Statistical Modelling 23(5-6), pp. 409-423, 2023.
DOI: 10.1177/1471082X231181173

(with R package ‘ordgam’)

Penalty parameter selection and asymmetry corrections to Laplace approximations in Bayesian P-splines models

Philippe Lambert [1,2], **Oswaldo Gressani** [3]

[1] Institut de Mathématique, Université de Liège, Belgium.
[2] Institut de Statistique, Biostatistique et Sciences Actuarielles (ISBA), Université catholique de Louvain, Belgium.
[3] Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat), Data Science Institute, Hasselt University, Belgium.

Address for correspondence: Philippe Lambert, Institut de Mathématique, Université de Liège, Allée de la Découverte 12 (B37), B-4000 Liège, Belgium.
E-mail: `p.lambert@uliege.be`.
Phone: +32 4 366 59 90.

Abstract: Laplace P-splines (LPS) combine the P-splines smoother and the Laplace approximation in a unifying framework for fast and flexible inference under the Bayesian paradigm. The Gaussian Markov random field prior assumed for penalized parameters and the Bernstein-von Mises theorem typically ensure a razor-sharp accuracy of the Laplace approximation to the posterior distribution of these quantities. This accuracy can be seriously compromised for some unpenalized parameters, especially when the information synthesized by the prior and the likelihood is sparse. Therefore, we propose a refined version of the LPS methodology by splitting the parameter space in two subsets. The first set involves parameters for which the joint posterior distribution is approached from a non-Gaussian perspective with an approximation scheme tailored to capture asymmetric patterns, while the posterior distribution for the penalized parameters in the complementary set undergoes the LPS treatment with Laplace approximations. As such, the dichotomization of the parameter space provides the necessary structure for a separate treatment of model parameters, yielding improved estimation accuracy as compared to a setting where posterior quantities are uniformly handled with Laplace. In addition, the proposed enriched version of LPS remains entirely sampling-free, so that it operates at a computing speed that is far from reach to any existing Markov chain Monte Carlo approach. The methodology is illustrated on the additive proportional odds model with an application on ordinal survey data.

Key words: Additive model ; Bayesian P-splines ; Laplace approximation ; Skewness.

1 Introduction

By publishing his *Mémoire sur la probabilité des causes par les événements* (Laplace, 1774), the young French polymath Pierre-Simon de Laplace (1749-1827) seeded an idea today known as the Laplace approximation. At that time, Laplace probably could not have imagined that almost two centuries later, his approximation technique would be resurrected (see e.g. Leonard, 1982; Tierney and Kadane, 1986; Rue et al., 2009) to play a pivotal role in the modern Bayesian literature. Essentially, the Laplace approximation is a Gaussian distribution centered at the maximum a posteriori (MAP) of the target distribution with a variance-covariance matrix that coincides with the inverse of the negative Hessian of the log-posterior target evaluated at the MAP estimate. Recently, the Laplace approximation has crossed the path of P-splines, the brainchild of Paul Eilers and Brian Marx (Eilers and Marx, 1996), to inaugurate a new approximate Bayesian methodology labelled "Laplace P-splines" (LPS) with promising applications in survival analysis (Gressani and Lambert, 2018; Gressani et al., 2022b; Lambert and Kreyenfeld, 2023), generalized additive models (Gressani and Lambert, 2021), nonparametric double additive location-scale models for censored data (Lambert, 2021) and infectious disease epidemiology (Gressani et al., 2022a,c). The sampling-free inference scheme delivered by Laplace approximations combined with the possibility of smoothing different model components with P-splines in a flexible fashion paves the way for a robust and much faster alternative to existing simulation-based methods.

Although LPS shares some methodological aspects with the popular integrated nested Laplace approximations (INLA) approach (Rue et al., 2009), there are fundamental points of divergence. First, the tools in INLA and its associated R-INLA software are originally built to compute approximate posteriors of univariate latent variables, contrary to LPS that natively delivers approximations to the (multivariate) joint posterior distribution of the latent vector. The key benefit of working with an approximate version of the joint posterior is that pointwise estimators and credible intervals for subsets of the latent vector (and functions thereof) can be straightforwardly constructed. Second, by working with closed-form expressions for the gradients and Hessians involved in the model, LPS is computationally more efficient than the numerical differentiation proposed in INLA. Third, while INLA can be combined with various techniques for smoothing nonlinear model components, LPS is entirely devoted to P-splines smoothers with the key advantage of having full control over the penalization scheme (as the approximate posterior distribution of the penalty parameter(s) is analytically available). In this regard, LPS has more in common with the work of Wood and Fasiolo (2017) than with INLA, especially in the class of (generalized) additive models (Wood, 2017).

The success of Laplace approximations in Bayesian statistics owes much to a central limit type argument. Under certain regularity conditions, the Bernstein-von Mises theorem (see e.g. Van der Vaart, 1998) ensures that posterior distributions in differentiable models converge to a Gaussian distribution under large samples. In situations involving small to medium sample sizes, the suitability of the Laplace approximation can be questioned as it does not take into account the potential skewness or kurtosis of posterior distributions (Ruli et al., 2016). Even under relatively large samples, the Laplace approximation might fail in scenarios involving binary data as the latter are poorly informative for the model parameters and can result in a flat log-likelihood function, thus complicating inference (Ferkingstad and Rue, 2015; Gressani and Lambert, 2021).

Laplace P-splines belong to the class of latent Gaussian models, where model parameters are dichotomized between a vector of latent variables $\boldsymbol{\xi}$ (including penalized B-spline coefficients, regression coefficients and other parameters of interest) that are assigned a Gaussian prior and another vector of hyperparameters $\boldsymbol{\eta}$ that involves nuisance parameters, such as the smoothing parameter inherent to P-splines, and for which prior assumptions need not be Gaussian. Combining Bayes' rule and a simplified Laplace approximation, the conditional posterior distribution of $\boldsymbol{\xi}$ under the LPS framework is approximated by a Gaussian distribution denoted by $\widetilde{p}_G(\boldsymbol{\xi}|\widehat{\boldsymbol{\eta}}, \mathcal{D})$, where $\widehat{\boldsymbol{\eta}}$ is a summary statistic of the posterior hyperparameter vector (e.g. the MAP estimate, the posterior mean or median) and $\mathcal{D}$ denotes the observed data. Although the latter approximation is typically accurate for penalized B-spline coefficients, it might be less appropriate for other candidates in $\boldsymbol{\xi}$ with large prior variance. In that case, the misfit between the Laplace approximation and a potentially asymmetric (or heavy-tailed) target posterior distribution for a parameter can have a detrimental effect on posterior summary statistics and on any results relying on the generated approximation for the posterior distribution of the model parameters. This motivates us to develop an approach that corrects for potential posterior misfits provided by the Laplace approximation.

A recent technique proposed by Chiuchiolo et al. (2022) in the INLA framework consists in using a skew Gaussian copula to correct for skewness when posterior latent variables have a non-negligible deviation from Gaussianity. Our proposal in models involving P-splines consists in splitting the latent vector $\boldsymbol{\xi}$ into a set of parameters $\boldsymbol{\gamma}$ for which the posterior distribution (conditional on the hyperparameters) is approximated in a non-Gaussian fashion with an emphasis on capturing asymmetries, and a set of parameters $\boldsymbol{\theta}$ (that typically involves penalized B-spline coefficients) for which the conditional posterior is approached with Laplace approximations. Our refined LPS approach thus allows to obtain an approximation to the joint posterior distribution of $\boldsymbol{\xi}$ (given $\boldsymbol{\eta}$) together with an approximation to the posterior of the hyperparameters $\boldsymbol{\eta}$ without relying on an MCMC sampling scheme.

