds to the year from which we are taking information, 1992. Right panel: fit and forecast of the log mortality rates until 2019, the data that are between the red and the black lines correspond to the data that contributes to the prediction.

## 5.1 Properties of the memory of a P-spline

Although we do not have yet analytic proof of the properties of the memory, we have performed the following simulation study that shows the behaviour of the memory.

We have applied the missing value approach with B-spline basis and second-order penalty to several simulated data sets by using different prediction horizons and bases of different sizes.

We have simulated from $\boldsymbol{y}_i = f(\boldsymbol{x}_i) + \boldsymbol{\epsilon}_i$, $i = 1, ..., n$, $\boldsymbol{x}_i \sim \text{Unif}[0, 1]$ with:

a) Smooth functions and errors:

   i) $f(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i)$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma = 0.1)$.

   ii) $f(\boldsymbol{x}_i) = 2 + \boldsymbol{x}_i$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma = 0.1)$.

   iii) $f(\boldsymbol{x}_i) = 2 + \sin(4\pi\boldsymbol{x}_i)$, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma = 0.4)$.

b) Basis: B-spline bases with dimensions $n \times k$, with $k = seq(40, 1000, by = 5)$.

c) Prediction horizons between 1 and 30 in steps of 1.

From the obtained results, we concluded:

1. The memory, like the effective dimension, only depends on the smoothing parameter and not on the size of the B-spline basis (provided that the basis is sufficiently large).
2. The memory does not depend on the prediction horizon.

3. The memory depends on the smoothing parameter. The smaller (larger) the smoothing parameter is, the smaller (greater) the influence of the past on the predicted values is.

In order to illustrate property 2, we use the previous mortality data set, Figure 7 shows the vector of weights for different prediction horizons, as we can see the memory is always the same and data prior to 1992 do not contribute to the prediction. To illustrate the third property we fit and forecast up to 2019 the log mortality rates by using different smoothing parameters, depending on the value of the smoothing parameter the memory is smaller or greater. As we can see in Figure 8, as the value of the smoothing parameter increases, the memory also increases.

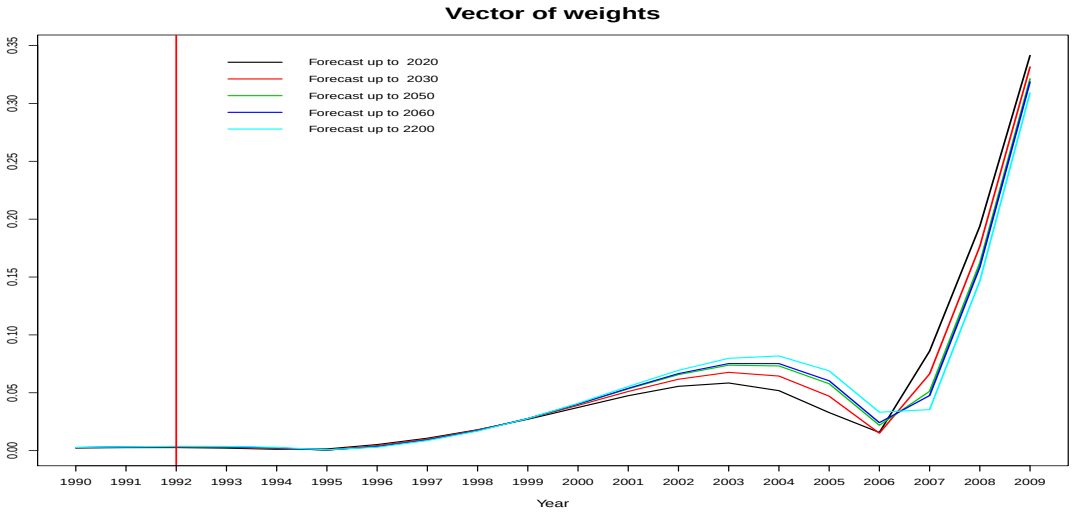


Figure 7: Vector of weights for different prediction horizons when we fit and forecast the log mortality rates of Spanish men aged 73.

