Penalized complexity priors for degrees of freedom in Bayesian P-splines

Penalized complexity priors for degrees of freedom in Bayesian P-splines

Massimo Ventrucci*and Håvard Rue

Department of Statistical Sciences, University of Bologna, Bologna, Italy

Department of Mathematics, Norwegian Institute of Technology, Trondheim, Norway

Abstract

Bayesian P-splines assume an intrinsic Gaussian Markov random field prior on the spline coefficients, conditional on a precision hyper-parameter τ. Prior elicitation of τ is difficult. To overcome this issue we aim to building priors on an interpretable property of the model, indicating the complexity of the smooth function to be estimated. Following this idea, we propose Penalized Complexity (PC) priors for the number of effective degrees of freedom. We present the general ideas behind the construction of these new PC priors, describe their properties and show how to implement them in P-splines for Gaussian data.

Keywords: Bayesian P-splines; degrees of freedom; Penalized complexity priors; penalized spline regression

1 Introduction

Penalized spline (P-spline) regression is a well established and numerically stable approach for smoothing (Eilers and Marx, 1996; Ruppert et al., 2003). Typically, P-spline components are implemented in Bayesian additive regression models (Fahrmeir et al., 2013) to fit non linear covariate effects or higher dimensional effects such as spatial and spatio-temporal smooth trends. The P-spline approach proposed by Eilers and Marx (1996) uses equally-spaced B-splines and constructs a smooth function as the sum of these B-splines scaled by spline coefficients. A regularizing penalty is assumed on these coefficients to control the degree of smoothness of the fitted function. A common approach is to penalize the sum of second order squared differences between adjacent spline coefficients, but specific penalties can be designed to drive the fit towards desired features (Eilers and Marx, 2010). This is a very useful strategy in presence of prior information about the shape, or degree of smoothness, of the function to be estimated.

The Bayesian approach to P-spline by Lang and Brezger (2004) assumes an Intrinsic Gaussian Markov Random Field (IGMRF) prior on the spline coefficients. An IGMRF is a multivariate

*Corresponding author. Email: massimo.ventrucci@unibo.it

normal distribution with rank deficient precision matrix $\boldsymbol{Q}(\tau)$, depending on a precision hyper-parameter τ. Similarly to a regularizing penalty, the IGMRF forces the spline coefficients to be shrunk towards an infinite smooth model, which we will denote as the *base model*. The degree of smoothness of the base model depends on the order of the IGMRF; for instance, an IGMRF of (polynomial) order 2 forces shrinkage towards a linear trend, i.e. a polynomial of degree one (Rue and Held, 2005).

The amount of shrinkage towards the corresponding base model depends on the IGMRF precision τ. The prior $\pi(\tau)$ can have a substantial impact on the posterior distribution of the spline coefficients and hence, to some extent, on the shape of the fit. A common strategy in Bayesian P-splines is to adopt the conjugate Gamma family, i.e. Gamma(a, b), with shape a and rate b (Fahrmeir and Kneib, 2009; Lang and Brezger, 2004). Lang and Brezger (2004) suggest to choose $a = 1$ and small b, e.g., $b = 5 \cdot 10^{-4}$, leading to a diffuse prior for τ^{-1}. Jullion and Lambert (2007) note that the choice of b clearly affects the smoothness of the fitted curve, when sample size or signal-to-noise ratio is small, and propose a mixture of Gamma distributions with different b values. Another popular choice is the Gamma(ϵ, ϵ), with small ϵ (e.g. $\epsilon = 0.001$, which is the default option in the software *BayesX* (Belitz et al., 2000)) as an attempt of vagueness on τ^{-1}. The suitability of the Gamma family as a noninformative prior for the scale parameters in hierarchical models has been debated in the literature (Gelman, 2006); overfitting due to Gamma priors has been demonstrated in Frühwirth-Schnatter and Wagner (2010, 2011); Simpson et al. (2014). In particular, in Bayesian P-splines, the main difficulty with using a Gamma prior on τ is that τ scales differently according to the amount of noise present in the data and the number (and location) of knots selected by the user.

The present work proposes a new prior for τ which is informative about model complexity and implicitly accounts for different choices about number (and location) of knots. A suitable measure of complexity of the P-spline model is the number of *effective degrees of freedom*, in the following denoted as d, calculated as the trace of the hat matrix (Hastie and Tibshirani, 1990). The value d relates to the degree of a polynomial equivalent to the smooth function to be estimated. An expert user who has a prior guess about the shape of this function may find easy to elicit d. As an example, for a monotonic cubic trend one may elicit d in a range between 3 and 5 and assign very low prior probability to $d > 5$. The key point is that, in presence of this prior information, elicitation of a range for d is intuitive and immediate, whereas elicitation of a distribution for τ, directly, is very difficult.

