Final article published in *Computational Statistics & Data Analysis,* Volume 154, February 2021.

https://doi.org/10.1016/j.csda.2020.107088

Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines

Oswaldo Gressani[a,*], 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

Generalized additive models (GAMs) are a well-established statistical tool for modeling complex nonlinear relationships between covariates and a response assumed to have a conditional distribution in the exponential family. To make inference in this model class, a fast and flexible approach is considered based on Bayesian P-splines and the Laplace approximation. The proposed Laplace-P-spline model contributes to the development of a new methodology to explore the posterior penalty space by considering a deterministic grid-based strategy or a Markov chain sampler, depending on the number of smooth additive terms in the predictor. The approach has the merit of relying on a simple Gaussian approximation to the conditional posterior of latent variables with closed form analytical expressions available for the gradient and Hessian of the approximate posterior penalty vector. This enables to construct accurate posterior pointwise and credible set estimators for (functions of) regression and spline parameters at a relatively low computational budget even for a large number of smooth additive components. The performance of the Laplace-P-spline model is confirmed through different simulation scenarios and the method is illustrated on two real datasets.

Keywords: Laplace approximation; Generalized additive models; Fast Bayesian computation; P-splines

*Corresponding author `oswaldo_gressani@hotmail.fr`

1. Introduction

Generalized additive models (GAMs) (Hastie & Tibshirani, 1986, 1987) extend generalized linear models (Nelder & Wedderburn, 1972) by having nonlinear smooth functions of quantitative covariates entering the linear predictor: they enable to relate in a flexible way covariates to the mean of a conditional distribution in the exponential family. The monograph of Hastie & Tibshirani (1990) gives a thorough introduction to additive regression structures and largely contributed to the dissemination of this model class. Ruppert et al. (2003) and Wood (2017) provide a complete and comprehensive treatment of GAMs, emphasizing on semiparametric methods and penalized regression splines.

There exists a large variety of regression splines in the literature for modeling the smooth terms in a GAM, for instance P-splines (Eilers & Marx, 1996), thin plate splines (Wood, 2003), O'Sullivan penalized splines (Wand & Ormerod, 2008) or adaptive splines (Krivobokova et al., 2008) to cite some popular instances. P-splines refer to a penalized B-splines basis, i.e. a basis defined on a compact support and constructed from a set of polynomial pieces joined together by "knots", where the penalty acts upon differences of adjacent B-spline coefficients. This article focuses exclusively on P-spline smoothers for two main reasons. First, the penalty matrix can be effortlessly constructed from basic difference formulas, keeping the penalization scheme simple and the P-spline approach numerically stable. Second, the attractiveness of P-splines lies in its rather natural extension to a Bayesian setting (Lang & Brezger, 2004) and from the efficiency of working with sparse bases and penalties for sampling-free approximate Bayesian inference (without requiring stochastic draws from a target distribution such as in Markov chain Monte Carlo (MCMC) methods).

As MCMC techniques can be subject to poor chain convergence and tend to carry a heavy computational burden, Rue et al. (2009) introduced an approximate Bayesian methodology based on Laplace approximations termed Integrated Nested Laplace Approximations (INLA), a completely sampling-free framework that delivers accurate and fast approximations of posterior marginals

in structured additive regression models. More recent articles on fast approximate likelihood or Bayesian-based inference include Luts et al. (2014), Wand (2017) and Hui et al. (2019) among others. Taken separately, P-splines and INLA have made an impressive impact in the statistical community and initiated a flourishing literature in diversified domains (see e.g. Eilers et al., 2015; Rue et al., 2017), yet a few references attempted to unify the strength of both approaches. Fraaije et al. (2015) designed field experiments to study vegetation patterns and plant diversity in riparian areas and relied on P-splines to model the response of plant species, with INLA as the underlying fitting mechanism. Ventrucci & Rue (2016) focus on the prior choice for the precision hyperparameter that controls the amount of smoothness in a Bayesian P-spline setting. They propose penalized complexity priors, an alternative prior to the classic Gamma family and use INLA to derive the posterior of spline coefficients. In survival analysis, Gressani & Lambert (2018) combine P-splines with Laplace approximations to develop an inferential tool in the class of promotion time cure models.

In the present article, we borrow some ideas from INLA and combine them with P-splines to design the Laplace-P-spline (LPS) methodology, a novel unified approach for approximate Bayesian inference in generalized additive models. Although INLA is a well-tailored approach for making inference in a variety of statistical models, there is room for further computational improvements when considering the specific class of GAMs. In particular, the use of numerical differentiation techniques in INLA to obtain finite difference approximations to the gradient and Hessian matrix of the posterior penalty vector can be replaced by their exact analytical expressions, yielding more efficient algorithms for model fitting. Furthermore, as the computational cost grows exponentially with the dimension of the penalty vector, in grid-based derivation of the marginal posterior of the regression parameters, alternative strategies are required to explore the posterior penalty space when the number of additive terms is large. Our methodology is free of the numerical differentiation scheme found in INLA, as it relies on closed analytical expressions for the gradient and Hessian required

during computation. It enables not only to fasten our code, but also offers a clear insight on the equations governing the implementation of the model. Moreover, we exploit this analytical availability to develop a novel cost-effective grid exploration algorithm to explore the posterior of the hyperparameters corresponding, in our specific context, to the penalty parameters controlling the smoothness of each additive term. The method accounts for possible asymmetries in the posterior hyperparameter space by applying a moment-matching technique with reference to the skew-normal family. Finally, in response to the "curse of dimensionality" related to the increase in computational resources with the hyperparameter dimension, we suggest to embed a regular MCMC algorithm to explore the hyperparameter posterior instead of the classic grid exploration when the dimension grows above a certain threshold.

The remainder of the article is outlined as follows. In Section 2, the Bayesian Laplace-P-spline generalized additive model is formulated and the Laplace approximation to the conditional posterior of regression and spline parameters is derived. Section 3 is devoted to posterior inference on these parameters. To efficiently explore the approximate marginal posterior of the penalty parameters, we propose a strategy that alternates between a deterministic grid and an independence Metropolis-Hastings sampler depending on the number of smooth additive components. The chosen penalty values are then used to approximate the marginal posterior for the vector of regression and spline parameters with their associated pointwise credible intervals. A detailed simulation study is presented in Section 4 with comparisons against a popular benchmark method. Section 5 illustrates the LPS model on two real datasets and Section 6 closes the paper with concluding remarks and sketches future research prospects.

2. The Laplace-P-spline generalized additive model

2.1. Flexible modeling with Bayesian P-splines

We consider a GAM where the response variable has a distribution belonging to the one-parameter exponential family $y_i \sim \text{EF}(\gamma_i, \varkappa)$ characterized by densities of the form:

$$p(y_i;\gamma_i,\varkappa) = \exp\left(\frac{y_i\gamma_i - s(\gamma_i)}{\varkappa} + c(y_i,\varkappa)\right), \tag{1}$$

where $s(\cdot)$ is a twice continuously differentiable real-valued function and $c(\cdot,\cdot)$ another real function, $\varkappa > 0$ is a known scale or dispersion parameter and γ_i is the natural or canonical parameter. Using well-known properties of the score function (McCullagh & Nelder, 1989), one can show that the mean and variance of the response are $\mathbb{E}(y_i) := \mu_i = s'(\gamma_i)$ and $\mathrm{Var}(y_i) = \varkappa s''(\gamma_i)$ respectively. Let $\mathcal{D} = \{(y_i, \mathbf{x}_i, \mathbf{z}_i) : i = 1,\ldots,n\}$ be a sample of n independent observations, where $\mathbf{x}_i = (x_{i1},\ldots,x_{iq})^T$ is a vector of continuous covariates and $\mathbf{z}_i = (z_{i1},\ldots,z_{ip})^T$ a vector of additional covariates (possibly categorical). The link function $g(\cdot)$ relates the mean response to the additive predictor as follows:

$$g(\mu_i) := \varrho_i = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + f_1(x_{i1}) + \cdots + f_q(x_{iq}), \quad i = 1,\ldots,n, \tag{2}$$

with regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$. In the spirit of the P-spline approach proposed in Eilers & Marx (1996), the unknown smooth functions f_j, $j = 1,\ldots,q$ are modeled with rich cubic B-spline bases and a discrete penalty on neighboring spline coefficients is imposed for controlling the roughness of the fit. Mathematically:

$$f_j(x_{ij}) = \sum_{k=1}^{K} \theta_{jk} b_{jk}(x_{ij}), \quad j = 1,\ldots,q, \tag{3}$$

where for simplicity the same number K of basis functions $b_{jk}(\cdot)$ is assumed for every f_j. The vector of B-spline coefficients associated to function f_j is $\boldsymbol{\theta}_j = (\theta_{j1},\ldots,\theta_{jK})^T$, while the collection of all spline coefficients present in the model is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T,\ldots,\boldsymbol{\theta}_q^T)^T$ and the vector of B-spline functions at x_{ij} is written as $\mathbf{b}_j(x_{ij}) = (b_{j1}(x_{ij}),\ldots,b_{jK}(x_{ij}))^T$. Model flexibility is compensated by a roughness penalty on finite differences of the coefficients of contiguous B-splines, $\boldsymbol{\theta}^T\mathcal{P}(\boldsymbol{\lambda})\boldsymbol{\theta}$, with block diagonal matrix $\mathcal{P}(\boldsymbol{\lambda})$ expressed compactly using a Kronecker product:

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

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_q)^T$ is a vector of positive penalty parameters and $P = D_r^T D_r + \epsilon I_K$ is a penalty matrix resulting from the product of rth order difference matrices D_r of dimension $(K - r) \times K$ to which a diagonal perturbation ϵI_K is added (with $\epsilon = 10^{-6}$, say), so that P is full rank. From a Bayesian perspective, Lang & Brezger (2004) suggest to obtain the roughness penalty by imposing a multivariate Gaussian prior on the spline amplitudes $\boldsymbol{\theta}|\boldsymbol{\lambda} \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\big(0, \mathcal{P}^{-1}(\boldsymbol{\lambda})\big)$. Furthermore, a Gaussian prior is assumed on the regression coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$, more specifically $\boldsymbol{\beta} \sim \mathcal{N}_{\dim(\boldsymbol{\beta})}(0, V_{\boldsymbol{\beta}}^{-1})$ with matrix $V_{\boldsymbol{\beta}} = \zeta I_{p+1}$ and small precision (say $\zeta = 10^{-5}$). The latent vector of the model is written as $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ and includes the regression and spline coefficients with prior distribution $\boldsymbol{\xi}|\boldsymbol{\lambda} \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\big(0, \big(Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\big)^{-1}\big)$ and precision 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 centered around their mean value. 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 for $j = 1, \ldots, q$ as follows:

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

To reach an identifiable model, we impose the following centering on the B-spline matrices $\widetilde{B}_j = B_j - (\mathbf{1}_n \mathbf{1}_L^T / L)\breve{B}_j,\ j = 1, \ldots, q$, where $\mathbf{1}_n$ and $\mathbf{1}_L$ are column vector of ones of length n and L respectively and $\breve{B}_j$ is a B-spline matrix computed

on a fine grid of equidistant values on the domain of f_j. This identifiability constraint centers the additive functional components around their average value. To ensure that all 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 vector of regression and spline parameters has dimension $\dim(\boldsymbol{\xi}) = q \times (K-1) + p + 1$.

Following Jullion & Lambert (2007), robust priors are specified on the roughness penalty parameters with a conjugate Gamma family having a hierarchical structure:

$$\lambda_j|\delta_j \sim \mathcal{G}(\nu/2, (\nu\delta_j)/2), \quad j = 1, \ldots, q.$$

An uninformative distribution is imposed on the hyperparameter δ_j :

$$\delta_j \sim \mathcal{G}(a_\delta, b_\delta), \ j = 1, \ldots, q,$$

with mean a_δ/b_δ and variance a_δ/b_δ^2. The authors show that when $a_\delta = b_\delta$ are calibrated to a small value (say 10^{-4}), the fitted curves are not sensitive to the value taken by ν (here $\nu = 3$). The penalty parameters are gathered in the vector $\boldsymbol{\eta} = (\boldsymbol{\lambda}^T, \boldsymbol{\delta}^T)^T$ and the vector of (additive) predictor variables is $\boldsymbol{\varrho} = (\varrho_1, \ldots, \varrho_n)^T$. Taking into account the identifiability constraint, the additive predictor in (2) can be expressed compactly as $\boldsymbol{\varrho} = B\boldsymbol{\xi}$, 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. The Bayesian model is summarized as follows:

$$\begin{aligned}
&y_i|\boldsymbol{\xi} \sim \mathrm{EF}(\gamma_i, \varkappa), \quad i = 1, \ldots, n, \\
&\boldsymbol{\xi}|\boldsymbol{\lambda} \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\big(0, \big(Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\big)^{-1}\big), \\
&\lambda_j|\delta_j \sim \mathcal{G}(\nu/2, (\nu\delta_j)/2), \quad j = 1, \ldots, q, \\
&\delta_j \sim \mathcal{G}(a_\delta, b_\delta), \quad j = 1, \ldots, q.
\end{aligned}$$