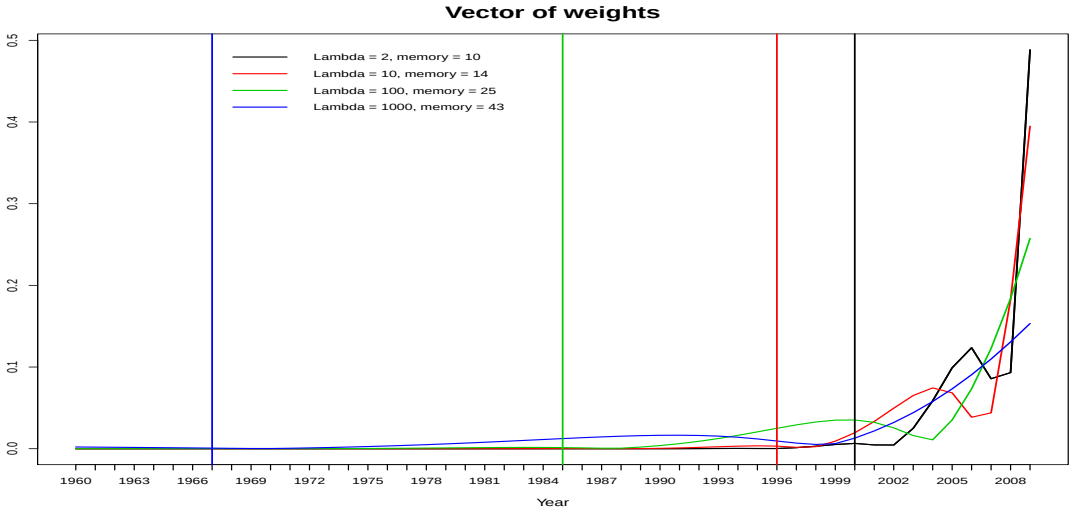


Figure 8: Vector of weights for different values of the smoothing parameter when we forecast 10 new observations of the log mortality rates of Spanish men aged 73.

## 6 Conclusions

Smoothing techniques have become a very popular tool for estimation of functions. 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 frameworks in which smoothing is carried out:

- Extend the basis used for regression and the penalty to control the smoothness in the framework of penalized regression based on quadratic penalties.
- Extend the fixed and random components in the context of mixed models.
- Define a Gaussian process for the extended set of random effects.

In the context of penalties based on differences between adjacent coefficients, we have proved the equivalence of all methods and we have seen that the order of penalty function, which is less relevant in the smoothing of data, is now critical, because the penalty function determines the form of the prediction.

We have also introduced the concept of "memory of a P-spline" as a tool to know how much known information we are using to predict. Through a simulation study we have been able to conclude that the memory just depends on the smoothing parameter, provided that the regression basis is sufficiently large.

To illustrate the methodology, the proved results and the concept of "memory of a P-spline", we have showed the performance of the missing value approach method with B-spline basis and penalties based on differences by using two examples based on real data sets.

Finally, the presented methodology is general and as further work we will focus on the extension of it to more general cases, for example, for correlated errors, for Poisson response variables or for multidimensional setting.

## Appendix

## A Proof of Theorem 1

*Proof.* Since with the transformation matrix (23) the extended fixed and random parts are the same in both methods, we just need to show that the fixed and random effects are equal in both methods.

