"The Laplace-P-spline methodology for fast approximate Bayesian inference in additive partial linear models"

Gressani, Oswaldo ; Lambert, Philippe

ABSTRACT

Multiple linear regression is among the cornerstones of statistical model building. Whether from a descriptive or inferential perspective, it is certainly the most widespread approach to analyze the inuence of a collection of explanatory variables on a response. The straightforward interpretability in conjunction with the simple and elegant mathematics of least squares created room for a well appreciated toolbox with an ubiquitous presence in various scientific fields. In this article, the linear dependence assumption of the response variable with respect to the covariates is relaxed and replaced by an additive architecture of univariate smooth functions of predictor variables. An approximate Bayesian approach combining Laplace approximations and P-splines is used for inference in this additive partial linear model class. The analytical availability of the gradient and Hessian of the posterior penalty vector allows for a fast and efficient exploration of the penalty space, which in turn yields accurate point and set estimates of latent field variables. Different simulation settings confirm the statistical performance of the Laplace-P-spline approach and the methodology is applied on mortality data.

THE LAPLACE-P-SPLINE METHODOLOGY FOR FAST APPROXIMATE BAYESIAN INFERENCE IN ADDITIVE PARTIAL LINEAR MODELS

Gressani, O. and P. LAMBERT

DISCUSSION PAPER | 2020 / 20

UCLouvain

ISBA | LiDAM
Louvain Institute of Data Analysis and Modeling in economics and statistics

The Laplace-P-spline methodology for fast approximate Bayesian inference in additive partial linear models

Oswaldo Gressani [a,*] and Philippe Lambert [a,b]

[a] Institute of Statistics, Biostatistics and Actuarial Sciences (ISBA),
Université catholique de Louvain, Voie du Roman Pays 20,
B-1348, Louvain-la-Neuve, Belgium

[b] Institut de Recherche en Sciences Sociales (IRSS),
Méthodes Quantitatives en Sciences Sociales,
Université de Liège, Place des Orateurs 3,
B-4000, Liège, Belgium

Abstract

Multiple linear regression is among the cornerstones of statistical model building. Whether from a descriptive or inferential perspective, it is certainly the most widespread approach to analyze the influence of a collection of explanatory variables on a response. The straightforward interpretability in conjunction with the simple and elegant mathematics of least squares created room for a well appreciated toolbox with an ubiquitous presence in various scientific fields. In this article, the linear dependence assumption of the response variable with respect to the covariates is relaxed and replaced by an additive architecture of univariate smooth functions of predictor variables. An approximate Bayesian approach combining Laplace approximations and P-splines is used for inference in this additive partial linear model class. The analytical availability of the gradient and Hessian of the posterior penalty vector allows for a fast and efficient exploration of the penalty space, which in turn yields accurate point and set estimates of latent field variables. Different simulation settings confirm the statistical performance of the Laplace-P-spline approach and the methodology is applied on mortality data.

*Corresponding author. E-mail address: *oswaldo_gressani@hotmail.fr*

1 Introduction

The dawn of additive models traces back to Friedman and Stuetzle (1981) who suggest a projection pursuit regression technique in which the response is approximated by a sum of univariate functions of one-dimensional projections of the vector of covariates. The paper by Buja et al. (1989) investigates a class of smoothers in additive models and studies the properties of the iterative backfitting algorithm proposed in Breiman and Friedman (1985) as the *Alternating Conditional Expectation* algorithm. Backfitting is a well-known tool for estimating the additive components of the model and imposed itself as a benchmark strategy in the literature with successful applications. Tjøstheim and Auestad (1994) and Linton and Nielsen (1995) independently suggested an alternative non-recursive estimation plan that consists in estimating the regression surface by a multidimensional smoother in a first step and integrate it in a second step to obtain an estimator of the marginal smooth function of interest, a method coined "marginal integration". Complete book-length treatment of additive models are found in Hastie and Tibshirani (1990) and Wood (2017).

We adapt the Laplace-P-spline (LPS) approach to additive models with Gaussian errors and develop a fast and flexible methodology for approximate Bayesian inference in this model class. Great efforts have been invested in the derivation of analytical formulas for the gradient and Hessian of the posterior penalty vector, which offers a nonnegligible computational gain when exploring the posterior penalty space. Moments of a skew-normal family of random variables are used to accurately approximate the posterior distribution of penalty parameters, thereby capturing the inherent asymmetric patterns. In Section 2.1, the Bayesian-P-spline additive model is introduced and a method is proposed to overcome identifiability problems. In Section 2.2 the priors on the penalty parameters are defined and the likelihood function is derived together with the conditional posterior distribution of the latent field vector. Section 3.1 is dedicated to the posterior of the hyperparameter vector. The nuisance parameters are integrated out in Section 3.2 and the gradient and Hessian of the penalty vector are obtained in closed-form in Section 3.3. Section 3.4 proposes a strategy to explore the posterior penalty vector based on skew-normal matching moments. In Section 4 the approximate posterior of the latent field is derived and Section 4.1 covers the derivation of pointwise credible intervals for marginal latent field elements and smooth functions. Section 5 implements a simulation study to assess the performance of the proposed methodology and Section 6 illustrates the LPS approach on mortality data before concluding.

2 The Bayesian P-spline additive model

2.1 Additive structure and latent field prior

Let us consider the set $\mathcal{D} = \{(y_i, \mathbf{x}_i, \mathbf{z}_i)_{i=1}^n\}$ of n independent observations, where y_i is a response variable, $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^\top$ a vector of continuous covariates and $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^\top$ a vector of additional continuous or categorical covariates. Each covariate group is assumed deterministic such that we are in a fixed design. The additive model is written as follows:

$$y_i = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + f_1(x_{i1}) + \cdots + f_q(x_{iq}) + \varepsilon_i, \tag{1}$$

for $i = 1, \ldots, n$, with regression coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$ and $\{\varepsilon_i\}_{i=1}^n$ a sequence of independent and Gaussian errors with mean 0, unknown variance $\sigma^2 < +\infty$ and precision $\tau = 1/\sigma^2$. The above model is also referred to as the additive partial linear model (explored among others in Opsomer and Ruppert, 1999; Fan and Li, 2003; Liang et al., 2008; Ma and Yang, 2011) as one part is specified parametrically and the remaining additive components are unknown smooth functions. Following the P-spline approach of Eilers and Marx (1996), the additive smooth components $f_j,\ j = 1, \ldots, q$ are approximated by a large number of cubic B-splines and a discrete penalty on neighboring spline coefficients is imposed to counterbalance the roughness of the fit:

$$f_j(x_{ij}) = \sum_{k=1}^{K} \theta_{jk} b_{jk}(x_{ij}), \quad j = 1, \ldots, q, \tag{2}$$

where the number K of basis functions $b_{jk}(\cdot)$ is the same for every f_j. The vector of B-spline amplitudes associated to function f_j is given by $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK})^\top$, while the set of all spline coefficients in the additive model is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_q^\top)^\top$ and the vector of B-spline basis functions at x_{ij} is $\mathbf{b}_j(x_{ij}) = (b_{j1}(x_{ij}), \ldots, b_{jK}(x_{ij}))^\top$. The roughness penalty on finite differences of the coefficients of adjacent B-spline coefficients is $\boldsymbol{\theta}^\top \mathcal{P}(\boldsymbol{\lambda})\boldsymbol{\theta}$, with block diagonal matrix $\mathcal{P}(\boldsymbol{\lambda})$ that can be expressed compactly using a Kronecker product:

$$\mathcal{P}(\boldsymbol{\lambda}) := \operatorname{diag}(\lambda_1, \ldots, \lambda_q) \otimes P = \begin{pmatrix} \lambda_1 P & 0 & \ldots & 0 \\ 0 & \lambda_2 P & \ldots & 0 \\ \vdots & \ldots & \ddots & 0 \\ 0 & \ldots & 0 & \lambda_q P \end{pmatrix},$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^\top$ is a vector of positive penalty parameters and $P = D_r^\top D_r + \epsilon I_K$ is a penalty matrix resulting from the product of rth order difference matrices D_r of dimension $(K-r) \times K$. Adding a diagonal perturbation ϵI_K (with $\epsilon = 10^{-6}$, say) ensures that P is a full rank matrix. In a Bayesian setting, Lang and Brezger (2004) suggest to interpret the roughness penalty as a multivariate Gaussian prior on the spline coefficients $\boldsymbol{\theta}|\boldsymbol{\lambda}, \tau \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\big(0, (\tau \mathcal{P}(\boldsymbol{\lambda}))^{-1}\big)$. Also, a Gaussian prior is imposed on the regression coefficients $\boldsymbol{\beta}|\tau \sim \mathcal{N}_{\dim(\boldsymbol{\beta})}(0, (\tau V_{\boldsymbol{\beta}})^{-1})$ (see for instance Jackman, 2009 p.104 or O'Hagan et al., 2004) with matrix $V_{\boldsymbol{\beta}} = \zeta I_{p+1}$ and small precision (say $\zeta = 10^{-5}$). The latent field of the model is written as $\boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ and includes the regression and spline coefficients with prior distribution $\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\big(0, (\tau Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\big)$ and the following matrix:

$$Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}} := Q_{\boldsymbol{\xi}}(\boldsymbol{\lambda}) = \begin{pmatrix} V_{\boldsymbol{\beta}} & 0 \\ 0 & \mathcal{P}(\boldsymbol{\lambda}) \end{pmatrix}.$$

Without loss of generality, the covariates $\mathbf{z}_i$ are mean centered. Let $\bar{z}_l = n^{-1} \sum_{i=1}^{n} z_{il},\ l = 1, \ldots, p$ and write the centered design matrix Z and B-spline matrices $B_j,\ j = 1, \ldots, q$ as:

$$Z = \begin{pmatrix} 1 & (z_{11} - \bar{z}_1) & \ldots & (z_{1p} - \bar{z}_p) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (z_{n1} - \bar{z}_1) & \ldots & (z_{np} - \bar{z}_p) \end{pmatrix}, B_j = \begin{pmatrix} b_{j1}(x_{1j}) & \ldots & b_{jK}(x_{1j}) \\ \vdots & \vdots & \vdots \\ b_{j1}(x_{nj}) & \ldots & b_{jK}(x_{nj}) \end{pmatrix}.$$

The additive model in (1) suffers from an identifiability issue. This can be easily illustrated through the simple model $E(y) = \beta_0 + f(x)$. Assume our goal is to estimate the expected value $E(y)$ from a sample $\{(x_i, y_i)\}_{i=1}^n$. Let c be any arbitrary constant and denote by $\tilde{\beta}_0 = \beta_0 - c$ and $\tilde{f}(x) = f(x) + c$. It follows that $E(y) = \tilde{\beta}_0 + \tilde{f}(x)$ for any c, such that there exists an infinite number of configurations for $\tilde{\beta}_0$ and $\tilde{f}$ yielding the same expected value, meaning that the model "parameters" cannot be uniquely identified and estimated for a given data set. To reach an identifiable model, we follow an approach similar to Durbán and Currie (2003) and define the centered B-spline matrices:

$$\widetilde{B}_j = B_j - (\mathbf{1}_n \mathbf{1}_L^\top / L)\breve{B}_j, \quad j = 1, \ldots, q,$$

where $\mathbf{1}_n$ and $\mathbf{1}_L$ are column vectors of ones of length n and L respectively and $\breve{B}_j$ is a B-spline matrix computed on a fine grid $\breve{x}_{1j}, \ldots, \breve{x}_{Lj}$ of equidistant values on the domain of f_j. The centered matrix can be written as:

$$\widetilde{B}_j = \begin{pmatrix} b_{j1}(x_{1j}) - \frac{1}{L}\sum_{l=1}^L b_{j1}(\breve{x}_{lj}) & \ldots & b_{jK}(x_{1j}) - \frac{1}{L}\sum_{l=1}^L b_{jK}(\breve{x}_{lj}) \\ \vdots & \vdots & \vdots \\ b_{j1}(x_{nj}) - \frac{1}{L}\sum_{l=1}^L b_{j1}(\breve{x}_{lj}) & \ldots & b_{jK}(x_{nj}) - \frac{1}{L}\sum_{l=1}^L b_{jK}(\breve{x}_{lj}) \end{pmatrix}.$$

We denote by $\widetilde{\mathbf{b}}_j(x_{ij})^\top$ the ith row of matrix $\widetilde{B}_j$. Hence, the ith entry of the vector $\widetilde{B}_j \boldsymbol{\theta}_j$ is given by:

$$\widetilde{\mathbf{b}}_j(x_{ij})^\top \boldsymbol{\theta}_j = \sum_{k=1}^K \theta_{jk} b_{jk}(x_{ij}) - \frac{1}{L}\sum_{l=1}^L \sum_{k=1}^K \theta_{jk} b_{jk}(\breve{x}_{lj})$$

and according to (2), the identifiability constraint is translated as $\tilde{f}_j(x_{ij}) = f_j(x_{ij}) - L^{-1}\sum_{l=1}^L f_j(\breve{x}_{lj})$, i.e. the additive functional components are centered around their average value (computed over a fine equidistant grid). To see how this solves the identifiability problem consider again the simple model $E(y) = \beta_0 + f(x) - \bar{f}$, with $\bar{f} = L^{-1}\sum_{l=1}^L f(\breve{x}_l)$ the average of f over a fine grid. Adding c to the intercept and subtracting the same amount from f yields:

$$\begin{aligned} \widetilde{E}(y) &= \beta_0 + c + f(x) - c - L^{-1}\sum_{l=1}^L (f(\breve{x}_l) - c) \\ &= \beta_0 + c + (f(x) - \bar{f}), \end{aligned}$$

such that $E(y) \neq \widetilde{E}(y)$. Centering the B-spline matrices implies a rank reduction as stated in the following proposition:

Proposition (Rank-reduction due to centering)
The rank of the centered B-spline matrix $\widetilde{B}_j$ is $K - 1$.