2.2. Approximated conditional posterior of the latent vector

Let us denote by $\ell(\boldsymbol{\xi};\mathcal{D}) = (1/\varkappa)\sum_{i=1}^{n}(y_i\gamma_i - s(\gamma_i)) + c$, with $c := \sum_{i=1}^{n} c(y_i,\varkappa)$ (for ease of notation) the log-likelihood function following from Equation (1). From the standard theory of exponential families, we know that the score vector is given by $\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\xi};\mathcal{D}) = B^T W D_g(\boldsymbol{y}-\boldsymbol{\mu})$, where $W := \text{diag}(w_1,\ldots,w_n)$ is a diagonal matrix with weights on the diagonal defined as $w_i := \left(\text{Var}(y_i)[g'(\mu_i)]^2\right)^{-1}$ and $D_g = \text{diag}(g'(\mu_1),\ldots,g'(\mu_n))$. Moreover, the observed Fisher information matrix (equal to the negative Hessian of the log-likelihood) is given by $-\nabla^2_{\boldsymbol{\xi}}\ell(\boldsymbol{\xi};\mathcal{D}) = B^T W B$. Using Bayes' theorem, the conditional posterior of $\boldsymbol{\xi}$ is proportional to the product of the likelihood and prior, which can be written as $p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D}) \propto \exp\left(\ell(\boldsymbol{\xi};\mathcal{D}) - (1/2)\boldsymbol{\xi}^T Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}}\boldsymbol{\xi}\right)$. Using the Newton-Raphson algorithm, we compute the mode $\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}$ of the conditional posterior $p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$ and use Laplace's method to approximate the latter by a normal density denoted by $\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$. Mathematically, Laplace's method for approximating a multivariate (and differentiable) posterior distribution, say $p(\mathbf{x}|\mathcal{D})$, consists in, first, computing the posterior mode $\widehat{\mathbf{x}}$ by maximizing either analytically or numerically $\log p(\mathbf{x}|\mathcal{D})$, and, second, computing the Hessian matrix of $\log p(\mathbf{x}|\mathcal{D})$ evaluated at $\widehat{\mathbf{x}}$, i.e. $\mathcal{H}(\widehat{\mathbf{x}})$. The resulting Laplace approximation to $p(\mathbf{x}|\mathcal{D})$ is a Gaussian distribution with mean $\widehat{\mathbf{x}}$ and variance-covariance matrix equal to $-(\mathcal{H}(\widehat{\mathbf{x}}))^{-1}$, see e.g. Bornkamp (2011). After convergence of the Newton-Raphson algorithm, the Laplace approximation to the conditional posterior latent vector is a Gaussian distribution with mean $\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} = (B^T\widetilde{W}B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}})^{-1}\widetilde{\boldsymbol{\varpi}}$ and covariance matrix $\widehat{\Sigma}_{\boldsymbol{\lambda}} = (B^T\widetilde{W}B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}})^{-1}$, where $\widetilde{W}$ is the weight matrix at convergence and $\widetilde{\boldsymbol{\varpi}}$ is the vector at convergence that results from the sequence $\boldsymbol{\varpi}^{(0)}, \boldsymbol{\varpi}^{(1)}, \boldsymbol{\varpi}^{(2)}, \ldots$, with $\boldsymbol{\varpi}^{(0)} := (1/\varkappa)B^T\left(\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\xi}^{(0)})\right) + B^T W\left(\boldsymbol{\xi}^{(0)}\right)B\boldsymbol{\xi}^{(0)}$ computed from an initial guess $\boldsymbol{\xi}^{(0)}$, e.g. a zero vector for the regression and spline parameters. The Laplace approximation $\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$ will be used to approximate the integrand entering the computation of the marginal posterior for $\boldsymbol{\xi}$:

$$p(\boldsymbol{\xi}|\mathcal{D}) = \int_{\mathbb{R}^q_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})\; p(\boldsymbol{\lambda}|\mathcal{D})\; d\boldsymbol{\lambda}. \tag{4}$$

2.3. Approximated marginal posterior of the penalty parameters

An indispensable intermediate step to reach an approximated version for the marginal posterior of $\boldsymbol{\xi}$ in (4) is to obtain the marginal posterior of the vector of penalty parameters $p(\boldsymbol{\lambda}|\mathcal{D})$. In that endeavor, we first derive an approximation of $p(\boldsymbol{\eta}|\mathcal{D})$ in the philosophy of Leonard (1982), Tierney & Kadane (1986) and Rue et al. (2009) and show how $\boldsymbol{\delta}$ can be integrated out, resulting in an approximation of the marginal posterior for the roughness penalty vector $\boldsymbol{\lambda}$. The gradient and Hessian of that log posterior are analytically derived and will be very useful to explore the support of the posterior distribution of the penalty vector. The posterior of the hyperparameter vector is given by:

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

where $\mathcal{L}(\boldsymbol{\xi};\mathcal{D})$ is the likelihood function. An approximation $\widetilde{p}(\boldsymbol{\eta}|\mathcal{D})$ to the above marginal posterior of $\boldsymbol{\eta}$ is obtained by substituting the Laplace approximation to $p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$ (cf. Section 2.2) and by evaluating the resulting expression at the posterior mode $\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}$. Let us express the natural parameter in the generalized additive model as $\gamma_i = \varrho_i = \mathbf{b}_i^T\boldsymbol{\xi}$, with $\mathbf{b}_i^T$ the row vector corresponding to the ith row of matrix B. Using the previous suggestion and noting that the determinant of the block diagonal matrix involved in the prior $p(\boldsymbol{\xi}|\boldsymbol{\lambda})$ is given by $|Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}} \propto \prod_{j=1}^{q}\lambda_j^{(K-1)/2}$, we obtain:

$$\begin{aligned} \widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) \propto\ & \exp\left(\frac{1}{\varkappa}\sum_{i=1}^{n}\left[y_i\mathbf{b}_i^T\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} - s\left(\mathbf{b}_i^T\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right)\right] - \frac{1}{2}\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}^T Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right) \\ &\times \left(\prod_{j=1}^{q}\delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\left(-\delta_j\left(b_\delta+\frac{\nu}{2}\lambda_j\right)\right)\right)\left(\prod_{j=1}^{q}\lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right) \\ &\times |B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}. \end{aligned} \tag{5}$$

As Gamma priors have been chosen for the penalty parameters λ_j and δ_j, one recognizes in (5) the conditional conjugacy for δ_j, as $\delta_j|\lambda_j, \mathcal{D} \sim \mathcal{G}\big(\frac{\nu}{2} + a_\delta, b_\delta + \frac{\nu}{2}\lambda_j\big)$. Under these prior specifications, the integration of (5) with respect to $\boldsymbol{\delta}$ is tractable and yields the (approximate) marginal penalty posterior:

$$\begin{aligned}
\widetilde{p}(\boldsymbol{\lambda}|\mathcal{D}) &= \int_0^{+\infty} \cdots \int_0^{+\infty} \widetilde{p}(\boldsymbol{\eta}|\mathcal{D})\ d\delta_1 \ldots d\delta_q \\
&\propto |B^T \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} \exp\left(\frac{1}{\varkappa}\sum_{i=1}^{n}\left[y_i \mathbf{b}_i^T \widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} - s\left(\mathbf{b}_i^T \widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right)\right] - \frac{1}{2}\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}^T Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}} \widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right) \\
&\quad \times \left(\prod_{j=1}^{q} \lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right)\left(\prod_{j=1}^{q}\left(b_\delta + \frac{\nu}{2}\lambda_j\right)^{-\left(\frac{\nu}{2}+a_\delta\right)}\right). \qquad (6)
\end{aligned}$$

Applying a log transform on the penalty parameters $v_j = \log(\lambda_j),\ j = 1, \ldots, q$ and using the multivariate Jacobian formula on (6), we obtain the following expression for the (log-) posterior of the log penalty vector:

$$\begin{aligned}
\log \widetilde{p}(\mathbf{v}|\mathcal{D}) &\doteq -\frac{1}{2}\log|B^T \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}| + \frac{\nu + K - 1}{2}\sum_{j=1}^{q} v_j + \frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^T \widehat{\boldsymbol{\xi}}_{\mathbf{v}} \\
&\quad -\frac{1}{\varkappa}\sum_{i=1}^{n} s\left(\mathbf{b}_i^T \widehat{\boldsymbol{\xi}}_{\mathbf{v}}\right) - \frac{1}{2}\widehat{\boldsymbol{\xi}}_{\mathbf{v}}^T Q_{\boldsymbol{\xi}}^{\mathbf{v}} \widehat{\boldsymbol{\xi}}_{\mathbf{v}} - \left(\frac{\nu}{2} + a_\delta\right)\sum_{j=1}^{q}\log\left(b_\delta + \frac{\nu}{2}\exp(v_j)\right), \qquad (7)
\end{aligned}$$

where $\doteq$ denotes equality up to an additive constant, $Q_{\boldsymbol{\xi}}^{\mathbf{v}}$ is the symmetric block diagonal matrix:

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

and $\widehat{\boldsymbol{\xi}}_{\mathbf{v}} := \left(B^T \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\right)^{-1} \widetilde{\boldsymbol{\varpi}}$. The gradient $\nabla_{\mathbf{v}} \log \widetilde{p}(\mathbf{v}|\mathcal{D})$ and Hessian $\nabla_{\mathbf{v}}^2 \log \widetilde{p}(\mathbf{v}|\mathcal{D})$ of expression (7) can be analytically derived, see Appendix A for full details. These expressions will turn to be useful to explore the marginal posterior of the penalty parameters.

Although some similarities are apparent between LPS and INLA, especially in the approach for approximating the hyperparameter vector $p(\boldsymbol{\eta}|\mathcal{D})$, there are noteworthy methodological differences. Classic INLA is inherently focusing on posterior marginals of univariate latent variables, while LPS is natively multivariate and emphasizes on approximating the marginal joint posterior latent vector (4). Also, the smooth terms of the GAM model are exclusively modeled with P-splines, with full-fledged analytical availability of the gradient and Hessian of the posterior penalty vector, whereas INLA uses numerical differentiation techniques. Another fundamental difference lies in the specification of the vector $\boldsymbol{\xi}$: INLA works with a latent vector having a dimension proportional to the sample size n, while in LPS it is independent of n.

3. Posterior inference on the marginal latent vector

3.1. Exploration of the posterior penalty vector

An approximation to the marginal posterior of the latent vector $\boldsymbol{\xi}$ (including the regression and spline parameters in the generalized additive model) can be obtained by integrating out the penalty parameters as in (4). Obtaining such a quadrature requires to explore the posterior of the penalty parameters $\boldsymbol{\lambda} = \exp(\mathbf{v})$. Two strategies are suggested according to the dimension q of the penalty vector. When q is small or moderate (say $q \leq 4$), a grid strategy is proposed that is sensitive to asymmetries in the response surface $\widetilde{p}(\mathbf{v}|\mathcal{D})$, with the skew-normal family of distributions forming the backbone to handle asymmetry. As the computational cost of constructing a grid grows with dimension q, we suggest, when q is large, an alternative strategy relying on MCMC (Yoon & Wilson, 2011; Gómez-Rubio & Rue, 2017) to draw a set of points in the domain of the posterior of the penalty parameters.

This hybrid approach alternates between a deterministic grid and a sampling scheme, giving to the end-user a complete and rapid tool to fit GAMs in a full Bayesian framework even when the number of smooth functions is large. A preliminary milestone for both strategies is to find the posterior mode $\hat{\mathbf{v}}$ of

$\log \widetilde{p}(\mathbf{v}|\mathcal{D})$ as it represents the "center of gravity" from which the exploration will depart. To this end, a Newton-Raphson algorithm is implemented in which we take advantage of the analytical forms for the gradient and Hessian of $\log \widetilde{p}(\mathbf{v}|\mathcal{D})$ to speed up the computational process. Once $\hat{\mathbf{v}}$ is obtained, we proceed with posterior exploration.

3.2. Hybrid exploration alternating between grids and independence sampling

An elementary approach to explore $\widetilde{p}(\mathbf{v}|\mathcal{D})$ could rely on a multivariate Gaussian approximation to the posterior of the log penalty parameters $\mathbf{v}$, i.e. $\widetilde{p}_G(\mathbf{v}|\mathcal{D}) = \mathcal{N}_{\dim(\mathbf{v})}\left(\widehat{\mathbf{v}}, \left(-\mathcal{H}^*\right)^{-1}\right)$, where the covariance matrix is obtained from the Hessian $\mathcal{H}^* = \nabla^2_{\mathbf{v}} \log \widetilde{p}(\widehat{\mathbf{v}}|\mathcal{D})$ evaluated at the mode $\widehat{\mathbf{v}}$. 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. 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:

$$p(x) = \frac{2}{\varsigma}\varphi\left(\frac{x-\mu}{\varsigma}\right)\ \Phi\left(\rho\frac{(x-\mu)}{\varsigma}\right), \tag{8}$$

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 $(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$ $(j = 1, ..., q)$ with a skew-normal distribution by matching its first three empirical moments with the theoretical ones for the density in (8), where $\widehat{\mathbf{v}}_{-j}$ denotes the vector $\widehat{\mathbf{v}}$ without the jth entry. Appendix B shows the derivations to obtain μ^*, ς^* and ρ^* in the approximating skew-normal distribution $\mathrm{SN}_j(\mu^*, \varsigma^{*2}, \rho^*)$ through moment matching.

Once a skew-normal distribution $\mathrm{SN}_j(\mu^*, \varsigma^{*2}, \rho^*)$ has been adjusted to the conditional $\widetilde{p}(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$, 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 $\mathrm{SN}_{j,0.025}$ and $\mathrm{SN}_{j,0.975}$ 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 $R(\mathbf{v}) = \widetilde{p}(\mathbf{v}|\mathcal{D})/\widetilde{p}(\widehat{\mathbf{v}}|\mathcal{D})$ 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. Figure 1 highlights the skew-normal match and the final grid in an example with $q = 2$ nonlinear smooth functions in the additive predictor and data generated from a Poisson response with sample size $n = 250$.

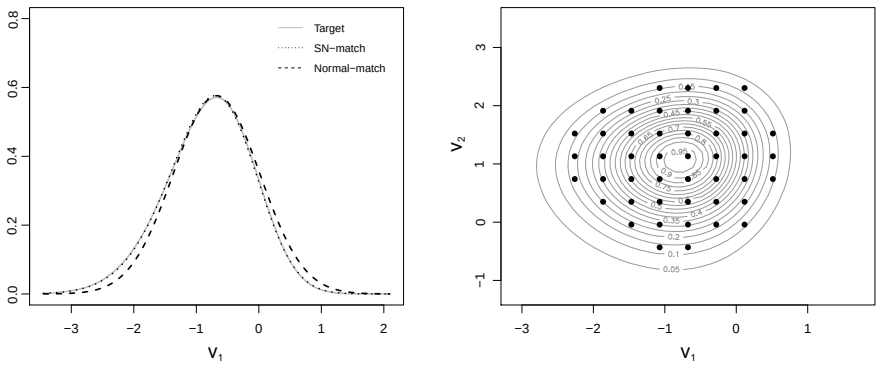

Figure 1: Left: Skew-normal fit (dotted) and naive Gaussian match (dashed) to the conditional $\widetilde{p}(v_1|\widehat{v}_2, \mathcal{D})$ (gray). The skew-normal fit is closer to the target and captures the lack of symmetry present in the target. Right: Final grid construction to explore $\widetilde{p}(\mathbf{v}|\mathcal{D})$.