Let us compute the covariance matrix $\boldsymbol{G}_+$ of the augmented random effects $\boldsymbol{\alpha}_+$, (22):

$$\boldsymbol{G}_+ = \sigma^2_{\boldsymbol{\alpha}}(\boldsymbol{\Omega}'_{+_r}\boldsymbol{D}'_+\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1} = \sigma^2_{\boldsymbol{\alpha}}\begin{bmatrix}\boldsymbol{G} & \boldsymbol{G}_{op}\\ \boldsymbol{G}_{po} & \boldsymbol{G}_{pp}\end{bmatrix}, \tag{A.1}$$

$\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r}$ is a squared matrix, so the inverse of $((\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})'\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1}$ is

$$((\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})'\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1} = (\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1}(\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{'-1} = (\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1}(\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1'}.$$

Using Lemma 8.5.4 of Harville (2000), we have that:

$$(\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r})^{-1} = \begin{bmatrix} \boldsymbol{D}\boldsymbol{\Omega}_r & \boldsymbol{0} \\ \boldsymbol{D}_1\boldsymbol{\Omega}_r & \boldsymbol{D}_2\boldsymbol{\Omega}_{p_r} \end{bmatrix}^{-1} = \begin{bmatrix} (\boldsymbol{D}\boldsymbol{\Omega}_r)^{-1} & \boldsymbol{0} \\ -(\boldsymbol{D}_2\boldsymbol{\Omega}_{p_r})^{-1}\boldsymbol{D}_1\boldsymbol{\Omega}_r(\boldsymbol{D}\boldsymbol{\Omega}_r)^{-1} & (\boldsymbol{D}_2\boldsymbol{\Omega}_{p_r})^{-1} \end{bmatrix}.$$

Therefore,

$$\begin{aligned} \boldsymbol{G} &= \boldsymbol{I}, \\ \boldsymbol{G}_{op} &= -\boldsymbol{\Omega}_r'\boldsymbol{D}_1', \\ \boldsymbol{G}_{po} &= -\boldsymbol{D}_1\boldsymbol{\Omega}_r, \\ \boldsymbol{G}_{pp} &= \boldsymbol{I} + \boldsymbol{D}_1\boldsymbol{\Omega}_r\boldsymbol{\Omega}_r'\boldsymbol{D}_1'. \end{aligned}$$

Notice that its inverse is:

$$\boldsymbol{G}_+^{-1} = \frac{1}{\hat{\sigma}^2_{\boldsymbol{\alpha}}}\begin{bmatrix} \boldsymbol{G}^{oo} & \boldsymbol{G}^{op} \\ \boldsymbol{G}^{po} & \boldsymbol{G}^{pp} \end{bmatrix} = \frac{1}{\hat{\sigma}^2_{\boldsymbol{\alpha}}}\begin{bmatrix} \boldsymbol{I} + \boldsymbol{\Omega}_r'\boldsymbol{D}_1'\boldsymbol{D}_1\boldsymbol{\Omega}_r & -\boldsymbol{\Omega}_r'\boldsymbol{D}_1' \\ \boldsymbol{D}_1\boldsymbol{\Omega}_r & \boldsymbol{I} \end{bmatrix}.$$

Now that we know $\boldsymbol{G}_+$, we just need to compute $\boldsymbol{V}_+^{-1}$ to know the expression of the extended fixed effects. We have that,

$$\boldsymbol{G}_+^{-1} + \boldsymbol{Z}_+'\boldsymbol{M}\boldsymbol{Z}_+ = \begin{bmatrix} \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}(\boldsymbol{I} + \boldsymbol{\Omega}_r'\boldsymbol{D}_1'\boldsymbol{D}_1\boldsymbol{\Omega}_r) + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}(\boldsymbol{B}\boldsymbol{\Omega}_r)'\boldsymbol{B}\boldsymbol{\Omega}_r & \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{\Omega}_r'\boldsymbol{D}_1' \\ \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{D}_1\boldsymbol{\Omega}_r & \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{K}_2 \\ \boldsymbol{K}_3 & \boldsymbol{K}_4 \end{bmatrix},$$

and that,

$$\boldsymbol{M}\boldsymbol{Z}_+ = \begin{bmatrix} \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{B}\boldsymbol{\Omega}_r & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}.$$

