# Laplace Approximations and Bayesian P-splines for Statistical Inference

**Oswaldo Gressani**

*A thesis submitted to the University of Louvain in partial fulfillment of the requirements for the degree of*

Doctor of Sciences


Thesis committee:

| | |
|---|---|
| **Philippe Lambert**, (Supervisor) | Université de Liège, Université catholique de Louvain |
| **Catherine Legrand** | Université catholique de Louvain |
| **Christian Ritter** | Université catholique de Louvain |
| **María Durbán** | Universidad Carlos III de Madrid |
| **Paul Eilers** | Erasmus Universiteit Rotterdam |

Louvain-la-Neuve, October 2020

Je suis de ceux qui pensent que la science a une grande beauté. [...] Je ne crois pas [...] que dans notre monde, l'esprit d'aventure risque de disparaître. Si je vois autour de moi quelque chose de vital, c'est précisément l'esprit d'aventure qui paraît indéracinable et s'apparente à la curiosité.

---

Marie Curie, *Discours sur l'avenir de la culture*, Madrid 1933.[1]

[1] Bibliothèque Nationale de France. Département des Manuscrits. https://gallica.bnf.fr/ark:/12148/btv1b9080310x/f29.item

# Acknowledgments

En débutant ce projet il y a environ cinq ans, j'étais loin de me douter que j'allais embarquer pour un voyage des plus enrichissants autant d'un point de vue scientifique que personnel. Afin de prendre du recul, je m'imagine tel un projectionniste, qui, émerveillé devant la dynamique de l'animation, fait défiler sur une surface immaculée ces années de recherche à l'institut de statistique au cours desquelles j'ai eu la chance d'aiguiser mes connaissances. Durant cette aventure formidable, l'Université catholique de Louvain, que je peux désormais considérer fièrement comme mon *alma mater*, aura été un lieu catalyseur d'idées où j'ai rencontré un grand nombre de personnes érudites qui ont joué un rôle plus ou moins important dans le développement de ce doctorat. J'espère que ces quelques lignes pourront expliciter à leur juste degré ma profonde gratitude et reconnaissance envers celles et ceux qui ont contribué à l'aboutissement de cette thèse.

Je tiens tout d'abord à adresser mes remerciements à mon superviseur de thèse, le Professeur Philippe Lambert pour m'avoir fait confiance tout au long de mon parcours. Merci infiniment de m'avoir initié au raisonnement bayésien qui m'a fait découvir à quel point l'horizon de recherche dans le domaine de la statistique pouvait être fascinant. J'ai fortement apprécié l'atmosphère de bonne humeur ainsi que le continuel enthousiasme qui était omniprésent lors de nos échanges. J'aimerais également souligner toute ma gratitude pour le temps que vous avez investi dans ce projet (les sujets abordés lors de nos rencontres étaient si captivants que je ne voyais pas les heures passer). Il va sans dire que l'apport de votre expertise et vos précieux conseils ont permis de me forger une base forte sur le plan scientifique et ont largement contribué à l'accomplissement de ce projet.

Ensuite, j'aimerais exprimer une chaleureuse reconnaissance à tous les autres membres du jury qui ont porté un intérêt à ce travail. La version finale de ce manuscrit a amplement bénéficié de leurs commentaires pendant la défense privée. Plus particulièrement, je remercie la présidente du jury et Professeure Catherine Legrand pour ses conseils éclairés. Un grand merci aussi au Professeur Christian Ritter pour sa perspective générale de la méthodologie sous-jacente à cette thèse. I would like to wholeheartedly thank Professor María Durbán from the Universidad Carlos III de Madrid who kindly accepted to be a member of the jury. I am indebted to her for her fruitful advice and constructive comments, especially regarding P-spline smoothers. I owe a similar debt of gratitude to Professor Paul Eilers of the Erasmus Universiteit Rotterdam for his encouragement and support. Thank you for having taken the time to test the different routines of the blapsr package and for having provided short and clear examples.

Je voudrais aussi évoquer le cadre accueillant qu'a été l'ISBA au cours de ma formation doctorale. De mémoire, il me semble qu'une ambiance chaleureuse a toujours été au rendez-vous, rythmée par divers événements festifs, comme la fameuse « crêpes-party » dont je garde un très bon souvenir. Merci beaucoup à toute l'équipe administrative : Nadja, Nancy, Maguy, Sophie et Tatiana pour votre travail exemplaire. Un immense merci à tous mes collègues et aux membres de l'ISBA qui ont partagé mon quotidien. Une pensée particulière pour mon ami Gautier Attanasi avec qui j'ai partagé un nombre incalculable de litres de café pendant les pauses de midi.

Quelques mots pour ma famille et en particulier mes parents qui m'ont inlassablement soutenu tout au long de mon parcours. Merci à mes grands-parents, à mes oncles et tantes pour leurs mots d'encouragement. Par rapport à ma recherche ils me demandaient souvent « Est-ce que tu as trouvé ? ». J'espère que ce manuscrit leur apportera au moins une réponse partielle. Finalement, je tiens à terminer en exprimant ma plus sincère gratitude à Sandra Rodrigues. Tout le temps à mes côtés, tu as été une source incommensurable de motivation qui m'a aidé à aller au bout de ce projet. Merci du fond du coeur de m'avoir tant donné et d'avoir confiance en moi pour l'avenir.

Oswaldo Gressani, Octobre 2020

# Contents

# Figures

# Tables

# Notation and Abbreviations

## List of symbols

| | |
|---|---|
| $\mathbb{N}$ | Set of natural numbers ($0 \in \mathbb{N}$) |
| $\mathbb{N}_+$ | Set of positive natural numbers ($0 \notin \mathbb{N}_+$) |
| $\mathbb{R}$ | Set of real numbers |
| $\mathbb{R}_+$ | Set of nonnegative real numbers ($0 \in \mathbb{R}_+$) |
| $\mathbb{R}_{++}$ | Set of positive real numbers ($0 \notin \mathbb{R}_{++}$) |
| $\otimes$ | Kronecker product |
| $\sim$ | Distributed as |
| $\circ$ | Hadamard product |
| $\propto$ | Equality up to a multiplicative constant |
| $\doteq$ | Equality up to an additive constant |
| $\text{sign}(\cdot)$ | Sign of a real number |
| $\text{Tr}(\cdot)$ | Trace of a matrix |
| $\text{rank}(\cdot)$ | Rank of a matrix |
| $\text{adj}(\cdot)$ | Adjoint of a matrix |
| $\partial/\partial y$ | Partial derivative with respect to $y$ |
| $\mathbf{1}_n$ | $n$-dimensional vector of ones |
| $\dim(\cdot)$ | Dimension of a vector |
| $A^\top$ | Transpose of matrix A |
| $A^{-1}$ | Inverse of matrix A |
| $\Gamma(\cdot)$ | Gamma function |
| $I_m$ | Identity matrix of dimension $m$ |
| $\sigma$ | Standard deviation |
| $\nabla f$ | Gradient of a scalar function $f$ |
| $\nabla^2 f$ | Hessian of a scalar function $f$ |
| $\mathcal{D}$ | Information set or observable data |
| $\|\cdot\|$ | Euclidean norm |

| | |
|---|---|
| $\mathcal{L}(\cdot, \mathcal{D})$ | Likelihood function |
| $\ell(\cdot, \mathcal{D})$ | Log-likelihood function |
| $\text{diag}(v_1, \ldots, v_n)$ | Diagonal matrix with elements $v_1, \ldots, v_n$ on the main diagonal |
| $\varphi(\cdot)$ | Density function of a standard Normal random variable |
| $\Phi(\cdot)$ | Cumulative distribution function of a standard Normal random variable |
| $\mathbb{I}(\cdot)$ | Indicator function |
| $\perp$ | Independence of random variables |
| $\approx$ | Approximately equal to |
| $\mathcal{N}_q(\mu, \Sigma)$ | $q$-variate Gaussian distribution with mean vector $\mu$ and variance-covariance matrix $\Sigma$ |
| $\mathcal{G}(a, b)$ | Gamma distribution with mean $a/b$ and variance $a/b^2$ |
| $\text{Bin}(n, p)$ | Binomial distribution with $n \in \mathbb{N}_+$ and $p \in (0, 1)$ |
| $\text{Beta}(a, b)$ | Beta distribution with $a > 0$ and $b > 0$ |
| $\text{Bern}(p)$ | Bernoulli distribution with $p \in (0, 1)$ |
| $\text{Weibull}(a, b)$ | Weibull distribution with shape $a > 0$ and scale $b > 0$ |
| $\text{Poisson}(\mu)$ | Poisson distribution with rate $\mu > 0$ |
| $\text{SN}(\cdot, \cdot, \cdot)$ | Skew-normal distribution |
| $\mathcal{U}(a, b)$ | Continuous uniform distribution on $(a, b)$ |
| $t_\vartheta(\cdot, \cdot)$ | Student-$t$ distribution with $\vartheta > 0$ degrees of freedom |

## List of abbreviations

| | |
|---|---|
| GAM(s) | Generalized additive model(s) |
| INLA | Integrated Nested Laplace Approximations |
| i.i.d. | Independent and identically distributed |
| LGM(s) | Latent Gaussian model(s) |
| LPS | Laplace-P-spline(s) |
| MCMC | Markov chain Monte Carlo |

# Introduction

Uncertainty and random events are two unavoidable facets of our daily life. Whether one is interested in the laws of physics governing the universe, economic decision making, biology or politics, the concept of a random phenomenon serves as a crucial building block for establishing a theory. Statistical science provides useful tools to understand and formalize randomness by means of probabilistic models. The scientific method in statistics can be implemented by following two philosophically different paths. From the frequentist or classical perspective, probability is defined as a relative frequency resulting from a theoretical infinite number of iterations of a random experience. In the present work, we take the other path based on the Bayesian paradigm in which probability has a subjective interpretation and is perceived as a degree of belief.

The contribution of this thesis is built upon the combination of Laplace's method and penalized regression splines. Laplace approximations to selected posterior distributions enables to bypass Markov chain Monte Carlo (MCMC) methods and yields accurate inferences at a low computational budget in a large class of models. Penalized regression splines is a popular and well established technique for curve fitting and allows flexible and smooth estimation of unknown regression functions. We build a novel approach that combines Laplace's method and penalized splines for fast approximate Bayesian inference in latent Gaussian models. The resulting "Laplace-P-spline" methodology (LPS) is developed within the framework of survival analysis and (generalized) additive models. A software package is developed in the **R** language and made available in a public repository. The thesis is organized in six chapters summarized below.

### Chapter 1

The objective of this chapter is to provide enough background material to familiarize the reader with Laplace approximations and penalized regression splines, the two fundamental concepts of the thesis. First, emphasis is placed on Laplace's method and its role as a posterior approximation scheme in a Bayesian setting. In addition to illustrative examples of posterior approximations, we also discuss in detail the importance of nested approximations in the Laplace-P-spline theory. Second, we provide the mathematical formulation of B-splines and their penalized version in both frequentist and Bayesian settings. The chapter ends by presenting a way of unifying Laplace's method and penalized splines in a Cox proportional hazards model which serves as smooth preliminary material to the next chapter. A small simulation setting highlights the computational benefits of our approach and encourages further extension of the Laplace-P-spline methodology.

### Chapter 2

The chapter borrows its content from the paper: **Fast Bayesian inference using Laplace approximations in a flexible promotion time cure model based on P-splines**[1] published in *Computational Statistics and Data Analysis*, August 2018, Volume 124, Pages 151-167 (Gressani and Lambert, 2018). The article is accessible at https://doi.org/10.1016/j.csda.2018.02.007. We develop a sampling-free approximate Bayesian inference methodology for fast inference in a promotion time cure model where an unknown fraction of cured subjects will never experience the event of interest. The LPS promotion time model goes beyond univariate marginal analysis by providing approximate joint posteriors for the spline and regression parameters and pointwise credible intervals even for relatively complicated functions of latent variables. The approximated multivariate posterior of the spline and regression coefficients is expressed as a finite mixture of Gaussian densities for which the mean and covariance matrix are available. A simulation study shows that our method performs well in different cure and censoring scenarios. The chapter ends with two real applications on malignant melanoma and oropharynx carcinoma data.

[1] 

### Chapter 3

The third chapter is inspired from the discussion paper: **The Laplace-P-spline methodology for fast approximate Bayesian inference in additive partial linear models** (Gressani and Lambert, 2020*a*). The aim is to develop the LPS methodology in the class of additive models, where the limiting assumption of a linear regression function is replaced by a sum of smooth functions of individual covariates. Analytic formulas for the gradient and Hessian of the posterior penalty vector are derived and used to construct an efficient algorithm for exploring the penalty space. When the number of smooth functions in the model is small to moderate, exploration of the posterior penalty domain relies on a moment-matching method based on the skew-normal family of distributions. In larger dimensions optimal smoothing is determined by the posterior mode of the penalty vector. We also address the construction of approximate quantile-based credible intervals for the vector of spline and regression parameters and credible intervals for smooth functions. The performance of our approach is assessed using a simulation study.

### Chapter 4

The fourth chapter is articulated around ideas found in the article: **Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines** [2] published in *Computational Statistics and Data Analysis*, February 2021, Volume 154 (Gressani and Lambert, 2021) accessible at https://doi.org/10.1016/j.csda.2020.107088. It is a generalization of Chapter 3 where Bayesian P-splines and Laplace approximations are coupled for inference in generalized additive models (GAMs). Our LPS-GAM is endowed with analytical forms for the gradient and Hessian of the posterior penalty vector and, hence, does not require numerical differentiation for inference. The main strength of our approach resides in the fast algorithm to estimate the model and the ability to accommodate any number of smooth terms. Furthermore, simulation results reveal good statistical perfomance and proves that our methodology is competitive against a widely used benchmark method. Finally, the LPS method is illustrated on two real datasets.

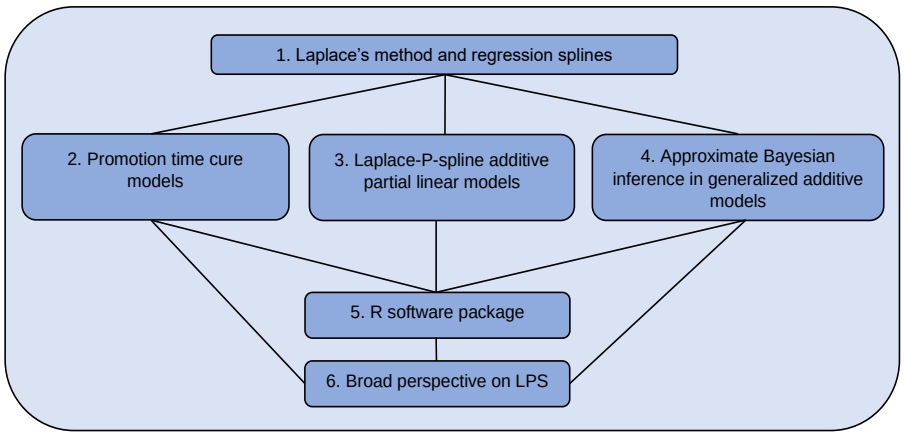
[2] 

### Chapter 5

This chapter is dedicated to the **blapsr** package (Gressani and Lambert, 2020*b*) written in the **R** language, a software created during the PhD project that implements Bayesian approximate inference in survival models and (generalized) additive models based on the Laplace-P-spline methodology presented in the previous chapters. A stable version of the package is on CRAN (`https://cran.r-project.org/package=blapsr`) and an in-development version can be found on GitHub (`https://github.com/oswaldogressani/blapsr`). A dedicated website is also available at `https://www.blapsr-project.org/`. Four main routines are presented that can be used to fit the classic Cox model and the promotion time cure model (for right censored survival data) as well as additive partial linear models and generalized additive models. The **R** functions are illustrated on simulated data and on real data examples.

### Chapter 6

The thesis concludes with a chapter that aims at giving the reader a broad perspective of the LPS methodology with additional complementary ideas. In particular, a general recipe explaining the key steps to implement LPS in a generic Bayesian setting is presented. The strengths and weaknesses of LPS are also summarized. Finally, several future research directions are proposed that open up new horizons for Laplace-P-splines.




Thesis outline.

# CHAPTER 1

# Laplace approximations and P-splines

## 1.1 Motivation

The present chapter aims at providing the reader with a broad conception of the methodological developments and challenges addressed in this thesis. The purpose is to draw the foundational aspects and principal ideas involved in subsequent chapters, as well as an outline of the difficulties to be surmounted by emphasizing on the benefits and limitations of the proposed methodology. We begin by explaining the role and use of the Laplace approximation scheme in a Bayesian framework and describe how it can be used to obtain a rapid and sampling-free tool for approximate inference in a general class of models. Penalized regression splines are another important facet of the thesis. Particular focus is placed on B-splines which are used to approximate an unknown regression function by specifying a linear combination of a chosen basis with a difference penalty on adjacent spline coefficients to prevent overfitting.

We propose to exploit the synergy between Laplace's method for fast posterior approximations and P-splines for flexible nonparametric modeling, giving birth to the "Laplace-P-spline" (LPS) model. The chapter

is structured as follows. Section 1.2 introduces the Laplace approximation and its role in a Bayesian framework. In addition, the concept of nested approximations is presented along with an illustration in a simple univariate model. Section 1.3 aims at familiarizing the reader with B-splines and their penalized version, both in a frequentist and Bayesian setting. In Section 1.4 the central idea of the thesis, namely the unification of Laplace's method and P-splines is presented and illustrated in a Cox proportional hazards model. Section 1.5 concludes the chapter.

## 1.2 The Laplace approximation

### 1.2.1 Historical note

At the forefront of Enlightenment thinkers, the French mathematician Pierre-Simon de Laplace (1749-1827) is recognized for his influential role in the development of probability theory and mathematical statistics. Although it is well known that Laplace contributed to build the theoretical background of least squares together with Legendre and Gauss (Plackett, 1949; Stigler, 1981), the idea that Laplace's work on inverse probability was largely responsible for disseminating the Bayesian paradigm is less widespread in the scientific community. Thus, the Bayesian method cannot be solely credited to Thomas Bayes, as Hogben (1968, p. 133) puts it: "The *fons et origo* of inverse probability is Laplace. For good or ill, the ideas commonly identified with the name of Bayes are largely his."

Among the vast scientific breakthroughs proposed by Laplace, the methodology developed in this thesis is largely based on the so-called Laplacian method of approximation introduced in his *Mémoire sur la probabilité des causes* (Laplace, 1774), see also Laplace (1986, pp. 366-367) for the English version. In the latter work, Laplace proposes a technique to approximate integrals whose integrand term has a single sharp peak (or mode) at a point $x_0$, implying that the entire integral can be well approximated by integration around a small neighborhood of $x_0$. The key idea behind Laplace's method is to implement a Taylor series expansion of the logarithm of the integrand around the mode (Azevedo-Filho and Shachter, 1994). This expansion will result in a term that shares similarities with a Gaussian distribution and hence can be integrated analytically.

The 1774 seminal work of Laplace served as a building block for other significant mathematical contributions in the nineteenth century and his approximation technique had an important impact in a wide variety of disciplines. The rebirth of Bayesian statistics after the Second World War, influenced by Leonard Savage in his book *The foundations of Statistics* (Savage, 1954), and the advent of computer age statistical inference triggered a revived interest in Laplace's method.

The resurrection of Laplace approximations in a Bayesian context is primarily attributed to Lindley (1961) and Mosteller and Wallace (1964), who use the method to approximate ratios of integrals. Later, Leonard (1982) proposes to approximate the denominator of a predictive distribution by using Laplace's method and Tierney and Kadane (1986) apply the ideas of Laplace to approximate posterior moments and marginal densities (see also Tierney et al., 1989).

During the first decade of the twenty-first century, a growing literature on Laplace approximations emerged along with the idea that it could be considered a serious challenger to existing Markov chain Monte Carlo (MCMC) methods, the dominant strategy in Bayesian analysis to characterize posterior distributions and perform statistical inference. These MCMC techniques are often plagued by several potential issues such as high posterior correlation between parameters, slow chain convergence, and foremost a strong computational cost. In an attempt to overcome the drawbacks inherent to MCMC sampling algorithms, an approximate Bayesian inference scheme coined Integrated Nested Laplace Approximations (INLA) has been proposed by Rue et al. (2009) to infer in a large subclass of structured additive regression models. The main advantage of this methodology is that accurate approximations of posterior marginals can be computed at a low computational budget.

### 1.2.2 Laplace approximation to the posterior

In the Bayesian paradigm, model parameters are treated as random variables and probability statements used to quantify parameter uncertainty should be interpreted as a degree of belief. Before data is collected, a subjective belief on unknown parameter values is formalized into a prior. After gathering the data, Bayes' theorem is used to update the initial beliefs by coupling the prior and the observed data to obtain a posterior

distribution. Let $\mathcal{D}$ denote the observed data and $\theta$ a scalar parameter of interest. Bayes' rule allows to write the posterior density as:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})},$$

and by omitting the normalizing constant $p(\mathcal{D}) = \int p(\mathcal{D}|\theta)p(\theta)d\theta$, usually called *the evidence* or *marginal likelihood*, we can write $p(\theta|\mathcal{D}) \propto p(\mathcal{D}|\theta)p(\theta)$, i.e. the posterior is proportional to the likelihood $p(\mathcal{D}|\theta)$ times the prior $p(\theta)$. In most situations, the posterior is complicated and hard to handle analytically. The ideas of Laplace can be used to approximate a complex posterior by computing a second-order Taylor expansion of the log-posterior, $\log p(\theta|\mathcal{D})$, around its posterior mode $\hat{\theta}$:

$$\begin{aligned}\log p(\theta|\mathcal{D}) \quad \approx \quad & \log p(\hat{\theta}|\mathcal{D}) + \frac{\partial \log p(\theta|\mathcal{D})}{\partial \theta}\bigg|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) \\ & + \frac{1}{2} \frac{\partial^2 \log p(\theta|\mathcal{D})}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2.\end{aligned}$$

Note that the first derivative of the log-posterior evaluated at the mode is equal to zero and can thus be discarded yielding:

$$\log p(\theta|\mathcal{D}) \approx \log p(\hat{\theta}|\mathcal{D}) - \frac{\tau}{2}(\theta - \hat{\theta})^2, \tag{1.1}$$

where $\tau = -(\partial^2 \log p(\theta|\mathcal{D})/\partial\theta^2)|_{\theta=\hat{\theta}}$. Recall that the logarithm of a Gaussian density for $\theta$ with mean $\mu$ and variance $\sigma^2$ is given by:

$$C - \frac{1}{2\sigma^2}(\theta - \mu)^2, \tag{1.2}$$

where $C$ is a normalization constant ensuring that the Gaussian density integrates to one. From (1.2), one recognizes that (1.1) is the Laplace approximation to $p(\theta|\mathcal{D})$ with mean $\mu = \hat{\theta}$ and variance equal to the inverse of the negative of the curvature of the posterior at the mode, i.e. $\sigma^2 = \tau^{-1}$.

To illustrate Laplace's method, assume that the posterior to be approximated follows a Maxwell-Boltzmann distribution (see Papoulis and Pillai, 2002, p. 26 and p. 149), a well-known density in the kinetic theory of gases given by:

$$p(\theta|\mathcal{D}) = \begin{cases} \sqrt{\frac{2}{\pi}} \frac{\theta^2 \exp\left(-\frac{\theta^2}{2a^2}\right)}{a^3} & \text{for } \theta \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

with $a > 0$ and mode $\hat{\theta} = \sqrt{2}a$. The log-posterior and its first and second derivatives are:

$$\begin{aligned} \log p(\theta|\mathcal{D}) &= \frac{1}{2}\log\left(\frac{2}{\pi}\right) - \log(a^3) + 2\log(\theta) - \left(\frac{\theta^2}{2a^2}\right), \\ \frac{\partial \log p(\theta|\mathcal{D})}{\partial \theta} &= \frac{2}{\theta} - \frac{\theta}{a^2}, \\ \frac{\partial^2 \log p(\theta|\mathcal{D})}{\partial \theta^2} &= -\left(\frac{2}{\theta^2} + \frac{1}{a^2}\right). \end{aligned}$$

Accordingly, the Laplace approximation to the posterior is a Gaussian with mean $\hat{\theta}$ and variance given by $\left((2/\hat{\theta}^2) + (1/a^2)\right)^{-1} = a^2/2$.

Laplace's method has its limitations and is not appropriate when the posterior probability distribution is not tightly concentrated around the mode. Consider for instance that the posterior has a Kumaraswamy distribution (see Michalowicz et al., 2013, p. 99), characterized by the following probability density function:

$$p(\theta|\mathcal{D}) = \begin{cases} ab\theta^{(a-1)}(1-\theta^a)^{(b-1)} & \text{for } 0 \leq \theta \leq 1 \\ 0 & \text{otherwise,} \end{cases}$$

with shape parameters $a > 0$ and $b > 0$. The mode is known to be $\hat{\theta} = ((a-1)/(ab-1))^{(1/a)}$. The log-posterior and the first and second derivatives are:

$$\begin{aligned} \log p(\theta|\mathcal{D}) &= \log(a) + \log(b) + (a-1)\log(\theta) + (b-1)\log(1-\theta^a), \\ \frac{\partial \log p(\theta|\mathcal{D})}{\partial \theta} &= \frac{a-1}{\theta} - a(b-1)\frac{\theta^{(a-1)}}{1-\theta^a}, \\ \frac{\partial^2 \log p(\theta|\mathcal{D})}{\partial \theta^2} &= -\frac{a-1}{\theta^2} - a(b-1)\frac{(a-1)\theta^{(a-2)}(1-\theta^a) + a\theta^{2(a-1)}}{(1-\theta^a)^2}. \end{aligned}$$

The Laplace approximation to the posterior will be a Gaussian centered around $\hat{\theta}$ with variance equal to $(-\partial^2 \log p(\theta|\mathcal{D})/\partial\theta^2)^{-1}$ evaluated at $\hat{\theta}$. Figure 1.1 illustrates the Laplace approximation in the Maxwell-Boltzmann and Kumaraswamy scenarios. In the left panel, Laplace's method works well as the target to be approximated has most of the posterior mass concentrated around the mode. In the right panel a Gaussian approximation is less appropriate.

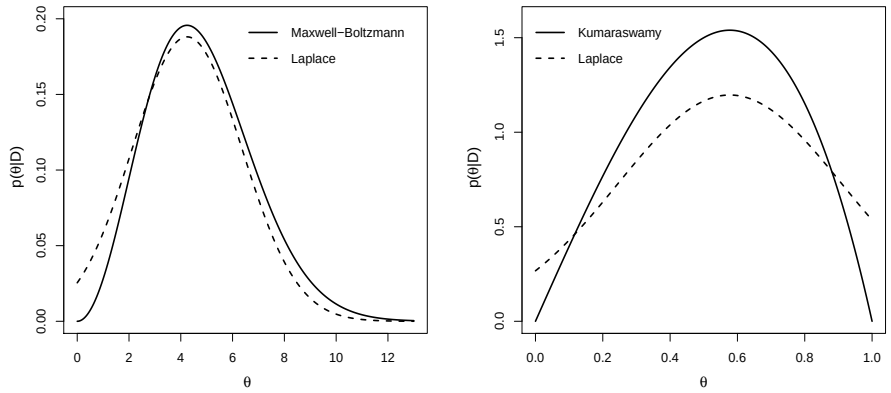



Figure 1.1: Laplace approximation (dashed curve) of the Maxwell-Boltzmann distribution (left) with $a = 3$ and of the Kumaraswamy distribution (right) with $a = 2$ and $b = 2$.

### 1.2.3 Ideas behind nested approximations

The Laplace approximation scheme can be used to explain the elegant ideas behind INLA. Assume for simplicity that a Bayesian model incorporates a single one-dimensional parameter $\theta$ for which we seek the posterior distribution. Furthermore, let $\eta$ be a nuisance, i.e. a parameter playing a crucial role in the modeling process but for which we have no direct interest. The posterior of $\theta$ can be obtained by solving the following integral:

$$\begin{aligned} p(\theta|\mathcal{D}) &= \int p(\theta, \eta|\mathcal{D})\, d\eta \\ &= \int p(\theta|\eta, \mathcal{D})\, p(\eta|\mathcal{D})\, d\eta. \end{aligned} \tag{1.3}$$

Laplace's method is used to approximate the conditional posterior in the above integrand $p(\theta|\eta, \mathcal{D})$ by $\widetilde{p}_G(\theta|\eta, \mathcal{D})$ and the latter Gaussian

expression is, in turn, nested in the approximating candidate for the posterior of the hyperparameter:

$$\widetilde{p}(\eta|\mathcal{D}) = \frac{p(\theta,\eta|\mathcal{D})}{\widetilde{p}_G(\theta|\eta,\mathcal{D})}\bigg|_{\theta=\hat{\theta}(\eta)},$$

where $\hat{\theta}(\eta)$ is the posterior mode of the conditional $p(\theta|\eta,\mathcal{D})$ and depends on $\eta$. After an appropriate choice of quadrature points $\{\eta^{(m)}\}$ in the domain of the nuisance posterior, (1.3) can be approximated by numerical integration resulting in the following expression which is exclusively a function of $\theta$:

$$\widetilde{p}(\theta|\mathcal{D}) = \sum_m \widetilde{p}_G(\theta|\eta^{(m)},\mathcal{D})\ \widetilde{p}(\eta^{(m)}|\mathcal{D})\ \Delta_m,$$

with quadrature weight $\Delta_m$. Once the above expression is computed, desired posterior moments and Bayesian credible intervals can be easily obtained as the posterior is fully characterized.

This nested approximation procedure, although simple at first sight comes with two major challenges. First, the Laplace approximation to the conditional posterior $p(\theta|\eta,\mathcal{D})$ requires to compute a second-order Taylor expansion of an expression that may have a certain degree of complexity. Hence, obtaining the gradient and Hessian usually requires an analytical effort. Second, a strategy has to be implemented to efficiently explore the approximated nuisance posterior. This entails among others, finding the posterior mode via an algorithm and choose an appropriate approach to select the points in the domain that captures most of the posterior mass.

### 1.2.4 Illustration of nested approximations

To show how nested approximations can be used to approximate a posterior $p(\theta|\mathcal{D})$, assume that $\mathcal{D} = (y_1,\ldots,y_n)$ is a sample of $n$ i.i.d. realizations from a Gumbel distribution with unknown location $\theta \in \mathbb{R}$ and known scale parameter $\beta > 0$, i.e. the contribution of the $i$th observation to the likelihood $\mathcal{L}(\theta;\mathcal{D})$ is $p(Y_i = y_i|\theta) = (1/\beta)\exp\left(-z_\theta - \exp(-z_\theta)\right)$, where $z_\theta = (y_i - \theta)/\beta$. We impose a Gaussian prior on $\theta$ with zero mean and variance $\eta^{-1}$, i.e. $p(\theta|\eta) = \sqrt{\eta/(2\pi)}\exp(-\eta\theta^2/2)$. Also, a Gamma prior with mean $a/b$ and variance $a/b^2$ is specified on the nuisance $\eta$,

namely $p(\eta) \propto \eta^{(a-1)} \exp(-b\eta)$. The conditional posterior of $\theta$ is:

$$\begin{aligned} p(\theta|\eta, \mathcal{D}) &\propto \mathcal{L}(\theta; \mathcal{D})\ p(\theta|\eta) \\ &\propto \exp\left(-\sum_{i=1}^{n}\left(\frac{(y_i-\theta)}{\beta} + \exp\left(-\frac{(y_i-\theta)}{\beta}\right)\right) - \frac{\eta}{2}\theta^2\right), \end{aligned}$$

with log-posterior, first and second derivatives given by:

$$\begin{aligned} \log p(\theta|\eta, \mathcal{D}) &\doteq -\sum_{i=1}^{n}\left(\frac{(y_i-\theta)}{\beta} + \exp\left(-\frac{(y_i-\theta)}{\beta}\right)\right) - \frac{\eta}{2}\theta^2, \\ \frac{\partial \log p(\theta|\eta, \mathcal{D})}{\partial\theta} &= \frac{1}{\beta}\left(n - \sum_{i=1}^{n}\exp\left(-\frac{(y_i-\theta)}{\beta}\right)\right) - \eta\theta, \\ \frac{\partial^2 \log p(\theta|\eta, \mathcal{D})}{\partial\theta^2} &= -\left(\frac{1}{\beta^2}\sum_{i=1}^{n}\exp\left(-\frac{(y_i-\theta)}{\beta}\right) + \eta\right). \end{aligned}$$

In this example, the mode of the conditional posterior is not analytically available but can be numerically obtained by solving the equation $n - \eta\beta\theta = \sum_{i=1}^{n}\exp\left(-(y_i-\theta)/\beta\right)$ for $\theta$. To obtain the variance of the Laplace approximation, we simply evaluate $(-\partial^2 \log p(\theta|\eta, \mathcal{D})/\partial\theta^2)^{-1}$ at the mode. Note that the mode and the variance will both depend on the value taken by $\eta$, so that the Laplace approximation to the conditional posterior can be written (by abuse of notation) as $\widetilde{p}_G(\theta|\eta, \mathcal{D}) = \mathcal{N}\left(\hat{\theta}(\eta), \sigma^2(\eta)\right)$. The approximate marginal posterior of the hyperparameter is computed as follows:

$$\begin{aligned} \widetilde{p}(\eta|\mathcal{D}) &\propto \left.\frac{\mathcal{L}(\theta; \mathcal{D})\ p(\theta|\eta)\ p(\eta)}{\widetilde{p}_G(\theta|\eta, \mathcal{D})}\right|_{\theta=\hat{\theta}(\eta)} \\ &\propto \sqrt{\eta}\ \sigma(\eta)\ \eta^{(a-1)} \exp(-b\eta) \exp\left(-\sum_{i=1}^{n}\left(\frac{(y_i-\hat{\theta}(\eta))}{\beta}\right.\right. \\ &\quad \left.\left. + \exp\left(-\frac{(y_i-\hat{\theta}(\eta))}{\beta}\right)\right) - \frac{\eta}{2}\hat{\theta}^2(\eta)\right). \end{aligned}$$

Figure 1.2, illustrates nested approximations when a sample of size $n = 50$ is generated from a Gumbel distribution with location $\theta = 1.5$ and scale $\beta = 5$. Parameters related to the Gamma prior are fixed to $a = 2$, $b = 2$. In Figure 1.2 (a), the conditional posterior of $\theta$

given $\eta = 0.2$ is shown together with its corresponding Laplace approximation. Figure 1.2 (b) shows the approximated hyperparameter posterior and the vertical ticks along the $x$-axis correspond to the arbitrary equidistant grid $\aleph_\eta = \{0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5, 1.7, 1.9\}$ of size $M = 10$ chosen to explore the posterior. Defining the weights $\omega_m = \left(\widetilde{p}(\eta^{(m)}|\mathcal{D})\Delta_m\right) / \left(\sum_{m=1}^{M} \widetilde{p}(\eta^{(m)}|\mathcal{D})\Delta_m\right)$ for $m = 1, \ldots, M$, with grid width $\Delta_m = 0.2$, the approximate posterior is given by the finite mixture $\widetilde{p}(\theta|\mathcal{D}) = \sum_{m=1}^{M} \omega_m \; \widetilde{p}_G(\theta|\eta^{(m)}, \mathcal{D})$ as illustrated by the solid curve in Figure 1.2 (d), which results from a sum of weighted Laplace approximations (dashed) computed for the hyperparameter values in $\aleph_\eta$. Taking the posterior mean $\hat{\theta} = \sum_{m=1}^{M} \omega_m \hat{\theta}\left(\eta^{(m)}\right)$ as a point estimate for $\theta$, one obtains $\hat{\theta} = 1.445$ with quantile-based 95% credible interval $\text{CI}^{\theta}_{95\%} = [0.156; 2.804]$.

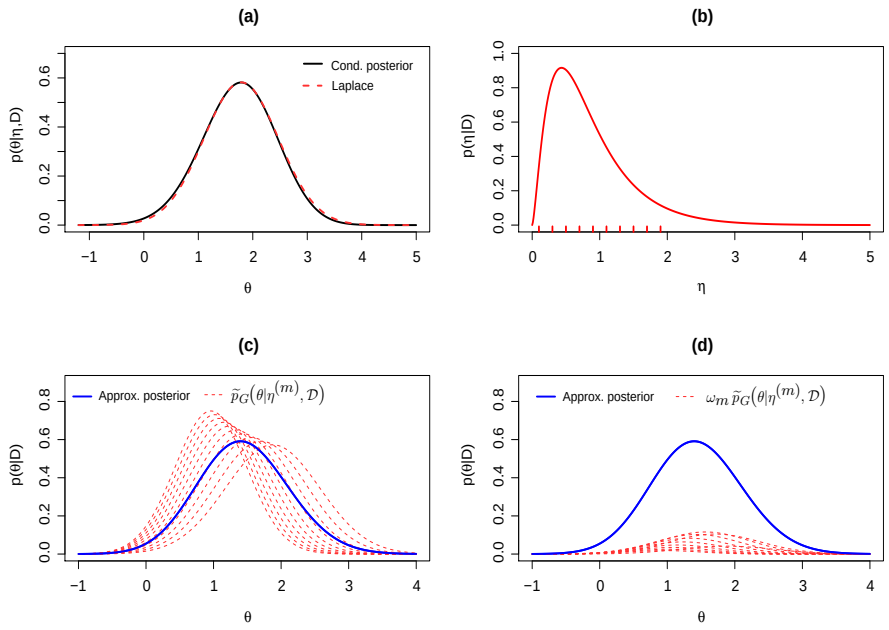



Figure 1.2: Illustration of nested approximations. Graph (a) corresponds to the conditional posterior of $\theta$ (solid) for $\eta = 0.2$ and its associated Laplace approximation (dashed). Graph (b) shows the approximated hyperparameter posterior and an arbitrary chosen grid. Graphs (c) and (d) show the approximated posterior for $\theta$ (solid) with the unweighted and weighted Laplace approximation terms (dashed) respectively.

## 1.3 P-splines

### 1.3.1 Origin of (B-)splines

The Oxford dictionary of statistics (Upton and Cook, 2014, p. 404), defines a spline as: "A set of polynomials, one for each sub-interval, that give an approximation to the function $f(x)$, ...". From a less mathematical perspective, the term "spline" is tightly bound to the activity of draftsmen in the automobile, aircraft and shipbuilding industry. Before the arrival of modern computer technology, a spline was used as a tool that consisted of a flexible piece of wood or metal which could be bent around lead weights (called "ducks" in the jargon) to form a smooth nonlinear shape around which the draftsmen traced the desired line.

The practical usefulness of splines in a large number of branches triggered a more theoretical interest in the topic and a rich variety of splines made their appearance in the literature, sometimes referred to as the *zoo of splines* (Lyche et al., 2018). A very popular species in this zoo is the B-spline, as it is endowed with attractive theoretical and computational properties. B-splines were pioneered by Schoenberg (1946*a*,*b*), even though the latter author suggests that the early emergence of B-splines can be traced back to the work of Laplace. Schoenberg's seminal work initiated a flourishing research trend in spline approximation theory and we refer the reader to Schumaker (2007) for a detailed bibliography and further historical notes related to (B-)splines.

### 1.3.2 Mathematical formulation of B-splines

The building blocks of a B-spline consist of a set of polynomial pieces assembled together at specific points called knots. The degree of a B-spline refers to the degree of the polynomial between adjacent knots. For a B-spline of degree $d$, there are $d+1$ polynomial segments tied together at $d$ inner knots and the B-spline is positive on a domain made of $d+2$ knots and zero everywhere else. When the knots are equidistant, the B-spline is said to be uniform. For the sake of illustration, consider a uniform cubic B-spline ($d=3$) defined on the following knots $\{-1, -0.5, 0, 0.5, 1\}$ with knot distance $h = 0.5$. The analytic formula (see e.g. Holmes, 2007, p. 64) is given in (1.4) and represented in Figure 1.3 (a).

In this thesis, B-splines will mainly be used in a regression context to approximate a smooth function. Consider for instance a simple homoscedastic model $y_i = f(x_i) + \varepsilon_i,\ i = 1, \ldots, n$ with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. To approximate the unknown function $f$ in an interval $[a, b]$, we use a (cubic) B-spline basis as shown in Figure 1.3 (b) and model the regression curve as a linear combination of these basis functions, i.e. $f(x) = \sum_{k=1}^{K} \theta_k b_k(x)$, where the spline parameters $\theta_k$'s are commonly referred to as the amplitudes of the B-splines and $K$ is the number of basis functions. Defining the $n \times K$ basis matrix $B$ for which the entry at the intersection of the $i$th row and $k$th column is $b_k(x_i)$, we can use the least squares criterion to find the amplitude vector that minimizes the sum of squares $\widehat{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \|\mathbf{y} - B\boldsymbol{\theta}\|^2$ ($\|\cdot\|$ is the Euclidean norm) with solution $\widehat{\boldsymbol{\theta}} = (B^\top B)^{-1} B^\top \mathbf{y}$ and hence fitted curve $\hat{f}(x) = \sum_{k=1}^{K} \hat{\theta}_k b_k(x)$.

$$b(x) = \begin{cases} \frac{1}{6h^3}(x+1)^3 & \text{if } -1 \le x \le -0.5 \\ \frac{1}{6} + \frac{1}{2h}(x+0.5) + \frac{1}{2h^2}(x+0.5)^2 & \\ \quad -\frac{1}{2h^3}(x+0.5)^3 & \text{if } -0.5 \le x \le 0 \\ \frac{1}{6} - \frac{1}{2h}(x-0.5) + \frac{1}{2h^2}(x-0.5)^2 & \\ \quad +\frac{1}{2h^3}(x-0.5)^3 & \text{if } 0 \le x \le 0.5 \\ -\frac{1}{6h^3}(x-1)^3 & \text{if } 0.5 \le x \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{1.4}$$

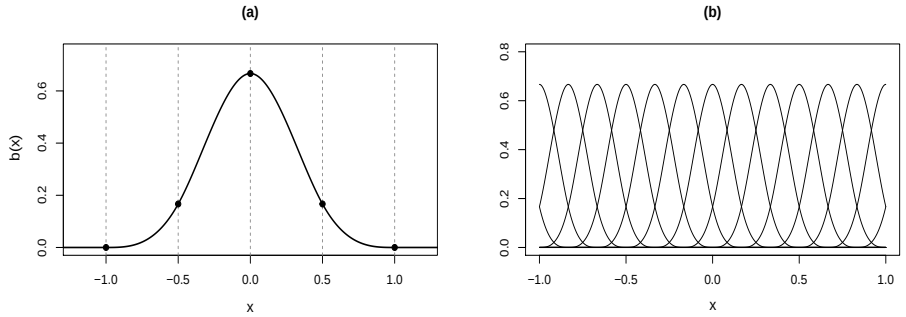



Figure 1.3: (a) A uniform cubic B-spline with knots at $\{-1, -0.5, 0, 0.5, 1\}$. (b) Cubic B-spline basis with 15 B-splines.

### 1.3.3 The role of the penalty

The main problem with the least squares criterion described in Section 1.3.2 is that the shape of the fitted curve will depend on the chosen number of B-spline basis functions. If the B-spline basis is too sparse, the resulting fit will fail to reproduce important patterns of the target, while overabundance of basis terms will produce a rough curve characterized by frequent fluctuations.

To overcome this problem, Eilers and Marx (1996) proposed to use the P-spline approach, which consists in specifying a large number of basis functions and counterbalance the flexibility of the fit by introducing a roughness penalty based on finite differences of adjacent B-spline coefficients. Mathematically, the estimated vector of B-spline amplitudes satisfies $\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \|\mathbf{y} - B\boldsymbol{\theta}\|^2 + \lambda \|D_r\boldsymbol{\theta}\|^2$, where $D_r$ is the $r$th order difference matrix with dimension $(K - r) \times K$ and $\lambda$ is a nonnegative smoothing (or penalty) parameter governing the smoothness of the fit. Let us define the $r$th order difference operator as $\Delta^r$, such that $\Delta^r\theta_k = \Delta^{r-1}\theta_k - \Delta^{r-1}\theta_{k-1}$ and $\Delta^1\theta_k = \theta_k - \theta_{k-1}$. Assuming a first-order penalty, the difference matrix is:

$$D_1 = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix},$$

and $D_1\boldsymbol{\theta} = (\Delta^1\theta_2, \dots, \Delta^1\theta_K)^\top$, such that the penalty can be written in terms of the difference operator $\|D_1\boldsymbol{\theta}\|^2 = \sum_{j=2}^{K}(\Delta^1\theta_j)^2$.

The solution to the penalized least squares problem is known to be $\widehat{\boldsymbol{\theta}} = (B^\top B + \lambda D_r^\top D_r)^{-1}B^\top\mathbf{y}$ and the resulting fit is $\widehat{\mathbf{y}} = S_\lambda\mathbf{y}$, where $S_\lambda = B(B^\top B + \lambda D_r^\top D_r)^{-1}B^\top$ is the smoothing matrix. Optimal smoothing is usually determined by a cross-validation argument, for instance the value of $\lambda$ is chosen to be the one that minimizes the generalized cross-validation criterion $\text{GCV}(\lambda) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 / (1 - n^{-1}\text{Tr}(S_\lambda))^2$, see Ruppert et al. (2003) p. 117.

To illustrate the role of the penalty assume that the regression function of the homoscedastic model in Section 1.3.2 is given by $f(x) = 2\cos(\pi x) +$

$\sin(x^3)$ and that our objective is to estimate the latter in the interval $[-2, 2]$ by use of $K = 40$ B-spline basis functions. The sample size is $n = 250$ and the standard deviation of the error is fixed at $\sigma = 0.8$. Estimation results are reported in Figure 1.4. In Figure 1.4 (a) the target regression function (black curve) is estimated without imposing any penalty and the large number of basis functions results in a complex fit with wiggly patterns (red curve). Figure 1.4 (b) shows the estimated target on the same dataset with a third order penalty ($r = 3$) imposed on finite differences of neighboring B-spline parameters. The smoothing parameter that minimizes the GCV criterion is $\lambda_{\text{GCV}} = 7.344$ and the P-spline fit is smoother, capturing more accurately the nonlinear trajectory of the target regression curve.

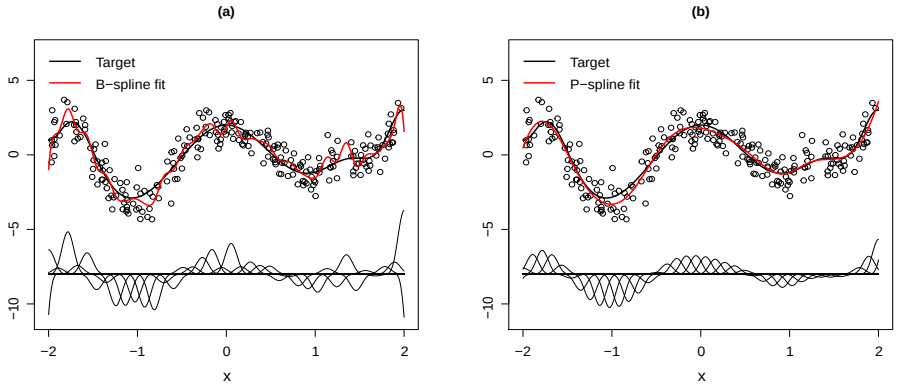



Figure 1.4: (a) Approximation of the regression function with $K = 40$ basis functions without penalty. (b) Penalized approximation of the target function with $K = 40$ and smoothing parameter $\lambda_{\text{GCV}} = 7.344$.

### 1.3.4 Bayesian P-splines

To work with P-splines in a Bayesian setting, Lang and Brezger (2004) proposed to replace the difference penalty on neighboring B-spline coefficients by its stochastic version corresponding to a random walk. For instance, a first order difference penalty is replaced by a random walk of first order, namely $\theta_k = \theta_{k-1} + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \lambda^{-1})$ and a diffuse prior on the initial value, i.e. $p(\theta_1) \propto \textit{constant}$. Using the difference matrix defined in the previous section, we can write $D_1\boldsymbol{\theta} \sim \mathcal{N}_{K-1}(0, \lambda^{-1} I_{K-1})$, where $I_{K-1}$ is an identity matrix of dimension $K - 1$. Furthermore, let us define the penalty matrix $P = D_1^\top D_1 + \epsilon I_K$, where $\epsilon$ is a small

number ($\epsilon = 10^{-6}$, say) added to the elements in the main diagonal of the rank deficient matrix $D_1^\top D_1$ to ensure $P$ is full rank. It follows that the (proper) prior for the vector of B-spline amplitudes is given by (see Brezger and Steiner, 2008) $p(\boldsymbol{\theta}|\lambda) \propto \exp\left(-0.5\lambda \boldsymbol{\theta}^\top P \boldsymbol{\theta}\right)$.

For a full Bayesian treatment, a prior is imposed on the smoothing parameter $\lambda$. Following Jullion and Lambert (2007), we use a robust specification for the roughness penalty prior $\lambda|\delta \sim \mathcal{G}(\nu/2, (\nu\delta)/2)$, where $\delta \sim \mathcal{G}(a_\delta, b_\delta)$. Fixing $\nu = 1$ and $a_\delta = b_\delta = 0.5$ yields a marginal prior density for $\lambda$ corresponding to a Beta-prime distribution (Lambert and Bremhorst, 2019); see Appendix A1 for details . Finally, Jeffreys' prior is imposed on the precision $\tau = 1/\sigma^2$. The Bayesian P-spline model for the regression setting of Section 1.3.2 is summarized as follows:

$$\begin{aligned}
(y_i|\boldsymbol{\theta}, \tau) &\sim \mathcal{N}(\boldsymbol{\theta}^\top \mathbf{b}(x_i), \tau^{-1}), \\
(\boldsymbol{\theta}|\lambda, \tau) &\sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\left(0, (\lambda\tau P)^{-1}\right), \\
(\lambda|\delta) &\sim \mathcal{G}(\nu/2, (\nu\delta)/2), \\
\delta &\sim \mathcal{G}(a_\delta, b_\delta), \\
p(\tau) &\propto \tau^{-1},
\end{aligned}$$

where $\mathbf{b}(x_i) = (b_1(x_i), \ldots, b_K(x_i))^\top$ is the $i$th row of matrix $B$. Let $\Sigma_{\boldsymbol{\theta}} := \tau^{-1}(B^\top B + \lambda P)^{-1}$, the conditional posterior distributions are given by (see Appendix A2):

$$\begin{aligned}
(\boldsymbol{\theta}|\lambda, \tau, \mathcal{D}) &\sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\left((B^\top B + \lambda P)^{-1} B^\top \mathbf{y}, \Sigma_{\boldsymbol{\theta}}\right), \\
(\tau|\boldsymbol{\theta}, \lambda, \mathcal{D}) &\sim \mathcal{G}\left(0.5(n+K), 0.5(\|\mathbf{y} - B\boldsymbol{\theta}\|^2 + \lambda \boldsymbol{\theta}^\top P \boldsymbol{\theta})\right), \\
(\delta|\lambda, \mathcal{D}) &\sim \mathcal{G}\left(0.5\nu + a_\delta, 0.5\nu\lambda + b_\delta\right), \\
(\lambda|\boldsymbol{\theta}, \tau, \delta, \mathcal{D}) &\sim \mathcal{G}\left(0.5(K+\nu), 0.5(\tau \boldsymbol{\theta}^\top P \boldsymbol{\theta} + \nu\delta)\right).
\end{aligned}$$

As full conditional posterior distributions are available, we use the Gibbs algorithm (Geman and Geman, 1984) to draw samples from the joint posterior $p(\boldsymbol{\theta}, \tau, \delta, \lambda|\mathcal{D})$. The Gibbs sampler is given in Algorithm 1.