When the number of smooth functions q in the additive model is above a certain threshold (say $q > 4$), the preceding computational strategy becomes too demanding as the number of quadrature points (following from the Cartesian product of the grid points for each penalty parameter $\exp(v_j)$ $(j = 1, \ldots, q)$) explodes. A cost-effective alternative relies on MCMC to sample values from the posterior $\widetilde{p}(\mathbf{v}|\mathcal{D})$. More thoroughly, an independence sampler is implemented using a multivariate Student-t proposal distribution $t_\vartheta(\widehat{\mathbf{v}}, (-\mathcal{H}^*)^{-1})$ with density $h(\mathbf{v}|\widehat{\mathbf{v}})$, degrees of freedom ($\vartheta = 3$, say), a mean set at the posterior mode $\widehat{\mathbf{v}}$, and variance-covariance matrix $(\vartheta/(\vartheta - 2))(-\mathcal{H}^*)^{-1}$.

Algorithm 1 summarizes the strategy to explore $\widetilde{p}(\mathbf{v}|\mathcal{D})$. When $q \leq 4$, a grid is constructed using a Cartesian product of marginal grids delimited by quantiles of approximating skew-normal densities. Exploration in larger dimensions relies on the independence Metropolis-Hastings sampler. This algorithm is used in the next section to approximate the marginal posterior of the vector of regression and spline coefficients.

Algorithm 1: Exploration of $\widetilde{p}(\mathbf{v}|\mathcal{D})$

If $q \leq 4$ **do** (Grid strategy, cf. Section 2.5.1)
  for $j = 1, \ldots, q$ **do**
    Compute the skew-normal match $\text{SN}_j(\mu^*, \varsigma^{*2}, \rho^*)$ to $\widetilde{p}(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$.
    Construct a Cartesian grid $\{v_{jm}\}_{m=1}^M$ from $\text{SN}_{j,0.025}$ to $\text{SN}_{j,0.975}$.
  end for
  Compute the Cartesian product of the univariate grids $\mathcal{C} = \times_{j=1}^q \{v_{jm}\}_{m=1}^M$.
  Choose α and keep the $\widetilde{M}$ values in $\mathcal{C}$ such that $R(\mathbf{v}) \geq \exp\left(-.5\chi^2_{q;1-\alpha}\right)$.
else do (Independence sampling, cf. Section 2.5.2)
  Choose an initial value $\mathbf{v}^{(0)} = \widehat{\mathbf{v}}$.
  for $m = 1, \ldots, \widetilde{M}$ **do**
    Generate $\mathbf{v}^{(\text{prop})} \sim h(\mathbf{v}|\widehat{\mathbf{v}})$.
    Compute the acceptance probability $\alpha = \min\left(1, \frac{\widetilde{p}\left(\mathbf{v}^{(\text{prop})}|\mathcal{D}\right) h\left(\mathbf{v}^{(m-1)}|\widehat{\mathbf{v}}\right)}{\widetilde{p}\left(\mathbf{v}^{(m-1)}|\mathcal{D}\right) h\left(\mathbf{v}^{(\text{prop})}|\widehat{\mathbf{v}}\right)}\right)$.
    Draw $u \sim \mathcal{U}(0, 1)$.
    If $u \leq \alpha$, set $\mathbf{v}^{(m)} = \mathbf{v}^{(\text{prop})}$, else set $\mathbf{v}^{(m)} = \mathbf{v}^{(m-1)}$.
  end for

3.3. *Approximate marginal posterior of the latent vector*

Using the Laplace approximation discussed in Section 2.2, the posterior of the latent vector $\boldsymbol{\xi}$ can be obtained as follows:

$$\begin{aligned} p(\boldsymbol{\xi}|\mathcal{D}) &= \int_{\mathbb{R}^q_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})\; p(\boldsymbol{\lambda}|\mathcal{D})\; d\boldsymbol{\lambda} \\ &\approx \int_{\mathbb{R}^q_{++}} \widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})\; \widetilde{p}(\boldsymbol{\lambda}|\mathcal{D})\; d\boldsymbol{\lambda} \\ &\approx \int_{\mathbb{R}^q} \widetilde{p}_G(\boldsymbol{\xi}|\exp(\mathbf{v}),\mathcal{D})\; \widetilde{p}(\mathbf{v}|\mathcal{D})\; d\mathbf{v}, \end{aligned} \tag{9}$$

where the last line follows from the change of variable in log-scale. Using Algorithm 1, we get a set of quadrature points $\{\mathbf{v}^{(m)}\}_{m=1}^{\widetilde{M}}$. Defining:

$$\omega_m = \frac{\widetilde{p}(\mathbf{v}^{(m)}|\mathcal{D})}{\sum_{m=1}^{\widetilde{M}} \widetilde{p}(\mathbf{v}^{(m)}|\mathcal{D})}, \quad m = 1,\ldots,\widetilde{M}, \tag{10}$$

when $q \leq 4$ and $\omega_m = 1/\widetilde{M}$ otherwise, Equation (9) suggests to approximate $p(\boldsymbol{\xi}|\mathcal{D})$ by:

$$\widetilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m\; \mathcal{N}_{\dim(\boldsymbol{\xi})}\left(\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \widehat{\Sigma}_{\mathbf{v}^{(m)}}\right), \tag{11}$$

where $\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}} = \left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}^{(m)}}\right)^{-1}\widetilde{\boldsymbol{\varpi}}$ and $\widehat{\Sigma}_{\mathbf{v}^{(m)}} = \left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}^{(m)}}\right)^{-1}$ are the conditional posterior mode and variance-covariance matrix resulting from the iterative Laplace approximations proposed in Section 2.2. Note that the computational cost of reevaluating the conditional posterior mode and variance-covariance for each penalty $\exp(\mathbf{v}^{(m)})$ in the grid can be reduced by adding an extra layer of approximation by replacing $\widetilde{W}$ in the Newton-Raphson procedure by its value $\widetilde{W}_{\widehat{\mathbf{v}}}$ at the posterior mode. A point estimate for the latent vector is given by the posterior mean of (11), which is a mixture of the location components, i.e. $\widehat{\boldsymbol{\xi}} = \sum_{m=1}^{\widetilde{M}} \omega_m\; \widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}$.

Approximate pointwise credible intervals for latent elements $\xi_h,\ h = 1,\ldots,\dim(\boldsymbol{\xi})$ can be straightforwardly obtained by starting from the finite mixture given in (11). The approximate posterior for the hth latent element is $\widetilde{p}(\xi_h|\mathcal{D}) =$

$\sum_{m=1}^{\widetilde{M}} \omega_m \, \mathcal{N}_1\Big(\widehat{\xi}_{h,\mathbf{v}^{(m)}}, \widehat{\Sigma}_{hh,\mathbf{v}^{(m)}}\Big)$, where $\widehat{\xi}_{h,\mathbf{v}^{(m)}}$ is the hth entry of vector $\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}$ and $\widehat{\Sigma}_{hh,\mathbf{v}^{(m)}}$ is the hth entry on the diagonal of matrix $\widehat{\Sigma}_{\mathbf{v}^{(m)}}$. The latter expression can be used to construct a $(1-\alpha)\times 100\%$ 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 $\boldsymbol{\xi}$ corresponding to the spline vector $\boldsymbol{\theta_j} = (\theta_{j1}, \ldots, \theta_{jK-1})^T$. Also, denote by $\widetilde{\mathbf{b}}_l$ the vector of B-splines in the basis evaluated at x_l. The function f_j at point x_l is thus modeled as $f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j}) = \widetilde{\mathbf{b}}_l^T \boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ and from (11) the posterior of $\boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ is approximated by the finite mixture:

$$\widetilde{p}(\boldsymbol{\xi}_{\boldsymbol{\theta}_j}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \, \mathcal{N}_{K-1}\left(\widehat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}, \widehat{\Sigma}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}\right), \tag{12}$$

where $\widehat{\Sigma}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}$ is a submatrix of $\widehat{\Sigma}_{\mathbf{v}^{(m)}}$ 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 posterior $p(f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j})|\mathcal{D})$ is to use a mixture of univariate normals:

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

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

4. Simulations

The performance of the LPS approach (with cubic B-splines and a third order penalty) is assessed through different simulation scenarios and compared with results obtained using the `gam()` function from the **mgcv** package in **R** (Wood, 2017), a popular and established toolkit for estimating GAMs. Options of the `gam()` function are carefully chosen so that the generated results can be meaningfully compared to these obtained using our LPS approach. In particular, both approaches share the same dimension and order for the B-spline

basis, as well as the same order for the difference penalty. The restricted maximum likelihood (REML) method is used by the `gam()` routine for selecting the penalty parameters $\boldsymbol{\lambda}$. 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 & Wood, 2011; Wood et al., 2013). Newton's method is used to numerically optimize the REML smoothing parameter estimation criterion. A detailed description of the `gam()` estimation procedure can be found in Wood (2011, 2017).

4.1. Estimation of linear parameters and additive functional components

The simulation setting entails $S = 500$ replications of a data set 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 $z_{i3} \sim \mathcal{N}(0,1)$. The full model is $g(\mu_i) = -1.50 + 0.70z_{i1} - 0.80z_{i2} + 0.40z_{i3} + f_1(x_{i1}) + f_2(x_{i2}) + f_3(x_{i3})$, for $i = 1, \ldots, n$ with regression coefficients $\beta_0 = -1.50$, $\beta_1 = 0.70$, $\beta_2 = -0.80$, $\beta_3 = 0.40$ and smooth additive terms $f_1(x_{i1}) = -4x_{i1}^6 + 2x_{i1}^2 + \cos(2\pi x_{i1}) - 0.1$, $f_2(x_{i2}) = 3x_{i2}^5 + 2\sin(4x_{i2}) + 1.5x_{i2}^2 - 0.5$, and $f_3(x_{i3}) = \sin(3\pi x_{i3})$. The covariates for the smooth functions are independent draws from the Uniform distribution on the domain $[-1, 1]$. The above functions are specified as 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. Four scenarios are considered for the response variable, namely (I) Generation from a Poisson distribution $y_i \sim \text{Poisson}(\mu_i)$, with $\mu_i = \exp(\varrho_i)$ to illustrate the case of count data, (II) Generation from a Gaussian $y_i \sim \mathcal{N}(\mu_i, \sigma^2 = 0.3)$, with $\mu_i = \varrho_i$, (III) Generation from a Binomial $y_i \sim \text{Bin}(15, p_i)$ and (IV) Generation from a Bernoulli $y_i \sim \text{Bern}(p_i)$ to illustrate the case of binary responses with success probability $p_i = \exp(\varrho_i)/(1 + \exp(\varrho_i))$ for Binomial and Bernoulli cases.

Table 1 shows the simulation results and comparisons with the `gam()` function. For all the considered data types, the Laplace-P-spline approach exhibits non-significant biases and the estimated coverage probabilities are consistent with their nominal level. Also, the ESE and RMSE show a behavior comparable to what is observed with the `gam()` output. For the Bernoulli scenario, ESEs are smaller with LPS, but biases are slightly larger than with `gam()`. The frequentist coverage of credible intervals remain compatible whatever the method used.

Data	Parameters	Bias	$CP_{90\%}$	$CP_{95\%}$	ESE	RMSE
Poisson	$\beta_1 = 0.70$	0.001 (0.003)	87.4 (88.2)	94.0 (94.6)	0.122 (0.122)	0.122 (0.121)
	$\beta_2 = -0.80$	0.006 (0.003)	91.0 (90.8)	95.8 (95.6)	0.061 (0.061)	0.062 (0.061)
	$\beta_3 = 0.40$	-0.001 (0.000)	90.0 (90.0)	95.8 (96.4)	0.060 (0.060)	0.060 (0.059)
Normal	$\beta_1 = 0.70$	0.001 (0.001)	90.6 (90.0)	96.4 (96.4)	0.065 (0.065)	0.065 (0.065)
	$\beta_2 = -0.80$	-0.001 (-0.001)	89.0 (89.4)	94.8 (95.0)	0.033 (0.033)	0.033 (0.033)
	$\beta_3 = 0.40$	0.000 (0.000)	89.6 (90.2)	94.8 (95.2)	0.034 (0.034)	0.033 (0.034)
Binomial	$\beta_1 = 0.70$	0.004 (0.006)	89.8 (90.8)	94.8 (95.0)	0.090 (0.090)	0.090 (0.091)
	$\beta_2 = -0.80$	0.011 (0.008)	88.8 (88.6)	93.6 (94.2)	0.047 (0.048)	0.049 (0.048)
	$\beta_3 = 0.40$	-0.003 (-0.001)	92.6 (92.6)	96.4 (96.8)	0.042 (0.042)	0.042 (0.042)
Bernoulli	$\beta_1 = 0.70$	-0.077 (-0.008)	87.4 (87.8)	93.0 (93.0)	0.320 (0.349)	0.329 (0.349)
	$\beta_2 = -0.80$	0.082 (0.005)	87.6 (91.8)	93.0 (96.4)	0.155 (0.175)	0.175 (0.174)
	$\beta_3 = 0.40$	-0.038 (0.003)	88.6 (89.8)	93.2 (94.0)	0.159 (0.176)	0.163 (0.176)

Table 1: Simulation results with the LPS method for $S = 500$ replicates of sample size $n = 300$ for different types of response (Poisson, Normal, Binomial and Bernoulli). The values in parentheses are estimation results from the `gam()` function.

Coverage properties of approximate 90% pointwise credible intervals for the additive terms f_1, f_2 and f_3 are reported in Table 2 for selected values of the covariate on $[-1, 1]$. An asterisk superscript is added to the estimated coverage to point a statistically significant difference with the nominal value. Results of the `gam()` function are labeled "MGCV". In addition to the LPS approach, Table 2 also highlights the coverage performance of LPSMAP, where each penalty parameter is replaced by its posterior mode $\widehat{\boldsymbol{\lambda}} = \exp(\widehat{\mathbf{v}})$ in our Laplace-P-spline method. For LPSMAP the uncertainty in the selection of $\boldsymbol{\lambda}$ is ignored (like in

Wood's approach), such that the mixture in Equation (12) is omitted and the point estimate of the latent vector and its associated variance-covariance matrix become $\widehat{\boldsymbol{\xi}}_{\widehat{\mathbf{v}}} = \left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\widehat{\mathbf{v}}}\right)^{-1}\widetilde{\boldsymbol{\varpi}}$ and $\widehat{\Sigma}_{\widehat{\mathbf{v}}} = \left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\widehat{\mathbf{v}}}\right)^{-1}$ respectively. With LPSMAP, an approximate $(1-\alpha)\times 100\%$ credible interval for function f_j at point x_l is computed from a frequentist perspective, $\widehat{f}_j(x_l) \pm z_{\alpha/2}\sqrt{\widetilde{\mathbf{b}}_l^T\widehat{\Sigma}_{\boldsymbol{\theta}_j,\widehat{\mathbf{v}}}\widetilde{\mathbf{b}}_l}$.