The challenge is to design a prior distribution on a model property (i.e., d) rather than on a parameter of the model (i.e., τ). To achieve this, we follow the Penalized Complexity (PC) prior approach proposed by Simpson et al. (2014). Within this framework, we derive the PC prior for d and calibrate it by two intuitive parameters: U, an upper bound for d and α, the prior

probability assigned to $d > U$. In the example aforementioned, the user would only need to set U equal to 5 and $\alpha = 0.01$, or some other small value. As a further challenging point, note that d depends on the noise variance characterizing the dataset. Thus, implementing the proposed PC prior for degrees of freedom in real datasets, where the noise variance is typically unknown, implies defining a joint prior on two quantities, the IGMRF precision and the noise precision.

The plan of the paper is as follows. In section 2.1, the Bayesian P-spline approach is revised with focus on the challenges to be addressed in defining a prior for τ. The principles behind the construction of a PC prior for τ are revised in section 3. In section 4, the PC prior for d is derived and its properties described in the case of known noise variance. In section 5, we show how to implement the PC prior for d when the noise variance is unknown, focusing on an additive P-splines model framework. Results from a simulation study assessing the impact of the proposed prior compared to standard Gamma priors and other PC priors proposed in the recent literature are illustrated in section 6. An application of these new priors in a P-spline model for nitrate concentrations observed in river *Oglio*, Lombardia Region, Italy, is illustrated in section 7. The paper closes with a discussion in section 8.

2 Background on P-splines

Let $\boldsymbol{y} = [y_1, ..., y_n]^\mathsf{T}$ be observations of a response variable, $\boldsymbol{x}$ be a continuous covariate, f be a smooth function describing the effect of the covariate on the response and $\boldsymbol{\epsilon}$ be independent errors with zero mean and variance τ_ϵ^{-1}. The P-spline model (Eilers and Marx, 1996) is $y = f(\boldsymbol{x}) + \boldsymbol{\epsilon}$, $f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\beta}$, where $\boldsymbol{B}$ is a n x K basis matrix containing K B-spline functions built on a set of equally-spaced (for simplicity) knots within the covariate domain, while $\boldsymbol{\beta}$ is a K x 1 vector of unknown spline coefficients. The method requires to select a generous number of knots to over-fit the data, to then add a penalty on $\boldsymbol{\beta}$ which smooths adjacent spline coefficients. In the frequentist approach, $\boldsymbol{\beta}$ is estimated via penalized maximum likelihood, conditional on a tuning parameter regulating the degree of smoothness of f. The optimal tuning can be found via cross validation (Wood, 2006) or estimated via restricted maximum likelihood in a mixed model representation (Ruppert et al. (2003), ch. 4). P-splines are widely used in generalized additive models (Hastie and Tibshirani, 1990; Wood, 2006) or structured additive regression models (Fahrmeir et al., 2004, 2013). Higher-dimensional smooth functions can also be represented as P-splines, using the tensor product of marginal B-spline bases (Eilers et al., 2006; Currie et al., 2006). For a systematic presentation of the different approaches to penalized spline regression see Ruppert et al. (2003); for an excellent review of spline methods and their applications in statistical modelling see Hastie et al. (2009), Wakefield (2013) and Wood (2006).

2.1 Bayesian P-splines

The Bayesian approach to P-splines (Lang and Brezger, 2004) assumes an IGMRF prior on the spline coefficients,

$$\pi(\boldsymbol{\beta}|\tau_\beta) = (2\pi)^{-\mathtt{rank}(\boldsymbol{R})/2}(|\tau_\beta \boldsymbol{R}|^*)^{1/2} \exp\left\{-\frac{\tau_\beta}{2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{R}\boldsymbol{\beta}\right\} \tag{1}$$

where the precision $\boldsymbol{Q}(\tau_\beta)$ is given by $\tau_\beta \boldsymbol{R}$. Matrix $\boldsymbol{R}$ is denoted as the structure of the IGMRF, i.e. a K x K sparse matrix with non-zero entries indicating conditional dependencies among the spline coefficients, τ_β is a scalar precision hyper-parameter and $|\tau_\beta \boldsymbol{R}|^*$ is the generalized determinant. Throughout the paper we will assume $\boldsymbol{R} = \boldsymbol{D}_r^\mathsf{T}\boldsymbol{D}_r$, where $\boldsymbol{D}_r$ is a $(K-r)$ x K matrix such that $\boldsymbol{D}_r\boldsymbol{\beta} = \Delta^r\boldsymbol{\beta}$ (Eilers et al., 2006), with Δ^r the r^{th}-order difference operator. In this form, $\boldsymbol{R}$ is the structure of an r^{th}-order random walk on $\boldsymbol{\beta}$ (Rue and Held (2005) ch. 3) with $\text{rank}(\boldsymbol{R}) = K - r$, where r indicates the order of the IGMRF (1).