Let us apply the Bayesian P-spline approach to the simulation setting of Section 1.3.3, where the aim is to estimate $f(x) = 2\cos(\pi x) + \sin(x^3)$ in $[-2, 2]$ by using 40 B-spline basis functions and a third order penalty.

**Algorithm 1: Gibbs sampler to draw from $p(\boldsymbol{\theta}, \tau, \delta, \lambda|\mathcal{D})$**

1: Fix initial values $\lambda^{(0)}$ and $\tau^{(0)}$.
2: **for** $m = 1, \ldots, M$ **do**
3: $\quad \boldsymbol{\theta}^{(m)} \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\left((B^\top B + \lambda^{(m-1)} P)^{-1} B^\top \mathbf{y}, \Sigma_{\boldsymbol{\theta}}^{(m-1)}\right)$.
4: $\quad \tau^{(m)} \sim \mathcal{G}\left(0.5(n+K), 0.5\left(\|\mathbf{y} - B\boldsymbol{\theta}^{(m)}\|^2 + \lambda^{(m-1)} \boldsymbol{\theta}^{(m)\top} P \boldsymbol{\theta}^{(m)}\right)\right)$.
5: $\quad \delta^{(m)} \sim \mathcal{G}\left(0.5\nu + a_\delta, 0.5\nu\lambda^{(m-1)} + b_\delta\right)$.
6: $\quad \lambda^{(m)} \sim \mathcal{G}\left(0.5(K+\nu), 0.5\left(\tau^{(m)} \boldsymbol{\theta}^{(m)\top} P \boldsymbol{\theta}^{(m)} + \nu\delta^{(m)}\right)\right)$.
7: **end for**

The Gibbs sampler (cf. Algorithm 1) is implemented with $M = 25,000$, a burn-in of length $10,000$ and initial parameters $\lambda^{(0)} = 3$ and $\tau^{(0)} = ((n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2)^{-1}$, where $\bar{y}$ is the sample mean of the response data. Geweke statistics (Geweke, 1992) are used as a diagnostic tool for chain convergence.

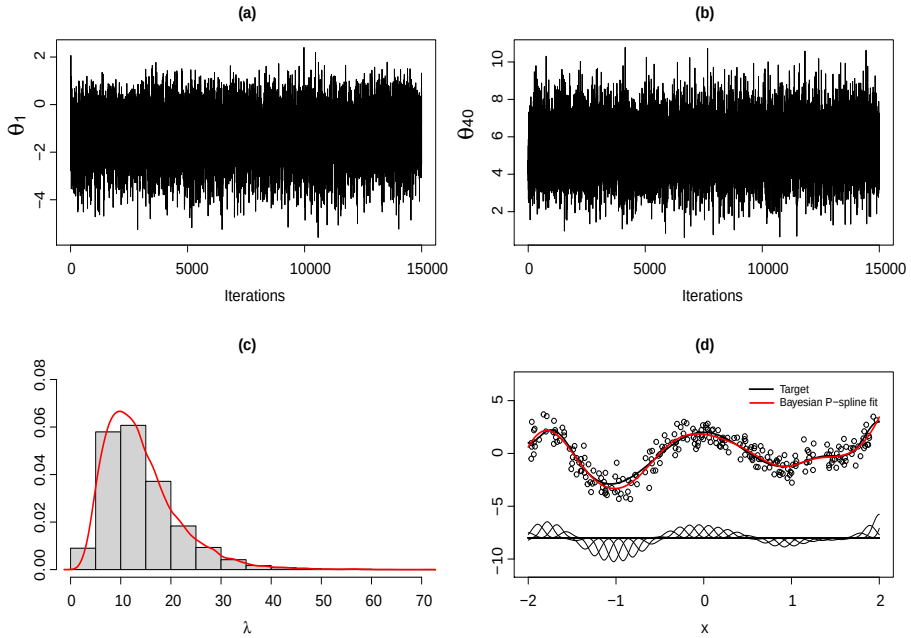



Figure 1.5: (a) Trace plot of $\theta_1$. (b) Trace plot of $\theta_{40}$. (c) Histogram of the smoothing parameter $\lambda$. (d) Penalized estimation of the target function with Bayesian P-splines.

All model parameters have Geweke statistics in the range $[-1.96, 1.96]$, suggesting that convergence of the chains is reached. Figure 1.5 displays the trace plots of $\theta_1$ and $\theta_{40}$ as well as the histogram of $\lambda$ after burn-in and the estimated target regression function with the Bayesian P-spline approach, where the mean of the posterior sample $\theta_j^{(m)}, m = 1, \ldots, 15\,000$ is used as a point estimate of the B-spline amplitude $\theta_j$.

## 1.4 Unifying Laplace's method and P-splines

This section aims at showing how Laplace approximations can be combined with penalized B-splines for fast approximate Bayesian inference in the Cox model (Cox, 1972), a popular modeling approach for survival data. The Cox-LPS methodology presented hereafter is summarized in the `coxlps()` routine of the **blapsr** package (cf. Chapter 5). The material presented here also serves as a smooth introduction to the next chapter devoted to the LPS method in promotion time cure models.

### 1.4.1 Basic elements of survival analysis

In survival analysis, the primary object of interest is a nonnegative random variable $T$ (assumed continuous here) representing the time from a well-defined origin to the occurrence of an event of interest. In the context of time-to-event data, the survival time (or failure time) $T$ is usually characterized by the survival function $S(t) = P(T > t)$, representing the probability that the event of interest will arise beyond time $t$ and satisfying $S(0) = 1$ and $S(+\infty) = 0$. Another important quantity is the hazard function $h(t) = \lim_{\Delta t \to 0^+} \{P(t \leq T < t + \Delta t \,|T \geq t)/\Delta t\}$ as it represents the instantaneous risk of experiencing the event at time $t$ given that the event did not occur prior to time $t$. Denoting by $f(t)$ the probability density function of $T$ and using the definition of conditional probability, we recover $h(t) = f(t)/S(t) = -d \log S(t)/dt$. The cumulative hazard function is $H(t) = \int_0^t h(u)\, du$, and since $h(t) = -S'(t)/S(t)$, we also have $H(t) = -\int_0^t S'(u)/S(u)\, du = -[\log S(u)|_0^t] = -\log S(t)$. Classic statistical approaches cannot be used to analyze time-to-event data because of a special feature called censoring. When a survival time of a unit under study is censored, it means that the event of interest has not been observed for that unit. A potential reason may be that the unit is lost to follow-up or experiences an event unrelated to the event of interest and is therefore outside the risk set. Let $C$ be a random censor-

ing time. Under right censoring, we observe the smallest between $T$ and $C > 0$, i.e. $T_{obs} = \min(T, C)$ and an indicator function $\tilde{\delta} = \mathbb{I}(T \leq C)$ satisfying $\tilde{\delta} = 1$ if $T \leq C$ (failure) and $\tilde{\delta} = 0$ if $T > C$ (censoring).

In practical applications, survival times are accompanied by further information about the group of subjects under study. This extra layer of data is expressed by a vector of covariates $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$. The Cox model postulates a relationship between explanatory variables and survival time by specifying the hazard as $h(t) = h_0(t) \exp(\boldsymbol{\beta}^\top \mathbf{x})$, where $h_0(t)$ is the baseline hazard and is solely a function of time, while $\exp(\boldsymbol{\beta}^\top \mathbf{x})$ incorporates covariate information without time dependency with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ a vector of regression coefficients. The model is also referred to as a proportional hazards model, as for two different covariate vector profiles $\mathbf{x}_i$ and $\mathbf{x}_j$, the hazard ratio $h_i(t)/h_j(t) = \exp(\boldsymbol{\beta}^\top(\mathbf{x}_i - \mathbf{x}_j))$ does not depend on time. To summarize the survival information, let us consider an i.i.d. sample of size $n$ and write the survival data as a set of triplets $\mathcal{D} = \{(t_i, \tilde{\delta}_i, \mathbf{x}_i)\}_{i=1}^n$, where $t_i$ is the failure or censoring time. These survival data will be used in the next section together with the Laplace-P-spline methodology for approximate Bayesian inference in a Cox proportional hazards model.

### 1.4.2 The Cox-Laplace-P-spline model

A flexible specification of the baseline hazard is obtained by writing the latter as a linear combination of cubic B-splines, namely $h_0(t) = \exp(\boldsymbol{\theta}^\top \mathbf{b}(t))$, with $\mathbf{b}(\cdot) = (b_1(\cdot), \ldots, b_K(\cdot))^\top$ a cubic B-spline basis with equidistant knots defined on $[0, t_u]$ and $t_u = \max(t_1, \ldots, t_n)$ the upper bound of the follow-up. Abrahamowicz et al. (1992) were among the first to use regression splines for density estimation in presence of censoring. They model the density as a linear combination of cubic M-splines and estimate the coefficients via pseudo maximum likelihood. Later, Rosenberg (1995) proposed another approach in which the hazard is modeled as a linear combination of cubic B-splines and the optimal amount of smoothness is determined by maximization of the Akaike information criterion (Akaike, 1973). Using the relationship between the survival and cumulative hazard functions, we can write:

$$S_0(t) \quad = \quad \exp\left(-\int_0^t h_0(s)\, ds\right)$$

and using the B-spline specification, we get:

$$\begin{aligned} S_0(t) &= \exp\left(-\int_0^t \exp\left(\boldsymbol{\theta}^\top \mathbf{b}(s)\right)\, ds\right) \\ &\approx \exp\left(-\sum_{j=1}^{j(t)} \exp\left(\boldsymbol{\theta}^\top \mathbf{b}(s_j)\right)\, \Delta_j\right), \end{aligned} \tag{1.5}$$

where the integral is approximated by the rectangle method with $[0, t_u]$ partitioned into $J$ (say 300) small width intervals $[\varphi_{j-1}, \varphi_j]$ and $0 = \varphi_0 < \varphi_1 < \cdots < \varphi_J = t_u$, where $s_j$ and $\Delta_j$ respectively denote the midpoint and width of $[\varphi_{j-1}, \varphi_j]$ and $j(t)$ is an index returning the interval containing $t$.

Let $\boldsymbol{\xi} = (\theta_1, \ldots, \theta_K, \beta_1, \ldots, \beta_p)^\top$ be the vector of B-spline amplitudes and regression coefficients with dimension $\dim(\boldsymbol{\xi}) = K + p$. Using Bayesian P-splines (Lang and Brezger, 2004), the prior on the spline vector is $\boldsymbol{\theta}|\lambda \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\left(0, \lambda^{-1}P^{-1}\right)$ and we further assume the following prior for the vector of regression coefficients $\boldsymbol{\beta} \sim \mathcal{N}_{\dim(\boldsymbol{\beta})}\left(0, \zeta^{-1}I_p\right)$ with small precision (say $\zeta = 10^{-5}$). Hence, the prior for the spline and regression parameters is $\boldsymbol{\xi}|\lambda \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\left(0, Q_{\boldsymbol{\xi}}^{-1}\right)$, with precision matrix:

$$Q_{\boldsymbol{\xi}} := Q_{\boldsymbol{\xi}}(\lambda) = \begin{pmatrix} \lambda P & 0 \\ 0 & \zeta I_p \end{pmatrix}.$$

Furthermore, the following priors are imposed on the elements of the hyperparameter vector $\boldsymbol{\eta} = (\lambda, \delta)^\top$, $\lambda|\delta \sim \mathcal{G}(\nu/2, (\nu\delta)/2)$ and $\delta \sim \mathcal{G}(a_\delta, b_\delta)$ (cf. Section 1.3.4) with $a_\delta = b_\delta = 10^{-4}$ and $\nu = 3$.

### 1.4.3 Approximated conditional posterior for $\boldsymbol{\xi}$

Under right censoring, the likelihood of the Cox model is given by $\mathcal{L}(\boldsymbol{\beta}; \mathcal{D}) = \prod_{i=1}^n \left(h_0(t_i)\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)\right)^{\tilde{\delta}_i} \left(S_0(t_i)\right)^{\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)}$ (see e.g. Dey et al., 1998, p. 275) and so the log-likelihood function is:

$$\ell(\boldsymbol{\beta}; \mathcal{D}) = \sum_{i=1}^n \left\{\tilde{\delta}_i \left(\log h_0(t_i) + \boldsymbol{\beta}^\top \mathbf{x}_i\right) + \left(\log S_0(t_i)\right)\exp(\boldsymbol{\beta}^\top \mathbf{x}_i)\right\}$$

$$\Leftrightarrow \ell(\boldsymbol{\xi};\mathcal{D}) \approx \sum_{i=1}^{n}\left\{\tilde{\delta}_i\left(\boldsymbol{\theta}^\top\mathbf{b}(t_i)+\boldsymbol{\beta}^\top\mathbf{x}_i\right) - \left(\sum_{j=1}^{j(t_i)}\exp\left(\boldsymbol{\theta}^\top\mathbf{b}(s_j)\right)\Delta_j\right)\exp(\boldsymbol{\beta}^\top\mathbf{x}_i)\right\}, \quad (1.6)$$

where the approximation in (1.6) follows from using (1.5). Let us denote the contribution of the $i$th observation to the log-likelihood by $g_i(\boldsymbol{\xi}) = \tilde{\delta}_i\left(\boldsymbol{\theta}^\top\mathbf{b}(t_i)+\boldsymbol{\beta}^\top\mathbf{x}_i\right) - \left(\sum_{j=1}^{j(t_i)}\exp\left(\boldsymbol{\theta}^\top\mathbf{b}(s_j)\right)\Delta_j\right)\exp(\boldsymbol{\beta}^\top\mathbf{x}_i)$.

Using Bayes' rule, the conditional posterior of the vector of spline and regression parameters is:

$$p(\boldsymbol{\xi}|\lambda,\mathcal{D}) \propto \exp\left(\sum_{i=1}^{n} g_i(\boldsymbol{\xi}) - \frac{1}{2}\boldsymbol{\xi}^\top Q_\xi \boldsymbol{\xi}\right). \quad (1.7)$$

The Laplace approximation to (1.7) is obtained in an iterative fashion. First, a second-order Taylor expansion of $g_i(\boldsymbol{\xi})$ is computed around an arbitrary chosen point $\boldsymbol{\xi}^{(0)}$:

$$\begin{aligned} g_i(\boldsymbol{\xi}) &\approx g_i(\boldsymbol{\xi}^{(0)}) + (\boldsymbol{\xi}-\boldsymbol{\xi}^{(0)})^\top\nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} \\ &\quad + \frac{1}{2}(\boldsymbol{\xi}-\boldsymbol{\xi}^{(0)})^\top\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}(\boldsymbol{\xi}-\boldsymbol{\xi}^{(0)}) \\ &\approx \left(g_i(\boldsymbol{\xi}^{(0)}) + \frac{1}{2}\boldsymbol{\xi}^{(0)^\top}\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)} - \boldsymbol{\xi}^{(0)^\top}\nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}\right) \\ &\quad + \boldsymbol{\xi}^\top\nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} + \frac{1}{2}\boldsymbol{\xi}^\top\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi} \\ &\quad - \boldsymbol{\xi}^\top\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)} \\ &\approx constant + \boldsymbol{\xi}^\top\left(\nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} - \nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)}\right) \\ &\quad + \frac{1}{2}\boldsymbol{\xi}^\top\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}, \end{aligned} \quad (1.8)$$

with gradient $\nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}$ and Hessian matrix $\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}$ given by:

$$\nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} = \left(\frac{\partial}{\partial\theta_1}g_i(\boldsymbol{\xi}),\ldots,\frac{\partial}{\partial\theta_K}g_i(\boldsymbol{\xi}),\frac{\partial}{\partial\beta_1}g_i(\boldsymbol{\xi}),\ldots,\frac{\partial}{\partial\beta_p}g_i(\boldsymbol{\xi})\right)^\top_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}$$

$$\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} = \begin{pmatrix} \underbrace{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} g_i(\boldsymbol{\xi})}_{K\times K} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\beta}^\top} g_i(\boldsymbol{\xi})}_{K\times p} \\ \underbrace{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}^\top} g_i(\boldsymbol{\xi})}_{p\times K} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} g_i(\boldsymbol{\xi})}_{p\times p} \end{pmatrix}_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}.$$

The first $K$ entries of the gradient are given by:

$$\frac{\partial}{\partial\theta_k} g_i(\boldsymbol{\xi}) = \tilde{\delta}_i b_k(t_i) - \left(\sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j)\Delta_j\right) \exp(\boldsymbol{\beta}^\top \mathbf{x}_i), \quad k = 1, \ldots, K$$

and the last $p$ elements of the gradient are:

$$\frac{\partial}{\partial\beta_m} g_i(\boldsymbol{\xi}) = \tilde{\delta}_i x_{im} - \left(\sum_{j=1}^{j(t_i)} h_0(s_j)\Delta_j\right) \exp(\boldsymbol{\beta}^\top \mathbf{x}_i)\ x_{im}, \quad m = 1, \ldots, p.$$

To obtain the upper left block of the Hessian matrix first note that:

$$\frac{\partial^2}{\partial\theta_k\partial\theta_l} g_i(\boldsymbol{\xi}) = -\left(\sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j) b_l(s_j)\Delta_j\right) \exp(\boldsymbol{\beta}^\top \mathbf{x}_i). \tag{1.9}$$

Since $j(t_i)$ indicates the bin containing observation $t_i$, we can define the following matrices:

$$B_s := \begin{pmatrix} b_1(s_1) & \ldots & b_K(s_1) \\ \vdots & \ddots & \vdots \\ b_1(s_J) & \ldots & b_K(s_J) \end{pmatrix},\ J_{j(t_i)} := \begin{pmatrix} \mathbf{1} & \mathbf{0} & \ldots & \mathbf{0} & 0 & \ldots & 0 \\ \mathbf{0} & \mathbf{1} & \ldots & \mathbf{0} & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \ldots & \ldots & \mathbf{1} & 0 & \ldots & 0 \\ 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \ldots & \ldots & 0 & 0 & \ldots & 0 \end{pmatrix},$$

where $J_{j(t_i)}$ is a $J \times J$ matrix with upper left block a $j(t_i) \times j(t_i)$ identity matrix. Hence, we can write more compactly:

$$\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}g_i(\boldsymbol{\xi}) = -B_s^\top\Big(\text{diag}\left\{h_0(s_j)\right\}_{j=1}^{J}\ J_{j(t_i)}\Big)B_s\exp(\boldsymbol{\beta}^\top\mathbf{x}_i)\ \Delta_j,$$

where $\text{diag}\left\{h_0(s_j)\right\}_{j=1}^{J}$ is a diagonal matrix of dimension $J\times J$ with diagonal elements $h_0(s_1),\ldots,h_0(s_J)$. The upper right (and lower left) block of the Hessian matrix is:

$$\frac{\partial^2}{\partial\theta_k\partial\beta_m}g_i(\boldsymbol{\xi}) = -\left(\sum_{j=1}^{j(t_i)}h_0(s_j)b_k(s_j)\Delta_j\right)\exp(\boldsymbol{\beta}^\top\mathbf{x}_i)\ x_{im}$$

for $k=1,\ldots,K$ and $m=1,\ldots,p$. Finally, the lower right block is:

$$\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}g_i(\boldsymbol{\xi}) = -\left(\sum_{j=1}^{j(t_i)}h_0(s_j)\Delta_j\right)\exp(\boldsymbol{\beta}^\top\mathbf{x}_i)\ \mathbf{x}_i\mathbf{x}_i^\top.$$

Defining $\sum_{i=1}^{n}\nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} := \nabla g_{\boldsymbol{\xi}^{(0)}}$ and $\sum_{i=1}^{n}\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} := \nabla^2 g_{\boldsymbol{\xi}^{(0)}}$, we obtain the following expression for the sum of the functions $g_i(\cdot)$ omitting the constant:

$$\sum_{i=1}^{n}g_i(\boldsymbol{\xi}) \approx \boldsymbol{\xi}^\top\Big(\nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)}\Big) + \frac{1}{2}\boldsymbol{\xi}^\top\nabla^2 g_{\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}.$$

Plugging the above result in (1.7) yields the Laplace approximation:

$$\widetilde{p}_G(\boldsymbol{\xi}|\lambda,\mathcal{D}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\xi}^\top\left(Q_{\boldsymbol{\xi}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\right)\boldsymbol{\xi} + \boldsymbol{\xi}^\top\left(\nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)}\right)\right).$$

Solving $\nabla\log\widetilde{p}_G(\boldsymbol{\xi}|\lambda,\mathcal{D}) = 0$ and computing $\left(-\nabla^2\log\tilde{p}_G(\boldsymbol{\xi}|\lambda,\mathcal{D})\right)^{-1}$, we find the mean and variance-covariance matrix of the Laplace approximation, namely $\boldsymbol{\xi}^{(1)}(\lambda) = \left(Q_{\boldsymbol{\xi}}(\lambda) - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\right)^{-1}\left(\nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)}\right)$ and $\Sigma^{(1)}(\lambda) = \left(Q_{\boldsymbol{\xi}}(\lambda) - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\right)^{-1}$ respectively.

We repeat the above Laplace approximation in an iterative algorithm until convergence to a Gaussian approximation centered around the posterior mode of $p(\boldsymbol{\xi}|\lambda,\mathcal{D})$. We will denote by $\boldsymbol{\xi}^*_\lambda = (Q_{\boldsymbol{\xi}}(\lambda) - \widetilde{\mathcal{H}})^{-1}\widetilde{\boldsymbol{\varpi}}$ and $\Sigma^*_\lambda = (Q_{\boldsymbol{\xi}}(\lambda) - \widetilde{\mathcal{H}})^{-1}$ the posterior mode and covariance respectively to-

wards which the iterative Laplace approximation scheme has converged, with $\widetilde{\mathcal{H}}$ the Hessian of the log-likelihood at convergence and $\widetilde{\boldsymbol{\varpi}}$ the vector $\left(\nabla g_{\boldsymbol{\xi}} - \widetilde{\mathcal{H}}\boldsymbol{\xi}\right)$ at convergence. By abuse of notation, we can write the Laplace approximation as $\widetilde{p}_G(\boldsymbol{\xi}|\lambda, \mathcal{D}) = \mathcal{N}_{\dim(\boldsymbol{\xi})}\left(\boldsymbol{\xi}^*_\lambda, \Sigma^*_\lambda\right)$.

### 1.4.4 Marginal posterior of the penalty parameter

The Laplace approximation to the conditional posterior $p(\boldsymbol{\xi}|\lambda, \mathcal{D})$ is a crucial component for approximating the marginal posterior of the hyperparameter vector. The latter posterior is given by:

$$p(\boldsymbol{\eta}|\mathcal{D}) = \frac{p(\boldsymbol{\eta}, \boldsymbol{\xi}|\mathcal{D})}{p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})} = \frac{p(\mathcal{D}|\boldsymbol{\eta}, \boldsymbol{\xi})\ p(\boldsymbol{\eta}, \boldsymbol{\xi})}{p(D)p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})} \propto \frac{\mathcal{L}(\boldsymbol{\xi}; \mathcal{D})\ p(\boldsymbol{\xi}|\lambda)\ p(\lambda|\delta)\ p(\delta)}{p(\boldsymbol{\xi}|\lambda, \mathcal{D})}.$$

The above expression is approximated as follows:

$$\begin{aligned}
\widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) &= \left.\frac{\mathcal{L}(\boldsymbol{\xi}; \mathcal{D})\ p(\boldsymbol{\xi}|\lambda)\ p(\lambda|\delta)\ p(\delta)}{\widetilde{p}_G(\boldsymbol{\xi}|\lambda, \mathcal{D})}\right|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*_\lambda} \\
&= |\Sigma^*_\lambda|^{\frac{1}{2}}\ \exp\left(\ell(\boldsymbol{\xi}^*_\lambda; \mathcal{D}) - \frac{1}{2}\boldsymbol{\xi}^{*\top}_\lambda Q_{\boldsymbol{\xi}}\boldsymbol{\xi}^*_\lambda\right)\ |Q_{\boldsymbol{\xi}}|^{\frac{1}{2}}\ \lambda^{\frac{\nu}{2}-1}\ \delta^{\frac{\nu}{2}+a_\delta-1} \\
&\quad \times \exp\left(-\delta\left(\frac{\nu\lambda}{2} + b_\delta\right)\right) \\
&= |\Sigma^*_\lambda|^{\frac{1}{2}}\ \exp\left(\ell(\boldsymbol{\xi}^*_\lambda; \mathcal{D}) - \frac{1}{2}\boldsymbol{\xi}^{*\top}_\lambda Q_{\boldsymbol{\xi}}\boldsymbol{\xi}^*_\lambda\right)\lambda^{\frac{K+\nu}{2}-1}\ \delta^{\frac{\nu}{2}+a_\delta-1} \\
&\quad \times \exp\left(-\delta\left(\frac{\nu\lambda}{2} + b_\delta\right)\right),
\end{aligned}$$

where the last line follows from $|Q_{\boldsymbol{\xi}}|^{\frac{1}{2}} = \lambda^{\frac{K}{2}}|P|^{\frac{1}{2}}\zeta^{\frac{p}{2}} \propto \lambda^{\frac{K}{2}}$ as the determinant of a block diagonal matrix is equal to the product of the determinants of the diagonal blocks. The chosen priors for $\lambda$ and $\delta$ leads to conditional conjugacy for $\delta$, i.e. $(\delta|\lambda, \mathcal{D}) \sim \mathcal{G}(0.5\nu + a_\delta, 0.5\nu\lambda + b_\delta)$. Hence, $\int_0^{+\infty} \delta^{\frac{\nu}{2}+a_\delta-1}\ \exp\left(-\delta\left(\frac{\nu\lambda}{2} + b_\delta\right)\right) d\delta = \Gamma\left(\frac{\nu}{2} + a_\delta\right)\left(\frac{\nu\lambda}{2} + b_\delta\right)^{-\left(\frac{\nu}{2}+a_\delta\right)}$, so that the (approximated) marginal posterior of $\lambda$ is:

$$\begin{aligned}
\widetilde{p}(\lambda|\mathcal{D}) &= \int_0^{+\infty} \widetilde{p}(\lambda, \delta|\mathcal{D})\ d\delta \\
&= |\Sigma^*_\lambda|^{\frac{1}{2}} \exp\left(\ell(\boldsymbol{\xi}^*_\lambda; \mathcal{D}) - \frac{1}{2}\boldsymbol{\xi}^{*\top}_\lambda Q_{\boldsymbol{\xi}}\boldsymbol{\xi}^*_\lambda\right)\lambda^{\frac{K+\nu}{2}-1}\left(\frac{\nu\lambda}{2} + b_\delta\right)^{-\left(\frac{\nu}{2}+a_\delta\right)}.
\end{aligned}$$

For greater numerical stability it is preferable to work on a log-scale for the penalty parameter. We therefore propose the change of variable $v = \log(\lambda)$ and use the method of transformations to obtain:

$$\begin{aligned}\widetilde{p}(v|\mathcal{D}) &= |\Sigma_v^*|^{\frac{1}{2}} \exp\left(\ell(\boldsymbol{\xi}_v^*;\mathcal{D}) - \frac{1}{2}\boldsymbol{\xi}_v^{*\top} Q_{\boldsymbol{\xi}}^v \boldsymbol{\xi}_v^*\right) \exp(v)^{\frac{K+\nu}{2}} \\ &\quad \times \left(\frac{\nu \exp(v)}{2} + b_\delta\right)^{-\left(\frac{\nu}{2}+a_\delta\right)}, \end{aligned} \tag{1.10}$$

with the following matrix being a function of $v$:

$$Q_{\boldsymbol{\xi}}^v := \begin{pmatrix} \exp(v)P & 0 \\ 0 & \zeta I_p \end{pmatrix},$$

vector $\boldsymbol{\xi}_v^* = (Q_{\boldsymbol{\xi}}^v - \widetilde{\mathcal{H}})^{-1}\widetilde{\boldsymbol{\varpi}}$ and covariance matrix $\Sigma_v^* = (Q_{\boldsymbol{\xi}}^v - \widetilde{\mathcal{H}})^{-1}$. The approximated posterior (1.10) is a univariate function of $v$ for which we can numerically compute the maximum a posteriori $\hat{v}$, as well as an equidistant grid $\aleph_v = \left\{v^{(m)}\right\}_{m=1}^M$ that will serve as a set of quadrature points to compute the approximate posterior of the vector $\boldsymbol{\xi}$ (cf. Section 1.2.3). Figure 1.6 illustrates the shape of $\widetilde{p}(v|\mathcal{D})$ for a sample of size $n = 300$ and survival times generated from a Weibull distribution. For this particular dataset the posterior mode is $\hat{v} = \text{argmax}_v\ \widetilde{p}(v|\mathcal{D}) = 7.232$.

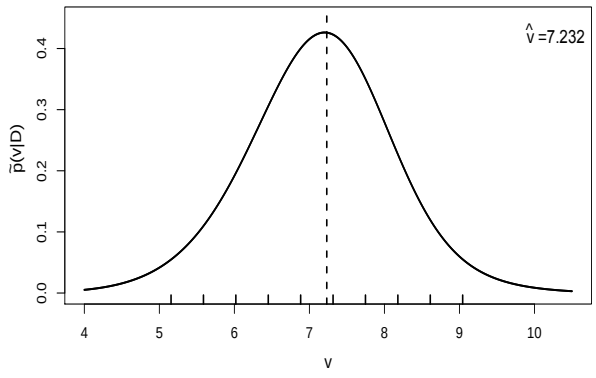



Figure 1.6: Approximated marginal posterior $\widetilde{p}(v|\mathcal{D})$ obtained with a sample of $n = 300$ survival times governed by a Weibull distribution. Vertical tick marks correspond to quadrature points and the dashed line is the maximum a posteriori.

The vertical tick marks along the $x$-axis is a set of (equidistant) quadrature points $\aleph_v = \{5.16, 5.59, \ldots, 9.04\}$ that will be used to approximate $p(\boldsymbol{\xi}|\mathcal{D})$. The quadrature points are chosen so that the posterior mass between the lower bound (equal to 5.16) and the upper bound (equal to 9.04) is approximately 95%.

### 1.4.5 Approximate marginal posterior for $\boldsymbol{\xi}$

The marginal posterior for $\boldsymbol{\xi}$ is obtained by integration:

$$\begin{aligned}
p(\boldsymbol{\xi}|\mathcal{D}) &= \int_0^{+\infty}\int_0^{+\infty} p(\boldsymbol{\xi},\lambda,\delta|\mathcal{D})\ d\delta\ d\lambda \\
&= \int_0^{+\infty} p(\boldsymbol{\xi}|\lambda,\mathcal{D}) \left(\int_0^{+\infty} p(\lambda,\delta|\mathcal{D})\ d\delta\right)\ d\lambda \\
&\approx \int_0^{+\infty} \widetilde{p}_G(\boldsymbol{\xi}|\lambda,\mathcal{D}) \left(\int_0^{+\infty} \widetilde{p}(\lambda,\delta|\mathcal{D})\ d\delta\right)\ d\lambda \\
&\approx \int_0^{+\infty} \widetilde{p}_G(\boldsymbol{\xi}|\lambda,\mathcal{D})\ \widetilde{p}(\lambda|\mathcal{D})\ d\lambda \\
&\approx \int_{\mathbb{R}} \widetilde{p}_G(\boldsymbol{\xi}|\exp(v),\mathcal{D})\ \widetilde{p}(v|\mathcal{D})\ dv \\
&\approx \sum_{m=1}^{M} \widetilde{p}_G\left(\boldsymbol{\xi}|\exp\left(v^{(m)}\right),\mathcal{D}\right)\ \widetilde{p}(v^{(m)}|\mathcal{D})\ \Delta_v, \qquad (1.11)
\end{aligned}$$

where $\Delta_v$ is the grid width of the equidistant grid $\aleph_v$ (cf. Section 1.4.4), $v^{(m)} \in \aleph_v$ and $M$ is the total number of grid points. Defining the following weights:

$$\omega_m := \frac{\widetilde{p}(v^{(m)}|\mathcal{D})\ \Delta_v}{\sum_{m=1}^{M} \widetilde{p}(v^{(m)}|\mathcal{D})\ \Delta_v}, \quad m = 1, \ldots, M \qquad (1.12)$$

and dividing (1.11) by the denominator of (1.12) results in the following approximation of the marginal posterior of the spline and regression coefficients:

$$\widehat{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{M} \omega_m\ \mathcal{N}_{\dim(\boldsymbol{\xi})}\left(\boldsymbol{\xi}^*_{v^{(m)}}, \Sigma^*_{v^{(m)}}\right), \qquad (1.13)$$

which corresponds to a finite mixture of multivariate Gaussian densities with mean $\boldsymbol{\xi}^*_{v^{(m)}} = (Q^{v^{(m)}}_{\boldsymbol{\xi}} - \widetilde{\mathcal{H}})^{-1}\widetilde{\boldsymbol{\varpi}}$, covariance matrix $\Sigma^*_{v^{(m)}} = (Q^{v^{(m)}}_{\boldsymbol{\xi}} - \widetilde{\mathcal{H}})^{-1}$ and $Q^{v^{(m)}}_{\boldsymbol{\xi}}$ is matrix $Q^{v}_{\boldsymbol{\xi}}$ evaluated at the quadrature point $v^{(m)}$.

A point estimate of the vector of spline and regression coefficients is given by the mixture mean $\widehat{\boldsymbol{\xi}} = \left(\widehat{\boldsymbol{\theta}}^\top, \widehat{\boldsymbol{\beta}}^\top\right)^\top = \sum_{m=1}^{M} \omega_m \, \boldsymbol{\xi}^*_{v^{(m)}}$ and replacing $\widehat{\boldsymbol{\theta}}$ in (1.5) yields the estimated baseline survival $\widehat{S}_0(t)$. The posterior variance-covariance matrix arising from the mixture is (see Frühwirth-Schnatter, 2006) $V(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{M} \omega_m \Sigma^*_{v^{(m)}} + \sum_{m=1}^{M} \omega_m (\boldsymbol{\xi}^*_{v^{(m)}} - \widehat{\boldsymbol{\xi}})(\boldsymbol{\xi}^*_{v^{(m)}} - \widehat{\boldsymbol{\xi}})^\top$. The posterior standard deviation of the $h$th element $\xi_h$ is thus given by the square root of the $h$th diagonal entry of $V(\boldsymbol{\xi}|\mathcal{D})$. From the joint posterior in (1.13), one can easily obtain the (approximated) posterior of a single element, say $\xi_h$, which corresponds to a mixture of univariate Gaussians :

$$\widehat{p}(\xi_h|\mathcal{D}) = \sum_{m=1}^{M} \omega_m \, \mathcal{N}_1\left(\xi^*_{h,v^{(m)}}, \Sigma^*_{hh,v^{(m)}}\right), \tag{1.14}$$

where the scalar $\xi^*_{h,v^{(m)}}$ is the $h$th entry of the vector $\boldsymbol{\xi}^*_{v^{(m)}}$ and $\Sigma^*_{hh,v^{(m)}}$ is the variance component corresponding to the $h$th element in the main diagonal of the variance-covariance matrix $\Sigma^*_{v^{(m)}}$. The posterior (1.14) can be exploited to numerically construct a $(1-\alpha)\times 100\%$ quantile-based credible interval for $\xi_h$. Figure 1.7 illustrates (in red) the posterior (1.14) for two arbitrarily chosen B-spline coefficients $\theta_8$ and $\theta_{13}$. The blue and green curves correspond to the unweighted and weighted Gaussian components respectively when $M = 10$ quadrature points are chosen in the mixture.

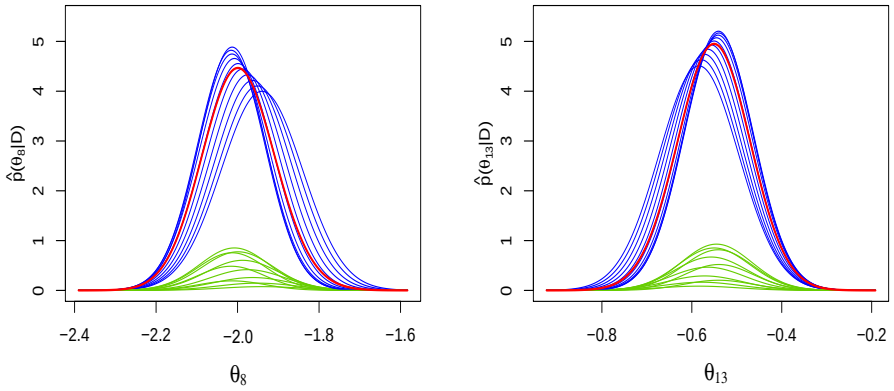



Figure 1.7: Approximate posterior distribution for two B-spline amplitudes $\theta_8$ and $\theta_{13}$ (red). Blue and green curves correspond to unweighted and weighted Gaussian density components respectively.

### 1.4.6 A small simulation study

To evaluate the performance of the Cox-Laplace-P-spline model, a small simulation study is implemented with three covariates, $X_1 \sim \mathcal{N}(0, 0.25)$, $X_2 \sim \mathcal{U}(0,1)$ and $X_3 \sim \text{Bern}(0.5)$ to which we subtract 0.5 to obtain a mean-centered covariate. To generate survival times from the Cox model, we follow Bender et al. (2005) with a Weibull distribution for the baseline characterized by the probability density function $f_0(t) = (a/b^a)t^{a-1}\exp(-(t/b)^a)$ for $t > 0$ and fix $a = 2.4$ and $b = 2.1$. The survival time is generated as $T_i \sim b\left(-\log\left(\mathcal{U}(0,1)\right)\exp(-\boldsymbol{\beta}^\top \mathbf{x}_i)\right)^{1/a}$, with $\beta_1 = 2.20$, $\beta_2 = 1.30$ and $\beta_3 = -0.90$.

Two scenarios are considered for the censoring scheme: $(C_{0\%})$ absence of censoring and $(C_{15\%})$ right censoring governed by a uniform distribution $C_i \sim \mathcal{U}(1,6)$ yielding approximately 15% of censored observations. To estimate the baseline survival $S_0(t) = \exp(-(t/b)^a)$, we specify 30 cubic B-splines (with a third order penalty) in $[0, t_u]$, where $t_u$ is the largest observed failure or censoring time for a given dataset. The simulation setting entails $S = 500$ replications of a sample of size $n = 300$.

To assess the frequentist properties of the Bayesian estimator of a regression coefficient $\beta_j$, the following measures of performance are computed. The empirical bias is defined as the average of the difference between the estimate and the true parameter value over $S$ replications:

$$\text{Bias}_{\hat{\beta}_j} := \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\beta}_j^{(s)} - \beta_j\right).$$

The empirical standard error (ESE) is taken to be the sample standard deviation of the estimates over $S$ replications:

$$\text{ESE}_{\hat{\beta}_j} := \left\{\frac{1}{S-1}\sum_{s=1}^{S}\left(\hat{\beta}_j^{(s)} - \bar{\hat{\beta}}_j\right)^2\right\}^{\frac{1}{2}},$$

where $\bar{\hat{\beta}}_j = S^{-1}\sum_{s=1}^{S}\hat{\beta}_j^{(s)}$. The root mean square error (RMSE) is defined as the square root of the average of the squared difference between the estimate and the true value over $S$ replications:

$$\text{RMSE}_{\hat{\beta}_j} := \left\{ \frac{1}{S} \sum_{s=1}^{S} \left( \hat{\beta}_j^{(s)} - \beta_j \right)^2 \right\}^{\frac{1}{2}}.$$

The coverage probability of the 90% and 95% (pointwise) credible intervals for $\beta_j$ is the average of an indicator function $\mathbb{I}(\cdot)$ that takes the value 1 if the constructed interval (denoted by $\text{CI}_{90\%,j}$ or $\text{CI}_{95\%,j}$) includes the true parameter value and 0 otherwise:

$$\begin{aligned} \text{CP}_{90\%,j} &:= \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\left( \beta_j \in \text{CI}_{90\%,j}^{(s)} \right), \\ \text{CP}_{95\%,j} &:= \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}\left( \beta_j \in \text{CI}_{95\%,j}^{(s)} \right). \end{aligned}$$

We also report the results obtained with the `coxph()` function from the **survival** package in **R**, where the $(1-\alpha) \times 100\%$ confidence interval for $\beta_j$ is computed as $\hat{\beta}_j \pm z_{\alpha/2} \sqrt{\widehat{V}(\hat{\beta}_j)}$ (asymptotically valid). Table 1.1 summarizes the results for the estimation of regression coefficients.

| $C_{0\%}$ | Parameter | Bias | $\text{CP}_{90\%}$ | $\text{CP}_{95\%}$ | ESE | RMSE |
|---|---|---|---|---|---|---|
| LPS | $\beta_1 = 2.20$ | -0.019 | 91.4 | 95.6 | 0.151 | 0.152 |
| | $\beta_2 = 1.30$ | -0.013 | 91.6 | 95.4 | 0.207 | 0.207 |
| | $\beta_3 = -0.90$ | 0.009 | 89.8 | 95.4 | 0.124 | 0.124 |
| `coxph()` | $\beta_1 = 2.20$ | 0.012 | 91.4 | 95.0 | 0.158 | 0.158 |
| | $\beta_2 = 1.30$ | 0.004 | 90.6 | 95.6 | 0.212 | 0.211 |
| | $\beta_3 = -0.90$ | 0.000 | 88.6 | 95.4 | 0.126 | 0.126 |
| $C_{15\%}$ | Parameter | Bias | $\text{CP}_{90\%}$ | $\text{CP}_{95\%}$ | ESE | RMSE |
| LPS | $\beta_1 = 2.20$ | -0.001 | 90.4 | 95.0 | 0.166 | 0.166 |
| | $\beta_2 = 1.30$ | -0.015 | 90.6 | 95.4 | 0.231 | 0.231 |
| | $\beta_3 = -0.90$ | 0.001 | 87.8 | 94.4 | 0.140 | 0.140 |
| `coxph()` | $\beta_1 = 2.20$ | 0.015 | 89.8 | 94.0 | 0.171 | 0.171 |
| | $\beta_2 = 1.30$ | -0.001 | 90.6 | 94.8 | 0.234 | 0.234 |
| | $\beta_3 = -0.90$ | -0.002 | 88.4 | 94.0 | 0.141 | 0.141 |

Table 1.1: Simulation results for $S = 500$ replicates of sample size $n = 300$ with the Laplace-P-spline approach and the `coxph()` function under absence of censoring $C_{0\%}$ and presence of right censoring $C_{15\%}$.

The above simulation results show that the LPS and `coxph()` approaches behave similarly. The estimated bias is nearly zero under the two censoring scenarios and the estimated coverage probabilities are relatively close to their nominal level. In terms of ESE and RMSE the LPS approach exhibits slightly lower values than `coxph()`. The LPS methodology is computationally inexpensive as under the specified simulation setting it takes approximately 160 milliseconds for a dataset of size $n = 300$ to fit the model with a total of 33 parameters to be estimated. Figures 1.8 and 1.9 show the estimated baseline survival and hazard respectively under the two considered censoring scenarios.

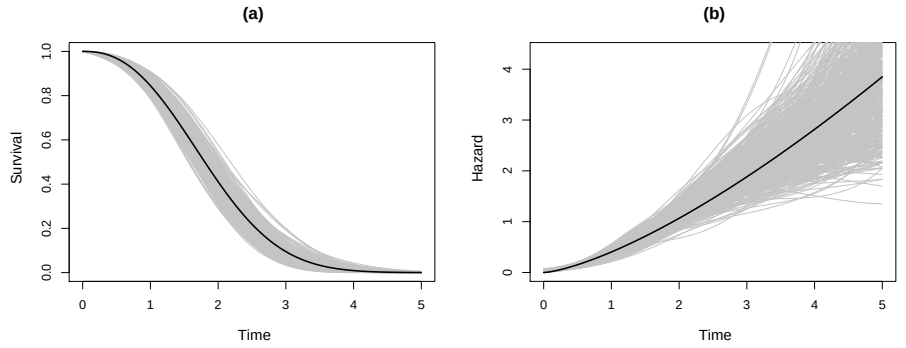



Figure 1.8: Estimation of the baseline survival (a) and baseline hazard (b) with the Laplace-P-spline method (one gray curve per dataset) under absence of censoring. The black curves are the target baseline functions.

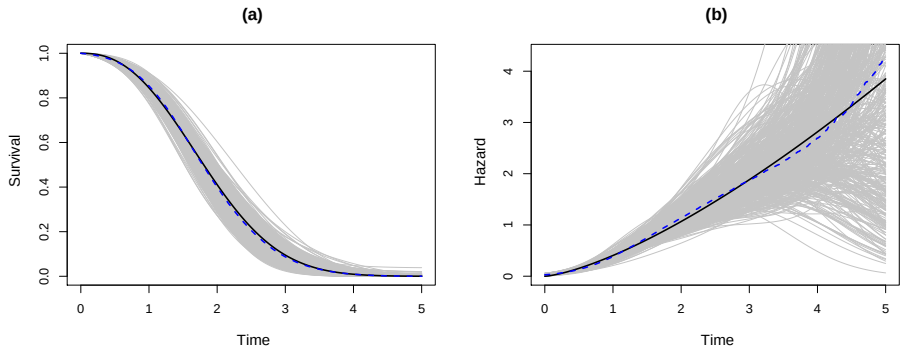



Figure 1.9: Estimation of the baseline survival (a) and baseline hazard (b) with the Laplace-P-spline method (one gray curve per dataset) with 15% right censoring. Black curves are the target baseline functions and dashed curves are the pointwise median of the 500 estimated curves.

It is also worth noting that the above simulation study uses the following parameterization for the prior hyperparameters $a_\delta = b_\delta = 10^{-4}$ and $\nu = 3$. Further simulations (not reported here) have been implemented with the alternative parameterization proposed in Section 1.3.4, namely $a_\delta = b_\delta = 0.5$ and $\nu = 1$. It turns out that there is little sensitivity of posterior estimates with regard to the chosen parameterization as the simulation results are similar to those reported in Table 1.1.

## 1.5 Conclusion

In this chapter, we presented the groundwork of the thesis by focusing on Laplace's method and P-splines. After having introduced basic concepts of Laplace approximations, we showed how they can be embedded in the Bayesian paradigm to approximate posterior quantities of interest and hence serve as a surrogate to classic MCMC methods for posterior exploration. Furthermore, we explained the notion of nested approximations and its usefulness in Bayesian modeling. With regard to the spline facet, B-splines and their penalized version have been introduced with a particular emphasis on the important role played by the roughness penalty. Finally, we combined Laplace's method and Bayesian P-splines for fast inference in a Cox model and measured the performance of the LPS approach in a small simulation exercise.

# CHAPTER 2

# Fast Bayesian inference in a flexible promotion time cure model based on Laplace-P-splines

This chapter is based on the paper: 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 and Data Analysis*, Volume 124, pages 151-167. https://doi.org/10.1016/j.csda.2018.02.007

## 2.1 Motivation

Bayesian methods for flexible time-to-event models usually rely on the theory of Markov chain Monte Carlo (MCMC) to sample from posterior distributions and perform statistical inference. In this chapter, a novel methodology is proposed to overcome the inconvenient facets inherent to MCMC sampling (e.g. convergence problems, huge computational resources) with the major advantage that posterior distributions of latent variables can rapidly be approximated with high accuracy. This is achieved by exploiting the synergy between Laplace's method for posterior approximations and P-splines, a flexible tool for nonparametric modeling. The methodology is developed in the class of cure survival models, a useful extension of standard time-to-event models where it is assumed that an unknown proportion of unidentified (cured) units will never experience the monitored event.

An attractive feature of this approach is that point estimators and credible intervals can be straightforwardly constructed even for complex functionals of latent model variables. The properties of the proposed methodology are evaluated using simulations and illustrated on two real datasets. The fast computational speed and accurate results suggest that the combination of P-splines and Laplace approximations can be considered as a serious competitor of MCMC to make inference in semiparametric models, as illustrated on survival models with a cure fraction.

## 2.2 Introduction

There is a growing interest for cure rate models in survival analysis as witnessed by the number of published papers on that topic in statistical journals. These models have gained in popularity as they intrinsically account for long-term survivors that will never experience the event of interest even when followed-up for an extended time period. The promotion time (cure) model introduced by Yakovlev et al. (1996) is motivated by cancer tumor kinetics, the biological mechanism underlying the proliferation and growth of carcinogenic cells. In particular, let $N \sim \text{Poisson}\big(\phi(\mathbf{x})\big)$ be the number of carcinogenic cells affecting a given subject with mean $\phi(\mathbf{x}) = \exp(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta})$. To the $i$th cell is associated a latent event time $T_i \geq 0$ representing the duration necessary for the cell to grow to a detectable tumor mass. Latent event times $\{T_1, \ldots, T_N\}$ are assumed to be independently and identically distributed with common cumulative distribution function $F(t)$ and the observed survival time is defined as $T = \min\{T_1, \ldots, T_N\}$.

When a Cox proportional hazards model (Cox, 1972) is used to model the $N$ conditional latent distributions $F(t_i|\mathbf{z}) = 1 - S_0(t_i)^{\exp(\mathbf{z}^\top \boldsymbol{\gamma})}$, $i = 1, \ldots, N$ one can show that the resulting survival function of $T$ is (see Tsodikov, 1998; Chen et al., 1999):

$$\begin{aligned} S_p(t|\mathbf{x}, \mathbf{z}) &= \exp\left(-\phi(\mathbf{x}) F(t|\mathbf{z})\right) \\ &= \exp\left(-\exp\left(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}\right)\left(1 - S_0(t)^{\exp(\mathbf{z}^\top \boldsymbol{\gamma})}\right)\right). \end{aligned} \quad (2.1)$$

In this model, a subject is cured when $N = 0$, an event arising with a probability given by $P(N = 0\ |\mathbf{x}, \mathbf{z}) = \lim_{t\to\infty} S_p(t|\mathbf{x}, \mathbf{z}) = \exp(-\phi(\mathbf{x}))$. Alternative specifications are proposed in the literature to model the dis-

tribution of latent event times $F(t_i)$, for example Ibrahim et al. (2001) propose a semiparametric form for the latent distribution involving a smoothing parameter controlling the degree of parametricity in the right tail of the population survival function, while Zeng et al. (2006) introduce a semiparametric class of cure models taking into account a subject-specific frailty.

Model (2.1) can be estimated by maximum likelihood methods in a frequentist setting (see Tsodikov, 2002, 2003). From a Bayesian perspective, Yin and Ibrahim (2005) assume a piecewise exponential model for the baseline survival function with a tradeoff between model flexibility and the number of partitions of the time axis. More recently, Bremhorst and Lambert (2016) use a large number of B-splines to specify the baseline hazard and, following Eilers and Marx (1996), counterbalance the flexibility of the model by using a roughness penalty based on finite differences of adjacent B-spline coefficients.

The rather complex structure of the posterior distributions in the latter Bayesian frameworks requires the use of MCMC techniques. For such models, the MCMC toolbox is usually accompanied by a large computational burden and challenging convergence problems under the original parameterization. A crucial component explaining the inefficiency of rejection sampling techniques is a strong posterior correlation appearing firstly among latent variables and secondly between latent variables and hyperparameters of the model, thus having a global impact on convergence speed and autocorrelation. Integrated Nested Laplace Approximations (INLA) is a sampling-free Bayesian methodology that allows to obtain marginal posteriors in the class of latent Gaussian models and has been recognized to be an interesting alternative to standard MCMC methods. In this dimension, Martino (2007) and Rue et al. (2009) are the pioneering references showing how to perform approximate Bayesian inference in latent Gaussian models via Laplace approximations.