As can be seen from Table 2, the LPS and LPSMAP methods perform well in the Poisson, Normal and Binomial scenarios as estimated frequentist coverage probabilities are close to the nominal level at almost all selected covariate values. The `gam()` results also show a similar performance across all scenarios. Comparing LPS and LPSMAP, we observe that omitting the penalty uncertainty globally translates into a slight decrease in percentage points for the estimated coverage probability. Yet, the LPSMAP approach still exhibits close to nominal coverage for all the functions. In terms of computational speed, the LPSMAP approach is approximately four times faster than the LPS approach and five times slower than `gam()` (≈ 0.05 seconds vs 0.26 seconds).

In the Bernoulli setting where the information content for a given sample size is much smaller than under the other simulation scenarios, all the considered methods exhibit effective frequentist coverages below the nominal value as illustrated in Table 3 with $n = 300$. It corresponds to situations where the estimates of the additive terms provided by LPS(MAP) or `gam()` can be inaccurate. The pronounced undercoverage in this setting is explained by the poor information conveyed by a binary random variable that translates into oversmoothing of the additive functional components as highlighted in Figure 2. However, as expected, increasing the sample size in the Bernoulli scenario yields frequentist coverage probabilities close to their nominal value (cf. Table 3 with $n = 2000$) both for the LPS(MAP) and `gam()` methods.

The effective frequentist coverages of 90%, 95% and 99% pointwise credible intervals averaged over 200 uniformly distributed values of the covariate on $[-1, 1]$ and $S = 500$ dataset replications in the Poisson, Normal and Binomial settings are reported in Appendix C. Again, the LPS and LPSMAP methodologies display estimated coverages close to their nominal value in all scenarios.

Note that `gam()` and LPSMAP rely on a similar approach for selecting the optimal posterior penalty value. Hence, the simulation results suggest that our penalty selection scheme is at least as efficient as what is implemented in `gam()` for estimating the smooth components in the additive part of the model.

Data	f Method	-0.95	-0.70	-0.50	-0.20	0.00	0.20	0.50	0.70	0.95
Poisson	f_1 LPS	86.0*	89.8	91.6	91.2	88.2	91.4	87.0	88.4	87.6
	f_1 LPSMAP	85.8*	89.2	89.6	90.8	88.2	91.4	86.0*	87.6	87.0
	f_1 MGCV	87.8	91.6	92.0	90.6	90.6	92.0	89.4	92.2	89.0
	f_2 LPS	93.2	82.8*	89.2	84.4*	91.2	89.2	86.2*	92.6	87.4
	f_2 LPSMAP	92.4	81.4*	87.4	81.4*	90.2	89.0	85.2*	92.4	86.8
	f_2 MGCV	92.6	87.6	90.8	89.8	92.4	91.0	89.8	92.2	89.0
	f_3 LPS	89.8	87.8	87.2	88.6	90.2	86.2*	86.8	90.4	90.6
	f_3 LPSMAP	88.8	87.2	86.0*	87.6	90.2	86.0*	86.0*	89.2	90.6
	f_3 MGCV	90.4	88.6	90.8	90.6	91.2	88.4	88.6	91.8	91.0
Normal	f_1 LPS	90.2	92.8	92.0	91.0	91.6	92.4	92.4	92.6	90.2
	f_1 LPSMAP	90.0	92.2	91.6	91.0	91.6	92.0	91.6	92.6	89.8
	f_1 MGCV	90.4	92.8	91.4	91.4	91.8	91.6	92.4	92.0	90.4
	f_2 LPS	91.6	90.4	91.2	94.8*	92.2	93.6*	91.2	90.0	89.4
	f_2 LPSMAP	91.2	89.4	90.0	94.6*	91.6	94.0*	90.8	90.0	89.2
	f_2 MGCV	92.0	90.4	90.8	94.4*	92.0	93.8*	92.0	91.2	89.6
	f_3 LPS	90.4	92.0	90.6	92.4	90.8	87.4	89.4	92.6	89.6
	f_3 LPSMAP	90.4	92.2	90.4	92.2	90.6	88.0	89.0	92.4	89.2
	f_3 MGCV	89.8	92.4	91.8	91.6	90.0	88.8	89.8	92.4	89.6
Binomial	f_1 LPS	88.4	94.0*	89.2	93.0	91.0	96.0*	91.6	90.8	88.2
	f_1 LPSMAP	87.6	93.0	87.6	92.8	90.6	96.0*	91.4	91.0	88.0
	f_1 MGCV	88.6	93.8*	89.4	93.4*	90.6	96.2*	93.2	91.4	89.0
	f_2 LPS	89.8	92.6	86.8	90.8	93.6*	92.8	86.8	92.0	84.2*
	f_2 LPSMAP	89.2	91.8	85.4*	90.2	93.6*	92.2	86.8	91.0	83.8*
	f_2 MGCV	90.0	94.4*	87.6	92.2	93.8*	92.4	90.4	91.6	86.8
	f_3 LPS	87.8	91.0	87.8	90.6	90.6	86.8	87.4	92.4	90.4
	f_3 LPSMAP	87.6	90.6	87.2	89.8	90.6	86.6	86.2*	92.2	90.0
	f_3 MGCV	88.6	91.0	89.4	91.8	89.8	89.4	89.4	92.6	90.6

Table 2: Effective frequentist coverages of 90% pointwise credible intervals for the functions f_1, f_2, f_3 at selected domain points for Poisson, Normal and Binomial data over $S = 500$ replications of sample size $n = 300$ for the Laplace-P-spline (LPS), the LPS omitting the mixture (LPSMAP) and `gam()` (MGCV) methods. An asterisk points a statistically significant difference with the nominal value.

Data	f	Method	-0.95	-0.70	-0.50	-0.20	0.00	0.20	0.50	0.70	0.95
Bernoulli (n=300)	f_1	LPS	85.4*	78.0*	0.6*	35.0*	1.4*	47.0*	1.0*	84.0*	82.2*
	f_1	LPSMAP	86.2*	78.2*	0.6*	25.6*	0.6*	46.0*	0.4*	84.6*	82.2*
	f_1	MGCV	84.8*	77.6*	42.0*	76.4*	38.2*	77.4*	42.0*	82.2*	85.2*
	f_2	LPS	86.8	82.6*	62.0*	34.4*	86.6	52.4*	58.6*	89.6	73.0*
	f_2	LPSMAP	83.2*	72.8*	60.6*	26.8*	84.2*	42.6*	58.0*	84.8*	66.6*
	f_2	MGCV	87.8	77.0*	84.8*	66.0*	90.0	72.2*	83.8*	79.6*	83.2*
	f_3	LPS	88.0	80.4*	2.6*	1.2*	96.0*	1.2*	2.2*	71.0*	77.8*
	f_3	LPSMAP	87.6	82.0*	2.2*	1.2*	92.8	1.2*	1.8*	65.0*	62.6*
	f_3	MGCV	87.4	84.2*	52.0*	51.0*	90.0	48.8*	49.0*	83.6*	86.8
Bernoulli (n=2000)	f_1	LPS	90.0	89.8	87.4	94.2*	87.4	91.8	87.6	89.8	86.6
	f_1	LPSMAP	89.4	90.2	87.0	94.0*	87.6	92.0	86.8	88.6	86.6
	f_1	MGCV	89.8	91.2	90.6	93.2	90.8	91.6	90.6	89.2	87.8
	f_2	LPS	88.8	90.8	87.0	89.8	93.0	90.8	86.6	91.2	86.8
	f_2	LPSMAP	87.6	90.6	86.2*	89.0	92.6	90.6	86.6	90.4	86.6
	f_2	MGCV	89.2	91.8	88.8	90.6	93.2	91.4	90.0	90.6	91.2
	f_3	LPS	90.2	88.2	86.0*	87.6	93.2	84.8*	84.4*	89.2	91.2
	f_3	LPSMAP	90.4	87.8	84.8*	87.2	93.0	83.8*	83.0*	89.2	90.6
	f_3	MGCV	90.8	88.6	89.6	91.4	92.2	88.6	87.0	90.2	91.2

Table 3: Effective frequentist coverages of 90% pointwise credible intervals for the functions f_1, f_2, f_3 at selected domain points for Bernoulli data over $S = 500$ replications of sample size $n = 300$ and $n = 2000$ for the Laplace-P-spline (LPS), the LPS omitting the mixture (LPSMAP) and `gam()` (MGCV) methods. An asterisk points a statistically significant difference with the nominal value.

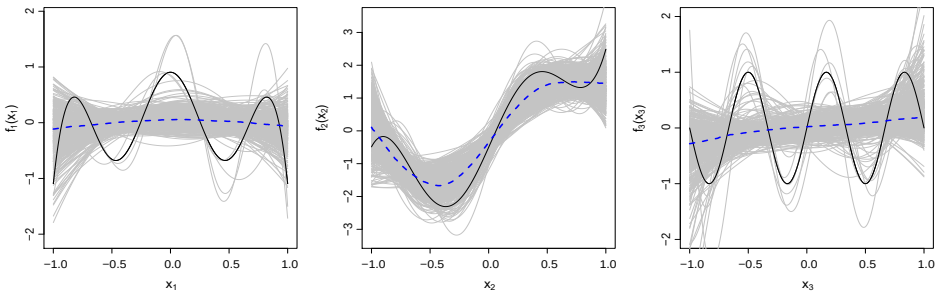

Figure 2: Estimation of smooth additive terms (gray curves) for $S = 500$ dataset replications of size $n = 300$ in the Bernoulli scenario with LPS. The dashed line is the pointwise median of the gray curves and the black curves are the target functions.

The simulation results confirm the attractiveness of the Laplace-P-spline model for pointwise and set estimation of the regression parameters in the linear part as well as of the smooth additive components. To enhance the estimation accuracy of our approach in the case of extremely discrete responses such as, for example, Bernoulli data, a possibility is to improve the approximation to the conditional posterior $\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D})$ by correcting for location and skewness as suggested in Rue et al. (2009). Beyond such extreme binary data configurations, the simple Laplace approximation underlying LPS and LPSMAP suffices for precise inference.

To complete the simulation study, we compare the LPSMAP methodology against BayesX (Umlauf et al., 2015), a fully Bayesian contender that can be used to fit structured additive regression models with MCMC. In particular, we use the **R2BayesX** package and fit the GAM in the Poisson scenario with the `bayesx()` routine using a chain of size $10,000$ and a burn-in of size $1,000$. Cubic B-spline bases are used to model the smooth terms with a second order penalty. In Table 4, we report the estimated 95% coverage of credible intervals for the smooth additive components of the model on selected points in the interval $[-1, 1]$ for $S = 200$ replicates with sample size $n = 300$.

Data	f	Method	-0.95	-0.70	-0.50	-0.20	0.00	0.20	0.50	0.70	0.95
Poisson	f_1	LPSMAP	89.0*	96.0	94.5	97.0	91.0	97.0	93.5	96.5	91.5
	f_1	BAYESX	94.0	98.5	91.5	95.5	94.5	94.5	94.0	95.5	88.5*
	f_2	LPSMAP	95.5	96.5	94.5	91.0	97.0	92.0	93.5	98.0	92.5
	f_2	BAYESX	93.0	94.0	95.5	93.5	96.5	92.5	91.0	94.0	84.5*
	f_3	LPSMAP	92.5	96.0	92.5	94.5	96.0	95.0	95.5	97.0	92.5
	f_3	BAYESX	93.0	97.5	94.5	93.5	96.5	94.0	95.5	96.5	95.5

Table 4: Effective frequentist coverages of 95% pointwise credible intervals for the functions f_1, f_2, f_3 at selected domain points for Poisson data over $S = 200$ replications of sample size $n = 300$ for LPSMAP and BayesX. An asterisk points a statistically significant difference with the nominal value.

The estimated frequentist coverage probabilities are close to the 95% nominal level for both methods. There is however a notable difference in terms of computational cost for model fitting. While the routines underlying BayesX take on average 6.53 seconds to fit the GAM for each dataset, the LPSMAP

methodology requires only 0.26 seconds (on average) for the fit. In other words, LPSMAP is approximately 25 times faster than BayesX while maintaining the same coverage performance for credible intervals on the smooth terms. With more additive terms ($q = 6$), the computational gain is maintained and we measured that LPSMAP is faster than BayesX by a factor of (approximately) 7. When q increases, most of the computational budget underlying LPSMAP to fit the GAM is dedicated to the Newton-Raphson algorithm to compute the posterior mode $\widehat{\mathbf{v}}$. Coding that optimization part in $\mathbf{C}_{++}$ (the language underlying BayesX) would further improve the speed of LPSMAP.

4.2. Computational costs

A notable feature of the Laplace-P-spline methodology is its low computational cost despite being fully Bayesian. In fact, our algorithm (underlying a fully Bayesian approach) is purely written in **R** (without any parallelization) and takes approximately 0.26 seconds per dataset in the above scenario as compared to 0.05 seconds for the `gam()` function (coding an empirical Bayes approach) for simulations performed on a machine equipped with an Intel Xeon E-2186M CPU running at a clock speed of 2.90 GHz. Considering that the `gam()` algorithm is neither fully Bayesian nor entirely written in **R** (as most of the script relies on **C** code which is much faster), the Laplace-P-spline toolkit can be considered a serious competitor for approximate full Bayesian inference in GAMs when smooth functions are modeled with P-splines. To illustrate the computational behavior of LPS and LPSMAP against sample size for fixed dimension $q = 3$, we consider an increasing sequence of sample sizes from $n = 200$ to $n = 3000$ in steps of 200 and for each considered sample size compute the average wall clock time (elapsed real time) in seconds with the `proc.time()` function in **R** over 10 different samples. In Figure 3 (a) the elapsed time to estimate the GAM model with LPS and LPSMAP is plotted against sample size to depict the involved computational resources. Both curves show a linear increase with sample size. LPSMAP is faster than LPS as it does not require a grid construction to explore the support of the marginal posterior of the penalty parameters, but rather fix

them at their posterior mode. Figure 3 (b) highlights the computational time of LPS(MAP) against sample size n on a log scale.