The IGMRF (1) describes deviation from a base model, which is a polynomial of degree $(r-1)$. The amount of deviation depends on τ_β. A fully Bayesian specification requires priors on τ_β and τ_ϵ. Since we usually have enough information in the data to estimate τ_ϵ, the prior $\pi(\tau_\epsilon)$ has less impact on the fit. The hyper-parameter τ_β enters at a lower level in the hierarchy, the data bring little information on it and the prior $\pi(\tau_\beta)$ can have a substantial impact on the posterior distribution of $\boldsymbol{\beta}$ and, as a consequence, on the smoothness of f. Therefore, specification of $\pi(\tau_\beta)$ should be as consistent as possible with the prior information actually available about the smoothness of the function to be estimated.

The marginal variance of the IGMRF (1), given by the diagonal elements of $\Sigma^* = \tau_\beta^{-1}\boldsymbol{R}^{-1}$, depends on K. We denote this as the *scaling issue* (Srbye and Rue, 2014), meaning that the amount of deviation from the base model depends on the number of knots. This is illustrated in Figure 1, where the two panels report the marginal standard deviation of the smooth $f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\beta}$, for $K = \{50, 100\}$. On the other hand, results (not shown here) show that the degree of the B-splines has little or no impact on the marginal variance of $\boldsymbol{B}\boldsymbol{\beta}$, especially when K is large enough, say $K > 50$.

2.2 Degrees of freedom

The scaling issue can be avoided if we consider building priors on the number of effective degrees of freedom (Hastie and Tibshirani, 1990), $d = \text{tr}\left\{\left(\boldsymbol{B}^\mathsf{T}\boldsymbol{B} + \frac{\tau_\beta}{\tau_\epsilon}\boldsymbol{R}\right)^{-1}\boldsymbol{B}^\mathsf{T}\boldsymbol{B}\right\}$. If we think of the smooth $f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\beta}$ as a polynomial, then d can be thought of as the degree of this polynomial. In presence of prior information on the degree of an equivalent polynomial, it seems a sensible approach to design a prior for d, $\pi(d)$, instead of τ_β.

A fundamental issue when building $\pi(d)$ is that d depends on both precisions τ_β and τ_ϵ. The former regulates the number of effective degrees of freedom, conditionally on the latter. When τ_ϵ is known, the construction of $\pi(d)$ can be based on the prior $\pi(\tau_\beta)$ (see section 4). When τ_ϵ is unknown, the prior $\pi(d)$ will be specified in terms of the joint $\pi(\tau_\beta|\tau_\epsilon)\pi(\tau_\epsilon)$, following a fully Bayesian approach (see section 5).

The degrees of freedom can be reduced to

$$d = \text{tr}\left\{\left(\boldsymbol{I} + \frac{\tau_\beta}{\tau_\epsilon}\boldsymbol{R}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}\right)^{-1}\right\} = \sum_{k=1}^{K}\frac{1}{1+\frac{\tau_\beta}{\tau_\epsilon}v_k}, \tag{2}$$

where $v_1, ..., v_K$ are the eigenvalues of $\boldsymbol{R}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}$, whose null space has dimension r (the rank deficiency of $\boldsymbol{R}$). When the factor τ_β/τ_ϵ goes to infinity we obtain the minimum number of degrees of freedom, $d = r$. When τ_β/τ_ϵ goes to zero we obtain the maximum number of degrees of freedom, K, corresponding to the most flexible model under the assumed IGMRF.

The prior $\pi(d)$ depends on the eigenvalues of $\boldsymbol{R}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}$, hence on the choice of $\boldsymbol{B}$. Hereafter, $\boldsymbol{B}$ will be referred to simply as *design*, because it is determined by both the assumptions made by the user (location and number of knots, order of the B-splines) and the assumptions purely made by design (location and number of observations along the covariate domain). Since the degrees of freedom depend on $\boldsymbol{B}$, it follows that $\pi(d)$ automatically accounts for the design. This will be discussed in detail in section 4.2.