A simple motivating example inspired by the infectious disease model of Gressani et al. (2022c) helps framing the problem. Let $\mathcal{D} = \{(x_i, y_i) : i = 1, \ldots, n\}$ be a sample of $n = 120$ independent pairs where $x_i = i$ and y_i has a negative binomial distribution $\mathrm{NB}(\mu(x_i), \phi)$ following the parametrization of Piegorsch (1990) with mean $\mathrm{E}(y|x) = \mu(x)$, variance $\mathrm{V}(y|x) = \mu(x) + \phi\, \mu(x)^2$ and overdispersion parameter $\phi > 0$.

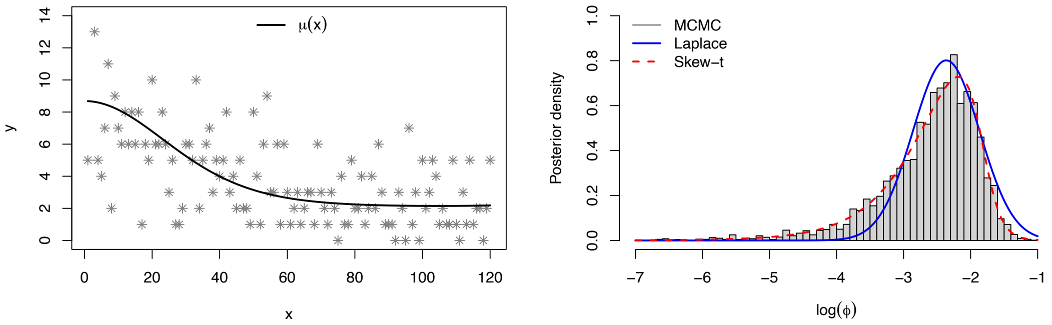

Figure 1: Left panel: Count data ($n = 120$) generated using a negative binomial $\mathrm{NB}(\mu(x), \phi)$ with $\phi^{-1} = 6$. Right panel: Histogram of a MCMC sample for $\log \phi$ compared to Laplace (solid) and skew-t (dashed) approximations.

We model the conditional mean with P-splines $\log \mu(x) = \boldsymbol{\theta}^\top \mathbf{b}(x)$, where $\mathbf{b}(\cdot)$ is a cubic B-spline basis on the interval $[1, 120]$ and $\boldsymbol{\theta}$ is a vector of B-spline coefficients. Following Lang and Brezger (2004), a global smoothness prior is assumed for the B-spline parameters, $p(\boldsymbol{\theta}|\lambda) \propto \exp\left(-\frac{\lambda}{2}\boldsymbol{\theta}^\top \mathbf{P}\boldsymbol{\theta}\right)$, where $\mathbf{P}$ is a penalty matrix and λ is the penalty parameter to which we assign a weakly informative Gamma prior (with mean a/b and variance a/b^2), denoted by $\lambda \sim \mathcal{G}(a, b)$. To complete the model specification, a Gamma prior with large variance is assumed for the overdispersion parameter ϕ^{-1}. The left panel of Figure 1 shows a data set of size $n = 120$ simulated from the above negative binomial model with a nonlinear function for $\mu(x)$ and $\phi^{-1} = 6$. The histogram for $\log(\phi)$ on the right panel of Figure 1 is obtained from a long MCMC chain with a Metropolis-within-Gibbs algorithm. There is an important misfit between the Laplace approximation (solid curve) and the MCMC output, so that quantities like the posterior standard deviation or selected posterior quantiles for $\log \phi$ will be poorly estimated using a straightforward Laplace approximation to $p(\phi|\mathcal{D})$. The dashed curve represents the skewed distribution that we propose as an alternative candidate to the Laplace approximation and that will be thoroughly discussed in the next section within a Bayesian P-splines context. The fitted distribution is able to capture the asymmetry that is apparent in the MCMC sample, improving the precision of posterior estimates for $\log \phi$ as compared to Laplace and at a much lower computational cost than MCMC. The article is organized as follows. Section 2 presents the Bayesian Laplace P-spline model and gives a detailed description of the proposed asymmetric posterior approximation methodology for non-penalized parameters. In Section 3, we illustrate the method in an additive proportional odds model for ordinal data. Finally, Section 4 concludes with a discussion.

2 Laplace approximation and Bayesian P-splines

2.1 Model specification

Consider a regression model describing the conditional distribution of a response y for given covariates $\boldsymbol{x}$. Denote by $\boldsymbol{\xi}$ the model parameters: it includes the regression and spline parameters, plus possibly the (log of the) scale and (unconstrained transformed) shape parameters. Denote by $p(\boldsymbol{\xi}|\boldsymbol{\eta})$ the joint prior density of $\boldsymbol{\xi}$ conditionally on a vector of hyperparameters $\boldsymbol{\eta}$. In the context of a P-spline model, the latter parameters can come from J unknown smooth functions specified as $f_j(\cdot) = \sum_{\ell=1}^{L} \theta_{\ell j} b_{j\ell}(\cdot)$ $(j = 1, \ldots, J)$ where $\mathcal{B}_j = \{b_{j\ell}(\cdot) : \ell = 1, \ldots, L\}$ denotes a B-spline basis with equidistant knots spanning the argument range. Vector $\boldsymbol{\eta}$ would typically include positive roughness penalty parameters λ_j with prior density $p(\lambda_j)$ for the jth function. The frequentist penalty on changes in differences of neighbour spline parameters (Eilers and Marx, 1996) can be translated in a Bayesian context using a conditional prior on $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jL})^\top$ (Lang and Brezger, 2004), $p(\boldsymbol{\theta}_j|\lambda_j) \propto \exp\left(-\frac{1}{2}\,\boldsymbol{\theta}_j{}^\top(\lambda_j \mathbf{P})\boldsymbol{\theta}_j\right)$, with $\mathbf{P} = \mathbf{D}_r^\top \mathbf{D}_r$ denoting a penalty matrix corresponding to a finite difference penalty matrix $\mathbf{D}_r$ of order r. For example, when $r = 2$, one has $\boldsymbol{\theta}_j^\top \mathbf{P} \boldsymbol{\theta}_j = ||\mathbf{D}_r \boldsymbol{\theta}_j||_2^2 = \sum_{\ell=1}^{L-2}(\theta_{\ell+2,j} - 2\theta_{\ell+1,j} + \theta_{\ell,j})^2$. The penalty parameter $\lambda_j > 0$ is used to tune the smoothness of the associated additive term with, at the limit when $\lambda_j \to +\infty$, a polynomial of order $r-1$ for $f_j(\cdot)$. Different prior distributions could be chosen for λ_j with Brezger and Lang (2006) suggesting to take Gamma priors $\lambda_j \sim \mathcal{G}\,(a_j, b_j)$ (with mean a_j/b_j and variance a_j/b_j^2). A small value for b_j $(= 10^{-4}$, say) combined with $b_j = a_j$ or $a_j = 1$ ensures a large prior variance with some more weight set on small or large values of λ_j, respectively. Mixtures of Gamma densities were also investigated in Jullion and Lambert (2007) with $(\lambda|\delta) \sim \mathcal{G}\,(\nu/2, \nu\delta/2)$ and $\delta \sim \mathcal{G}\,(a_\delta, b_\delta)$ yielding, in the special case $a_\delta = b_\delta = .5$ and $\nu = 1$, a Beta prime distribution $\mathcal{B}'(.5, .5)$ for λ or equivalently a half-Cauchy prior for $\sqrt{\lambda}$ with $p(\lambda) \propto \lambda^{-.5}(1+\lambda)^{-1}$ (Lambert and Bremhorst, 2019).