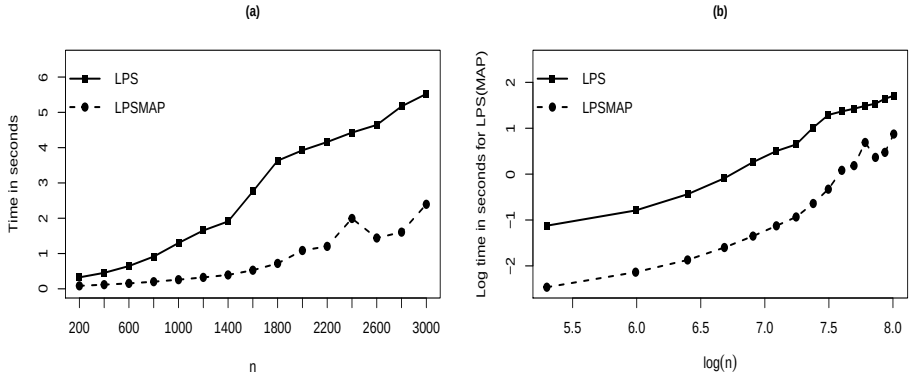

Figure 3: (a) Real elapsed time in seconds as a function of sample size for LPS and LPSMAP. (b) Log of computational time (in seconds) of LPS(MAP) against log sample size.

4.3. Simulation study with more additive terms.

A large number q of smooth functions in the additive predictor implies an increased computational burden. Algorithm 1 suggests to prefer independence sampling over a grid construction to explore the marginal posterior of the penalty parameters when $q > 4$, see Section 3.2 for details. To illustrate how the Laplace-P-spline model performs with a larger number of smooth functions, we simulate $S = 500$ datasets of size $n = 300$ and a Markov chain sample of size 500 for each replicate with the following additive terms:

$$\begin{aligned} f_1(x_1) &= 0.5(2x_1^5 + 3x_1^2 + \cos(3\pi x_1) - 1), \\ f_2(x_2) &= 1.3x_2^5 + \sin(4x_2) + 0.75x_2^2 - 0.25, \\ f_3(x_3) &= \sin(4\pi x_3), \\ f_4(x_4) &= \exp(-x_4^3)\sin(2\pi x_4^2) - 0.1, \end{aligned}$$

$$
\begin{aligned}
f_5(x_5) &= 0.8x_5^2(x_5^3 + 2\exp(-3x_5^4 + \log(2x_5 + \pi))) - 0.65, \\
f_6(x_6) &= 1.5\ \big(0.1\sin(2\pi x_6) + 0.2\cos(2\pi x_6) + 0.3\sin^2(2\pi x_6) \\
&\quad +0.4\cos^3(2\pi x_6) + 0.5\sin^3(2\pi x_6)\big) - 0.22.
\end{aligned}
$$

There are three additional covariates specified as in Section 4.1 with regression coefficients $\beta_0 = -1.20$, $\beta_1 = 0.50$, $\beta_2 = -0.40$ and $\beta_3 = 0.70$. The covariates of the smooth functions are drawn independently from the Uniform distribution on the domain $[-1, 1]$. Each smooth function is modeled using a linear combination of 15 cubic B-splines associated to equidistant knots on $[-1, 1]$ and a third order penalty to control smoothness. Two scenarios are considered for the generating process of the response, namely (1) a Gaussian model $y_i \sim \mathcal{N}(\mu_i, \sigma^2 = 0.5)$ and (2) a Binomial model $y_i \sim \text{Bin}(20, p_i)$, with p_i the success probability and a logit link function. Table 5 shows the simulation results of the Laplace-P-spline approach combined with MCMC (cf. Section 3.2). The estimation results obtained with the `gam()` function from the **mgcv** package are shown in parenthesis.

Estimated biases shown in Table 5 are almost similar for the two different approaches and nearly equal to zero in the considered data scenarios. In addition, the reported coverage probabilities are close to their corresponding nominal value and analogous results appear for the ESE and RMSE with the LPS and `gam()` algorithms. Figure 4 illustrates the estimation results for the six additive smooth terms with the proposed Laplace-P-spline methodology in the Binomial case. For each graph, there are $S = 500$ gray curves representing the estimates of the corresponding unknown smooth function (black) entering the additive predictor. The dashed curve represents the pointwise median of the 500 estimated curves. For each smooth term, the observed estimates are close to the target, even with highly oscillating functions (e.g. f_3 and f_6). For function f_6, small bumps arising near main curvatures can be better captured by increasing the number of B-splines in the basis.

Data	Parameters	Bias	$CP_{90\%}$	$CP_{95\%}$	ESE	RMSE
Normal	$\beta_1 = 0.50$	0.001 (0.001)	87.8 (87.4)	94.0 (94.6)	0.096 (0.095)	0.096 (0.095)
	$\beta_2 = -0.40$	0.003 (0.003)	86.8 (87.4)	94.8 (95.0)	0.047 (0.047)	0.047 (0.047)
	$\beta_3 = 0.70$	0.003 (0.003)	86.2 (86.8)	93.2 (92.2)	0.049 (0.049)	0.049 (0.049)
Binomial	$\beta_1 = 0.50$	-0.007 (-0.003)	89.6 (89.6)	93.4 (94.0)	0.078 (0.078)	0.079 (0.078)
	$\beta_2 = -0.40$	0.003 (0.000)	88.8 (89.6)	94.4 (94.4)	0.041 (0.041)	0.041 (0.041)
	$\beta_3 = 0.70$	-0.009 (-0.003)	87.8 (88.2)	94.2 (95.0)	0.043 (0.043)	0.044 (0.043)

Table 5: Simulation results for $S = 500$ replicates of sample size $n = 300$ for Normal and Binomial data when independence sampling is used to draw samples from $\widetilde{p}(\mathbf{v}|\mathcal{D})$. The values in parentheses are estimation results from the `gam()` function.

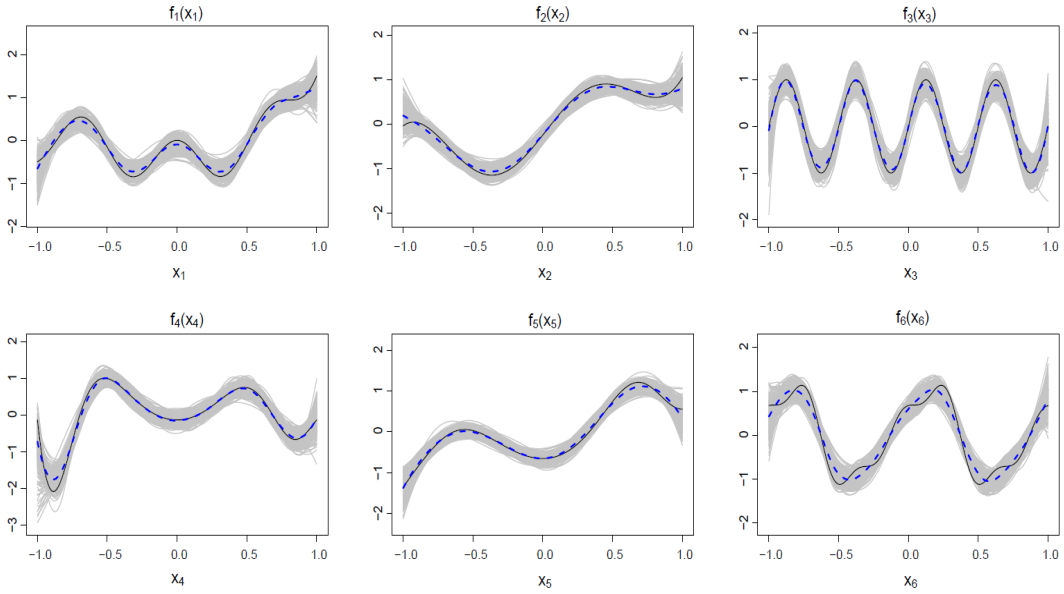

Figure 4: Estimation of smooth additive terms $f_1, \ldots, f_6$ (gray curves) for $S = 500$ dataset replications of size $n = 300$ in the Binomial scenario. The dashed line is the pointwise median of the gray curves.

With $q = 6$, our LPS methodology coupled with MCMC (LPS-MCMC) requires (to build a chain of length 500) on average 4.70 seconds for a dataset of size $n = 300$. In Table 6, we provide computation times of the LPS-MCMC algorithm to estimate the GAM for different dimensions q and sample sizes. As expected the computation time increases with q and n. Figure 5 gives an overview of the

average computational times required to estimate the GAM with the LPS and LPS-MCMC algorithms for an increasing number of additive terms. When $q \leq 4$ the LPS approach is faster, but in larger dimensions the LPS-MCMC algorithm (with an independence sample of length 500) requires less computational budget than the grid construction in LPS.

Dimension	Average computation time (in seconds)		
	$n = 300$	$n = 1000$	$n = 3000$
$q = 1$	1.86	2.78	7.00
$q = 2$	2.10	3.46	11.60
$q = 3$	2.51	4.66	15.09
$q = 4$	3.04	6.53	21.04
$q = 5$	3.82	8.83	27.55
$q = 6$	4.70	11.46	36.08

Table 6: Average computation time (in seconds) of the LPS-MCMC algorithm over $S = 20$ samples of size $n \in \{300, 1000, 3000\}$ for different dimensions $q \in \{1, 2, 3, 4, 5, 6\}$.

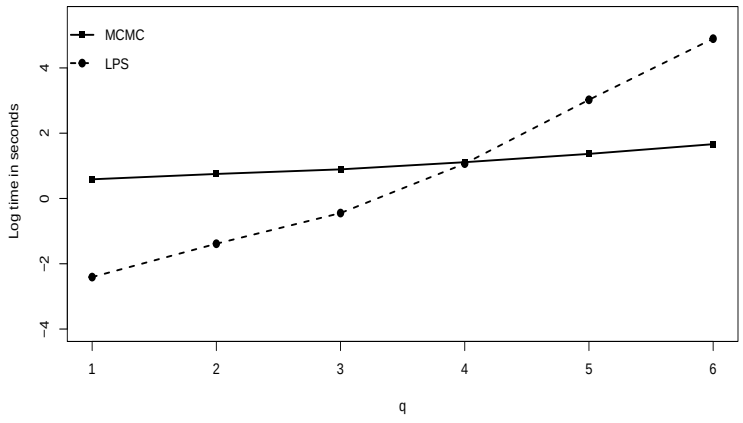

Figure 5: Logarithm of the average computation time (in seconds) of LPS (dahsed) and LPS-MCMC (solid) over $S = 20$ samples of size $n = 300$ and dimensions $q \in \{1, 2, 3, 4, 5, 6\}$.

5. Applications

5.1. Model for the number of doctor visits

We apply our Laplace-P-spline model in the context of a health-care study on Medicaid eligibles. The data are from the 1986 Medicaid Consumer Survey sponsored by the Health Care Financing Administration in the USA. This Medicaid database has first been studied by Gurmu (1997) in the framework of a semi-parametric hurdle model and later by Sapra (2013) as an econometric application of generalized additive models using the **mgcv** package in **R**. Our analysis will focus on a sample of $n = 485$ adults who meet the requirement for eligibility in the Aid to Families with Dependent Children (AFDC) program. The response variable is the number of doctor visits (office/clinic and health center) over a period of 120 days. The explanatory variables included in the linear part of the GAM are *Children* (Total number of children in the household), *Race* (0=other; 1=white) and *Maritalstatus* (0=other; 1=married). The variables modeled in the smooth nonlinear part are taken to be *Age*, the household annual *Income* (in US dollars), a variable measuring the ease of *Access* to health services with values in the interval (0=low access; 100=high access) and the first principal component built from three health-status variables (functional limitations, acute conditions, chronic conditions) denoted by *PC1* with larger positive numbers meaning poorer health. Descriptive statistics of these variables are detailed in Gurmu (1997). The GAM model with a Poisson conditional distribution Poisson(μ_i) ($i = 1, ..., n$) for the number of doctor visits can be written as follows:

$$\begin{aligned} g(\mu_i) &= \beta_0 + \beta_1 \mathit{Children}_i + \beta_2 \mathit{Race}_i + \beta_3 \mathit{Maritalstatus}_i \\ &\quad + f_1(\mathit{Age}_i) + f_2(\mathit{Income}_i) + f_3(\mathit{Access}_i) + f_4(\mathit{PC1}_i), \quad i = 1, \ldots, n, \end{aligned}$$

where $g(\cdot)$ is the log-link and the smooth functions f_j are modeled using a linear combination of 15 cubic B-splines penalized by a third order penalty. The B-spline bases are defined over the domain $[x_{j,\min}, x_{j,\max}]$, where $x_{j,\min}$ ($x_{j,\max}$) is the minimum (maximum) of the covariate values on which f_j is defined. Given

the moderate number of additive terms ($q = 4$), the posterior penalty space is explored via the grid strategy of Section 2.5.1.

Table 7 summarizes the estimation results for the parametric linear part of the GAM. The results highlight a negative and significant relationship between the number of children in a household and the (mean) number of doctor visits. The demographic variable *Race* has a non-significant effect on the the mean response, while a negative and significant relationship between *Maritalstatus* and the (mean) number of doctor visits is observed. Figure 6 displays the estimated smooth functions (solid curves) and the associated 95% approximate pointwise credible intervals (gray surfaces).

Parameters	**Estimates**	**CI 90%**	$\mathbf{sd}_{post}$
β_1 (*Children*)	-0.179	[-0.239; -0.122]	0.036
β_2 (*Race*)	-0.127	[-0.263; 0.005]	0.081
β_3 (*Maritalstatus*)	-0.234	[-0.431; -0.043]	0.118

Table 7: Estimation results for the parametric linear part of the GAM. The second column is the parameter estimate, the third column gives the associated 90% credible interval and the last column is the posterior standard deviation.

As in Gurmu (1997), we observe a concave relationship between the mean response and Age with a peak in the average number of visits arising around *Age*=28. As most of the AFDC beneficiaries are women the concave pattern of Age may be explained by pregnancy-related visits during fertile periods and less frequent visits in later periods of life. The socio-economic variable *Income* exhibits no significant effect on the mean number of doctor visits when *Income* is below $10,000. Hence an increase in income for poor households with an annual income below $10,000 is (on average) not reflected by an increase in the number of doctor visits. However, when the annual income goes above $10,000 individuals tend to care more about their health and the (average) number of medical visits increases. Furthermore for the variable *Access*, we observe a strong oscillation of the mean response around a linear trend in the domain $[0, 70]$, suggesting that for low to moderate health service availability, the mean number of doctor visits remains stable.