Proof:
Let us first use the property that $\mathbf{1}_n = B_j \mathbf{1}_K$, i.e. the sum over the rows of matrix B_j is equal to one, and write the centered matrix as follows:

$$\begin{aligned}\widetilde{B}_j &= B_j - B_j(\mathbf{1}_K \mathbf{1}_L^\top / L)\breve{B}_j \\ &= B_j(I_K - \mathcal{B}),\end{aligned}$$

where $\mathcal{B} = (L^{-1}\mathbf{1}_K \mathbf{1}_L^\top)\breve{B}_j$ is a $K \times K$ idempotent matrix. Indeed:

$$\begin{aligned}\mathcal{B}\mathcal{B} &= L^{-1}L^{-1}\mathbf{1}_K \mathbf{1}_L^\top \breve{B}_j \mathbf{1}_K \mathbf{1}_L^\top \breve{B}_j \\ &= L^{-1}L^{-1}\mathbf{1}_K(\mathbf{1}_L^\top \mathbf{1}_L)\mathbf{1}_L^\top \breve{B}_j \text{ using } \mathbf{1}_L = \breve{B}_j \mathbf{1}_K \\ &= (L^{-1}\mathbf{1}_K \mathbf{1}_L^\top)\breve{B}_j \\ &= \mathcal{B}.\end{aligned}$$

Provided the Schoenberg-Whitney conditions are satisfied, the B-spline matrix B_j will have full column rank K (see Ma and Kruth, 1995). Using the product property of ranks, it follows that $\text{rank}(\widetilde{B}_j) = \text{rank}(I_K - \mathcal{B})$. As $\mathcal{B}$ is idempotent, $(I_K - \mathcal{B})$ is also idempotent and so its rank is equal to its trace:

$$\begin{aligned}\text{rank}(\widetilde{B}_j) &= \text{rank}(I_K - \mathcal{B}) \\ &= \text{Tr}(I_K - \mathcal{B}) \\ &= \text{Tr}(I_K) - \text{Tr}(L^{-1}\mathbf{1}_K \mathbf{1}_L^\top \breve{B}_j) \\ &= K - L^{-1}\text{Tr}(\breve{B}_j \mathbf{1}_K \mathbf{1}_L^\top) \\ &= K - L^{-1}\text{Tr}(\mathbf{1}_L \mathbf{1}_L^\top) \\ &= K - 1. \quad \square\end{aligned}$$

To ensure that all the spline coefficients can be estimated in a unique way, we follow Wood (2017) and fix the Kth element of each spline vector $\boldsymbol{\theta}_j$ to zero and delete the Kth column in $\widetilde{B}_j$ and difference matrix D_r. Hence $\widetilde{B}_j$ has $K-1$ columns and the latent vector has dimension $\dim(\boldsymbol{\xi}) = q \times (K-1) + p + 1$. Taking the identifiability constraint into account, the ith entry of the vector $\widetilde{B}_j \boldsymbol{\theta}_j$ becomes:

$$\tilde{b}_j(x_{ij})^\top \boldsymbol{\theta}_j = \sum_{k=1}^{K-1} \theta_{jk} b_{jk}(x_{ij}) - \frac{1}{L} \sum_{l=1}^{L} \sum_{k=1}^{K-1} \theta_{jk} b_{jk}(\breve{x}_{lj}). \tag{3}$$

With the identifiability constraint and the centered Z matrix, the additive model in (1) can be expressed compactly as:

$$\begin{aligned}\mathbf{y} &= Z\boldsymbol{\beta} + \widetilde{B}_1 \boldsymbol{\theta}_1 + \cdots + \widetilde{B}_q \boldsymbol{\theta}_q + \boldsymbol{\varepsilon} \\ &= B\boldsymbol{\xi} + \boldsymbol{\varepsilon},\end{aligned} \tag{4}$$

where B is a side by side configuration of design matrices, $B = [Z : \widetilde{B}_1 : \cdots : \widetilde{B}_q]$ and corresponds to the full design matrix of the model. In the next section, we summarize the full Bayesian model and proceed with the derivation of the conditional posterior distribution of the latent field.

2.2 Latent field conditional posterior

The following priors are used for the penalty parameters $\lambda_j|\delta_j \sim \mathcal{G}(\nu/2,(\nu\delta_j)/2)$, $j = 1,\ldots,q$ and $\delta_j \sim \mathcal{G}(a_\delta, b_\delta)$, $\quad j = 1,\ldots,q$ with $a_\delta = b_\delta = 10^{-4}$ and $\nu = 3$. Moreover, we use Jeffreys' prior for the precision $p(\tau) \propto \tau^{-1}$ and write the hyperparameter vector as $\boldsymbol{\eta} = (\boldsymbol{\lambda}^\top, \boldsymbol{\delta}^\top, \tau)^\top$, where $\boldsymbol{\delta} = (\delta_1,\ldots,\delta_q)^\top$. The full Bayesian model is:

$$
\begin{aligned}
&y_i|\boldsymbol{\xi},\tau \sim \mathcal{N}_1\Big(\beta_0 + \sum_{l=1}^{p}\beta_l z_{il} + \sum_{j=1}^{q} b_j(x_{ij})^\top \boldsymbol{\theta}_j, \tau^{-1}\Big), \quad i = 1,\ldots,n,\\
&\boldsymbol{\theta}|\boldsymbol{\lambda},\tau \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\big(0,(\tau\mathcal{P}(\boldsymbol{\lambda}))^{-1}\big),\\
&\boldsymbol{\xi}|\boldsymbol{\lambda},\tau \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\big(0,(\tau Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\big),\\
&\lambda_j|\delta_j \sim \mathcal{G}(\nu/2,(\nu\delta_j)/2), \quad j = 1,\ldots,q,\\
&\delta_j \sim \mathcal{G}(a_\delta, b_\delta), \quad j = 1,\ldots,q,\\
&p(\tau) \propto \tau^{-1}.
\end{aligned}
$$

Taking into account the centering of the covariates in the linear part and the identifiability constraint of the smooth functions, the likelihood of the model is written as:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\xi},\tau;\mathcal{D}) &= \prod_{i=1}^{n} \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\Big\{-\frac{\tau}{2}\Big(y_i - \Big(\beta_0 + \sum_{l=1}^{p}\beta_l(z_{il}-\bar{z}_l)\\
&\qquad + \sum_{j=1}^{q}\tilde{b}_j(x_{ij})^\top\boldsymbol{\theta}_j\Big)\Big)^2\Big\}\\
&\propto \tau^{\frac{n}{2}} \exp\Big\{-\frac{\tau}{2}(\mathbf{y}-B\boldsymbol{\xi})^\top(\mathbf{y}-B\boldsymbol{\xi})\Big\}.
\end{aligned}
$$

The conditional posterior distribution of the latent field can be obtained as follows:

$$
\begin{aligned}
p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) &= \frac{\mathcal{L}(\boldsymbol{\xi},\tau;\mathcal{D})p(\boldsymbol{\xi},\boldsymbol{\lambda},\tau)}{p(\boldsymbol{\lambda},\tau,\mathcal{D})}\\
&\propto \mathcal{L}(\boldsymbol{\xi},\tau;\mathcal{D})\ p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau).
\end{aligned}
$$

Using the previously specified latent field prior and likelihood we get:

$$
\begin{aligned}
p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) &\propto \exp\Big(-\frac{\tau}{2}\big(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top B\boldsymbol{\xi} + \boldsymbol{\xi}^\top B^\top B\boldsymbol{\xi}\big) - \frac{\tau}{2}\boldsymbol{\xi}^\top Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\boldsymbol{\xi}\Big)\\
&\propto \exp\Big(\tau\mathbf{y}^\top B\boldsymbol{\xi} - \frac{\tau}{2}\boldsymbol{\xi}^\top(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})\boldsymbol{\xi}\Big). \qquad (5)
\end{aligned}
$$

Note that (5) is the exponential of a quadratic form in $\boldsymbol{\xi}$ and can be written as a Gaussian distribution. To find the mean vector we solve $\nabla_{\boldsymbol{\xi}} \log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) = 0$ and obtain $\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} = (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}B^\top\mathbf{y}$. The precision is $-\nabla_{\boldsymbol{\xi}}^2 \log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) = \tau(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})$ and so the conditional posterior of the latent field is characterized by the following Gaussian distribution:

$$
(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}, \tau^{-1}(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\right). \qquad (6)
$$

3 Posterior of the penalty vector

The aim of this section is to derive the posterior of the hyperparameter vector $\boldsymbol{\eta}$, an essential step to obtain the joint posterior of the latent field. First, we give the expression of $p(\boldsymbol{\eta}|\mathcal{D})$ and show how it can be integrated with respect to the nuisance hyperparameters $\boldsymbol{\delta}$ and τ resulting in a posterior for the roughness penalty vector. The gradient and Hessian of the posterior penalty are then analytically derived and used to compute the posterior mode through a Newton-Raphson algorithm.

3.1 Posterior of the full hyperparameter vector

The posterior of the full hyperparameter vector $\boldsymbol{\eta}$ is:

$$\begin{aligned} p(\boldsymbol{\eta}|\mathcal{D}) &= \frac{p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathcal{D})}{p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})} \\ &= \frac{\mathcal{L}(\boldsymbol{\xi}, \tau; \mathcal{D}) p(\boldsymbol{\xi}, \boldsymbol{\eta})}{p(\mathcal{D}) p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})} \\ &= \frac{\mathcal{L}(\boldsymbol{\xi}, \tau; \mathcal{D}) p(\boldsymbol{\xi}|\boldsymbol{\eta}) p(\boldsymbol{\lambda}, \boldsymbol{\delta}|\tau) p(\tau)}{p(\mathcal{D}) p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})}, \end{aligned}$$

where $p(\boldsymbol{\xi}|\boldsymbol{\eta}) = p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \boldsymbol{\delta}, \tau) = p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau)$ as $\boldsymbol{\xi} \perp \boldsymbol{\delta}|\boldsymbol{\lambda}, \tau$ and $p(\boldsymbol{\lambda}, \boldsymbol{\delta}|\tau) = p(\boldsymbol{\lambda}, \boldsymbol{\delta})$ as $\boldsymbol{\lambda}, \boldsymbol{\delta} \perp \tau$. Hence, the expression becomes:

$$p(\boldsymbol{\eta}|\mathcal{D}) \propto \frac{\mathcal{L}(\boldsymbol{\xi}, \tau; \mathcal{D}) p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau) \left(\prod_{j=1}^{q} p(\lambda_j|\delta_j) \right) \left(\prod_{j=1}^{q} p(\delta_j) \right) p(\tau)}{p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau, \mathcal{D})},$$

where $p(\lambda_j|\delta_j) \propto \delta_j^{\frac{\nu}{2}} \lambda_j^{(\frac{\nu}{2}-1)} \exp\big(-\frac{\nu}{2}\delta_j\lambda_j \big)$ and $p(\delta_j) \propto \delta_j^{a_\delta - 1} \exp(-b_\delta \delta_j)$. Note also that:

$$\begin{aligned} \left(\prod_{j=1}^{q} p(\lambda_j|\delta_j) \right) \left(\prod_{j=1}^{q} p(\delta_j) \right) \propto & \left(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j \Big(b_\delta + \frac{\nu}{2}\lambda_j \Big) \Big) \right) \\ & \times \left(\prod_{j=1}^{q} \lambda_j^{(\frac{\nu}{2}-1)} \right). \end{aligned}$$

Following Rue et al. (2009), the posterior of the hyperparameter vector can be evaluated around the mode of the conditional posterior of the latent field, namely $p(\boldsymbol{\eta}|\mathcal{D})\big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}}$. Using the previously derived expressions of the model:

$$\begin{aligned} p(\boldsymbol{\eta}|\mathcal{D})\big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}} \propto\ & \tau^{\frac{n}{2}} \exp\Big(-\frac{\tau}{2}\mathbf{y}^\top\mathbf{y} + \tau\mathbf{y}^\top B\boldsymbol{\xi} - \frac{\tau}{2}\boldsymbol{\xi}^\top B^\top B\boldsymbol{\xi} \Big) \Big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}} \\ & \times \tau^{\frac{\dim(\boldsymbol{\xi})}{2}} |Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} \exp\Big(-\frac{\tau}{2}\boldsymbol{\xi}^\top Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\boldsymbol{\xi} \Big) \Big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}} \\ & \times \left(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j \Big(b_\delta + \frac{\nu}{2}\lambda_j \Big) \Big) \right) \left(\prod_{j=1}^{q} \lambda_j^{(\frac{\nu}{2}-1)} \right) \\ & \times \tau^{-1} \tau^{-\frac{\dim(\boldsymbol{\xi})}{2}} |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}. \end{aligned}$$

Replacing $\boldsymbol{\xi}$ by $\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} = (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top \mathbf{y}$ in the above expression, one obtains:

$$
\begin{aligned}
p(\boldsymbol{\eta}|\mathcal{D})\big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}} \propto\, & \tau^{(\frac{n}{2}-1)} |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} |Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} \Bigg(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \\
& \times \exp\Big(-\delta_j \Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big)\Bigg) \left(\prod_{j=1}^{q} \lambda_j^{(\frac{\nu}{2}-1)} \right) \\
& \times \exp\Big(-\frac{\tau}{2}\mathbf{y}^\top \mathbf{y} + \tau \mathbf{y}^\top B (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top \mathbf{y} - \frac{\tau}{2} \mathbf{y}^\top B \\
& \qquad \times (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}) (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top \mathbf{y}\Big) \\
\propto\, & \tau^{(\frac{n}{2}-1)} |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} |Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} \Bigg(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \\
& \times \exp\Big(-\delta_j \Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big)\Bigg) \times \left(\prod_{j=1}^{q} \lambda_j^{(\frac{\nu}{2}-1)} \right) \exp\Big(-\frac{\tau}{2}\mathbf{y}^\top \mathbf{y} \\
& + \tau \mathbf{y}^\top B (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top \mathbf{y} - \frac{\tau}{2} \mathbf{y}^\top B (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top \mathbf{y}\Big)
\end{aligned}
$$

$$
\begin{aligned}
\propto \quad & \tau^{(\frac{n}{2}-1)} |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} |Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} \left(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j \Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big) \right) \\
& \times \left(\prod_{j=1}^{q} \lambda_j^{(\frac{\nu}{2}-1)} \right) \exp\Big(-\frac{\tau}{2}\mathbf{y}^\top \mathbf{y} + \frac{\tau}{2} \mathbf{y}^\top B (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top \mathbf{y}\Big) \\
\propto \quad & \tau^{(\frac{n}{2}-1)} |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} |Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} \left(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j \Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big) \right) \\
& \times \left(\prod_{j=1}^{q} \lambda_j^{(\frac{\nu}{2}-1)} \right) \exp\Big(-\frac{\tau}{2}\mathbf{y}^\top \big(I_n - B (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top\big) \mathbf{y}\Big).
\end{aligned}
$$

