Complementary material

A general framework for prediction in penalized regression

Alba Carballo[1] Maria Durban[1] Göeran Kauermann[2] 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

This document contains complementary material to the paper "A general framework for prediction in penalized regression". The proofs of Corollary 2 and Theorem 3 are given.

1 Proof of Corollary 2

Proof. Let us consider the proof for different penalty orders.

- 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 (2.9) is:

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

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 (2.9) is:

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

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 (2.9):

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

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. □

2 Proof of Theorem 3

Proof. Since with the transformation matrix (3.10) 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 $\boldsymbol{\alpha}_+$, (3.9):

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

$\boldsymbol{D}_+\boldsymbol{\Omega}_{+_r}$ 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,

$$\boldsymbol{G} = \boldsymbol{I}, \quad \boldsymbol{G}_{op} = -\boldsymbol{\Omega}_r^{'}\boldsymbol{D}_1^{'}, \quad \boldsymbol{G}_{po} = -\boldsymbol{D}_1\boldsymbol{\Omega}_r, \quad \boldsymbol{G}_{pp} = \boldsymbol{I} + \boldsymbol{D}_1\boldsymbol{\Omega}_r\boldsymbol{\Omega}_r^{'}\boldsymbol{D}_1^{'}.$$

Notice that its inverse is:

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

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

We have that,

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

and that,

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

Defining $\left(\frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{G}_+^{-1} + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\right)^{-1} = \begin{bmatrix} \boldsymbol{J}_1 & \boldsymbol{J}_2 \\ \boldsymbol{J}_3 & \boldsymbol{J}_4 \end{bmatrix}$, it follows that:

$$\frac{1}{\sigma^4_{\boldsymbol{\epsilon}}}\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\left(\frac{1}{\sigma^2_{\boldsymbol{\alpha}}}\boldsymbol{G}_+^{-1} + \frac{1}{\sigma^2_{\boldsymbol{\epsilon}}}\boldsymbol{Z}'_+\boldsymbol{R}_+^{-1}\boldsymbol{Z}_+\right)^{-1}\boldsymbol{Z}'_+\boldsymbol{R}_+^{-1} \quad = \quad \begin{bmatrix} \frac{1}{\sigma^4_{\boldsymbol{\epsilon}}}\boldsymbol{B\Omega}_r\boldsymbol{J}_1(\boldsymbol{B\Omega}_r)' & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} \end{bmatrix}.$$

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

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

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:

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

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

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

By Theorem 18.2.8 given in Harville (2000),

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

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

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

Moreover, by the extended mixed model approach we have that,

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

As we wanted to show solutions given by extended mixed model approach and mixed model approach are the same.

Let us prove that the variance components $(\sigma^2_{\boldsymbol{\epsilon}},\,\sigma^2_{\boldsymbol{\alpha}})$ that maximize the approximate restricted maximum likelihoods (3.11) and (3.12) are equal. Consider the parts of both expressions as follows:

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

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

since $\tilde{\boldsymbol{V}}^-_+=\begin{bmatrix}\boldsymbol{V}^{-1} & \boldsymbol{O}\\ \boldsymbol{O} & \boldsymbol{O}\end{bmatrix}$, it is straightforward to prove that Part II and Part III of both restricted maximum likelihoods are equal. As $\tilde{\boldsymbol{V}}_+\neq\boldsymbol{V}$, Part I of (3.11) and (3.12) 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 $\sigma^2_{\boldsymbol{\epsilon}}$:

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

and

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

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

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

and

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

□

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.