Penalties can also be combined and extended in multiple ways, see e.g. the book by Eilers and Marx (2021) for inspiring examples. More generally, we assume that the joint conditional prior for the vector $\boldsymbol{\theta}$, stacking all the vectors of penalized B-spline coefficients in the model, can be written as

$$p(\boldsymbol{\theta}|\boldsymbol{\lambda}) \propto \exp\left(-\frac{1}{2}\,\boldsymbol{\theta}^\top \boldsymbol{\mathcal{P}}_\lambda\, \boldsymbol{\theta}\right), \tag{2.1}$$

where $\boldsymbol{\mathcal{P}}_\lambda$ is a positive semi-definite matrix. The vector of model parameters can be reorganized as follows, $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top)^\top \in \mathrm{I\!R}^{k_1+k_2}$, where $\boldsymbol{\gamma} \in \mathrm{I\!R}^{k_1}$ denotes the vector of non-penalized parameters. If $\mathcal{D}$ generically denotes the available data and if $\boldsymbol{\lambda}$ stands for the vector of hyperparameters $\boldsymbol{\eta}$ in the specific context of P-spline models, then the joint posterior for $\boldsymbol{\xi}$ directly follows from Bayes' theorem,

$$p(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D}) \propto \mathcal{L}(\boldsymbol{\xi}|\mathcal{D})\, p(\boldsymbol{\gamma})\, p(\boldsymbol{\theta}|\boldsymbol{\lambda})\, p(\boldsymbol{\lambda}),$$

where $\mathcal{L}(\boldsymbol{\xi}|\mathcal{D})$ denotes the likelihood. It is typically explored using Markov chain

Monte Carlo methods (MCMC). In this paper, we build up on the methodology described in Gressani and Lambert (2018, 2021) and in Lambert (2021), where Laplace approximations to the conditional posterior of $(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$ and an additional approximation to the marginal posterior of $(\boldsymbol{\lambda}|\mathcal{D})$ enable to bypass sampling algorithms, see Section 2.2.

2.2 Laplace approximation and penalty parameter selection

Assume that closed form expressions can be derived for the gradient and Hessian of $\log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$,

$$\mathbf{U}_\lambda = \mathbf{U}_\lambda(\boldsymbol{\xi}) = \partial \log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})/\partial\boldsymbol{\xi} \;\; ; \;\; \mathbf{H}_\lambda = \mathbf{H}_\lambda(\boldsymbol{\xi}) = \partial^2 \log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\top \; .$$

The conditional posterior mode $\hat{\boldsymbol{\xi}}_\lambda$ of $\boldsymbol{\xi}$ can be quickly obtained using the Newton-Raphson (NR) algorithm with the substitution, $\boldsymbol{\xi} \longleftarrow \boldsymbol{\xi} - \mathbf{H}_\lambda^{-1}\mathbf{U}_\lambda$, repeated until convergence. The Levenberg-Marquardt algorithm (Marquardt, 1963) could be preferred if good initial conditions are not easily found to ensure convergence. A Laplace approximation to the conditional posterior distribution of $\boldsymbol{\xi}$ directly follows: $(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D}) \mathrel{\dot\sim} \mathcal{N}_{k_1+k_2}(\hat{\boldsymbol{\xi}}_\lambda,\Sigma_\lambda)$ where $\Sigma_\lambda = -\mathbf{H}_\lambda^{-1}$. Thanks to the Gaussian Markov random field (GMRF) prior (Rue and Held, 2005), $p(\boldsymbol{\theta}|\boldsymbol{\lambda})$, assumed in (2.1) for the penalized parameters, the Gaussian approximation to the conditional posterior of $(\boldsymbol{\theta}|\boldsymbol{\lambda},\mathcal{D})$ is usually excellent, see Rue et al. (2009) for the same argument in latent Gaussian models. However this might not be true for some non-penalized parameters in $\boldsymbol{\xi}$, especially when the combined information coming from their prior and the likelihood is sparse, see Section 2.3 for a specific handling.

The preceding Laplace approximation can be used to approximate the marginal posterior distribution of the penalty parameters $\boldsymbol{\lambda}$ with the Normal approximation substituted in the denominator of the following identity, $p_\lambda(\boldsymbol{\lambda}|\mathcal{D}) = p(\boldsymbol{\xi},\boldsymbol{\lambda}|\mathcal{D})/p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$, yielding

$$\begin{aligned}\widetilde{p}_\lambda(\boldsymbol{\lambda}|\mathcal{D}) = \frac{p(\boldsymbol{\xi},\boldsymbol{\lambda}|\mathcal{D})}{\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})} &\propto p(\hat{\boldsymbol{\xi}}_\lambda,\boldsymbol{\lambda}|\mathcal{D})\big|\widehat{\Sigma}_\lambda\big|^{\frac{1}{2}} \\ &\propto \underbrace{\mathcal{L}(\hat{\boldsymbol{\xi}}_\lambda|\mathcal{D})\, p(\hat{\boldsymbol{\xi}}_\lambda|\boldsymbol{\lambda},\mathcal{D})|\widehat{\Sigma}_\lambda|^{\frac{1}{2}}}_{\text{Marginal likelihood}} \times p(\boldsymbol{\lambda})\,,\end{aligned} \tag{2.2}$$

see Tierney and Kadane (1986) for the same strategy in the approximation of a marginal distribution. One might prefer to work with $\boldsymbol{v} = \log\boldsymbol{\lambda}$ and its approximate marginal posterior,

$$\widetilde{p}_v(\boldsymbol{v}|\mathcal{D}) = \tilde{p}_\lambda(\mathrm{e}^{\boldsymbol{v}}|\mathcal{D}) \prod_j \mathrm{e}^{v_j} \,. \tag{2.3}$$

The maximization of (2.2) or of the marginal likelihood (as with ‘empirical Bayes’ methods) can be used to select a specific value for $\boldsymbol{\lambda}$. Alternatively, it could be derived from the log-penalty using (2.3), yielding larger penalty values when the selection is made from the marginal posterior instead of the (parametrization invariant) marginal likelihood.

2.3 Asymmetric posterior for non-penalized parameters

Assume that $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top)^\top \in \mathbb{R}^{k_1+k_2}$ with $\boldsymbol{\gamma} \in \mathbb{R}^{k_1}$ is suspected to have a non-symmetric marginal posterior distribution. Let $\hat{\boldsymbol{\xi}}_\lambda = (\hat{\boldsymbol{\gamma}}_\lambda^\top, \hat{\boldsymbol{\theta}}_\lambda^\top)^\top$ denote the posterior mode of $p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D})$ and $\widehat{\Sigma}_\lambda^{-1}$ the observed information matrix structured in blocks as follows,

$$\widehat{\Sigma}_\lambda = \begin{bmatrix} \widehat{\Sigma}_\lambda^{\gamma\gamma} & \widehat{\Sigma}_\lambda^{\gamma\theta} \\ \widehat{\Sigma}_\lambda^{\theta\gamma} & \widehat{\Sigma}_\lambda^{\theta\theta} \end{bmatrix} .$$