Let us define the scalar function $\phi(\boldsymbol{\lambda}) := \frac{1}{2}\mathbf{y}^\top \big(I_n - B(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top\big)\mathbf{y}$ and write compactly:

$$
\begin{aligned}
p(\boldsymbol{\eta}|\mathcal{D})\big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}} \propto\, & |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} |Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} \left(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j \Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big) \right) \\
& \times \left(\prod_{j=1}^{q} \lambda_j^{(\frac{\nu}{2}-1)} \right) \tau^{(\frac{n}{2}-1)} \exp\big(-\tau\phi(\boldsymbol{\lambda})\big).
\end{aligned} \tag{7}
$$

3.2 Integration with respect to the nuisance parameters

The nuisance parameter τ can be integrated out from (7) as expression $\tau^{(\frac{n}{2}-1)}\exp\big(-\tau\phi(\boldsymbol{\lambda})\big)$ is up to a multiplicative constant the density of a Gamma distribution parameterized by $\mathcal{G}(n/2,\phi(\boldsymbol{\lambda}))$. Hence, $\int_0^{+\infty}\tau^{(\frac{n}{2}-1)}\exp\big(-\tau\phi(\boldsymbol{\lambda})\big)\,d\tau = \Gamma\big(\frac{n}{2}\big)\phi(\boldsymbol{\lambda})^{-\frac{n}{2}}$, where $\Gamma(\cdot)$ is the Gamma function. Using this property, the integral is given by:

$$\begin{aligned} p(\boldsymbol{\lambda},\boldsymbol{\delta}|\mathcal{D}) &= \int_0^{+\infty} p(\boldsymbol{\eta}|\mathcal{D})\big|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}}\,d\tau \\ &\propto |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}|Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}}\left(\prod_{j=1}^q \delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\Big(-\delta_j\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)\Big)\right) \\ &\quad\times\left(\prod_{j=1}^q \lambda_j^{(\frac{\nu}{2}-1)}\right)\phi(\boldsymbol{\lambda})^{-\frac{n}{2}}. \end{aligned} \tag{8}$$

The above expression can be further simplified using the property that the determinant of a block diagonal matrix is equal to the product of the determinants of the blocks:

$$|Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} = \left(\zeta^{(p+1)}\;|I_{p+1}|\;|P|^q\prod_{j=1}^q\lambda_j^{(K-1)}\right)^{\frac{1}{2}} = \underbrace{\zeta^{\frac{(p+1)}{2}}|P|^{\frac{q}{2}}}_{constant}\prod_{j=1}^q\lambda_j^{\frac{(K-1)}{2}},$$

such that (8) becomes:

$$\begin{aligned} p(\boldsymbol{\lambda},\boldsymbol{\delta}|\mathcal{D}) \;&\propto\; |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}\left(\prod_{j=1}^q \lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right) \\ &\times\left(\prod_{j=1}^q \delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\Big(-\delta_j\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)\Big)\right)\;\phi(\boldsymbol{\lambda})^{-\frac{n}{2}}. \end{aligned} \tag{9}$$

The posterior in (9) can be integrated with respect to δ_j successively for $j=1,\ldots,q$ since $\delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\Big(-\delta_j\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)\Big)$ is (up to a multiplicative constant) a Gamma density parameterized by $\mathcal{G}(\frac{\nu}{2}+a_\delta, b_\delta+\frac{\nu}{2}\lambda_j)$, so:

$$\begin{aligned} &\int_0^{+\infty}\cdots\int_0^{+\infty}\left(\prod_{j=1}^q \delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\Big(-\delta_j\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)\Big)\right)d\delta_1\ldots d\delta_q \\ &= \prod_{j=1}^q\left(\int_0^{+\infty}\delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\Big(-\delta_j\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)\Big)\,d\delta_j\right) \\ &= \left(\Gamma\Big(\frac{\nu}{2}+a_\delta\Big)\right)^q\left(\prod_{j=1}^q\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)^{-(\frac{\nu}{2}+a_\delta)}\right) \end{aligned} \tag{10}$$

and the posterior of the penalty vector is:

$$\begin{aligned} p(\boldsymbol{\lambda}|\mathcal{D}) &= \int_0^{+\infty}\cdots\int_0^{+\infty}p(\boldsymbol{\lambda},\boldsymbol{\delta}|\mathcal{D})\;d\delta_1\ldots d\delta_q \\ &\propto\; |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}\left(\prod_{j=1}^q\lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right)\left(\prod_{j=1}^q\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)^{-(\frac{\nu}{2}+a_\delta)}\right)\phi(\boldsymbol{\lambda})^{-\frac{n}{2}}. \end{aligned} \tag{11}$$

One can easily compute the ratio:

$$
\begin{aligned}
p(\tau|\boldsymbol{\lambda},\mathcal{D}) &= \frac{p(\tau,\boldsymbol{\lambda}|\mathcal{D})}{p(\boldsymbol{\lambda}|\mathcal{D})} \\
&\propto \tau^{\left(\frac{n}{2}-1\right)}\exp(-\tau\phi(\boldsymbol{\lambda})),
\end{aligned}
$$

such that the conditional posterior distribution for τ is $(\tau|\boldsymbol{\lambda},\mathcal{D}) \sim \mathcal{G}\left(n/2, \phi(\boldsymbol{\lambda})\right)$.

3.3 Gradient and Hessian of the posterior penalty

The analytical gradient and Hessian of the penalty vector can be derived to find its posterior mode via a Newton-Raphson algorithm. The posterior mode as a measure of central tendency is essential to construct a grid for exploring $p(\boldsymbol{\lambda}|\mathcal{D})$. To ensure numerical stability, the penalty parameters are log transformed, $v_j = \log(\lambda_j), \quad j = 1,\ldots,q$, and the associated vector is $\mathbf{v} = (v_1,\ldots,v_q)^\top$. Using the multivariate transformation method on (11), the posterior becomes:

$$
\begin{aligned}
p(\mathbf{v}|\mathcal{D}) &\propto |B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|^{-\frac{1}{2}}\left(\prod_{j=1}^{q}\exp(v_j)^{\left(\frac{\nu+K-3}{2}\right)}\right) \\
&\quad\times\left(\prod_{j=1}^{q}\left(b_\delta + \frac{\nu}{2}\exp(v_j)\right)^{-\left(\frac{\nu}{2}+a_\delta\right)}\right)\phi(\mathbf{v})^{-\frac{n}{2}} \\
&\quad\times\left(\prod_{j=1}^{q}\exp(v_j)\right),
\end{aligned}
\tag{12}
$$

where $\prod_{j=1}^{q}\exp(v_j)$ is the Jacobian of the transformation, $\phi(\mathbf{v})$ is the following function of the log penalty vector $\phi(\mathbf{v}) = \frac{1}{2}\mathbf{y}^\top\left(I_n - B(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}B^\top\right)\mathbf{y}$ and $Q_{\boldsymbol{\xi}}^{\mathbf{v}}$ is a symmetric block diagonal matrix given by:

$$
Q_{\boldsymbol{\xi}}^{\mathbf{v}} = \begin{pmatrix} \zeta I_{p+1} & 0_{p+1,q\times(K-1)} \\ 0_{q\times(K-1),p+1} & \operatorname{diag}(\exp(v_1),\ldots,\exp(v_q))\otimes P \end{pmatrix}.
$$

Taking the log of (12) yields:

$$
\begin{aligned}
\log p(\mathbf{v}|\mathcal{D}) &\doteq -\frac{1}{2}\underbrace{\log|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}_{\text{Term I}} + \underbrace{\left(\frac{\nu+K-1}{2}\right)\sum_{j=1}^{q}v_j}_{\text{Term II}} - \frac{n}{2}\underbrace{\log\phi(\mathbf{v})}_{\text{Term III}} \\
&\quad -\left(\frac{\nu}{2}+a_\delta\right)\underbrace{\sum_{j=1}^{q}\log\left(b_\delta + \frac{\nu}{2}\exp(v_j)\right)}_{\text{Term IV}}.
\end{aligned}
\tag{13}
$$

Gradient

Using Jacobi's formula for the partial derivatives of the determinant with respect to v_j (see Harville, 1997, Chapter 15), in Term I:

$$
\begin{aligned}
\frac{\partial \log |B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}{\partial v_j} &= \frac{1}{|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|} \frac{\partial}{\partial v_j} |B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}| \\
&= \frac{1}{|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|} \mathrm{Tr}\Big(\mathrm{adj}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}) \frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\Big) \\
&= \frac{1}{|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|} \mathrm{Tr}\Big(|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}| \; (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} \\
&\quad \times \frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\Big) \\
&= \mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big), \qquad (14)
\end{aligned}
$$

where $\mathrm{adj}(\cdot)$ is the adjoint of a matrix (transpose of the cofactor matrix), $\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} := (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}$ is a symmetric matrix and P_{v_j} is a (symmetric) block diagonal matrix defined as:

$$
\begin{aligned}
P_{v_j} &:= \frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}) \\
&= \begin{pmatrix} 0_{p+1,p+1} & 0_{p+1,q\times(K-1)} \\ 0_{q\times(K-1),p+1} & \mathrm{diag}(0,\ldots,\exp(v_j),\ldots,0) \otimes P \end{pmatrix},
\end{aligned}
$$

where $\mathrm{diag}(0,\ldots,\exp(v_j),\ldots,0)$ is a $q \times q$ diagonal matrix, whose jth diagonal element is $\exp(v_j)$ and all other diagonal elements are zero. Derivation of Term II with respect to v_j is trivial:

$$
\frac{\partial}{\partial v_j}\left(\frac{\nu + K - 1}{2}\right) \sum_{j=1}^{q} v_j = \left(\frac{\nu + K - 1}{2}\right). \qquad (15)
$$

The partial derivative of Term III is:

$$
\begin{aligned}
\frac{\partial}{\partial v_j} \log(\phi(\mathbf{v})) &= \frac{1}{\phi(\mathbf{v})} \frac{\partial \phi(\mathbf{v})}{\partial v_j} \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\frac{\partial}{\partial v_j}\Big(\mathbf{y}^\top B (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} B^\top \mathbf{y}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\frac{\partial}{\partial v_j}\mathrm{Tr}\Big(\mathbf{y}^\top B (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} B^\top \mathbf{y}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\frac{\partial}{\partial v_j}\mathrm{Tr}\Big(B^\top \mathbf{y}\mathbf{y}^\top B (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathrm{Tr}\Big(B^\top \mathbf{y}\mathbf{y}^\top B \frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}\Big)\Big)
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathrm{Tr}\Big(B^\top \mathbf{y}\mathbf{y}^\top B\big(-(B^\top B+Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j} \\
&\quad \times(B^\top B+Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}\big)\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathrm{Tr}\Big(\mathbf{y}^\top B\big(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\big)B^\top\mathbf{y}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathbf{y}^\top B\big(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\big)B^\top\mathbf{y}\Big) \\
&= \frac{1}{2\phi(\mathbf{v})}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}. \qquad (16)
\end{aligned}
$$

Taking the derivative of Term IV with respect to v_j gives:

$$
\begin{aligned}
\frac{\partial}{\partial v_j}\sum_{j=1}^{q}\log\Big(b_\delta+\frac{\nu}{2}\exp(v_j)\Big) &= \frac{\frac{\nu}{2}\exp(v_j)}{b_\delta+\frac{\nu}{2}\exp(v_j)} \\
&= \frac{1}{1+\frac{2b_\delta}{\nu\exp(v_j)}}. \qquad (17)
\end{aligned}
$$

Finally, taking (14), (15), (16) and (17), the gradient $\nabla_{\mathbf{v}}\log p(\mathbf{v}|\mathcal{D})$ has entries:

$$
\begin{aligned}
\frac{\partial \log p(\mathbf{v}|\mathcal{D})}{\partial v_j} &= -\frac{1}{2}\mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)+\left(\frac{\nu+K-1}{2}\right) \\
&\quad -\frac{n}{4\phi(\mathbf{v})}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y} \\
&\quad -\frac{\big(\frac{\nu}{2}+a_\delta\big)}{1+\frac{2b_\delta}{\nu\exp(v_j)}}, \quad j=1,\dots,q.
\end{aligned}
$$

Hessian

To obtain the diagonal elements of the Hessian, the following differentiation is required:

$$
\begin{aligned}
\frac{\partial}{\partial v_j}\mathrm{Tr}\Big((B^\top B+Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j}\Big) &= \mathrm{Tr}\Big(\frac{\partial}{\partial v_j}(B^\top B+Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j}\Big) \\
&= \mathrm{Tr}\Big(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}+\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big) \\
&= -\mathrm{Tr}\Big(\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)^2-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big). \qquad (18)
\end{aligned}
$$

In addition, recall from (16) that:

$$
\frac{\partial\phi(\mathbf{v})}{\partial v_j}=\frac{1}{2}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}. \qquad (19)
$$

Furthermore, note the following differentiation result:

$$\begin{aligned}
&\frac{\partial}{\partial v_j}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\\
&=\frac{\partial}{\partial v_j}\mathrm{Tr}\Big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\Big)\\
&=\frac{\partial}{\partial v_j}\mathrm{Tr}\Big(B^\top\mathbf{y}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\Big)\\
&=\mathrm{Tr}\Big(B^\top\mathbf{y}\mathbf{y}^\top B\frac{\partial}{\partial v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\Big)\\
&=\mathrm{Tr}\left(B^\top\mathbf{y}\mathbf{y}^\top B\left(\frac{\partial\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}+\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial P_{v_j}}{\partial v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}+\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\frac{\partial\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}\right)\right)
\end{aligned}$$

$$\begin{aligned}
&= \mathrm{Tr}\left(B^\top\mathbf{y}\mathbf{y}^\top B\left(-2\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)^2\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}+\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right)\\
&= \mathrm{Tr}\left(\mathbf{y}^\top B\left(-2\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)^2\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}+\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)B^\top\mathbf{y}\right)\\
&= -2\mathbf{y}^\top B\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)^2\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}+\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}. \qquad (20)
\end{aligned}$$

Using (19), (20) and the quotient rule for derivatives yields:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\frac{\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}}{\phi(\mathbf{v})} &= \frac{1}{\phi^2(\mathbf{v})}\Big(-2\phi(\mathbf{v})\mathbf{y}^\top B\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)^2\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\\
&\quad+\phi(\mathbf{v})\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\\
&\quad-\frac{1}{2}\big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\big)^2\Big). \qquad (21)
\end{aligned}$$