While INLA has been shown to work well in a large variety of applications like stochastic volatility models (Martino, Aas, Lindqvist, Neef and Rue, 2011), generalized dynamic linear models (Ruiz-Cárdenas et al., 2012) and spatio-temporal disease mapping models (Schrödle and Held, 2011), there seems to be little work related to survival analysis or penalized B-spline models.

Among the contributions on the subject, we can cite Fong et al. (2010) who combine INLA and O'Sullivan splines in a nonparametric smoothing setting. Martino, Akerkar and Rue (2011) investigate the use of INLA with the R-INLA package (www.r-inla.org) by considering a Cox model where the baseline hazard has a parametric or semiparametric specification. Also, Jiang et al. (2014) study the effect of environmental radiation on cancer by using a cure fraction mixture survival model with a Weibull distribution for event times.

We investigate how Laplace approximations can be extended and combined with penalized B-splines in the context of a semiparametric promotion time cure model. Bridging the gap between Lapace's method and regression splines brings a twofold advantage. First, it provides a fast computational approach to approximate posterior distributions and second, the spline dimension allows for a flexible specification of the baseline distribution yielding smooth estimates of survival quantities. Another crucial point is that in contrast to the classic INLA approach which focuses mainly on posterior marginal univariate distributions, our methodology permits to compute reliable approximations to the posterior joint distributions of latent variables including regression parameters, with the implication that set estimators can be derived even for complicated functions of spline and regression parameters such as the baseline or conditional population survival functions.

Accordingly, the end user will be endowed with a powerful and rapid tool for making inference in the promotion time cure model. Furthermore, while the code design underlying INLA assumes a one-to-one connection between data points and a subset of the latent vector, implying that the dimension of the latter grows with the sample size $n$, our modeling strategy choice is more efficient as it involves a latent vector of a dimension unaffected by the number of observations. Hence, given that the number of B-splines is fixed (to a large value and counterbalanced by a roughness penalty) in the P-spline approach (Eilers and Marx, 2010), the latent vector dimension grows only with the number of regressors in the model and not with $n$.

This chapter is organized as follows. In Section 2.3, the Laplace-P-spline promotion time cure model is defined and the gradient and Hessian of the log-likelihood are computed to obtain a Gaussian approximation of the

conditional posterior distribution of the vector of spline and regression parameters. A strategy is proposed to explore the posterior distribution of the hyperparameter vector and the joint posterior of spline and regression coefficients is derived. The construction of credible intervals for the baseline and population survival functions is also addressed here. In Section 2.4, the merits of the proposed methodology will be assessed by extensive simulations with different scenarios regarding the percentages of cured individuals and right censored subjects. Coverage properties of credible intervals will also be considered. In Section 2.5, we apply the model to two real datasets and Section 2.6 concludes with a discussion.

## 2.3 Laplace-P-spline promotion time model

### 2.3.1 Flexible modeling of the baseline hazard

Following Rosenberg (1995), the log-hazard corresponding to the baseline survival function $S_0(t)$ in (2.1) is specified as a linear combination of cubic B-splines $h_0(t) = \exp\left(\boldsymbol{\theta}^\top \mathbf{b}(t)\right)$, where $\mathbf{b}(\cdot) = (b_1(\cdot), \ldots, b_K(\cdot))^\top$ is a cubic B-spline basis obtained by taking equidistant knots on the compact set $[0, t_u]$, with $t_u$ the upper bound of the follow-up and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$ is the vector of B-spline coefficients. Under this specification and using the rectangle method, the baseline survival function in (2.1) can be approximated as follows (cf. Chapter 1 Equation 1.5):

$$S_0(t) \approx \exp\left( -\sum_{j=1}^{j(t)} \exp\left(\boldsymbol{\theta}^\top \mathbf{b}(s_j)\right) \Delta_j \right), \tag{2.2}$$

where $[0, t_u]$ is divided into $J$ (say 300) small intervals of equal width $\Delta_j$ with midpoint $s_j$. The term $j(t) \in \{1, 2 \ldots, J\}$ is a number corresponding to the interval containing $t$.

### 2.3.2 Latent variables and priors

The vector $\boldsymbol{\xi} = (\theta_1, \ldots, \theta_K, \beta_0, \ldots, \beta_p, \gamma_1, \ldots, \gamma_l)^\top$ of dimension $\dim(\boldsymbol{\xi}) = K + (p+1) + l$ gathers all the latent variables of the model: it contains the B-spline coefficients $\{\theta_k : k = 1, \ldots, K\}$, the regression coefficients $\{\beta_m : m = 0, \ldots, p\}$ used to model the expected number of carcinogenic cells and the regression parameters $\{\gamma_s : s = 1, \ldots, l\}$ in the Cox model describing the incubation time of a given cell.

The key idea behind P-splines (Eilers and Marx, 1996) is to use a fixed large number of B-spline basis functions and to compensate the flexibility by a roughness penalty on finite differences of contiguous B-spline coefficients. The Bayesian analogue (Lang and Brezger, 2004) translates the roughness penalty into a multivariate normal prior distribution for the spline coefficients $\boldsymbol{\theta}|\lambda \sim \mathcal{N}_K(0, \lambda^{-1}P^{-1})$, with $P = D_r^\top D_r + \epsilon I_K$ where $D_r$ is a $(K-r)\times K$ matrix yielding $r$th order differences when applied on a $K$-vector, and $\lambda$ is a nonnegative roughness penalty parameter. For an arbitrary small $\epsilon$ $\big($say $\epsilon = 10^{-6}\big)$, the diagonal perturbation $\epsilon I_K$ makes $P$ full rank. Then, the prior for the full vector $\boldsymbol{\xi}$ given $\lambda$ can be written as:

$$\boldsymbol{\xi}|\lambda \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\big(\boldsymbol{\mu}_{\boldsymbol{\xi}}, \Sigma_{\boldsymbol{\xi}}(\lambda)\big), \qquad \Sigma_{\boldsymbol{\xi}}(\lambda) = \begin{pmatrix} \lambda^{-1}P^{-1} & 0 \\ 0 & \Sigma_{\boldsymbol{\beta},\boldsymbol{\gamma}} \end{pmatrix},$$

where the vector $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ attributes a zero mean to the B-spline coefficients and a potential informative prior mean on the regression coefficients with (prior) positive-definite variance-covariance matrix $\Sigma_{\boldsymbol{\beta},\boldsymbol{\gamma}}$. Whenever a priori knowledge on central tendency or correlation measures is available for the regression coefficient vector, it can be incorporated into the prior $\boldsymbol{\xi}|\lambda$ through the mean and covariance structure.

The hyperparameters of the model are given by $\boldsymbol{\eta} = (\lambda, \delta)^\top$ as, following Jullion and Lambert (2007), we use a robust specification for the roughness penalty prior $\lambda|\delta \sim \mathcal{G}\big(\nu/2, (\nu\delta)/2\big)$ with an uninformative proper distribution on parameter $\delta \sim \mathcal{G}(a_\delta, b_\delta)$. The latter reference shows that when $a_\delta = b_\delta$ are set to a small value (say $10^{-4}$), the estimated curve is not sensitive to the choice of $\nu$ (here set equal to 3).

### 2.3.3 Conditional posterior and Laplace approximation

Let $\mathcal{D}_i = (t_i, \tilde{\delta}_i, \mathbf{x}_i, \mathbf{z}_i)$ denote the observables for unit $i$, with $t_i$ the failure or censoring time, $\tilde{\delta}_i$ a dichotomous event indicator and $\mathbf{x}_i$, $\mathbf{z}_i$ the covariates. The log-likelihood function of the promotion time cure model is $\ell(\boldsymbol{\xi}; \mathcal{D}) = \sum_{i=1}^{n} \Big\{\tilde{\delta}_i \log h_p(t_i|\mathbf{x}_i, \mathbf{z}_i) + \log S_p(t_i|\mathbf{x}_i, \mathbf{z}_i)\Big\}$, where $\mathcal{D} = \bigcup_{i=1}^n \mathcal{D}_i$ and $h_p(\cdot|\mathbf{x}, \mathbf{z})$ is the conditional population hazard function, $h_p(t|\mathbf{x}, \mathbf{z}) = \phi(\mathbf{x})\ \exp(\mathbf{z}^\top\boldsymbol{\gamma})\ S_0(t)^{\exp(\mathbf{z}^\top\boldsymbol{\gamma})} h_0(t)$. Using the B-spline specification of the baseline hazard, we can write more compactly $\ell(\boldsymbol{\xi}; \mathcal{D}) \approx \sum_{i=1}^n g_i(\boldsymbol{\xi})$.

The scalar-valued function $g_i : \mathbb{R}^{\dim(\boldsymbol{\xi})} \to \mathbb{R}$ gives the contribution of the $i$th unit to the log-likelihood and is given by:

$$\begin{aligned} g_i(\boldsymbol{\xi}) = \tilde{\delta}_i \Big( & \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma} + \boldsymbol{\theta}^\top \mathbf{b}(t_i) \\ & - \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) \sum_{j=1}^{j(t_i)} \exp\Big(\boldsymbol{\theta}^\top \mathbf{b}(s_j)\Big) \Delta_j \Big) \\ & - \exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) \left( 1 - \exp\left( - \sum_{j=1}^{j(t_i)} \exp\Big(\boldsymbol{\theta}^\top \mathbf{b}(s_j)\Big) \Delta_j \right)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \right). \end{aligned}$$

The first step of our procedure is to derive the Laplace approximation of the conditional posterior distribution of the spline and regression parameters, namely:

$$p(\boldsymbol{\xi}|\lambda, \mathcal{D}) \propto \exp\left( \sum_{i=1}^{n} g_i(\boldsymbol{\xi}) - \frac{1}{2} \boldsymbol{\xi}^\top Q(\lambda) \boldsymbol{\xi} + \boldsymbol{\xi}^\top Q(\lambda) \boldsymbol{\mu}_{\boldsymbol{\xi}} \right), \tag{2.3}$$

where $Q(\lambda) = \Sigma_{\boldsymbol{\xi}}^{-1}(\lambda)$ is the precision matrix. One major difference with the theoretical set-up described in Rue et al. (2009) is the dimension of the latent vector assumed there to be larger than the number of observations; it is usually much smaller here with $\dim(\boldsymbol{\xi}) << n$. With non-Gaussian responses, $p(\boldsymbol{\xi}|\lambda, \mathcal{D})$ is non-Gaussian and unknown. To make it tractable, we use Laplace's method and compute a second-order Taylor expansion of $g_i(\boldsymbol{\xi})$ around an arbitrary point $\boldsymbol{\xi}^{(0)} \in \mathbb{R}^{\dim(\boldsymbol{\xi})}$:

$$\begin{aligned} g_i(\boldsymbol{\xi}) \quad \approx \quad & constant + \boldsymbol{\xi}^\top \Big( \nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} - \nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} \boldsymbol{\xi}^{(0)} \Big) \\ & + \frac{1}{2} \boldsymbol{\xi}^\top \nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} \boldsymbol{\xi}, \end{aligned} \tag{2.4}$$

where the gradient and Hessian of $g_i(\boldsymbol{\xi})$ are given by:

$$\begin{aligned} \nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} = \Bigg( & \frac{\partial g_i(\boldsymbol{\xi})}{\partial \theta_1} \cdots \frac{\partial g_i(\boldsymbol{\xi})}{\partial \theta_K} \; \frac{\partial g_i(\boldsymbol{\xi})}{\partial \beta_0} \cdots \\ & \cdots \frac{\partial g_i(\boldsymbol{\xi})}{\partial \beta_p} \; \frac{\partial g_i(\boldsymbol{\xi})}{\partial \gamma_1} \cdots \frac{\partial g_i(\boldsymbol{\xi})}{\partial \gamma_l} \Bigg)^\top_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}, \end{aligned}$$

$$\nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} = \begin{pmatrix} \underbrace{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} g_i(\boldsymbol{\xi})}_{K\times K} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\beta}^\top} g_i(\boldsymbol{\xi})}_{K\times(p+1)} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\gamma}^\top} g_i(\boldsymbol{\xi})}_{K\times l} \\ \underbrace{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}^\top} g_i(\boldsymbol{\xi})}_{(p+1)\times K} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} g_i(\boldsymbol{\xi})}_{(p+1)\times(p+1)} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}^\top} g_i(\boldsymbol{\xi})}_{(p+1)\times l} \\ \underbrace{\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\theta}^\top} g_i(\boldsymbol{\xi})}_{l\times K} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\beta}^\top} g_i(\boldsymbol{\xi})}_{l\times(p+1)} & \underbrace{\frac{\partial^2}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^\top} g_i(\boldsymbol{\xi})}_{l\times l} \end{pmatrix}_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}}.$$

### 2.3.4 Computation of the gradient

To avoid heavy notation, we define the following scalar quantities:

$$\begin{aligned} \sum_{j=1}^{j(t_i)} h_0(s_j)\Delta_j &:= \omega_{0i}, \\ \sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j)\Delta_j &:= \omega_{0i}^k, \\ \sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j) b_l(s_j)\Delta_j &:= \omega_{0i}^{kl}. \end{aligned}$$

Deriving with respect to the B-spline coefficients gives us:

$$\begin{aligned} \frac{\partial}{\partial\theta_k} g_i(\boldsymbol{\xi}) \;=\; & \tilde{\delta}_i \left( b_k(t_i) - \exp(\mathbf{z}_i^\top\boldsymbol{\gamma}) \sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j)\Delta_j \right) \\ & + \exp(\beta_0 + \mathbf{x}_i^\top\boldsymbol{\beta}) \exp(\mathbf{z}_i^\top\boldsymbol{\gamma}) \exp\left( - \sum_{j=1}^{j(t_i)} h_0(s_j)\Delta_j \right)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})-1} \\ & \times \exp\left( - \sum_{j=1}^{j(t_i)} h_0(s_j)\Delta_j \right) \left( - \sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j)\Delta_j \right) \end{aligned}$$

$$
\begin{aligned}
= \quad & \tilde{\delta}_i \left( b_k(t_i) - \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) \sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j) \Delta_j \right) \\
& - \exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}) \exp\left( - \sum_{j=1}^{j(t_i)} h_0(s_j) \Delta_j \right)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \\
& \times \left( \sum_{j=1}^{j(t_i)} h_0(s_j) b_k(s_j) \Delta_j \right),
\end{aligned}
$$

so finally we have:

$$
\begin{aligned}
\frac{\partial}{\partial \theta_k} g_i(\boldsymbol{\xi}) \quad = \quad & \tilde{\delta}_i \left( b_k(t_i) - \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) \omega_{0i}^k \right) - \exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}) \\
& \times \exp\big( - \omega_{0i} \big)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \omega_{0i}^k, \quad k = 1, \ldots, K.
\end{aligned}
$$

The derivatives with respect to the $\boldsymbol{\beta}$ coefficients are:

$$
\begin{aligned}
\frac{\partial}{\partial \beta_m} g_i(\boldsymbol{\xi}) \quad = \quad & \tilde{\delta}_i x_{im} - \exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) \left( 1 - \exp\big( - \omega_{0i} \big)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \right) x_{im}, \\
& m = 0, \ldots, p \ \text{ with } x_{i0} = 1.
\end{aligned}
$$

For the derivatives with respect to the $\boldsymbol{\gamma}$ coefficients, we use the rule:

$$
\frac{d}{dx} a^{u(x)} = a^{u(x)} \log(a) \frac{d}{dx} u(x), \ a > 0.
$$

$$
\begin{aligned}
\frac{\partial}{\partial \gamma_s} g_i(\boldsymbol{\xi}) \quad = \quad & \tilde{\delta}_i \left( z_{is} - \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) z_{is} \omega_{0i} \right) + \exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}) \\
& \times \exp\big( - \omega_{0i} \big)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \big( - \omega_{0i} \big) \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) z_{is}
\end{aligned}
$$

and more compactly:

$$
\begin{aligned}
\frac{\partial}{\partial \gamma_s} g_i(\boldsymbol{\xi}) \quad = \quad & \tilde{\delta}_i z_{is} (1 - \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) \omega_{0i}) - \exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}) \\
& \times \exp\big( - \omega_{0i} \big)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \omega_{0i} z_{is}, \quad s = 1, \ldots, l.
\end{aligned}
$$

### 2.3.5 Computation of the Hessian

To compute the Hessian, we require the block matrices given below. Blocks 21, 31 and 32 are obtained by transposing blocks 12, 13 and 23.

$$\begin{aligned}
\text{Block 11} &: \frac{\partial^2}{\partial\theta_k\partial\theta_l}g_i(\boldsymbol{\xi}) \quad k=1,\ldots,K \;\; l=1,\ldots,K.\\
\text{Block 12} &: \frac{\partial^2}{\partial\theta_k\partial\beta_m}g_i(\boldsymbol{\xi}) \;\; k=1,\ldots,K \;\; m=0,\ldots,p.\\
\text{Block 13} &: \frac{\partial^2}{\partial\theta_k\partial\gamma_s}g_i(\boldsymbol{\xi}) \quad k=1,\ldots,K \;\; s=1,\ldots,l.\\
\text{Block 22} &: \frac{\partial^2}{\partial\beta_m\partial\beta_l}g_i(\boldsymbol{\xi}) \;\; m=0,\ldots,p \;\; l=0,\ldots,p.\\
\text{Block 23} &: \frac{\partial^2}{\partial\beta_m\partial\gamma_s}g_i(\boldsymbol{\xi}) \;\; m=0,\ldots,p \;\; s=1,\ldots,l.\\
\text{Block 33} &: \frac{\partial^2}{\partial\gamma_s\partial\gamma_v}g_i(\boldsymbol{\xi}) \quad s=1,\ldots,l \;\; v=1,\ldots,l.
\end{aligned}$$

**Block 11**

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_k\partial\theta_l}g_i(\boldsymbol{\xi}) &= \tilde{\delta}_i\left(-\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})\omega_{0i}^{kl}\right)-\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\boldsymbol{\gamma})\\
&\quad\times\left(\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})-1}\exp\big(-\omega_{0i}\big)\right.\\
&\qquad\left.\times\big(-\omega_{0i}^l\big)\omega_{0i}^k+\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\omega_{0i}^{kl}\right)
\end{aligned}$$

and more compactly for $k=1,\ldots,K$ and $l=1,\ldots,K$ we have:

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_k\partial\theta_l}g_i(\boldsymbol{\xi}) &= -\tilde{\delta}_i\,\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})\omega_{0i}^{kl}+\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\boldsymbol{\gamma})\\
&\quad\times\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\big(\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})\omega_{0i}^l\omega_{0i}^k-\omega_{0i}^{kl}\big).
\end{aligned}$$

**Block 12**

$$\begin{aligned}
\frac{\partial^2}{\partial\theta_k\partial\beta_m}g_i(\boldsymbol{\xi}) &= -\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\boldsymbol{\gamma})\,\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\gamma)}\omega_{0i}^k x_{im},\\
&\quad k=1,\ldots,K \;\; m=0,\ldots,p,\;\; x_{i0}=1.
\end{aligned}$$

**Block 13**

$$\begin{aligned}\frac{\partial^2}{\partial\theta_k\partial\gamma_s}g_i(\boldsymbol{\xi}) &= \tilde{\delta}_i\left(-\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})z_{is}\omega_{0i}^k\right)-\omega_{0i}^k\left(\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\boldsymbol{\gamma})\right.\\ &\quad\times z_{is}\,\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}+\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\boldsymbol{\gamma})\\ &\quad\left.\times\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\big(-\omega_{0i}\big)\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})z_{is}\right)\end{aligned}$$

and more compactly for $k=1,\ldots,K$ and $s=1,\ldots,l$ we have:

$$\begin{aligned}\frac{\partial^2}{\partial\theta_k\partial\gamma_s}g_i(\boldsymbol{\xi}) &= -\tilde{\delta}_i\,\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})z_{is}\omega_{0i}^k-\omega_{0i}^k\,\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\gamma)\\ &\quad\times\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\Big(z_{is}-\omega_{0i}\,\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})\;z_{is}\Big).\end{aligned}$$

**Block 22**

$$\begin{aligned}\frac{\partial^2}{\partial\beta_m\partial\beta_l}g_i(\boldsymbol{\xi}) &= -\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta})\left(1-\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\right)x_{im}x_{il},\\ &\quad m,l=0,\ldots,p\;\;x_{i0}=1.\end{aligned}$$

**Block 23**

$$\begin{aligned}\frac{\partial^2}{\partial\beta_m\partial\gamma_s}g_i(\boldsymbol{\xi}) &= \exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta})\,\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\big(-\omega_{0i}\big)\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})\\ &\quad\times z_{is}x_{im},\end{aligned}$$

and using the short notation:

$$\begin{aligned}\frac{\partial^2}{\partial\beta_m\partial\gamma_s}g_i(\boldsymbol{\xi}) &= -\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\boldsymbol{\gamma})\,\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\omega_{0i}z_{is}x_{im},\\ &\quad m=0,\ldots,p\;\;s=1,\ldots,l\;\;x_{i0}=1.\end{aligned}$$

**Block 33**

$$\begin{aligned}\frac{\partial^2}{\partial\gamma_s\partial\gamma_v}g_i(\boldsymbol{\xi}) &= -\tilde{\delta}_i\;z_{is}\,\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})z_{iv}\omega_{0i}-\omega_{0i}z_{is}\\ &\quad\times\left(\exp(\beta_0+\mathbf{x}_i^\top\boldsymbol{\beta}+\mathbf{z}_i^\top\boldsymbol{\gamma})z_{iv}\,\exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top\boldsymbol{\gamma})}\right.\end{aligned}$$

$$
\begin{aligned}
&+\exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \boldsymbol{z}_i^\top \boldsymbol{\gamma})\, \exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})}\big(-\omega_{0i}\big) \\
&\times \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) z_{iv}\Bigg)
\end{aligned}
$$

and more compactly:

$$
\begin{aligned}
\frac{\partial^2}{\partial\gamma_s \partial\gamma_v} g_i(\boldsymbol{\xi}) \;=\; & -\tilde{\delta}_i\, \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) \omega_{0i} z_{is} z_{iv} - \omega_{0i} z_{is}\, \exp(\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{\gamma}) \\
& \times \exp\big(-\omega_{0i}\big)^{\exp(\mathbf{z}_i^\top \boldsymbol{\gamma})} \Big( z_{iv} - \omega_{0i}\, \exp(\mathbf{z}_i^\top \boldsymbol{\gamma}) z_{iv} \Big), \\
& s, v = 1, \ldots, l.
\end{aligned}
$$

Defining the short notation $\sum_{i=1}^n \nabla g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(\mathbf{0})}} := \nabla g_{\boldsymbol{\xi}^{(0)}}$ for the sum of gradients and $\sum_{i=1}^n \nabla^2 g_i(\boldsymbol{\xi})|_{\boldsymbol{\xi}=\boldsymbol{\xi}^{(0)}} := \nabla^2 g_{\boldsymbol{\xi}^{(0)}}$ for the sum of Hessians, we obtain the following expression for the sum of the functions $g_i(\cdot)$ in (2.4) omitting the constant:

$$
\sum_{i=1}^n g_i(\boldsymbol{\xi}) = \boldsymbol{\xi}^\top \Big( \nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}} \boldsymbol{\xi}^{(0)} \Big) + \frac{1}{2} \boldsymbol{\xi}^\top \nabla^2 g_{\boldsymbol{\xi}^{(0)}} \boldsymbol{\xi}. \tag{2.5}
$$

Introducing (2.5) into (2.3), we recover:

$$
\begin{aligned}
\widetilde{p}_G(\boldsymbol{\xi}|\lambda, \mathcal{D}) \;\propto\; & \exp\Bigg( -\frac{1}{2} \boldsymbol{\xi}^\top \Big( Q(\lambda) - \nabla^2 g_{\boldsymbol{\xi}^{(0)}} \Big) \boldsymbol{\xi} \\
& + \boldsymbol{\xi}^\top \Big( \nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}} \boldsymbol{\xi}^{(0)} + Q(\lambda) \boldsymbol{\mu}_{\boldsymbol{\xi}} \Big) \Bigg).
\end{aligned} \tag{2.6}
$$

The above expression is a Gaussian density (up to a multiplicative constant) with mean and variance-covariance matrix that can be derived as follows. First take the logarithm of (2.6):

$$
\begin{aligned}
\log \widetilde{p}_G(\boldsymbol{\xi}|\lambda, \mathcal{D}) \;\doteq\; & -\frac{1}{2} \boldsymbol{\xi}^\top \Big( Q(\lambda) - \nabla^2 g_{\boldsymbol{\xi}^{(0)}} \Big) \boldsymbol{\xi} \\
& + \boldsymbol{\xi}^\top \Big( \nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}} \boldsymbol{\xi}^{(0)} + Q(\lambda) \boldsymbol{\mu}_{\boldsymbol{\xi}} \Big),
\end{aligned}
$$

where the symbol $\doteq$ denotes equality up to an additive constant.

To obtain the mean, we solve $\nabla_{\boldsymbol{\xi}} \log \widetilde{p}_G(\boldsymbol{\xi}|\lambda, \mathcal{D}) = 0$:

$$-\Big(Q(\lambda) - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\Big)\boldsymbol{\xi} + \Big(\nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)} + Q(\lambda)\boldsymbol{\mu}_{\boldsymbol{\xi}}\Big) = 0,$$

so the mean is:

$$\boldsymbol{\xi}^{(1)} = \Big(Q(\lambda) - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\Big)^{-1}\Big(\nabla g_{\boldsymbol{\xi}^{(0)}} - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\boldsymbol{\xi}^{(0)} + Q(\lambda)\boldsymbol{\mu}_{\boldsymbol{\xi}}\Big).$$

The precision is obtained by computing the negative of the Hessian matrix:

$$Q(\lambda)^{(1)} = -\nabla^2_{\boldsymbol{\xi}} \log \widetilde{p}_G(\boldsymbol{\xi}|\lambda, \mathcal{D}) = \Big(Q(\lambda) - \nabla^2 g_{\boldsymbol{\xi}^{(0)}}\Big).$$

Next, we repeat the Taylor expansion around $\boldsymbol{\xi}^{(1)}$ and continue to implement this iterative process in a Newton-Raphson type algorithm to converge towards a Gaussian approximation centered around the posterior mode of $p(\boldsymbol{\xi}|\lambda, \mathcal{D})$. Note that in each iteration, the inversion of a large dimensional matrix of the form $Q(\lambda) - \nabla^2 g$ is required. This is achieved by Householder transformations (Householder, 1958; Golub and Van Loan, 2012), a numerically stable tool for matrix inversion that achieves a QR decomposition through a sequence of orthogonal transformations of the input matrix.

### 2.3.6 Exploring the hyperparameter posterior

The next step consists in exploring the posterior distribution of the hyperparameter vector:

$$p(\boldsymbol{\eta}|\mathcal{D}) \propto \frac{\mathcal{L}(\boldsymbol{\xi}; \mathcal{D}) p(\boldsymbol{\xi}|\lambda) p(\lambda|\delta) p(\delta)}{p(\boldsymbol{\xi}|\lambda, \mathcal{D})}, \tag{2.7}$$

where $\mathcal{L}(\boldsymbol{\xi}; \mathcal{D})$ is the likelihood function. In order to avoid identifiability issues, we follow Bremhorst and Lambert (2016) and fix the last B-spline coefficient $\xi_K = \theta_K$ to a large value (say 10), denoted $c$. This forces the baseline survival function $S_0(\cdot)$ to be virtually zero at the end of the follow-up. Taking this constraint into account and using the Gaussian approximation scheme proposed in Section 2.3.3, we can approximate (2.7) as follows:

$$\widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) = \frac{\mathcal{L}(\boldsymbol{\xi};\mathcal{D})p(\boldsymbol{\xi}|\lambda)p(\lambda|\delta)p(\delta)|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*_{cc}(\lambda)}}{\widetilde{p}_G(\boldsymbol{\xi}_{-K}|\xi_K = c, \lambda, \mathcal{D})|_{\boldsymbol{\xi}_{-K}=\boldsymbol{\xi}^*_c(\lambda)}}, \tag{2.8}$$

where $\boldsymbol{\xi}^*_c(\lambda) \in \mathbb{R}^{\dim(\boldsymbol{\xi})-1}$ is the conditional posterior mean of the Gaussian approximation given $\xi_K = c$, and $\boldsymbol{\xi}^*_{cc}(\lambda) \in \mathbb{R}^{\dim(\boldsymbol{\xi})}$ corresponds to the vector $\boldsymbol{\xi}^*_c(\lambda)$ to which we add the constraint $c$ at position $K$. These two quantities are derived in Appendix B1. All the factors in (2.8) have mathematically closed forms, so that the approximated posterior of the hyperparameter vector can be extensively written as:

$$\begin{aligned}\widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) \propto \quad & \exp\left(\sum_{i=1}^{n} g_i(\boldsymbol{\xi}^*_{cc}(\lambda)) - \frac{1}{2}\Big(\boldsymbol{\xi}^*_{cc}(\lambda) - \boldsymbol{\mu}_{\boldsymbol{\xi}}\Big)^\top Q(\lambda)\Big(\boldsymbol{\xi}^*_{cc}(\lambda) - \boldsymbol{\mu}_{\boldsymbol{\xi}}\Big)\right) \\ & \times |Q(\lambda)|^{\frac{1}{2}} |\Sigma^*_c(\lambda)|^{\frac{1}{2}} \lambda^{\frac{\nu}{2}-1} \delta^{\frac{\nu}{2}+a_\delta-1} \exp\big(-\delta(b_\delta + \nu\lambda/2)\big). \end{aligned} \tag{2.9}$$

Note that $\delta$ can be integrated out from (2.9) to obtain the following approximated marginal posterior density of the penalty parameter:

$$\begin{aligned}\widetilde{p}(\lambda|\mathcal{D}) \propto \quad & \exp\left(\sum_{i=1}^{n} g_i(\boldsymbol{\xi}^*_{cc}(\lambda)) - \frac{1}{2}\Big(\boldsymbol{\xi}^*_{cc}(\lambda) - \boldsymbol{\mu}_{\boldsymbol{\xi}}\Big)^\top Q(\lambda)\Big(\boldsymbol{\xi}^*_{cc}(\lambda) - \boldsymbol{\mu}_{\boldsymbol{\xi}}\Big)\right) \\ & \times |Q(\lambda)|^{\frac{1}{2}} |\Sigma^*_c(\lambda)|^{\frac{1}{2}} \lambda^{\frac{\nu}{2}-1} \Big(b_\delta + \nu\lambda/2\Big)^{-(\nu/2+a_\delta)}. \end{aligned} \tag{2.10}$$

In addition, the conditional posterior of $\delta$ is given by $(\delta|\lambda, \mathcal{D}) \sim \mathcal{G}(\nu/2 + a_\delta, b_\delta + (\nu\lambda)/2)$ and does not directly depend on the data.

Next, our aim is to find a sub-region in the domain of $\widetilde{p}(\boldsymbol{\eta}|\mathcal{D})$ that supports most of the posterior probability mass. In that endeavor, we use an equidistant grid $\aleph_\lambda = \{\lambda_j\}_{j=1}^{m_1}$ of size $m_1 = 10$ in the domain of $\widetilde{p}(\lambda|\mathcal{D})$ that supports approximately 95% of the posterior mass. Then, for each point $\lambda_j \in \aleph_\lambda$, we construct a regular grid of length $m_2 = 5$ with starting and ending values corresponding to the 2.5th and 97.5th percentiles respectively of the $\mathcal{G}(\nu/2 + a_\delta, b_\delta + (\nu\lambda_j)/2)$ distribution. This enables us to construct a grid $\aleph_{\lambda,\delta} = \big(\lambda^{(m)}, \delta^{(m)}\big)_{m=1}^{M} \in \mathbb{R}^2_{++}$ with $M = m_1 \times m_2$ points that will be used to approximate the posterior distribution of spline and regression parameters.

### 2.3.7 Multivariate posterior of latent variables

In Sections 2.3.3-2.3.5, we have seen that the conditional posterior distribution of the vector of spline and regression parameters for a given $\xi_K = c$ and $\lambda$ can be approximated by a Gaussian density. By abuse of notation $\widetilde{p}_G(\boldsymbol{\xi}_{-K}|\xi_K = c, \lambda, \mathcal{D}) = \widetilde{p}_G(\boldsymbol{\xi}_{-K}|\lambda, \mathcal{D}) = \mathcal{N}_{\dim(\boldsymbol{\xi})-1}\big(\boldsymbol{\xi}_c^*(\lambda), \Sigma_c^*(\lambda)\big)$. The posterior joint distribution of $\boldsymbol{\xi}_{-K}$ can be written as:

$$\begin{aligned} p(\boldsymbol{\xi}_{-K}|D) &= \int_0^{+\infty}\int_0^{+\infty} p(\boldsymbol{\xi}_{-K}, \lambda, \delta|\mathcal{D})\ d\lambda\ d\delta \\ &= \int_0^{+\infty}\int_0^{+\infty} p(\boldsymbol{\xi}_{-K}|\lambda, \mathcal{D})\ p(\lambda, \delta|\mathcal{D})\ d\lambda\ d\delta. \quad (2.11) \end{aligned}$$

Given our grid coordinates and their associated weights $\Delta_m = \Delta_{\lambda^{(m)}} \times \Delta_{\delta^{(m)}}$ being the area of the parallelograms in the grid, we can approximate (2.11) by numerical integration:

$$\widetilde{p}(\boldsymbol{\xi}_{-K}|\mathcal{D}) = \sum_m \widetilde{p}_G(\boldsymbol{\xi}_{-K}|\lambda^{(m)}, \mathcal{D})\ \widetilde{p}(\lambda^{(m)}, \delta^{(m)}|\mathcal{D})\ \Delta_m.$$

The weights of the Gaussian densities in the sum can be normalized:

$$\omega_m = \frac{\widetilde{p}(\lambda^{(m)}, \delta^{(m)}|\mathcal{D})\ \Delta_m}{\sum_m \widetilde{p}(\lambda^{(m)}, \delta^{(m)}|\mathcal{D})\ \Delta_m},$$

to improve the approximation to the approximate joint posterior distribution of $\boldsymbol{\xi}_{-K}$, yielding:

$$\widehat{p}(\boldsymbol{\xi}_{-K}|\mathcal{D}) = \sum_m \omega_m\ \mathcal{N}_{\dim(\boldsymbol{\xi})-1}\left(\boldsymbol{\xi}_c^*(\lambda^{(m)}), \Sigma_c^*\big(\lambda^{(m)}\big)\right). \quad (2.12)$$

Equation (2.12) is a Gaussian mixture density with mean and variance-covariance matrix given by (see e.g. Frühwirth-Schnatter, 2006):

$$\begin{aligned} E(\boldsymbol{\xi}_{-K}|\mathcal{D}) &= \sum_m \omega_m\ \boldsymbol{\xi}_c^*(\lambda^{(m)}), \\ V(\boldsymbol{\xi}_{-K}|\mathcal{D}) &= \sum_m \omega_m\ \Sigma_c^*\big(\lambda^{(m)}\big) + \sum_m \omega_m\ \left(\boldsymbol{\xi}_c^*(\lambda^{(m)}) - E(\boldsymbol{\xi}_{-K}|\mathcal{D})\right) \\ &\quad \times \left(\boldsymbol{\xi}_c^*(\lambda^{(m)}) - E(\boldsymbol{\xi}_{-K}|\mathcal{D})\right)^\top. \end{aligned}$$

These quantities can be used to compute pointwise estimates and approximate credible intervals for variables in $\boldsymbol{\xi}$. In the next section, we show how to derive credible intervals for complex functionals of the vector of spline and regression parameters.

### 2.3.8 Credible intervals for latent variables

The flexibility of the Laplace-P-spline model can be exploited to compute pointwise credible intervals for complicated functions of spline and regression parameters and thus go beyond a marginal analysis. Construction of joint credible bands for subsets of $\boldsymbol{\xi}$ is also discussed in Sørbye and Rue (2011). In this section, we focus on the derivation of pointwise credible intervals for the baseline survival function $S_0(t)$ and for the population survival function given in (2.1). The "Delta method" will serve as the main mechanism to derive approximate credible intervals. Using a $\log(-\log(\cdot))$ transform of the baseline survival function, we recover:

$$G_0(\boldsymbol{\theta}^c|t) = \log\big(-\log S_0(t)\big) = \log\left(\sum_{j=1}^{j(t)} \exp\big(\boldsymbol{\theta}^\top \mathbf{b}(s_j)\big)\Delta_j\right), \quad (2.13)$$

where $\boldsymbol{\theta}^c = (\theta_1, \ldots, \theta_{K-1})^\top$ and $\theta_K = c$ as the last B-spline coefficient is fixed for identifiability purposes in the cure promotion time model. Using the strategy presented in Section 2.3.3, one has a Gaussian approximation to the conditional posterior of $\boldsymbol{\xi}$, namely $\widetilde{p}_G(\boldsymbol{\xi}|\lambda, \mathcal{D}) = \mathcal{N}_{\dim(\boldsymbol{\xi})}\big(\boldsymbol{\xi}^*(\lambda), \Sigma^*(\lambda)\big)$. Taking into account the constraint on the last B-spline coefficient $\theta_K = c$, we recover the following conditional posterior distribution for the vector $\boldsymbol{\theta}^c$:

$$\widetilde{p}_G(\boldsymbol{\theta}^c|\lambda, \mathcal{D}) = \mathcal{N}_{K-1}\big(\boldsymbol{\mu}_{\boldsymbol{\theta}^c}(\lambda), \Sigma_{\boldsymbol{\theta}^c}(\lambda)\big),$$

where $\boldsymbol{\mu}_{\boldsymbol{\theta}^c}(\lambda) = (\boldsymbol{\xi}^*_{c,1}(\lambda), \ldots, \boldsymbol{\xi}^*_{c,K-1}(\lambda))^\top$ and $\Sigma_{\boldsymbol{\theta}_c}(\lambda)$ is a $K-1$ dimensional square matrix corresponding to the first $K-1$ rows and columns of $\Sigma^*_c(\lambda)$. Using similar techniques as in Section 2.3.7, we can show that the approximated joint posterior distribution for the spline vector is:

$$\widehat{p}(\boldsymbol{\theta}^c|\mathcal{D}) = \sum_m \omega_m \, \mathcal{N}_{K-1}\big(\boldsymbol{\mu}_{\boldsymbol{\theta}^c}(\lambda^{(m)}), \Sigma_{\boldsymbol{\theta}^c}\big(\lambda^{(m)}\big)\big). \quad (2.14)$$

We thus recover in (2.14) a multivariate Gaussian mixture for which the mean and variance-covariance matrix are analytically known (cf. Section 2.3.7). Let us denote by $\boldsymbol{\theta}_{c,0}^m = \boldsymbol{\mu}_{\boldsymbol{\theta}^c}\big(\lambda^{(m)}\big)$ the mean of mixture component $m$. Using a first-order Taylor expansion of $G_0(\cdot|t)$ around $\boldsymbol{\theta}_{c,0}^m$, one gets:

$$G_{0,m}(\boldsymbol{\theta}^c|t) \approx G_0(\boldsymbol{\theta}_{c,0}^m|t) + (\boldsymbol{\theta}^c - \boldsymbol{\theta}_{c,0}^m)^\top \nabla_{\boldsymbol{\theta}^c} G_0(\boldsymbol{\theta}_{c,0}^m|t), \tag{2.15}$$

where $\nabla_{\boldsymbol{\theta}^c} G_0(\cdot|t)$ denotes the gradient of $G_0$ with respect to $\boldsymbol{\theta}^c$. Combining (2.14) with (2.15) suggests to approximate the marginal posterior of $G_{0,m}(\boldsymbol{\theta}^c|t)$ by the following univariate Gaussian distribution $(G_{0,m}(\boldsymbol{\theta}^c|t)|\mathcal{D}) \sim \mathcal{N}_1\Big(G_0(\boldsymbol{\theta}_{c,0}^m|t), \nabla_{\boldsymbol{\theta}^c} G_0(\boldsymbol{\theta}_{c,0}^m|t)^\top \Sigma_{\boldsymbol{\theta}^c}\big(\lambda^{(m)}\big) \nabla_{\boldsymbol{\theta}^c} G_0(\boldsymbol{\theta}_{c,0}^m|t)\Big)$. Accordingly, the posterior density of $G_0(\boldsymbol{\theta}^c|t)$ across all mixture components can in turn be approximated by the mixture $\Big(G_0(\boldsymbol{\theta}^c|t)|\mathcal{D}\Big) \approx \sum_m \omega_m \, \mathcal{N}_1\Big(G_0(\boldsymbol{\theta}_{c,0}^m|t), \nabla_{\boldsymbol{\theta}^c} G_0(\boldsymbol{\theta}_{c,0}^m|t)^\top \Sigma_{\boldsymbol{\theta}^c}\big(\lambda^{(m)}\big) \nabla_{\boldsymbol{\theta}^c} G_0(\boldsymbol{\theta}_{c,0}^m|t)\Big)$. It follows that a $(1-\alpha)\times 100\%$ credible interval can be obtained numerically by finding $C$ such that:

$$\int_C p\big(G_0(\boldsymbol{\theta}^c|t)|\mathcal{D}\big) dG_0(\boldsymbol{\theta}^c|t) = 1-\alpha.$$

To construct credible intervals for the population survival function given in (2.1) with a given profile of covariates, the procedure is the same except that the function of interest has the following form after a $\log(-\log(\cdot))$ transform:

$$\begin{aligned} G_0(\boldsymbol{\xi}_c|\mathbf{x},\mathbf{z},t) &= \beta_0 + \mathbf{x}^\top\boldsymbol{\beta} \\ &\quad + \log\left(1 - \exp\left(-\sum_{j=1}^{j(t)} \exp\big(\boldsymbol{\theta}^\top \mathbf{b}(s_j)\big)\Delta_j\right)^{\exp(\mathbf{z}^\top\boldsymbol{\gamma})}\right), \end{aligned}$$

where $\boldsymbol{\xi}_c = (\theta_1,\ldots,\theta_{K-1},\beta_0,\ldots,\beta_p,\gamma_1,\ldots,\gamma_l)^\top$. Using a first-order Taylor expansion of $G_0$ about the mean of each mixture component $\boldsymbol{\xi}_{c,0}^m = \boldsymbol{\xi}_c^*(\lambda^{(m)})$, it can be shown using our previous arguments that:

$$\begin{aligned} \big(G_0(\boldsymbol{\xi}_c|\mathbf{x},\mathbf{z},t)|\mathcal{D}\big) &\approx \sum_m \omega_m \, \mathcal{N}_1\Big(G_0(\boldsymbol{\xi}_{c,0}^m|\mathbf{x},\mathbf{z},t), \\ &\quad \nabla_{\boldsymbol{\xi}_c} G_0(\boldsymbol{\xi}_{c,0}^m|\mathbf{x},\mathbf{z},t)^\top \Sigma_c^*\big(\lambda^{(m)}\big) \nabla_{\boldsymbol{\xi}_c} G_0(\boldsymbol{\xi}_{c,0}^m|\mathbf{x},\mathbf{z},t)\Big). \end{aligned}$$

### 2.3.9 Cure prediction

Another quantity of interest in the promotion time cure model is the probability that a subject is cured given that (s)he has survived until a certain point in time $t$, say. Mathematically, one has:

$$P(T = +\infty | T \geq t, \mathbf{x}, \mathbf{z}) = \exp\Big( - \exp(\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}) S_0(t)^{\exp(\mathbf{z}^\top \boldsymbol{\gamma})} \Big).$$

Again, taking a $\log(-\log(\cdot))$ transform and using a first-order Taylor expansion, it can be shown that the resulting approximation to the posterior distribution is a Gaussian mixture, the only difference being in the gradient $\nabla_{\boldsymbol{\xi}_c} G_0(\boldsymbol{\xi}_{c,0}^m | \mathbf{x}, \mathbf{z}, t)$. The gradients required to obtain the above credible intervals have been computed with analytic forms provided in Appendix B2. The reader is also referred to the `curelps()` function of the **blapsr** package in Chapter 5 which implements the LPS method for inference in promotion time cure models.

## 2.4 Simulation study

The aim of this section is to implement a simulation study to assess the statistical performance of the LPS approach (cf. Section 2.3) in the promotion time cure model. The simulation setting is exactly the same as in Bremhorst and Lambert (2016) when the follow-up is sufficiently long except that we choose different cure and censoring rates, as well as more B-splines in the basis as enabled by the numerical efficiency of our method. Our methodology can also be applied when the follow-up period is not sufficiently long, provided that we account for identifiability issues. Indeed, as suggested in Bremhorst and Lambert (2016), when the follow-up of any susceptible subject is not long enough to observe its failure, then covariate effects are identifiable under the condition that the covariates are not simultaneously present in the probability to be cured and in the proportional hazards model parts.

The regressors consist in normal variates $x_{i1} = z_{i1} \sim \mathcal{N}(0, 1)$ for $i = 1, \ldots, n$ and discrete covariates following a Bernoulli distribution $x_{i2} = z_{i2} \sim \text{Bern}(0.5)$, $i = 1, \ldots, n$ to which we subtract 0.5 to obtain mean-centered covariates. The baseline distribution to generate latent event times is chosen to be a Weibull with mean 8 and variance 17.47. The regression coefficients in the Cox proportional hazards model are set to

$\gamma_1 = 0.40$ and $\gamma_2 = -0.40$, while the coefficients $\beta_0$, $\beta_1$ and $\beta_2$ are calibrated to get two different percentages for the proportion of cured subjects, namely around 20% and 30%. Finally, censoring is either governed by a uniform distribution on $[20, 25]$ or by a Weibull with shape parameter 3 and scale parameter 25. We redirect the reader to Bremhorst and Lambert (2016), Section 5.1 for more details concerning the generation of latent event times and censoring times. We use the Laplace-P-spline model with the above covariates and 25 cubic B-splines in $[0, t_u]$ where the upper bound of the follow-up is fixed to $t_u = 25$. A third order penalty on the coefficients of adjacent B-splines is used to counterbalance their flexibility. Furthermore, the last B-spline coefficient is fixed to $\theta_K = 10$ to translate the "sufficiently long follow-up hypothesis" in cure models, thereby avoiding identifiability problems. Simulations are performed on $S = 500$ replicates of sample size $n = 300$ and $n = 600$ with results reported in Tables 2.1 and 2.2.

| Cure | Setting | IMSE | Parameter | Bias | $CP_{90\%}$ | $CP_{95\%}$ | ESE | RMSE |
|---|---|---|---|---|---|---|---|---|
| 20% | 1 | 0.023 | $\beta_0 = 0.75$ | 0.022 | 90.4 | 95.4 | 0.101 | 0.103 |
| | | | $\beta_1 = 0.80$ | 0.035 | 89.8 | 94.0 | 0.119 | 0.124 |
| | | | $\beta_2 = -0.50$ | -0.039 | 89.4 | 93.8 | 0.175 | 0.179 |
| | | | $\gamma_1 = 0.40$ | -0.056 | 89.0 | 93.6 | 0.146 | 0.156 |
| | | | $\gamma_2 = -0.40$ | 0.050 | 89.0 | 93.2 | 0.218 | 0.223 |
| | 2 | 0.030 | $\beta_0 = 0.75$ | 0.016 | 90.8 | 95.4 | 0.112 | 0.113 |
| | | | $\beta_1 = 0.80$ | 0.045 | 91.8 | 96.0 | 0.137 | 0.144 |
| | | | $\beta_2 = -0.50$ | -0.036 | 93.0 | 97.2 | 0.200 | 0.203 |
| | | | $\gamma_1 = 0.40$ | -0.071 | 90.4 | 93.6 | 0.173 | 0.187 |
| | | | $\gamma_2 = -0.40$ | 0.047 | 92.4 | 97.0 | 0.248 | 0.252 |
| 30% | 1 | 0.017 | $\beta_0 = 0.30$ | 0.010 | 89.4 | 94.6 | 0.092 | 0.092 |
| | | | $\beta_1 = 1.00$ | 0.034 | 89.6 | 95.0 | 0.123 | 0.127 |
| | | | $\beta_2 = -0.75$ | -0.015 | 89.8 | 94.2 | 0.173 | 0.173 |
| | | | $\gamma_1 = 0.40$ | -0.057 | 88.6 | 94.6 | 0.143 | 0.154 |
| | | | $\gamma_2 = -0.40$ | 0.042 | 88.8 | 95.8 | 0.210 | 0.214 |
| | 2 | 0.025 | $\beta_0 = 0.30$ | -0.001 | 91.2 | 94.8 | 0.103 | 0.103 |
| | | | $\beta_1 = 1.00$ | 0.047 | 91.2 | 95.6 | 0.136 | 0.143 |
| | | | $\beta_2 = -0.75$ | -0.032 | 91.2 | 96.8 | 0.194 | 0.197 |
| | | | $\gamma_1 = 0.40$ | -0.072 | 89.0 | 94.0 | 0.175 | 0.189 |
| | | | $\gamma_2 = -0.40$ | 0.038 | 92.6 | 95.8 | 0.242 | 0.245 |

Table 2.1: Simulation results for $S = 500$ and $n = 300$. Setting 1: Censoring times generated from a $\mathcal{U}(20, 25)$ distribution; Setting 2: Censoring times generated from a Weibull$(3, 25)$ distribution.

We compute the integrated mean square error (IMSE), bias, empirical standard error (ESE) and root mean square error (RMSE) of the posterior (mixture) mean taken as a pointwise estimator of the regression coefficients in the cure probability and survival parts. Coverage probabilities (CP) of 90% and 95% credible intervals are also given. A negligible bias is observed across the different cure and censoring settings. Furthermore, the estimated coverage probabilities are reasonably close to the nominal values of 90% and 95% in each setting. We also notice that, as expected, the ESE and RMSE decrease with sample size.

| Cure | Setting | IMSE | Parameter | Bias | $CP_{90\%}$ | $CP_{95\%}$ | ESE | RMSE |
|---|---|---|---|---|---|---|---|---|
| 20% | 1 | 0.010 | $\beta_0 = 0.75$ | 0.016 | 88.4 | 94.2 | 0.069 | 0.071 |
| | | | $\beta_1 = 0.80$ | 0.029 | 91.0 | 95.4 | 0.076 | 0.081 |
| | | | $\beta_2 = -0.50$ | -0.016 | 91.0 | 94.2 | 0.119 | 0.119 |
| | | | $\gamma_1 = 0.40$ | -0.054 | 87.4 | 93.2 | 0.099 | 0.112 |
| | | | $\gamma_2 = -0.40$ | 0.039 | 92.0 | 95.8 | 0.143 | 0.148 |
| | 2 | 0.014 | $\beta_0 = 0.75$ | 0.009 | 91.6 | 96.6 | 0.074 | 0.074 |
| | | | $\beta_1 = 0.80$ | 0.033 | 89.2 | 94.8 | 0.099 | 0.104 |
| | | | $\beta_2 = -0.50$ | -0.020 | 89.6 | 95.6 | 0.140 | 0.141 |
| | | | $\gamma_1 = 0.40$ | -0.054 | 88.0 | 94.4 | 0.120 | 0.131 |
| | | | $\gamma_2 = -0.40$ | 0.037 | 89.6 | 95.6 | 0.173 | 0.177 |
| 30% | 1 | 0.010 | $\beta_0 = 0.30$ | 0.002 | 90.0 | 95.0 | 0.064 | 0.064 |
| | | | $\beta_1 = 1.00$ | 0.021 | 90.6 | 94.0 | 0.087 | 0.089 |
| | | | $\beta_2 = -0.75$ | -0.013 | 90.0 | 94.6 | 0.123 | 0.123 |
| | | | $\gamma_1 = 0.40$ | -0.037 | 88.6 | 93.8 | 0.104 | 0.110 |
| | | | $\gamma_2 = -0.40$ | 0.028 | 90.2 | 95.4 | 0.147 | 0.149 |
| | 2 | 0.015 | $\beta_0 = 0.30$ | 0.001 | 90.6 | 94.8 | 0.074 | 0.074 |
| | | | $\beta_1 = 1.00$ | 0.030 | 90.0 | 95.2 | 0.099 | 0.104 |
| | | | $\beta_2 = -0.75$ | -0.014 | 90.8 | 95.8 | 0.140 | 0.140 |
| | | | $\gamma_1 = 0.40$ | -0.055 | 85.8 | 92.6 | 0.125 | 0.137 |
| | | | $\gamma_2 = -0.40$ | 0.030 | 89.4 | 94.0 | 0.179 | 0.181 |

Table 2.2: Simulation results for $S = 500$ and $n = 600$. Setting 1: Censoring times generated from a $\mathcal{U}(20, 25)$ distribution; Setting 2: Censoring times generated from a Weibull$(3, 25)$ distribution.