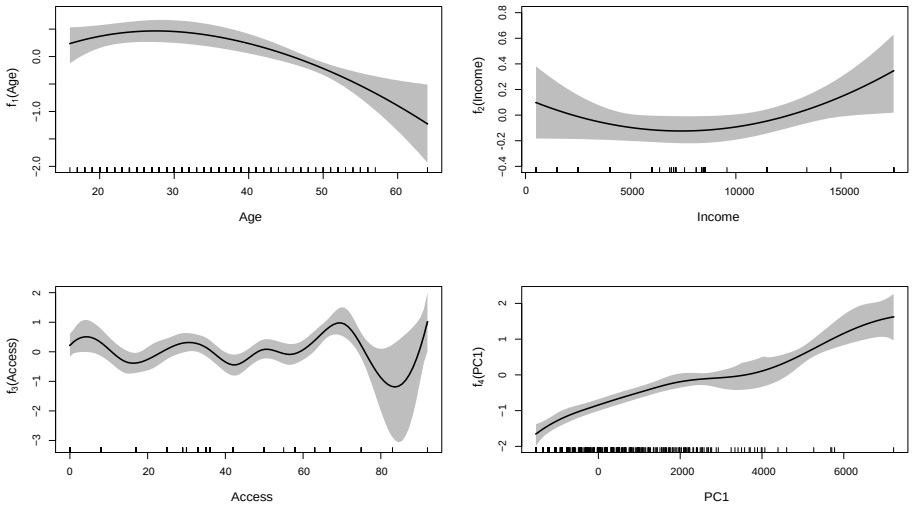

Figure 6: Estimated smooth functions (solid curve) and 95% approximate pointwise credible intervals (gray surface) for variables *Age*, *Income*, *Access* and *PC1*.

With regard to health-status variables gathered in *PC1* the results are as expected. Indeed, we observe a clear upward trend, i.e. the average number of medical visits increases with poorer health conditions.

5.2. Nutritional study

In a second application, we implement our methodology to analyze data from a nutritional epidemiology study. More thoroughly, we are interested in modeling the relationship between the plasma beta-carotene level and several explanatory variables related to individual factors and dietary characteristics. Human cells are driven by an important dynamic called the oxidation process, an energy delivery mechanism that is crucial for a proper functioning at the cellular level. By-products of the oxidation process are molecules known as free radicals. An imbalance between free radicals and antioxidant defenses generates oxidative stress which in turn triggers carcinogenesis. Beta-carotene is an antioxidant acting as a free radical scavenger and has been shown to prevent various cancer types and other diseases (Comstock et al., 1992; Rimm et al., 1993 and Zhang et al., 1999).

The dataset provided by Stukel (2008) on plasma beta-carotene levels has $n = 314$ observations on 14 variables. Factors influencing beta-carotene plasma concentration levels have been studied by Nierenberg et al. (1989), who found that beta-carotene level had a positive relationship with dietary beta-carotene consumption and tends to be larger for females, whereas a negative relation appeared with current smoker status. The dataset was also analyzed by Liu et al. (2011) who develop a variable selection procedure to identify the significant linear components in a semiparametric additive partial linear model. The LPS model is implemented on the data to study the relationship between the logarithm of beta-carotene plasma level (in ng/ml) and various explanatory variables retained as significant by the analysis in Liu et al. (2011).

The linear part of the additive model will include the *BMI* or Quetelet index (weight/height^2), the dietary beta-carotene consumption (*Betadiet*) (in mg/day), *Gender* (0=Male; 1=Female), a binary indicator *Smoking* status (0=non smoker; 1=current smoker) and the covariates *Fiber* and *Fat* indicating the hectograms of fiber and fat respectively consumed on a daily basis. The nonlinear part of the model will encompass the variables *Age* (in years) and the log of *Cholesterol* consumption (in mg/day). To summarize, the GAM model with an identity link is given by $y_i = \log(Betaplasma_i) \sim \mathcal{N}(\mu_i, s^2)$ where $s^2 = 0.559$ is the empirical variance of the response and the mean is modeled as:

$$\begin{aligned}\mu_i &= \beta_0 + \beta_1 BMI_i + \beta_2 Betadiet_i + \beta_3 Gender_i + \beta_4 Smoking_i + \beta_5 Fiber \\ &\quad + \beta_6 Fat + f_1(Age_i) + f_2(\log(Cholesterol_i)),\ i = 1, \ldots, n.\end{aligned}$$

In Table 8, we report the estimation results of the linear part. All variables are significant, except *Betadiet*. There is a negative association between *BMI* and the mean log plasma beta-carotene level meaning that for a fixed height, individuals with lower weight tend to have (on average) higher plasma beta-carotene concentrations. As in Nierenberg et al. (1989), we find that females and non-smokers tend to have a significantly larger beta-response level. A possible explanation is that smoke actually deteriorates beta-carotene molecules through an oxidation process. Finally, fiber consumption increases the mean

plasma beta-carotene level, with the consumption of vegetables on a daily basis helping to maintain antioxidants at a high level, while a high-fat diet tends to have a negative effect on the mean response.

Figure 7 highlights the estimated smooth functions for *Age* and log *Cholesterol*. For variable *Age* the shape of the estimated function is similar to what is observed in Liu et al. (2011). There is a positive association with the mean response when *Age* is smaller than 45 years or greater than 65 years. On the other hand, the relation of the mean response to the log-cholesterol level does not appear significant.

Parameters	**Estimates**	**CI 90%**	$\mathbf{sd}_{post}$
β_1 (BMI)	-0.034	[-0.046; -0.022]	0.007
β_2 $(Betadiet)$	0.047	[-0.009; 0.101]	0.033
β_3 $(Gender)$	0.300	[0.076; 0.520]	0.135
β_4 $(Smoking)$	-0.301	[-0.515; -0.093]	0.128
β_5 $(Fiber)$	2.396	[0.804; 3.938]	0.956
β_6 (Fat)	-0.245	[-0.493; -0.003]	0.149

Table 8: Estimation results for the parametric linear part of the GAM for the nutritional study. The second column is the parameter estimate, the third column gives the associated 90% credible interval and the last column is the posterior standard deviation.

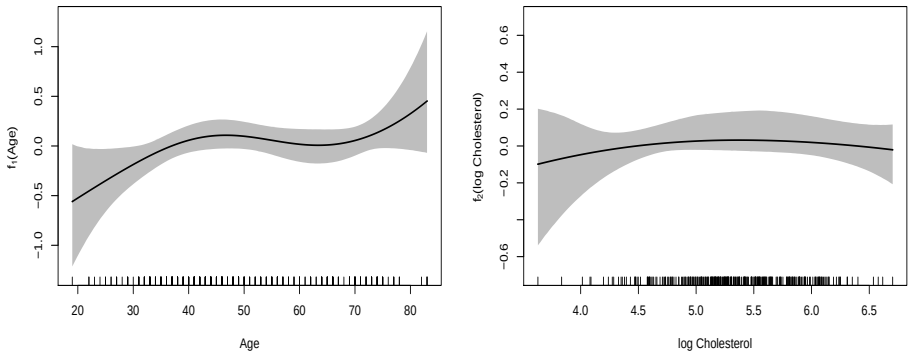

Figure 7: Estimated smooth functions (solid curve) and 95% approximate pointwise credible intervals (gray surface) for variables *Age* and log(*Cholesterol*) of the nutritional study dataset.

6. Concluding remarks

In this article, we have put forward a new methodology for approximate Bayesian estimation in Generalized additive models (GAMs) by unifiying P-splines and Laplace approximations. The Laplace-P-spline model is endowed with closed form expressions for the gradient and Hessian of the log posterior penalty vector. These analytical forms constitute a valuable asset for a computationally efficient and precise exploration strategy of the posterior penalty space that in turn leads to an accurate approximation of the joint posterior latent vector (including the regression and spline parameters in the generalized additive model) even when the number of smooth functions is large.

Extensive simulation studies show that the algorithms underlying LPS and LPSMAP exhibit good estimation quality with respect to the considered performance metrics, as shown for instance by non-significant biases or frequentist coverage probability of credible intervals appreciably close to their nominal value. Furthermore, our approximate Bayesian approach has proved to be reliable in terms of estimation performance with respect to smooth additive terms.

Finally, even though the Laplace-P-spline approach works from a complete Bayesian perspective, the computational budget required for inference is relatively low as compared to existing methods fully relying on MCMC algorithms. A future research challenge will be to summarize the algorithms in a software package to disseminate the LPS and LPSMAP approaches. Moreover, it would be interesting to explore the idea to handle models for spatial data or with additional hierarchy levels.

Conflict of interest

The authors declare no conflicts of interest.

Acknowledgments

Gressani Oswaldo wants to thank the Luxembourgish Ministry of higher education and research for a PhD program grant. The authors are also grateful to Dr. Thérèse Stukel for granting permission to use the nutritional study data in this article.

Appendix A

This appendix provides in full details the analytical derivations of the gradient and Hessian associated to the (log-) posterior of the log penalty vector:

$$\begin{aligned}
\log \widetilde{p}(\mathbf{v}|\mathcal{D}) \quad \doteq \quad & -\frac{1}{2}\underbrace{\log|B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}_{\text{Term I}} + \underbrace{\left(\frac{\nu+K-1}{2}\right)\sum_{j=1}^{q} v_j}_{\text{Term II}} \\
& + \underbrace{\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}}_{\text{Term III}} \\
& - \underbrace{\frac{1}{\varkappa}\sum_{i=1}^{n} s\left(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)}_{\text{Term IV}} - \frac{1}{2}\underbrace{\widetilde{\boldsymbol{\varpi}}^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} Q_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}}_{\text{Term V}} \\
& - \left(\frac{\nu}{2} + a_\delta\right)\underbrace{\sum_{j=1}^{q}\log\left(b_\delta + \frac{\nu}{2}\exp(v_j)\right)}_{\text{Term VI}}, \qquad (13)
\end{aligned}$$

where for notational convenience, we define $\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} := \left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\right)^{-1}$.

Gradient associated to the penalty in a GAM

To obtain the gradient of $\log \widetilde{p}(\mathbf{v}|\mathcal{D})$, the partial derivatives of the latter quantity with respect to $v_j,\ j = 1, \ldots, q$ are required. The partial derivative of Term I in (13) can be obtained using Jacobi's formula:

$$\begin{aligned}
\frac{\partial \log|B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}{\partial v_j} &= \frac{1}{|B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}\frac{\partial}{\partial v_j}|B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}| \\
&= \frac{1}{|B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}\mathrm{Tr}\left(\mathrm{adj}(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\frac{\partial}{\partial v_j}(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\right) \\
&= \frac{1}{|B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}\mathrm{Tr}\Big(|B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|\ (B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} \\
&\qquad\qquad \frac{\partial}{\partial v_j}(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\Big) \\
&= \mathrm{Tr}\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right),
\end{aligned}$$

where $\widetilde{P}_{v_j}$ is a (symmetric) block diagonal matrix defined as:

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

Derivation of Term II with respect to v_j simply equals the scalar $(\nu+K-1)/2$:

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

Partial derivatives of Term III and Term IV are obtained using:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} &= \frac{\partial}{\partial v_j}\left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\right)^{-1} \\
&= -\left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\right)^{-1}\widetilde{P}_{v_j}\left(B^T\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\right)^{-1} \\
&= -\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}.
\end{aligned}$$

For Term III, recall that the trace is invariant under cyclic permutations:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} y_i\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) &= \frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} y_i\text{Tr}\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\right) \\
&= \frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} y_i\text{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right) \\
&= \frac{1}{\varkappa}\sum_{i=1}^{n} y_i\frac{\partial}{\partial v_j}\text{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right) \\
&= \frac{1}{\varkappa}\sum_{i=1}^{n} y_i\text{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\frac{\partial}{\partial v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i\text{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i\text{Tr}\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}. \qquad (14)
\end{aligned}$$

For Term IV we use the chain rule and obtain:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} s\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\right) &= \frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\frac{\partial}{\partial v_j}\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\\
&= \frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\frac{\partial}{\partial v_j}\mathrm{Tr}\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\\
&= \frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\frac{\partial}{\partial v_j}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}.
\end{aligned}$$

The partial derivative of Term V is obtained as follows:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\left(\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) &= \frac{\partial}{\partial v_j}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\\
&= \frac{\partial}{\partial v_j}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\\
&= \mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\frac{\partial}{\partial v_j}\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right)\\
&= \mathrm{Tr}\Bigg(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\Bigg(\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial Q_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\\
&\quad +\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}\Bigg)\Bigg)
\end{aligned}$$

$$\begin{aligned}
&= \mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\left(-\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right)\\
&= \mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}^T\left(-\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\widetilde{\boldsymbol{\varpi}}\right)\\
&= -\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} + \widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\\
&= -\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} - \left(\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)^T\\
&\quad +\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\\
&= -\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\\
&\quad +\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\\
&= -2\widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} + \widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}.
\end{aligned}$$

With regard to the derivative of Term VI we have:

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

For notational convenience we define $\widetilde{\Upsilon}^j_{\mathbf{v}} := \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}$. From all the above intermediate results for Terms I-VI, the gradient $\nabla_{\mathbf{v}}\log\widetilde{p}(\mathbf{v}|\mathcal{D})$ has the following entries:

$$\begin{aligned}\frac{\partial\log\widetilde{p}(\mathbf{v}|\mathcal{D})}{\partial v_j} &= \underbrace{-\frac{1}{2}\mathrm{Tr}\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)}_{\text{Term VII}}+\left(\frac{\nu+K-1}{2}\right)\underbrace{-\frac{1}{\varkappa}\sum_{i=1}^{n}y_i\mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}}_{\text{Term VIII}}\\ &\quad\underbrace{+\frac{1}{\varkappa}\sum_{i=1}^{n}s'\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}\right)\mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}}_{\text{Term IX}}+\underbrace{\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}}_{\text{Term X}}\\ &\quad\underbrace{-\frac{1}{2}\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}}_{\text{Term XI}}\underbrace{-\frac{\left(\frac{\nu}{2}+a_\delta\right)}{1+\frac{2b_\delta}{\nu\exp(v_j)}}}_{\text{Term XII}},\ j=1,\ldots,q.\end{aligned}$$

Hessian associated to the penalty in a GAM

First, we focus on the diagonal entries. The derivative of Term VII is:

$$\begin{aligned}\frac{\partial}{\partial v_j}\mathrm{Tr}\left((B^T\widetilde{W}B+Q^{\mathbf{v}}_{\boldsymbol{\xi}})^{-1}\widetilde{P}_{v_j}\right) &= \mathrm{Tr}\left(\frac{\partial}{\partial v_j}(B^T\widetilde{W}B+Q^{\mathbf{v}}_{\boldsymbol{\xi}})^{-1}\widetilde{P}_{v_j}\right)\\ &= \mathrm{Tr}\left(-\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}+\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)\\ &= -\mathrm{Tr}\left(\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2-\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right).\end{aligned}$$