The conditional posterior of $(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})$ has an approximate Normal distribution resulting from the GMRF prior in (2.1) for $(\boldsymbol{\theta}|\boldsymbol{\lambda})$. One has

$$(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}) \mathrel{\dot\sim} \mathcal{N}_{k_2}\left(\mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}), \widehat{\Sigma}_\lambda^{\theta|\gamma}\right) ,$$

where

$$\begin{aligned} \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}) &= \hat{\theta}_\lambda + \widehat{\Sigma}_\lambda^{\theta\gamma}\left(\widehat{\Sigma}_\lambda^{\gamma\gamma}\right)^{-1}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_\lambda) ; \\ \widehat{\Sigma}_\lambda^{\theta|\gamma} &= \widehat{\Sigma}_\lambda^{\theta\theta} - \widehat{\Sigma}_\lambda^{\theta\gamma}\left(\widehat{\Sigma}_\lambda^{\gamma\gamma}\right)^{-1}\widehat{\Sigma}_\lambda^{\gamma\theta} . \end{aligned} \tag{2.4}$$

Hence, starting from the following identity, $p_\gamma(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D}) = p(\boldsymbol{\gamma}, \boldsymbol{\theta}|\boldsymbol{\lambda}, \mathcal{D})/p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})$, one gets the approximation

$$p_\gamma(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D}) \approx \frac{p(\boldsymbol{\gamma}, \boldsymbol{\theta}|\boldsymbol{\lambda}, \mathcal{D})}{\widetilde{p}_G(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})} \propto p\left(\boldsymbol{\gamma}, \mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})|\boldsymbol{\lambda}, \mathcal{D}\right)\left|\widehat{\Sigma}_\lambda^{\theta|\gamma}\right|^{\frac{1}{2}} , \tag{2.5}$$

see Eq. (2) in Tierney et al. (1989) for a similar expression. We propose to reparametrize $\boldsymbol{\gamma}$ by projecting it on the eigenvectors of the singular value decomposition (SVD) of $\widehat{\Sigma}_\lambda^{\gamma\gamma} = \mathbf{V}\mathrm{Z}\mathbf{V}^\top$ where $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_{k_1}]$ denotes the matrix of orthonormal eigenvectors, $\boldsymbol{\zeta}$ the eigenvalues and $\mathrm{Z} = \mathrm{diag}(\boldsymbol{\zeta})$. It yields $\tilde{\boldsymbol{\gamma}} = \mathrm{Z}^{-\frac{1}{2}}\mathbf{V}^\top(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_\lambda)$ and $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}}_\lambda + \mathbf{V}\mathrm{Z}^{\frac{1}{2}}\tilde{\boldsymbol{\gamma}}$, with

$$p_{\tilde{\gamma}}(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D}) \propto p_\gamma(\hat{\boldsymbol{\gamma}}_\lambda + \mathbf{V}\mathrm{Z}^{\frac{1}{2}}\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D}). \tag{2.6}$$

The posterior dependence between the components of $\tilde{\boldsymbol{\gamma}}$ is expected to be milder than under the original $\boldsymbol{\gamma}$ parametrization. Therefore, conditionally on $\boldsymbol{\lambda}$, we propose to approximate the joint posterior density of $(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D})$ by the product of the marginal densities of its components:

$$p_{\tilde{\gamma}}(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D}) \approx \prod_{s=1}^{k_1} p_{\tilde{\gamma}_s}(\tilde{\gamma}_s|\boldsymbol{\lambda}, \mathcal{D}). \tag{2.7}$$

Under that working independence hypothesis, each univariate marginal in the product in (2.7) is equal to its conditional with the other components set equal to an arbitrary value. Combined with (2.6), it implies that

$$\begin{aligned} p_{\tilde{\gamma}_s}(\tilde{\gamma}_s|\boldsymbol{\lambda}, \mathcal{D}) &= p_{\tilde{\gamma}_s|\tilde{\gamma}_{-s}}(\tilde{\gamma}_s|\tilde{\boldsymbol{\gamma}}_{-s} = 0, \boldsymbol{\lambda}, \mathcal{D}) \\ &\propto p_{\tilde{\gamma}}(\tilde{\gamma}_s\mathbf{e}_s|\boldsymbol{\lambda}, \mathcal{D}) \\ &\propto p_\gamma(\hat{\boldsymbol{\gamma}}_\lambda + \tilde{\gamma}_s\sqrt{\zeta_s}\mathbf{v}_s|\boldsymbol{\lambda}, \mathcal{D}) , \end{aligned} \tag{2.8}$$

where $\mathbf{e}_s$ denotes the sth unit vector in $\mathbb{R}^{k_1}$ such that $[\mathbf{e}_s]_k = \delta_{ks}$. The univariate marginal posterior density in (2.8) can be evaluated for any value of $\tilde{\gamma}_s$ using (2.5). We suggest to approximate it using a skew-normal (SN) or a skew-t (ST) distribution to also handle kurtosis. Here, we provide details for the ST distribution. By definition, $X \sim \mathrm{ST}(\psi, \omega^2, \alpha, \upsilon)$ if $X \in \mathbb{R}$ and has density

$$t(x|\psi, \omega^2, \alpha, \upsilon) = \frac{2}{\omega}\, t_\upsilon\left(\frac{x-\psi}{\omega}\right) T_\upsilon\left(\alpha\frac{x-\psi}{\omega}\right) \;,$$

with location parameter $\psi \in \mathbb{R}$, scale parameter $\omega > 0$, slant (or skewness) parameter $\alpha \in \mathbb{R}$ and $\upsilon > 0$ degrees of freedom (d.f.), where t_υ, T_υ respectively denote the density and c.d.f. of a standard Student distribution with υ d.f., see e.g. Azzalini and Capitanio (2014) for more details on the ST definition and its properties. An approximation to the target distribution can for example be derived by minimizing its Jensen-Shannon divergence from a skew-t distribution, giving $(\tilde{\gamma}_s|\boldsymbol{\lambda}, \mathcal{D}) \dot{\sim} \mathrm{ST}(\tilde{\psi}_s, \tilde{\omega}_s^2, \tilde{\alpha}_s, \tilde{\upsilon}_s)$. Substituting these ST densities in (2.7) provides an approximation to the joint posterior of $(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D})$ and an efficient method to sample from it using the independence of its components. An analytic form for the joint posterior density of $(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D})$ follows from the combination of (2.7) with $\tilde{\boldsymbol{\gamma}} = \mathrm{Z}^{-\frac{1}{2}}\mathbf{V}^\top(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}_\lambda)$, giving

$$p_\gamma(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D}) = \prod_{s=1}^{k_1} \frac{1}{\sqrt{\zeta_s}}\, t(\tilde{\gamma}_s|\tilde{\psi}_s, \tilde{\omega}_s^2, \tilde{\alpha}_s, \tilde{\upsilon}_s).$$

An approximation to the marginal distribution of its sth component $(\gamma_s|\boldsymbol{\lambda}, \mathcal{D})$ can be obtained using fast Monte Carlo techniques. Indeed, an approximate large random sample from $(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D})$ can first be generated by sampling its independent skew-t components $\mathrm{ST}(\tilde{\psi}_s, \tilde{\omega}_s^2, \tilde{\alpha}_s, \tilde{\upsilon}_s)$, yielding $\{\tilde{\boldsymbol{\gamma}}^{(m)} : m = 1, \ldots, M\}$. The associated random sample for $(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D})$ is given by $\{\boldsymbol{\gamma}^{(m)} = \hat{\boldsymbol{\gamma}}_\lambda + \mathbf{V}\mathrm{Z}^{\frac{1}{2}}\tilde{\boldsymbol{\gamma}}^{(m)} : m = 1, \ldots, M\}$. Then, a skew-$t$ approximation to the marginal posterior of γ_s can be fitted to $\{\gamma_s^{(m)} : m = 1, \ldots, M\}$, yielding $(\gamma_s|\boldsymbol{\lambda}, \mathcal{D}) \dot{\sim} \mathrm{ST}(\psi_s, \omega_s^2, \alpha_s, \upsilon_s)$. Point estimates or credible regions for γ_s can be computed from it. These different steps provide a convenient approximation to the joint posterior of the model parameters. Indeed, based on the factorization of the joint posterior density, $p(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D}) = p(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D})\, p(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D})\, p(\boldsymbol{\lambda}|\mathcal{D})$, one has the following stochastic representation for $(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D})$,