Coverage estimates of 90% credible intervals for the baseline survival function are reported in Table 2.3 with an asterisk as superscript when the estimated coverage is incompatible with the nominal value at the 95% level. Globally, the estimated coverage probability across all quantiles is close to the 90% nominal value. Also, the poor coverage in the 5% quantile when $n = 300$ improves with growing sample size.

| n=300 | Cure | Cens. | Setting | 5% | 15% | 35% | 50% | 65% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | 20% | 1 | 90.6 | 91.6 | 89.8 | 88.2 | 89.0 | 88.4 | 88.6 | 88.0 |
| | 20% | 23% | 2 | 87.8 | 91.0 | 88.4 | 89.0 | 89.4 | 91.2 | 90.4 | 92.0 |
| | 30% | 30% | 1 | 87.8 | 89.0 | 91.0 | 91.4 | 90.6 | 90.0 | 88.2 | 87.2* |
| | 30% | 33% | 2 | 82.6* | 88.6 | 88.4 | 89.0 | 88.6 | 89.4 | 88.0 | 90.8 |
| n=600 | Cure | Cens. | Setting | 5% | 15% | 35% | 50% | 65% | 75% | 85% | 95% |
| | 20% | 20% | 1 | 90.4 | 92.8* | 88.4 | 90.4 | 91.2 | 91.4 | 88.8 | 88.4 |
| | 20% | 23% | 2 | 87.6 | 89.0 | 88.8 | 89.8 | 87.0* | 87.8 | 85.2* | 88.0 |
| | 30% | 30% | 1 | 91.8 | 92.6 | 90.0 | 89.8 | 91.0 | 91.8 | 90.2 | 90.4 |
| | 30% | 33% | 2 | 86.6* | 91.0 | 88.8 | 87.2* | 86.4* | 87.6 | 88.2 | 88.6 |

Table 2.3: Coverage estimates of 90% credible intervals using first-order Taylor approximations for the baseline survival function at selected quantiles $(5\%, 15\%, 35\%, 50\%, 65\%, 75\%, 85\%, 95\%)$ of $T$ under the promotion time cure model. Setting 1: Censoring times generated from a $\mathcal{U}(20, 25)$ distribution; Setting 2: Censoring times generated from a Weibull$(3, 25)$ distribution.

In Table 2.4, we report the coverage estimates of 90% credible intervals for the population survival function at selected quantiles with the continuous covariate fixed to 0.1 and the binary covariate to 0.5. Again, the constructed credible intervals show good performances even for the 5% and 95% quantiles.

| n=300 | Cure | Cens. | Setting | 5% | 15% | 35% | 50% | 65% | 75% | 85% | 95% |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 20% | 20% | 1 | 89.8 | 88.4 | 90.2 | 91.8 | 91.8 | 89.8 | 88.6 | 89.6 |
| | 20% | 23% | 2 | 87.6 | 89.2 | 90.0 | 90.8 | 91.2 | 90.8 | 90.6 | 92.8* |
| | 30% | 30% | 1 | 90.8 | 90.2 | 89.0 | 89.4 | 90.8 | 92.2 | 91.0 | 90.8 |
| | 30% | 33% | 2 | 90.2 | 89.4 | 89.6 | 89.6 | 91.0 | 90.6 | 90.2 | 91.4 |

Table 2.4: Coverage estimates of 90% credible intervals using first-order Taylor approximations for the population survival function at selected quantiles of $T$ when $x = 0.1$ and $z = 0.5$. Setting 1: Censoring times generated from a $\mathcal{U}(20, 25)$ distribution; Setting 2: Censoring times from a Weibull$(3, 25)$ distribution.

In Figure 2.1, the solid line is the target baseline survival distribution for the susceptible corresponding to the Weibull with mean 8 and standard deviation 4.18. The gray curves are estimates of $S_0(t)$ under each replicate and the dashed curve corresponds to the pointwise median of the 500 estimated baseline survival functions. Globally, we can say that the Laplace-P-spline approach provides accurate estimates of the baseline distribution with little variability around the target.

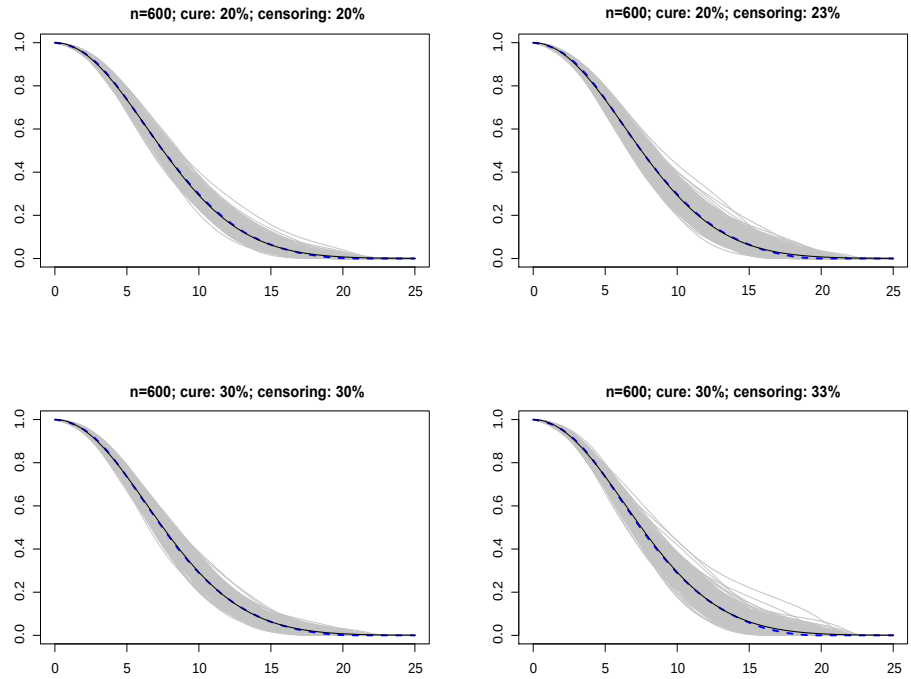



Figure 2.1: Estimation of the baseline distribution $S_0(t)$ for $S = 500$ replications, (one gray curve per dataset) and sample size $n = 600$. In the left column the censoring rate is governed by a $\mathcal{U}(20, 25)$ distribution and in the right column it is governed by a Weibull$(3, 25)$ distribution. The solid line is the true function and the dashed line is the pointwise median of the 500 estimated curves.

For the sake of assessing the algorithmic performance of our approach, we implement a computational speed comparison with a MCMC algorithm. The competitor is taken to be a Metropolis-within-Gibbs algorithm with blockwise sampling for which we compute a chain of length $23,000$ and a burnin of length $3,000$ to explore the joint posterior of $\boldsymbol{\xi}$.

Under the same simulation settings, we observe a computational speed-up of a factor 15 with the Laplace-P-spline approach. It is also worth noting that most of the computational intensive tasks in our MCMC algorithm are written in **Fortran** language and called via **R**, while our Laplace-P-spline algorithm is exclusively coded in **R** language, such that the mentioned computational gain is conservative and under-evaluated.

## 2.5 Real data analysis

### 2.5.1 Application to malignant melanoma data

In this section, we illustrate the LPS methodology with the analysis of a malignant melanoma survival dataset (Andersen et al., 1993). The dataset concerns 205 patients affected by skin cancer and operated for malignant melanoma at Odense University Hospital in Denmark during 1962-1977. The response of interest is the time (in years) elapsed between operation and death from malignant melanoma. The covariates are *Age* at operation (in years), *Gender* (1=M, 0=F), *Tumor Thickness* (in mm) and a dichotomous factor (*Ulcer*) indicating presence of ulceration (1=presence, 0=absence) at baseline. Among the 205 patients, 57 died from malignant melanoma while the remaining 148 are right censored. This dataset was first investigated using single-factor analysis techniques (Drzewiecki, Ladefoged and Christensen, 1980; Drzewiecki, Christensen, Ladefoged and Poulsen, 1980) and a Cox regression model (Drzewiecki and Andersen, 1982). More recently, Li and Lin (2009) used the melanoma dataset to illustrate a semiparametric mixture model, while Chyong-Mei and Chen-Hsin (2016) implemented it to highlight a heteroscedastic transformation cure model. We propose to use the promotion time approach in which the covariates will simultaneously affect the probability of being cured as well as the time to event for susceptible subjects. The use of the same covariates in the two parts of the model is not problematic when it comes to inference as a plateau is observed in the Kaplan-Meier curve, suggesting a sufficiently long follow-up.

We use 50 B-splines on $[0, t_u]$ and follow a common choice in the literature to specify $t_u$ as the largest observed survival time (here $t_u = 15.236$). The algorithm in pure **R** code takes $\approx 15$ seconds to obtain estimates for all B-spline coefficients, the posterior standard deviation ($sd_{post}$) and 95% quantile-based credible intervals for the regression coefficients.

The first estimation results (not detailed here) suggest that ulceration has a significant effect on the probability to be "cured", i.e. of not observing a relapse whatever the duration of the follow-up, while *Tumor Thickness* significantly affects the time to event for susceptible subjects. The model is then estimated a second time (see Table 2.5) by omitting *Age* and *Gender* as they have no significant effect in the model (conditionally on thickness and ulceration). The results suggest that ulceration has a negative effect on the probability to be cured. Furthermore, *Tumor Thickness* at time of surgery is an important factor affecting the time necessary to detect a new tumor. In fact, a large tumor at baseline may already be a sign of metastatic occurrence such that after an incomplete removal of cancer cells, a relapse is more likely to occur in a shorter period of time (conditionally on other covariates), although it has no significant effect on the long term risk of relapse.

| | Parameters | Estimates | CI 95% | $\text{sd}_{post}$ |
|---|---|---|---|---|
| $\phi(\mathbf{x})$ | Intercept | -1.589 | [-2.226; -0.948] | 0.326 |
| | Thickness | 0.067 | [-0.010; 0.142] | 0.039 |
| | Ulcer | 1.096 | [ 0.370; 1.819] | 0.370 |
| $1 - S_0(t)^{\exp(\mathbf{z}^\top \boldsymbol{\gamma})}$ | Thickness | 0.111 | [ 0.017; 0.201] | 0.047 |
| | Ulcer | 0.327 | [-0.619; 1.278] | 0.484 |

Table 2.5: Posterior mixture mean for each regression parameter using 50 B-splines for the baseline log-hazard in the reduced model, the 95% quantile-based approximate credible intervals (CI) and the posterior standard deviation. $\phi(\mathbf{x})$ is minus the log of the probability to be cured and $1 - S_0(t)^{\exp(\mathbf{z}^\top \boldsymbol{\gamma})}$ represents the time necessary for a cell to produce a detectable tumor mass.

Our analysis also investigates to what extent ulceration affects the probability that a patient is cured given that (s)he has survived until a given time reference $t$. This conditional probability is estimated in Table 2.6 for a median value of *Tumor Thickness* (1.94 mm) and approximate 90% credible intervals are also provided. Figure 2.2 gives a graphical representation of the pointwise and set estimates for these probabilities. We see that in presence of an ulcer, the estimated probability that a patient is cured given that (s)he has survived until $t$ is smaller than the estimate corresponding to ulcer absence, regardless of the reference time values. Also, we see that the estimated probabilities increase with

$t$, simply corroborating the idea that the longer a patient has survived (with or without an ulcer), the larger his/her chances of being cured.

| | Probability to be cured given that T≥ t | | | |
|---|---|---|---|---|
| | No Ulceration | | Ulceration | |
| $t$ | Estimates | CI 90% | Estimates | CI 90% |
| 2 | 0.812 | [0.697; 0.887] | 0.538 | [0.404; 0.676] |
| 4 | 0.855 | [0.735; 0.924] | 0.631 | [0.491; 0.799] |
| 6 | 0.904 | [0.773; 0.961] | 0.745 | [0.596; 0.912] |
| 8 | 0.944 | [0.793; 0.986] | 0.849 | [0.690; 0.974] |

Table 2.6: Pointwise estimates and approximate 90% credible intervals for the conditional probability to be cured given that $T \geq t$ for $t \in \{2, 4, 6, 8\}$ (in years) with and without ulceration and for a median value of *Tumor Thickness*.

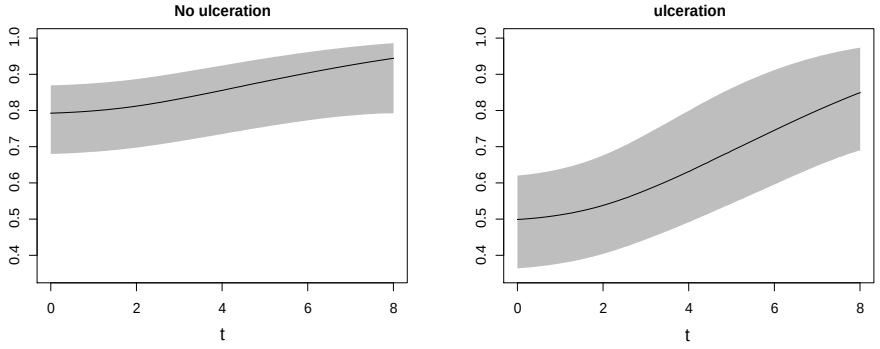



Figure 2.2: Evolution over time $t$ of the probability to be cured $P(T = +\infty | T \geq t, TT = 1.94)$ for a median *Tumor Thickness* (TT) represented by the solid line for two scenarios, no ulceration (left) and ulceration (right). The gray surface represents the approximate 90% pointwise credible intervals.

### 2.5.2 Application to oropharynx carcinoma data

We implement a second data analysis using data from Kalbfleisch and Prentice (2011) on oropharynx carcinoma. The dataset comes from a clinical trial achieved by the Radiation Therapy Oncology group involving patients from six clinics suffering from squamous cell carcinoma located in different sites of the mouth and throat. There are 195 patients randomly assigned in two arms at the moment of entry in the study:

(1) radiation therapy alone (standard) or (2) radiation therapy with a chemotherapeutic agent (special). To highlight the use of our model, we focus on 130 patients (among which 38 are censored) with cancer located in the pharyngeal tongue and tonsillar fossa part of the mouth. We retain the covariates *Age*, *Sex* (1=M, 0=F), *Treatment* (1=special or 0=standard) and tumor staging (*Tumor*) for explaining survival times. For tumor staging, we follow Lopes and Bolfarine (2012) and categorize the variable as *Tumor*=0 if primary tumor and *Tumor*=1 if massive tumor. As in the previous application, we use 50 B-splines in the interval ranging from 0 to the largest observed survival time measured in years (4.99).

The main objective of this analysis is to assess the effect of the two types of treatments on survival times of patients accounting for tumor staging. The estimated Kaplan-Meier curve given in Figure 2.3 (left panel) shows a plateau, indicating the presence of a cured fraction and thus justifying our choice to let the covariates influence jointly the probability to be cured and the time to event for susceptible patients. In Table 2.7, we report the posterior mixture mean, the 90% quantile-based approximate credible interval and the posterior standard deviation. We see that *Tumor* is the only variable having a negative and significant effect on the probability to be cured, such that presence of a massive tumor decreases the chances of being cured from oropharynx cancer. In addition, *Treatment* has a significant impact on a recurrence timing (conditionally on other covariates), but no significant effect on the risk of recurrence in the long term.

In Figure 2.3 (right panel), we show the estimated population survival functions when the model is estimated without *Age* and *Sex* and by only accounting for the effects of *Tumor* and *Treatment* on, respectively, the cancer recurrence probability and on its timing for susceptible subjects. Whether we consider a standard or special treatment, we see that the risk of cancer recurrence only changes with tumor status, with a higher risk when a massive tumor is present as compared to a primary tumor. In addition, we see that the type of treatment mainly impacts the speed at which the recurrence arises for susceptible patients. Finally, Figure 2.4 compares the estimated population survival functions obtained with the Laplace-P-spline model (blue curve) against Kaplan-Meier curves (in

black) for each tumor staging and treatment configuration. In each situation, the Laplace-P-spline model provides survival curves that appear to be appropriate smoothed versions of the Kaplan-Meier estimates.

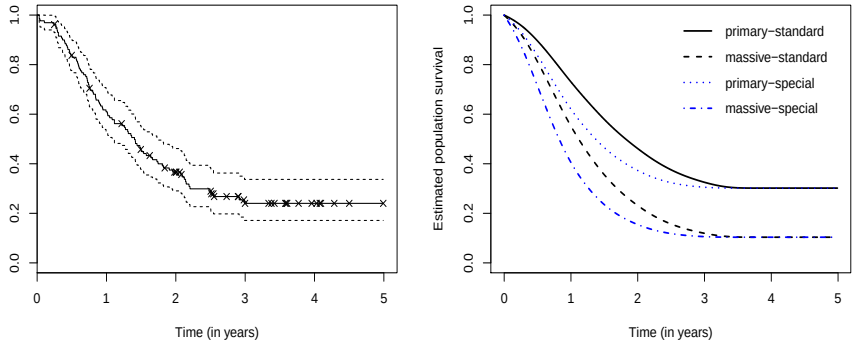



Figure 2.3: (Left panel) Kaplan-Meier estimated curve from the oropharynx dataset. A cross indicates a censored patient. (Right panel) Estimated population survival functions for different tumor-treatment configurations.

| | Parameters | Estimates | CI 90% | $\text{sd}_{post}$ |
|---|---|---|---|---|
| $\phi(\mathbf{x})$ | Intercept | -0.323 | [-1.436; 0.788] | 0.676 |
| | Age | 0.008 | [-0.010; 0.025] | 0.011 |
| | Sex | 0.291 | [-0.148; 0.727] | 0.266 |
| | Tumor | 0.510 | [ 0.020; 0.998] | 0.297 |
| | Treatment | -0.315 | [-0.733; 0.101] | 0.253 |
| $1-S_0(t)^{\exp(\mathbf{z}^\top\boldsymbol{\gamma})}$ | Age | 0.006 | [-0.012; 0.022] | 0.010 |
| | Sex | -0.704 | [-1.253; -0.156] | 0.334 |
| | Tumor | 0.356 | [-0.321; 1.031] | 0.411 |
| | Treatment | 0.763 | [ 0.231; 1.292] | 0.323 |

Table 2.7: Posterior mixture mean, 90% quantile-based approximate credible interval (CI) and posterior standard deviation for each regression parameter of the promotion time model.

## 2.6 Discussion

In this chapter, we introduced a novel methodology for fast Bayesian inference in semiparametric survival models by coupling P-splines with Laplace approximations. Our approach opens up promising perspectives

for inference in cure survival models as it enables to obtain pointwise and set estimators for non-trivial functions of latent variables with a drastic computational speed-up as compared to existing MCMC methods.

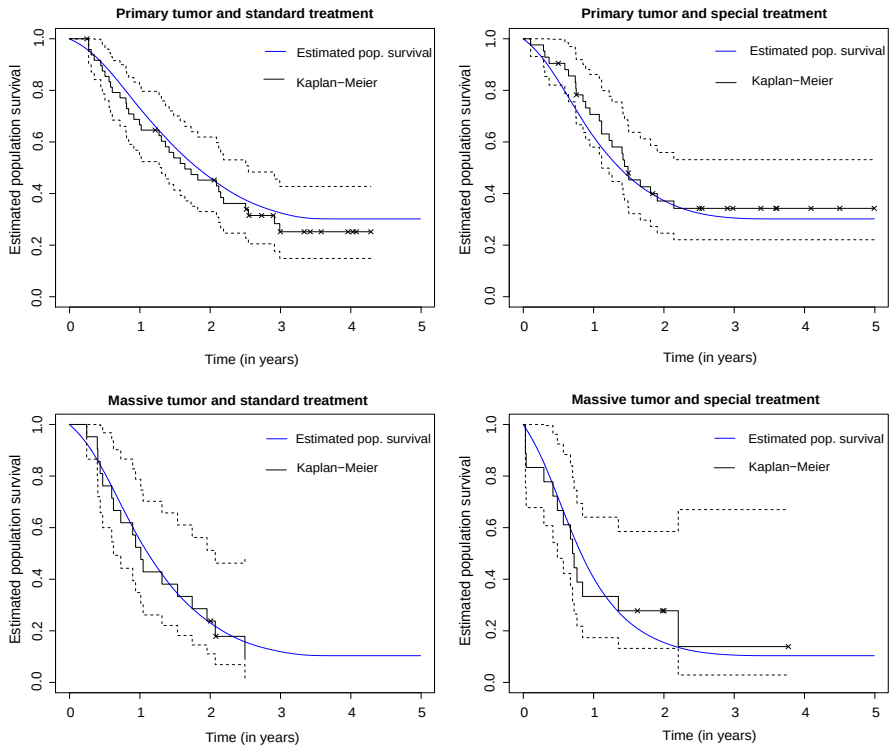



Figure 2.4: Estimated population survival functions from the Laplace-P-spline model (blue) versus Kaplan-Meier curves (black) and their 95% confidence interval (dashed) for different tumor status and treatment.

Even though the Laplace approximation mechanisms presented in this work share some similarities with the classic INLA approach (Rue et al., 2009), our methodology is sharply contrasted with the latter in many respects. In particular, our modeling strategy involves a specification of the prior of the roughness penalty parameter that is robust to the choice of hyperparameters (Jullion and Lambert, 2007). In the standard INLA approach, that concern is not addressed with the implication that posterior estimation can be sensitive towards the hyperparameter prior chosen by the user. In addition, our work goes beyond the treatment of univariate posterior marginal distributions by deriving reliable approximations to the joint posterior distributions of spline and regression parameters

for which the mean and covariance matrix have known analytic forms. Another major difference is that the dimension of our latent vector only grows with the number of regressors and not with sample size, impacting directly the underlying algorithmic efficiency when dealing with large datasets.

A practical limitation may arise when dealing with a hyperparameter vector of large dimension. This might be the case for instance in additive regression models, where the number of roughness penalty parameters is equal to the number of smooth functions to be estimated, implying a much larger computational cost for the grid strategy recommended in Section 2.3.6. However, even for a large number of hyperparameters, we expect our approach to be much faster than existing MCMC techniques, which would require long computation times in such situations.

# CHAPTER 3

# Laplace-P-splines for approximate Bayesian inference in additive models

This chapter is based on the discussion paper: Gressani, O. and Lambert, P. (2020a). The Laplace-P-spline methodology for fast approximate Bayesian inference in additive partial linear models, *ISBA Discussion papers, DP-2020/20.* http://hdl.handle.net/2078.1/230728

## 3.1 Motivation

Multiple linear regression is among the cornerstones of statistical model building. Whether from a descriptive or inferential perspective, it is certainly the most widespread approach to analyze the influence of a collection of explanatory variables on a response. The straightforward interpretability in conjunction with the simple and elegant mathematics of least squares created room for a well appreciated toolbox with an ubiquitous presence in various scientific fields. There are two rather strong assumptions underlying the multiple linear regression model that restrain their use in most practical applications. First a linear dependence is assumed between the mean response and each covariate. Second, the response variable is often assumed to be continuous with a Gaussian distribution. To circumvent these limitations, several extensions have been proposed in the literature giving birth to more general model classes. In this chapter, the linear dependence assumption of the response variable

with respect to the covariates is relaxed and replaced by an additive architecture of univariate smooth functions of predictor variables. We keep the assumption of a continuous and Gaussian response until Chapter 4, where generalized additive models are introduced.

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

We adapt the Laplace-P-spline (LPS) approach to additive models with Gaussian errors and develop a fast and flexible methodology for approximate Bayesian inference in this model class. Great efforts have been invested in the derivation of analytical formulas for the gradient and Hessian of the posterior penalty vector, which offers a nonnegligible computational gain when exploring the posterior penalty space. Moments of a skew-normal family of random variables are used to accurately approximate the posterior distribution of penalty parameters, thereby capturing the inherent asymmetric patterns. The `amlps()` routine of the **blapsr** package (cf. Chapter 5) is based on the methodology presented in this chapter and can be used to fit additive partial linear models with LPS. In Section 3.2, the Bayesian-P-spline additive model is introduced and a method is proposed to overcome identifiability problems. In Section 3.3 the priors on the penalty parameters are defined and the likelihood function is derived together with the conditional posterior distribution of

the vector of regression and spline parameters. Section 3.4 is dedicated to the posterior of the hyperparameter vector. The nuisance parameters are integrated out and the gradient and Hessian of the penalty vector are obtained in closed-form. Section 3.5 proposes a strategy to explore the posterior penalty vector based on skew-normal matching moments. In Section 3.6 the approximate posterior of the latent vector is derived and Section 3.7 covers the derivation of pointwise credible intervals for marginal latent variables and smooth functions. Section 3.8 implements a simulation study to assess the performance of the proposed methodology and Section 3.9 is devoted to the application of LPS on mortality data. Finally, Section 3.10 concludes the chapter.

## 3.2 The Bayesian P-spline additive model

### 3.2.1 Additive structure and priors

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

$$y_i = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + f_1(x_{i1}) + \cdots + f_q(x_{iq}) + \varepsilon_i, \qquad (3.1)$$

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

$$f_j(x_{ij}) = \sum_{k=1}^{K} \theta_{jk} b_{jk}(x_{ij}), \quad j = 1, \ldots, q, \qquad (3.2)$$

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

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

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

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

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

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

### 3.2.2 Identifiability

The additive model in (3.1) suffers from an identifiability issue. This can be easily illustrated through the simple model $E(y) = \beta_0 + f(x)$. Assume our goal is to estimate the expected value $E(y)$ from a sample $\{(x_i, y_i)\}_{i=1}^n$. Let $c$ be any arbitrary constant and denote by $\tilde{\beta}_0 = \beta_0 - c$ and $\tilde{f}(x) = f(x)+c$. It follows that $E(y) = \tilde{\beta}_0+\tilde{f}(x)$ for any $c$, such that there exists an infinite number of configurations for $\tilde{\beta}_0$ and $\tilde{f}$ yielding the same expected value, meaning that the model "parameters" cannot be uniquely identified and estimated for a given data set. To reach an identifiable model, we follow an approach similar to Durbán and Currie (2003) and define the centered B-spline matrices:

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

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

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

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

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

and according to (3.2), the identifiability constraint is translated as $\tilde{f}_j(x_{ij}) = f_j(x_{ij}) - L^{-1}\sum_{l=1}^L f_j(\breve{x}_{lj})$, i.e. the additive functional components are centered around their average value (computed over a fine equidistant grid). To see how this solves the identifiability problem, consider again the simple model $E(y) = \beta_0 + f(x) - \bar{f}$, with $\bar{f} = L^{-1}\sum_{l=1}^L f(\breve{x}_l)$ the average of $f$ over a fine grid. Adding $c$ to the intercept and subtracting the same amount from $f$ yields:

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

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

**Proposition (Rank-reduction due to centering)**

The rank of the centered B-spline matrix $\widetilde{B}_j$ is $K - 1$.

**Proof:**

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

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

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

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

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

$$\begin{aligned}
\text{rank}(\widetilde{B}_j) &= \text{rank}(I_K - \mathcal{B}) \\
&= \text{Tr}(I_K - \mathcal{B})
\end{aligned}$$

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

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

$$
\tilde{b}_j(x_{ij})^\top\boldsymbol{\theta}_j = \sum_{k=1}^{K-1}\theta_{jk}b_{jk}(x_{ij}) - \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K-1}\theta_{jk}b_{jk}(\breve{x}_{lj}). \tag{3.3}
$$

With the identifiability constraint and the centered $Z$ matrix, the additive model in (3.1) can be expressed compactly as:

$$
\begin{aligned}
\mathbf{y} &= Z\boldsymbol{\beta} + \widetilde{B}_1\boldsymbol{\theta}_1 + \cdots + \widetilde{B}_q\boldsymbol{\theta}_q + \boldsymbol{\varepsilon} \\
&= B\boldsymbol{\xi} + \boldsymbol{\varepsilon}, \qquad (3.4)
\end{aligned}
$$

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

## 3.3 Conditional posterior for $\boldsymbol{\xi}$

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

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

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

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

The conditional posterior distribution of the vector of regression and spline parameters can be obtained as follows:

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

Using the previously specified prior for $\boldsymbol{\xi}$ and likelihood, we get:

$$
\begin{aligned}
p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) &\propto \exp\Big(-\frac{\tau}{2}\big(\mathbf{y}^\top\mathbf{y} - 2\mathbf{y}^\top B\boldsymbol{\xi} + \boldsymbol{\xi}^\top B^\top B\boldsymbol{\xi}\big) - \frac{\tau}{2}\boldsymbol{\xi}^\top Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\boldsymbol{\xi}\Big)\\
&\propto \exp\Big(\tau\mathbf{y}^\top B\boldsymbol{\xi} - \frac{\tau}{2}\boldsymbol{\xi}^\top(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})\boldsymbol{\xi}\Big). \qquad (3.5)
\end{aligned}
$$

Note that (3.5) is the exponential of a quadratic form in $\boldsymbol{\xi}$ and can be written as a Gaussian distribution. To find the mean vector we solve $\nabla_{\boldsymbol{\xi}}\log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) = 0$ and obtain $\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} = (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}B^\top\mathbf{y}$. The precision is $-\nabla^2_{\boldsymbol{\xi}}\log p(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) = \tau(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})$ and so the conditional

posterior of the vector of regression and B-spline coefficients is characterized by the following Gaussian distribution:

$$(\boldsymbol{\xi}|\boldsymbol{\lambda},\tau,\mathcal{D}) \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}\left(\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}, \tau^{-1}(B^\top B + Q^{\boldsymbol{\lambda}}_{\boldsymbol{\xi}})^{-1}\right). \tag{3.6}$$

## 3.4 Posterior of the penalty vector

### 3.4.1 Objectives

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

### 3.4.2 Posterior of the full hyperparameter vector

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

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

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

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

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

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

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

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

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

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

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

Let us define the scalar function $\phi(\boldsymbol{\lambda}) := \frac{1}{2}\mathbf{y}^\top(I_n - B(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}B^\top)\mathbf{y}$ (see Appendix C1 for efficient evaluation of this function) and write compactly:

$$p(\boldsymbol{\eta}|\mathcal{D})\big|_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}}$$
$$\propto |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}|Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}}\left(\prod_{j=1}^{q}\delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\Big(-\delta_j\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)\Big)\right)$$
$$\times\left(\prod_{j=1}^{q}\lambda_j^{(\frac{\nu}{2}-1)}\right)\tau^{(\frac{n}{2}-1)}\exp\big(-\tau\phi(\boldsymbol{\lambda})\big). \tag{3.7}$$

### 3.4.3 Integration with respect to the nuisance parameters

The nuisance parameter $\tau$ can be integrated out from (3.7) as expression $\tau^{(\frac{n}{2}-1)}\exp\big(-\tau\phi(\boldsymbol{\lambda})\big)$ is up to a multiplicative constant the density of a Gamma distribution parameterized by $\mathcal{G}(n/2, \phi(\boldsymbol{\lambda}))$. Hence, $\int_0^{+\infty}\tau^{(\frac{n}{2}-1)}\exp\big(-\tau\phi(\boldsymbol{\lambda})\big)\,d\tau = \Gamma\big(\frac{n}{2}\big)\phi(\boldsymbol{\lambda})^{-\frac{n}{2}}$, where $\Gamma(\cdot)$ is the Gamma function. Using this property, the integral is given by:

$$p(\boldsymbol{\lambda},\boldsymbol{\delta}|\mathcal{D}) = \int_0^{+\infty} p(\boldsymbol{\eta}|\mathcal{D})\big|_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}}\,d\tau$$
$$\propto |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}|Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{\frac{1}{2}}\left(\prod_{j=1}^{q}\delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\Big(-\delta_j\Big(b_\delta+\frac{\nu}{2}\lambda_j\Big)\Big)\right)$$
$$\times\left(\prod_{j=1}^{q}\lambda_j^{(\frac{\nu}{2}-1)}\right)\phi(\boldsymbol{\lambda})^{-\frac{n}{2}}. \tag{3.8}$$

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

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

such that (3.8) becomes:

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{\delta}|\mathcal{D}) \quad &\propto \quad |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} \left(\prod_{j=1}^{q} \lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right) \\ &\times \left(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j\Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big)\right) \phi(\boldsymbol{\lambda})^{-\frac{n}{2}}. \quad (3.9) \end{aligned}$$

The posterior in (3.9) can be integrated with respect to $\delta_j$ successively for $j = 1, \ldots, q$ since $\delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j\Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big)$ is (up to a multiplicative constant) a Gamma density parameterized by $\mathcal{G}(\frac{\nu}{2} + a_\delta, b_\delta + \frac{\nu}{2}\lambda_j)$, so:

$$\begin{aligned} &\int_0^{+\infty} \cdots \int_0^{+\infty} \left(\prod_{j=1}^{q} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j\Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big)\right) d\delta_1 \ldots d\delta_q \\ = &\prod_{j=1}^{q} \left(\int_0^{+\infty} \delta_j^{(\frac{\nu}{2}+a_\delta-1)} \exp\Big(-\delta_j\Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)\Big)\ d\delta_j\right) \\ = &\left(\Gamma\Big(\frac{\nu}{2} + a_\delta\Big)\right)^q \left(\prod_{j=1}^{q} \Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)^{-(\frac{\nu}{2}+a_\delta)}\right) \quad (3.10) \end{aligned}$$

and the posterior of the penalty vector is:

$$\begin{aligned} p(\boldsymbol{\lambda}|\mathcal{D}) &= \int_0^{+\infty} \cdots \int_0^{+\infty} p(\boldsymbol{\lambda}, \boldsymbol{\delta}|\mathcal{D})\ d\delta_1 \ldots d\delta_q \\ &\propto |B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} \left(\prod_{j=1}^{q} \lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right) \left(\prod_{j=1}^{q} \Big(b_\delta + \frac{\nu}{2}\lambda_j\Big)^{-(\frac{\nu}{2}+a_\delta)}\right) \phi(\boldsymbol{\lambda})^{-\frac{n}{2}}. \end{aligned} \quad (3.11)$$

One can easily compute the ratio:

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

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

### 3.4.4 Gradient and Hessian of the posterior penalty

The analytical gradient and Hessian of the penalty vector can be derived to find its posterior mode via a Newton-Raphson algorithm. The posterior mode as a measure of central tendency is essential to construct a grid for exploring $p(\boldsymbol{\lambda}|\mathcal{D})$. To ensure numerical stability, the penalty parameters are log transformed, $v_j = \log(\lambda_j)$, for $j = 1,\ldots,q$, and the associated vector is $\mathbf{v} = (v_1,\ldots,v_q)^\top$. Using the multivariate transformation method on (3.11), the posterior becomes:

$$\begin{aligned} p(\mathbf{v}|\mathcal{D}) \propto\ & |B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|^{-\frac{1}{2}}\left(\prod_{j=1}^{q}\exp(v_j)^{\left(\frac{\nu+K-3}{2}\right)}\right) \\ & \times\left(\prod_{j=1}^{q}\left(b_\delta+\frac{\nu}{2}\exp(v_j)\right)^{-\left(\frac{\nu}{2}+a_\delta\right)}\right)\phi(\mathbf{v})^{-\frac{n}{2}} \\ & \times\left(\prod_{j=1}^{q}\exp(v_j)\right), \end{aligned} \tag{3.12}$$

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

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

Taking the log of (3.12) yields:

$$\log p(\mathbf{v}|\mathcal{D}) \doteq -\frac{1}{2}\underbrace{\log|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}_{\text{Term I}} + \underbrace{\left(\frac{\nu+K-1}{2}\right)\sum_{j=1}^{q} v_j}_{\text{Term II}} - \frac{n}{2}\underbrace{\log\phi(\mathbf{v})}_{\text{Term III}}$$
$$-\Big(\frac{\nu}{2}+a_\delta\Big)\underbrace{\sum_{j=1}^{q}\log\Big(b_\delta+\frac{\nu}{2}\exp(v_j)\Big)}_{\text{Term IV}}. \tag{3.13}$$

### 3.4.5 Gradient

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

$$\begin{aligned}
\frac{\partial \log|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}{\partial v_j} &= \frac{1}{|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}\frac{\partial}{\partial v_j}|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}| \\
&= \frac{1}{|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}\mathrm{Tr}\Big(\mathrm{adj}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\Big) \\
&= \frac{1}{|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}\mathrm{Tr}\Big(|B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|\ (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} \\
&\qquad \times\frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})\Big) \\
&= \mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big),
\end{aligned} \tag{3.14}$$

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

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

where $\mathrm{diag}(0,\ldots,\exp(v_j),\ldots,0)$ is a $q\times q$ diagonal matrix, whose $j$th diagonal element is $\exp(v_j)$ and all other diagonal elements are zero.

Derivation of Term II with respect to $v_j$ is trivial:

$$\frac{\partial}{\partial v_j}\left(\frac{\nu+K-1}{2}\right)\sum_{j=1}^{q} v_j = \left(\frac{\nu+K-1}{2}\right). \tag{3.15}$$

The partial derivative of Term III is:

$$\begin{aligned}
\frac{\partial}{\partial v_j}\log(\phi(\mathbf{v})) &= \frac{1}{\phi(\mathbf{v})}\frac{\partial \phi(\mathbf{v})}{\partial v_j} \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\frac{\partial}{\partial v_j}\Big(\mathbf{y}^\top B(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}B^\top \mathbf{y}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\frac{\partial}{\partial v_j}\mathrm{Tr}\Big(\mathbf{y}^\top B(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}B^\top \mathbf{y}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\frac{\partial}{\partial v_j}\mathrm{Tr}\Big(B^\top \mathbf{y}\mathbf{y}^\top B(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathrm{Tr}\Big(B^\top \mathbf{y}\mathbf{y}^\top B\frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathrm{Tr}\Big(B^\top \mathbf{y}\mathbf{y}^\top B\Big(-(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j} \\
&\quad \times (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}\Big)\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathrm{Tr}\Big(\mathbf{y}^\top B\big(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\big)B^\top \mathbf{y}\Big)\Big) \\
&= \frac{1}{\phi(\mathbf{v})}\Big(-\frac{1}{2}\mathbf{y}^\top B\big(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\big)B^\top \mathbf{y}\Big) \\
&= \frac{1}{2\phi(\mathbf{v})}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}B^\top \mathbf{y}.
\end{aligned} \tag{3.16}$$

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

$$\begin{aligned}
\frac{\partial}{\partial v_j}\sum_{j=1}^{q}\log\Big(b_\delta + \frac{\nu}{2}\exp(v_j)\Big) &= \frac{\frac{\nu}{2}\exp(v_j)}{b_\delta + \frac{\nu}{2}\exp(v_j)} \\
&= \frac{1}{1+\frac{2b_\delta}{\nu\exp(v_j)}}.
\end{aligned} \tag{3.17}$$

From (3.14), (3.15), (3.16) and (3.17), the gradient $\nabla_{\mathbf{v}}\log p(\mathbf{v}|\mathcal{D})$ has entries:

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

### 3.4.6 Hessian

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

$$\begin{aligned}\frac{\partial}{\partial v_j}\mathrm{Tr}\Big((B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j}\Big) &= \mathrm{Tr}\Big(\frac{\partial}{\partial v_j}(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1}P_{v_j}\Big) \\ &= \mathrm{Tr}\Big(-\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} + \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big) \\ &= -\mathrm{Tr}\Big(\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big)^2 - \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big). \quad (3.18)\end{aligned}$$

In addition, recall from (3.16) that:

$$\frac{\partial \phi(\mathbf{v})}{\partial v_j} = \frac{1}{2}\mathbf{y}^\top B\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y}. \quad (3.19)$$

Furthermore, note the following differentiation result:

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

$$
\begin{aligned}
&= \operatorname{Tr}\Big(B^\top \mathbf{y}\mathbf{y}^\top B\Big(-2\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big)^2 \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} + \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\Big)\Big) \\
&= \operatorname{Tr}\Big(\mathbf{y}^\top B\Big(-2\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big)^2 \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} + \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}}\Big) B^\top \mathbf{y}\Big) \\
&= -2\mathbf{y}^\top B\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big)^2 \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} + \mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y}. \quad (3.20)
\end{aligned}
$$

Using (3.19), (3.20) and the quotient rule for derivatives yields:

$$
\begin{aligned}
\frac{\partial}{\partial v_j} \frac{\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y}}{\phi(\mathbf{v})} &= \frac{1}{\phi^2(\mathbf{v})}\Big(-2\phi(\mathbf{v})\mathbf{y}^\top B\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big)^2 \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} \\
&\quad + \phi(\mathbf{v})\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} \\
&\quad - \frac{1}{2}\big(\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y}\big)^2\Big). \quad (3.21)
\end{aligned}
$$

Finally, note that:

$$
\frac{\partial}{\partial v_j} \frac{\big(\frac{\nu}{2} + a_\delta\big)}{\Big(1 + \frac{2b_\delta}{\nu \exp(v_j)}\Big)} = \frac{b_\delta\big(1 + \frac{2a_\delta}{\nu}\big)\exp(-v_j)}{\Big(1 + \frac{2b_\delta}{\nu \exp(v_j)}\Big)^2}, \quad (3.22)
$$

and using (3.18), (3.21), (3.22), the diagonal entries of the Hessian of $\log p(\mathbf{v}|\mathcal{D})$ are:

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

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

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

Furthermore, similarly to (3.20):

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

such that using the quotient rule, we have:

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

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

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

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

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

$$\begin{aligned}&\frac{\partial \log p(\mathbf{v}|\mathcal{D})}{\partial v_j} \\ &= -\frac{1}{2}\mathrm{Tr}\Big(\mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j}\Big) + \left(\frac{\nu + K - 1}{2}\right) - \frac{n}{4\phi(\mathbf{v})}\mathbf{y}^\top B \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} P_{v_j} \mathcal{M}_{\boldsymbol{\xi}}^{\mathbf{v}} B^\top \mathbf{y} \\ &- \frac{\left(\frac{\nu}{2} + a_\delta\right)}{1 + \frac{2b_\delta}{\nu \exp(v_j)}}. \end{aligned} \tag{3.23}$$

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

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

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

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

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

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

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

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

In Table 3.1, we show the largest difference (in absolute value) between the entries of the numerical and analytical gradients and Hessians respectively computed across 1000 randomly selected points $\mathbf{v}$ with entries $v_j \sim \mathcal{U}(-5, 5),\ j = 1, 2, 3$.

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

Table 3.1: Largest absolute difference between gradient and Hessian entries computed from our analytical formulas and the numerical derivatives from the **numDeriv** package.

## 3.5 Exploration of the posterior penalty space

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

### 3.5.1 Grid strategy with skew-normal match

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

**Algorithm 2: Newton-Raphson to locate the mode of $p(\mathbf{v}|\mathcal{D})$**

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

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

$$p(x) = \frac{2}{\varsigma}\varphi\left(\frac{x-\mu}{\varsigma}\right) \Phi\left(\rho\frac{(x-\mu)}{\varsigma}\right), \tag{3.24}$$

where $\mu \in \mathbb{R}$ is a location parameter, $\varsigma \in \mathbb{R}_{++}$ a scale parameter and $\rho \in \mathbb{R}$ a shape parameter regulating skewness. Also, $\varphi(\cdot)$ and $\Phi(\cdot)$ denote the standard Gaussian density function and its cumulative distribution function respectively, such that setting $\rho = 0$ yields the $\mathcal{N}(\mu, \varsigma^2)$ distribution.

We suggest to approximate the conditional posterior distribution of $(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$ $(j = 1, ..., q)$ with a skew-normal distribution by matching its first three empirical moments with the theoretical ones for the density in (3.24), where $\widehat{\mathbf{v}}_{-j}$ denotes the vector $\widehat{\mathbf{v}}$ without the $j$th entry. The derivations to obtain $\mu^*, \varsigma^*$ and $\rho^*$ in the approximating skew-normal distribution $\text{SN}_j(\mu^*, \varsigma^{*2}, \rho^*)$ to $p(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$ through moment matching are shown below.

### 3.5.2 Approximating skew-normal distribution

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

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

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

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

Fast evaluation of the above empirical moments is discussed in Appendix C2. The skew-normal fit to $p(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$ is found by matching the empirical and theoretical moments, i.e. the following system needs to be solved:

$$
m_{j1} = \mu + \varsigma \sqrt{\frac{2}{\pi}}\, \psi \tag{3.25}
$$

$$
m_{j2} = \varsigma^2 \left(1 - \frac{2}{\pi}\, \psi^2\right) \tag{3.26}
$$

$$
m_{j3} = \frac{1}{2}(4 - \pi)\, \varsigma^3 \left(\frac{2}{\pi}\right)^{\frac{3}{2}}\, \psi^3. \tag{3.27}
$$

From (3.26), we isolate $\varsigma$:

$$
\varsigma = \sqrt{\frac{m_{j2}}{(1 - \frac{2}{\pi}\, \psi^2)}} > 0. \tag{3.28}
$$

Plugging (3.28) in (3.27) yields:

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

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

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

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

$$\psi^* = \text{sign}(m_{j3})\frac{\sqrt{4\left(\kappa^2+\frac{2\kappa^4}{\pi}\right)}}{2+\frac{4\kappa^2}{\pi}}. \tag{3.29}$$

So, we have $\rho^* = \psi^*/\sqrt{1-(\psi^*)^2}$ and plugging (3.29) in (3.28), we recover:

$$\varsigma^* = \sqrt{\frac{m_{j2}}{\left(1 - \frac{2}{\pi}\,(\psi^*)^2\right)}}. \tag{3.30}$$

Finally, the location parameter is given by:

$$\mu^* = m_{j1} - \varsigma^* \sqrt{\frac{2}{\pi}}\,\psi^*. \tag{3.31}$$

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

Once a skew-normal distribution has been adjusted to the conditional $p(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$, 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 $\text{SN}_{j,0.025}$ and $\text{SN}_{j,0.975}$ respectively. This process is repeated across all dimensions $j = 1, \ldots, q$ and a Cartesian product of the univariate grids is taken, ending up with a total of $M^q$ (multivariate) grid points. Next, a filtering strategy is implemented to get rid of quadrature points associated to a small posterior mass.

Let us consider the normalized posterior $R(\mathbf{v}) = p(\mathbf{v}|\mathcal{D})/p(\widehat{\mathbf{v}}|\mathcal{D})$ 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 $R(\mathbf{v}) \geq \exp\left(-0.5\chi^2_{\dim(\mathbf{v});1-\alpha}\right)$. 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 $R(\mathbf{v}) < \exp(-0.5\chi^2_{2;0.95}) = 0.05$, leaving $\widetilde{M}$ grid points after filtering. Figure 3.1 highlights the difference between the skew-normal match and the naive Gaussian fit to the targets $p(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$, $j = 1, 2$ with $q = 2$ nonlinear smooth functions in the additive predictor and sample size $n = 300$. In Figure 3.2, the surface plot of $R(\mathbf{v})$ is shown. Finally, Figure 3.3 summarizes the strategy behind the grid construction. In (a), an equidistant univariate grid is constructed in each dimension resulting in a cross-shaped pattern with center $\widehat{\mathbf{v}}$.

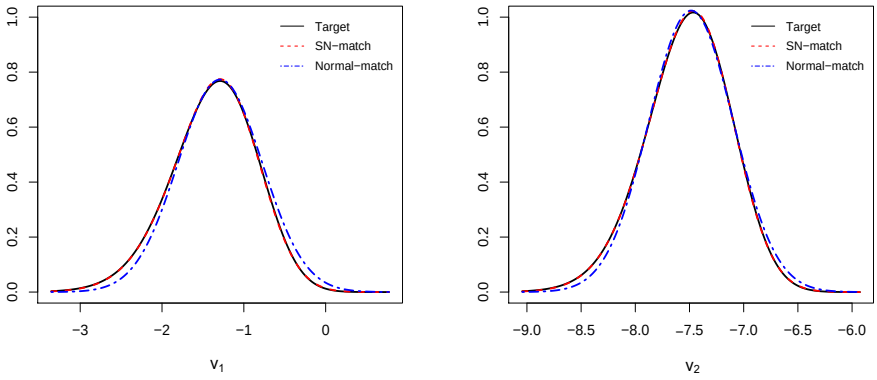



Figure 3.1: Skew-normal fit (dashed) and naive Gaussian match (dash-dotted) to the normalized conditional $p(v_1|\widehat{v}_2, \mathcal{D})$ (left) and $p(v_2|\widehat{v}_1, \mathcal{D})$ (right). The skew-normal fit is closer to the target and captures the lack of symmetry.

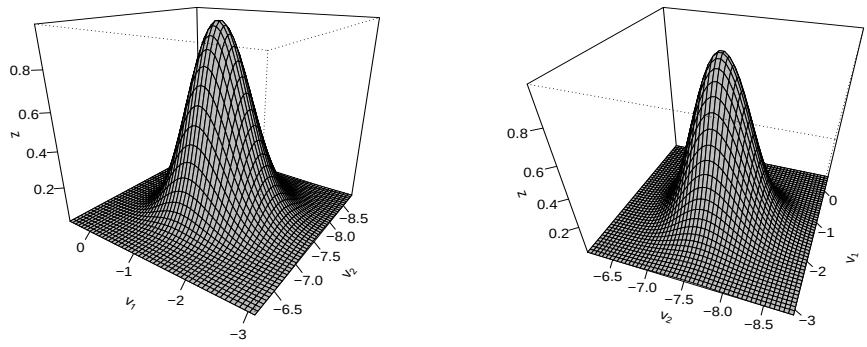



Figure 3.2: Surface plot of $R(\mathbf{v})$ when $q = 2$.

The Cartesian product of these univariate grids is computed and shown in (b). Following our filtering rule, we only keep a subset of the Cartesian product grid as shown by the blue points in (c). Figure 3.3 (d) shows the final grid that will be used for further inference in the additive model.