Let us derive the intermediate result:

$$\begin{aligned}
\frac{\partial \widetilde{\Upsilon}^j_{\mathbf{v}}}{\partial v_j} &= \frac{\partial}{\partial v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \\
&= \left(\frac{\partial \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}}{\partial v_j}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\frac{\partial \widetilde{P}_{v_j}}{\partial v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\frac{\partial \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}}{\partial v_j}\right) \\
&= \left(-\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} - \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\right) \\
&= \left(-2\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\Upsilon}^j_{\mathbf{v}}\right). \qquad (15)
\end{aligned}$$

Partial differentiation of Term VIII yields:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^n y_i \mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) &= \frac{\partial}{\partial v_j}\mathrm{Tr}\left(\frac{1}{\varkappa}\sum_{i=1}^n y_i \mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) \\
&= \frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^n y_i\mathrm{Tr}\left(\mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\right) \\
&= \frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^n y_i\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\right)\right) \\
&= \frac{1}{\varkappa}\sum_{i=1}^n y_i\frac{\partial}{\partial v_j}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\right) \\
&= \frac{1}{\varkappa}\sum_{i=1}^n y_i\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\left(\frac{\partial \widetilde{\Upsilon}^j_{\mathbf{v}}}{\partial v_j}\right)\right),
\end{aligned}$$

and using intermediate result (15), one obtains for Term VIII:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^n y_i \mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) &= -\frac{1}{\varkappa}\sum_{i=1}^n y_i\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\mathbf{b}_i^T\left(2\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} - \widetilde{\Upsilon}^j_{\mathbf{v}}\right)\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^n y_i\mathrm{Tr}\left(\mathbf{b}_i^T\left(2\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} - \widetilde{\Upsilon}^j_{\mathbf{v}}\right)\widetilde{\boldsymbol{\varpi}}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^n y_i\mathbf{b}_i^T\left(2\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} - \widetilde{\Upsilon}^j_{\mathbf{v}}\right)\widetilde{\boldsymbol{\varpi}}.
\end{aligned}$$

For Term IX, we have:

$$\frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right) = \frac{1}{\varkappa}\sum_{i=1}^{n}\Bigg(s''(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}) \frac{\partial}{\partial v_j}\mathrm{Tr}\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\left(\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right) + s'(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\frac{\partial}{\partial v_j}\mathrm{Tr}\left(\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right)\Bigg).$$

Using (14) and intermediate result (15) we have for Term IX:

$$\begin{aligned}\frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right) &= \frac{1}{\varkappa}\sum_{i=1}^{n}\Bigg(s''(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\left(-\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right)\\ &\quad\left(\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right) + s'(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^T\left(-2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\Upsilon}_{\mathbf{v}}^{j}\right)\widetilde{\boldsymbol{\varpi}}\Bigg)\\ = -\frac{1}{\varkappa}\sum_{i=1}^{n}\Bigg(&s'(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^T\left(2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\Upsilon}_{\mathbf{v}}^{j}\right)\widetilde{\boldsymbol{\varpi}}\\ &+s''(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right)^2\Bigg).\end{aligned}$$

The partial derivative of Term X is obtained as follows:

$$\begin{aligned}\frac{\partial}{\partial v_j}\left(\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) &= \frac{\partial}{\partial v_j}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\\ &= \frac{\partial}{\partial v_j}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\\ &= \mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\frac{\partial}{\partial v_j}\left(\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right)\\ &= \mathrm{Tr}\Bigg(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\Bigg(\frac{\partial\widetilde{\Upsilon}_{\mathbf{v}}^{j}}{\partial v_j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\Upsilon}_{\mathbf{v}}^{j}\frac{\partial Q_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\\ &\quad+\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}\Bigg)\Bigg)\\ &= \mathrm{Tr}\Bigg(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\Bigg(\Big(-2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\Upsilon}_{\mathbf{v}}^{j}\Big)Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\\ &\quad+\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\Upsilon}_{\mathbf{v}}^{j}\Bigg)\Bigg)\end{aligned}$$

$$
\begin{aligned}
&= \operatorname{Tr}\Bigg(\widetilde{\boldsymbol{\varpi}}^T\Bigg(-2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \\
&\quad +\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}-\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\Upsilon}_{\mathbf{v}}^{j}\Bigg)\widetilde{\boldsymbol{\varpi}}\Bigg) \\
&= -2\widetilde{\boldsymbol{\varpi}}^T\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}+\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\left(Q_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{P}_{v_j}\right)\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} \\
&\quad -\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}.
\end{aligned}
$$

Partial differentiation of Term XI gives us:

$$
\begin{aligned}
\frac{\partial}{\partial v_j}\left(\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right) &= \frac{\partial}{\partial v_j}\operatorname{Tr}\left(\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right) \\
&= \frac{\partial}{\partial v_j}\operatorname{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\right) \\
&= \operatorname{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\frac{\partial\widetilde{\Upsilon}_{\mathbf{v}}^{j}}{\partial v_j}\right) \\
&= \operatorname{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^T\left(-2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\Upsilon}_{\mathbf{v}}^{j}\right)\right) \\
&= \operatorname{Tr}\left(\widetilde{\boldsymbol{\varpi}}^T\left(-2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\Upsilon}_{\mathbf{v}}^{j}\right)\widetilde{\boldsymbol{\varpi}}\right) \\
&= -2\widetilde{\boldsymbol{\varpi}}^T\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}+\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}.
\end{aligned}
$$

Finally derivation of Term XII gives us:

$$
\frac{\partial}{\partial v_j}\frac{\left(\frac{\nu}{2}+a_\delta\right)}{\left(1+\frac{2b_\delta}{\nu\exp(v_j)}\right)}=\frac{b_\delta\left(1+\frac{2a_\delta}{\nu}\right)\exp(-v_j)}{\left(1+\frac{2b_\delta}{\nu\exp(v_j)}\right)^2}.
$$

Using the differentiation results for Terms VII-XII, the diagonal elements of the Hessian of $\log\widetilde{p}(\mathbf{v}|\mathcal{D})$ are:

$$
\begin{aligned}
\frac{\partial^2 \log \widetilde{p}(\mathbf{v}|\mathcal{D})}{\partial v_j^2} &= \frac{1}{2}\mathrm{Tr}\left(\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2 - \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right) \\
&+\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^T \left(2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2 \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\Upsilon}_{\mathbf{v}}^{j}\right)\widetilde{\boldsymbol{\varpi}} \\
&-\frac{1}{\varkappa}\sum_{i=1}^{n}\left(s'(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^T\left(2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\Upsilon}_{\mathbf{v}}^{j}\right)\widetilde{\boldsymbol{\varpi}}\right.\right. \\
&\left.\left.+s''(\mathbf{b}_i^T\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right)^2\right)\right. \\
&-2\widetilde{\boldsymbol{\varpi}}^T\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} \\
&+\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\left(Q_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{P}_{v_j}\right)\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}} \\
&+\widetilde{\boldsymbol{\varpi}}^T\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} - \frac{1}{2}\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}} \\
&-\frac{b_\delta\left(1+\frac{2a_\delta}{\nu}\right)\exp(-v_j)}{\left(1+\frac{2b_\delta}{\nu\exp(v_j)}\right)^2}, \quad j=1,\ldots,q.
\end{aligned}
$$

Regarding the off-diagonal components, note that for index $s \neq j$ we have for Term VII:

$$
\begin{aligned}
\frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right) &= \mathrm{Tr}\left(\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}\widetilde{P}_{v_j}\right) \\
&= -\mathrm{Tr}\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right).
\end{aligned}
$$

Let us define $\widetilde{\Upsilon}_{\mathbf{v}}^{s} := \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}$ and consider the following intermediate result:

$$
\begin{aligned}
\frac{\partial\widetilde{\Upsilon}_{\mathbf{v}}^{j}}{\partial v_s} &= \frac{\partial}{\partial v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \\
&= \left(\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial\widetilde{P}_{v_j}}{\partial v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}\right) \\
&= \left(-\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right) \\
&= -\left(\widetilde{\Upsilon}_{\mathbf{v}}^{s}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\Upsilon}_{\mathbf{v}}^{s}\right). \qquad (16)
\end{aligned}
$$

Result (16) can be used to obtain the differentiation of Term VIII:

$$\begin{aligned}
\frac{\partial}{\partial v_s}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right) &= \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right) \\
&= \frac{\partial}{\partial v_s}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathrm{Tr}\left(\mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right)\right) \\
&= \frac{1}{\varkappa}\sum_{i=1}^{n} y_i \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}} \mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j\right) \\
&= \frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}} \mathbf{b}_i^T \frac{\partial \widetilde{\Upsilon}_{\mathbf{v}}^j}{\partial v_s}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}} \mathbf{b}_i^T \left(\widetilde{\Upsilon}_{\mathbf{v}}^s \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\Upsilon}_{\mathbf{v}}^s\right)\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathrm{Tr}\left(\mathbf{b}_i^T \left(\widetilde{\Upsilon}_{\mathbf{v}}^s \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\Upsilon}_{\mathbf{v}}^s\right) \widetilde{\boldsymbol{\varpi}}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^T \left(\widetilde{\Upsilon}_{\mathbf{v}}^s \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\Upsilon}_{\mathbf{v}}^s\right) \widetilde{\boldsymbol{\varpi}}.
\end{aligned}$$

To derive Term IX, we also use result (16):

$$\begin{aligned}
\frac{\partial}{\partial v_s}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}\right) \mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right) &= \frac{1}{\varkappa}\sum_{i=1}^{n}\Bigg(s''(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}) \\
&\quad \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}\right)\left(\mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right) \\
&\quad + s'(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}) \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right)\Bigg) \\
&= \frac{1}{\varkappa}\sum_{i=1}^{n}\Bigg(s''(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}})\left(-\mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^s \widetilde{\boldsymbol{\varpi}}\right)\left(\mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right) \\
&\quad + s'(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}})\left(-\mathbf{b}_i^T \left(\widetilde{\Upsilon}_{\mathbf{v}}^s \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\Upsilon}_{\mathbf{v}}^s\right) \widetilde{\boldsymbol{\varpi}}\right)\Bigg) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n}\Bigg(s'(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}) \mathbf{b}_i^T \left(\widetilde{\Upsilon}_{\mathbf{v}}^s \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\Upsilon}_{\mathbf{v}}^s\right) \widetilde{\boldsymbol{\varpi}} \\
&\quad + s''(\mathbf{b}_i^T \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^s \widetilde{\boldsymbol{\varpi}}\right)\left(\mathbf{b}_i^T \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right)\Bigg).
\end{aligned}$$

Partial differentiation of Term X goes as follows:

$$\begin{aligned}
\frac{\partial}{\partial v_s}\left(\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\varpi}\right) &= \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\varpi}\right)\\
&= \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\varpi}\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\right)\\
&= \mathrm{Tr}\left(\widetilde{\varpi}\widetilde{\varpi}^T\frac{\partial}{\partial v_s}\left(\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\right)\right)\\
&= \mathrm{Tr}\Bigg(\widetilde{\varpi}\widetilde{\varpi}^T\Bigg(\frac{\partial\widetilde{\Upsilon}^j_{\mathbf{v}}}{\partial v_s}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}+\widetilde{\Upsilon}^j_{\mathbf{v}}\frac{\partial Q^{\mathbf{v}}_{\boldsymbol{\xi}}}{\partial v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\\
&\quad +\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\frac{\partial\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}}{\partial v_s}\Bigg)\Bigg)\\
&= \mathrm{Tr}\Bigg(\widetilde{\varpi}\widetilde{\varpi}^T\Bigg(-\left(\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}+\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}\right)Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\\
&\quad +\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}-\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\Bigg)\Bigg)\\
&= \mathrm{Tr}\Bigg(\widetilde{\varpi}^T\Bigg(-\left(\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}+\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}\right)Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\\
&\quad +\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}-\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\Bigg)\widetilde{\varpi}\Bigg)\\
&= -\widetilde{\varpi}^T\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\varpi}\\
&\quad -\widetilde{\varpi}^T\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\varpi}\\
&\quad +\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\varpi}-\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{\varpi}.
\end{aligned}$$

Partial differentiation of Term XI gives us:

$$\begin{aligned}
\frac{\partial}{\partial v_s}\left(\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\varpi}\right) &= \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\varpi}\right)\\
&= \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\varpi}\widetilde{\varpi}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\right)\\
&= \mathrm{Tr}\left(\widetilde{\varpi}\widetilde{\varpi}^T\frac{\partial\widetilde{\Upsilon}^j_{\mathbf{v}}}{\partial v_s}\right)\\
&= -\mathrm{Tr}\left(\widetilde{\varpi}\widetilde{\varpi}^T\left(\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}+\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}\right)\right)\\
&= -\mathrm{Tr}\left(\widetilde{\varpi}^T\left(\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}+\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}\right)\widetilde{\varpi}\right)
\end{aligned}$$

$$
\begin{aligned}
&= -\widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} - \left(\widetilde{\boldsymbol{\varpi}}^T \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{P}_{v_j} \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} \right)^T \\
&= -\widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} \\
&= -2\widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}}.
\end{aligned}
$$

Finally, using the above results, the off-diagonal elements $s = 1, \ldots, q$; $j = 1, \ldots, q$ and $s \neq j$ of the Hessian of $\log \widetilde{p}(\mathbf{v}|\mathcal{D})$ are:

$$
\begin{aligned}
\frac{\partial^2 \log \widetilde{p}(\mathbf{v}|\mathcal{D})}{\partial v_s \, \partial v_j} &= \frac{1}{2} \mathrm{Tr} \left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{P}_{v_s} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{P}_{v_j} \right) + \frac{1}{\varkappa} \sum_{i=1}^{n} y_i \mathbf{b}_i^T \left(\widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{P}_{v_j} \widetilde{\Upsilon}^s_{\mathbf{v}} \right) \widetilde{\boldsymbol{\varpi}} \\
&\quad - \frac{1}{\varkappa} \sum_{i=1}^{n} \Bigg(s'(\mathbf{b}_i^T \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}}) \mathbf{b}_i^T \left(\widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{P}_{v_j} \widetilde{\Upsilon}^s_{\mathbf{v}} \right) \widetilde{\boldsymbol{\varpi}} \\
&\quad + s''(\mathbf{b}_i^T \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}}) \left(\mathbf{b}_i^T \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} \right) \left(\mathbf{b}_i^T \widetilde{\Upsilon}^j_{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} \right) \Bigg) \\
&\quad - \widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} Q^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{P}_{v_j} \widetilde{\Upsilon}^s_{\mathbf{v}} Q^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} \\
&\quad + \widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^j_{\mathbf{v}} \widetilde{P}_{v_s} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^j_{\mathbf{v}} Q^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} + \widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^s_{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}}.
\end{aligned}
$$