$$(\boldsymbol{\xi}, \boldsymbol{\lambda}|\mathcal{D}) \dot{\sim} \mathcal{N}_{k_2}\left(\mathbb{E}(\boldsymbol{\theta}|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathcal{D}), \widehat{\Sigma}_\lambda^{\theta|\gamma}\right) \times \prod_{s=1}^{k_1} \mathrm{ST}(\tilde{\gamma}_s|\tilde{\psi}_s, \tilde{\omega}_s^2, \tilde{\alpha}_s, \tilde{\upsilon}_s) \times (\boldsymbol{\lambda}|\mathcal{D}) \;, \qquad (2.9)$$

with mean and variance-covariance matrix in the first factor given in (2.4). It can be used to generate an arbitrarily large number of independent copies from the joint posterior much faster than with MCMC. This is for example particularly useful to make inference on complicated functions of the model parameters or for predictive purposes.

3 Illustration

3.1 The additive proportional odds model for ordinal data

Denote by $\mathcal{E}_m = \{1, \ldots, m\}$ the set containing the first m positive integers. Assume that n independent units are observed with data $\mathcal{D} = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$ where y is an ordinal random variable taking values in $\mathcal{E}_R$ and $\mathbf{x}$ a vector of covariates. The proportional odds (PO) model is a popular choice when the response is ordinal (Agresti, 2010). It assumes that

$$\text{logit}[P(Y \leq r|\mathbf{x})] = \eta_r = \gamma_r + \mathbf{x}^\top \boldsymbol{\theta} \quad (r \in \mathcal{E}_{R-1}) \ ,$$

with a specific intercept γ_r for each cumulative logit, but a shared vector of regression parameters $\boldsymbol{\theta}$. Consequently,

$$\log \frac{\Pr(Y \leq r|\mathbf{x}_1)/\Pr(Y > r|\mathbf{x}_1)}{\Pr(Y \leq r|\mathbf{x}_2)/\Pr(Y > r|\mathbf{x}_2)} = \boldsymbol{\theta}(\mathbf{x}_1 - \mathbf{x}_2)$$

is independent of r and provides a clear interpretation to the regression parameters $\boldsymbol{\theta}$ with a change θ_k in the log-odds of Y taking values in the lower end of the ordinal scale for every unit increase in the kth component of $\mathbf{x}$. Let $F_{ir} = P(Y_i \leq r|\mathbf{x}_i) = \mathrm{e}^{\eta_{ir}}/(1+\mathrm{e}^{\eta_{ir}})$ for r in $\mathcal{E}_{R-1}$ and let $F_{i0} = 0$, $F_{iR} = 1$. Then, $\pi_{ir} = P(Y_i = r|\mathbf{x}_i) = F_{ir} - F_{i,r-1}$ for r in $\mathcal{E}_R$. Let $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top)^\top$ and assume a prior of the following form,

$$p(\boldsymbol{\xi}|\boldsymbol{\lambda}) \propto \exp\left(-\frac{1}{2}\,(\boldsymbol{\xi} - \mathbf{e})^\top \mathbf{K}_\lambda (\boldsymbol{\xi} - \mathbf{e})\right) \ , \tag{3.1}$$

conditionally on a vector of parameters $\boldsymbol{\lambda}$ and for a positive semi-definite matrix $\mathbf{K}_\lambda$. The log-likelihood takes a simple form, $\ell(\boldsymbol{\xi}|\mathcal{D}) = \sum_{i=1}^n \log \pi_{iy_i}(\boldsymbol{\xi})$, with the resulting conditional posterior for $\boldsymbol{\xi}$,

$$\log p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D}) = \ell(\boldsymbol{\xi}|\mathcal{D}) - \frac{1}{2}\,(\boldsymbol{\xi} - \mathbf{e})^\top \mathbf{K}_\lambda (\boldsymbol{\xi} - \mathbf{e}). \tag{3.2}$$

Explicit analytical forms can be derived for the associated gradient and Hessian matrix, see Appendix A. Assume now for simplicity a model with J continuous covariates $x_1, \ldots, x_J$ with smooth additive terms $f_j(x_j)$ $(j = 1, \ldots, J)$ describing their effects on the conditional log-odds,

$$\text{logit}[P(Y \leq r|\mathbf{x})] = \eta_r = \gamma_r + f_1(x_1) + \ldots + f_J(x_J) \ .$$

Following Eilers and Marx (1996), consider now a basis of $(L+1)$ cubic B-splines $\{s^*_{j\ell}(\cdot)\}_{\ell=1}^{L+1}$ associated to a generous number of equally spaced knots on the range $(x_j^{\min}, x_j^{\max})$ of values for x_j (Marx and Eilers, 1998). They are recentered for identification purposes in the additive model using $s_{j\ell}(\cdot) = s^*_{j\ell}(\cdot) - \frac{1}{x_j^{\max} - x_j^{\min}} \int_{x_j^{\min}}^{x_j^{\max}} s^*_{j\ell}(u)du$ $(\ell = 1, \ldots, L)$. Then, the additive terms in the conditional model can be approximated using linear combinations of these recentered B-splines, $f_j(x_j) = \sum_{\ell=1}^L s_{j\ell}(x_{ij})\theta_{\ell j}$. In a Bayesian framework, as reminded in Section 2.1, smoothness can be forced on these additive terms by taking GMRF priors for the vectors of spline coefficients

$$p(\boldsymbol{\theta}_j|\lambda_j) \propto \exp\left(-\frac{1}{2}\,{\boldsymbol{\theta}_j}^\top(\lambda_j\mathbf{P})\boldsymbol{\theta}_j\right) \ ,$$

where $\mathbf{P}$ stands for the penalty matrix. A multivariate Normal prior could be taken for $\boldsymbol{\gamma}$ to complete the model specification, $\boldsymbol{\gamma} \sim \mathcal{N}\left(\tilde{\mathbf{e}}, (\mathbf{Q})^{-1}\right)$. Under the general formulation in (3.1), one has $\boldsymbol{\xi} = (\boldsymbol{\gamma}^\top, \boldsymbol{\theta}^\top)^\top$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_J^\top)^\top$, $\mathbf{e} = (\tilde{\mathbf{e}}^\top, \mathbf{0}_{JL}^\top)^\top$, and a block-diagonal penalty matrix $\mathbf{K}_\lambda = \text{diag}\big(\mathbf{Q}, \boldsymbol{\mathcal{P}}_\lambda\big)$ where $\boldsymbol{\mathcal{P}}_\lambda = \boldsymbol{\Lambda} \otimes \mathbf{P}$ with $[\boldsymbol{\Lambda}]_{jj'} = \delta_{jj'}\lambda_j$. The conditional posterior for $\boldsymbol{\xi}$ is given by (3.2). With a Gamma prior for the penalty parameters, $\lambda_j \sim \mathcal{G}\,(a, b)$, one has $(\lambda_j|\boldsymbol{\xi}, \mathcal{D}) \sim \mathcal{G}\,\big(a + .5\,\rho(\mathbf{P}), b + .5\,\boldsymbol{\theta}_j^\top\mathbf{P}\boldsymbol{\theta}_j\big)$. Starting from these conditional posterior distributions, a Metropolis-within-Gibbs algorithm can be set up to generate a random sample from the joint posterior for $(\boldsymbol{\xi}, \boldsymbol{\lambda})$, with Gibbs steps for the penalty parameters $\boldsymbol{\lambda}$ and Metropolis steps for the regression and splines parameters $\boldsymbol{\xi}$. Alternatively, proposals for $\boldsymbol{\xi}$ could be made using the modified Langevin (Roberts and Tweedie, 1996; Lambert and Eilers, 2009) or the Metropolis-Hastings algorithm with proposals based on the local topological information provided by the explicit analytic forms for the gradient and Hessian matrix (Gamerman, 1997). Such a sampling approach based on MCMC will be compared to the strategy proposed in Section 2.