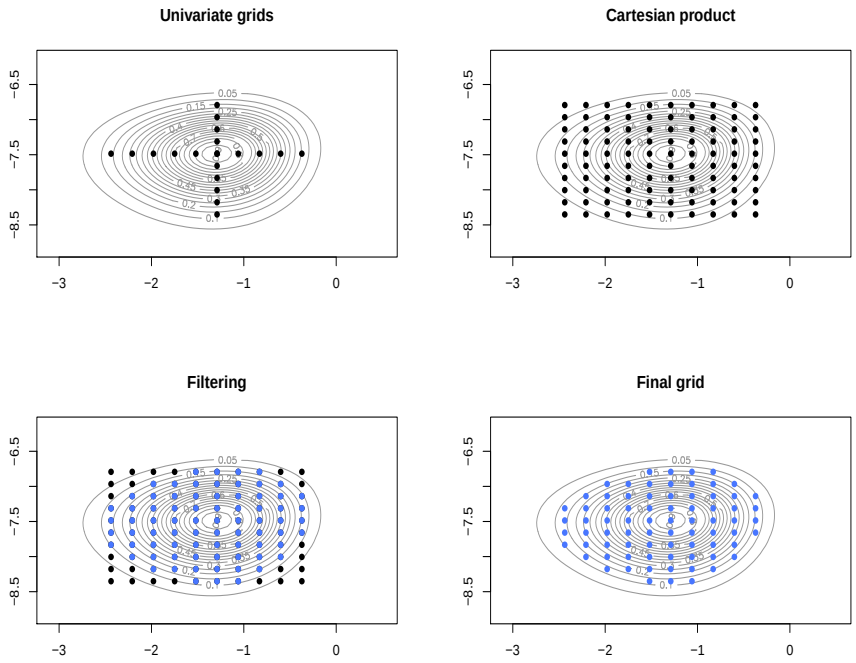



Figure 3.3: Grid strategy to explore $\log p(\mathbf{v}|\mathcal{D})$. (a) Equidistant univariate grid in each dimension. (b) Cartesian product. (c) Filtering out the points. (d) Final grid used for further inference in the additive model.

## 3.6 Approximate marginal posterior of vector $\boldsymbol{\xi}$

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

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

$$= \int_{\mathbb{R}^q_{++}} \left( \int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda}, \tau, \mathcal{D})\ p(\tau|\boldsymbol{\lambda}, \mathcal{D})\ d\tau \right)\ p(\boldsymbol{\lambda}|\mathcal{D})\ d\boldsymbol{\lambda} \quad (3.32)$$

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

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

The integrand is thus given by:

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

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

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

Using the above result, the integral is:

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

Note that:

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

so that the integral is finally given by:

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

The above formula is a multivariate Student distribution for $\boldsymbol{\xi}$ (see Jackman, 2009, p.508) with $n$ degrees of freedom denoted by $t_n\left(\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}, \widetilde{S}_{\boldsymbol{\lambda}}\right)$ with

location parameter $\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} = (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1} B^\top \mathbf{y}$ and symmetric, positive-definite matrix $\widetilde{S}_{\boldsymbol{\lambda}} = \frac{s_{\boldsymbol{\lambda}}}{n}\left(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\right)^{-1}$. Using the above integral result, the marginal posterior of $\boldsymbol{\xi}$ in (3.32) simplifies to:

$$p(\boldsymbol{\xi}|\mathcal{D}) = \int_{\mathbb{R}^q_{++}} t_n\left(\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}, \widetilde{S}_{\boldsymbol{\lambda}}\right) p(\boldsymbol{\lambda}|\mathcal{D})\ d\boldsymbol{\lambda}. \tag{3.33}$$

Using the log-transformation on the penalty parameters, (3.33) becomes:

$$p(\boldsymbol{\xi}|\mathcal{D}) = \int_{\mathbb{R}^q} t_n\left(\widehat{\boldsymbol{\xi}}_{\mathbf{v}}, \widetilde{S}_{\mathbf{v}}\right)\ p(\mathbf{v}|\mathcal{D})\ d\mathbf{v}, \tag{3.34}$$

where $\widehat{\boldsymbol{\xi}}_{\mathbf{v}} = (B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} B^\top \mathbf{y}$ and $\widetilde{S}_{\mathbf{v}} = (s_{\mathbf{v}}/n)\left(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\right)^{-1}$ with the scalar $s_{\mathbf{v}} = \mathbf{y}^\top \left(I_n - B(B^\top B + Q_{\boldsymbol{\xi}}^{\mathbf{v}})^{-1} B^\top\right)\mathbf{y}$. Let $\Delta_{v_j}$ be the width of the $j$th univariate grid and denote by $\Delta\mathbf{v} = \Delta_{v_1} \times \cdots \times \Delta_{v_q}$ the discretized version of $d\mathbf{v}$. Using the quadrature points from the grid strategy $\{\mathbf{v}^{(m)}\}_{m=1}^{\widetilde{M}}$, integral (3.34) can be approximated as follows:

$$\widetilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} t_n\left(\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \widetilde{S}_{\mathbf{v}^{(m)}}\right)\ p(\mathbf{v}^{(m)}|\mathcal{D})\ \Delta\mathbf{v}. \tag{3.35}$$

Furthermore, define the weights:

$$\omega_m = \frac{p(\mathbf{v}^{(m)}|\mathcal{D})\ \Delta\mathbf{v}}{\sum_{m=1}^{\widetilde{M}} p(\mathbf{v}^{(m)}|\mathcal{D})\ \Delta\mathbf{v}}, \quad m = 1, \ldots, \widetilde{M}. \tag{3.36}$$

Dividing (3.35) by the denominator of $\omega_m$, one obtains a mixture of multivariate Student distributions for the approximate posterior of the latent vector:

$$\widetilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m\ t_n\left(\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \widetilde{S}_{\mathbf{v}^{(m)}}\right). \tag{3.37}$$

Note that $\omega_m \geq 0$ and $\sum_{m=1}^{\widetilde{M}} \omega_m = 1$, such that (3.37) is a probability density function. Furthermore, $t_n\left(\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \widetilde{S}_{\mathbf{v}^{(m)}}\right)$ converges in law to $\mathcal{N}_{\dim(\boldsymbol{\xi})}\left(\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \widetilde{S}_{\mathbf{v}^{(m)}}\right)$ as $n \to +\infty$ (see Kroese et al., 2013, p. 147), so for $n$ sufficiently large, we can write (3.37) as a finite mixture of multivariate Gaussian densities:

$$\widetilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \, \mathcal{N}_{\dim(\boldsymbol{\xi})} \left( \widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \widetilde{S}_{\mathbf{v}^{(m)}} \right). \tag{3.38}$$

A point estimate for the latent vector is given by the posterior mean of (3.38) which is simply the mixture of the location components (see Frühwirth-Schnatter, 2006):

$$\widehat{\boldsymbol{\xi}} = \sum_{m=1}^{\widetilde{M}} \omega_m \, \widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}. \tag{3.39}$$

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

## 3.7 Credible intervals

### 3.7.1 Quantile-based credible intervals for latent variables

Approximate quantile-based credible intervals for latent variables $\xi_h$, $h = 1, \ldots, \dim(\boldsymbol{\xi})$ can be straightforwardly constructed. Starting from the joint marginal posterior in (3.38), we can write the univariate marginal posterior for element $\xi_h$ as:

$$\widetilde{p}(\xi_h|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \, \mathcal{N}_1 \left( \widehat{\xi}_{h,\mathbf{v}^{(m)}}, \widetilde{S}_{hh,\mathbf{v}^{(m)}} \right), \tag{3.40}$$

where $\widehat{\xi}_{h,\mathbf{v}^{(m)}}$ is the $h$th entry of vector $\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}$ and $\widetilde{S}_{hh,\mathbf{v}^{(m)}}$ is the $h$th entry on the diagonal of matrix $\widetilde{S}_{\mathbf{v}^{(m)}}$. Posterior (3.40) can then be used to numerically construct an approximate $(1-\alpha) \times 100\%$ quantile-based credible interval for $\xi_h$ as follows. Construct an equidistant fine grid, say $\{\xi_{hl}\}_{l=1}^{L}$ of width $\Delta_l$, and evaluate the posterior at each element of that grid, i.e. compute $\widetilde{p}(\xi_{hl}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \, \mathcal{N}_1 \left( \xi_{hl}; \widehat{\xi}_{h,\mathbf{v}^{(m)}}, \widetilde{S}_{hh,\mathbf{v}^{(m)}} \right)$, for $l = 1, \ldots, L$. Then, find the indices $q_{low} \in \{1, \ldots, L\}$ and $q_{up} \in \{1, \ldots, L\}$, such that $\sum_{l=1}^{q_{low}} \widetilde{p}(\xi_{hl}|\mathcal{D}) \; \Delta_l \approx \alpha/2$ and $\sum_{l=1}^{q_{up}} \widetilde{p}(\xi_{hl}|\mathcal{D}) \; \Delta_l \approx 1 - (\alpha/2)$. The resulting interval $[\xi_{hq_{low}}, \xi_{hq_{up}}]$ is an approximate $(1-\alpha) \times 100\%$ quantile-based credible interval for $\xi_h$.

### 3.7.2 Pointwise credible intervals for smooth functions

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 $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK-1})^\top$. Also, denote by $\widetilde{\mathbf{b}}_l^\top = (\tilde{b}_{j1}(x_l), \ldots, \tilde{b}_{jK-1}(x_l))$ the vector of B-splines in the basis evaluated at $x_l$. The function $f_j$ at point $x_l$ is thus modeled as $f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j}) = \widetilde{\mathbf{b}}_l^\top \boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ and from (3.38) the posterior of $\boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ is approximated by the finite mixture:

$$\widetilde{p}(\boldsymbol{\xi}_{\boldsymbol{\theta}_j}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \, \mathcal{N}_{K-1}\left(\widehat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}, \widetilde{S}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}\right), \tag{3.41}$$

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

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

A quantile-based credible interval for $f_j$ at point $x_l$ can easily be computed from the above (approximate) univariate posterior as in Section 3.7.1.

## 3.8 Simulation study

The performance of LPS in additive models (with cubic B-splines and a third order penalty) is assessed through different simulation scenarios and compared with results obtained using the `gam()` function of the **mgcv** package in **R** (Wood, 2017), a popular and established toolkit for estimating (generalized) additive models. Options of the `gam()` function are carefully chosen so that the generated results can be meaningfully compared to these obtained using our Laplace-P-spline approach. In particular, smooth terms are specified with the `gam()` function using *s(x, bs="ps", k=K, m=c(2,3))* , where $x$ is the vector of covariate values associated to the estimated smooth function and *ps* specifies a P-spline basis. The scalar $k$ is the basis dimension, the first entry in $m = c(\cdot, \cdot)$

refers to the order of the spline basis (with order 2 corresponding to cubic P-splines), while the second entry refers to the order of the difference penalty. Another chosen option in `gam()` is *method* = "$REML$", requiring an estimation of the penalty parameters $\boldsymbol{\lambda}$ by restricted maximum likelihood. It corresponds to an empirical Bayes approach in the sense that a Bayesian log marginal likelihood is maximized with respect to $\boldsymbol{\lambda}$ in a context where penalties come from Gaussian priors on the spline coefficients (Marra and Wood, 2011; Wood et al., 2013). The optimization method in `gam()` is chosen to be *optimizer=c("outer","newton")* as it provides reliable and stable computations.

### 3.8.1 Simulation results for parameters in the linear part

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

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

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

Figure 3.4 illustrates the shape of the functions $f_1$, $f_2$ and $f_3$ with a set of simulated data for $n = 300$ with medium signal to noise ratio ($\sigma = 0.40$).

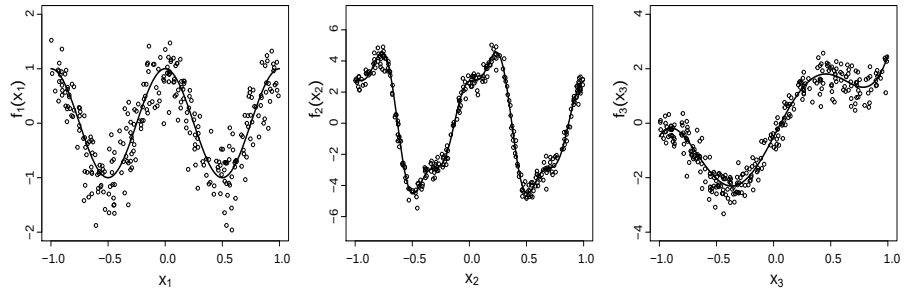



Figure 3.4: Illustration of functions $f_1$, $f_2$, $f_3$ (solid lines) and simulated data ($n = 300$) under medium signal to noise ratio ($\sigma = 0.40$).

The simulation results given in Table 3.2 show that our LPS estimation procedure exhibits good performance for the three different noise levels. Nonsignificant biases are observed and the estimated coverage probabilities are close to their nominal value in each setting. Furthermore, LPS and `gam()` have similar results regarding the ESE and RMSE.

In Figure 3.5, we show the LPS estimation of the smooth additive terms (gray curves) and the pointwise median (dashed) curves across all replications when 50 B-splines are used for each function. The estimated curves are close to their target on the entire domain except on the boundaries where the estimates exhibit larger variability.

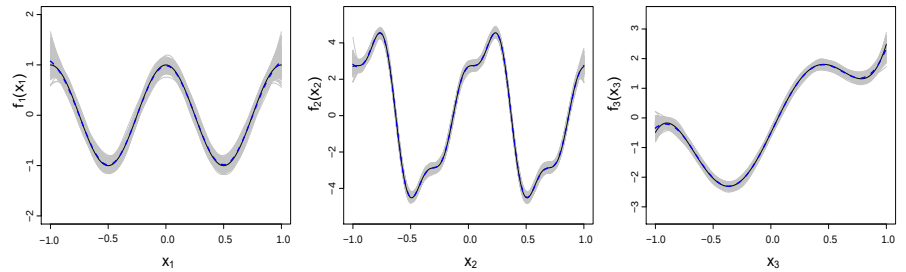



Figure 3.5: Estimation of the smooth functions $f_1$, $f_2$ and $f_3$ for $S = 500$ replications (one gray curve per dataset), sample size $n = 300$ and $\sigma = 0.40$ using 50 B-splines for each function. The solid (black) curve is the true function and the dashed curve is the pointwise median of the 500 estimated curves.

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

Table 3.2: Simulation results with the LPS method for $S = 500$ replicates of sample size $n = 300$ and $\sigma \in \{0.20, 0.40, 0.60\}$. The values in parentheses are estimation results from the `gam()` (MGCV) method.

### 3.8.2 Coverage of the smooth functions $f_j$

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

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

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

Tables 3.3 and 3.4 show the coverage estimates of 90% and 95% pointwise credible intervals for the functions $f_1, f_2$ and $f_3$ at selected points of their domain and for three different noise levels with 50 B-splines for each function. The frequentist coverage of credible intervals are compatible with their nominal value for all the considered noise levels for the LPS and `gam()` methods.

## 3.9 Application to Milan mortality data

In this section, the LPS methodology is illustrated on the Milan mortality data (Ruppert et al., 2003) available in the **SemiPar** package on CRAN (`https://CRAN.R-project.org/package=SemiPar`). The data contains observations on $n = 3652$ consecutive days between January 1st, 1980 and December 30th, 1989 for the city of Milan in Italy for air pollution indicators and health variables.

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

Table 3.3: Coverage estimates of 90% pointwise credible intervals of the functions $f_1, f_2, f_3$ at selected domain points for three noise levels $\sigma \in \{0.20, 0.40, 0.60\}$ over $S = 500$ replications of sample size $n = 300$ for the Laplace-P-spline approach (LPS) and `gam()` (MGCV) method. An asterisk indicates that the estimated coverage is incompatible with the nominal value at the 99% level.

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

Table 3.4: Coverage estimates of 95% pointwise credible intervals of the functions $f_1, f_2, f_3$ at selected domain points for three noise levels $\sigma \in \{0.20, 0.40, 0.60\}$ over $S = 500$ replications of sample size $n = 300$ for the Laplace-P-spline approach (LPS) and `gam()` (MGCV) method. An asterisk indicates that the estimated coverage is incompatible with the nominal value at the 99% level.

The objective is to study how air pollution and other meteorological indicators impact mortality using an additive partial linear model. In that endeavor, the square root of the total number of death (*Mortality*) is taken to be the response variable. Following Ruppert et al. (2003), the variable *TSP* measuring the total suspended particles in ambient air enters as a linear predictor. The dichotomous variable *Holiday* is an indicator of public holiday (1=public holiday; 0=otherwise) and is also naturally added in the linear part of the model. The remaining predictors are modeled as smooth functions, namely: the mean daily temperature in °C (*Temperature*), the relative humidity (*Humidity*), a measure of sulfur dioxide ($SO_2$) in ambient air and the number of days (*Numdays*) elapsed as from December 31st, 1979. Figure 3.6 provides a graphical illustration for some data variables.

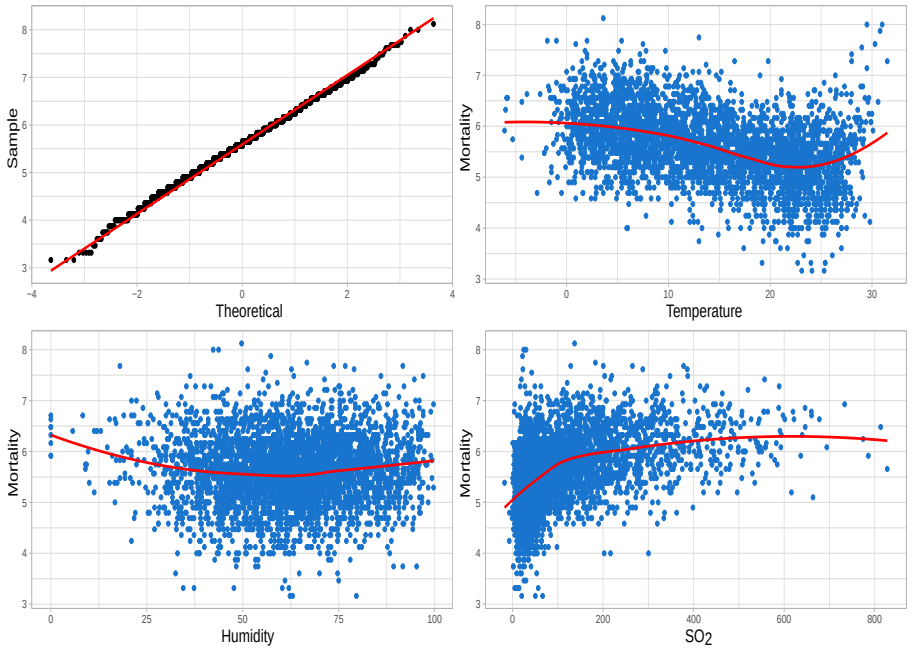



Figure 3.6: The Milan mortality data. Top-left: Q-Q plot of the response variable *Mortality*. Top-right: Scatter plot of *Mortality* and *Temperature*. Bottom-left: Scatter plot of *Mortality* across *Humidity*. Bottom-right: Scatter plot of the response and $SO_2$.

The quantile-quantile plot of the response variable on the top-left graph confirms that *Mortality* is approximately normally distributed. The scatter plots of the response with *Temperature*, *Humidity* and $SO_2$ and the associated locally estimated scatterplot smoothing (LOESS) fit in red suggest that the latter variables are nonlinearly related to *Mortality*. The additive model for the mortality data is written as:

$$\begin{aligned} Mortality_i = \beta_0 + \beta_1 TSP_i + \beta_2 Holiday_i + f_1(Temperature_i) \\ + f_2(Humidity_i) + f_3(SO_{2i}) + f_4(Numdays_i) + \varepsilon_i, \end{aligned}$$

for $i = 1, \ldots, n$, with i.i.d. errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. The smooth functions $f_j,\ j = 1, 2, 3$ are modeled with 35 cubic B-splines and a second-order penalty. The B-spline basis for a smooth term $f_j$ is defined over the domain $[x_{j,\min}, x_{j,\max}]$, i.e. over the range of its observed values $x_j$. Estimation results for *TSP* and *Holiday* are summarized in Table 3.5. *TSP* has a small positive and significant effect on the response, while *Holiday* has a negative and significant effect.

| **Parameters** | **Estimates** | **CI 95%** | $\mathbf{sd}_{post}$ |
|---|---|---|---|
| $\beta_1$ (*TSP*) | 0.0006 | [ 0.0001; 0.0010] | 0.0002 |
| $\beta_2$ (*Holiday*) | -0.1240 | [-0.2342; -0.0164] | 0.0558 |

Table 3.5: Estimation results for the parametric linear part of the additive model. The second column is the parameter estimate, the third column gives the associated 95% credible interval and the last column is the posterior standard deviation.

Figure 3.7 shows the estimated additive terms with approximate 95% pointwise credible intervals. We see that the conditional impact of *Temperature* on the mean response is slightly decreasing until approximately 25°C after which an explosive increase indicates that higher temperatures are associated to an important increase in the expected number of deaths. *Humidity* seems to have no significant impact on the response as it remains stable around zero. An increase in $SO_2$ levels from 0 to 180 is associated to an increase in average mortality. However, further increase of the $SO_2$ concentrations in ambient air seems to have negligible impact on the mean response as the smooth estimated term remains flat with a plausible zero value for the slope. For *Numdays*, we observe the

seasonal pattern already reported in Ruppert et al. (2003), i.e. average mortality fluctuates over seasons with spikes arising during winter.

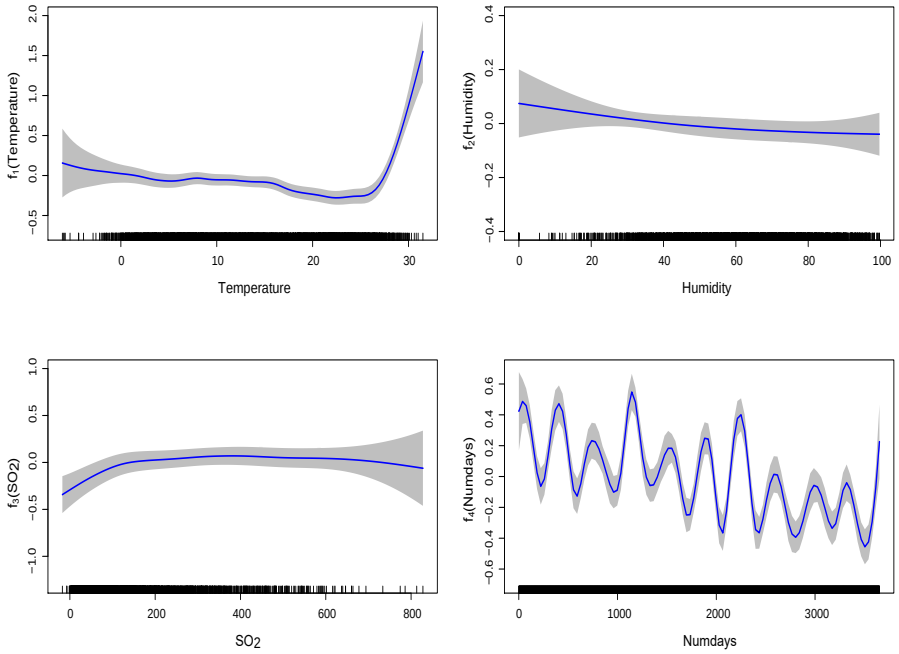



Figure 3.7: Estimates of the nonlinear predictors with 95% pointwise credible interval.

## 3.10 Conclusion

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

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

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

The main limitation behind a grid exploration of the posterior penalty space is an exponentially growing computational budget with the number $q$ of smooth functions in the additive model. To alleviate the problem, we propose a hybrid approach that alternates between a grid for small or moderate $q$ and a classic MCMC algorithm when $q$ is above a certain threshold. This is thoroughly discussed in Chapter 4 in the framework of generalized additive models. It is also worth noting that our LPS algorithm requires a low computational budget even though the modeling approach is fully Bayesian. An in-depth study of computational aspects is presented in the next chapter.

# CHAPTER 4

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

This chapter is based on: Gressani, O. and Lambert, P. (2021). Laplace approximations for fast Bayesian inference in generalized additive models based on P-splines, *Computational Statistics and Data Analysis*, Volume 154. https://doi.org/10.1016/j.csda.2020.107088

## 4.1 Motivation

Generalized additive models (GAMs) (Hastie and Tibshirani, 1986, 1987) extend generalized linear models (Nelder and 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 and 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 and Marx, 1996), thin plate splines (Wood, 2003), O'Sullivan penalized splines (Wand and Ormerod, 2008) or adaptive splines (Krivobokova et al., 2008) to cite the most popular instances. The material presented here 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 and Brezger, 2004) and from the efficiency of working with sparse bases and penalties for sampling-free approximate Bayesian inference or Markov chain Monte Carlo (MCMC) methods. Marx and Eilers (1998) are the first to revisit GAMs with P-splines. They developed the P-GAM technique where all the smooth terms are estimated simultaneously and the optimal penalty choice is controlled by information criterion or cross validation.

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. 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.

Taken separately, P-splines and INLA have made an impressive impact in the statistical community and initiated a flourishing literature in di-

versified domains (see e.g. Eilers et al., 2015; Rue et al., 2017), yet few references attempted to unify the strength of both approaches. 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 GAMs. 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 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 latter idea of combining Laplace approximations with MCMC can be found in Yoon and Wilson (2011) and more recently in Gómez-Rubio and Rue (2018).

The remainder of the chapter is outlined as follows. In Section 4.2 the Bayesian Laplace-P-spline GAM is formulated and the Laplace approximation to the conditional posterior of latent variables is derived. To efficiently explore the approximate posterior of the penalty vector, we propose a strategy that alternates between a deterministic grid and an independence Metropolis-Hastings sampler depending on the number of smooth components. The chosen penalty values are then used to approximate the marginal posterior for latent variables along with their associated pointwise credible intervals. A detailed simulation study is presented in Section 4.3 together with comparisons against a popular benchmark method. Section 4.4 illustrates the LPS model on two real datasets and Section 4.5 closes the chapter with concluding remarks and sketches future research prospects. The **blapsr** package (cf. Chapter 5) contains a routine called `gamlps()` to fit GAMs with LPS.

## 4.2 The Laplace-P-spline generalized additive model

### 4.2.1 Flexible modeling with 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:

$$p(y_i; \gamma_i, \varkappa) = \exp\left(\frac{y_i\gamma_i - s(\gamma_i)}{\varkappa} + c(y_i, \varkappa)\right), \tag{4.1}$$

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 $\gamma_i$ is the natural or canonical parameter. Using well-known properties of the score function (McCullagh and Nelder, 1989), one can show that the mean and variance of the response are $\mathbb{E}(y_i) := \mu_i = s'(\gamma_i)$ and $\text{Var}(y_i) = \varkappa s''(\gamma_i)$ respectively. Appendix D1 gives a detailed account of the one-parameter exponential family distributions used in this chapter. Let $\mathcal{D} = \{(y_i, \mathbf{x}_i, \mathbf{z}_i) : i = 1, \ldots, n\}$ be a sample of $n$ independent observations, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{iq})^\top$ is a vector of continuous covariates and $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip})^\top$ a vector of additional covariates (possibly categorical). The link function $g(\cdot)$ relates the mean response to the additive predictor as follows:

$$g(\mu_i) := \varrho_i = \beta_0 + \sum_{l=1}^{p} \beta_l z_{il} + \sum_{j=1}^{q} f_j(x_{ij}), \quad i = 1, \ldots, n. \tag{4.2}$$

In the spirit of the P-spline approach proposed in Eilers and 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:

$$f_j(x_{ij}) = \sum_{k=1}^{K} \theta_{jk} b_{jk}(x_{ij}), \quad j = 1, \ldots, q, \tag{4.3}$$

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 $\boldsymbol{\theta}_j = (\theta_{j1}, \ldots, \theta_{jK})^\top$, while the collection of all spline coefficients present in the model is $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \ldots, \boldsymbol{\theta}_q^\top)^\top$ and the vector of B-spline functions at $x_{ij}$ is written as $\mathbf{b}_j(x_{ij}) = (b_{j1}(x_{ij}), \ldots, b_{jK}(x_{ij}))^\top$. Model flexibility is compensated by a roughness penalty on finite differences of the coefficients of contiguous B-splines, $\boldsymbol{\theta}^\top \mathcal{P}(\boldsymbol{\lambda})\boldsymbol{\theta}$, with block diagonal matrix $\mathcal{P}(\boldsymbol{\lambda})$ written compactly using a Kronecker product (cf. Section 3.2.1), where $\boldsymbol{\lambda}$ is vector of positive penalty parameters. From a Bayesian perspective, Lang and 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 $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^\top$, 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 $\boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ 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:

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

Covariates $\mathbf{z}_i$ are centered around their mean value $\bar{z}_l = n^{-1}\sum_{i=1}^n z_{il}$, $l = 1, \ldots, p$ and identifiability constraints are imposed as in Section 3.2.2 yielding centered B-spline matrices $\widetilde{B}_j$ with $K-1$ columns and a latent vector of dimension $\dim(\boldsymbol{\xi}) = q \times (K-1) + p + 1$. This is to be contrasted with the model setting in INLA, where the latent field dimension grows with sample size $n$.

Following Jullion and Lambert (2007), robust priors are specified on the roughness penalty parameters with a conjugate Gamma family having a hierarchical structure $\lambda_j|\delta_j \sim \mathcal{G}(\nu/2, (\nu\delta_j)/2)$, $j = 1, \ldots, q$. An uninformative distribution is imposed on the hyperparameter $\delta_j \sim \mathcal{G}(a_\delta, b_\delta)$, $j = 1, \ldots, q$ with $a_\delta = b_\delta = 10^{-4}$ and $\nu = 3$. The penalty parameters are gathered in the vector $\boldsymbol{\eta} = (\boldsymbol{\lambda}^\top, \boldsymbol{\delta}^\top)^\top$. Taking into account the identifiability constraint, the additive predictor in (4.2) can be expressed compactly as $\boldsymbol{\varrho} = B\boldsymbol{\xi}$, where $B$ is a side by side configuration of design matrices, $B = [Z : \widetilde{B}_1 : \ldots : \widetilde{B}_q]$ and corresponds to the full design matrix of the model. The Bayesian model is summarized as follows:

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

### 4.2.2 Likelihood, Score function and Fisher information

The log-likelihood of a response variable having a density as in (4.1) is $\ell(\boldsymbol{\xi}; \mathcal{D}) = (1/\varkappa) \sum_{i=1}^{n} (y_i\gamma_i - s(\gamma_i)) + c$, with $c := \sum_{i=1}^{n} c(y_i, \varkappa)$ for ease of notation. The $h$th element of the score function for $\boldsymbol{\xi}$ is:

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\xi}; \mathcal{D})}{\partial \xi_h} &= \frac{1}{\varkappa} \sum_{i=1}^{n} \left( y_i \frac{\partial \gamma_i}{\partial \xi_h} - \frac{\partial s(\gamma_i)}{\partial \gamma_i} \frac{\partial \gamma_i}{\partial \xi_h} \right) \\
&= \frac{1}{\varkappa} \sum_{i=1}^{n} \big(y_i - s'(\gamma_i)\big) \frac{\partial \gamma_i}{\partial \xi_h} \\
&= \frac{1}{\varkappa} \sum_{i=1}^{n} \big(y_i - \mu_i\big) \frac{\partial \gamma_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \varrho_i} \frac{\partial \varrho_i}{\partial \xi_h}.
\end{aligned}
$$

As $\mu_i = s'(\gamma_i)$ and $s'(\cdot)$ is a strictly monotonic function, it holds from the inverse function rule that $\partial\gamma_i/\partial\mu_i = 1/(\partial\mu_i/\partial\gamma_i) = 1/s''(\gamma_i)$. Also, since $\varrho_i = g(\mu_i)$, we have $\partial\mu_i/\partial\varrho_i = 1/(\partial\varrho_i/\partial\mu_i) = 1/g'(\mu_i)$. Finally, $\partial\varrho_i/\xi_h = B_{ih}$ is the entry in the $i$th row and $h$th column of the design matrix $B$. Using these derivative results, one obtains:

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\xi}; \mathcal{D})}{\partial \xi_h} &= \frac{1}{\varkappa} \sum_{i=1}^{n} (y_i - \mu_i) \frac{1}{s''(\gamma_i)} \frac{1}{g'(\mu_i)} B_{ih} \\
&= \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(y_i)} \frac{1}{g'(\mu_i)} B_{ih} \\
&= \sum_{i=1}^{n} w_i (y_i - \mu_i) g'(\mu_i) B_{ih},
\end{aligned}
$$

with weights $w_i := \Big[Var(y_i)[g'(\mu_i)]^2\Big]^{-1}$. Also, define the following diagonal matrices $W := \mathrm{diag}(w_1, \ldots, w_n)$ and $D_g = \mathrm{diag}(g'(\mu_1), \ldots, g'(\mu_n))$,

so that the score is written as:

$$\nabla_{\boldsymbol{\xi}}\ell(\boldsymbol{\xi};\mathcal{D}) = B^\top W D_g(\boldsymbol{y}-\boldsymbol{\mu}). \tag{4.4}$$

Another quantity of interest is the expected Fisher information matrix defined as the expected value of the negative Hessian of the log-likelihood. The latter is obtained by computing the following partial derivatives for $s, h = 1, \ldots, \dim(\boldsymbol{\xi})$:

$$\begin{aligned}
&\frac{\partial^2\ell(\boldsymbol{\xi};\mathcal{D})}{\partial\xi_s\partial\xi_h} \\
&= \frac{\partial}{\partial\xi_s}\left(\frac{\partial\ell(\boldsymbol{\xi};\mathcal{D})}{\partial\xi_h}\right) \\
&= \frac{\partial}{\partial\xi_s}\left(\sum_{i=1}^n (y_i-\mu_i)w_i g'(\mu_i)B_{ih}\right) \\
&= \sum_{i=1}^n \left(\frac{\partial}{\partial\xi_s}(y_i-\mu_i)\right) w_i g'(\mu_i)B_{ih} + \sum_{i=1}^n (y_i-\mu_i)\left(\frac{\partial}{\partial\xi_s}w_i g'(\mu_i)\right)B_{ih}.
\end{aligned}$$

As $\partial(y_i-\mu_i)/\partial\xi_s = -(\partial\mu_i/\partial\varrho_i)\,(\partial\varrho_i/\partial\xi_s) = (-1/g'(\mu_i))B_{is}$, we recover:

$$\frac{\partial^2\ell(\boldsymbol{\xi};\mathcal{D})}{\partial\xi_s\partial\xi_h} = -\sum_{i=1}^n w_i B_{is}B_{ih} + \sum_{i=1}^n (y_i-\mu_i)\left(\frac{\partial}{\partial\xi_s}w_i g'(\mu_i)\right)B_{ih}. \tag{4.5}$$

The expected value of the second term in the right-hand side of (4.5) equals zero [1], so the expected Fisher information matrix has entries:

$$\mathcal{I}_{sh} = E\left(-\frac{\partial^2\ell(\boldsymbol{\xi};\mathcal{D})}{\partial\xi_s\partial\xi_h}\right) = \sum_{i=1}^n w_i B_{is}B_{ih}. \tag{4.6}$$

In matrix notation:

$$\mathcal{I}(\boldsymbol{\xi}) = E\left(-\frac{\partial^2\ell(\boldsymbol{\xi};\mathcal{D})}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\top}\right) = E\left(-\nabla^2_{\boldsymbol{\xi}}\ell(\boldsymbol{\xi};\mathcal{D})\right) = B^\top W B. \tag{4.7}$$

[1] Note that $(\partial w_i g'(\mu_i)/\partial\xi_s) = (\partial/\partial\xi_s)(\varkappa s''(\gamma_i)g'(\mu_i))^{-1}$. Recall that with a canonical link, $g(\mu_i) = (s')^{-1}(\mu_i) = \gamma_i$, hence $g'(\mu_i) = 1/(s''((s')^{-1}(\mu_i))) = 1/s''(\gamma_i)$ such that $(\partial w_i g'(\mu_i)/\partial\xi_s) = 0$ and the expected and observed Fisher information are identical, i.e. $E\left(-\partial^2\ell(\boldsymbol{\xi};\mathcal{D})/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\top\right) = -\partial^2\ell(\boldsymbol{\xi};\mathcal{D})/\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}^\top$.

### 4.2.3 Approximated conditional posterior of $\boldsymbol{\xi}$

Using Bayes' theorem, the conditional posterior of the latent vector is proportional to the product of the likelihood and prior, which can be written as $p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D}) \propto \exp\Big(\ell(\boldsymbol{\xi};\mathcal{D}) - 0.5\boldsymbol{\xi}^\top Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\boldsymbol{\xi}\Big)$. 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 $\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$. After convergence of the iterative algorithm, we recover a Gaussian centered around $\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} = (B^\top\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\widetilde{\boldsymbol{\varpi}}$ with variance-covariance matrix equal to the inverse of the sum of the negative Hessian of the log-likelihood and the precision matrix $Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}$, i.e. $\widehat{\Sigma}_{\boldsymbol{\lambda}} = (B^\top\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}$, 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 $\boldsymbol{\varpi}^{(0)} := (1/\varkappa)B^\top\big(\boldsymbol{y} - \boldsymbol{\mu}\big(\boldsymbol{\xi}^{(0)}\big)\big) + B^\top W\big(\boldsymbol{\xi}^{(0)}\big)B\boldsymbol{\xi}^{(0)}$ computed from an initial guess $\boldsymbol{\xi}^{(0)}$ of the latent vector. The Laplace approximation $\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})$ is used to approximate the integrand entering the computation of the marginal posterior for $\boldsymbol{\xi}$:

$$p(\boldsymbol{\xi}|\mathcal{D}) = \int_{\mathbb{R}^q_{++}} p(\boldsymbol{\xi}|\boldsymbol{\lambda},\mathcal{D})\; p(\boldsymbol{\lambda}|\mathcal{D})\; d\boldsymbol{\lambda}. \tag{4.8}$$

Quadrature points to compute (4.8) will be obtained in the next section using an approximation of the marginal posterior $p(\boldsymbol{\lambda}|\mathcal{D})$ for the vector of penalty parameters.

### 4.2.4 Marginal posterior of the penalty parameters

An indispensable intermediate step to reach an approximated version for the marginal posterior of the regression and spline variables $\boldsymbol{\xi}$ is to obtain the marginal posterior of the vector $\boldsymbol{\lambda}$ of penalty parameters. In that endeavor, we first derive an approximation of $p(\boldsymbol{\eta}|\mathcal{D})$ in the philosophy of Leonard (1982), Tierney and 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 prove to be very useful to explore the support of the posterior distribution of the penalty vector.

### 4.2.5 Approximation to the posterior 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 4.2.3) 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 $\gamma_i = \varrho_i = \mathbf{b}_i^\top\boldsymbol{\xi}$, with $\mathbf{b}_i^\top$ the row vector corresponding to the $i$th 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:

$$\begin{aligned} \widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) \propto\ & \exp\left(\frac{1}{\varkappa}\sum_{i=1}^{n}\left[y_i\mathbf{b}_i^\top\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} - s\left(\mathbf{b}_i^\top\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right)\right] - \frac{1}{2}\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}^\top Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}\hat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right) \\ &\times\left(\prod_{j=1}^{q}\delta_j^{(\frac{\nu}{2}+a_\delta-1)}\exp\left(-\delta_j\left(b_\delta+\frac{\nu}{2}\lambda_j\right)\right)\right)\left(\prod_{j=1}^{q}\lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right) \\ &\times|B^\top\widetilde{W}B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}}. \end{aligned} \tag{4.9}$$

As Gamma priors have been chosen for the penalty parameters $\lambda_j$ and $\delta_j$, one recognizes in (4.9) the conditional conjugacy for $\delta_j$, as $\delta_j|\lambda_j,\mathcal{D} \sim \mathcal{G}\left(\frac{\nu}{2}+a_\delta, b_\delta+\frac{\nu}{2}\lambda_j\right)$. Under these prior specifications, the integration of (4.9) with respect to $\boldsymbol{\delta}$ is tractable and yields the (approximate) marginal penalty posterior:

$$\widetilde{p}(\boldsymbol{\lambda}|\mathcal{D}) = \int_0^{+\infty} \cdots \int_0^{+\infty} \widetilde{p}(\boldsymbol{\eta}|\mathcal{D})\ d\delta_1 \ldots d\delta_q$$
$$\propto |B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}}|^{-\frac{1}{2}} \exp\left(\frac{1}{\varkappa}\sum_{i=1}^{n}\left[y_i \mathbf{b}_i^\top \widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}} - s\left(\mathbf{b}_i^\top \widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right)\right] - \frac{1}{2}\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}^\top Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}} \widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}\right)$$
$$\times \left(\prod_{j=1}^{q} \lambda_j^{\left(\frac{\nu+K-3}{2}\right)}\right)\left(\prod_{j=1}^{q}\left(b_\delta + \frac{\nu}{2}\lambda_j\right)^{-\left(\frac{\nu}{2}+a_\delta\right)}\right). \tag{4.10}$$

Using a log transform on the penalty parameters $v_j = \log(\lambda_j)$, $j = 1, \ldots, q$ and using the multivariate transformation method on (4.10), we obtain the following expression for the (log) posterior of the log penalty vector:

$$\begin{aligned}\log \widetilde{p}(\mathbf{v}|\mathcal{D}) \ \doteq\ & -\frac{1}{2}\log|B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}| + \frac{\nu+K-1}{2}\sum_{j=1}^{q} v_j \\ & + \frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^\top \widehat{\boldsymbol{\xi}}_{\mathbf{v}} - \frac{1}{\varkappa}\sum_{i=1}^{n} s\left(\mathbf{b}_i^\top \widehat{\boldsymbol{\xi}}_{\mathbf{v}}\right) - \frac{1}{2}\widehat{\boldsymbol{\xi}}_{\mathbf{v}}^\top Q_{\boldsymbol{\xi}}^{\mathbf{v}} \widehat{\boldsymbol{\xi}}_{\mathbf{v}} \\ & - \left(\frac{\nu}{2} + a_\delta\right)\sum_{j=1}^{q} \log\left(b_\delta + \frac{\nu}{2}\exp(v_j)\right),\end{aligned} \tag{4.11}$$

where $Q_{\boldsymbol{\xi}}^{\mathbf{v}}$ 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 $\widehat{\boldsymbol{\xi}}_{\mathbf{v}} := \left(B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}\right)^{-1} \widetilde{\boldsymbol{\varpi}}$. The gradient $\nabla_{\mathbf{v}} \log \widetilde{p}(\mathbf{v}|\mathcal{D})$ and Hessian $\nabla^2_{\mathbf{v}} \log \widetilde{p}(\mathbf{v}|\mathcal{D})$ of expression (4.11) are analytically derived in Appendix D2. These expressions will turn to be useful to explore the marginal posterior of the penalty parameters.

### 4.2.6 Strategy to explore the posterior penalty space

An approximation to the marginal posterior of the latent variables $\boldsymbol{\xi}$ (including the regression and spline parameters in the generalized additive model) can be obtained by integrating out the penalty parameters as in (4.8). 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 an alternative strategy relying on MCMC to draw a set of points in the domain of the posterior of the penalty parameters when $q$ is large. 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 $\widehat{\mathbf{v}}$ of $\log \widetilde{p}(\mathbf{v}|\mathcal{D})$ as it represents the "center of gravity" around 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 $\widehat{\mathbf{v}}$ is obtained, we proceed with posterior exploration. The skew-normal approximation method has already been presented in Section 3.5.1 and is therefore omitted here.

### 4.2.7 Independence sampling when $q$ is large

When the number of smooth functions $q$ in the GAM is above a certain threshold (say $q > 4$), a grid-based 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 $(\vartheta/(\vartheta-2))(-\mathcal{H}^*)^{-1}$, where $\mathcal{H}^* = \nabla^2_{\mathbf{v}} \log \widetilde{p}(\mathbf{v}|\mathcal{D})|_{\mathbf{v}=\widehat{\mathbf{v}}}$.

Algorithm 3 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 $\mathrm{SN}_{j,\alpha/2}$ of approximating skew-normal densities. The grid is also filtered by removing the points satisfying the inequality

$R(\mathbf{v}) = p(\mathbf{v}|\mathcal{D})/p(\widehat{\mathbf{v}}|\mathcal{D}) < \exp\big(-0.5\chi^2_{q;1-\alpha}\big)$ (cf. Section 3.5.2). Exploration in larger dimensions relies on the independence Metropolis-Hastings sampler. This algorithm will be used in the next section to approximate the marginal posterior of the latent vector.

---

**Algorithm 3: Exploration of $\widetilde{p}(\mathbf{v}|\mathcal{D})$**

---

**If** $q \leq 4$ **do** (Grid strategy)
  **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 Cartesian product of univariate grids $\mathcal{C} = \times_{j=1}^q \{v_{jm}\}_{m=1}^M$.
  Choose $\alpha$ and keep $\widetilde{M}$ points in $\mathcal{C}$ for which $R(\mathbf{v}) \geq \exp\big(-.5\chi^2_{q;1-\alpha}\big)$.
**else do** (Independence sampling)
  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

$$\alpha = \min\left(1, \frac{\widetilde{p}\big(\mathbf{v}^{(\text{prop})}|\mathcal{D}\big)h\big(\mathbf{v}^{(m-1)}|\widehat{\mathbf{v}}\big)}{\widetilde{p}\big(\mathbf{v}^{(m-1)}|\mathcal{D}\big)h\big(\mathbf{v}^{(\text{prop})}|\widehat{\mathbf{v}}\big)}\right).$$

    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**

---

### 4.2.8 Approximate posterior for the vector of regression and spline parameters

Using the Laplace approximation discussed in Section 4.2.3, the posterior of the vector $\boldsymbol{\xi}$ can be obtained as follows:

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

$$\begin{aligned} &\approx \int_{\mathbb{R}^q_{++}} \widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\lambda}, \mathcal{D})\ \widetilde{p}(\boldsymbol{\lambda}|\mathcal{D})\ d\boldsymbol{\lambda} \\ &\approx \int_{\mathbb{R}^q} \widetilde{p}_G(\boldsymbol{\xi}|\exp(\mathbf{v}), \mathcal{D})\ \widetilde{p}(\mathbf{v}|\mathcal{D})\ d\mathbf{v}, \end{aligned} \tag{4.12}$$

where the last line follows from the change of variable in log-scale. Using Algorithm 3, we get a set of quadrature points $\{\mathbf{v}^{(m)}\}_{m=1}^{\widetilde{M}}$. Defining:

$$\omega_m = \frac{\widetilde{p}(\mathbf{v}^{(m)}|\mathcal{D})}{\sum_{m=1}^{\widetilde{M}} \widetilde{p}(\mathbf{v}^{(m)}|\mathcal{D})}, \quad m = 1, \ldots, \widetilde{M}, \tag{4.13}$$

when $q \leq 4$ and $\omega_m = 1/\widetilde{M}$ otherwise, Equation (4.12) suggests to approximate $p(\boldsymbol{\xi}|\mathcal{D})$ by:

$$\widetilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m\ \mathcal{N}_{\dim(\boldsymbol{\xi})}\left(\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}, \widehat{\Sigma}_{\mathbf{v}^{(m)}}\right), \tag{4.14}$$

where $\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}} = \left(B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\mathbf{v}^{(m)}}\right)^{-1} \widetilde{\boldsymbol{\varpi}}$ and $\widehat{\Sigma}_{\mathbf{v}^{(m)}} = \left(B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\mathbf{v}^{(m)}}\right)^{-1}$ are the conditional posterior mode and variance-covariance matrix resulting from the iterative Laplace approximations proposed in Section 4.2.3. 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 that consists in 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 (4.14), which is a mixture of the location components, i.e. $\widehat{\boldsymbol{\xi}} = \sum_{m=1}^{\widetilde{M}} \omega_m\ \widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}$.

### 4.2.9 Credible intervals

Approximate pointwise credible intervals for $\xi_h,\ h = 1, \ldots, \dim(\boldsymbol{\xi})$ can be straightforwardly obtained by starting from the finite mixture given in (4.14). The approximate posterior for the $h$th element is $\widetilde{p}(\xi_h|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m\ \mathcal{N}_1\left(\widehat{\xi}_{h,\mathbf{v}^{(m)}}, \widehat{\Sigma}_{hh,\mathbf{v}^{(m)}}\right)$, where $\widehat{\xi}_{h,\mathbf{v}^{(m)}}$ is the $h$th entry of vector $\widehat{\boldsymbol{\xi}}_{\mathbf{v}^{(m)}}$ and $\widehat{\Sigma}_{hh,\mathbf{v}^{(m)}}$ is the $h$th 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 $\xi_h$. 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 $\boldsymbol{\theta_j} = (\theta_{j1}, \ldots, \theta_{jK-1})^\top$. 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 $f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j}) = \widetilde{\mathbf{b}}_l^\top \boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ and from (4.14) the posterior of $\boldsymbol{\xi}_{\boldsymbol{\theta}_j}$ is approximated by the finite mixture:

$$\widetilde{p}(\boldsymbol{\xi}_{\boldsymbol{\theta}_j}|\mathcal{D}) = \sum_{m=1}^{\widetilde{M}} \omega_m \, \mathcal{N}_{K-1}\left(\widehat{\boldsymbol{\xi}}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}, \widehat{\Sigma}_{\boldsymbol{\theta}_j,\mathbf{v}^{(m)}}\right), \tag{4.15}$$

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 $p(f_j(x_l|\boldsymbol{\xi}_{\boldsymbol{\theta}_j})|\mathcal{D})$ is to use a mixture of univariate normals:

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

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

## 4.3 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). The reader is referred to Section 3.8 for the set-up of the input arguments in the `gam()` routine.

### 4.3.1 Estimation of the parameters in the linear part

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 $z_{i3} \sim \mathcal{N}(0,1)$, for $i = 1, \ldots, n$ and coefficients $\beta_0 = -1.50$, $\beta_1 = 0.70$, $\beta_2 = -0.80$, $\beta_3 = 0.40$. The covariates for the smooth functions are independent draws from the uniform distribution on the domain $[-1, 1]$. The smooth additive terms coincide with the functions:

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

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 4.1 shows the simulation results and comparisons with the `gam()` function. For all the considered data types, the Laplace-P-spline approach exhibits nonsignificant 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. A notable feature of the Laplace-P-spline methodology is that it requires a 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 0.90 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.

| 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 4.1: 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.

Considering that the algorithm behind `gam()` 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.

### 4.3.2 Estimation of the additive terms $f_j$

The coverage properties of approximate 90% pointwise credible intervals for the additive terms $f_1, f_2$ and $f_3$ are reported in Table 4.2 for selected values of the covariate on $[-1, 1]$. An asterisk superscript is added to the estimated coverage to indicate incompatibility with the nominal value. Results of the `gam()` function are labeled "MGCV".

In addition to the LPS approach, Table 4.2 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 (as in Wood's approach), such that the mixture in Equation (4.15) is omitted and the point estimate of the latent vector and its associated variance-covariance matrix become $\widehat{\boldsymbol{\xi}}_{\widehat{\mathbf{v}}} = \left(B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\widehat{\mathbf{v}}}\right)^{-1} \widetilde{\boldsymbol{\varpi}}$ and $\widehat{\Sigma}_{\widehat{\mathbf{v}}} = \left(B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\widehat{\mathbf{v}}}\right)^{-1}$ 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, $\widehat{f}_j(x_l) \pm z_{\alpha/2} \sqrt{\widetilde{\mathbf{b}}_l^\top \widehat{\Sigma}_{\boldsymbol{\theta}_j, \widehat{\mathbf{v}}} \widetilde{\mathbf{b}}_l}$.