Defining $(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}_+'(\boldsymbol{W}\boldsymbol{R}_+\boldsymbol{W})^{-1}\boldsymbol{Z}_+)^{-1} = \begin{bmatrix} \boldsymbol{J}_1 & \boldsymbol{J}_2 \\ \boldsymbol{J}_3 & \boldsymbol{J}_4 \end{bmatrix}$, it follows that:

$$\boldsymbol{M}\boldsymbol{Z}_+(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}_+'\boldsymbol{M}\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}_+'\boldsymbol{M} = \begin{bmatrix} \frac{1}{\sigma^4_{\boldsymbol{\epsilon}}}\boldsymbol{B}\boldsymbol{\Omega}_r\boldsymbol{J}_1(\boldsymbol{B}\boldsymbol{\Omega}_r)' & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}.$$

Hence, we just need to know $\boldsymbol{J}_1$. Applying Theorem 8.5.11 given in Harville (2000):

$$\begin{aligned} \boldsymbol{J}_1^{-1} &= \boldsymbol{K}_1 - \boldsymbol{K}_2\boldsymbol{K}_4^{-1}\boldsymbol{K}_3 \\ &= \boldsymbol{K}_1 - \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{\Omega}_r'\boldsymbol{D}_1'(\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I})\frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{D}_1\boldsymbol{\Omega}_r \\ &= \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}(\boldsymbol{I} + \boldsymbol{\Omega}_r'\boldsymbol{D}_1'\boldsymbol{D}_1\boldsymbol{\Omega}_r) + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}(\boldsymbol{B}\boldsymbol{\Omega}_r)'\boldsymbol{B}\boldsymbol{\Omega}_r - \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{\Omega}_r'\boldsymbol{D}_1'\boldsymbol{D}_1\boldsymbol{\Omega}_r \\ &= \frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{I} + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}(\boldsymbol{B}\boldsymbol{\Omega}_r)'\boldsymbol{B}\boldsymbol{\Omega}_r, \end{aligned}$$

and, applying Theorem 18.2.8, given in Harville (2000) to compute $\boldsymbol{J}_1$:

$$\begin{aligned}\boldsymbol{J}_1 &= \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I} - \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I}(\boldsymbol{B\Omega}_r)'(\sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I} + \boldsymbol{B\Omega}_r\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I}(\boldsymbol{B\Omega}_r)')^{-1}\boldsymbol{B\Omega}_r\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I} \\ &= \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I} - (\sigma^2_{\boldsymbol{\alpha}})^2(\boldsymbol{B\Omega}_r)'(\sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I} + \boldsymbol{B\Omega}_r\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I}(\boldsymbol{B\Omega}_r)')^{-1}\boldsymbol{B\Omega}_r.\end{aligned}$$

Therefore:

$$\begin{aligned}\boldsymbol{V}_+^{-1} &= \boldsymbol{M} - \boldsymbol{M}\boldsymbol{Z}_+(\boldsymbol{G}_+^{-1} + \boldsymbol{Z}'_+\boldsymbol{M}\boldsymbol{Z}_+)^{-1}\boldsymbol{Z}'_+\boldsymbol{M} \\ &= \begin{bmatrix} \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{I} - \frac{1}{\sigma^4_{\boldsymbol{\epsilon}}}\boldsymbol{B\Omega}_r\left[\sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I} - \sigma^4_{\boldsymbol{\alpha}}(\boldsymbol{B\Omega}_r)'(\sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I} + \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{B\Omega}_r(\boldsymbol{B\Omega}_r)')^{-1}\boldsymbol{B\Omega}_r\right](\boldsymbol{B\Omega}_r)' & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{V}^*_{+11} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}.\end{aligned}$$

Moreover, as $\boldsymbol{V} = \sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}'$, with $\boldsymbol{G} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I}$:

$$\boldsymbol{V}^{-1} = \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{I} - \frac{1}{\sigma^4_{\boldsymbol{\epsilon}}}\boldsymbol{B\Omega}_r\left(\frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{I} + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}(\boldsymbol{B\Omega}_r)'\boldsymbol{B\Omega}_r\right)^{-1}(\boldsymbol{B\Omega}_r)'$$