3.2 Application on survey data

Consider now an illustration of the proposed methodology on data coming from the European Social Survey (ESS Round 9, 2018) with a specific focus on the French speaking respondents from Wallonia, one of the three regions in Belgium. Each of the participants (aged at least 15) was asked to react to the following statement, *Gay men and lesbians should be free to live their own life as they wish*, with a positioning on a Likert scale going from 1 (=*Agree strongly*) to 5 (=*Disagree strongly*), with 3 labelled as *Neither agree nor disagree* (with relative frequencies 1: 54.9% ; 2: 30.4% ; 3: 8.2% ; 4: 5.4% ; 5: 1.1%). That ordinal response effectively recorded on $n = 552$ respondents was analyzed using the proportional odds model described above with the number of completed years of education (14.1 ± 4.4 years) and age (47.3 ± 18.5 years) entering as additive terms with $L = 10$ recentered B-splines spanning each covariate range. Starting from Gamma priors, $\lambda_j \sim \mathcal{G}\,(1, 10^{-4})$ $(j = 1, 2)$, the penalty parameters λ_1 and λ_2 associated to $f_1(\text{eduyrs})$ and $f_2(\text{age})$, respectively, were selected by maximizing $p(\boldsymbol{\lambda}|\mathcal{D})$ in (2.2) using the Levenberg-Marquardt algorithm, yielding $\hat{\lambda}_1 = 191.8$ (e.d.f.=1.24), $\hat{\lambda}_2 = 18.4$ (e.d.f.=2.55), the value in brackets standing for the effective degrees of freedom. Alternatively the maximization of the marginal likelihood in (2.2) would yield a very large value for $\hat{\lambda}_1$ (e.d.f.=1.00), suggesting linearity for $f_1(\text{eduyrs})$, and $\hat{\lambda}_2 = 18.5$ (e.d.f.=2.55) practically unchanged. The fitted additive terms are visible on Fig. 2 with their pointwise 95% credible intervals, suggesting a statistically non-significant effect of `eduyrs`, but a tolerant perception of homosexuality tending to decrease with `age`, with a marked change in attitude revealed beyond age 60. To compare the merits of our proposal, a MCMC algorithm was

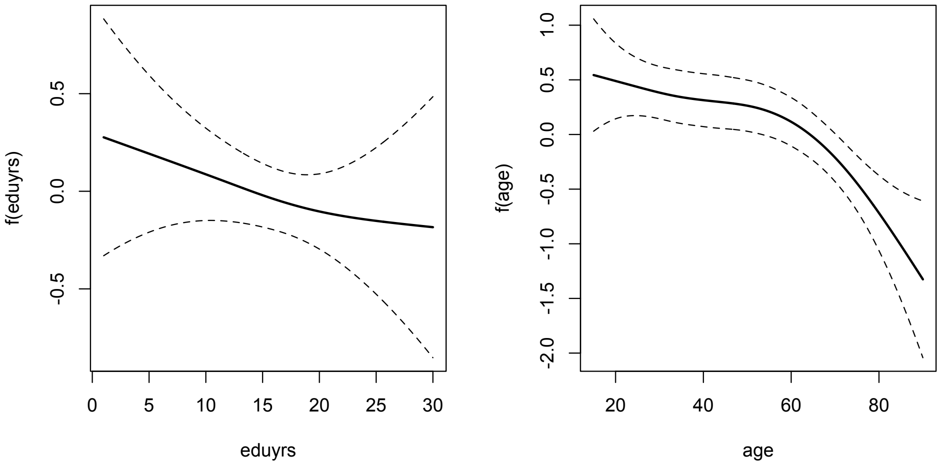

Figure 2: ESS dataset: fitted additive terms for `eduyrs` and `age` with pointwise 95% credible intervals: this suggests a growing hostility to homosexuality beyond the age of 60, while it appears that the number of completed years of education does not play a statistically significant role.

run to explore $p(\boldsymbol{\xi}|\hat{\boldsymbol{\lambda}},\mathcal{D})$, the generated samples and their properties being compared to the analytical approximations suggested in Section 2.3. The estimated additive terms and their 95% credible intervals based on MCMC are practically identical to our estimates in Fig. 2, confirming the excellent quality of the Laplace approximation to the conditional posterior distribution of $(\boldsymbol{\theta}|\boldsymbol{\lambda},\mathcal{D})$ underlying our calculations. Let us now focus on the non-penalized regression parameter in $\boldsymbol{\gamma}$ standing for the four intercepts in the proportional odds model. The scatterplot of the MCMC sample $\{\boldsymbol{\gamma}^{(m)}:m=1,\ldots,M\}$ generated from $p(\boldsymbol{\gamma}|\hat{\boldsymbol{\lambda}},\mathcal{D})$ using the modified Langevin algorithm can be found in the left panel of Fig. 3 where the posterior dependence between the vector components clearly stands out. The reparametrization suggested in Section 2.3 along the principal axes corresponding to the eigenvectors of the SVD decomposition of $\widehat{\Sigma}_{\lambda}^{\gamma\gamma}$ yields $\tilde{\boldsymbol{\gamma}}$, with the scatterplot of the associated MCMC sample $\{\tilde{\boldsymbol{\gamma}}^{(m)}=\mathrm{Z}^{-\frac{1}{2}}\mathbf{V}^{\top}(\boldsymbol{\gamma}^{(m)}-\hat{\boldsymbol{\gamma}}_{\lambda}):m=1,\ldots,M\}$, visible in the right panel of Fig. 3 confirming that the posterior dependence between the vector components of $\tilde{\boldsymbol{\gamma}}$ is very mild and probably negligible for most practical purposes. The suggested analytical approximation to $p_{\tilde{\gamma}_s}(\tilde{\gamma}_s|\boldsymbol{\lambda},\mathcal{D})$ in (2.8) was evaluated and added to the MCMC sample taken as a trustful proxy of the true marginal posterior distribution of $(\tilde{\gamma}_s|\boldsymbol{\lambda},\mathcal{D})$, see Fig. 4. The quality of the analytical approximations is excellent with a noticeable left asymmetry for $(\tilde{\gamma}_1|\boldsymbol{\lambda},\mathcal{D})$ in particular. When transformed back to the γ-parametrization, the approximating skew-Normal densities shown in Fig. 5 closely match the distribution of the MCMC samples for $(\gamma_s|\boldsymbol{\lambda},\mathcal{D})$. The positive skewness is non-negligible for the marginal posterior distribution of γ_4: that asymmetry would not be captured by a simple Laplace approximation. It is caused by the small proportion of respondents in the survey expressing a strong disagreement with the submitted statement on the freedom of gays and lesbians.