Finally, note that:

$$\frac{\partial}{\partial v_j}\frac{\big(\frac{\nu}{2}+a_\delta\big)}{\Big(1+\frac{2b_\delta}{\nu\exp(v_j)}\Big)}=\frac{b_\delta\big(1+\frac{2a_\delta}{\nu}\big)\exp(-v_j)}{\Big(1+\frac{2b_\delta}{\nu\exp(v_j)}\Big)^2}, \qquad (22)$$

and using (18), (21), (22), the diagonal entries of the Hessian of $\log p(\mathbf{v}|\mathcal{D})$ are:

$$\begin{aligned}
&\frac{\partial^2\log p(\mathbf{v}|\mathcal{D})}{\partial v_j^2}\\
&=\frac{1}{2}\mathrm{Tr}\Big(\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)^2-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)-\frac{n}{4\phi^2(\mathbf{v})}\left(-2\phi(\mathbf{v})\mathbf{y}^\top B\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)^2\right.\\
&\left.\times\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}+\phi(\mathbf{v})\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}-\frac{1}{2}\big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\big)^2\right)\\
&-\frac{b_\delta\big(1+\frac{2a_\delta}{\nu}\big)\exp(-v_j)}{\Big(1+\frac{2b_\delta}{\nu\exp(v_j)}\Big)^2},\quad j=1,\ldots,q.
\end{aligned}$$

To obtain the off-diagonal elements of the Hessian, note that for index $s\neq j$:

$$\begin{aligned}\frac{\partial}{\partial v_s}\mathrm{Tr}\Big((B^\top B+Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j}\Big) &= \mathrm{Tr}\Big(\frac{\partial}{\partial v_s}(B^\top B+Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j}\Big)\\ &= \mathrm{Tr}\Big(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)\\ &= -\mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big).\end{aligned}$$

Furthermore, similarly to (20):

$$\begin{aligned}&\frac{\partial}{\partial v_s}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\\ &=\frac{\partial}{\partial v_s}\mathrm{Tr}\Big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\Big)\\ &=\frac{\partial}{\partial v_s}\mathrm{Tr}\Big(B^\top\mathbf{y}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\Big)\\ &=\mathrm{Tr}\Big(B^\top\mathbf{y}\mathbf{y}^\top B\frac{\partial}{\partial v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\Big)\\ &=\mathrm{Tr}\left(B^\top\mathbf{y}\mathbf{y}^\top B\left(\frac{\partial\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}+\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial P_{v_j}}{\partial v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}+\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\frac{\partial\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}\right)\right)\\ &=\mathrm{Tr}\left(B^\top\mathbf{y}\mathbf{y}^\top B\left(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right)\\ &=\mathrm{Tr}\left(\mathbf{y}^\top B\left(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)B^\top\mathbf{y}\right)\\ &=-\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}-(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y})^\top\\ &=-2\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y},\end{aligned}$$

such that using the quotient rule, we have:

$$\begin{aligned}&\frac{\partial}{\partial v_s}\frac{\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}}{\phi(\mathbf{v})}\\ &=\frac{1}{\phi^2(\mathbf{v})}\Big(-2\phi(\mathbf{v})\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\\ &-\frac{1}{2}\big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\big)\big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\big)\Big).\end{aligned}$$

Hence, the off-diagonal elements $s=1,\ldots,q,\quad j=1,\ldots,q$ and $s\neq j$ of the Hessian are:

$$\begin{aligned}\frac{\partial^2\log p(\mathbf{v}|\mathcal{D})}{\partial v_s\,\partial v_j} &= \frac{1}{2}\mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\Big)\\ &+\frac{n}{4\phi^2(\mathbf{v})}\Big(2\phi(\mathbf{v})\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\\ &+\frac{1}{2}\big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\big)\big(\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_s}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top\mathbf{y}\big)\Big).\end{aligned}$$

To summarize, the gradient and Hessian entries of $\log p(\mathbf{v}|\mathcal{D})$ are:

Gradient $\nabla_{\mathbf{v}} \log p(\mathbf{v}|\mathcal{D})$ entries for $j = 1, \dots, q$:

$$\begin{aligned}
&\frac{\partial \log p(\mathbf{v}|\mathcal{D})}{\partial v_j} \\
&= -\frac{1}{2}\mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big) + \left(\frac{\nu + K - 1}{2}\right) - \frac{n}{4\phi(\mathbf{v})}\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} \\
&- \frac{\left(\frac{\nu}{2} + a_\delta\right)}{1 + \frac{2b_\delta}{\nu \exp(v_j)}}.
\end{aligned} \tag{23}$$

Hessian $\nabla^2_{\mathbf{v}} \log p(\mathbf{v}|\mathcal{D})$, diagonal elements $j = 1, \dots, q$:

$$\begin{aligned}
&\frac{\partial^2 \log p(\mathbf{v}|\mathcal{D})}{\partial v_j^2} \\
&= \frac{1}{2}\mathrm{Tr}\Big(\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big)^2 - \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big) - \frac{n}{4\phi^2(\mathbf{v})}\Bigg(- 2\phi(\mathbf{v})\mathbf{y}^\top B \Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big)^2 \\
&\times \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} + \phi(\mathbf{v})\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} - \frac{1}{2}\big(\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y}\big)^2\Bigg) \\
&- \frac{b_\delta\big(1 + \frac{2a_\delta}{\nu}\big)\exp(-v_j)}{\Big(1 + \frac{2b_\delta}{\nu \exp(v_j)}\Big)^2}.
\end{aligned}$$

Hessian $\nabla^2_{\mathbf{v}} \log p(\mathbf{v}|\mathcal{D})$, off-diagonal elements $s = 1, \dots, q,\ j = 1, \dots, q,\ j \neq s$:

$$\begin{aligned}
\frac{\partial^2 \log p(\mathbf{v}|\mathcal{D})}{\partial v_s\, \partial v_j} &= \frac{1}{2}\mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_s} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big) \\
&+ \frac{n}{4\phi^2(\mathbf{v})}\Big(2\phi(\mathbf{v})\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_s} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} \\
&+ \frac{1}{2}\big(\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y}\big)\big(\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_s} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y}\big)\Big).
\end{aligned}$$

The **R** output below compares (for $q = 3$) the analytical gradient and Hessian formulas with the numerical derivatives of $\log p(\mathbf{v}|\mathcal{D})$ obtained with the `grad()` and `hessian()` functions of the `numDeriv` package at a randomly selected point $\mathbf{v}$ with entries $v_j \sim \mathcal{U}(-5, 5),\ j = 1, 2, 3$.

```
---------------Gradient-------------
 "----------analytic----------"
 -3.747028 -25.223528  -9.407790
 "----------numeric-----------"
 -3.747036 -25.223532  -9.407792
```

```
--------------Hessian-------------
 "----------analytic----------"
         [,1]      [,2]      [,3]
[1,] -1.774439  0.849825  0.401218
[2,]  0.849825 -3.846784  1.759438
[3,]  0.401218  1.759438 -3.381276

 "----------numeric----------"
         [,1]      [,2]      [,3]
[1,] -1.774438  0.849825  0.401218
[2,]  0.849825 -3.846783  1.759438
[3,]  0.401218  1.759438 -3.381276
```

In Table 1, we show the largest difference (in absolute value) between the entries of the numerical and analytical gradients and Hessians respectively computed across 1000 randomly selected points $\mathbf{v}$ with entries $v_j \sim \mathcal{U}(-5, 5),\ j = 1, 2, 3$.

	v_1	v_2	v_3
Gradient entries	0.000298	0.000141	0.001738
Hessian diagonal entries	0.010067	0.004479	0.034679
Hessian off-diagonal entries	0.000042	0.000207	0.000127

Table 1: Largest absolute difference between gradient and Hessian entries computed from our analytical formulas and the numerical derivatives from the **numDeriv** package.

3.4 Exploration of the posterior penalty space

A crucial step to derive the approximate posterior of latent variables is to identify the behavior of $p(\mathbf{v}|\mathcal{D})$. This is similar to a design problem in the sense that a set of points has to be efficiently chosen in the domain of a response surface to capture the essence of the functional pattern. A grid strategy is proposed that is sensible to asymmetries in the response surface $p(\mathbf{v}|\mathcal{D})$, with the skew-normal family of distributions forming the backbone that manages the lack of symmetry. The grid will be constructed around the posterior mode $\hat{\mathbf{v}}$ of the target $\log p(\mathbf{v}|\mathcal{D})$ which can be obtained through a Newton-Raphson method summarized in Algorithm 1, which contains the previously derived gradient $\nabla_{\mathbf{v}} \log p(\mathbf{v}|\mathcal{D})$ and Hessian $\nabla^2_{\mathbf{v}} \log p(\mathbf{v}|\mathcal{D})$.

An elementary approach to explore $p(\mathbf{v}|\mathcal{D})$ could rely on a multivariate Gaussian approximation to the posterior of the log penalty parameters $\mathbf{v}$, namely $\widetilde{p}_G(\mathbf{v}|\mathcal{D}) = \mathcal{N}_{\dim(\mathbf{v})}\big(\hat{\mathbf{v}}, \big(-\mathcal{H}^*\big)^{-1}\big)$, where the covariance matrix is obtained from the Hessian $\mathcal{H}^* = \nabla^2_{\mathbf{v}} \log p(\hat{\mathbf{v}}|\mathcal{D})$ evaluated at the mode $\hat{\mathbf{v}}$. However, as already pointed in Martins et al. (2013), the presence of potential asymmetries would not be captured by a Gaussian approximation. Instead, to efficiently explore the posterior penalty space, a grid strategy is proposed, which implicitly takes into account asymmetries by using skew-normal distributions to approximate the conditional posterior of each penalty parameter through a moment-matching approach.

Algorithm 1: Newton-Raphson to locate the mode of $p(\mathbf{v}|\mathcal{D})$

1: Set tol=10^{-5}, dist=3, $\mathbf{v}^{(0)} = \left(v_1^{(0)}, \ldots, v_q^{(0)}\right)$ and m=0.
2: **while** dist $>$ tol **do**
3: $\quad \mathbf{v}^{(m+1)} = \mathbf{v}^{(m)} - \left(\nabla^2_{\mathbf{v}} \log p(\mathbf{v}^{(m)}|\mathcal{D})\right)^{-1} \nabla_{\mathbf{v}} \log p(\mathbf{v}^{(m)}|\mathcal{D})$.
4: $\quad$ dist=$\|\mathbf{v}^{(m+1)} - \mathbf{v}^{(m)}\|$.
5: **end while**
6: At convergence return $\hat{\mathbf{v}} = (\hat{v}_1, \ldots, \hat{v}_q)$.

The skew-normal family was first introduced by Azzalini (1985), see Azzalini (2014) for more details. In the univariate case, a random variable X has a skew-normal distribution denoted by $X \sim \mathrm{SN}(\mu, \varsigma^2, \rho)$ if its probability density function at $x \in \mathbb{R}$ is:

$$p(x) = \frac{2}{\varsigma}\varphi\left(\frac{x-\mu}{\varsigma}\right)\ \Phi\left(\rho\frac{(x-\mu)}{\varsigma}\right), \tag{24}$$

where $\mu \in \mathbb{R}$ is a location parameter, $\varsigma \in \mathbb{R}_{++}$ a scale parameter and $\rho \in \mathbb{R}$ a shape parameter regulating skewness. Also, $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the standard Gaussian density function and its cumulative distribution function respectively, such that setting $\rho = 0$ yields the $\mathcal{N}(\mu, \varsigma^2)$ distribution. We suggest to approximate the conditional posterior distribution of $(v_j|\hat{\mathbf{v}}_{-j}, \mathcal{D})$ $(j = 1, ..., q)$ with a skew-normal distribution by matching its first three empirical moments with the theoretical ones for the density in (24), where $\hat{\mathbf{v}}_{-j}$ denotes the vector $\hat{\mathbf{v}}$ without the jth entry. The derivations to obtain μ^*, ς^* and ρ^* in the approximating skew-normal distribution $\mathrm{SN}_j(\mu^*, \varsigma^{*2}, \rho^*)$ to $p(v_j|\hat{\mathbf{v}}_{-j}, \mathcal{D})$ through moment matching are shown below.

The first moment and the second and third central moments of $X \sim \mathrm{SN}(\mu, \varsigma^2, \rho)$ are given by:

$$\begin{aligned}
E(X) &= \mu + \varsigma \sqrt{\frac{2}{\pi}}\ \psi, \\
E\big((X - E(X))^2\big) &= \varsigma^2 \left(1 - \frac{2}{\pi}\ \psi^2\right), \\
E\big((X - E(X))^3\big) &= \frac{1}{2}(4-\pi)\ \varsigma^3\ \left(\frac{2}{\pi}\right)^{\frac{3}{2}}\ \psi^3,
\end{aligned}$$

where $\psi = \rho/\sqrt{1+\rho^2} \in (-1, 1)$. These theoretical moments will be matched with the empirical moments of the the conditional distributions $p(v_j|\hat{\mathbf{v}}_{-j}, \mathcal{D})$. The empirical moments of the conditionals are computed on an equidistant grid $\{v_{jl}\}_{l=1}^L$ with interval length Δ_l:

$$\begin{aligned}
m_{j1} &= \sum_{l=1}^{L} v_{jl}\ p(v_{jl}|\hat{\mathbf{v}}_{-j}, \mathcal{D})\ \Delta_l, \\
m_{j2} &= \sum_{l=1}^{L} (v_{jl} - m_{j1})^2\ p(v_{jl}|\hat{\mathbf{v}}_{-j}, \mathcal{D})\ \Delta_l, \\
m_{j3} &= \sum_{l=1}^{L} (v_{jl} - m_{j1})^3\ p(v_{jl}|\hat{\mathbf{v}}_{-j}, \mathcal{D})\ \Delta_l.
\end{aligned}$$

The skew-normal fit to $p(v_j|\hat{\mathbf{v}}_{-j},\mathcal{D})$ is found by matching the empirical and theoretical moments, i.e. the following system needs to be solved:

$$\begin{aligned} m_{j1} &= \mu+\varsigma\sqrt{\frac{2}{\pi}}\;\psi && (25)\\ m_{j2} &= \varsigma^2\left(1-\frac{2}{\pi}\;\psi^2\right) && (26)\\ m_{j3} &= \frac{1}{2}(4-\pi)\;\varsigma^3\;\left(\frac{2}{\pi}\right)^{\frac{3}{2}}\;\psi^3. && (27)\end{aligned}$$

From (26), we isolate ς:

$$\varsigma=\sqrt{\frac{m_{j2}}{\left(1-\frac{2}{\pi}\;\psi^2\right)}}>0. \tag{28}$$