By Theorem 18.2.8 given in Harville (2000),

$$\left(\frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{I} + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}(\boldsymbol{B\Omega}_r)'\boldsymbol{B\Omega}_r\right)^{-1} = \sigma^2_{\boldsymbol{\alpha}}\boldsymbol{I} - \sigma^4_{\boldsymbol{\alpha}}(\boldsymbol{B\Omega}_r)'\left(\sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{I} + \boldsymbol{B\Omega}_r\sigma^2_{\boldsymbol{\alpha}}(\boldsymbol{B\Omega}_r)'\right)^{-1}\boldsymbol{B\Omega}_r$$

i.e., $\boldsymbol{V}^{-1} = \boldsymbol{V}^*_{+11}$.

As we have proved that $\boldsymbol{V}_+^{-1} = \begin{bmatrix}\boldsymbol{V}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}$ it is straightforward to show that $\boldsymbol{\beta}_+ = \boldsymbol{\beta}$.

Moreover, by the MMM method we have that,

$$\begin{aligned}\boldsymbol{\alpha}_{+MMM} &= \boldsymbol{G}_+\boldsymbol{Z}'_+\boldsymbol{V}_+^{-1}(\boldsymbol{y}_+ - \boldsymbol{X}_+\boldsymbol{\beta}_+) \\ &\underset{\boldsymbol{\beta}=\boldsymbol{\beta}_+}{=} \sigma^2_{\boldsymbol{\alpha}}\begin{bmatrix}(\boldsymbol{B\Omega}_r)'\boldsymbol{V}^{-1} & \boldsymbol{O} \\ -\boldsymbol{D}_1\boldsymbol{\Omega}_r(\boldsymbol{B\Omega}_r)'\boldsymbol{V}^{-1} & \boldsymbol{O}\end{bmatrix}(\boldsymbol{y}_+ - \boldsymbol{X}_+\boldsymbol{\beta}) \\ &= \begin{bmatrix}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\boldsymbol{G}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\end{bmatrix} \\ &= \begin{bmatrix}\boldsymbol{\alpha}_{MM} \\ \boldsymbol{G}_{po}\boldsymbol{G}^{-1}\boldsymbol{\alpha}_{MM}\end{bmatrix}.\end{aligned}$$

As we wanted to show solutions given by MMM and MM are the same.

□

## B Proof of the equivalence of the restricted maximum likelihoods (24) and (25)

By the proof of theorem 1 we knok that $\boldsymbol{V}_+^{-1} = \begin{bmatrix}\boldsymbol{V}^{-1} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}$, therefore it is straightforward to prove that Part II and III of approximate restricted maximum likelihoods

(24) and (25) are equal. As $\boldsymbol{V}_+ \neq \boldsymbol{V}$, Part I of (24) and (25) are not equal, but its derivatives with respect to the parameters $(\sigma^2_{\boldsymbol{\epsilon}}, \sigma^2_{\boldsymbol{\alpha}})$ are equal:

Derivatives of Part I with respect to $\sigma^2$:

$$\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}|\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} = \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\right)$$

and

$$\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}_+|\right)}{\partial\sigma^2_{\boldsymbol{\epsilon}}} = \frac{1}{2}\text{trace}\left(\boldsymbol{V}_+^{-1}\frac{\partial\sigma^2_{\boldsymbol{\epsilon}}\boldsymbol{W}}{\partial\sigma^2_{\boldsymbol{\epsilon}}}\right) = \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\right),$$

the derivatives with respect to the correlation parameter, $\rho$, are analogous.