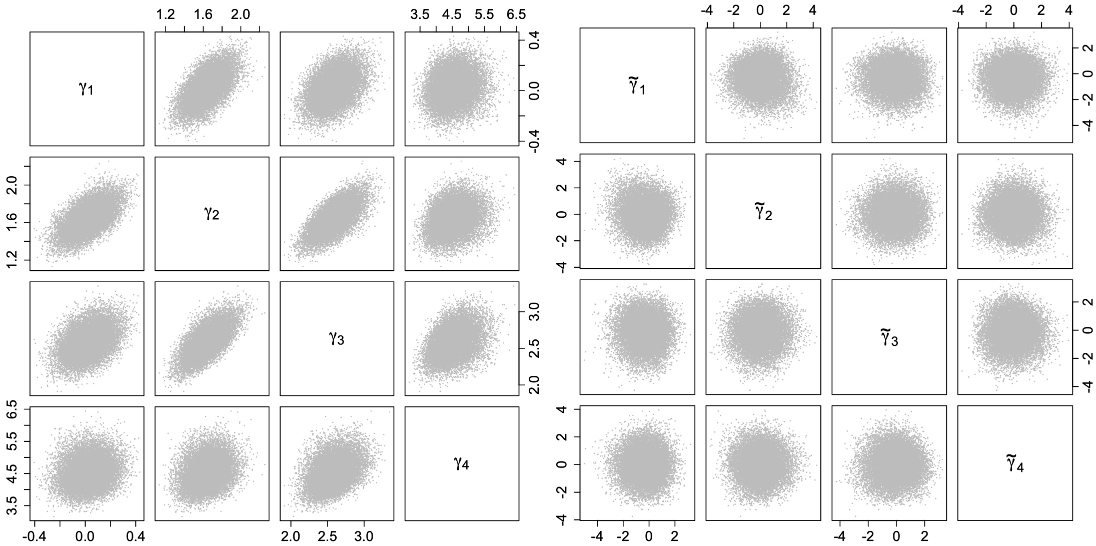

Figure 3: ESS dataset: scatterplots of the MCMC samples for $(\boldsymbol{\gamma}|\boldsymbol{\lambda}, \mathcal{D})$ and $(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D})$ when $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$.

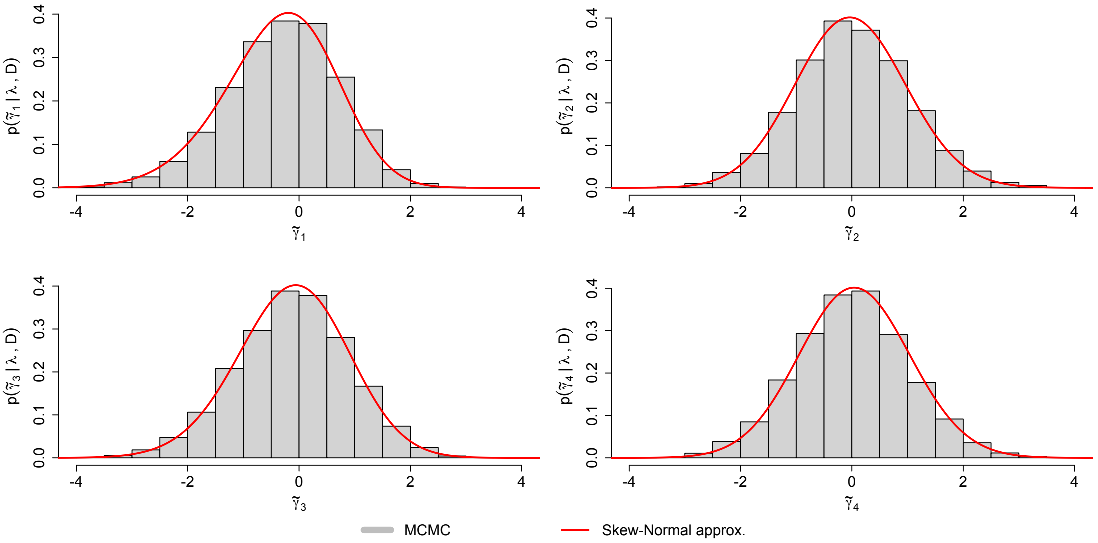

Figure 4: ESS dataset: approximated marginal posterior density for $(\tilde{\boldsymbol{\gamma}}|\boldsymbol{\lambda}, \mathcal{D})$ compared to MCMC samples when $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}$.

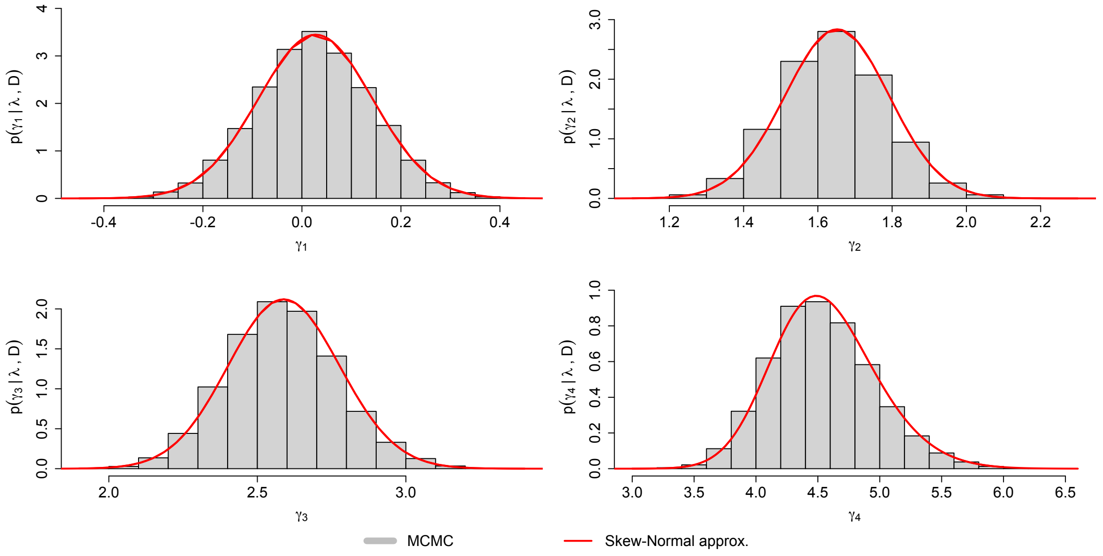

Figure 5: ESS dataset: approximated marginal posterior density for $(\boldsymbol{\gamma}|\boldsymbol{\lambda},\mathcal{D})$ compared to MCMC samples when $\boldsymbol{\lambda}=\hat{\boldsymbol{\lambda}}$.

4 Discussion

In this paper, the Laplace P-spline (LPS) approach has been extended to improve the accuracy of inference in a Bayesian framework. Indeed, when information is sparse, the posterior distribution of non-penalized parameters may exhibit a non-negligible skewness that can have adverse effects on inference or predictions when ignored. The proposed approximation to the joint posterior density in (2.9) takes a simple form that can be used in a much faster way than MCMC to make predictions or inference on functions of the model parameters.

An approximation to the marginal posterior distribution of the penalty parameters $\boldsymbol{\lambda}$ keeps playing an important role in the procedure. Point estimates for $\boldsymbol{\lambda}$ can be derived from it with a subsequent empirical Bayes approach (Carlin and Louis, 2000) to handle these hyperparameters, see Section 3 for an illustration. Alternatively, the uncertainty in the selection of $\boldsymbol{\lambda}$ could be accounted for by marginalizing over it with a Monte Carlo or a grid-based integration in $p(\boldsymbol{\xi}|\mathcal{D})=\int_{\lambda}p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})\,p(\boldsymbol{\lambda}|\mathcal{D})\,d\boldsymbol{\lambda}$. However, in the context of generalized additive models (Gressani and Lambert, 2021) and nonparametric double additive location-scale models (Lambert, 2021), simulation studies suggest that coverages of credible intervals resulting from an empirical Bayes approach for model parameter estimation are already close to their nominal values, even with moderate sample sizes.