Plugging (28) in (27) yields:

$$\begin{aligned} m_{j3} &= \frac{1}{2}(4-\pi)\frac{m_{j2}^{\frac{3}{2}}}{\left(1-\frac{2}{\pi}\;\psi^2\right)^{\frac{3}{2}}}\left(\frac{2}{\pi}\right)^{\frac{3}{2}}\psi^3\\ &\Leftrightarrow \frac{\psi^3}{\left(1-\frac{2}{\pi}\;\psi^2\right)^{\frac{3}{2}}}=\frac{2m_{j3}\pi^{\frac{3}{2}}}{(4-\pi)m_{j2}^{\frac{3}{2}}2^{\frac{3}{2}}}\\ &\Leftrightarrow \frac{\psi^3}{\left(1-\frac{2}{\pi}\;\psi^2\right)^{\frac{3}{2}}}=\frac{m_{j3}\pi^{\frac{3}{2}}}{(4-\pi)\sqrt{2}\;m_{j2}^{\frac{3}{2}}}\\ &\Leftrightarrow \frac{\psi}{\left(1-\frac{2}{\pi}\;\psi^2\right)^{\frac{1}{2}}}=\frac{m_{j3}^{\frac{1}{3}}\pi^{\frac{1}{2}}}{(4-\pi)^{\frac{1}{3}}2^{\frac{1}{6}}\;m_{j2}^{\frac{1}{2}}}.\end{aligned}$$

Let $\kappa:=m_{j3}^{\frac{1}{3}}\pi^{\frac{1}{2}}/(4-\pi)^{\frac{1}{3}}2^{\frac{1}{6}}\;m_{j2}^{\frac{1}{2}}$, so that the above equation becomes:

$$\begin{aligned} &\psi=\kappa\;\left(1-\frac{2}{\pi}\psi^2\right)^{\frac{1}{2}}\\ &\Leftrightarrow \psi^2+\frac{2\kappa^2}{\pi}\;\psi^2-\kappa^2=0\\ &\Leftrightarrow \psi^2\left(1+\frac{2\kappa^2}{\pi}\right)-\kappa^2=0.\end{aligned}$$

The discriminant of the above quadratic equation in ψ is given by $\Delta=4\left(1+\frac{2\kappa^2}{\pi}\right)\kappa^2>0$. Even though there are two solutions, the only solution retained is the one whose sign is the same as the sign of the third empirical central moment. Indeed, if m_{j3} is negative/positive, ψ^* (and by extension ρ^*) should also be negative/positive to capture the negatively/positively skewed pattern of $p(v_j|\hat{\mathbf{v}}_{-j},\mathcal{D})$. Hence, using the sign$(\cdot)$ function:

$$\psi^*=\operatorname{sign}(m_{j3})\frac{\sqrt{4\left(\kappa^2+\frac{2\kappa^4}{\pi}\right)}}{2+\frac{4\kappa^2}{\pi}}. \tag{29}$$

So, we have $\rho^*=\psi^*/\sqrt{1-(\psi^*)^2}$ and plugging (29) in (28), we recover:

$$\varsigma^* = \sqrt{\frac{m_{j2}}{\left(1 - \frac{2}{\pi}\ (\psi^*)^2\right)}}. \tag{30}$$

Finally, the location parameter is given by:

$$\mu^* = m_{j1} - \varsigma^* \sqrt{\frac{2}{\pi}}\ \psi^*. \tag{31}$$

The skew-normal fit to the conditional $p(v_j|\hat{\mathbf{v}}_{-j}, \mathcal{D})$ is written as follows $\text{SN}_j(\mu^*, \varsigma^{*2}, \rho^*)$ and can be used for the grid construction strategy.

Once a skew-normal distribution has been adjusted to the conditional $p(v_j|\hat{\mathbf{v}}_{-j}, \mathcal{D})$, we construct an equidistant grid $\{v_{jm}\}_{m=1}^{M}$ of size M from the 2.5th to the 97.5th quantiles of the skew-normal fit denoted by $\text{SN}_{j,0.025}$ and $\text{SN}_{j,0.975}$ respectively. This process is repeated across all dimensions $j = 1, \ldots, q$ and a Cartesian product of the univariate grids is taken, ending up with a total of M^q (multivariate) grid points. Next, a filtering strategy is implemented to get rid of quadrature points associated to a small posterior mass.
Let us consider the normalized posterior $R(\mathbf{v}) = p(\mathbf{v}|\mathcal{D})/p(\hat{\mathbf{v}}|\mathcal{D})$ and use the property that $-2\log R(\mathbf{v})$ is approximately distributed as a chi-square distribution with $\dim(\mathbf{v})$ degrees of freedom denoted by $\chi^2_{\dim(\mathbf{v})}$. Then, an approximate $(1-\alpha)$ credible region for $\mathbf{v}$ is defined by the set of values in $\mathbb{R}^{\dim(\mathbf{v})}$ such that $R(\mathbf{v}) \geq \exp\left(-.5\chi^2_{\dim(\mathbf{v});1-\alpha}\right)$. As an illustration, take $\alpha = 0.05$ and $\dim(\mathbf{v}) = 2$. If we decide to concentrate on quadrature points in the 95% credible region for $\mathbf{v}$, then the preceding result would suggest to discard values $\mathbf{v}$ in the bivariate grid for which $R(\mathbf{v}) < \exp(-.5\chi^2_{2;0.95}) = .05$, leaving $\widetilde{M}$ grid points after filtering. Figure 1 highlights the difference between the skew-normal match and the naive Gaussian fit to the targets $p(v_j|\hat{\mathbf{v}}_{-j}, \mathcal{D}),\ j = 1, 2$ with $q = 2$ nonlinear smooth functions in the additive predictor and sample size $n = 300$. Figure 2 shows the surface plot of $R(\mathbf{v})$.

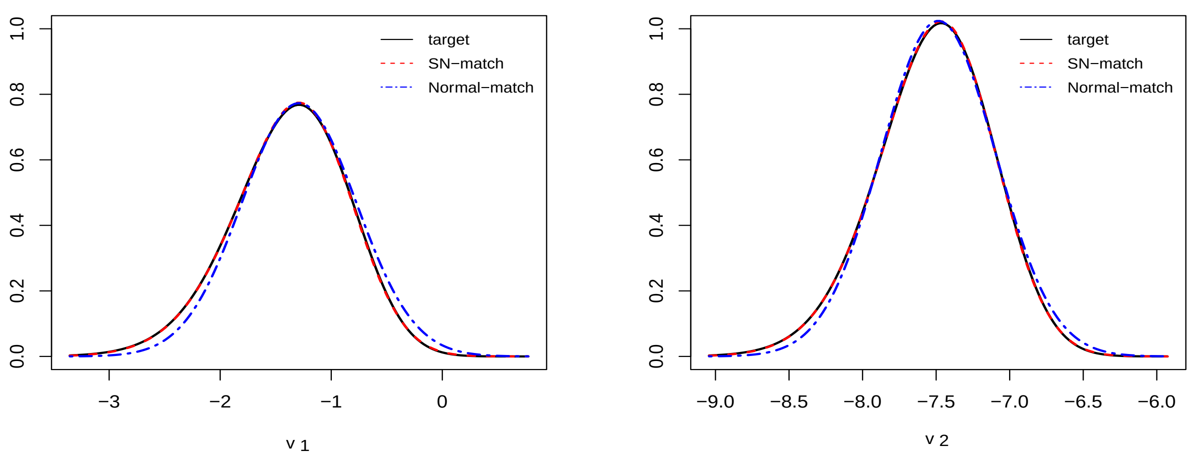

Figure 1: Skew-normal fit (dashed) and naive Gaussian match (dash-dotted) to the normalized conditional $p(v_1|\hat{v}_2, \mathcal{D})$ (left) and $p(v_2|\hat{v}_1, \mathcal{D})$ (right). The skew-normal fit is closer to the target and captures the lack of symmetry.

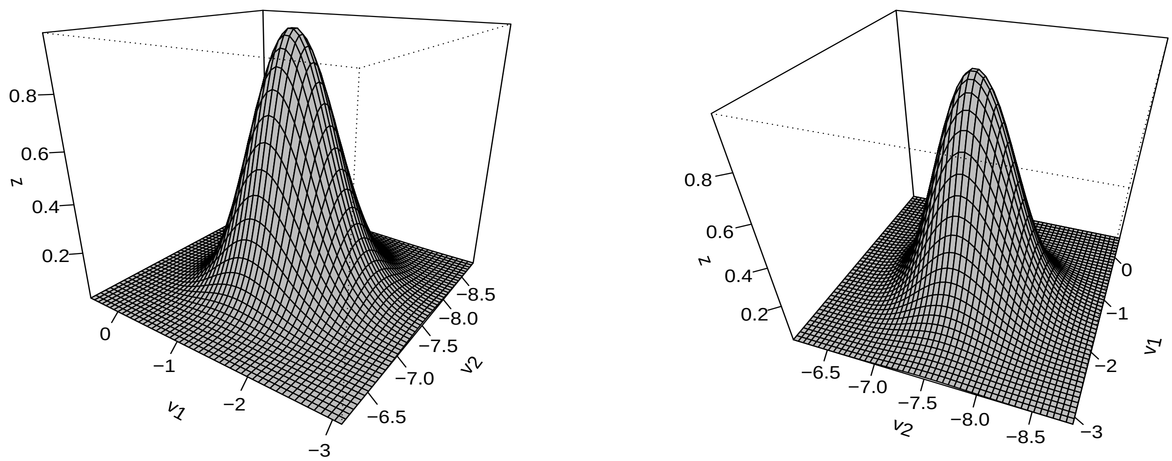

Figure 2: Surface plot of $R(\mathbf{v})$ when $q = 2$.

Finally, Figure 3 summarizes the strategy behind the grid construction. In (a), an equidistant univariate grid is constructed in each dimension resulting in a cross-shaped pattern with center $\hat{\mathbf{v}}$. The Cartesian product of these univariate grids is computed and shown in (b). Following our filtering rule, we only keep a subset of the Cartesian product grid as shown by the blue points in (c). Figure 3 (d) shows the final grid which will be used for further inference in the additive model.

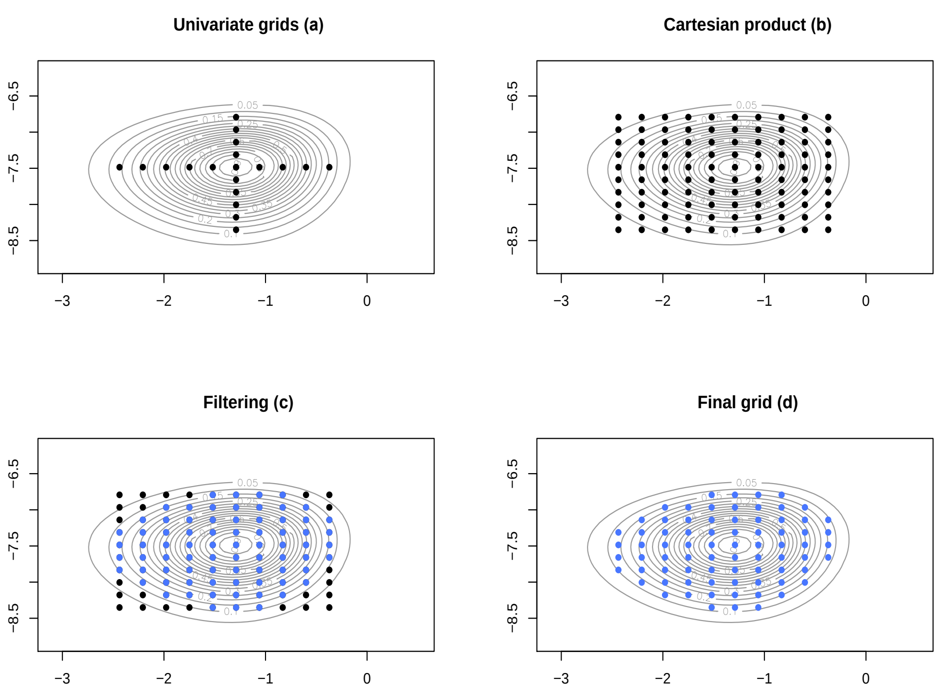

Figure 3: Grid strategy to explore $\log p(\mathbf{v}|\mathcal{D})$. (a) Equidistant univariate grid in each dimension. (b) Cartesian product. (c) Filtering out the points. (d) Final grid used for further inference in the additive model.