Derivatives of Part I with respect to $\sigma^2_{\boldsymbol{\alpha}}$:

$$\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}|\right)}{\partial\sigma^2_{\boldsymbol{\alpha}}} = \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right)$$

and

$$\begin{aligned}
\frac{\partial\left(\frac{1}{2}\log|\boldsymbol{V}_+|\right)}{\partial\sigma^2_{\boldsymbol{\alpha}}} &= \frac{1}{2}\text{trace}\left(\boldsymbol{V}_+^{-1}\boldsymbol{Z}_+\frac{\partial\boldsymbol{G}_+}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'_+\right) \\
&= \frac{1}{2}\text{trace}\left(\begin{bmatrix}\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}' & \boldsymbol{V}^{-1}\boldsymbol{Z}\left(\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'_1 + \frac{\partial\boldsymbol{G}_{op}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'_2\right) \\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}\right) \\
&= \frac{1}{2}\text{trace}\left(\boldsymbol{V}^{-1}\boldsymbol{Z}\frac{\partial\boldsymbol{G}}{\partial\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{Z}'\right).
\end{aligned}$$

## C Proof of corollary 2

*Proof.* • Differences of order 1.
Suppose a difference matrix with first order penalty $\boldsymbol{D}_+$ of dimensions $(c_+ - 1) \times c_+$,

$$\boldsymbol{D}_+ = \begin{bmatrix}\boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_1 & \boldsymbol{D}_2\end{bmatrix} = \begin{bmatrix}-1 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & -1 & 1 & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1\end{bmatrix},$$

where $\boldsymbol{D}_1$ has dimension $c_p \times c$, with $c_p$ the additional number of parameters in $\boldsymbol{\theta}_+$, and $\boldsymbol{D}_2$ has dimension $c_p \times c_p$:

$$\boldsymbol{D}_1 = \begin{bmatrix}0 & 0 & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0\end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix}1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & -1 & 1\end{bmatrix}.$$

Then, the additional vector of coefficients in (30) is:

$$\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}} = -\begin{bmatrix} 0 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & 0 & -1 \\ 0 & 0 & \cdots & 0 & -1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}_1 \\ \hat{\boldsymbol{\theta}}_2 \\ \vdots \\ \hat{\boldsymbol{\theta}}_{c-1} \\ \hat{\boldsymbol{\theta}}_c \end{bmatrix} = \hat{\boldsymbol{\theta}}_c \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix}.$$

Therefore, using differences of order 1 the new coefficients are equal to the last coefficient.

• Differences of order 2.
Suppose a difference matrix with second order penalty $\boldsymbol{D}_+$ of dimensions $(c_+-2)\times c_+$,

$$\boldsymbol{D}_+ = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{O} \\ \boldsymbol{D}_1 & \boldsymbol{D}_2 \end{bmatrix} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 1 & -2 & 1 & 0 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix},$$

where $\boldsymbol{D}_1$ has dimension $c_p \times c$, with $c_p$ the additional number of parameters in $\boldsymbol{\theta}_+$, and $\boldsymbol{D}_2$ has dimension $c_p \times c_p$:

$$\boldsymbol{D}_1 = \begin{bmatrix} 0 & 0 & \cdots & 1 & -2 \\ 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ -2 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & 2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

Then, the additional vector of coefficients in (30) is:

$$\hat{\boldsymbol{\theta}}_p = -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}} = -\begin{bmatrix} 0 & 0 & \cdots & 1 & -2 \\ \vdots & \vdots & \cdots & 2 & -3 \\ \vdots & \vdots & \cdots & 3 & -4 \\ 0 & 0 & \cdots & 4 & -5 \\ \vdots & \vdots & \cdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}_1 \\ \hat{\boldsymbol{\theta}}_2 \\ \vdots \\ \hat{\boldsymbol{\theta}}_{c-1} \\ \hat{\boldsymbol{\theta}}_c \end{bmatrix} = \hat{\boldsymbol{\theta}}_c \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} + (\hat{\boldsymbol{\theta}}_c - \hat{\boldsymbol{\theta}}_{c-1}) \begin{bmatrix} 1 \\ 2 \\ 3 \\ \vdots \end{bmatrix}.$$

Therefore, using differences of order 2 the new coefficients are a linear combination of the two last coefficients obtained after fitting the observed data.

• Differences of order 3.