As can be seen from Table 4.2, 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 four times slower than `gam()` ($\approx 0.05$ seconds vs 0.26 seconds).

| 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 4.2: Effective frequentist coverages of 90% pointwise credible intervals for $f_1, f_2, f_3$ at selected domain points over $S = 500$ replications of sample size $n = 300$ for LPS, LPSMAP and MGCV methods. An asterisk indicates incompatibility with the nominal value.

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 4.3 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 4.1. However, as expected, increasing the sample size in the Bernoulli scenario yields frequentist coverage probabilities close to their nominal value (cf. Table 4.3 with $n = 2000$) both for the LPS(MAP) and `gam()` methods.

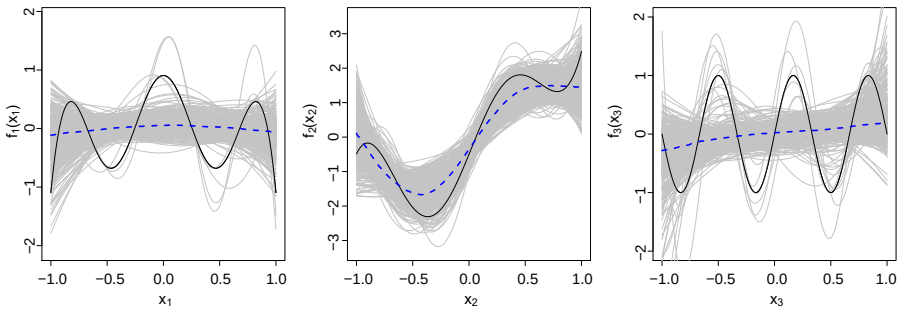



Figure 4.1: Estimation of smooth additive terms (gray curves) for $S = 500$ dataset replications of size $n = 300$ in the Bernoulli scenario with LPS. The dashed line is the pointwise median of the gray curves and the black curves are the target functions.

Table 4.4 reports 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. Again, the LPS and LPSMAP methodologies display estimated coverages close to their nominal value in all scenarios. The `gam()` results show similar performance when coverages are averaged over the covariate support. Note that `gam()` and LPSMAP rely on a similar approach for selecting the optimal posterior penalty value. Hence, the simulation results presented in this section 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 4.3: 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 indicates incompatibility with the nominal value.

| | | 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 |

Table 4.4: 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.

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. Further simulations (not reported here) have been implemented to assess the estimation performance of the additive smooth terms in small samples (with n=150). Regardless whether the LPS(MAP) or MGCV method is used, moderate to severe undercoverage is observed as the small sample size does not convey enough information for a precise reconstruction of the additive components. Even with a MCMC sampler to explore the joint posterior $p(\boldsymbol{\xi}, \mathbf{v}|\mathcal{D})$ (and without making Laplace approximations), the reported coverages remain unsatisfying in a small sample setting.

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 4.5, 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 $n = 300$. 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.

| $f$ | Method | -0.95 | -0.70 | -0.50 | -0.20 | 0.00 | 0.20 | 0.50 | 0.70 | 0.95 |
|---|---|---|---|---|---|---|---|---|---|---|
| $f_1$ | 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 4.5: 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.

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.

### 4.3.3 Computational costs

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 4.2 (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 4.2 (b) highlights the computational time of LPS(MAP) against sample size $n$ on a log scale.

### 4.3.4 Simulation study with more additive terms.

A large number $q$ of smooth functions in the additive predictor implies an increased computational burden.

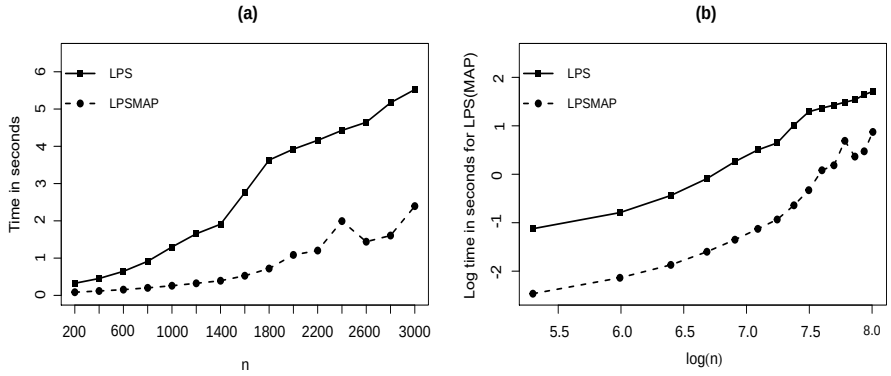



Figure 4.2: (a) Real elapsed time in seconds as a function of sample size for LPS and LPSMAP. (b) Log of computational time (in seconds) of LPS(MAP) against log sample size.

Algorithm 3 suggests to prefer independence sampling over a grid construction to explore the marginal posterior of the penalty parameters when $q > 4$, see Section 4.2.7 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:

$$\begin{aligned}
f_1(x_1) &= 0.5(2x_1^5 + 3x_1^2 + \cos(3\pi x_1) - 1),\\
f_2(x_2) &= 1.3x_2^5 + \sin(4x_2) + 0.75x_2^2 - 0.25,\\
f_3(x_3) &= \sin(4\pi x_3),\\
f_4(x_4) &= \exp(-x_4^3)\sin(2\pi x_4^2) - 0.1,\\
f_5(x_5) &= 0.8x_5^2(x_5^3 + 2\exp(-3x_5^4 + \log(2x_5 + \pi))) - 0.65,\\
f_6(x_6) &= 1.5\,\big(0.1\sin(2\pi x_6) + 0.2\cos(2\pi x_6) + 0.3\sin^2(2\pi x_6)\\
&\quad +0.4\cos^3(2\pi x_6) + 0.5\sin^3(2\pi x_6)\big) - 0.22.
\end{aligned}$$

There are three additional covariates specified as in Section 4.3.1 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 4.6 shows the simulation results of the Laplace-P-spline approach combined with MCMC. The estimation results obtained with the `gam()` function from the **mgcv** package are shown in parenthesis. Estimated biases shown in Table 4.6 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 **mgcv** algorithms.

Figure 4.3 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 are better captured by increasing the number of B-splines in the basis.

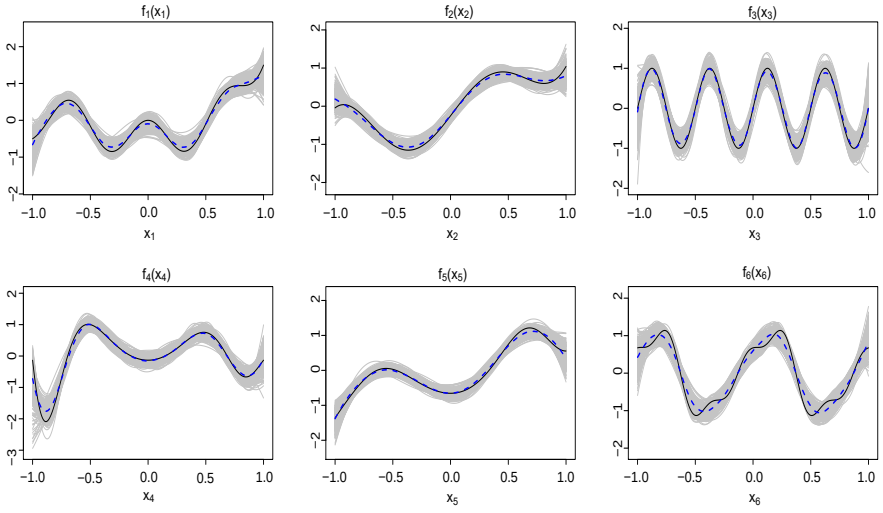



Figure 4.3: Estimation of smooth additive terms $f_1, \ldots, f_6$ (gray curves) for $S = 500$ dataset replications of size $n = 300$ in the Binomial scenario. The dashed line is the pointwise median of the gray curves.

| 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 4.6: 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.

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 4.7, 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 4.4 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.

| | 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 4.7: Average computation time (in seconds) of the LPS-MCMC algorithm over $S = 20$ samples of size $n \in \{300, 1000, 3000\}$ for different dimensions $q \in \{1, 2, 3, 4, 5, 6\}$.

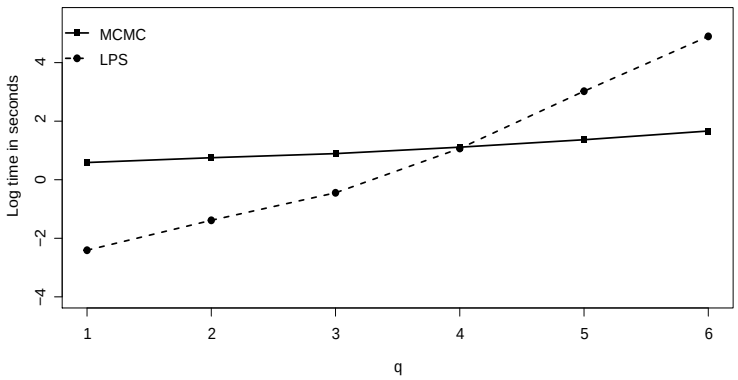



Figure 4.4: Logarithm of the average computation time (in seconds) of LPS (dahsed) and LPS-MCMC (solid) over $S = 20$ samples of size $n = 300$ and dimensions $q \in \{1, 2, 3, 4, 5, 6\}$.

## 4.4 Applications

### 4.4.1 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 semiparametric 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($\mu_i$) ($i = 1, ..., n$) for the number of doctor visits can be written as follows:

$$\begin{aligned} g(\mu_i) &= \beta_0 + \beta_1 Children_i + \beta_2 Race_i + \beta_3 Maritalstatus_i + f_1(Age_i) + \\ &\quad f_2(Income_i) + f_3(Access_i) + f_4(PC1_i), \quad i = 1, \ldots, n, \end{aligned}$$

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.

Table 4.8 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 4.5 displays the estimated smooth functions (solid curves) and the associated 95% approximate pointwise credible intervals (gray surfaces).

| **Parameters** | **Estimates** | **CI 90%** | $\mathbf{sd}_{post}$ |
|---|---|---|---|
| $\beta_1$ (*Children*) | -0.179 | [-0.239; -0.122] | 0.036 |
| $\beta_2$ (*Race*) | -0.127 | [-0.263; 0.005] | 0.081 |
| $\beta_3$ (*Maritalstatus*) | -0.234 | [-0.431; -0.043] | 0.118 |

Table 4.8: 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.

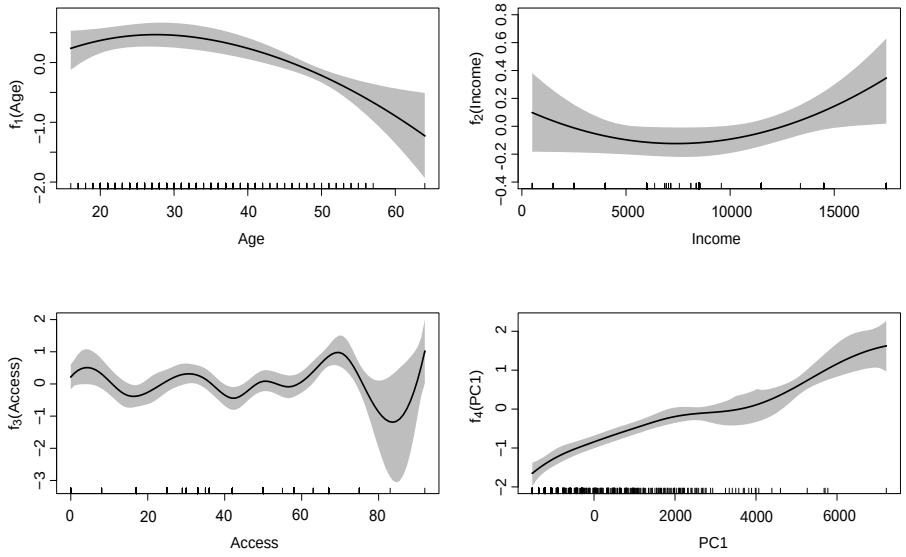



Figure 4.5: Estimated smooth functions (solid curve) and 95% approximate pointwise credible intervals (gray surface) for variables *Age*, *Income*, *Access* and *PC1*.

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. 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.

### 4.4.2 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 rela-

tionship 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:

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

In Table 4.9, 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.

| Parameters | Estimates | CI 90% | $\mathbf{sd}_{post}$ |
|---|---|---|---|
| $\beta_1$ $(BMI)$ | -0.034 | [-0.046; -0.022] | 0.007 |
| $\beta_2$ $(Betadiet)$ | 0.047 | [-0.009; 0.101] | 0.033 |
| $\beta_3$ $(Gender)$ | 0.300 | [ 0.076; 0.520] | 0.135 |
| $\beta_4$ $(Smoking)$ | -0.301 | [-0.515; -0.093] | 0.128 |
| $\beta_5$ $(Fiber)$ | 2.396 | [ 0.804; 3.938] | 0.956 |
| $\beta_6$ $(Fat)$ | -0.245 | [-0.493; -0.003] | 0.149 |

Table 4.9: 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 4.6 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.

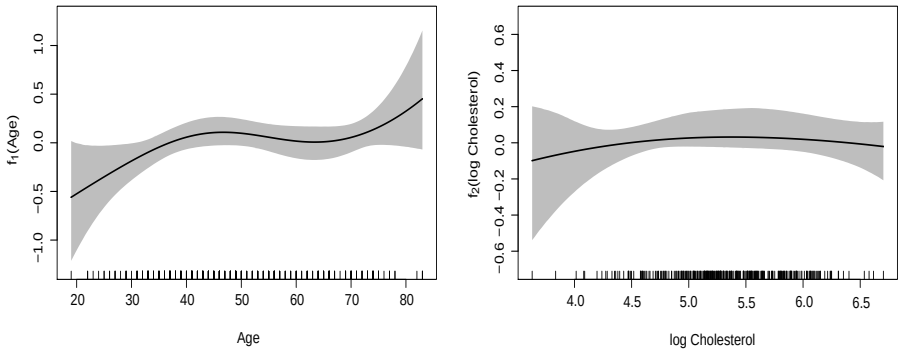



Figure 4.6: Estimated smooth functions (solid curve) and 95% approximate pointwise credible intervals (gray surface) for variables *Age* and log(*Cholesterol*) of the nutritional study dataset.

## 4.5 Concluding remarks

In this chapter, we have put forward a new methodology for approximate Bayesian estimation in generalized additive models (GAMs) by unifying 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.

# CHAPTER 5

# The blapsr package for approximate Bayesian inference with LPS

This chapter is based on: Gressani, O. and Lambert, P. (2020b). The blapsr package for fast Bayesian inference in latent Gaussian models by combining Laplace approximations and P-splines. Version 0.5.1, published on *CRAN*.

## 5.1 Motivation

The **blapsr** package implements Bayesian-based approximate inference in survival models and (generalized) additive models by coupling P-splines for flexible modeling of functional structures and Laplace approximations to bypass Markov chain Monte Carlo (MCMC) methods for fast derivation of posterior distributions. In particular, the routines allow **R** users to fit the Cox model and promotion time cure model for right censored event-times, as well as the additive partial linear model and generalized additive model (GAM) with a response belonging to the one-parameter exponential family. Penalized Bayesian B-splines are used to model smooth functional components. In presence of multiple smooth terms, the optimal penalty choice is driven by analytically available first and second derivatives of the posterior penalty vector. For each of these models, the syntax is intuitive and various options are available to evaluate, analyze and visualize posterior quantities of interest. The aim of this chapter is to describe the main functionalities of the package.

## 5.2 Introduction

Latent Gaussian models (LGMs) encompass a wide range of popular statistical models that flexibly relate a set of covariates to a response, allowing to apprehend nonlinear and complex relationships often encountered in practical applications. A LGM has a hierarchical structure comprising three layers. The first two layers consist in a (potentially high-dimensional) latent vector of model parameters $\boldsymbol{\xi}$ and another vector $\boldsymbol{\eta}$ of smaller dimension that contains the hyperparameters. Conditionally on the hyperparameter vector, the latent vector is assigned a Gaussian prior, while the prior distributional assumptions on the components of $\boldsymbol{\eta}$ are not restricted to Gaussianity. At the bottom of the hierarchy, we find the third layer in which the parameters of the prior distributions of $\boldsymbol{\eta}$ are gathered.

The **blapsr** package focuses on four model classes that belong to LGMs, namely Cox proportional hazards models (Cox, 1972), promotion time cure models (Yakovlev et al., 1996; Tsodikov, 1998; Chen et al., 1999), additive partial linear models (Opsomer and Ruppert, 1999; Fan and Li, 2003) and generalized additive models (Hastie and Tibshirani, 1986, 1987). P-splines (Eilers and Marx, 1996) serve as the main smoother in the LPS approach as they enjoy the following three elegant properties: (1) the penalty matrix is sparse and easily constructed from simple difference matrices, (2) P-splines are categorized as low-rank (or reduced-knot) smoothers, such that the number of knots used to construct the spline basis is usually considerably smaller than the sample size, keeping the computational budget minimal and (3) P-splines can be straightforwardly adapted to a Bayesian setting by translating the roughness penalty in a multivariate Gaussian distribution for the spline amplitudes (Lang and Brezger, 2004).

The Comprehensive **R** Archive Network (`https://cran.r-project.org/`) hosts an incredibly large number of packages dedicated to statistical inference in structured additive regression models. Among those, the **mgcv** package (Wood, 2017) is probably the most appreciated toolkit to fit generalized additive models, as it provides a rich collection of smoothers to choose from, with stable and fast routines. Also, the **R-INLA** package that can be found on the website (`http://www.r-inla.org/`) is based on the Integrated Nested Laplace Approximations tech-

nique pioneered by Rue et al. (2009) which has proved to work well in a variety of applications and has therefore been largely recognized in the statistical community. For the analysis of survival data, established routines to fit Cox models are provided in the **survival** package (Therneau, 2020). Few **R** functions are dedicated to cure models; among others, the **smcure** package Cai et al. (2012) can be used for inference in semiparametric mixture cure models and the more recent package **miCoPTCM** (Bertrand et al., 2017) implements functionalities to fit promotion time cure models taking into account mis-measured covariates. The **blapsr** project can be seen as an extension to the aforementioned packages with a particular emphasis on the combination of Laplace approximations and penalized regression splines serving as the main mechanism for approximate Bayesian inference. The key advantage of our LPS approach lies in its analytical tractability. Focusing on a single type of smoother (P-splines) gives us the opportunity to write down the full likelihood of the model and hence obtain full-fledged analytical expressions for the conditional posterior of the latent vector and the posterior penalty vector. This in turn, permits exact derivation of the gradient and Hessian of the posterior penalty vector, which offers a non-negligible computational advantage to explore the posterior penalty space with optimization methods and grid-based (or sampling) approaches. Furthermore, even for complex functions of latent variables, it is possible to obtain accurate point and set estimates. Finally, the LPS methodology has shown to have excellent statistical properties and is relatively fast despite being a fully Bayesian approach.

This chapter illustrates the use of the **blapsr** package through **R** code on simulated and real data. The purpose is not to provide a stand-alone text of the package, nor is it to give a long-winded exposition of all the options available. Rather, the emphasis is placed on a first "hands-on" experience with the available routines and how they can be used to extrapolate important information from the data. The reader interested in a deeper understanding of the routines can consult the package documentation in which all functionalities are thoroughly documented as well as the website dedicated to the project `https://www.blapsr-project.org/` that has many examples and explanations. The rest of this chapter is organized as follows. In Section 5.3, a compact formulation of latent Gaussian models is presented with a brief discussion on the challenges

to be surmounted for model fitting. Section 5.4 is devoted to the core functions for inference in the Cox model and the promotion time cure model. The latter routines are illustrated on a simulated example and on colon cancer data. Section 5.5 presents the available functions to fit (generalized) additive models. Finally, the chapter closes with a discussion in Section 5.6.

## 5.3 Laplace-P-splines in latent Gaussian models

The latent vector of the models involved in the **blapsr** package is of the form $\boldsymbol{\xi} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$, where $\boldsymbol{\beta}$ is a vector of regression parameters including (or not) an intercept and $\boldsymbol{\theta}$ is a vector of B-spline coefficients. In additive models, this vector is often high-dimensional. For instance, in a GAM with, say 5 smooth terms, each modeled with a basis containg 30 B-splines and 3 covariates in the linear part, there are $\dim(\boldsymbol{\xi}) = (5 \times 29) + 3 + 1 = 149$ latent variables. The number 29 is due to the elimination of the last column of the B-spline matrix because of the identifiability constraint. In the promotion time cure model, an additional vector $\boldsymbol{\gamma}$ is added to the latent vector to model the covariates in the so-called "short-term survival" part (Lambert and Bremhorst, 2019). Model hyperparameters are organized in a vector $\boldsymbol{\eta} = (\boldsymbol{\lambda}^T, \boldsymbol{\delta}^T)^T$, where $\boldsymbol{\lambda}$ is a vector of positive penalty parameters and $\boldsymbol{\delta}$ further hyperparameters present in the robust Gamma prior specification for the roughness penalty parameters following Jullion and Lambert (2007). A Gaussian prior is imposed on the vector of regression and spline parameters conditional on the hyperparameter vector, i.e. $\boldsymbol{\xi}|\boldsymbol{\eta} \sim \mathcal{N}_{\dim(\boldsymbol{\xi})}(0, Q_{\boldsymbol{\xi}}^{-1})$, where $Q_{\boldsymbol{\xi}}$ is a block diagonal (sparse) precision matrix. The nested structure underlying the LPS approach can be written in terms of the conditional posterior of $\boldsymbol{\xi}$ and the marginal posterior of the hyperparameter vector:

$$p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D}) \propto \mathcal{L}(\boldsymbol{\xi}; \mathcal{D})\ p(\boldsymbol{\xi}|\boldsymbol{\eta}), \tag{5.1}$$

$$\widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) \propto \left. \frac{\mathcal{L}(\boldsymbol{\xi}; \mathcal{D})\ p(\boldsymbol{\xi}|\boldsymbol{\eta})\ p(\boldsymbol{\eta})}{\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})} \right|_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}}, \tag{5.2}$$

where $\mathcal{D} = \cup_{i=1}^n \mathcal{D}_i$ is the set of observables for a sample of size $n$, $\widetilde{p}_G(\cdot)$ denotes a Gaussian approximation to the conditional latent vector posterior with posterior mode $\widehat{\boldsymbol{\xi}}_{\boldsymbol{\lambda}}$, where the subscript explicitly indicates

that the mode depends on the penalty vector $\boldsymbol{\lambda}$. Note that for an additive partial linear model with normal errors, $p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})$ is Gaussian, so that the expression for the posterior in (5.2) is exact and not an approximation. The Gaussian approximation to the conditional posterior of $\boldsymbol{\xi}$ is obtained through Newton-based methods. To avoid numerical pitfalls in Newton-Raphson approaches, it is important to guarantee that the negative Hessian matrix of the concerned posterior distribution to be optimized is positive definite at each step of the algorithm to ensure an ascent direction (for maximization). Even if we proceed in an ascent direction, it does not necessarily mean that an ascent will be reached in a given step. Therefore, a step-halving strategy is highly recommended as a warranty to increase the objective function at each step.

After appropriate integration of the posterior in (5.2), one obtains the marginal (approximate) posterior distribution $\widetilde{p}(\mathbf{v}|\mathcal{D})$, where $\mathbf{v}$ is the vector of log penalty parameters. Using a grid-based method (or a sampling scheme), a collection of quadrature points $\{\mathbf{v}^{(m)}\}$ is used to approximate the joint posterior distribution of $\boldsymbol{\xi}$:

$$\begin{aligned} p(\boldsymbol{\xi}|\mathcal{D}) &= \int_{\mathbb{R}_{++}} p(\boldsymbol{\xi}|\mathbf{v}, \mathcal{D})\; p(\mathbf{v}|\mathcal{D})\; d\mathbf{v} \\ &\approx \sum \widetilde{p}_G(\boldsymbol{\xi}|\mathbf{v}^{(m)}, \mathcal{D})\; \widetilde{p}(\mathbf{v}^{(m)}|\mathcal{D})\; \Delta\mathbf{v}. \end{aligned} \tag{5.3}$$

Posterior estimates and credible intervals for $\boldsymbol{\xi}$ (or functions of latent variables) can then be constructed from (5.3).

For a given model, the effective dimension (ED) also known as the "effective degrees of freedom" (Ye, 1998; Eilers et al., 2015; Wood et al., 2016; Eilers, 2018) can be computed and serves as a measure of model complexity. Let $\mathcal{I} = -\nabla^2_{\boldsymbol{\xi}} \ell(\boldsymbol{\xi}; \mathcal{D})|_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}}$ be the negative Hessian of the log-likelihood $\ell(\cdot)$ evaluated at the posterior estimate of the vector $\boldsymbol{\xi}$. The ED of the whole model is obtained by summing the $\dim(\boldsymbol{\xi})$ elements on the main diagonal of the matrix $\mathcal{H} = (\mathcal{I} + Q_{\boldsymbol{\xi}})^{-1}\mathcal{I}$. In (generalized) additive models, the ED of a given smooth term is computed by summing up the appropriate elements of $\text{diag}(\mathcal{H})$ that correspond to the B-spline coefficients used to approximate the smooth. We refer to Chapters 3 and 4 for theoretical details on the LPS methodology in (generalized) additive models.

## 5.4 The blapsr package for survival analysis

The **blapsr** package can be installed and loaded by typing:

```
R> install.packages("blapsr")
R> library("blapsr")
```

The routines presented in this section are summarized in Table 5.1 and can be used for the analysis of right censored survival data.

| Function name | Description |
|---|---|
| `simsurvdata()` | Simulation of right censored survival data |
| `coxlps()` | Fit a Cox proportional hazards model |
| `plot.coxlps()` | Plot baseline hazard and survival curves |
| `curelps()` | Fit a promotion time cure model |
| `curelps.extract()` | Computation of posterior survival quantities |
| `plot.curelps()` | Plot of posterior survival quantities |

Table 5.1: Routines for survival analysis

### 5.4.1 The `coxlps()` function to fit Cox models

The `coxlps()` function is illustrated on simulated data obtained with the `simsurvdata()` routine which generates right censored time-to-event data with latent event times drawn from a Weibull distribution parameterized by a shape $a > 0$ and scale $b > 0$ parameter. Latent event times are generated following Bender et al. (2005) and censoring times are drawn from an exponential distribution. The user can specify a sample size, a vector of regression coefficients and a target censoring percentage.

```
R> require("survival")
R> set.seed(3)
R> simul <- simsurvdata(a = 3.5, b = 4, n = 250,
+ betas = c(0.7, -0.8, 0.4), censperc = 15)
R> simul

Sample size:          250
Censoring:            Exponential
Number of events:     210
Censoring percentage: 16%
```

```
Weibull mean:         3.60
Weibull variance:     1.30

R> simdat <- simul$survdata
R> head(simdat, 5)

    time     delta      x1         x2          x3
1 6.166957     1   -0.96193342 -0.5350631 -2.13984191
2 2.046183     0   -0.29252572  1.3681062 -1.26347924
3 4.993415     1    0.25878822  0.1418443  0.08330797
4 1.918469     0   -1.15213189 -0.7828150  0.18832513
5 4.509665     1    0.19578283  1.8815187  0.18981419

R> plot(simul)
```

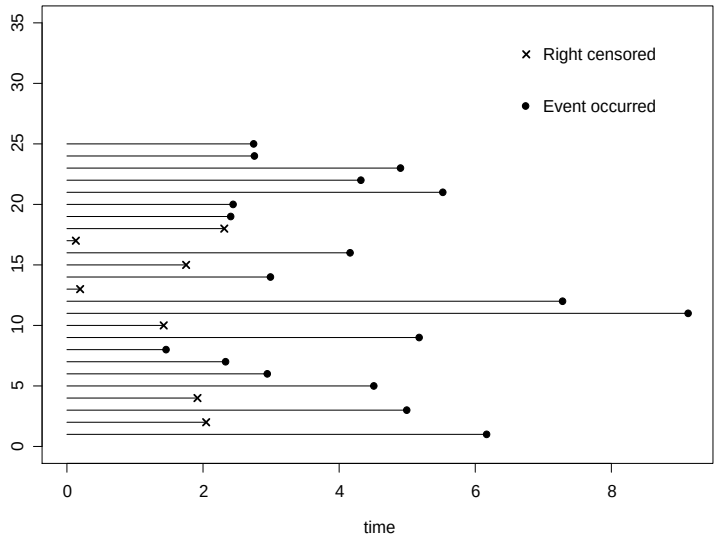



Figure 5.1: Overview of the first 25 observed follow-up times.

The `coxlps()` routine models the (log) baseline hazard as a linear combination of cubic B-splines computed with the routine `cubicbs()` in the interval $[0, t_u]$, where $t_u$ is the upper bound of the follow-up (here $t_u = 11.834$). The formula syntax is the same as for the `coxph()` formula of the **survival** package. Fitting the Cox model with 25 cubic B-splines and a second order penalty (the default) yields:

```
R> fit <- coxlps(Surv(time, delta) ~ x1 + x2 + x3,
+ data = simdat, K = 25)
R> fit

Formula:
Surv(time, delta) ~ x1 + x2 + x3
Object class: "coxlps"

Number of B-splines in basis: 25
Number of parametric coeffs.: 3
Latent vector dimension     : 28
Penalty order               : 2
Sample size                 : 250
Number of events:           : 210
Effective dimension (ED)    : 7.80

Estimated model coefficients:
       coef   exp(coef) sd.post    z
x1   0.8210    2.2729   0.0809  10.1507
x2  -0.9059    0.4042   0.0870 -10.4171
x3   0.3106    1.3643   0.0804   3.8659

     exp(coef) exp(-coef) lower.95 upper.95
x1    2.2729     0.4400   1.9365   2.6607
x2    0.4042     2.4741   0.3403   0.4788
x3    1.3643     0.7330   1.1636   1.5955
---
AIC.p = 377.5761 AIC.ED = 387.1704
BIC.p = 387.6174 BIC.ED = 413.2684
```

The first few lines of output contains information on the structure of the vector $\boldsymbol{\xi}$ which can be decomposed into B-spline coefficients and parametric terms associated to the covariates of the model. The penalty order, sample size, number of events and effective model dimension are also summarized. The table of estimated model coefficients gives a detailed account of posterior pointwise and set estimates of the regression coefficients. The last lines of output provide the Akaike information criterion (Akaike, 1973) and the Bayesian information criterion or Schwarz

criterion (Schwarz et al., 1978) that are useful for model selection. The AIC and BIC are given by AIC.p $= -2\ell(\widehat{\boldsymbol{\xi}}, \mathcal{D}) + 2p$, where $p$ is the number of regression coefficients; BIC.p $= -2\ell(\widehat{\boldsymbol{\xi}}, \mathcal{D}) + p\log(n_e)$, where $n_e$ is the number of non-censored event times. Formulas to compute AIC.ED and BIC.ED are analogous and use the ED instead of $p$.

The `plot.coxlps()` routine allows the user to plot the estimated baseline hazard and survival functions with an associated credible interval at level $(1-\alpha)\times 100\%$. It is also possible to overlay the estimated Kaplan-Meier curve by setting the `overlay.km` option to `TRUE`. The code below plots the smooth estimated baseline survival function with a 90% approximate pointwise credible interval and makes a comparison with the true baseline survival used to simulate the data as shown in Figure 5.2.

```
R> domt <- seq(0, 7, length = 200)
R> plot(fit, h0 = FALSE, cred.int = 0.90, overlay.km = F,
+ plot.cred = T, xlim = c(0, 7), show.legend = F)
R> lines(domt, simul$S0(domt), type = "l", col = "red",
+ lty = 2, lwd = 2)
R> legend("topright", col = c("black", "gray", "red"),
+ lty = c(1, 1, 2), c("Bayesian LPS", "90% CI", "Target"),
+ cex = 0.8, bty = "n")
```

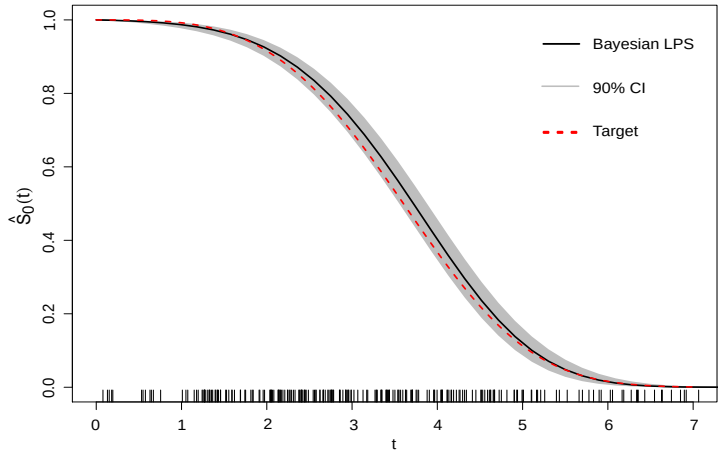



Figure 5.2: Estimated baseline survival function (black curve). The gray surface corresponds to a 90% credible interval and the dashed curve is the target baseline survival.

### 5.4.2 The promotion time cure model with `curelps()`

A phenomenon often encountered in the analysis of time-to-event data is the existence of a fraction of long-term survivors that will never experience the monitored event irrespective of the duration of the follow-up. The promotion time cure model is an extension of the Cox model that endorses the existence of an unidentified proportion of cured subjects. Although the model was initially motivated by a biological context involving cancer cells (Yakovlev et al., 1996; Tsodikov, 1998; Chen et al., 1999) it can be extrapolated to a more generic framework in which the model skeleton is divided in two parts. The first part involves covariates influencing the cure probability or "long-term survival" and the second part incorporates covariates affecting the population hazard dynamics or "short-term survival" (Bremhorst and Lambert, 2016; Gressani and Lambert, 2018; Lambert and Bremhorst, 2019). The `curelps()` function follows this semantic in the formula argument through the terms `st()` and `lt()`, offering the user an intuitive and simple syntax. For instance, `Surv(time, event)` $\sim$ `lt(x1 + x2) + st(x1 + x3)` is a formula specifying a promotion time cure model with covariate `x1` affecting survival jointly in the long- and short term, while `x2` and `x3` only affect long-term and short-term survival respectively. Bayesian Laplace-P-splines have good statistical properties when applied to such survival models with a cured fraction (cf. Chapter 2) and drastically outperform MCMC methods from a computational perspective.

The `curelps()` routine is illustrated on colon cancer data available in the **survival** package. The study goes back to Laurie et al. (1989) in which eligible patients were assigned either to observation alone (no adjuvant therapy), to treatment with Levamisole or to a combination of Levamisole and Fluorouracil (5-FU). The data has further been investigated in Moertel et al. (1990) and Moertel et al. (1995) with encouraging results for the treatment Levasimole plus 5-FU to reduce cancer recurrence. More recently, Lambert and Bremhorst (2019) fitted a promotion time cure model to these data by combining P-splines for flexible estimation of the (log) baseline hazard and a Metropolis-within-Gibbs algorithm to sample the joint posterior of the model. With the few **R** lines below, one obtains the Kaplan-Meier estimates of the survival curves for the recurrence times (in years) for each treatment.

```
R> library(survival)
R> data("colon")
R> colondat <- subset(colon, (etype==1) & (!is.na(nodes)) &
+ (!is.na(differ)))
R> colondat$time <- colondat$time / 365
R> plot(survfit(Surv(time, status) ~ rx, data = colondat),
+ col = c("black", "red", "blue"), lty = c(1,2,3),
+ lwd = c(2,2,2), mark.time = T, mark = "x",
+ xlab ="Time t (in years)", ylab = expression(S[p](t)))
R> legend("topright", c("Obs", "Lev", "Lev+5FU"),
+ lty = c(1,2,3), lwd =  + c(2,2,2), bty = "n",
+ col = c("black", "red", "blue"), cex = 0.8)
```

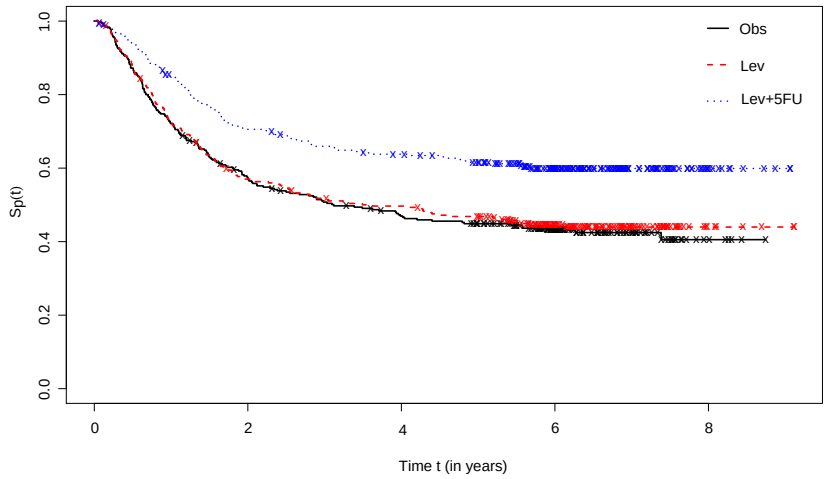



Figure 5.3: Kaplan-Meier curves for the time to recurrence in each treatment group.

The treatment (`rx`), number of lymph nodes (`nodes`) and cancer extent (`extent`) are the three (factor) covariates entering the long-term survival, while the short-term survival part includes the number of lymph nodes and the tumour differentiation (`differ`). Before fitting the model, we proceed to a recategorization of covariates (cf. Lambert and Bremhorst, 2019, Section 4):

```
R> levels(colondat$rx) <- c("Obs", "Obs", "Lev+5FU")
R> colondat$nodes[colondat$nodes <= 2] <- 1
R> colondat$nodes[colondat$nodes >= 3 &
```

```
+ colondat$nodes <= 5] <- 2
R> colondat$nodes[colondat$nodes >= 6] <- 3
R> colondat$nodes <- as.factor(colondat$nodes)
R> levels(colondat$nodes) <- c("<=2", "[3-5]", ">=6")
R> colondat$extent <- as.factor(colondat$extent)
R> levels(colondat$extent) <- c("Submucosa/muscle",
+ "Submucosa/muscle", "Serosa", "Contig.structures")
R> colondat$extent <- factor(colondat$extent,
+ levels(colondat$extent)[c(2,1,3)])
R> colondat$differ <- as.factor(colondat$differ)
R> levels(colondat$differ) <- c("Well/moderate",
+ "Well/moderate", "Poor")
```

The promotion time cure model described in Chapter 2 is fitted using the following syntax with 20 B-splines and a third-order penalty:

```
R> fit <- curelps(Surv(time,status) ~ lt(rx + nodes + extent)
+ st(nodes + differ), data = colondat, K = 20, penorder = 3)
R> fit

Formula:
Surv(time, status) ~ lt(rx + nodes + extent)
+ st(nodes + differ)
Object class: "curelps"

Number of B-splines in basis: 20
Number of parametric coeffs.: 9
Latent vector dimension:      29
Penalty order:                3
Sample size:                  888
Number of events:             446
Effective model dimension:    11.66

Coefficients influencing the cure probability
(long-term survival):
                    coef   sd.post   z     lower.95 upper.95
(Intercept)       -0.3306  0.0541 -6.1149  -0.4376  -0.2253
Lev+5FU           -0.5026  0.1091 -4.6072  -0.7186  -0.2901
```

```
[3-5]              0.4348  0.1217  3.5742   0.1939   0.6718
>=6                0.8422  0.1281  6.5749   0.5886   1.0917
Submucosa/muscle  -0.5631  0.1713 -3.2865  -0.9023  -0.2293
Contig.structures  0.4811  0.2108  2.2824   0.0637   0.8916

Coefficients affecting the population hazard dynamics
(short-term survival):
       coef   exp(coef) sd.post   z
[3-5] 0.2849    1.3297  0.1539 1.8508
>=6   0.2890    1.3351  0.1626 1.7771
Poor  0.6979    2.0096  0.1444 4.8323
---
        exp(coef) exp(-coef) lower.95  upper.95
[3-5]    1.3297     0.7521    0.9803   1.7946
>=6      1.3351     0.7490    0.9675   1.8328
Poor     2.0096     0.4976    1.5098   2.6626
---
AIC.p = 2415.9369 AIC.ED = 2421.2658
BIC.p = 2452.8398 BIC.ED = 2469.0936
```

The `curelps.extract()` routine is used to compute the cure prediction for a given profile of covariates, i.e. the probability that a subject is cured given that (s)he has survived until a certain time point $t$. We compare the difference in cure probability at times $t = (0.5, 1, 2)$ between groups receiving Levamisole or no treatment versus Levamisole plus 5-FU for subjects having [3-5] lymph nodes with extent of local spread in Serosa and a poor differentiation of tumour.

```
R> profileConTLEV <- c(0, 1, 0, 0, 0, 1, 0, 1)
R> profileLEV5FU <- c(1, 1, 0, 0, 0, 1, 0, 1)
R> curelps.extract(fit, time = c(0.5, 1 ,2),
+ curvetype = "probacure", covar.profile = profileConTLEV)

Estimated cure prediction at specified time points (*):
      Time  Cure.prob(**) Cure.low  Cure.up
[1,] 0.5000    0.5353      0.4008    0.6523
[2,] 1.0000    0.7029      0.5333    0.8206
[3,] 2.0000    0.9013      0.7435    0.9642
```

```
---
* Bounds correspond to a 95.00% credible interval.
** Cure prediction for covariate profile: 0, 1, 0, 0, 0, 1,
0, 1 .

R> curelps.extract(fit, time = c(0.5, 1 ,2),
+ curvetype = "probacure", covar.profile = profileLEV5FU)

Estimated cure prediction at specified time points (*):
      Time  Cure.prob(**)  Cure.low Cure.up
[1,] 0.5000    0.6852      0.5553    0.7842
[2,] 1.0000    0.8079      0.6725    0.8917
[3,] 2.0000    0.9391      0.8321    0.9787
---
* Bounds correspond to a 95.00% credible interval.
** Cure prediction for covariate profile: 1, 1, 0, 0, 0, 1,
0, 1 .
```

The output shows the estimated cure prediction for time values summarized in the first column along with 95% approximate quantile-based credible intervals. The `plot.curelps()` function is used to plot smooth estimates of the cure prediction as shown in Figure 5.4.

```
R> par(mfrow = c(1, 2))
R> plot(fit, curvetype = "probacure",
+ covar.profile = profileConTLEV,
plot.cred = T, ylim = c(0, 1), xlim = c(0, 4),
main = "ConTLEV", cex.main = 0.8, show.legend = F)
R> legend("bottomright", c("Cure proba.", "95% CI"),
+ lty = c(1, 1), col = c("black", "gray75"), bty = "n")

R> plot(fit, curvetype = "probacure",
+ covar.profile = profileLEV5FU,
plot.cred = T, ylim = c(0, 1), xlim = c(0, 4),
main = "LEV + 5FU", cex.main = 0.8, show.legend = F)
R> legend("bottomright", c("Cure proba.", "95% CI"),
+ lty = c(1, 1), col = c("black", "gray75"), bty = "n")
```

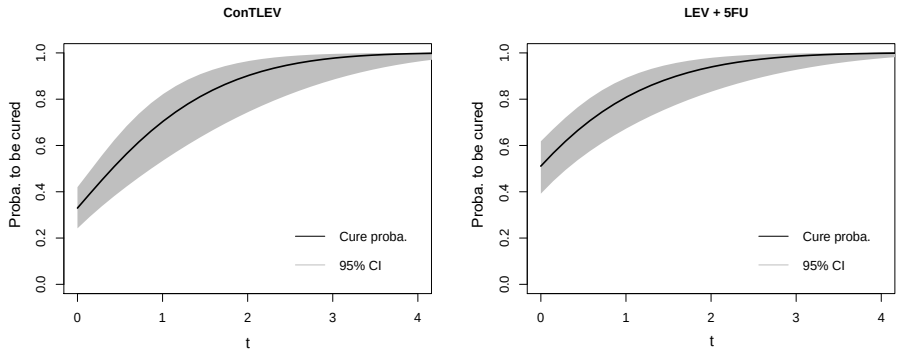



Figure 5.4: Estimated cure prediction for groups receiving no adjuvant therapy or Levamisole alone (left) and Levamisole plus 5-FU (right). The gray surface represents approximate 95% pointwise credible intervals.

## 5.5 Routines for (generalized) additive models

The aim of this section is to present the routines of the **blapsr** package dedicated to the analysis of additive models and generalized additive models. Additive models play an important role in the statistical literature as they provide well-tailored regression tools to capture nonlinearities in the data. They also allow to deviate from the (often restrictive) assumption of a response being governed by Gaussianity, by considering distributions belonging to a more general class. The functions for additive modeling using the LPS approach are summarized in Table 5.2.

| Function name | Description |
|---|---|
| `amlps()` | Additive partial linear modeling with LPS |
| `gamlps()` | Generalized additive models with LPS |
| `plot.amlps()` | Plot smooth terms for additive models |
| `simgamdata()` | Data simulation for GAMs |
| `plot.gamlps()` | Plot smooth terms for GAMs |

Table 5.2: Routines for (generalized) additive modeling

### 5.5.1 Additive partial linear models with normal errors

The ozone data is a benchmark dataset in the GAM literature and has extensively been used to illustrate nonparametric regression techniques. It has originally been analyzed by Breiman and Friedman (1985) to study

the relationship between ozone concentration in the atmosphere and a set of meteorological covariates measured in the Los Angeles area. The `amlps()` routine is illustrated on the ozone data obtained from the **ibr** package using the log of ozone concentration as a response with $n = 330$ observations and eight covariates summarized in Table 5.3.

| Variable name | Description |
|---|---|
| `vh` | 500 millibar pressure height (m) |
| `wind` | Wind speed (mph) |
| `humidity` | Humidity (in %) |
| `temp` | Temperature (°F) measured at Sandburg, CA |
| `ibh` | Inversion base height (feet) |
| `dpg` | Pressure gradient (mmHg) |
| `ibt` | Inversion base temperature (°F) |
| `vis` | Visibility (miles) |

Table 5.3: Meteorological covariates for the ozone data

The formula syntax of `amlps()` closely mimics the syntax used in the `gam()` function of the **mgcv** package (Wood, 2017) to specify smooth terms. For instance, the formula `y ~ z1 + z2 + sm(x1) + sm(x2)` specifies an additive partial linear model with continuous or categorical covariates `z1` and `z2` in the linear part and two smooth terms depending on the continuous covariates `x1` and `x2` respectively. The following code illustrates the use of the `amlps()` routine to fit the ozone data with all covariates in the smooth part of the model, 25 cubic B-splines in the basis and a second order penalty:

```
R> library("ibr")
R> data("ozone")
R> ozonedat <- ozone
R> colnames(ozonedat) <- c("ozone", "vh", "wind", "humidity",
+ "temp", "ibh", "dpg", "ibt", "vis")
R> fit <- amlps(log(ozone) ~ sm(vh) + sm(wind) +
+ sm(humidity) + sm(temp) + sm(ibh) + sm(dpg) + sm(ibt) +
+ sm(vis), data = ozonedat, K = 25, penorder = 2)
R> fit
```

```
Formula:
log(ozone) ~ sm(vh) + sm(wind) + sm(humidity) + sm(temp) +
sm(ibh) + sm(dpg) + sm(ibt) + sm(vis)

Sample size:                  330
Number of B-splines in basis: 25
Number of smooth terms:       8
Penalty order:                2
Latent vector dimension:      193
Model degrees of freedom:     23.49

Linear coefficients:
             Estimate sd.post z-score  lower .95 upper .95
(Intercept)   1.9447   0.0686  92.0280   1.9033    1.9862
---

Effective degrees of freedom of smooth terms:
             edf     lower.95  upper.95   Tr       p-value
sm(vh)       1.6900   1.0000    9.6173    1.5426   0.4900151
sm(wind)     2.3603   1.0783    4.5525    4.0142   0.2582988
sm(humidity) 2.3467   1.2131    4.2485    5.8127   0.1175616
sm(temp)     3.0910   1.0477    5.8101   29.2759   6.125e-06 ***
sm(ibh)      3.2234   1.2246    5.5781   15.6525   0.0033446 **
sm(dpg)      4.0310   2.1824    6.4819   22.9305   0.0003743 ***
sm(ibt)      2.2326   1.0000   12.7350    2.3126   0.4858104
sm(vis)      3.5165   1.0739    6.9757   17.0811   0.0025989 **

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
---
Posterior interval corresponds to a 95% HPD interval

Estimated standard deviation of error: 0.3839
Adjusted R-squared: 0.7657
```

The output starts with a couple of lines summarizing the model specified by the user, the number of latent variables to be estimated and the degrees of freedom of the model as a proxy for model complexity. Next, a table containing point and set estimates of the coefficients in the linear

part of the model are shown (here only the intercept). Finally, the last table displays results on the estimated effective degrees of freedom (edf) of the smooth terms together with a 95% highest posterior density (HPD) interval. A test statistic $T_r$ (Wood, 2013) and its associated p-value is also provided to test the presence of a significant effect of the corresponding covariate.

The theoretical effective degrees of freedom of a smooth term varies in the range $[r-1, \min(n, K-1)]$, where $r$ denotes the penalty order and $K$ the number of B-splines in the basis. Note that one degree of freedom is lost due to the identifiability constraint inherent in additive models. Values in the edf column measure the complexity of the fitted smooth functions. An estimated edf value for a smooth term close to unity means that the true function to be estimated is close to linearity. The 95% HPD credible interval is based on a sample of effective degrees of freedom computed from a random sample of log roughness penalty parameters generated from a Gaussian approximation to the (log) posterior penalty vector around its posterior mode, see Section 3.5.1. Despite being a crude approximation, it gives us an idea of the uncertainty associated to the estimation of the edf. In the above results, we see that the smallest edf values arise for the variables `vh`, `wind`, `humidity` and `ibt`.

The statistic $T_r$ is a Wald-type statistic used to test the null hypothesis $H_0 : f_j(x) = 0 \ \forall x \in \mathcal{X}_j$ versus $H_0 : f_j(x) \neq 0$ for some $x \in \mathcal{X}_j$, where $\mathcal{X}_j$ denotes the range of $f_j$. Non-rejection of $H_0$ suggests to drop the $j$th covariate from the model. Let $\widetilde{B}_j$ be the B-spline basis such that $\widehat{f}_j = \widetilde{B}_j \widehat{\boldsymbol{\theta}}_j$. The covariance of $\widehat{f}_j$ is the $n \times n$ matrix $V_{\widehat{f}_j} = \widetilde{B}_j \widehat{\Sigma}_{\boldsymbol{\theta}_j} \widetilde{B}_j^T$, where $\widehat{\Sigma}_{\boldsymbol{\theta}_j}$ is the estimated variance-covariance matrix of the B-spline coefficients associated to the $j$th smooth. A well-behaved Wald statistic proposed in Wood (2013) is $T_r = \widehat{f}_j^T V_{\widehat{f}_j}^{r-} \widehat{f}_j$, where $r$ is the estimated effective degrees of freedom of the concerned smooth term and $V_{\widehat{f}_j}^{r-}$ is a rank-$r$ Moore-Penrose inverse of $V_{\widehat{f}_j}$. The p-values are computed using that under the null hypothesis, $T_r$ is approximately Gamma distributed $\mathcal{G}(r/2, 1/2)$, so that $E(T_r) = r$ and $V(T_r) = 2r$. According to the computed p-values, we decide to remove `vh`, `wind`, `humidity` and `ibt` from the smooth part of the model. The `plot.amlps()` routine is used to plot the smooth terms using the following commands:

```
R> par(mfrow = c(2,3))
R> for(j in 1:8) plot(fit, smoo.index = j, ylim = c(-1, 1))
```

The plot is shown in Figure 5.5. The vertical ticks on the abscissa correspond to the observed covariate values. On the vertical axis, the variable names are given together with their associated degrees of freedom.