4 Approximate posterior of the latent field

The quadrature points derived in the previous section will serve to approximate the posterior of the latent vector $\boldsymbol{\xi}$ and to construct pointwise estimators and credible intervals of latent field elements. The posterior of the latent vector can be written as:

$$\begin{aligned}
p(\boldsymbol{\xi}|\mathcal{D}) &= \int_{\mathbb{R}_{++}} \cdots \int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\delta}, \tau|\mathcal{D}) \; d\lambda_1 \dots d\lambda_q \; d\delta_1 \dots d\delta_q \; d\tau \\
&= \int_{\mathbb{R}^q_{++}} \int_{\mathbb{R}^q_{++}} \int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau, \mathcal{D}) \; p(\tau|\boldsymbol{\lambda}, \mathcal{D}) \; p(\boldsymbol{\delta}, \boldsymbol{\lambda}|\mathcal{D}) \; d\boldsymbol{\lambda} \; d\boldsymbol{\delta} \; d\tau \\
&= \int_{\mathbb{R}^q_{++}} \left(\int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau, \mathcal{D}) \; p(\tau|\boldsymbol{\lambda}, \mathcal{D}) \; d\tau \right) \left(\int_{\mathbb{R}^q_{++}} p(\boldsymbol{\delta}, \boldsymbol{\lambda}|\mathcal{D}) \; d\boldsymbol{\delta} \right) \; d\boldsymbol{\lambda}
\end{aligned}$$

$$= \int_{\mathbb{R}^q_{++}} \left(\int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau, \mathcal{D}) \; p(\tau|\boldsymbol{\lambda}, \mathcal{D}) \; d\tau \right) \; p(\boldsymbol{\lambda}|\mathcal{D}) \; d\boldsymbol{\lambda} \tag{32}$$

The integral with respect to τ results in a function of $\boldsymbol{\xi}$ that corresponds to a multivariate Student distribution with n degrees of freedom. Indeed, let us reparameterize the conditional posterior of the precision as $(\tau|\boldsymbol{\lambda}, \mathcal{D}) \sim \mathcal{G}\left(n/2, (ns_{\boldsymbol{\lambda}})/(2n)\right)$, with the following scalar quantity $s_{\boldsymbol{\lambda}} = \mathbf{y}^\top \left(I_n - B(B^\top B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}})^{-1} B^\top \right) \mathbf{y}$, so that the integrand can be written as the product of the two distributions:

$$\begin{aligned}
p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau, \mathcal{D}) &= (2\pi)^{-\frac{\dim(\boldsymbol{\xi})}{2}} \tau^{\frac{\dim(\boldsymbol{\xi})}{2}} |B^\top B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}}|^{\frac{1}{2}} \\
&\quad \times \exp\left(-\frac{\tau}{2} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})^\top (B^\top B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}}) (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}) \right) \\
p(\tau|\boldsymbol{\lambda}, \mathcal{D}) &= \frac{\left(\frac{s_{\boldsymbol{\lambda}}}{n}\right)^{\frac{n}{2}} \left(\frac{n}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} \tau^{\left(\frac{n}{2}-1\right)} \exp\left(-\tau \frac{s_{\boldsymbol{\lambda}}}{n} \frac{n}{2} \right),
\end{aligned}$$

The integrand is thus given by:

$$\begin{aligned}
&p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau, \mathcal{D}) \; p(\tau|\boldsymbol{\lambda}, \mathcal{D}) \\
&= \frac{|B^\top B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}}|^{\frac{1}{2}} \left(\frac{s_{\boldsymbol{\lambda}}}{n}\right)^{\frac{n}{2}} \left(\frac{n}{2}\right)^{\frac{n}{2}}}{(2\pi)^{\frac{\dim(\boldsymbol{\xi})}{2}} \Gamma\left(\frac{n}{2}\right)} \tau^{\left(\frac{n+\dim(\boldsymbol{\xi})}{2}-1\right)} \exp\left(-\tau \left(\frac{1}{2} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})^\top \right.\right. \\
&\left.\left. \times (B^\top B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}}) (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}) + \frac{s_{\boldsymbol{\lambda}}}{n} \frac{n}{2} \right) \right).
\end{aligned}$$

Let $u := \left(\frac{1}{2} (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})^\top (B^\top B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}}) (\boldsymbol{\xi} - \hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}) + (s_{\boldsymbol{\lambda}}/n)(n/2) \right)$ and consider the integral:

$$\int_{\mathbb{R}_{++}} \tau^{\left(\frac{n+\dim(\boldsymbol{\xi})}{2}-1\right)} \exp(-\tau u) \; d\tau = \Gamma\left(\frac{n+\dim(\boldsymbol{\xi})}{2} \right) u^{-\frac{(n+\dim(\boldsymbol{\xi}))}{2}}.$$

Using the above result the integral is:

$$\begin{aligned}
&\int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D})\ p(\tau|\boldsymbol{\lambda},\mathcal{D})\ d\tau \\
&= \frac{\Gamma\left(\frac{n+\dim(\boldsymbol{\xi})}{2}\right)\ |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}}\left(\frac{s_{\boldsymbol{\lambda}}}{n}\right)^{\frac{n}{2}}\left(\frac{n}{2}\right)^{\frac{n}{2}}}{(2\pi)^{\frac{\dim(\boldsymbol{\xi})}{2}}\Gamma\left(\frac{n}{2}\right)} \\
&\times\left(\frac{1}{2}(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})^\top(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}) + \frac{s_{\boldsymbol{\lambda}}}{n}\frac{n}{2}\right)^{-\frac{(n+\dim(\boldsymbol{\xi}))}{2}} \\
&= \frac{\Gamma\left(\frac{n+\dim(\boldsymbol{\xi})}{2}\right)\ |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}}\left(\frac{s_{\boldsymbol{\lambda}}}{n}\right)^{\frac{n}{2}}\left(\frac{n}{2}\right)^{\frac{n}{2}}}{(2\pi)^{\frac{\dim(\boldsymbol{\xi})}{2}}\Gamma\left(\frac{n}{2}\right)} \\
&\times\left(\frac{s_{\boldsymbol{\lambda}}}{n}\frac{n}{2}\left(1+\frac{1}{n}(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})^\top\left(n\ s_{\boldsymbol{\lambda}}^{-1}(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})\right)(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})\right)\right)^{-\frac{(n+\dim(\boldsymbol{\xi}))}{2}} \\
&= \frac{\Gamma\left(\frac{n+\dim(\boldsymbol{\xi})}{2}\right)\ \left(\frac{n}{2}\right)^{-\frac{\dim(\boldsymbol{\xi})}{2}}|B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}}\left(\frac{s_{\boldsymbol{\lambda}}}{n}\right)^{-\frac{\dim(\boldsymbol{\xi})}{2}}}{(2\pi)^{\frac{\dim(\boldsymbol{\xi})}{2}}\Gamma\left(\frac{n}{2}\right)} \\
&\times\left(1+\frac{1}{n}(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})^\top\left(n\ s_{\boldsymbol{\lambda}}^{-1}(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})\right)(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})\right)^{-\frac{(n+\dim(\boldsymbol{\xi}))}{2}}.
\end{aligned}$$

Note that:

$$\left|B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\right|^{\frac{1}{2}}\left(\frac{s_{\boldsymbol{\lambda}}}{n}\right)^{-\frac{\dim(\boldsymbol{\xi})}{2}} = \left|\left(\frac{s_{\boldsymbol{\lambda}}}{n}\right)(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\right|^{-\frac{1}{2}},$$

so that the integral is finally given by:

$$\begin{aligned}
&\int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D})\ p(\tau|\boldsymbol{\lambda},\mathcal{D})\ d\tau \\
&= \frac{\Gamma\left(\frac{n+\dim(\boldsymbol{\xi})}{2}\right)}{\Gamma\left(\frac{n}{2}\right)n^{\frac{\dim(\boldsymbol{\xi})}{2}}\pi^{\frac{\dim(\boldsymbol{\xi})}{2}}\left|\frac{s_{\boldsymbol{\lambda}}}{n}(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\right|^{\frac{1}{2}}} \\
&\times\left(1+\frac{1}{n}(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})^\top\left(\frac{s_{\boldsymbol{\lambda}}}{n}(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\right)^{-1}(\boldsymbol{\xi}-\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}})\right)^{-\frac{(n+\dim(\boldsymbol{\xi}))}{2}}.
\end{aligned}$$

The above formula is a multivariate Student distribution for $\boldsymbol{\xi}$ (see Jackman, 2009, p.508) with n degrees of freedom denoted by $t_n\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}, \tilde{S}_{\boldsymbol{\lambda}}\right)$ with location parameter $\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} = (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}B^\top\mathbf{y}$ and symmetric, positive-definite matrix $\tilde{S}_{\boldsymbol{\lambda}} = \frac{s_{\boldsymbol{\lambda}}}{n}(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}$. Using the above integral result, the posterior of the latent field in (32) simplifies to:

$$p(\boldsymbol{\xi}|\mathcal{D}) = \int_{\mathbb{R}_{++}^q} t_n\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}, \tilde{S}_{\boldsymbol{\lambda}}\right) p(\boldsymbol{\lambda}|\mathcal{D})\ d\boldsymbol{\lambda}. \tag{33}$$

Using the log-transformation on the penalty parameters, (33) becomes:

$$p(\boldsymbol{\xi}|\mathcal{D}) = \int_{\mathbb{R}^q} t_n\left(\hat{\boldsymbol{\xi}}_{\mathbf{v}}, \tilde{S}_{\mathbf{v}}\right)\ p(\mathbf{v}|\mathcal{D})\ d\mathbf{v}, \tag{34}$$

where $\hat{\boldsymbol{\xi}}_{\mathbf{v}} = (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} B^\top \mathbf{y}$ and $\tilde{S}_{\mathbf{v}} = (s_{\mathbf{v}}/n)\big(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\big)^{-1}$ with the scalar $s_{\mathbf{v}} = \mathbf{y}^\top \left(I_n - B(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} B^\top\right) \mathbf{y}$. Let Δ_{v_j} be the width of the jth univariate grid and denote by $\Delta \mathbf{v} = \Delta_{v_1} \times \cdots \times \Delta_{v_q}$ the discretized version of $d\mathbf{v}$. Using the quadrature points from the grid strategy $\{\mathbf{v}^{(m)}\}_{m=1}^{\widetilde{M}}$, integral (34) can be approximated as follows:

$$\tilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} t_n \left(\hat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \tilde{S}_{\mathbf{v}^{(m)}}\right) \; p(\mathbf{v}^{(m)}|\mathcal{D}) \; \Delta \mathbf{v}. \tag{35}$$

Furthermore, define the weights:

$$\omega_m = \frac{p(\mathbf{v}^{(m)}|\mathcal{D}) \; \Delta \mathbf{v}}{\sum_{m=1}^{\widetilde{M}} p(\mathbf{v}^{(m)}|\mathcal{D}) \; \Delta \mathbf{v}}, \quad m = 1, \ldots, \widetilde{M}. \tag{36}$$

Dividing (35) by the denominator of ω_m, one obtains a mixture of multivariate Student distributions for the approximate posterior of the latent field:

$$\tilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \; t_n \left(\hat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \tilde{S}_{\mathbf{v}^{(m)}}\right). \tag{37}$$

Note that $\omega_m \geq 0$ and $\sum_{m=1}^{\widetilde{M}} \omega_m = 1$, such that (37) is a probability density function. Furthermore, $t_n\big(\hat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \tilde{S}_{\mathbf{v}^{(m)}}\big)$ converges in law to $\mathcal{N}_{\dim(\boldsymbol{\xi})}\big(\hat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \tilde{S}_{\mathbf{v}^{(m)}}\big)$ as $n \to +\infty$ (see Kroese et al., 2013, p.147), so for n sufficiently large, we can write (37) as a finite mixture of multivariate Gaussian densities:

$$\tilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \; \mathcal{N}_{\dim(\boldsymbol{\xi})} \left(\hat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \tilde{S}_{\mathbf{v}^{(m)}}\right). \tag{38}$$

A point estimate for the latent vector is given by the posterior mean of (38) which is simply the mixture of the location components (see Frühwirth-Schnatter, 2006):

$$\hat{\boldsymbol{\xi}} = \sum_{m=1}^{\widetilde{M}} \omega_m \; \hat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}. \tag{39}$$

From $(\tau|\mathbf{v}, \mathcal{D}) \sim \mathcal{G}\left(n/2, \phi(\mathbf{v})\right)$, a point estimate of the precision can be obtained by computing the posterior mean of the Gamma at the posterior mode $\hat{\mathbf{v}}$ of $\log p(\mathbf{v}|\mathcal{D})$, i.e. $\hat{\tau} = 0.5 \; n \; (\phi(\hat{\mathbf{v}}))^{-1}$. Hence, a point estimate of the standard deviation of the error is $\hat{\sigma} = \hat{\tau}^{-0.5}$.

4.1 Credible intervals

Approximate quantile-based credible intervals for latent field elements $\xi_h, \; h = 1, \ldots, \dim(\boldsymbol{\xi})$ can be straightforwardly constructed. Starting from the joint marginal posterior in (38), we can write the univariate marginal posterior for element ξ_h as:

$$\tilde{p}(\xi_h|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \; \mathcal{N}_1\Big(\hat{\xi}_{h,\mathbf{v}^{(m)}}, \tilde{S}_{hh,\mathbf{v}^{(m)}}\Big), \tag{40}$$

where $\hat{\xi}_{h,\mathbf{v}^{(m)}}$ is the hth entry of vector $\hat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}$ and $\tilde{S}_{hh,\mathbf{v}^{(m)}}$ is the hth entry on the diagonal of matrix $\tilde{S}_{\mathbf{v}^{(m)}}$. Posterior (40) can then be used to numerically construct an approximate $(1-\alpha)\times 100\%$ quantile-based credible interval for ξ_h as follows. Construct an equidistant fine grid, say $\{\xi_{hl}\}_{l=1}^{L}$ of width Δ_l and evaluate the posterior at each element of that grid, i.e. compute $\tilde{p}(\xi_{hl}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m\ \mathcal{N}_1\Big(\xi_{hl}; \hat{\xi}_{h,\mathbf{v}^{(m)}}, \tilde{S}_{hh,\mathbf{v}^{(m)}}\Big)$, for $l = 1,\dots,L$. Then, find the indices $q_{low} \in \{1,\dots,L\}$ and $q_{up} \in \{1,\dots,L\}$, such that $\sum_{l=1}^{q_{low}} \tilde{p}(\xi_{hl}|\mathcal{D})\ \Delta_l \approx \alpha/2$ and $\sum_{l=1}^{q_{up}} \tilde{p}(\xi_{hl}|\mathcal{D})\ \Delta_l \approx 1-(\alpha/2)$. The resulting interval $[\xi_{hq_{low}}, \xi_{hq_{up}}]$ is an approximate $(1-\alpha)\times 100\%$ quantile-based credible interval for ξ_h.

To obtain pointwise set estimates of a smooth function f_j, let $\{x_l\}_{l=1}^{L}$ be an equidistant (fine) grid on the domain of f_j and $\boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ be the subvector of the latent field corresponding to the spline vector $\boldsymbol{\theta}_j = (\theta_{j1},\dots,\theta_{jK-1})^\top$. Also, denote by $\widetilde{\mathbf{b}}_l^\top = (\tilde{b}_{j1}(x_l),\dots,\tilde{b}_{jK-1}(x_l))$ the vector of B-splines in the basis evaluated at x_l. The function f_j at point x_l is thus modeled as $f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j}) = \widetilde{\mathbf{b}}_l^\top \boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ and from (38) the posterior of $\boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ is approximated by the finite mixture:

$$\tilde{p}(\boldsymbol{\xi}_{\boldsymbol{\theta}_j}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m\ \mathcal{N}_{K-1}\left(\hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}, \tilde{S}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}\right), \tag{41}$$

where $\tilde{S}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}$ is a submatrix of $\tilde{S}_{\mathbf{v}^{(m)}}$ corresponding to the variance-covariance matrix of $\boldsymbol{\xi}_{\boldsymbol{\theta}_j}$. As $f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j})$ is a linear combination of the spline vector, a natural candidate to approximate the following posterior $p(f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j})|\mathcal{D})$ is to use a mixture of univariate normals:

$$\tilde{p}(f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j})|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m\ \mathcal{N}_1\Big(\widetilde{\mathbf{b}}_l^\top \hat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}, \widetilde{\mathbf{b}}_l^\top \tilde{S}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}} \widetilde{\mathbf{b}}_l\Big).$$

A quantile-based credible interval for f_j at point x_l can easily be computed from the above (approximate) univariate posterior.