Suppose a difference matrix with third order penalty, $\boldsymbol{D}_+$ of dimensions $(c_+-3)\times c_+$,

$$\boldsymbol{D}_+ = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -1 & 3 & -3 & 1 \end{bmatrix}.$$

In this case, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are:

$$\boldsymbol{D}_1 = \begin{bmatrix} 0 & 0 & 0 & \cdots & -1 & 3 & -3 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 3 \\ 0 & 0 & 0 & \cdots & 0 & 0 & -1 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{D}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ -3 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ 3 & -3 & 1 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 3 & -3 & 1 \end{bmatrix}.$$

Therefore, by (30):

$$\begin{aligned} \hat{\boldsymbol{\theta}}_p &= -\boldsymbol{D}_2^{-1}\boldsymbol{D}_1\hat{\boldsymbol{\theta}} = - \begin{bmatrix} 0 & \cdots & 0 & -1 & 3 & -3 \\ 0 & \cdots & 0 & -3 & 8 & -6 \\ 0 & \cdots & 0 & -6 & 15 & -10 \\ 0 & \cdots & 0 & -10 & 24 & -15 \\ 0 & \cdots & 0 & -15 & 35 & -21 \\ 0 & \cdots & 0 & -21 & 48 & -28 \\ 0 & \cdots & 0 & -28 & 63 & -36 \\ 0 & \cdots & 0 & -36 & 80 & -45 \\ 0 & \cdots & 0 & -45 & 99 & -55 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\theta}}_1 \\ \hat{\boldsymbol{\theta}}_2 \\ \hat{\boldsymbol{\theta}}_3 \\ \vdots \\ \hat{\boldsymbol{\theta}}_{c-2} \\ \hat{\boldsymbol{\theta}}_{c-1} \\ \hat{\boldsymbol{\theta}}_c \end{bmatrix} \\ &= \hat{\boldsymbol{\theta}}_c \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} + \frac{3\hat{\boldsymbol{\theta}}_c - 4\hat{\boldsymbol{\theta}}_{c-1} + \hat{\boldsymbol{\theta}}_{c-2}}{2} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{bmatrix} + \frac{\hat{\boldsymbol{\theta}}_c - 2\hat{\boldsymbol{\theta}}_{c-1} + \hat{\boldsymbol{\theta}}_{c-2}}{2} \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ \vdots \end{bmatrix}^2, \end{aligned}$$

in this case, the new coefficients are a linear combination of the last three coefficients obtained after fitting the observed values. The prediction is a quadratic polynomial.

□

## References

Ba, A., Sinn, M., Goude, Y., & Pompey, P. 2012. Adaptive Learning of Smoothing Functions: Application to Electricity Load Forecasting. *In:* Pereira, F., Burges, C.J.C., Bottou, L., & Weinberger., K.Q. (eds), *Advances in Neural Information Processing Systems 25.*

Caudel, K., & Frey, M. 2012 (June). Continuous Updates of Penalized Spline Regression for Flow Field Forecasting. *In: Proceedings of the 32th Annual International Symposium on Forecasting.*

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

Currie, I. D., & Durbán, M. 2002. Flexible Smoothing with $P$-Splines: A Unified Approach. *Statistical Modelling*, **2**, 333–349.

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

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

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

Girosi, F., Jones, M., & Poggio, T. 1995. Regularization Theory and Neural Networks Architectures. *Neural Computation*, 219–269.

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.

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

Jones, D. R., Schonlau, M, & William, J. 1998. Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimiztion*, 455–492.

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

Poggio, T., & Girosi, F. 1990. Networks for approximation and learning. *Proceedings of IEEE*, 1481–1497.

Pollice, A., & Bilancia, M. 2001. Kriging with mixed effects models. *Statistica*, 405–429.

Sacks, J., J., Welch W., J., Mitchell T., & H., Wynn P. 1989. Design and Analysis of Computer Experiments. *Statistical Science*, 409–435.

Yi, G., Shi, J.Q., & Choi, T. 2011. Penalized Gaussian Process Regression and Classification for High-Dimensional Nonlinear Data. *Biometrics*, 1285–1294.