The proposed methodology diverges from the proposal made by Rue et al. (2009) and underlying INLA where the size of the latent vector $\boldsymbol{\xi}$ increases with sample size and where the marginal distribution of the scalar components of $\boldsymbol{\xi}$ are the research focus.

The joint distribution of $(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$ with asymmetric forms for the non-penalized components is here available for the whole vector $\boldsymbol{\xi}$ and completed by an approximation to the marginal posterior distribution of the hyperparameters $\boldsymbol{\lambda}$.

The code necessary to reproduce the results in the paper can be downloaded from https://github.com/plambertULiege/ordgam.

This paper and, more broadly, our research on smoothing methods, owe much to Brian Marx, who left us too soon. His joint work with Paul Eilers will continue to endure and shape the field for years to come.

Acknowledgements

Philippe Lambert acknowledges the support of the ARC project IMAL (grant 20/25-107) financed by the Wallonia-Brussels Federation and granted by the Académie Universitaire Louvain.

A Gradient and Hessian in the PO model

Consider the proportional odds model defined in Section 3 and the notations therein. Let $v_{ir} = F_{ir}(1-F_{ir})$, $w_{ir} = (1+\pi_{ir}-2F_{ir})$, $z_{ir} = (1-2F_{ir})v_{ir}$ for r in $\mathcal{E}_R$ and take $s,t \in \mathcal{E}_{R-1}$. One has:

$$\begin{aligned}
\frac{\partial \ell}{\partial \gamma_s} &= \sum_i \frac{\partial \log \pi_{iy_i}}{\partial \gamma_s} \quad ; \quad \frac{\partial \log \pi_{ir}}{\partial \gamma_s} = \frac{1}{\pi_{ir}}(\delta_{rs}v_{ir} - \delta_{r-1,s}v_{i,r-1}) \\
\frac{\partial \ell}{\partial \theta_k} &= \sum_i \frac{\partial \log \pi_{iy_i}}{\partial \theta_k} \quad ; \quad \frac{\partial \log \pi_{ir}}{\partial \theta_k} = x_{ik}w_{ir}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial \gamma_s \partial \gamma_t} &= \sum_i \frac{1}{\pi_{iy_i}}\left\{\delta_{y_i,s,t}z_{i,y_i} - \delta_{y_i-1,s,t}z_{i,y_i-1}\right\} - \sum_i \frac{\partial \log \pi_{iy_i}}{\partial \gamma_s}\frac{\partial \log \pi_{iy_i}}{\partial \gamma_t} \ ; \\
\frac{\partial^2 \ell}{\partial^2 \theta_k \theta_\ell} &= \sum_i x_{ik}x_{i\ell}\left(\pi_{iy_i}w_{iy_i} - 2\sum_{j=1}^{y_i}\pi_{ij}w_{ij}\right) \ ; \\
\frac{\partial^2 \ell}{\partial \theta_k \partial \gamma_s} &= -\sum_i x_{ik}(\delta_{y_is}v_{iy_i} + \delta_{y_i-1,s}v_{i,y_i-1}) \ .
\end{aligned}$$

Therefore, given $\boldsymbol{\lambda}$, one has

$$\mathbf{U}_\lambda = \frac{\partial \log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})}{\partial \boldsymbol{\xi}} = \frac{\partial \ell}{\partial \boldsymbol{\xi}} - \mathbf{K}_\lambda(\boldsymbol{\xi} - \mathbf{e}) \ ; \ \mathbf{H}_\lambda = \frac{\partial^2 \log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^\top} = \frac{\partial^2 \ell}{\partial \boldsymbol{\xi}\partial \boldsymbol{\xi}^\top} - \mathbf{K}_\lambda \ .$$

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Azzalini, A. and Capitanio, A. (2014). *The Skew-Normal and Related Families*. New-York: Cambridge University Press.

Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, **50**, 967–991.

Carlin, B. and Louis, T. (2000). *Bayes and Empirical Bayes Methods for Data Analysis*. New-York: Chapman and Hall / CRC Press.

Chiuchiolo, C., van Niekerk, J., and Rue, H. (2022). Joint posterior inference for latent Gaussian models with R-INLA. *Journal of Statistical Computation and Simulation*. doi: 10.1080/00949655.2022.2117813.

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

Eilers, P. H. C. and Marx, B. D. (2021). *Practical Smoothing: The Joys of P-splines*. Cambridge University Press.

ESS Round 9 (2018). European Social Survey Round 9. Data file edition 3.1. Sikt - Norwegian Agency for shared services in education and research, Norway - Data Archive and distributor of ESS data for ESS ERIC. doi: 10.21338/NSD-ESS9-2018.

Ferkingstad, E. and Rue, H. (2015). Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electronic Journal of Statistics*, **9** (2), 2706–2731.

Gamerman, D. (1997). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing*, **7**, 57–68.

Gressani, O. and 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.

Gressani, O. and Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines. *Computational Statistics & Data Analysis*, **154**, 107088.

Gressani, O., Faes, C., and Hens, N. (2022a). An approximate Bayesian approach for estimation of the instantaneous reproduction number under misreported epidemic data. *Biometrical Journal*. doi: 10.1002/bimj.202200024.

Gressani, O., Faes, C., and Hens, N. (2022b). Laplacian-P-splines for Bayesian inference in the mixture cure model. *Statistics in Medicine*, **41**(14), 2602–2626.

Gressani, O., Wallinga, J., Althaus, C., Hens, N., and Faes, C. (2022c). EpiLPS: A fast and flexible Bayesian tool for estimation of the time-varying reproduction number. *PLoS Computational Biology*, **18**(10), 1010618.

Jullion, A. and Lambert, P. (2007). Robust specification of the roughness penalty prior distribution in spatially adaptive Bayesian P-splines models. *Computational Statistics and Data Analysis*, **51**(5), 2542–2558.

Lambert, P. (2021). Fast Bayesian inference using Laplace approximations in nonparametric double additive location-scale models with right- and interval-censored data. *Computational Statistics & Data Analysis*, **161**, 107250.

Lambert, P. and Bremhorst, V. (2019). Estimation and identification issues in the promotion time cure model when the same covariates influence long- and short-term survival. *Biometrical Journal*, **61**(2), 275–289.

Lambert, P. and Eilers, P. H. C. (2009). Bayesian density estimation from grouped continuous data. *Computational Statistics and Data Analysis*, **53**, 1388–1399.

Lambert, P. and Kreyenfeld, M. (2023). Exogenous time-varying covariates in double additive cure survival model with application to fertility. arXiv:2302.00331.

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

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie Royale des Sciences de Paris (Savants étrangers)*, **6**, 621–656.

Leonard, T. (1982). A simple predictive density function: Comment. *Journal of the American Statistical Association*, **77**(379), 657–658.

Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Applied Mathematics*, **11**, 431–441.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics and Data Analysis*, **28**, 193–209.

Piegorsch, W. W. (1990). Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*, **46**(3), 863–867.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, **2**, 341–363.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* Boca Raton: CRC Press.

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

Ruli, E., Sartori, N., and Ventura, L. (2016). Improved Laplace approximation for marginal likelihoods. *Electronic Journal of Statistics*, **10**(2), 3986–4009.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**(393), 82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84**(407), 710–716.

Van der Vaart, A. W. (1998). *Asymptotic statistics.* Cambridge University Press.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R (2nd ed.).* Boca Raton: CRC Press.

Wood, S. N. and Fasiolo, M. (2017). A generalized Fellner-Schall method for smoothing parameter optimization with application to Tweedie location, scale and shape models. *Biometrics*, **73**, 1071–1081.