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

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

$$
\begin{aligned}
\frac{\partial \log \widetilde{p}(\mathbf{v}|\mathcal{D})}{\partial v_j} &= -\frac{1}{2} \mathrm{Tr} \left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{P}_{v_j} \right) + \left(\frac{\nu + K - 1}{2} \right) - \frac{1}{\varkappa} \sum_{i=1}^{n} y_i \mathbf{b}_i^T \widetilde{\Upsilon}^j_{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} \\
&\quad + \frac{1}{\varkappa} \sum_{i=1}^{n} s' \left(\mathbf{b}_i^T \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} \right) \mathbf{b}_i^T \widetilde{\Upsilon}^j_{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} + \widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^j_{\mathbf{v}} Q^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} \widetilde{\boldsymbol{\varpi}} \\
&\quad - \frac{1}{2} \widetilde{\boldsymbol{\varpi}}^T \widetilde{\Upsilon}^j_{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} - \frac{\left(\frac{\nu}{2} + a_\delta \right)}{1 + \frac{2b_\delta}{\nu \exp(v_j)}}.
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{\partial^2 \log \widetilde{p}(\mathbf{v}|\mathcal{D})}{\partial v_j^2} &= \frac{1}{2}\mathrm{Tr}\left(\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2 - \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right) \\
&+\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^T \left(2\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2 \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} - \widetilde{\Upsilon}^j_{\mathbf{v}}\right)\widetilde{\boldsymbol{\varpi}} \\
&-\frac{1}{\varkappa}\sum_{i=1}^{n}\left(s'(\mathbf{b}_i^T\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^T\left(2\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} - \widetilde{\Upsilon}^j_{\mathbf{v}}\right)\widetilde{\boldsymbol{\varpi}}\right.\right. \\
&\left.\left.+s''(\mathbf{b}_i^T\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)^2\right)\right. \\
&-2\widetilde{\boldsymbol{\varpi}}^T\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}} \\
&+\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\left(Q^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{P}_{v_j}\right)\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} \\
&+\widetilde{\boldsymbol{\varpi}}^T\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}} - \frac{1}{2}\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} \\
&-\frac{b_\delta\left(1+\frac{2a_\delta}{\nu}\right)\exp(-v_j)}{\left(1+\frac{2b_\delta}{\nu\exp(v_j)}\right)^2}, \quad j = 1, \ldots, q.
\end{aligned}
$$

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

$$
\begin{aligned}
\frac{\partial^2 \log \widetilde{p}(\mathbf{v}|\mathcal{D})}{\partial v_s\, \partial v_j} &= \frac{1}{2}\mathrm{Tr}\left(\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\right) + \frac{1}{\varkappa}\sum_{i=1}^{n} y_i\mathbf{b}_i^T\left(\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}\right)\widetilde{\boldsymbol{\varpi}} \\
&-\frac{1}{\varkappa}\sum_{i=1}^{n}\left(s'(\mathbf{b}_i^T\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^T\left(\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}} + \widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}\right)\widetilde{\boldsymbol{\varpi}}\right.\right. \\
&\left.\left.+s''(\mathbf{b}_i^T\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^T\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\left(\mathbf{b}_i^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\right)\right. \\
&-\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}} \\
&+\widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}} - \widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}} + \widetilde{\boldsymbol{\varpi}}^T\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}.
\end{aligned}
$$

To assess the accuracy of the above gradient and Hessian equations associated to $\log \widetilde{p}(\mathbf{v}|\mathcal{D})$, we have implemented a procedure in **R** that compares the analytical results with the numerical derivatives of $\log \widetilde{p}(\mathbf{v}|\mathcal{D})$ obtained with the `grad()` and `hessian()` functions of the **numDeriv** package at 50 randomly selected

points $\mathbf{v} \in \mathbb{R}^3$ with $v_j \sim \text{Uniform}(-4, 8),\ j = 1, 2, 3$ and the response generated from a Poisson distribution. Numerical and analytical derivative results turn out to be very similar, a clear indication that the derived analytical results are accurate.

Appendix B

In this appendix, we show the derivations related to the skew-normal fit to the conditional $\widetilde{p}(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$. The skew-normal distribution denoted by $X \sim \text{SN}(\mu, \varsigma^2, \rho)$ has probability density function:

$$p(x) = \frac{2}{\varsigma}\varphi\left(\frac{x-\mu}{\varsigma}\right)\ \Phi\left(\rho\frac{(x-\mu)}{\varsigma}\right). \tag{17}$$

The first moment and the second and third central moments of X 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 $\widetilde{p}(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$, where $\widehat{\mathbf{v}}_{-j}$ is the vector $\widehat{\mathbf{v}}$ without the j^{th} entry. The empirical moments of the conditionals are computed on an equidistant grid $\{v_{jl}\}_{l=1}^{L}$ with interval length Δ_l and correspond to:

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

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

$$\mathcal{M}_{j1} = \mu + \varsigma \sqrt{\frac{2}{\pi}}\, \psi \tag{18}$$

$$\mathcal{M}_{j2} = \varsigma^2 \left(1 - \frac{2}{\pi}\, \psi^2\right) \tag{19}$$

$$\mathcal{M}_{j3} = \frac{1}{2}(4-\pi)\, \varsigma^3 \left(\frac{2}{\pi}\right)^{\frac{3}{2}}\, \psi^3. \tag{20}$$

From (19), we isolate ς:

$$\varsigma = \sqrt{\frac{\mathcal{M}_{j2}}{\left(1 - \frac{2}{\pi}\, \psi^2\right)}} > 0. \tag{21}$$

Plugging (21) in (20) yields:

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

Let $\kappa := \mathcal{M}_{j3}^{\frac{1}{3}}\pi^{\frac{1}{2}}/(4-\pi)^{\frac{1}{3}}2^{\frac{1}{6}}\, \mathcal{M}_{j2}^{\frac{1}{2}}$, 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 is $\Delta = 4\left(1 + \frac{2\kappa^2}{\pi}\right)\kappa^2 > 0$. 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 $\mathcal{M}_{j3}$ is negative/positive, ψ^* (and by extension ρ^*) should also be negative/positive to capture the negatively/positively skewed pattern of $\tilde{p}(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$. Hence using the sign$(\cdot)$ function:

$$\psi^* = \text{sign}(\mathcal{M}_{j3}) \frac{\sqrt{4\big(\kappa^2 + \frac{2\kappa^4}{\pi}\big)}}{2 + \frac{4\kappa^2}{\pi}}. \tag{22}$$

So, we have $\rho^* = \psi^*/\sqrt{1-(\psi^*)^2}$ and plugging (22) in (21), we recover:

$$\varsigma^* = \sqrt{\frac{\mathcal{M}_{j2}}{\big(1 - \frac{2}{\pi}\ (\psi^*)^2\big)}}. \tag{23}$$

Finally, the location parameter is given by:

$$\mu^* = \mathcal{M}_{j1} - \varsigma^* \sqrt{\frac{2}{\pi}}\ \psi^*. \tag{24}$$

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

Appendix C

		90%			95%			99%		
Data	Method	f_1	f_2	f_3	f_1	f_2	f_3	f_1	f_2	f_3
Poisson	LPS	87.6	87.0	89.1	93.0	92.6	94.4	98.0	98.1	98.9
	LPSMAP	86.7	85.6	88.7	92.4	91.6	94.0	97.7	97.4	98.7
	MGCV	89.8	89.6	90.3	94.4	94.4	95.1	98.8	98.7	99.1
Normal	LPS	90.8	91.1	91.0	95.6	95.8	95.8	99.2	99.0	99.3
	LPSMAP	90.6	90.7	90.9	95.4	95.4	95.6	99.2	99.0	99.3
	MGCV	91.1	91.5	91.2	95.8	95.8	95.8	99.3	99.1	99.3
Binomial	LPS	90.2	89.3	90.3	95.0	94.5	95.3	98.8	98.8	99.1
	LPSMAP	89.9	88.8	90.1	94.7	94.1	95.1	98.7	98.6	99.1
	MGCV	91.2	90.2	90.9	95.4	95.1	95.6	99.0	98.9	99.2

Effective frequentist coverages of 90%, 95% and 99% pointwise credible intervals averaged over 200 uniformly distributed values of the covariate x in $[-1, 1]$ for Poisson, Normal and Binomial data with $S = 500$ replications of sample size $n = 300$ for the Laplace-P-spline (LPS), the LPS omitting the mixture (LPSMAP) and `gam()` (MGCV) methods.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, *12*, 171–178. URL: www.jstor.org/stable/4615982.

Azzalini, A. (2014). *The skew-normal and related families* volume 3. Cambridge University Press.

Bornkamp, B. (2011). Approximating probability densities by iterated Laplace approximations. *Journal of Computational and Graphical Statistics*, *20*, 656–669. doi:https://doi.org/10.1198/jcgs.2011.10099.

Comstock, G. W., Bush, T. L., & Helzlsouer, K. (1992). Serum retinol, beta-carotene, vitamin E, and selenium as related to subsequent cancer of specific sites. *American Journal of Epidemiology*, *135*, 115–121. doi:10.1093/oxfordjournals.aje.a116264.

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*, 89–121. doi:10.1214/ss/1038425655.

Eilers, P. H. C., Marx, B. D., & Durbán, M. (2015). Twenty years of P-splines. *SORT: Statistics and operations research transactions*, *39*, 149–186.

Fraaije, R. G. A., ter Braak, C. J. F., Verduyn, B., Breeman, L. B. S., Verhoeven, J. T. A., & Soons, M. B. (2015). Early plant recruitment stages set the template for the development of vegetation patterns along a hydrological gradient. *Functional Ecology*, *29*, 971–980.

Gómez-Rubio, V., & Rue, H. (2017). Markov chain Monte Carlo with the Integrated Nested Laplace Approximation. *Statistics and Computing*, *28*, 1033–1051. doi:10.1007/s11222-017-9778-y.

Gressani, O., & Lambert, P. (2018). Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines. *Computational Statistics & Data Analysis*, *124*, 151–167.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. *Journal of Applied Econometrics*, *12*, 225–242.

Hastie, T., & Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, *1*, 297–310. doi:10.1214/ss/1177013604.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*, 371–386. doi:10.2307/2289439.

Hastie, T. J., & Tibshirani, R. J. (1990). Generalized additive models, volume 43 of Monographs on Statistics and Applied Probability.

Hui, F. K. C., You, C., Shang, H. L., & Müller, S. (2019). Semiparametric regression using variational approximations. *Journal of the American Statistical Association*, *114*, 1765–1777. doi:10.1080/01621459.2018.1518235.

Jullion, A., & Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics & Data Analysis*, *51*, 2542–2558. doi:10.1016/j.csda.2006.09.027.

Krivobokova, T., Crainiceanu, C. M., & Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics*, *17*, 1–20. URL: www.jstor.org/stable/27594289.

Lang, S., & Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics*, *13*, 183–212. doi:10.1198/1061860043010.

Leonard, T. (1982). Comment on "A Simple Predictive Density Function," by M. Lejeune and G.D. Faulkenberry. *Journal of the American Statistical Association*, *77*, 657–658.

Liu, X., Wang, L., & Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, *21*, 1225–1248. doi:10.5705/ss.2009.140.

Luts, J., Broderick, T., & Wand, M. P. (2014). Real-time semiparametric regression. *Journal of Computational and Graphical Statistics*, *23*, 589–615. doi:10.1080/10618600.2013.810150.

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, *55*, 2372–2387. doi:10.1016/j.csda.2011.02.004.

Martins, T. G., Simpson, D., & Lindgren, H., F.and Rue (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis*, *67*, 68–83. doi:10.1016/j.csda.2013.04.014.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* volume 37. CRC press.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A*, *135*, 370–384. doi:10.2307/2344614.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R., & Group, S. C. P. S. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, *130*, 511–521. doi:10.1093/oxfordjournals.aje.a115365.

Rimm, E. B., Stampfer, M. J., Ascherio, A., Giovannucci, E., Colditz, G. A., & Willett, W. C. (1993). Vitamin E consumption and the risk of coronary heart disease in men. *New England Journal of Medicine*, *328*, 1450–1456. doi:10.1056/NEJM199305203282004.

Rue, H., Martino, S., & 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*, 319–392. doi:10.1111/j.1467-9868.2008.00700.x.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *An-*

nual Review of Statistics and its Application, *4*, 395–421. doi:10.1146/annurev-statistics-060116-054045.

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

Sapra, S. K. (2013). Generalized additive models in business and economics. *International Journal of Advanced Statistics and Probability*, *1*. doi:10.14419/ijasp.v1i3.1022.

Stukel, T. (2008). Determinants of plasma retinol and beta-carotene levels. *StatLib Datasets Archive. URL http://lib. stat. cmu. edu/datasets/Plasma_Retinol*, .

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86. doi:10.1080/01621459.1986.10478240.

Umlauf, N., Adler, D., Kneib, T., Lang, S., & Zeileis, A. (2015). Structured additive regression models: An R interface to BayesX. *Journal of Statistical Software*, *63*, 1–46. doi:10.18637/jss.v063.i21.

Ventrucci, M., & Rue, H. (2016). Penalized complexity priors for degrees of freedom in Bayesian P-splines. *Statistical Modelling*, *16*, 429–453.

Wand, M., & Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, *50*, 179–198. doi:https://doi.org/10.1111/j.1467-842X.2008.00507.x.

Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association*, *112*, 137–168. doi:10.1080/01621459.2016.1197833.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B*, *65*, 95–114. doi:https://doi.org/10.1111/1467-9868.00374.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*, 3–36.

Wood, S. N. (2017). *Generalized additive models: an introduction with R (Second edition)*. CRC press.

Wood, S. N., Scheipl, F., & Faraway, J. J. (2013). Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, *23*, 341–360. doi:10.1007/s11222-012-9314-z.

Yoon, J. W., & Wilson, S. P. (2011). Inference for latent variable models with many hyperparameters. In *Proceedings of the 58th World Statistical Congress of the International Statistical Institute, Dublin*.

Zhang, S., Hunter, D. J., Forman, M. R., Rosner, B. A., Speizer, F. E., Colditz, G. A., Manson, J. E., Hankinson, S. E., & Willett, W. C. (1999). Dietary carotenoids and vitamins A, C, and E and risk of breast cancer. *Journal of the National Cancer Institute*, *91*, 547–556. doi:10.1093/jnci/91.6.547.