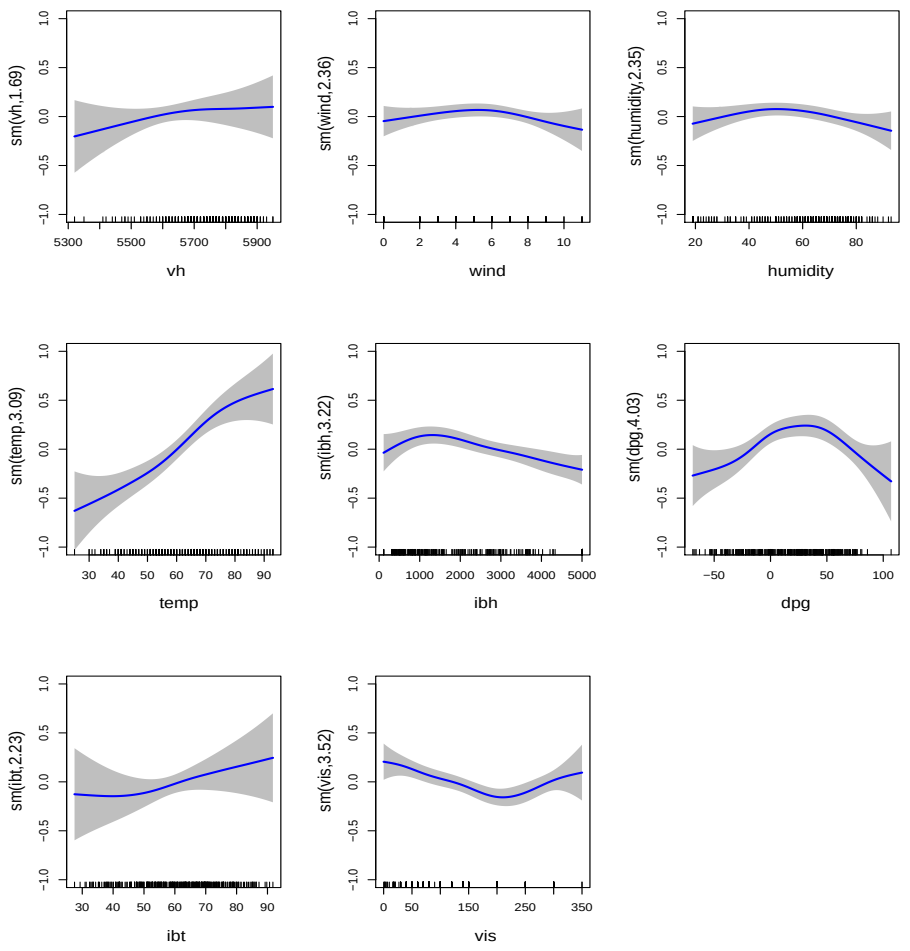



Figure 5.5: Estimated smooth terms for the ozone dataset with approximate 95% pointwise credible intervals. Vertical ticks on the abscissa correspond to observed covariate values.

We fit a second model (not reported here) with the nonsignificant smooth variables entering as linear components. As none of the linear components were significant, we decide to fit a simplified model with only two covariates:

```
R> fit3 <- amlps(log(ozone) ~  temp  + sm(dpg),
+ data = ozonedat, penorder = 2)
R> fit3

Formula:
log(ozone) ~ temp + sm(dpg)

Sample size:                  330
Number of B-splines in basis: 30
Number of smooth terms:       1
Penalty order:                2
Latent vector dimension:      31
Model degrees of freedom:     6.74

Linear coefficients:
             Estimate sd.post z-score lower .95 upper .95
(Intercept)  -0.2193  0.0376 -2.0246   -0.4316   -0.0070
temp          0.0374  0.0017 21.8821    0.0341    0.0407
---
Effective degrees of freedom of smooth terms:
          edf    lower.95  upper.95  Tr       p-value
sm(dpg) 4.7385   2.8194    6.9970    54.4669  5.621e-10 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
---
Posterior interval corresponds to a 95% HPD interval
Estimated standard deviation of error: 0.4358
Adjusted R-squared: 0.6698
```

The above results suggest that `temp` and `dpg` are significant variables in explaining the ozone concentration level. Similar conclusions are found in Donnell et al. (1994) and Gu et al. (2010).

### 5.5.2 Generalized additive models: a simulated example

In many practical applications the assumption of Gaussian errors is rather restrictive. GAMs provide a useful extension of generalized linear models (Nelder and Wedderburn, 1972) in the sense that covariates are flexibly related to the mean of a conditional distribution in the exponential family. The `gamlps()` routine can be used to fit GAMs with the LPS methodology for a response belonging to the one-parameter exponential family (cf. Chapter 4). The input components are the same as for `amlps()` with an additional option to specify the family which can be either Gaussian, Poisson, Bernoulli or Binomial. As an illustration, consider the following scenario with a Binomial response $y_i \sim \text{Bin}(15, p_i),\ i = 1, \ldots, 450$ and success probability $p_i$, which can be simulated using the `simgamdata()` routine:

```
R> set.seed(8)
R> sim <- simgamdata(n = 450, dist = "binomial", scale = 0.4)

Setting      : 1
Sample size n: 450
Distribution : binomial
------------------------
Covariates generated:
z1 ~ Bern(0.5)
z2 ~ N(0,1)
xj ~ U(-1,1), j = 1,2,3
True linear coefficients: -1.45 0.25 -0.9

R> simdat <- sim$data
R> head(simdat, 5)
   y  z1     z2         x1         x2          x3
1 13  0  0.19505235 -0.5995711 -0.9276131 -0.0008430188
2  3  0 -0.08121796  0.3704372 -0.1711766 -0.0236410098
3  6  0  1.40425504  0.8337515 -0.3656346 -0.6208308893
4  6  1  0.66141818 -0.4312011 -0.0916422 -0.1154869045
5  8  1  0.52990335 -0.7906997 -0.6233590  0.7552377293
```

The dataset has a binary covariate $z_1 \sim \text{Bern}(0.5)$ and a continuous covariate $z_2 \sim \mathcal{N}(0, 1)$, both of which will enter the linear part of the

model. The remaining covariates $x_j \sim \mathcal{U}(-1,1),\ j = 1,2,3$ have the following associated functions $f_1(x_1) = 0.5(2x_1-1)^2$, $f_2(x_2) = \cos(2\pi x_2)$ and $f_3(x_3) = 4\sin(2\pi x_3)/(2-\sin(2\pi x_3))$. The code below fits a GAM with `gamlps()` on the simulated data and plots the fitted smooth terms with `plot.gamlps()` along with the true target functions.

```
R> fit <- gamlps(y ~ z1 + z2 + sm(x1) + sm(x2) + sm(x3),
+ data = simdat, family = "binomial", nbinom = 15,
+ penorder = 2)
R> fit
Formula:
y ~ z1 + z2 + sm(x1) + sm(x2) + sm(x3)

Family:                         binomial
Link function:                  logit
Sample size:                    450
Number of B-splines in basis: 30
Number of smooth terms:       3
Penalty order:                2
Latent vector dimension:      90
Model degrees of freedom:     33.93

Linear coefficients:
             Estimate  sd.post  z-score lower.95 upper.95
(Intercept)   0.3601   0.0393   9.1647   0.2825   0.4358
z1            0.1929   0.0753   2.5628   0.0443   0.3380
z2           -0.8878   0.0409 -21.6808  -0.9686  -0.8088
---
Effective degrees of freedom of smooth terms:
         edf     lower.95  upper.95    Tr       p-value
sm(x1)  4.9539    3.8587    6.6960   790.6542 < 2.2e-16 ***
sm(x2)  9.4361    7.9371   11.5889   306.5613 < 2.2e-16 ***
sm(x3) 16.5439   15.0965   18.5881  1433.4472 < 2.2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
 ‘ ’ 1,
---
Posterior interval corresponds to a 95% HPD interval
```

```
Adjusted R-squared: 0.9443
R> par(mfrow = c(2,2))
R> domx <- seq(-1, 1, length = 300)
R> f1target <- sim$f[[1]](domx) - mean(sim$f[[1]](domx))
R> f2target <- sim$f[[2]](domx) - mean(sim$f[[2]](domx))
R> f3target <- sim$f[[3]](domx) - mean(sim$f[[3]](domx))
R> plot(sim)
R> plot(fit, smoo.index = 1, cred.int = 0.90,
+ fit.col = "red", ylim = c(-1.6, 3.2))
R> lines(domx, f1target, type ="l", lty = 2, lwd = 2)
R> plot(fit, smoo.index = 2, cred.int = 0.90,
+ fit.col = "red", ylim = c(-1.8, 2))
R> lines(domx, f2target, type ="l", lty = 2, lwd = 2)
R> plot(fit, smoo.index = 3, cred.int = 0.90,
+ fit.col = "red", ylim = c(-2.5, 3.8))
R> lines(domx, f3target, type ="l", lty = 2, lwd = 2)
```

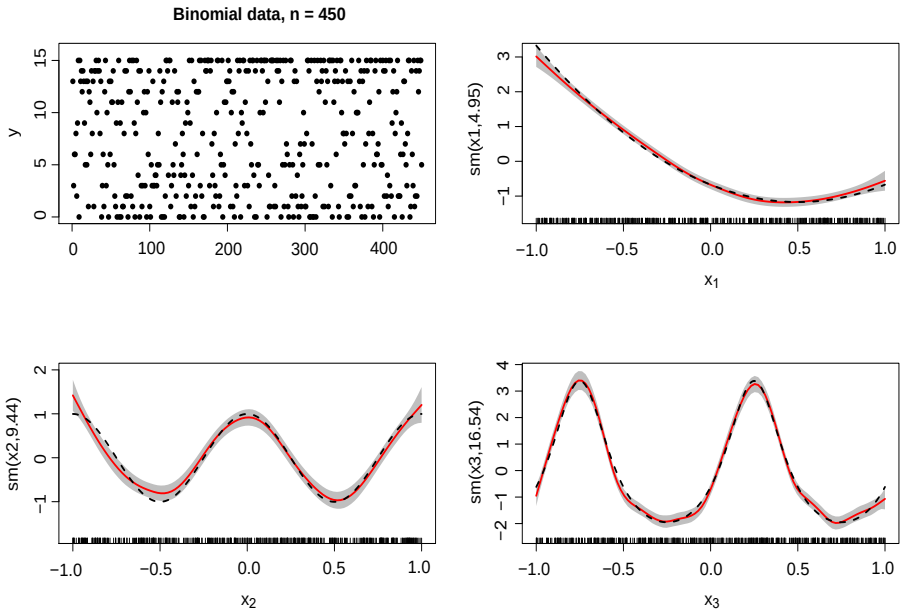



Figure 5.6: Estimated smooth terms for the Binomial simulated dataset with approximate 90% pointwise credible intervals. Dashed curves are the true functions.

## 5.6 Discussion

The **blapsr** package provides **R** functionalities to fit survival models and GAMs based on an approximate Bayesian inference technique relying on a combination of Laplace approximations to conditional posteriors of latent vectors and P-spline smoothers for a flexible specification of nonlinear model terms. The optimal amount of smoothing underlying the routines is automatically determined either through a grid-based strategy or more simply through the posterior maximum penalty value. Being based on a fully Bayesian approach, the **blapsr** package provides the user with routines implicitly taking into account the uncertainty surrounding the smoothing parameters. Furthermore, numerical differentiation to obtain gradients and Hessians of the likelihood or the posterior penalty vector is completely avoided as exact analytical versions have been derived for the considered models, thus reducing the computational cost for model fitting.

From here, several proposals can be taken into account to extend and enhance the **blapsr** package in the future. First of all, even though the routines are already relatively fast as compared to existing fully Bayesian methods, the computational speed can be further enhanced by coding the most costly sub-routines into a faster language; for example $\mathbf{C}_{++}$ or **Fortran**. Second, the analytical availability of the (approximate) posterior penalty vector $p(\mathbf{v}|\mathcal{D})$ and its posterior mode (and variance-covariance matrix) can be considered a good starting point for MCMC sampling. Instead of relying on a grid-based approach, one can think of a routine that allows to sample directly from $p(\mathbf{v}|\mathcal{D})$ using for instance a Metropolis or independence sampler to construct the chains (as in Chapter 4, Section 4.2.7). The latter samples of penalty vectors can then be used to estimate the joint posterior of the spline and regression parameters and any required credible interval on latent variables (or functions thereof). Finally, the good statistical performance behind the LPS methodology encourages its extension to other models, for instance, (bivariate) density estimation, frailty models for survival data, Cox or cure models with time-varying covariates and spatial models.

# CHAPTER 6

# Conclusion

## 6.1 Motivation

This concluding chapter aims at giving a global perspective on the Laplace-P-spline (LPS) methodology developed in this thesis. Taking the time to reflect on LPS will certainly help the reader to grasp the "big picture" sketched by the ideas in the previous chapters and to have a clear overview of the future research prospects to be investigated.

To begin with, a compact recipe is presented in Section 6.2 that emphasizes on the main ingredients and the *modus operandi* employed for implementing the LPS methodology in a general Bayesian setting. It highlights the structure on which LPS is based and is a useful starting point for a researcher wishing to use LPS as an inference instrument in other model classes. In Section 6.3 focus will be placed on the strengths and weaknesses of LPS and a summary about the advantages (and disadvantages) of using LPS over existing competitors will be formulated. The **blapsr** package is also further explored in Section 6.4 by providing additional simple examples in specific settings. Finally, in Section 6.5 we conclude with an extensive but not exhaustive set of possibilities for future research revolving around Laplace-P-splines.

## 6.2 Laplace-P-splines in a nutshell

Ideas behind LPS were strongly motivated by the influential articles of Eilers and Marx (1996) and Rue et al. (2009). The former paper presents a simple, yet powerful approach for flexible modeling of smooth terms in a regression context based on P-splines, with relatively simple formulas to obtain the estimator of the vector of spline amplitudes and a rather intuitive interpretation of the parameter controlling the smoothness of the fit. The latter paper targets the class of latent Gaussian models and proposes a sampling-free methodology based on nested Laplace approximations to approximate the posterior marginals of the latent variables.

### 6.2.1 Numerical considerations behind Laplace approximations

Incorporating the concepts of both papers in a unified framework does not come without challenges. After having specified the Bayesian model, one usually starts by computing the Laplace approximation to the conditional posterior of the regression and spline parameter vector $\boldsymbol{\xi}$ (conditionally on $\boldsymbol{\eta} = (\boldsymbol{\lambda}^\top, \boldsymbol{\delta}^\top)^\top$, where $\boldsymbol{\lambda}$ denotes the vector of penalty parameters). Mathematically, Laplace's method for approximating a multivariate (and differentiable) conditional posterior distribution, say $p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})$, consists in, first, computing the posterior mode $\widehat{\boldsymbol{\xi}}$ by maximizing either analytically or numerically $\log p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})$, and, second, computing the Hessian matrix of $\log p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})$ evaluated at $\widehat{\boldsymbol{\xi}}$, i.e. $\mathcal{H}(\widehat{\boldsymbol{\xi}})$. The resulting Laplace approximation to $p(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})$ is a Gaussian distribution with mean $\widehat{\boldsymbol{\xi}}$ and variance-covariance matrix equal to $-(\mathcal{H}(\widehat{\boldsymbol{\xi}}))^{-1}$, see e.g. Bornkamp (2011).

Maximization of the (log) conditional posterior latent vector is in rare circumstances analytically attainable, such that an iterative numerical approach (e.g. Newton-Raphson) is required. The algorithm must be constructed with extreme caution to reach the desired posterior mode $\widehat{\boldsymbol{\xi}}$ and to avoid numerical pitfalls lurking around the corner. Initializing the iterative optimization process with a "good" starting point for the regression and spline parameters is of crucial importance. Most of the time a zero vector works just fine and leads to convergence in a few steps, but more subtle choices may sometimes be necessary, for instance an initial guess based on a maximum likelihood estimate of the latent vector

or a hybrid specification with a zero vector for the B-spline parameters and another specification for the regression parameters. In any case, the researcher should diagnose whether the final point towards which the algorithm has converged is truly a global maximum.

Another recommendation with respect to the Newton-Raphson algorithm is to check that, at each iteration, the objective function to be maximized is explored in an ascent direction. This can be achieved by ensuring that the negative Hessian matrix of the objective function is positive definite (or equivalently that the Hessian is negative definite). Even then, an additional tuning of the step size can be necessary to avoid a deterioration of the function to be maximized after a given iteration of the algorithm. One should also keep in mind that in order to accelerate the computation of the posterior mode via Newton-Raphson, analytical forms for the gradient and Hessian of the conditional posterior of $\boldsymbol{\xi}$ (given the hyperparameters) should be available. Depending on the complexity of the model likelihood, important efforts need to be invested to deal with possibly cumbersome formulas.

### 6.2.2 Optimal smoothing

The next step consists in using the previously derived Laplace approximation to the conditional posterior of the vector of spline and regression coefficients to identify the region in $p(\boldsymbol{\lambda}|\mathcal{D})$ where most of the probability mass is concentrated. A variety of techniques exist to explore the approximated posterior $\widetilde{p}(\boldsymbol{\lambda}|\mathcal{D})$ and the researcher has a lot of freedom to achieve this. The classic approach taken in this thesis when $\dim(\boldsymbol{\lambda}) > 1$ is to start with the computation of the marginal posterior mode of the penalty vector, using a Newton-Raphson algorithm relying on analytical forms for the gradient and Hessian of $\log \widetilde{p}(\boldsymbol{\lambda}|\mathcal{D})$. These analytical derivations are probably the highest price to pay to set up the LPS strategy, but it is an essential element to guarantee a fast selection of the penalty parameters tuning the smoothness of the functionals modeled with B-splines.

In more simple models where $\dim(\boldsymbol{\lambda}) = 1$, the exploration of the penalty posterior is easier to achieve since it takes place in a one-dimensional space. For instance, one could implement a root finding algorithm to find the mode of $\log \widetilde{p}(\lambda|\mathcal{D})$ by computing its first derivative to approach

numerically the point where it becomes zero. Alternatively, an informal way of tackling the problem is to make a visual inspection of the graph of $\widetilde{p}(\lambda|\mathcal{D})$ and place (equidistant) grid points in a close neighborhood of the peak of the posterior.

It is also worth mentioning that, regardless of the dimension of the penalty vector, setting $\boldsymbol{\lambda}$ at its marginal posterior mode as if it were a non stochastic quantity, usually suffices to guarantee reliable inference, as suggested by the simulation results in Chapter 4. The researcher wishing even more accuracy can rely on a grid-based strategy for exploring $\widetilde{p}(\boldsymbol{\lambda}|\mathcal{D})$ or even MCMC samplers. Finally, as the penalty parameter(s) are positive, it is usually (numerically) advisable to work with log transformed penalties.

### 6.2.3 Final approximation to the marginal posterior of $\boldsymbol{\xi}$

The final step consists in using the quadrature points selected during the exploration of $\widetilde{p}(\boldsymbol{\lambda}|\mathcal{D})$ to approximate the marginal posterior of the regression and spline parameter vector $\boldsymbol{\xi}$. Typically, the latter posterior is approximated using a mixture of (multivariate) Gaussian distributions, for which the mean and variance-covariance matrix are analytically available. Alternatively, when the uncertainty in the estimation of $\boldsymbol{\lambda}$ is ignored and the penalty vector is fixed at its posterior mode, the final approximation to the marginal posterior of $\boldsymbol{\xi}$ is simply a multivariate Gaussian. From there, point estimates and credible intervals can be readily constructed. The five steps below summarize the recipe to use LPS for inference in a generic model.

**Recipe for inference with LPS**

Denote by $\boldsymbol{\xi}$ the vector containing the regression and spline parameters. Let $\boldsymbol{\eta} = (\boldsymbol{\lambda}^\top, \boldsymbol{\delta}^\top)^\top$, where $\boldsymbol{\lambda}$ denotes the vector of penalty parameters in the P-spline model and $\boldsymbol{\delta}$ is a set of hyperparameters.

I. Specify the (Gaussian) conditional prior of $\boldsymbol{\xi}$ given $\boldsymbol{\eta}$, as well as the prior of the penalty vector $\boldsymbol{\lambda}$ given $\boldsymbol{\delta}$.

II. Compute the Laplace approximation to the conditional posterior of the latent vector $\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})$ (cf. Section 6.2.1) with posterior mode $\widehat{\boldsymbol{\xi}}(\boldsymbol{\eta})$.

III. Use the Laplace approximation in (II) to approximate the posterior of hyperparameters $\widetilde{p}(\boldsymbol{\eta}|\mathcal{D}) = \Big(p(\boldsymbol{\xi}, \boldsymbol{\eta}|\mathcal{D})/\widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\eta}, \mathcal{D})\Big)\Big|_{\boldsymbol{\xi}=\widehat{\boldsymbol{\xi}}(\boldsymbol{\eta})}$. If possible, integrate out nuisance parameters to obtain the approximated posterior $\widetilde{p}(\boldsymbol{\lambda}|\mathcal{D})$ for the penalty vector (cf. Section 6.2.2).

IV. Explore $\widetilde{p}(\boldsymbol{\eta}|\mathcal{D})$ through grid-based approaches or MCMC samplers to obtain quadrature points $\{\boldsymbol{\eta}^{(m)}\}$. Alternatively, set $\boldsymbol{\eta}$ equal to the mode $\widehat{\boldsymbol{\eta}}$.

V. Use the quadrature points in (IV) to approximate the marginal posterior of the vector of spline and regression coefficients $\widetilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \sum_m \widetilde{p}_G(\boldsymbol{\xi}|\boldsymbol{\eta}^{(m)}, \mathcal{D})\ p(\boldsymbol{\eta}^{(m)}|\mathcal{D})\ \Delta_m$. Alternatively, the mode $\widehat{\boldsymbol{\eta}}$ can be used to obtain a Gaussian approximation to the posterior $\widetilde{p}(\boldsymbol{\xi}|\mathcal{D}) = \widetilde{p}_G(\boldsymbol{\xi}|\widehat{\boldsymbol{\eta}}, \mathcal{D})$.

## 6.3 Merits and limitations of Laplace-P-splines

### 6.3.1 Strengths and weaknesses of LPS

Although the simulation results of the previous chapters convey a clear message, namely that LPS exhibits excellent frequentist properties for the considered Bayesian estimators and that LPSMAP is almost as performant as LPS in terms of estimation accuracy despite ignoring the uncertainty in the selection of the penalty, it is a good exercise to highlight the positive and negative facets of LPS. This is done in Table 6.1 where the strengths and weaknesses of the methodology are identified.

| **Strengths** |
|---|
| + LPS is natively built to approximate the joint marginal posterior of the vector of spline and regression parameters. |
| + LPS is not a black box and all the steps leading to the final approximated posterior latent vector are clearly explained. |

| | |
|---|---|
| + | Availability of the analytical gradient/Hessian permits fast and efficient exploration of the posterior penalty vector. |
| + | LPS is fully Bayesian. If required, the methodology can be extended to inject additional prior information. |
| + | The latent field dimension is independent of the sample size. |
| + | Construction of (approximate) pointwise credible intervals is relatively straightforward, even for complex functions of regression and spline parameters. |
| + | LPSMAP is much faster than LPS with more or less the same estimation accuracy. |
| + | LPS(MAP) is generally much faster than existing Bayesian methods fully relying on MCMC. |
| | **Weaknesses** |
| – | LPS currently focuses on a single smoother (P-splines). |
| – | The requirement to obtain explicit expressions for the gradient and Hessian of $\log p(\boldsymbol{\xi}\|\boldsymbol{\eta}, \mathcal{D})$ for a given model class. |
| – | The deterioration of the frequentist properties of the parameters and functional estimates when information is very sparse. |

Table 6.1: Strengths and weaknesses of LPS(MAP).

To complete the merits and limits of LPS, a summary of the arguments relating LPS to other competitors considered in this thesis is provided in the next sections.

### 6.3.2 LPS vs INLA

*Arguments in favor of LPS*

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 of the latent vector. From there, pointwise and set estimators for functions of the latent vector can be relatively easily constructed. Furthermore, as smooth terms in (generalized) additive models are exclusively modeled with P-splines, full-fledged analytical formulas are available for the gradient and Hessian of the posterior penalty vector, whereas INLA relies on numerical differentiation techniques.

Another fundamental difference lies in the specification of the latent vector: INLA works with a latent field having a dimension proportional to the sample size $n$, while in LPS it is independent of $n$. Finally, an important argument in favor of LPS is that the methodological outline employed is relatively simple and reflected in the organized structure underlying the routines of the **blapsr** package. The structure of the INLA package is less intuitive in that regard and is reported by many users as a black box.

*Arguments in favor of INLA*

The main advantage of INLA over LPS is its generality. In fact, INLA can be used for Bayesian regression in various models (see e.g. Gómez-Rubio, 2020) such as (spatio-)temporal models, mixed-effects models, multilevel models and more. In addition, several books have already been written on the topic and many illustrative datasets and applications of the **R-INLA** package are available. Finally, INLA offers more options for the selection of smoothers and is very efficient from a computational view point as it naturally takes advantage of parallelization possibilities offered by modern multi-core processors (Mantovan and Secchi, 2010).

### 6.3.3 LPS vs BayesX

*Argument in favor of LPS*

When comparing LPS with a fully Bayesian competitor such as BayesX, the main argument in favor of LPS(MAP) is its computational speed (cf. simulation results in Chapter 4). Indeed, as BayesX (and its **R** interface **R2BayesX**) is based on MCMC simulation techniques to fit models in the wide class of structured additive regression models, the cost of drawing samples from the target posterior distribution (of the vector of spline and regression parameters) often outweighs the cost of approximating the posterior latent vector with iterated Laplace approximations.

*Argument in favor of BayesX*

As LPS is a methodology developed for the class of latent Gaussian models, the Gaussian prior imposed on the latent variables translates in a "near-Gaussian" and (often) symmetric posterior for the vector of regression and spline parameters as the likelihood will usually mildly affect the bell-shaped prior. In that direction, MCMC based methods may be preferable as they are able to reconstruct posterior targets with a stronger degree of asymmetry. This would be particularly interesting for non-penalized regression parameters, as strongly asymmetric posteriors are sometimes observed.

### 6.3.4 LPS vs MGCV

*Argument in favor of LPS*

As LPS is a fully Bayesian approach, the joint posterior of the penalty vector can be characterized and the uncertainty in the selection of the penalty vector can be taken into account. This is not the case for MGCV based methods as they rely on an empirical Bayes approach that selects a single value for the penalty vector (usually the posterior mode) and, hence, ignores the uncertainty surrounding the penalty parameter selection.

*Argument in favor of MGCV*

The **mgcv** package developed by Simon Wood is already very mature and well documented. It also gathers robust and extremely fast routines

for inference that are far more general than the routines provided in the **blapsr** package which is still in its infancy.

## 6.4 Additional (simple) examples with blapsr

### 6.4.1 Density estimation

The `gamlps()` routine of the **blapsr** package is used here[1] for smoothing a histogram of the Old Faithful Geyser Data (see e.g. Weisberg, 1980; Silverman, 1986) obtained from the **datasets** package in **R**. The dataset consists of $n = 272$ observations of eruption times (in minutes). The following lines of code yield a smoothed version of the histogram as show in Figure 6.1.

```
R> library("blapsr")
R> data("faithful")
R> erupt <- faithful$eruptions
R> xl <- 1.3
R> xr <- 5.5
R> brk <- seq(xl, xr, by = 0.05)
R> hst <- hist(erupt, breaks = brk, plot = T,
+ col = "lightgrey", xlab = "Eruption time (in minutes)",
+ main = "", freq = F)
R> x <- hst$mids
R> y <- hst$counts
R> h <- x[2] - x[1]
R> fit <- gamlps(y ~ 1 + sm(x), K = 30, penorder = 3,
+ family = "poisson")
R> mu <- fit$fitted.values
R> lines(x, mu * h, col = "red", lwd = 2)
```

### 6.4.2 Scatterplot smoothing

In the context of scatterplot smoothing, the motorcycle data analyzed in Silverman (1985) can be used as a simple example to illustrate the use of the `amlps()` routine of the **blapsr** package. The data comes from the **MASS** package of Venables and Ripley (2002) with a sample of size $n = 133$ that consists in measurements of head acceleration in units of

[1] Special thanks to Paul Eilers for providing this smoothing histogram example.

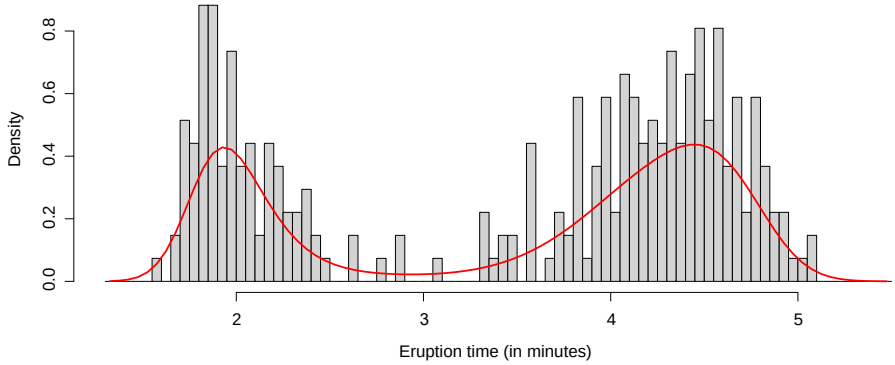



Figure 6.1: Smoothed version of the histogram for the Old Faithful Geyser Data with `gamlps()`.

gravity (g) at different times in milliseconds (ms) after impact to test crash helmets. The code below provides a smoothed version of the scatterplot with 20 cubic B-splines and a second order penalty. Figure 6.2 gives a graphical representation of the motorcycle data with the smooth fitted curve.

```
R> library("blapsr")
R> library("MASS")
R> fit <- amlps(accel ~ sm(times), data = mcycle, K = 20,
+ penorder = 2, cred.int = 0.95)
R> xgrid <- seq(min(mcycle$times), max(mcycle$times),
+ length = 200)
R> smoo.fit <- plot(fit, xp = xgrid, smoo.index = 1,
+ show.info = FALSE, show.plot = FALSE)
R> plot(mcycle, ylim = c(-150, 100),
+ ylab = "Acceleration (g)", xlab = "Time (ms)")
R> lines(smoo.fit$xp, fit$linear.coeff[1] + smoo.fit$sm.xp,
+ lwd = 2)
R> abline(h = 0)
```

### 6.4.3 Count data regression

We provide a further example based on the female horseshoe crabs data analyzed in Agresti (2013). The dataset can be downloaded from http://users.stat.ufl.edu/~aa/cda/cda.html and includes $n = 173$ observations on several crab characteristics.

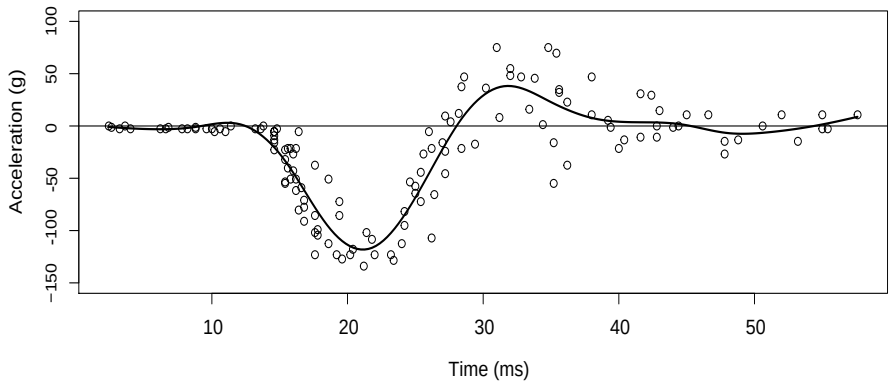



Figure 6.2: Smoothed version of the motorcycle data with `amlps()`.

The response variable of interest is the number of male crabs (called "satellites") that gather around females during spawning season to fertilize eggs. For simplicity, the only explanatory variable considered here is the carapace width (in cm) of the female crab.

We use the `gamlps()` routine to fit a Poisson model with a log link that specifies the log mean number of satellites as a smooth function of the female carapace width. The code given below loads the dataset and fits a Poisson model with 15 cubic B-splines and a third order penalty. Figure 6.3 (a) is a scatterplot of the number of satellites versus female carapace width. Figure 6.3 (b) shows the smoothing curve along with a set of points for 8 width categories for which the coordinate on the $x$-axis is the mean carapace width and the coordinate on the $y$-axis is the mean number of satellites (see Agresti, 2013, p. 124). These quantities are encoded in the vectors `mean.widths` and `mean.satellites` respectively. The upward trend reveals that the mean cluster size of satellites gathering around a female increase with the carapace width.

```
R> library("blapsr")
R> # Read dataset
R> crabs <- read.table("Crabs.dat", header = TRUE)
R> fit <- gamlps(sat ~ sm(width), data = crabs,
+ family = "poisson", K = 15, penorder = 3)
R> mean.widths <- c(22.69286, 23.84286, 24.77500, 25.83846,
+ 26.79091, 27.73750, 28.66667, 30.40714)
R> mean.satellites <- c(1.000000, 1.428571, 2.392857,
```

```
+ 2.692308, 2.863636, 3.875000, 3.944444, 5.142857)
R> xx <- seq(min(crabs$width), 32, length = 100)
R> yfit <- plot(fit, smoo.index = 1, xp = xx,
+ show.info = FALSE, show.plot = FALSE)
R> par(mfrow = c(1, 2))
R> plot(crabs$width, crabs$sat, type = "p", pch = 16,
+ xlab = "Carapace width (cm)",
+ ylab = "Number of satellites", main = "(a)")
R> plot(xx, exp(fit$linear.coeff[1] + yfit$sm.xp),
+ type = "l", col = "blue",
+ ylab = "Number of satellites",
+ xlab = "Carapace width (cm)", ylim = c(0, 5.5),
+ xlim = c(21, 32), main = "(b)")
R> lines(mean.widths, mean.satellites, type ="p", pch = 16)
R> legend("topleft", pch = 16,
+ "Mean for 8 carapace width categories", bty = "n")
```

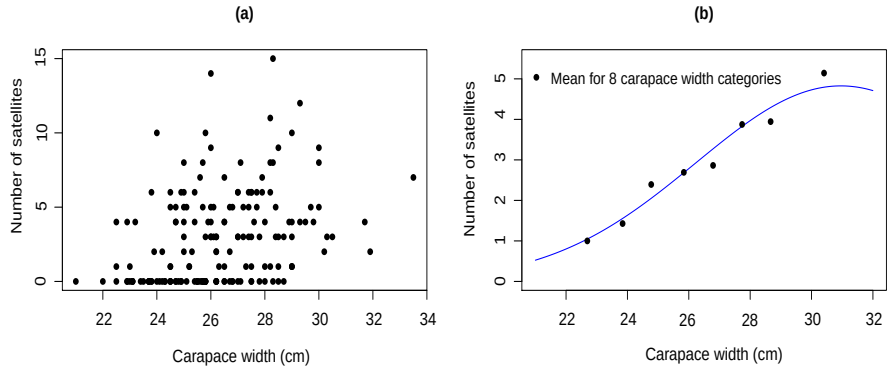



Figure 6.3: Poisson model with `gamlps()` to fit the crabs data.

In Table 6.2, we report the sample mean and variance of the number of satellites for the 8 carapace width categories considered in Agresti (2013). From this table, it is easy to see that the conditional variance of the response variable (conditional on a given category for the carapace width) exceeds the conditional mean. This phenomenon is known in the literature as overdispersion. As the Poisson model presented above assumes that the mean of the response is equal to its variance (equidispersion assumption), mathematically $E(Y) := \mu = V(Y)$, it may not be the best option to fit the crabs data. A better alternative would be

to extend LPS in the framework of a negative Binomial model where the variance of the response is specified as a quadratic function of the mean response $V(Y) = \mu + \delta\mu^2$, where $\delta > 0$ is usually interpreted as a dispersion parameter.

| Category | Mean | Variance |
|---|---|---|
| 1 | 1.00 | 2.77 |
| 2 | 1.43 | 8.88 |
| 3 | 2.39 | 6.54 |
| 4 | 2.69 | 11.38 |
| 5 | 2.86 | 6.89 |
| 6 | 3.88 | 8.81 |
| 7 | 3.94 | 16.88 |
| 8 | 5.14 | 8.29 |

Table 6.2: Sample mean and variance of the number of satellites for the 8 categories considered in Agresti (2013).

## 6.5 Final discussion and future research

The aim of this thesis is to bridge the gap between Laplace approximations and P-splines for fast Bayesian inference in survival models and (generalized) additive models. The proposed Laplace-P-spline methodology is a much faster alternative for inference in latent Gaussian models than existing MCMC methods that usually require more computational resources to sample from (often complex) posterior distributions. All the chapters are built from the ground up and are constructed around simulation scenarios and real data applications. It should be emphasized that in the different applications of LPS, the choice of the number of B-splines in the basis was completely arbitrary. We simply followed the philosophy of using a "large" number of B-splines with a penalty to counterbalance the flexibility. Whenever an unnecessary large number of B-splines was used to model a smooth term, the only purpose was to charge the model with more parameters than necessary to confirm that LPS was able to (numerically) cope with the situation. The same argument holds for the order of the penalty used in the text. Alternating (arbitrarily) between a second and third order penalty allowed us to monitor the LPS behavior under different penalty structures. The material presented in this thesis is a good starting point to understand in

detail the LPS methodology for further use with other classes of models. A non-exhaustive presentation of the possible extensions and applications of this methodology is proposed in the following sub-sections.

### 6.5.1 Reaching analytically tractable penalty posteriors

Undoubtedly, the most severe computational burden in the LPS estimation procedure comes from the exploration of the posterior penalty vector. The latter requires an iterative algorithm (e.g. Newton-Raphson) to find its posterior mode and eventually a grid (or MCMC sampler) to compute the marginal posterior distribution of the regression and spline parameters. Hence, the following question arises naturally: "Is it possible to somehow arrive at a posterior penalty distribution that is analytically tractable?". Asked differently, is it possible to bypass Newton methods and to obtain a reliable analytical approximation to the mode of the posterior penalty vector? If the answer is positive, then the estimation of a complex model with LPS would almost be possible in real time and results would be available in a few milliseconds. A tractable form for the posterior distribution of the penalty vector would also facilitate the construction of credible regions for these parameters and, hence, the selection of the necessary grid points to obtain $p(\boldsymbol{\xi}|\mathcal{D})$. In particular, this would be interesting for models with penalty vectors in high dimension.

### 6.5.2 Extending LPS to spatial models

Another direction completely ignored in this thesis is the implementation of LPS in spatial models. Datasets with spatial and geographical characteristics are available in many different fields, for instance in epidemiology where data are collected to understand the dynamics in the propagation of a given disease. Geo-referenced data is also of interest in meteorology, agriculture, ecology and demography among others. Extension of the LPS methodology in a spatial framework requires to extend the one-dimensional B-spline smoothers to higher-dimensional smoothers based on tensor products of B-splines with difference penalties imposed on neighboring coefficients of the tensor products (Eilers et al., 2006). The idea would be to start writing the LPS spatial model for data in two dimensions and derive the analytical formula for the posterior penalty vector that will be explored to determine the amount

of smoothness. Once this part is fully mastered, one can proceed with an extension of the LPS formulas in larger dimensions.

### 6.5.3 Refinements and extensions in survival analysis

The LPS promotion time cure model presented in Chapter 2 can be further refined by incorporating cluster-specific effects such as in frailty models (see e.g. Wienke, 2010). When data are clustered like in the oropharynx carcinoma dataset in Section 2.5.2, it may be important to include this information in the model to account for possible cluster effects (e.g. effect of a clinic) on the event time distribution for susceptible subjects and on cure probabilities. The extension of cure models is investigated in Gallardo et al. (2016), where two random effects are used for each cluster, one that explains the effect (of the cluster) on the survival time for susceptible subjects and the other that explains the effect on the cure fraction. The joint distribution of the random effects vector is taken to be a bivariate normal distribution.

To extend even further the LPS methodology in the family of cure regression models, one can consider an adaptation to the class of mixture cure models proposed by Boag (1949) and Berkson and Gage (1952). This model class directly specifies the population survival function as a mixture of two types of subjects, namely the cured and the susceptibles. A modeling possibility would be to specify the conditional survival function of the susceptibles via a Cox proportional hazards model where the baseline survival function is approximated with (cubic) B-splines.

There are many other research directions in which to extend LPS for time-to-event data. One could for instance consider an extension of the classic Cox model, to handle time-varying covariates. Introducing this dynamic flavor is important whenever a study involves subjects for which some covariate variables change over time (e.g. tumor size, blood pressure, glucose level, ...). This could be done within the framework of joint models for survival and endogenous longitudinal data (see for instance Ibrahim et al., 2010; Rizopoulos, 2012). Also, instead of considering the classic right censoring scheme, one could extend LPS to data with left censoring, interval censoring and even consider different truncation scenarios (see e.g. Lambert, 2020).

### 6.5.4 Improving blapsr

The **blapsr** package is currently in its early development phase and thus is far from being completely satisfactory. In fact, there is still a long way ahead to reach a versatile and general toolbox to use LPS for approximate Bayesian inference. First of all, it would be a good idea to rewrite the numerically demanding parts of the code using $\mathbf{C}_{++}$ or **Fortran**. Another track to improve the speed of the package procedures would be to parallelize the exploration of the penalty posterior in large dimensions and take advantage of the multi-core processors that are omnipresent in the computer market. Taken together, parallelized algorithms and subroutines rewritten in a more efficient language will yield a lightning fast package even if the underlying LPS framework is fully Bayesian.

Moreover, we may say a few words regarding the extension of the smoother considered in this thesis. In fact, at this stage of its development, the package only allows the user to work with cubic B-splines and a second or third order penalty. Giving the user the possibility to specify his own penalty matrix, his own B-spline basis with personalized knot positions, or even alternative function bases would be desirable. Finally, a further layer of sophistication can be brought to the routines underlying **blapsr** by allowing factor-by-curve interactions, i.e. the possibility to fit different smooth terms for each level of a categorical covariate.

# Appendix A (Chapter 1)

### A1. Marginal prior of the penalty parameter $\lambda$

Let us derive the marginal prior for $\lambda$ starting from the hierarchical priors $\lambda|\delta \sim \mathcal{G}(\nu/2, (\nu\delta)/2)$ and $\delta \sim \mathcal{G}(a_\delta, b_\delta)$:

$$\begin{aligned}
p(\lambda) &= \int_0^{+\infty} p(\lambda, \delta)\ d\delta \\
&= \int_0^{+\infty} p(\lambda|\delta)\ p(\delta)\ d\delta \\
&= \int_0^{+\infty} \frac{(\nu\delta)^{\frac{\nu}{2}}}{2^{\frac{\nu}{2}}} \frac{1}{\Gamma(\frac{\nu}{2})} \lambda^{\frac{\nu}{2}-1} \exp\left(-\delta\frac{\nu\lambda}{2}\right) \frac{b_\delta^{a_\delta}}{\Gamma(a_\delta)} \delta^{a_\delta-1} \exp(-\delta b_\delta)\ d\delta \\
&\propto \lambda^{\frac{\nu}{2}-1} \int_0^{+\infty} \delta^{\frac{\nu}{2}+a_\delta-1} \exp\left(-\delta\left(\frac{\nu\lambda}{2}+b_\delta\right)\right)\ d\delta \\
&\propto \lambda^{\frac{\nu}{2}-1} \left(\frac{\nu\lambda}{2}+b_\delta\right)^{-(\frac{\nu}{2}+a_\delta)} \\
&\propto \lambda^{\frac{\nu}{2}-1} \left(b_\delta\left(1+\frac{\nu\lambda}{2b_\delta}\right)\right)^{-(\frac{\nu}{2}+a_\delta)} \\
&\propto \lambda^{\frac{\nu}{2}-1} \left(1+\frac{\nu\lambda}{2b_\delta}\right)^{-(\frac{\nu}{2}+a_\delta)}.
\end{aligned}$$

Fixing $a_\delta = b_\delta = 0.5$ and $\nu = 1$, we thus have $p(\lambda) \propto (\sqrt{\lambda}\ (1+\lambda))^{-1}$. Recall that a random variable $X$ with a Beta-prime distribution $X \sim \text{BetaPrime}(a, b), a > 0,\ b > 0$ has probability density function (see e.g. Dos Passos, 2009, p.330):

$$p(x) = \begin{cases} \frac{x^{a-1}(1+x)^{-a-b}}{B(a,b)} & \text{for } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where the denominator $B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\ dt$ is the beta function. Hence replacing $a = b = 0.5$ in the above density, we recover up to a multiplicative constant $p(x) \propto (\sqrt{x}\ (1+x))^{-1}$, and so we have $\lambda \sim \text{BetaPrime}(0.5, 0.5)$. In the latter case, the normalizing constant $c_{norm}$ can be obtained analytically by computing the inverse of the beta function at $a = b = 0.5$:

$$\begin{aligned} c_{norm} &= (B(0.5,0.5))^{-1} \\ &= \left(\int_0^1 \frac{1}{\sqrt{t(1-t)}}\ dt\right)^{-1} \\ &= \left(2\arcsin(\sqrt{t})\Big|_0^1\right)^{-1} \\ &= (2\arcsin(1) - 2\arcsin(0))^{-1} \\ &= \pi^{-1} \end{aligned}$$

Below, a graphical illustration is shown for $p(\lambda)$ under the two configurations used in this thesis namely $a_\delta = b_\delta = 0.5,\ \nu = 1$ and $a_\delta = b_\delta = 10^{-4},\ \nu = 3$.

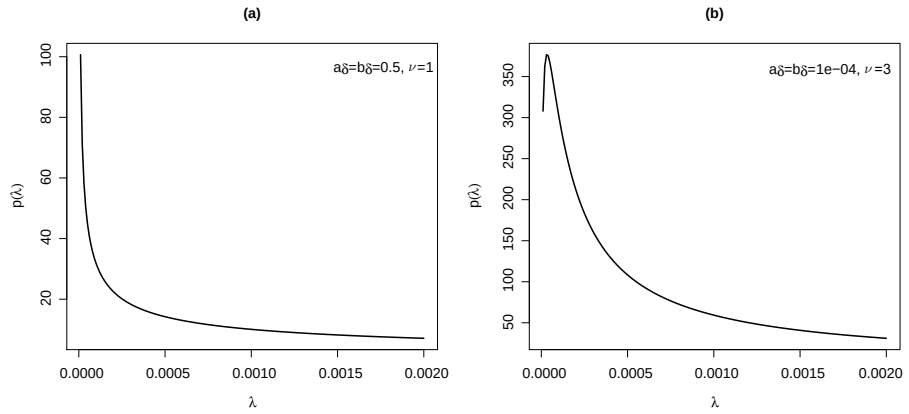



Figure A1: Marginal prior of the penalty parameter $\lambda$ under two parameterizations for the hyperparameters. (a) $a_\delta = b_\delta = 0.5,\ \nu = 1$ and (b) $a_\delta = b_\delta = 10^{-4},\ \nu = 3$.

## A2. Conditional posterior distributions of the Bayesian P-spline model

For the homoscedastic model in Section 1.3.2, the likelihood is:

$$\mathcal{L}(\boldsymbol{\theta}, \tau; \mathcal{D}) \propto \tau^{\frac{n}{2}} \exp\left(-0.5\tau \left\|\mathbf{y} - B\boldsymbol{\theta}\right\|^2\right).$$

The conditional posterior of the vector of B-spline amplitudes is:

$$\begin{aligned} p(\boldsymbol{\theta}|\lambda, \tau, \mathcal{D}) &\propto \mathcal{L}(\boldsymbol{\theta}, \tau; \mathcal{D})\ p(\boldsymbol{\theta}|\lambda, \tau) \\ &\propto \exp\left(-0.5\tau \left\|\mathbf{y} - B\boldsymbol{\theta}\right\|^2\right)\ \exp\left(-0.5\lambda\tau\boldsymbol{\theta}^\top P\boldsymbol{\theta}\right) \\ &\propto \exp\left(\tau\mathbf{y}^\top B\boldsymbol{\theta} - 0.5\tau\ \boldsymbol{\theta}^\top(B^\top B + \lambda P)\boldsymbol{\theta}\right). \end{aligned}$$

The above expression is the exponential of a quadratic form in $\boldsymbol{\theta}$ and is (up to a multiplicative constant) the density of a Gaussian with mean $\mu_{\boldsymbol{\theta}}$ and variance-covariance matrix $\Sigma_{\boldsymbol{\theta}}$ obtained as follows:

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\lambda, \tau, \mathcal{D}) &= 0 \\ \Leftrightarrow \tau B^\top \mathbf{y} - \tau(B^\top B + \lambda P)\boldsymbol{\theta} &= 0 \\ \Leftrightarrow (B^\top B + \lambda P)^{-1} B^\top \mathbf{y} &= \mu_{\boldsymbol{\theta}}. \end{aligned}$$

The covariance matrix corresponds to the inverse of the negative Hessian, i.e. $\Sigma_{\boldsymbol{\theta}} = \left(-\nabla^2_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta}|\lambda, \tau, \mathcal{D})\right)^{-1} = \tau^{-1}(B^\top B + \lambda P)^{-1}$ and hence finally, $(\boldsymbol{\theta}|\lambda, \tau, \mathcal{D}) \sim \mathcal{N}_{\dim(\boldsymbol{\theta})}\left((B^\top B + \lambda P)^{-1} B^\top \mathbf{y}, \tau^{-1}(B^\top B + \lambda P)^{-1}\right)$. Let us now focus on the conditional posterior of the precision:

$$\begin{aligned} p(\tau|\boldsymbol{\theta}, \lambda, \mathcal{D}) &\propto \mathcal{L}(\boldsymbol{\theta}, \tau; \mathcal{D})\ p(\boldsymbol{\theta}|\lambda, \tau)\ p(\tau) \\ &\propto \tau^{\frac{n+K}{2}-1} \exp\left(-0.5\tau \left(\left\|\mathbf{y} - B\boldsymbol{\theta}\right\|^2 + \lambda\boldsymbol{\theta}^\top P\boldsymbol{\theta}\right)\right), \\ (\tau|\boldsymbol{\theta}, \lambda, \mathcal{D}) &\sim \mathcal{G}\left(0.5(n+K), 0.5\left(\left\|\mathbf{y} - B\boldsymbol{\theta}\right\|^2 + \lambda\boldsymbol{\theta}^\top P\boldsymbol{\theta}\right)\right). \end{aligned}$$

The conditional posterior for the hyperparameter $\delta$ is:

$$\begin{aligned} p(\delta|\lambda, \mathcal{D}) &\propto p(\lambda|\delta)\ p(\delta) \\ &\propto \delta^{\frac{\nu}{2}+a_\delta-1} \exp\left(-\delta(0.5\nu\lambda + b_\delta)\right) \\ \text{so, } (\delta|\lambda, \mathcal{D}) &\sim \mathcal{G}\left(0.5\nu + a_\delta, 0.5\nu\lambda + b_\delta\right). \end{aligned}$$

Finally, he conditional posterior of the penalty parameter is given by:

$$\begin{aligned} p(\lambda|\boldsymbol{\theta},\tau,\delta,\mathcal{D}) &\propto p(\boldsymbol{\theta}|\lambda,\tau)\ p(\lambda|\delta) \\ &\propto \lambda^{\frac{K+\nu}{2}-1}\exp\left(-0.5\lambda(\tau\boldsymbol{\theta}^\top P\boldsymbol{\theta}+\nu\delta)\right) \\ \text{so, } (\lambda|\boldsymbol{\theta},\tau,\delta,\mathcal{D}) &\sim \mathcal{G}\left(0.5(K+\nu), 0.5(\tau\boldsymbol{\theta}^\top P\boldsymbol{\theta}+\nu\delta)\right). \end{aligned}$$

# Appendix B (Chapter 2)

## B1. Conditional mean

The vector $\boldsymbol{\xi}_c^*(\lambda) \in \mathbb{R}^{\dim(\boldsymbol{\xi})-1}$ is the conditional posterior mean of the Gaussian approximation for a given $\xi_K = c$ and should not be confused with $\boldsymbol{\xi}_{cc}^*(\lambda)$. To obtain $\boldsymbol{\xi}_c^*(\lambda)$, we compute the Gaussian approximation around the posterior mode of $p(\boldsymbol{\xi}|\lambda, \mathcal{D})$ as described in Section 2.3.3 and find a multivariate ($\dim(\boldsymbol{\xi})$-dimensional) Gaussian distribution with mean $\boldsymbol{\xi}^*(\lambda)$ and covariance matrix $\Sigma^*(\lambda)$.