3 PC priors for P-splines

The PC prior framework by Simpson et al. (2014) introduces a new concept for building priors in hierarchical additive models, where the latent structure is given by the sum of a number of model components described by a small number of flexibility parameters. Each model component is seen as a flexible extension of a base model. For instance, τ_β is a flexibility parameter for the IGMRF component $\pi(\boldsymbol{\beta}|\tau_\beta)$ and a natural base model corresponds to $\pi(\boldsymbol{\beta}|\tau_\beta^{-1} = 0)$. Below, the four principles underpinning the construction of a PC prior for τ_β are reviewed, following Simpson et al. (2014).

- *Parsimony.* A simple model should be preferred unless there is enough evidence for a more flexible one. Under this principle, the prior probability mass assigned to models of increasing complexity should decay as their distance from the base model (measured in terms of model complexity) increases. In Bayesian P-splines, the IGMRF prior operates on $\boldsymbol{\beta}$ (the object of inference) but we can extend the notion of base and flexible model to the *spline-modelled* function; we denote with $\boldsymbol{f}_0 = \boldsymbol{B}\boldsymbol{\beta}_0$ the base model, which is a polynomial

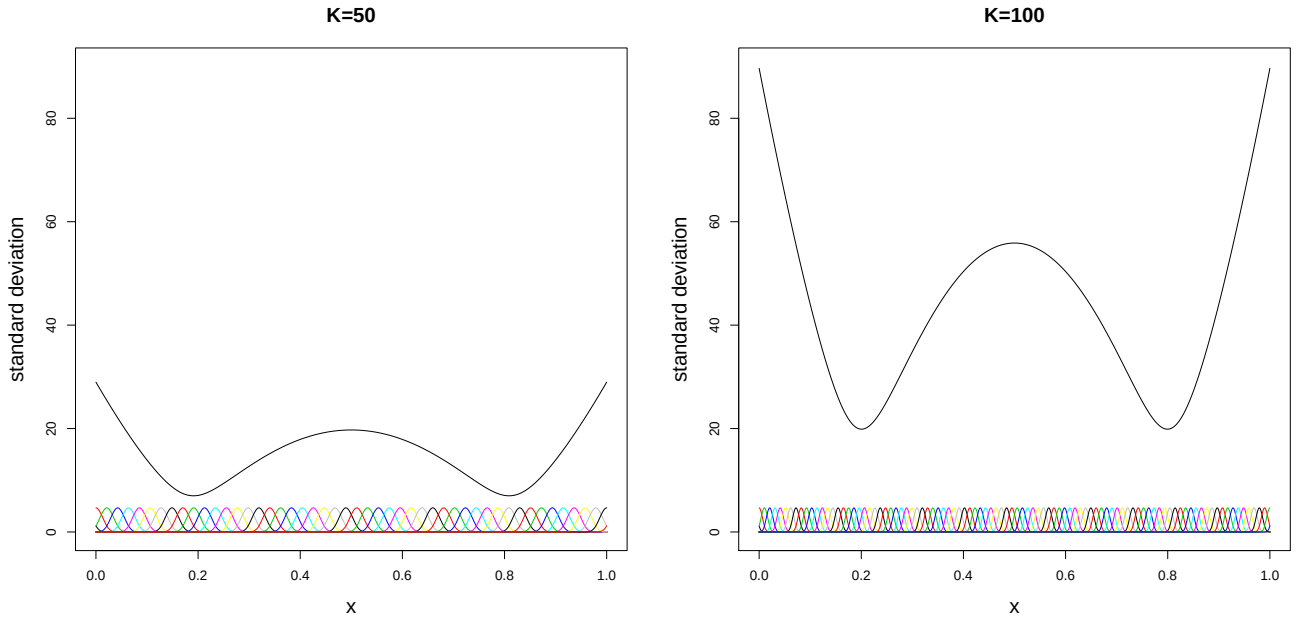

Figure 1: The scaling issue. The two panels show the marginal standard deviation of $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\beta}$ for varying dimension of the basis $\boldsymbol{B}$, $K = \{50, 100\}$, where $\boldsymbol{B}$ is a matrix of cubic B-splines (coloured lines) defined over the interval $(0, 1)$ and $\boldsymbol{\beta}$ is an IGMRF of order 2. The standard deviation (black line) is calculated as the squared root diagonal entries of $\boldsymbol{B}\boldsymbol{\Sigma}^*\boldsymbol{B}^\mathsf{T}$, with $\tau_\beta = 1$.

of degree $r - 1$, and with $\boldsymbol{f} = \boldsymbol{B}\boldsymbol{\beta}$ the flexible model, which reflects any deviation from such polynomial.