5 Simulation study

The performance of LPS in additive models (with cubic B-splines and a third order penalty) is assessed through different simulation scenarios and compared with results obtained using the `gam()` function of the **mgcv** package in **R** (Wood, 2017), a popular and established toolkit for estimating (generalized) additive models. Options of the `gam()` function are carefully chosen so that the generated results can be meaningfully compared to these obtained using our Laplace-P-spline approach. In particular, smooth terms are specified with the `gam()` function using *s(x, bs="ps", k=K, m=c(2,3))* , where x is the vector of covariate values associated to the estimated smooth function and *ps* specifies a P-spline basis. The scalar k is the basis dimension, the first entry in $m = c(\cdot,\cdot)$ refers to the order of the spline basis (with order 2 corresponding to cubic P-splines), while the second entry refers to the order of the difference penalty. Another chosen option in `gam()` is $method$ = "$REML$", requiring an estimation of the penalty parameters $\boldsymbol{\lambda}$ by restricted maximum likelihood. It corresponds to an empirical Bayes approach in the sense that a Bayesian log marginal likelihood is maximized with respect to $\boldsymbol{\lambda}$ in a context where penalties come from Gaussian priors on the spline coefficients (Marra and Wood, 2011; Wood et al., 2013). The optimization method in `gam` is chosen to be *optimizer=c("outer", "newton")* as it provides reliable and stable computations.

5.1 Simulation results for parameters in the linear part

The first set of simulations consists in $S = 500$ replications of a sample of size $n = 300$ with three covariates in the linear part generated independently as $z_{i1} \sim \text{Bern}(0.5)$, $z_{i2} \sim \mathcal{N}(0,1)$ and $z_{i3} \sim \mathcal{N}(0,1)$, for $i = 1, \ldots, n$ and coefficients $\beta_0 = 0.50,\ \beta_1 = 1.60,\ \beta_2 = -0.80,\ \beta_3 = 0.40$. The covariates for the smooth functions are independent draws from the Uniform distribution on the domain $[-1, 1]$. The functions of interest are partly inspired from Antoniadis et al. (2012) and are given by:

$$
\begin{aligned}
f_1(x_1) &= \cos(2\pi x_1), \\
f_2(x_2) &= 6\ \big(0.1\sin(2\pi x_2) + 0.2\cos(2\pi x_2) + 0.3\sin^2(2\pi x_2) \\
&\quad +0.4\cos^3(2\pi x_2) + 0.5\sin^3(2\pi x_2)\big) - 0.9, \\
f_3(x_3) &= 3x_3^5 + 2\sin(4x_3) + 1.5x_3^2 - 0.5.
\end{aligned}
$$

Three noise levels are considered, namely $\sigma \in \{0.20, 0.40, 0.60\}$, corresponding to a high, medium and low signal to noise ratio. Each smooth function is modeled by a linear combination of cubic B-splines with a third order penalty and $K = 15$ B-splines in $[-1, 1]$. The frequentist properties of the Bayesian estimators are measured by the bias, the empirical standard error (ESE), the root mean square error (RMSE) and coverage probability (CP) of the 90% and 95% (pointwise) credible intervals for the linear coefficients. Figure 4 illustrates the shape of the functions f_1, f_2 and f_3 with a set of simulated data for $n = 300$ with medium signal to noise ratio ($\sigma = 0.40$).

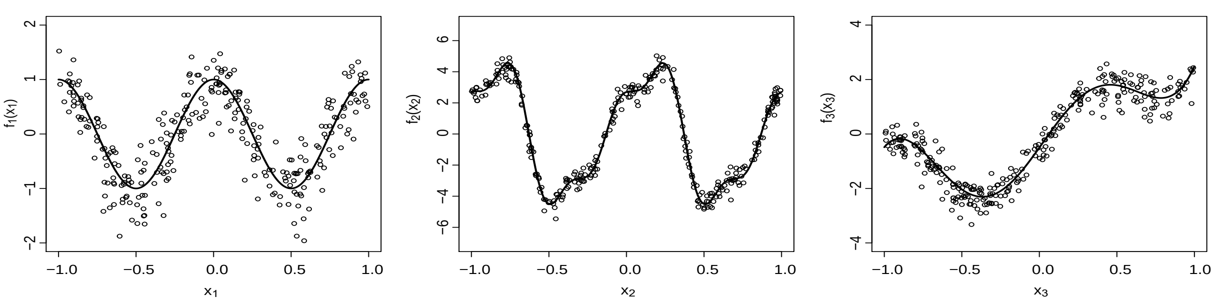

Figure 4: Illustration of functions f_1, f_2, f_3 (solid lines) and simulated data ($n = 300$) under medium signal to noise ratio ($\sigma = 0.40$).

The simulation results given in Table 2 show that our LPS estimation procedure exhibits good performance for the three different noise levels. Nonsignificant biases are observed and the estimated coverage probabilities are close to their nominal value in each setting. Furthermore, LPS and `gam()` have similar results regarding the ESE and RMSE.

In Figure 5, we show the LPS estimaton of the smooth additive terms (gray curves) and the pointwise median (dashed) curves across all replications when 50 B-splines are used for each function. The estimated curves are close to their target on the entire domain except on the boundaries where the estimates exhibit larger variability.

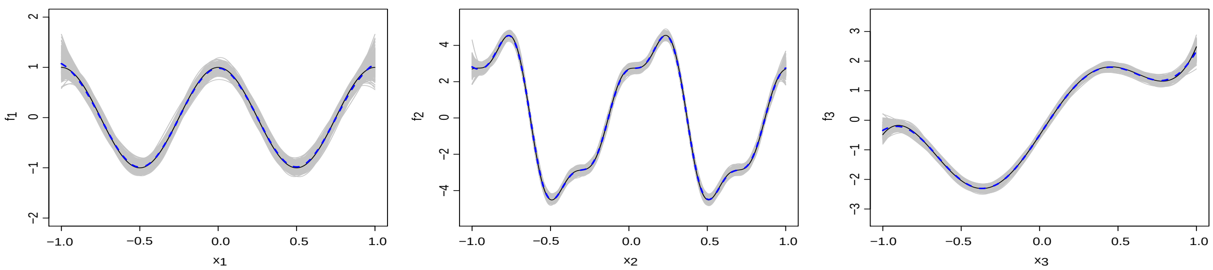

Figure 5: Estimation of the smooth functions f_1, f_2 and f_3 for $S = 500$ replications (one gray curve per dataset), sample size $n = 300$ and $\sigma = 0.40$ using 50 B-splines for each function. The solid (black) curve is the true function and the dashed curve is the pointwise median of the 500 estimated curves.

5.2 Coverage of the smooth functions f_j

To assess the quality of approximate pointwise credible intervals for a function f_j, one can work from a Bayesian perspective and consider a Uniform prior on the probability π_{sj} that the function f_j at point x_{sj} will be contained in the constructed $(1-\alpha) \times 100\%$ credible interval. This is denoted by $\pi_{sj} \sim \mathcal{U}(0,1)$. In addition, let S_{num} denote the number of constructed credible intervals at x_{sj} containing the value $f_j(x_{sj})$ among S datasets. The variable S_{num} follows a Binomial distribution, i.e. $S_{\text{num}} \sim \text{Bin}(S, \pi_{sj})$. From Bayes' rule:

$$\begin{aligned} p(\pi_{sj}|\mathcal{D}) &\propto P(\mathcal{D}|\pi_{sj})\; p(\pi_{sj}) \\ &\propto \pi_{sj}^{s_{\text{num}}}(1-\pi_{sj})^{S-s_{\text{num}}}. \end{aligned}$$

Hence, a posteriori $(\pi_{sj}|\mathcal{D}) \sim \text{Beta}(1+s_{\text{num}}, 1+S-s_{\text{num}})$. We say that the constructed credible interval at x_{sj} is compatible with the nominal value $(1-\alpha)\times 100\%$ at the 99% level provided $(1-\alpha)$ falls within the 0.5th and 99.5th quantiles of the $\text{Beta}(1+s_{\text{num}}, 1+S-s_{\text{num}})$ distribution. This method is equivalent to the hypothesis test $H_0 : \pi_{sj} = (1-\alpha)$ versus $H_1 : \pi_{sj} \neq (1-\alpha)$. If $(1-\alpha)$ falls within the 99% posterior credible interval for π_{sj}, then we do not reject the null. Note also that the posterior mode of the Beta distribution $(\pi_{sj}|\mathcal{D})_{mode} = \frac{s_{\text{num}}}{S}$ corresponds to the point estimator of the coverage probability.

Tables 3 and 4 show the coverage estimates of 90% and 95% pointwise credible intervals for the functions f_1, f_2 and f_3 at selected points of their domain and for three different noise levels with 50 B-splines for each function. The frequentist coverage of credible intervals are compatible with their nominal value for all the considered noise levels for the LPS and `gam`() methods.

σ	Parameters	Bias	$CP_{90\%}$	$CP_{95\%}$	ESE	RMSE
	$\beta_1 = 1.60$	0.003 (0.002)	88.6 (88.6)	94.6 (94.8)	0.040 (0.040)	0.040 (0.040)
0.20	$\beta_2 = -0.80$	0.000 (0.000)	87.6 (89.0)	95.2 (95.2)	0.020 (0.020)	0.020 (0.020)
	$\beta_3 = 0.40$	0.000 (0.000)	88.6 (89.4)	94.8 (95.0)	0.020 (0.020)	0.020 (0.020)
	$\beta_1 = 1.60$	-0.002 (-0.002)	91.0 (91.4)	96.4 (96.4)	0.056 (0.056)	0.056 (0.056)
0.40	$\beta_2 = -0.80$	0.000 (0.000)	89.6 (90.6)	94.2 (93.6)	0.030 (0.029)	0.030 (0.029)
	$\beta_3 = 0.40$	0.000 (-0.001)	89.2 (90.0)	94.8 (95.2)	0.029 (0.029)	0.029 (0.029)
	$\beta_1 = 1.60$	0.000 (0.000)	90.6 (89.8)	95.0 (95.0)	0.079 (0.079)	0.079 (0.079)
0.60	$\beta_2 = -0.80$	-0.003 (-0.003)	88.2 (89.0)	94.8 (95.8)	0.041 (0.040)	0.041 (0.040)
	$\beta_3 = 0.40$	0.000 (0.000)	88.8 (89.0)	94.6 (94.8)	0.042 (0.042)	0.042 (0.042)

Table 2: Simulation results with the LPS method for $S = 500$ replicates of sample size $n = 300$ and $\sigma \in \{0.20, 0.40, 0.60\}$. The values in parentheses are estimation results from the `gam()` (MGCV) method.

σ	f	Method	-0.95	-0.70	-0.50	-0.20	0.00	0.20	0.50	0.70	0.95
0.20	f_1	LPS	89.4	91.6	92.2	92.6	92.2	93.6*	93.6*	94.0*	94.0*
	f_1	MGCV	89.4	91.8	91.4	92.0	92.6	93.2	93.8*	93.6*	94.2*
	f_2	LPS	89.6	92.0	93.2	92.6	93.0	91.8	92.0	90.2	90.8
	f_2	MGCV	90.0	91.8	92.6	92.6	93.4*	91.0	93.0	91.4	90.8
	f_3	LPS	89.0	90.0	91.8	92.0	94.0*	93.2	90.6	93.0	91.6
	f_3	MGCV	88.6	91.0	91.8	92.0	94.0*	93.0	90.8	92.8	91.0
0.40	f_1	LPS	89.6	92.6	90.8	92.2	94.4*	92.0	89.2	92.4	91.2
	f_1	MGCV	89.6	93.0	91.0	92.6	93.8*	92.6	90.2	92.8	91.4
	f_2	LPS	88.2	89.6	93.6*	91.2	89.2	92.0	91.8	91.6	90.6
	f_2	MGCV	88.4	90.2	93.6*	91.0	89.8	92.0	91.8	91.6	90.0
	f_3	LPS	88.8	91.6	93.6*	93.0	94.8*	92.0	91.6	92.2	85.6*
	f_3	MGCV	89.2	90.6	93.8*	93.4*	94.2*	92.0	91.6	91.6	85.6*
0.60	f_1	LPS	90.8	90.4	90.8	94.2*	90.6	93.0	92.2	92.4	88.2
	f_1	MGCV	90.4	90.8	91.6	94.4*	91.6	92.8	92.8	94.0*	88.8
	f_2	LPS	91.4	91.4	90.4	90.0	93.2	90.8	91.2	92.8	94.0*
	f_2	MGCV	90.6	91.2	90.6	90.6	93.0	90.8	91.8	93.2	94.0*
	f_3	LPS	87.8	91.0	91.0	94.2*	93.8*	92.2	92.2	92.8	88.0
	f_3	MGCV	88.0	92.2	90.4	94.2*	93.2	92.2	92.0	93.2	89.2

Table 3: Coverage estimates of 90% pointwise credible intervals of the functions f_1, f_2, f_3 at selected domain points for three noise levels $\sigma \in \{0.20, 0.40, 0.60\}$ over $S = 500$ replications of sample size $n = 300$ for the Laplace-P-spline approach (LPS) and `gam()` (MGCV) method. An asterisk indicates that the estimated coverage is incompatible with the nominal value at the 99% level.