Next, using classic properties of the Normal density, we derive the distribution of $\boldsymbol{\xi}_{-K} = (\xi_1, \ldots, \xi_{K-1}, \xi_{K+1}, \ldots, \xi_{\dim(\boldsymbol{\xi})}) \in \mathbb{R}^{\dim(\boldsymbol{\xi})-1}$ given the constraint $\xi_K = c$. The resulting distribution is Gaussian with mean vector $\boldsymbol{\xi}_c^*(\lambda) = \boldsymbol{\xi}_{-K}^*(\lambda) + \widetilde{\Sigma}_{2,1}(\lambda)\widetilde{\Sigma}_{1,1}^{-1}(\lambda)\big(c - \xi_K^*(\lambda)\big)$ and covariance matrix $\Sigma_c^*(\lambda) = \widetilde{\Sigma}_{2,2}(\lambda) - \widetilde{\Sigma}_{2,1}(\lambda)\widetilde{\Sigma}_{1,1}^{-1}(\lambda)\widetilde{\Sigma}_{1,2}(\lambda)$ with the following elements:

$$\begin{aligned}
\widetilde{\Sigma}_{1,1}(\lambda) &= \Sigma_{K,K}^*(\lambda), \\
\widetilde{\Sigma}_{1,2}(\lambda) &= (\widetilde{\Sigma}_{2,1}(\lambda))^\top \\
&= (\Sigma_{K,1}^*(\lambda), \ldots, \Sigma_{K,(K-1)}^*(\lambda), \Sigma_{K,(K+1)}^*(\lambda), \ldots, \Sigma_{K,\dim(\boldsymbol{\xi})}^*(\lambda))
\end{aligned}$$

and $\widetilde{\Sigma}_{22}(\lambda)$ is the matrix $\Sigma^*(\lambda)$ without row and column $K$. The vector $\boldsymbol{\xi}_{cc}^*(\lambda) \in \mathbb{R}^{\dim(\boldsymbol{\xi})}$ corresponds to $\boldsymbol{\xi}_c^*(\lambda)$ to which we add $\xi_K = c$ at position $K$, i.e. $\boldsymbol{\xi}_{cc}^*(\lambda) = (\xi_{c,1}^*(\lambda), \ldots, \xi_{c,K-1}^*(\lambda), c, \xi_{c,K}^*(\lambda), \ldots, \xi_{c,\dim(\boldsymbol{\xi})-1}^*(\lambda))$, where $\xi_{c,i}^*(\lambda)$ denotes the $i$th entry of $\boldsymbol{\xi}_c^*(\lambda)$.

## B2. Gradients for credible intervals

### Gradient associated to the baseline survival function

$$\nabla_{\boldsymbol{\theta}_c} G_0(\boldsymbol{\theta}^m_{c,0}|t) = \left(\sum_{j=1}^{j(t)} h_0(s_j)\Delta_j\right)^{-1} \times \begin{pmatrix} \sum_{j=1}^{j(t)} \exp\left(\sum_{k=1}^{K} \theta_k b_k(s_j)\right)\ b_1(s_j)\Delta_j \\ \vdots \\ \sum_{j=1}^{j(t)} \exp\left(\sum_{k=1}^{K} \theta_k b_k(s_j)\right)\ b_{K-1}(s_j)\Delta_j \end{pmatrix}_{\boldsymbol{\theta}_c=\boldsymbol{\theta}^m_{c,0}}.$$

### Gradient associated to the population survival function

$$\nabla_{\boldsymbol{\xi}_c} G_0(\boldsymbol{\xi}^m_{c,0}|\mathbf{x},\mathbf{z},t) = \begin{pmatrix} v(\boldsymbol{\theta},\boldsymbol{\gamma})^{-1}\exp(\mathbf{z}^\top\boldsymbol{\gamma})S_0(t)^{\exp(\mathbf{z}^\top\boldsymbol{\gamma})}\left(\sum_{j=1}^{j(t)} h_0(s_j)b_1(s_j)\Delta_j\right) \\ \vdots \\ v(\boldsymbol{\theta},\boldsymbol{\gamma})^{-1}\exp(\mathbf{z}^\top\boldsymbol{\gamma})S_0(t)^{\exp(\mathbf{z}^\top\boldsymbol{\gamma})}\left(\sum_{j=1}^{j(t)} h_0(s_j)b_{K-1}(s_j)\Delta_j\right) \\ 1 \\ x_1 \\ \vdots \\ x_p \\ v(\boldsymbol{\theta},\boldsymbol{\gamma})^{-1}\exp(\mathbf{z}^\top\boldsymbol{\gamma})S_0(t)^{\exp(\mathbf{z}^\top\boldsymbol{\gamma})}\left(\sum_{j=1}^{j(t)} h_0(s_j)\Delta_j\right)z_1 \\ \vdots \\ v(\boldsymbol{\theta},\boldsymbol{\gamma})^{-1}\exp(\mathbf{z}^\top\boldsymbol{\gamma})S_0(t)^{\exp(\mathbf{z}^\top\boldsymbol{\gamma})}\left(\sum_{j=1}^{j(t)} h_0(s_j)\Delta_j\right)z_l \end{pmatrix}_{\boldsymbol{\xi}_c=\boldsymbol{\xi}^m_{c,0}},$$

with $v(\boldsymbol{\theta},\boldsymbol{\gamma}) = 1 - \exp\Big(-\sum_{j=1}^{j(t)} \exp\left(\sum_{k=1}^{K} \theta_k b_k(s_j)\right)\Delta_j\Big)^{\exp\left(\mathbf{z}^\top\boldsymbol{\gamma}\right)}$.

**Gradient associated to the conditional probability** $P(T = +\infty | T \geq t, \mathbf{x}, \mathbf{z})$

$$\nabla_{\boldsymbol{\xi}_c} G_0(\boldsymbol{\xi}_{c,0}^m | \mathbf{x}, \mathbf{z}, t) = \begin{pmatrix} -\exp(\mathbf{z}^\top \boldsymbol{\gamma}) \sum_{j=1}^{j(t)} h_0(s_j) b_1(s_j) \Delta_j \\ \vdots \\ -\exp(\mathbf{z}^\top \boldsymbol{\gamma}) \sum_{j=1}^{j(t)} h_0(s_j) b_{K-1}(s_j) \Delta_j \\ 1 \\ x_1 \\ \vdots \\ x_p \\ -z_1 \exp(\mathbf{z}^\top \boldsymbol{\gamma}) \sum_{j=1}^{j(t)} h_0(s_j) \Delta_j \\ \vdots \\ -z_l \exp(\mathbf{z}^\top \boldsymbol{\gamma}) \sum_{j=1}^{j(t)} h_0(s_j) \Delta_j \end{pmatrix}_{\boldsymbol{\xi}_c = \boldsymbol{\xi}_{c,0}^m} .$$

# Appendix C (Chapter 3)

### C1. Efficient evaluation of $\phi(\boldsymbol{\lambda})$

The scalar function $\phi(\boldsymbol{\lambda}) := \frac{1}{2}\,\mathbf{y}^\top \big(I_n - B(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}B^\top\big)\mathbf{y}$ has to be frequently evaluated in our Laplace-P-spline approach as it is present in the gradient and Hessian of $\log p(\mathbf{v}|\mathcal{D})$. To efficiently evaluate $\phi(\cdot)$ one can first take the trace :

$$\mathrm{Tr}(\phi(\boldsymbol{\lambda})) = \frac{1}{2}\left(\mathrm{Tr}(\mathbf{y}^\top\mathbf{y}) - \mathrm{Tr}\left(\mathbf{y}^\top B(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}B^\top\mathbf{y}\right)\right).$$

Since the trace is invariant under cyclic permutations, one has:

$$\mathrm{Tr}(\phi(\boldsymbol{\lambda})) = \frac{1}{2}\left(\mathrm{Tr}(\mathbf{y}^\top\mathbf{y}) - \mathrm{Tr}\left(B^\top\mathbf{y}\mathbf{y}^\top B(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\right)\right).$$

Furthermore, $B^\top\mathbf{y}\mathbf{y}^\top B$ is a symmetric matrix, so:

$$\mathrm{Tr}(\phi(\boldsymbol{\lambda})) = \frac{1}{2}\left(\mathrm{Tr}(\mathbf{y}^\top\mathbf{y}) - \mathrm{Tr}\left(\left(B^\top\mathbf{y}\mathbf{y}^\top B\right)^\top (B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\right)\right).$$

Using the Hadamard product $\circ$, the trace becomes:

$$\mathrm{Tr}(\phi(\boldsymbol{\lambda})) = \frac{1}{2}\left(\mathrm{Tr}(\mathbf{y}^\top\mathbf{y}) - \sum_{i,j}\left(\left(B^\top\mathbf{y}\mathbf{y}^\top B\right)\circ(B^\top B + Q_{\boldsymbol{\xi}}^{\boldsymbol{\lambda}})^{-1}\right)_{i,j}\right),$$

where $\Sigma_{i,j}$ is the sum over the entries of a matrix. Computing with the Hadamard product is much faster than taking the trace of matrix products.

## C2. Fast evaluation of empirical moments

The skew-normal match to the $j$th conditional $p(v_j|\widehat{\mathbf{v}}_{-j}, \mathcal{D})$ requires to compute the empirical moments of the latter expression using an equidistant grid. For each element of the grid $v_{jl}$, evaluating the conditional posterior involves the computation of a (potentially large) matrix inverse $\big(B^\top B + Q_{\boldsymbol{\xi}}^{\widetilde{\mathbf{v}}_j}\big)^{-1}$ through the scalar function $\phi(v_{jl}|\widehat{\mathbf{v}}_{-j}) = \frac{1}{2}\mathbf{y}^\top\Big(I_n - B\big(B^\top B + Q_{\boldsymbol{\xi}}^{\widetilde{\mathbf{v}}_j}\big)^{-1}B^\top\Big)\mathbf{y}$, where $\widetilde{\mathbf{v}}_j$ is a $q$-dimensionsal vector with all entries fixed at $\widehat{\mathbf{v}}$, except entry $j$ which is $v_{jl}$, that is $\widetilde{\mathbf{v}}_j = (\widehat{v}_1, \ldots, \widehat{v}_{j-1}, v_{jl}, \widehat{v}_{j+1}, \ldots, \widehat{v}_q)$. To circumvent the matrix inverse, let us write $Q_{\boldsymbol{\xi}}^{\widetilde{\mathbf{v}}_j}$ as follows:

$$
\begin{aligned}
Q_{\boldsymbol{\xi}}^{\widetilde{\mathbf{v}}_j} &= \begin{pmatrix} \zeta I_{p+1} & 0_{p+1,q\times(K-1)} \\ 0_{q\times(K-1),p+1} & \operatorname{diag}\big(\exp(\widehat{v}_1), \ldots, \exp(v_{jl}), \ldots, \exp(\widehat{v}_q)\big) \otimes P \end{pmatrix} \\
&= \begin{pmatrix} \zeta I_{p+1} & 0_{p+1,q\times(K-1)} \\ 0_{q\times(K-1),p+1} & \operatorname{diag}\big(\exp(\widehat{v}_1), \ldots, 0, \ldots, \exp(\widehat{v}_q)\big) \otimes P \end{pmatrix} \\
&\quad + \exp(v_{jl}) \begin{pmatrix} 0_{p+1,p+1} & 0_{p+1,q\times(K-1)} \\ 0_{q\times(K-1),p+1} & \operatorname{diag}\big(0, \ldots, 1, \ldots, 0\big) \otimes P \end{pmatrix}.
\end{aligned}
$$

Also, define the $\dim(\boldsymbol{\xi}) \times \dim(\boldsymbol{\xi})$ matrices:

$$
\begin{aligned}
\widetilde{Q}_{\boldsymbol{\xi}}^{-j} &:= \begin{pmatrix} \zeta I_{p+1} & 0_{p+1,q\times(K-1)} \\ 0_{q\times(K-1),p+1} & \operatorname{diag}\big(\exp(\widehat{v}_1), \ldots, 0, \ldots, \exp(\widehat{v}_q)\big) \otimes P \end{pmatrix}, \\
\widetilde{P}_j &:= \begin{pmatrix} 0_{p+1,p+1} & 0_{p+1,q\times(K-1)} \\ 0_{q\times(K-1),p+1} & \operatorname{diag}\big(0, \ldots, 1, \ldots, 0\big) \otimes P \end{pmatrix}.
\end{aligned}
$$

Since $\widetilde{P}_j$ is of rank $(K-1)$, we perturb the main diagonal by $\epsilon = 10^{-10}$ and define the full rank symmetric matrix $\breve{P}_j := \widetilde{P}_j + \epsilon I_{\dim(\boldsymbol{\xi})}$ so that $Q_{\boldsymbol{\xi}}^{\widetilde{\mathbf{v}}_j} \approx \widetilde{Q}_{\boldsymbol{\xi}}^{-j} + \exp(v_{jl})\breve{P}_j$ and hence:

$$
\begin{aligned}
\big(B^\top B + Q_{\boldsymbol{\xi}}^{\widetilde{\mathbf{v}}_j}\big)^{-1} &\approx \big(B^\top B + \widetilde{Q}_{\boldsymbol{\xi}}^{-j} + \exp(v_{jl})\breve{P}_j\big)^{-1} \\
&\approx \big(\mathbb{B}_j + \exp(v_{jl})\breve{P}_j\big)^{-1}, \qquad \big(\text{with } \mathbb{B}_j = B^\top B + \widetilde{Q}_{\boldsymbol{\xi}}^{-j}\big) \\
&\approx \big(\mathbb{B}_j + \exp(v_{jl})V_j\Lambda_j V_j^\top\big)^{-1},
\end{aligned}
$$

where $V_j \Lambda_j V_j^\top$ is the spectral decomposition of $\breve{P}_j$ with $V_j$ the orthogonal matrix of eigenvectors satisfying $V_j V_j^\top = I_{\dim(\boldsymbol{\xi})}$ and $V_j^{-1} = V_j^\top$. Matrix $\Lambda_j$ is a diagonal matrix with the eigenvalues of $\breve{P}_j$ on the main diagonal. One can further decompose:

$$\begin{aligned}
\left(B^\top B + Q_{\boldsymbol{\xi}}^{\tilde{\mathbf{v}}_j}\right)^{-1} &\approx \left(\mathbb{B}_j + \exp(v_{jl}) V_j \Lambda_j V_j^\top\right)^{-1} \\
&\approx \left(V_j V_j^\top \mathbb{B}_j V_j V_j^\top + V_j \exp(v_{jl}) \Lambda_j V_j^\top\right)^{-1} \\
&\approx \left(V_j \left(V_j^\top \mathbb{B}_j V_j + \exp(v_{jl}) \Lambda_j\right) V_j^\top\right)^{-1} \\
&\approx V_j \left(V_j^\top \mathbb{B}_j V_j + \exp(v_{jl}) \Lambda_j\right)^{-1} V_j^\top \\
&\approx V_j \left(V_j^\top \mathbb{B}_j V_j + \exp(v_{jl}) \Lambda_j^{\frac{1}{2}} \Lambda_j^{\frac{1}{2}}\right)^{-1} V_j^\top \\
&\approx V_j \left(\Lambda_j^{\frac{1}{2}} \left(\Lambda_j^{-\frac{1}{2}} V_j^\top \mathbb{B}_j V_j \Lambda_j^{-\frac{1}{2}} + \exp(v_{jl}) I_{\dim(\boldsymbol{\xi})}\right) \Lambda_j^{\frac{1}{2}}\right)^{-1} V_j^\top \\
&\approx V_j \Lambda_j^{-\frac{1}{2}} \left(\Lambda_j^{-\frac{1}{2}} V_j^\top \mathbb{B}_j V_j \Lambda_j^{-\frac{1}{2}} + \exp(v_{jl}) I_{\dim(\boldsymbol{\xi})}\right)^{-1} \Lambda_j^{-\frac{1}{2}} V_j^\top.
\end{aligned}$$

Using the spectral decomposition $\Lambda_j^{-\frac{1}{2}} V_j^\top \mathbb{B}_j V_j \Lambda_j^{-\frac{1}{2}} = U_j D_j U_j^\top$, where $U_j$ is an orthogonal matrix and $D_j$ a symmetric matrix with eigenvalues on the main diagonal, one has:

$$\begin{aligned}
\left(B^\top B + Q_{\boldsymbol{\xi}}^{\tilde{\mathbf{v}}_j}\right)^{-1} &\approx V_j \Lambda_j^{-\frac{1}{2}} \left(U_j D_j U_j^\top + \exp(v_{jl}) I_{\dim(\boldsymbol{\xi})}\right)^{-1} \Lambda_j^{-\frac{1}{2}} V_j^\top \\
&\approx V_j \Lambda_j^{-\frac{1}{2}} \left(U_j D_j U_j^\top + U_j \exp(v_{jl}) U_j^\top\right)^{-1} \Lambda_j^{-\frac{1}{2}} V_j^\top \\
&\approx V_j \Lambda_j^{-\frac{1}{2}} \left(U_j \left(D_j + \exp(v_{jl}) I_{\dim(\boldsymbol{\xi})}\right) U_j^\top\right)^{-1} \Lambda_j^{-\frac{1}{2}} V_j^\top \\
&\approx V_j \Lambda_j^{-\frac{1}{2}} U_j \left(D_j + \exp(v_{jl}) I_{\dim(\boldsymbol{\xi})}\right)^{-1} U_j^\top \Lambda_j^{-\frac{1}{2}} V_j^\top,
\end{aligned}$$

where $\left(D_j + \exp(v_{jl}) I_{\dim(\boldsymbol{\xi})}\right)^{-1}$ is the inverse of a diagonal matrix and is equal to a diagonal matrix with the inverse of the diagonal entries:

$$\begin{aligned}
\mathbb{D}_j &:= \left(D_j + \exp(v_{jl}) I_{\dim(\boldsymbol{\xi})}\right)^{-1} \\
&= \operatorname{diag}\Big(\left(D_{j11} + \exp(v_{jl})\right)^{-1}, \ldots, \left(D_{j\dim(\boldsymbol{\xi})\dim(\boldsymbol{\xi})} + \exp(v_{jl})\right)^{-1}\Big).
\end{aligned}$$

Finally, the inverse is approximated as:

$$\left(B^\top B + Q_{\boldsymbol{\xi}}^{\widetilde{\mathbf{v}}_j}\right)^{-1} \approx V_j \Lambda_j^{-\frac{1}{2}} U_j \mathbb{D}_j U_j^\top \Lambda_j^{-\frac{1}{2}} V_j^\top.$$

From a computational perspective, using the above approximating formula is much faster than computing the direct inverse. In fact, one can compute the matrices $V_j, \Lambda_j, U_j$ and $D_j$ once and for all across dimensions $j = 1, \ldots, q$ and then simply evaluate the above approximate version of the inverse on the chosen grid points $\{v_{jl}\}_{l=1}^{L}$ to compute the desired empirical moments.

# Appendix D (Chapter 4)

## D1. One-parameter exponential family distributions

### Poisson distribution

Let $Y \sim \text{Poisson}(\mu)$ with probability mass function:

$$\begin{aligned} p(y) &= \frac{\mu^y \exp(-\mu)}{y!},\ y \in \mathbb{N} \\ \Leftrightarrow \exp\left(\log(p(y))\right) &= \exp\left(y\log(\mu) - \mu - \log(y!)\right). \end{aligned}$$

Let $\gamma = \log(\mu)$, $s(\gamma) = \exp(\gamma)$, $\varkappa = 1$ and $c(y, \varkappa) = -\log(y!)$. The above equation becomes:

$$p(y) = \exp\left(\frac{y\gamma - s(\gamma)}{\varkappa} + c(y, \varkappa)\right),$$

such that the Poisson distribution belongs to the one-parameter exponential family and has the following associated functions $s'(\gamma) = s''(\gamma) = \exp(\gamma)$, the canonical link is $g(\mu) = \log(\mu)$ with $g'(\mu) = (s''(\gamma))^{-1} = 1/\exp(\gamma)$ and weight $w = \left(\text{Var}(y)(g'(\mu))^2\right)^{-1} = s''(\gamma) = \exp(\gamma)$.

### Gaussian distribution

Let us now consider the random variate $Y \sim \mathcal{N}(\mu, \sigma^2)$ with known variance $\sigma^2$ and probability density function:

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)$$

$$\begin{aligned}
\Leftrightarrow \exp\left(\log(p(y))\right) &= \exp\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right) \\
&= \exp\left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} + \frac{y\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{y\mu - \left(\frac{\mu^2}{2}\right)}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}\right).
\end{aligned}$$

Let $\gamma = \mu$, $s(\gamma) = \gamma^2/2$, $\varkappa = \sigma^2$ and $c(y, \varkappa) = -(1/2)\log(2\pi\sigma^2) - y^2/(2\sigma^2)$. The above equation becomes:

$$p(y) = \exp\left(\frac{y\gamma - s(\gamma)}{\varkappa} + c(y, \varkappa)\right),$$

such that the Normal distribution with known variance belongs to the one-parameter exponential family and has the following associated functions $s'(\gamma) = \gamma$, $s''(\gamma) = 1$. The canonical link is the identity link, $g(\mu) = \mu$ with $g'(\mu) = 1$ and weight $w = \left(\mathrm{Var}(y)(g'(\mu))^2\right)^{-1} = 1/\varkappa$.

### Binomial distribution

For the Binomial distribution $Y \sim \mathrm{Bin}(m, p), m \in \mathbb{N}_+,\ p \in (0, 1)$, with $y \in \{0, 1 \ldots, m\}$, we have the following probability mass function:

$$\begin{aligned}
p(y) &= \binom{m}{y} p^y (1-p)^{(m-y)} \\
\Leftrightarrow \exp\left(\log(p(y))\right) &= \exp\Bigg(y\log(p) + (m-y)\log(1-p) + \\
&\qquad \log\left(\frac{m!}{y!(m-y)!}\right)\Bigg) \\
&= \exp\Bigg(y\log(p) - y\log(1-p) + m\log(1-p) + \\
&\qquad \log\left(\frac{m!}{y!(m-y)!}\right)\Bigg)
\end{aligned}$$

$$= \exp\left(y \log\left(\frac{p}{1-p}\right) + m \log(1-p) + \log\left(\frac{m!}{y!(m-y)!}\right)\right).$$

Let $\gamma = \log(p/(1-p)) = \text{logit}(p)$, $s(\gamma) = m \log(1 + \exp(\gamma))$, $\varkappa = 1$ and $c(y, \varkappa) = \log(m!/(y!(m-y)!))$. The above equation becomes:

$$p(y) = \exp\left(\frac{y\gamma - s(\gamma)}{\varkappa} + c(y, \varkappa)\right),$$

so, the Binomial distribution belongs to the one-parameter exponential family with $s'(\gamma) = (m \exp(\gamma))/(1 + \exp(\gamma))$, $s''(\gamma) = (m \exp(\gamma))/(1 + \exp(\gamma))^2$. The mean is related to $p$ as follows $\mu = mp$ or $p = \mu/m$, such that with a canonical link $g(\mu) = \gamma = \log(p/(1-p)) = \log(\mu/(m-\mu))$ and weight $w = \left(\text{Var}(y)(g'(\mu))^2\right)^{-1} = \mu(m-\mu)/m$.

### Bernoulli distribution

Let $Y \sim \text{Bern}(p)$ be a Bernoulli random variable with $p \in (0, 1)$, $y \in \{0, 1\}$ and probability mass function:

$$\begin{aligned} p(y) &= p^y (1-p)^{1-y} \\ \Leftrightarrow \exp(\log(p(y))) &= \exp\left(y \log(p) + (1-y) \log(1-p)\right) \\ &= \exp\left(y \log(p) - y \log(1-p) + \log(1-p)\right) \\ &= \exp\left(y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right). \end{aligned}$$

Let $\gamma = \log(p/(1-p)) = \text{logit}(p)$, $s(\gamma) = \log(1 + \exp(\gamma)) = -\log(1-p)$, $\varkappa = 1$ and $c(y, \varkappa) = 0$. The above equation becomes:

$$p(y) = \exp\left(\frac{y\gamma - s(\gamma)}{\varkappa} + c(y, \varkappa)\right),$$

hence, the Bernoulli distribution belongs to the one-parameter exponential family with $s'(\gamma) = \exp(\gamma)/(1 + \exp(\gamma))$, $s''(\gamma) = \exp(\gamma)/(1 + \exp(\gamma))^2$. The mean of $Y$ is $\mu = p$; the canonical link is $g(\mu) = \gamma = \log(\mu/(1-\mu))$ and weight $w = \left(\text{Var}(y)(g'(\mu))^2\right)^{-1} = \mu(1-\mu)$.

## D2. Gradient and Hessian of $\log \widetilde{p}(\mathbf{v}|\mathcal{D})$

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

$$
\begin{aligned}
&\log \widetilde{p}(\mathbf{v}|\mathcal{D}) \\
&\doteq -\frac{1}{2}\underbrace{\log|B^\top \widetilde{W} B + Q_{\boldsymbol{\xi}}^{\mathbf{v}}|}_{\text{Term I}} + \underbrace{\left(\frac{\nu+K-1}{2}\right)\sum_{j=1}^{q} v_j}_{\text{Term II}} + \underbrace{\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}}_{\text{Term III}} \\
&\underbrace{-\frac{1}{\varkappa}\sum_{i=1}^{n} s\left(\mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}\right)}_{\text{Term IV}} - \frac{1}{2}\underbrace{\widetilde{\boldsymbol{\varpi}}^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} Q_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}}_{\text{Term V}} \\
&-\left(\frac{\nu}{2}+a_\delta\right)\underbrace{\sum_{j=1}^{q}\log\left(b_\delta + \frac{\nu}{2}\exp(v_j)\right)}_{\text{Term VI}}, \qquad \text{(D2.1)}
\end{aligned}
$$

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

### 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 (D2.1) can be obtained using Jacobi's formula:

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

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

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

Derivation of Term II with respect to $v_j$ simply equals a scalar:

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

Partial derivatives of Term III and Term IV are obtained using the following result:

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

Hence for Term III, using the property that the trace is invariant under cyclic permutations:

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

$$
\begin{aligned}
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathrm{Tr}\left(\mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n} y_i \mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}. \qquad \text{(D2.2)}
\end{aligned}
$$

For Term IV we use the chain rule and obtain:

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

The partial derivative of Term V is obtained as follows:

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

$$
\begin{aligned}
&= -\widetilde{\boldsymbol{\varpi}}^\top \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}}^\top \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}} \\
&\quad + \widetilde{\boldsymbol{\varpi}}^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} \\
&= -\widetilde{\boldsymbol{\varpi}}^\top \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}}^\top \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)^\top \\
&\quad + \widetilde{\boldsymbol{\varpi}}^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} \\
&= -\widetilde{\boldsymbol{\varpi}}^\top \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}}^\top \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}}^\top \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}}^\top \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}}^\top \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 $\widetilde{\Upsilon}_{\mathbf{v}}^{j} := \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}$. 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:

$$
\begin{aligned}
&\frac{\partial \log \widetilde{p}(\mathbf{v}|\mathcal{D})}{\partial v_j} \\
&= -\frac{1}{2} \underbrace{\operatorname{Tr}\left( \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{P}_{v_j} \right)}_{\text{Term VII}} + \left( \frac{\nu + K - 1}{2} \right) \underbrace{- \frac{1}{\varkappa} \sum_{i=1}^{n} y_i \mathbf{b}_i^\top \widetilde{\Upsilon}_{\mathbf{v}}^{j} \widetilde{\boldsymbol{\varpi}}}_{\text{Term VIII}} \\
&+ \underbrace{\frac{1}{\varkappa} \sum_{i=1}^{n} s'\left( \mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}} \right) \mathbf{b}_i^\top \widetilde{\Upsilon}_{\mathbf{v}}^{j} \widetilde{\boldsymbol{\varpi}}}_{\text{Term IX}} + \underbrace{\widetilde{\boldsymbol{\varpi}}^\top \widetilde{\Upsilon}_{\mathbf{v}}^{j} Q_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} \widetilde{\boldsymbol{\varpi}}}_{\text{Term X}} - \frac{1}{2} \underbrace{\widetilde{\boldsymbol{\varpi}}^\top \widetilde{\Upsilon}_{\mathbf{v}}^{j} \widetilde{\boldsymbol{\varpi}}}_{\text{Term XI}} \\
&- \underbrace{\frac{\left( \frac{\nu}{2} + a_\delta \right)}{1 + \frac{2b_\delta}{\nu \exp(v_j)}}}_{\text{Term XII}}, \; j = 1, \ldots, q.
\end{aligned}
$$

## Hessian associated to the penalty in a GAM

### Diagonal elements

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

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

Let us derive the intermediate result:

$$\begin{aligned}\frac{\partial\widetilde{\Upsilon}_{\mathbf{v}}^j}{\partial v_j} &= \frac{\partial}{\partial v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\\ &= \left(\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial\widetilde{P}_{v_j}}{\partial v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_j}\right)\\ &= \left(-\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}-\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\\ &= \left(-2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\Upsilon}_{\mathbf{v}}^j\right). \qquad \text{(D2.3)}\end{aligned}$$

Partial differentiation of Term VIII yields:

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

and using intermediate result (D2.3), one obtains for Term VIII:

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

For Term IX, we have:

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

Using (D2.2) and intermediate result (D2.3) we have for Term IX:

$$
\begin{aligned}
&\frac{\partial}{\partial v_j}\left(\frac{1}{\varkappa}\sum_{i=1}^{n} s'\left(\mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\mathbf{b}_i^\top \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right) \\
&= \frac{1}{\varkappa}\sum_{i=1}^{n}\left(s''(\mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\left(-\mathbf{b}_i^\top \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right)\left(\mathbf{b}_i^\top \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right) + s'(\mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^\top\right. \\
&\quad \left. \times\left(-2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2 \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\Upsilon}_{\mathbf{v}}^j\right)\widetilde{\boldsymbol{\varpi}}\right) \\
&= -\frac{1}{\varkappa}\sum_{i=1}^{n}\left(s'(\mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^\top \left(2\left(\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\right)^2 \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\Upsilon}_{\mathbf{v}}^j\right)\widetilde{\boldsymbol{\varpi}}\right. \\
&\quad \left. + s''(\mathbf{b}_i^\top \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^\top \widetilde{\Upsilon}_{\mathbf{v}}^j \widetilde{\boldsymbol{\varpi}}\right)^2\right).
\end{aligned}
$$

The partial derivative of Term X is obtained as follows:

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

Partial differentiation of Term XI gives us:

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

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

Finally derivation of Term XII is simply:

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

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}\operatorname{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)\\
&\quad+\frac{1}{\varkappa}\sum_{i=1}^{n}y_i\mathbf{b}_i^{\top}\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}}\\
&\quad-\frac{1}{\varkappa}\sum_{i=1}^{n}\left(s'(\mathbf{b}_i^{\top}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^{\top}\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.\\
&\quad\left.\left.+s''(\mathbf{b}_i^{\top}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^{\top}\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\right)^2\right)-2\widetilde{\boldsymbol{\varpi}}^{\top}\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}}\right.\\
&\quad+\widetilde{\boldsymbol{\varpi}}^{\top}\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}}^{\top}\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\\
&\quad+\widetilde{\boldsymbol{\varpi}}^{\top}\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}}^{\top}\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{\boldsymbol{\varpi}}\\
&\quad-\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}
$$

**Off-diagonal elements**

Note that for index $s \neq j$ we have for Term VII:

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

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 intermediate result:

$$\begin{aligned}\frac{\partial\widetilde{\Upsilon}_{\mathbf{v}}^{j}}{\partial v_s} &= \frac{\partial}{\partial v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\\ &= \left(\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial\widetilde{P}_{v_j}}{\partial v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}\right)\\ &= \left(-\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} - \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\\ &= -\left(\widetilde{\Upsilon}_{\mathbf{v}}^{s}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}} + \widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\Upsilon}_{\mathbf{v}}^{s}\right). \qquad \text{(D2.4)}\end{aligned}$$

Result (D2.4) can be used to obtain the differentiation of Term VIII:

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

To derive Term IX, we also use result (D2.4):

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

Partial differentiation of Term X goes as follows:

$$
\begin{aligned}
&\frac{\partial}{\partial v_s}\left(\widetilde{\boldsymbol{\varpi}}^\top \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_s}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}^\top \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_s}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^\top \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}}^\top \frac{\partial}{\partial v_s}\left(\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right)\\
&=\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^\top\left(\frac{\partial\widetilde{\Upsilon}_{\mathbf{v}}^{j}}{\partial v_s}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_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\frac{\partial\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}}{\partial v_s}\right)\right)\\
&=\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^\top\left(-\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)Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}+\widetilde{\Upsilon}_{\mathbf{v}}^{j}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right.\right.\\
&\quad\left.\left.-\widetilde{\Upsilon}_{\mathbf{v}}^{j}Q_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}_{\boldsymbol{\xi}}^{\mathbf{v}}\right)\right)
\end{aligned}
$$

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

Partial differentiation of Term XI gives us:

$$
\begin{aligned}
\frac{\partial}{\partial v_s}\left(\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right) &= \frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)=\frac{\partial}{\partial v_s}\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^j_{\mathbf{v}}\right)\\
&= \mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^\top\frac{\partial\widetilde{\Upsilon}^j_{\mathbf{v}}}{\partial v_s}\right)\\
&= -\mathrm{Tr}\left(\widetilde{\boldsymbol{\varpi}}\widetilde{\boldsymbol{\varpi}}^\top\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{\boldsymbol{\varpi}}^\top\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)\\
&= -\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}-\left(\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{P}_{v_j}\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)^\top\\
&= -\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}-\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}\\
&= -2\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}.
\end{aligned}
$$

Finally, 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)\\
&\quad +\frac{1}{\varkappa}\sum_{i=1}^{n}y_i\mathbf{b}_i^\top\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^\top\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}})\mathbf{b}_i^\top\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^\top\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}})\left(\mathbf{b}_i^\top\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\left(\mathbf{b}_i^\top\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}\right)\Bigg)\\
&\quad -\widetilde{\boldsymbol{\varpi}}^\top\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}}^\top\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}}^\top\widetilde{\Upsilon}^j_{\mathbf{v}}\widetilde{P}_{v_s}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}-\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^j_{\mathbf{v}}Q^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{\boldsymbol{\varpi}}+\widetilde{\boldsymbol{\varpi}}^\top\widetilde{\Upsilon}^s_{\mathbf{v}}\widetilde{P}_{v_j}\widetilde{\mathcal{M}}^{\mathbf{v}}_{\boldsymbol{\xi}}\widetilde{\boldsymbol{\varpi}}.
\end{aligned}
$$

# Bibliography

Abrahamowicz, M., Clampl, A. and Ramsay, J. O. (1992). Nonparametric density estimation for censored survival data: Regression-spline approach. *Canadian Journal of Statistics* **20**(2), 171–185.

Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *in* '2nd International Symposium on Information Theory'. Akademiai Kaido.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer New York.

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

Azevedo-Filho, A. and Shachter, R. D. (1994). Laplace's method approximations for probabilistic inference in belief networks with continuous variables. *Uncertainty Proceedings* pp. 28–36.

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

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

Bender, R., Augustin, T. and Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine* **24**(11), 1713–1723.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* **47**(259), 501–515.

Bertrand, A., Legrand, C., Léonard, D. and Van Keilegom, I. (2017). Robustness of estimation methods in a survival cure model with mismeasured covariates. *Computational Statistics & Data Analysis* **113**, 3–18.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)* **11**(1), 15–53.

Bornkamp, B. (2011). Approximating probability densities by iterated Laplace approximations. *Journal of Computational and Graphical Statistics* **20**(3), 656–669.

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

Bremhorst, V. and Lambert, P. (2016). Flexible estimation in cure survival models using Bayesian P-splines. *Computational Statistics & Data Analysis* **93**, 270–284.

Brezger, A. and Steiner, W. J. (2008). Monotonic regression based on Bayesian P-splines: An application to estimating price response functions from store-level scanner data. *Journal of Business & Economic Statistics* **26**(1), 90–104.

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

Cai, C., Zou, Y., Peng, Y. and Zhang, J. (2012). smcure: An R-package for estimating semiparametric mixture cure models. *Computer Methods and Programs in Biomedicine* **108**(3), 1255–1260.

Chen, M.-H., Ibrahim, J. G. and Sinha, D. (1999). A new Bayesian model for survival data with a surviving fraction. *Journal of the American Statistical Association* **94**(447), 909–919.

Chyong-Mei, C. and Chen-Hsin, C. (2016). Heteroscedastic transformation cure regression models. *Statistics in Medicine* **35**(14), 2359–2376.

Comstock, G. W., Bush, T. L. and Helzlsouer, K. (1992). Serum retinol, beta-carotene, vitamin E, and selenium as related to subsequent cancer of specific sites. *American Journal of Epidemiology* **135**(2), 115–121.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34**(2), 187–220.

Dey, D. D., MüIler, P. and Sinha, D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. Vol. 133. Springer New York.

Donnell, D. J., Buja, A. and Stuetzle, W. (1994). Analysis of additive dependencies and concurvities using smallest additive principal components. *The Annals of Statistics* **22**(4), 1635–1668.

Dos Passos, W. (2009). *Numerical Methods, Algorithms and Tools in C*. CRC Press.

Drzewiecki, K. T. and Andersen, P. (1982). Survival with malignant melanoma: A regression analysis of prognostic factors. *Cancer* **49**(11), 2414–2419.

Drzewiecki, K. T., Christensen, H. E., Ladefoged, C. and Poulsen, H. (1980). Clinical course of cutaneous malignant melanoma related to histopathological criteria of primary tumour. *Scandinavian Journal of Plastic and Reconstructive Surgery* **14**(3), 229–234.

Drzewiecki, K. T., Ladefoged, C. and Christensen, H. E. (1980). Biopsy and prognosis for cutaneous malignant melanomas in clinical stage I. *Scandinavian Journal of Plastic and Reconstructive Surgery* **14**(2), 141–144.

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

Eilers, P. H. C. (2018). The truth about the effective dimension. *Statistica Neerlandica* **72**(3), 201–209.

Eilers, P. H. C., Currie, I. D. and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis* **50**(1), 61–76.

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

Eilers, P. H. C. and Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**(6), 637–653.

Eilers, P. H. C., Marx, B. D. and Durbán, M. (2015). Twenty years of P-splines. *SORT: Statistics and Operations Research Transactions* **39**(2), 149–186.

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

Fong, Y., Rue, H. and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics* **11**(3), 397–412.

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

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

Gallardo, D. I., Bolfarine, H. and Pedroso-De-Lima, A. C. (2016). Promotion time cure rate model with bivariate random effects. *Communications in Statistics-Simulation and Computation* **45**(2), 603–624.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**(6), 721–741.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. in: Bernardo, J.M., Berger, J.0., Dawid, A.P., Smith, A.F.M. (eds.). *Bayesian Statistics* **4**, 641–649.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix Computations*. Vol. 3. John Hopkins University Press.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Gómez-Rubio, V. and Rue, H. (2018). Markov chain Monte Carlo with the Integrated Nested Laplace Approximation. *Statistics and Computing* **28**, 1033–1051.

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. https://doi.org/10.1016/j.csda.2018.02.007.

Gressani, O. and Lambert, P. (2020*a*). The Laplace-P-spline methodology for fast approximate Bayesian inference in additive partial linear models. *ISBA Discussion papers, DP-2020/20* . http://hdl.handle.net/2078.1/230728.

Gressani, O. and Lambert, P. (2020*b*). *The blapsr package for fast Bayesian inference in latent Gaussian models by combining Laplace approximations and P-splines.* Version 0.5.1, https://www.blapsr-project.org/.

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

Gu, H., Kenney, T. and Zhu, M. (2010). Partial generalized additive models: An information-theoretic approach for dealing with concurvity and selecting variables. *Journal of Computational and Graphical Statistics* **19**(3), 531–551.

Gurmu, S. (1997). Semi-parametric estimation of hurdle regression models with an application to medicaid utilization. *Journal of Applied Econometrics* **12**(3), 225–242.

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

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

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science* **1**(3), 297–310.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association* **82**(398), 371–386.

Hogben, L. T. (1968). *Statistical Theory: The relationship of probability, credibility, and error; an examination of the contemporary crisis in statistical theory from a behaviourist viewpoint.* W.W. Norton & Company.

Holmes, M. (2007). *Introduction to Numerical Methods in Differential Equations.* Texts in Applied Mathematics. Springer New York.

Householder, A. S. (1958). Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM (JACM)* **5**(4), 339–342.

Hui, F. K. C., You, C., Shang, H. L. and Müller, S. (2019). Semiparametric regression using variational approximations. *Journal of the American Statistical Association* **114**(528), 1765–1777.

Ibrahim, J. G., Chen, M.-H. and Sinha, D. (2001). Bayesian semiparametric models for survival data with a cure fraction. *Biometrics* **57**(2), 383–388.

Ibrahim, J. G., Chu, H. and Chen, L. M. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology* **28**(16), 2796.

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

Jiang, H., Brown, P. E., Rue, H. and Shimakura, S. (2014). Geostatistical survival models for environmental risk assessment with large retrospective cohorts. *Journal of the Royal Statistical Society: Series A* **177**(3), 679–695.

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

Kalbfleisch, J. D. and Prentice, R. L. (2011). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons.

Krivobokova, T., Crainiceanu, C. M. and Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics* **17**(1), 1–20.

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

Lambert, P. (2020). Fast Bayesian inference in nonparametric double additive location-scale models with right- and interval-censored data. *ArXiv:2005.05156* .

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.

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 évènements. *Mémoires de Mathématique et de Physique, Presentés à l'Académie Royale des Sciences, par divers Savans & lûs dans ses Assemblées* **6**, 621–656.

Laplace, P. S. (1986). Memoir on the probability of the causes of events. *Statistical Science* **1**(3), 364–378.

Laurie, J. A., Moertel, C. G., Fleming, T. R., Wieand, H. S., Leigh, J. E., Rubin, J., McCormack, G. W., Gerstner, J. B., Krook, J. E., Malliard, J. et al. (1989). Surgical adjuvant therapy of large-bowel carcinoma: an evaluation of levamisole and the combination of levamisole and fluorouracil. *Journal of Clinical Oncology* **7**(10), 1447–1456.

Leonard, T. (1982). Comment on "A Simple Predictive Density Function," by M. Lejeune and G.D. Faulkenberry. *Journal of the American Statistical Association* **77**(379), 657–658.

Li, G. and Lin, C. (2009). Analysis of two-sample censored data using a semiparametric mixture model. *Acta Mathematicae Applicatae Sinica, English Series* **25**, 389–398.

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

Lindley, D. V. (1961). The use of prior probability distributions in statistical inference and decision. *in* 'Proc. 4th Berkeley Symp. on Math. Stat. and Prob'. pp. 453–468.

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

Liu, X., Wang, L. and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica* **21**(3), 1225–1248.

Lopes, C. M. C. and Bolfarine, H. (2012). Random effects in promotion time cure rate models. *Computational Statistics & Data Analysis* **56**(1), 75–87.

Luts, J., Broderick, T. and Wand, M. P. (2014). Real-time semiparametric regression. *Journal of Computational and Graphical Statistics* **23**(3), 589–615.

Lyche, T., Manni, C. and Speleers, H. (2018). *Foundations of Spline Theory: B-splines, Spline Approximation, and Hierarchical Refinement*. Springer.

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

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

Mantovan, P. and Secchi, P. (2010). *Complex Data Modeling and Computationally Intensive Statistical Methods*. Springer.

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

Martino, S. (2007). Approximate Bayesian inference for latent Gaussian models. *Fakultet for Informasjonsteknologi, Matematikk og Elektroteknikk* .

Martino, S., Aas, K., Lindqvist, O., Neef, L. R. and Rue, H. (2011). Estimating stochastic volatility models using Integrated Nested Laplace Approximations. *The European Journal of Finance* **17**(7), 487–503.

Martino, S., Akerkar, R. and Rue, H. (2011). Approximate Bayesian inference for survival models. *Scandinavian Journal of Statistics* **38**(3), 514–528.

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

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis* **28**(2), 193–209.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear models*. Vol. 37. CRC press.

Michalowicz, J. V., Nichols, J. M. and Bucholtz, F. (2013). *Handbook of Differential Entropy*. Chapman and Hall/CRC.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Goodman, P. J., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., Glick, J. H. et al. (1990). Levamisole and fluorouracil for adjuvant therapy of resected colon carcinoma. *New England Journal of Medicine* **322**(6), 352–358.

Moertel, C. G., Fleming, T. R., Macdonald, J. S., Haller, D. G., Laurie, J. A., Tangen, C. M., Ungerleider, J. S., Emerson, W. A., Tormey, D. C., Glick, J. H. et al. (1995). Fluorouracil plus levamisole as effective adjuvant therapy after resection of stage iii colon carcinoma: a final report. *Annals of Internal Medicine* **122**(5), 321–326.

Mosteller, F. and Wallace, D. (1964). *Inference and Disputed Authorship: The Federalist*. Addison-Wesley.

Nelder, J. A. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A* **135**(3), 370–384.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R. and Group, S. C. P. S. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology* **130**(3), 511–521.

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

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

Papoulis, A. and Pillai, S. U. (2002). *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill Education.

Plackett, R. L. (1949). A historical note on the method of least squares. *Biometrika* **36**(3/4), 458–460.

Rimm, E. B., Stampfer, M. J., Ascherio, A., Giovannucci, E., Colditz, G. A. and Willett, W. C. (1993). Vitamin E consumption and the risk of coronary heart disease in men. *New England Journal of Medicine* **328**(20), 1450–1456.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. CRC press.

Rosenberg, P. (1995). Hazard function estimation using B-splines. *Biometrics* **51**(3), 874–887.

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

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. K. (2017). Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application* **4**(1), 395–421.

Ruiz-Cárdenas, R., Krainski, E. and Rue, H. (2012). Direct fitting of dynamic models using Integrated Nested Laplace Approximations-INLA. *Computational Statistics & Data Analysis* **56**(6), 1808–1828.

Ruppert, D., Wand, M. P. and 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**(3), 64–81.

Savage, L. (1954). *The Foundations of Statistics*. New York, John Wiley and Sons.

Schoenberg, I. J. (1946*a*). Contributions to the problem of approximation of equidistant data by analytic functions: Part a.—on the problem of smoothing or graduation. a first class of analytic approximation formulae. *Quarterly of Applied Mathematics* **4**(1), 45–99.

Schoenberg, I. J. (1946*b*). Contributions to the problem of approximation of equidistant data by analytic functions: Part b—on the problem of osculatory interpolation. a second class of analytic approximation formulae. *Quarterly of Applied Mathematics* **4**(2), 112–141.

Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using INLA. *Environmetrics* **22**(6), 725–734.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**(2), 461–464.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological)* **47**(1), 1–52.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Vol. 26. CRC press.

Sørbye, S. H. and Rue, H. (2011). Simultaneous credible bands for latent Gaussian models. *Scandinavian Journal of Statistics* **38**(4), 712–725.

Stigler, S. M. (1981). Gauss and the invention of least squares. *The Annals of Statistics* **9**(3), 465–474.

Stukel, T. (2008). 'Determinants of plasma retinol and beta-carotene levels'. http://lib.stat.cmu.edu/datasets/Plasma_Retinol.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.1-12.

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.

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

Tsodikov, A. D. (1998). A proportional hazards model taking account of long-term survivors. *Biometrics* **54**(4), 1508–1516.

Tsodikov, A. D. (2002). Semi-parametric models of long-and short-term survival: an application to the analysis of breast cancer survival in Utah by age and stage. *Statistics in Medicine* **21**(6), 895–920.

Tsodikov, A. D. (2003). Semiparametric models: a generalized self-consistency approach. *Journal of the Royal Statistical Society: Series B* **65**(3), 759–774.

Umlauf, N., Adler, D., Kneib, T., Lang, S. and Zeileis, A. (2015). Structured additive regression models: An R interface to BayesX. *Journal of Statistical Software* **63**(21), 1–46.

Upton, G. and Cook, I. (2014). *A Dictionary of Statistics (Third edition)*. Oxford University Press.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Fourth edn. Springer. New York.

Wand, M. and Ormerod, J. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics* **50**(2), 179–198.

Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing. *Journal of the American Statistical Association* **112**(517), 137–168.

Weisberg, S. (1980). *Applied Linear Regression*. New York, Wiley.

Wienke, A. (2010). *Frailty Models in Survival Analysis*. CRC press.

Wood, S. N. (2003). Thin plate regression splines. *Journal of the Royal Statistical Society: Series B* **65**(1), 95–114.

Wood, S. N. (2013). On p-values for smooth components of an extended generalized additive model. *Biometrika* **100**(1), 221–228.

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

Wood, S. N., Pya, N. and Säfken, B. (2016). Smoothing parameter and model selection for general smooth models. *Journal of the American Statistical Association* **111**(516), 1548–1563.

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

Yakovlev, A. Y., Tsodikov, A. D. and Asselain, B. (1996). *Stochastic models of tumor latency and their Biostatistical applications*. Vol. 1 of Mathematical Biology and Medicine. World Scientific. Singapore.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association* **93**(441), 120–131.

Yin, G. and Ibrahim, J. G. (2005). Cure rate models: a unified approach. *Canadian Journal of Statistics* **33**(4), 559–570.

Yoon, J. W. and Wilson, S. P. (2011). Inference for latent variable models with many hyperparameters. *in* 'Proc. 58th World Statistical Congress, Dublin'.

Zeng, D., Yin, G. and Ibrahim, J. G. (2006). Semiparametric transformation models for survival data with a cure fraction. *Journal of the American Statistical Association* **101**(474), 670–684.

Zhang, S., Hunter, D. J., Forman, M. R., Rosner, B. A., Speizer, F. E., Colditz, G. A., Manson, J. E., Hankinson, S. E. and Willett, W. C. (1999). Dietary carotenoids and vitamins A, C, and E and risk of breast cancer. *Journal of the National Cancer Institute* **91**(6), 547–556.