σ	f	Method	-0.95	-0.70	-0.50	-0.20	0.00	0.20	0.50	0.70	0.95
	f_1	LPS	95.0	96.6	96.6	96.8	96.6	97.4	97.0	97.8*	96.8
	f_1	MGCV	95.2	96.2	97.0	96.2	96.8	97.0	97.0	97.8*	97.0
0.20	f_2	LPS	95.2	96.4	96.6	97.0	98.2*	95.4	96.6	96.6	94.6
	f_2	MGCV	95.6	96.8	96.8	97.0	98.0*	96.0	96.0	96.6	94.8
	f_3	LPS	94.4	94.8	97.0	96.8	97.8*	95.6	96.4	96.0	96.8
	f_3	MGCV	94.2	94.4	97.0	96.2	97.6*	96.0	96.2	96.2	96.6
	f_1	LPS	94.8	98.4*	96.4	97.0	97.0	95.8	95.2	96.4	96.2
	f_1	MGCV	94.8	98.0*	96.6	97.2	97.4	96.0	95.2	96.2	95.6
0.40	f_2	LPS	93.4	95.2	96.4	95.4	94.8	96.6	96.4	96.2	95.8
	f_2	MGCV	93.6	95.6	96.6	95.0	95.0	96.4	96.4	96.2	95.6
	f_3	LPS	94.2	96.6	97.4	97.0	98.4*	96.6	96.4	96.2	92.2*
	f_3	MGCV	94.4	96.8	96.8	96.6	98.4*	96.4	96.6	96.2	92.8
	f_1	LPS	94.4	94.6	94.6	96.8	96.0	97.4	96.4	96.6	93.4
	f_1	MGCV	94.8	95.6	95.2	96.8	96.4	97.2	96.4	97.0	93.2
0.60	f_2	LPS	95.8	95.4	95.8	96.4	97.2	96.2	95.8	97.4	96.6
	f_2	MGCV	96.6	96.2	95.8	96.8	97.0	96.2	96.2	97.4	96.8
	f_3	LPS	92.4*	95.4	96.4	96.6	98.0*	96.4	96.8	97.0	93.6
	f_3	MGCV	92.6	95.8	95.6	96.4	97.4	96.0	97.2	97.4	94.6

Table 4: Coverage estimates of 95% pointwise credible intervals of the functions f_1, f_2, f_3 at selected domain points for three noise levels $\sigma \in \{0.20, 0.40, 0.60\}$ over $S = 500$ replications of sample size $n = 300$ for the Laplace-P-spline approach (LPS) and `gam()` (MGCV) method. An asterisk indicates that the estimated coverage is incompatible with the nominal value at the 99% level.

6 Application to Milan mortality data

In this section, the LPS methodology is illustrated on the Milan mortality data (Ruppert et al., 2003) available in the **SemiPar** package on CRAN (`https://CRAN.R-project.org/package=SemiPar`). The dataset contains observations on $n = 3652$ consecutive days between January 1st, 1980 and December 30th, 1989 for the city of Milan in Italy for air pollution indicators and health variables. The objective is to study how air pollution and other meteorological indicators impact mortality using an additive partial linear model. In that endeavor, the square root of the total number of death (*Mortality*) is taken to be the response variable. Following Ruppert et al. (2003), the variable *TSP* measuring the total suspended particles in ambient air enters as a linear predictor. The dichotomous variable *Holiday* is an indicator of public holiday (1=public holiday; 0=otherwise) and is also naturally added in the linear part of the model. The remaining predictors are modeled as smooth functions, namely: the mean daily temperature in °C (*Temperature*), the relative humidity (*Humidity*), a measure of sulfur dioxide (SO_2) in ambient air and the number of days (*Numdays*) elapsed as from December 31st, 1979. Figure 6 provides a graphical illustration for some data variables. The quantile-quantile plot of the response variable on the top-left graph confirms that *Mortality* is approximately normally distributed. The scatter plots of the response with *Temperature*, *Humidity* and SO_2 and the associated locally estimated scatterplot smoothing (LOESS) fit in red suggest that the latter variables are nonlinearly related to *Mortality*.

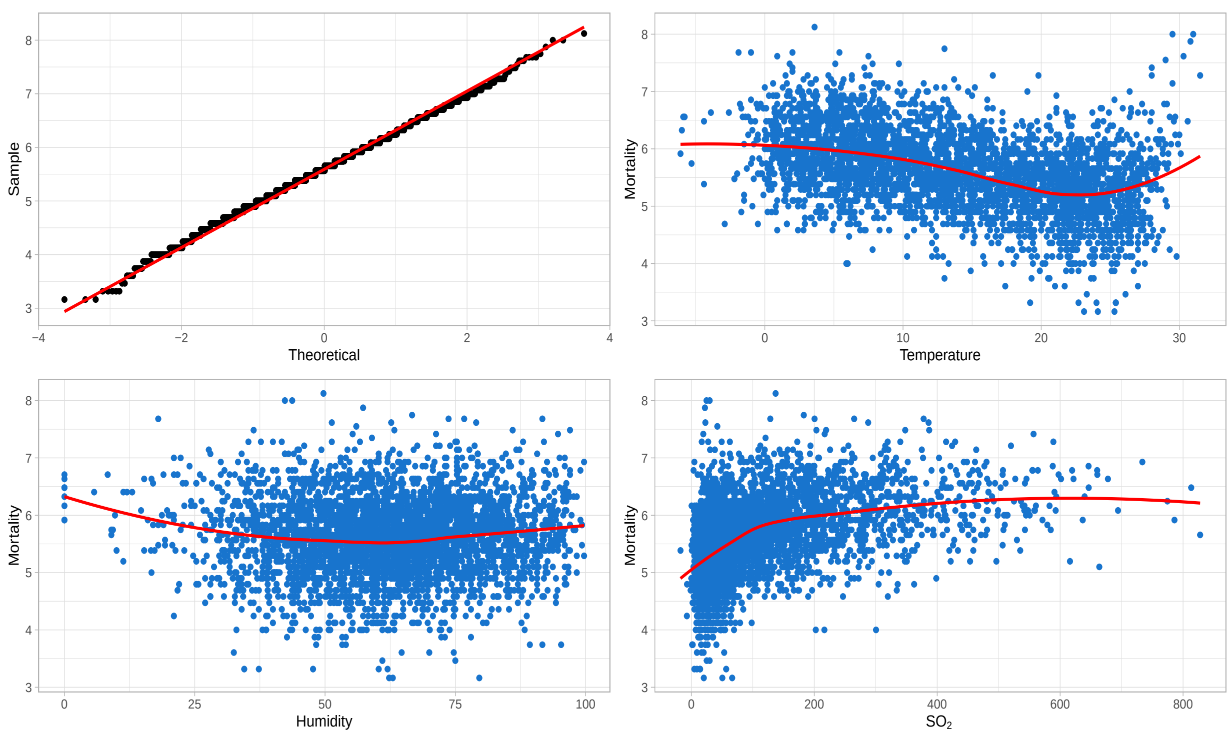

Figure 6: The Milan mortality data. Top-left: Q-Q plot of the response variable *Mortality*. Top-right: Scatter plot of *Mortality* and *Temperature*. Bottom-left: Scatter plot of *Mortality* across *Humidity*. Bottom-right: Scatter plot of the response and SO_2.

The additive model for the mortality data is written as:

$$\begin{aligned} Mortality_i &= \beta_0 + \beta_1 TSP_i + \beta_2 Holiday_i + f_1(Temperature_i) \\ &\quad + f_2(Humidity_i) + f_3(SO_{2i}) + f_4(Numdays) + \varepsilon_i,\ i = 1, \ldots, n, \end{aligned} \tag{42}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and smooth terms $f_j,\ j = 1, 2, 3$ modeled with 35 cubic B-splines and a second order penalty. The B-spline basis for a smooth term f_j is defined over the domain $[x_{j,\min}, x_{j,\max}]$, i.e. over the range of its observed values x_j. Estimation results for *TSP* and *Holiday* are summarized in Table 5. *TSP* has a small positive and significant effect on the response, while *Holiday* has a negative and significant effect.

Parameters	**Estimates**	**CI 95%**	$\mathbf{sd}_{post}$
β_1 (*TSP*)	0.0006	[0.0001; 0.0010]	0.0002
β_2 (*Holiday*)	-0.1240	[-0.2342; -0.0164]	0.0558

Table 5: Estimation results for the parametric linear part of the additive model. The second column is the parameter estimate, the third column gives the associated 95% credible interval and the last column is the posterior standard deviation.

Figure 7 shows the estimated additive terms with approximate 95% pointwise credible intervals. We see that the conditional impact of *Temperature* on the mean response is slightly decreasing until approximately 25°C after which an explosive increase indicates that higher temperatures are associated to an important increase in the expected number of deaths. *Humidity* seems to have no significant impact on the response as it remains stable around zero.

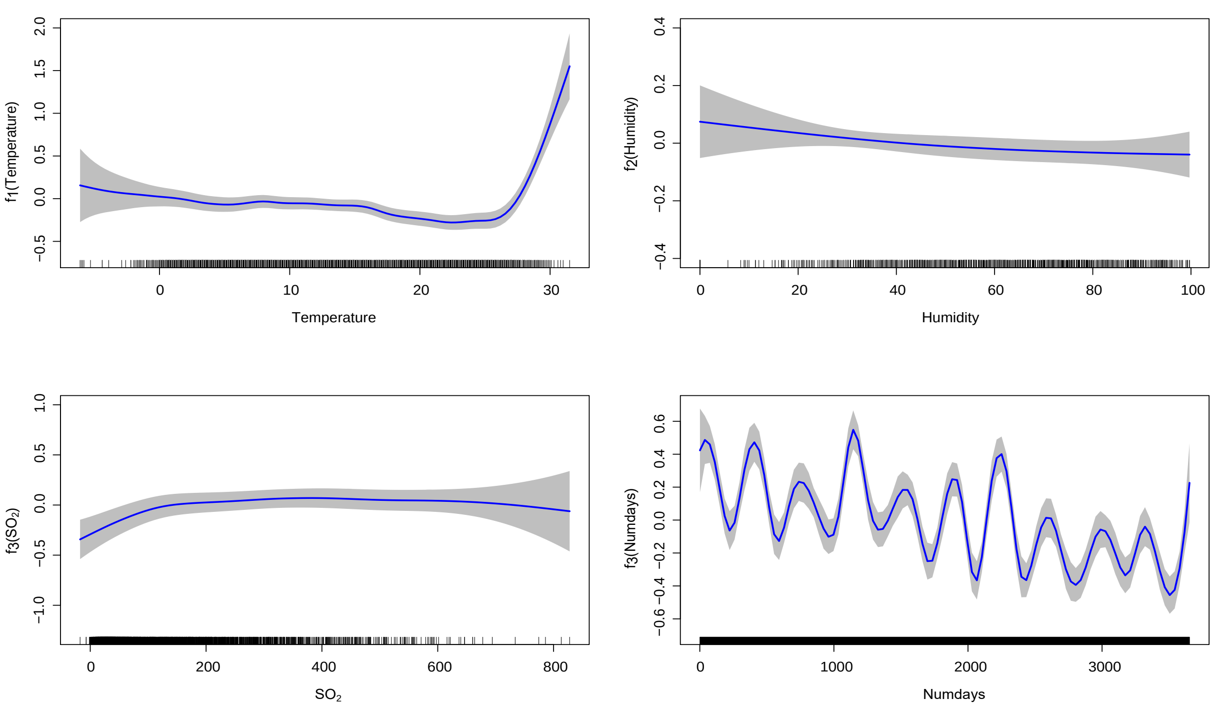

Figure 7: Estimates of the nonlinear predictors with 95% pointwise credible interval.

An increase in SO_2 levels from 0 to 180 is associated to an increase in average mortality. However, further increase of the SO_2 concentrations in ambient air seems to have negligible impact on the mean response as the smooth estimated term remains flat with a plausible zero value for the slope. For *Numdays*, we observe the seasonal pattern already reported in Ruppert et al. (2003), i.e. average mortality fluctuates over seasons with spikes arising during winter.

Conclusion

The core contribution of this paper is to adapt the Laplace-P-spline (LPS) methodology for fast approximate Bayesian inference in additive models with Gaussian errors. Working from a Bayesian perspective, we model the smooth additive terms with penalized B-splines and impose a Gaussian prior on the latent field, which is composed of linear regression coefficients and spline amplitudes.

After having introduced the theoretical foundations of the model, we derive the conditional posterior of the latent vector and use the latter to obtain an expression of the marginal posterior of the penalty vector. Important efforts have been invested in the derivation of the gradient and Hessian of the log posterior of the (log-) penalty vector as it enables to avoid numerical differentiation to obtain its posterior mode and hence accelerates the computational process behind Newton-Raphson.

To efficiently explore the posterior penalty space, we develop a strategy which consists in adjusting a skew-normal distribution to the conditional posterior of the (log-) penalty parameters at their modal value. This method has the merit of capturing potential asymmetries in the posterior penalty and hence allows a precise grid-based exploration. The constructed grid is then used to compute an approximate version of the joint posterior latent vector resulting in a finite mixture of multivariate Gaussian distributions from which point and set estimators can be derived.

The main limitation behind a grid exploration of the posterior penalty space is an exponentially growing computational budget with the number q of smooth functions in the additive model. To alleviate the problem, a possibility is to implement a hybrid approach that alternates between a grid for small or moderate q and a classic MCMC algorithm when q is above a certain threshold. It is also worth noting that our LPS algorithm requires a low computational budget even though the modeling approach is fully Bayesian.

References

Antoniadis, A., Gijbels, I., and Verhasselt, A. (2012). Variable selection in additive models using P-splines. *Technometrics*, 54(4):425–438.

Azzalini, A. (1985). A class of distributions which includes the Normal ones. *Scandinavian Journal of Statistics*, 12(2):171–178.

Azzalini, A. (2014). *The Skew-Normal and Related Families*, volume 3. Cambridge University Press.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510.

Durbán, M. and Currie, I. D. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics*, 18(2):251–262.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–102.

Fan, Y. and Li, Q. (2003). A kernel-based method for estimating additive partially linear models. *Statistica Sinica*, 13(3):739–762.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American Statistical Association*, 76(376):817–823.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer New York.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models, volume 43 of Monographs on Statistics and Applied Probability*. Chapman & Hall, London.

Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. John Wiley & Sons.

Kroese, D. P., Taimre, T., and Botev, Z. I. (2013). *Handbook of Monte Carlo Methods*. John Wiley & Sons.

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

Liang, H., Thurston, S. W., Ruppert, D., Apanasovich, T., and Hauser, R. (2008). Additive partial linear models with measurement errors. *Biometrika*, 95(3):667–678.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100.

Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *Journal of Statistical Planning and Inference*, 141(1):204–219.

Ma, W. and Kruth, J.-P. (1995). Parameterization of randomly measured points for least squares fitting of B-spline curves and surfaces. *Computer-Aided Design*, 27(9):663–675.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

Martins, T. G., Simpson, D., and Lindgren, F.and Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, 67:68–83.

O'Hagan, A., Kendall, M. G., and Forster, J. (2004). Kendall's Advanced Theory of Statistics: Bayesian Statistics. Vol. 2B.

Opsomer, J. D. and Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, 8(4):715–732.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society, Series B*, 71(2):319–392.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

Tjøstheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *Journal of the American Statistical Association*, 89(428):1398–1409.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R (Second edition)*. CRC press.

Wood, S. N., Scheipl, F., and Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, 23(3):341–360.