- *Measure of complexity.* The Kullback-Leibler divergence (KLD, Kullback and Leibler, 1951) is assumed to evaluate the distance, δ, between the complexities of two different models. We use $\text{KLD}(\boldsymbol{f}||\boldsymbol{f}_0)$ to denote the increased complexity of the flexible model $\boldsymbol{f}$ with respect to the base model $\boldsymbol{f}_0$. Since $\boldsymbol{B}$ is fixed by design, it is enough to evaluate $\text{KLD}(\boldsymbol{\beta}||\boldsymbol{\beta}_0)$. Let τ_{β_0} and τ_β be the precisions of the base and flexible model, respectively, it can be shown that $\text{KLD}(\boldsymbol{\beta}||\boldsymbol{\beta}_0)$ goes to $\frac{\tau_{\beta_0} K}{2\tau_\beta}$, for τ_β much lower than τ_{β_0} and $\tau_{\beta_0} \to \infty$; see a proof in Simpson et al. (2014). Finally, for convenience we take the transformation $\delta = \sqrt{2\text{KLD}(\boldsymbol{\beta}||\boldsymbol{\beta}_0)} = \sqrt{\tau_{\beta_0} K/\tau_\beta}$.

- *Constant rate penalization.* Flexible models are penalized by a constant decay rate for increasing δ. Following this principle, the PC prior is defined as an exponential distribution on the distance scale, $\pi_{PC}(\delta) = \lambda \exp(-\lambda\delta)$, with constant rate λ. It follows that the mode of a PC prior is always at the base model. By a change of variable and setting the rate

$\lambda = \theta/\sqrt{K\tau_{\beta_0}}$, Simpson et al. (2014) obtain the PC prior for τ_β as,

$$\begin{aligned} \pi_{PC}(\tau_\beta) &= \lambda \exp\left(-\lambda\sqrt{\tau_{\beta_0}K/\tau_\beta}\right)\left|\frac{\partial\sqrt{\tau_{\beta_0}K/\tau_\beta}}{\partial\tau_\beta}\right| \\ &= \frac{\theta}{2}\tau_\beta^{-3/2}\exp\left(-\theta/\sqrt{\tau_\beta}\right), \end{aligned} \tag{3}$$

which is a Gumbel$(1/2, \theta)$ type 2 distribution, $\theta > 0$.

- *User-defined scaling.* Often, the user has an idea about the size of an interpretable transformation of the original parameter τ_β, say $h(\tau_\beta)$ (e.g. degrees of freedom). In this case the user may elicit an upper bound U for $h(\tau_\beta)$ and set a prior probability α for the tail event, i.e. $\alpha = Pr(h(\tau_\beta) > U)$. Simpson et al. (2014) suggest to bound the marginal standard deviation, $1/\sqrt{\tau_\beta}$. To obtain θ in (3) it is enough to specify (U, α) and solve $Pr(1/\sqrt{\tau_\beta} > U) = \alpha$ for θ, which yields $\theta = -\log(\alpha)/U$.

PC priors can be helpful as *default* priors in complex hierarchical models where, typically, "it is difficult to elicit information about structural parameters that are further down the model hierarchy" (Simpson et al., 2014). In addition, the user-defined scaling approach enables to build *informative* priors for the original parameter τ_β or for a property of the associated model component, by tuning two intuitive parameters U and α. In the next section we introduce a new scaling approach to derive the PC prior for the degrees of freedom of a P-spline model component $\boldsymbol{B\beta}$. Other approaches might be possible though. For instance, recently Klein and Kneib (2015) proposed PC priors for the scale (or range of variation) of $\boldsymbol{B\beta}$, and showed via simulation that these outperformed the Gamma family in cases where the data are weakly informative and/or the size of the effects is close to the base model.

4 PC-priors for degrees of freedom

4.1 A new scaling approach

With no loss of generality, we derive the PC prior for degrees of freedom and study its properties under the assumption that the noise precision τ_ϵ is known. Given τ_ϵ, denote as $d(\tau_\beta) = \sum_{k=1}^{K} \frac{1}{1+\frac{\tau_\beta}{\tau_\epsilon}v_k}$ the function mapping the precision τ_β into the number of effective degrees of freedom, following (2). (Hereafter $d(\tau_\beta)$ will be referred to as the mapping). Figure 2 shows the mapping d in the log precision scale for $\tau_\epsilon = 1$ and various designs (left panel) and for a specific design and varying τ_ϵ (right panel).

We introduce a new user-defined scaling operating not directly on τ_β, but on $d(\tau_\beta)$. Let U be an upper bound for $d(\tau_\beta)$ and α a (small) probability associated to the tail event,

$$\alpha = Pr(d(\tau_\beta) > U) = Pr(\tau_\beta < d^{-1}(U)) = F(d^{-1}(U))$$

where F is the c.d.f of the Gumbel$(1/2, \theta)$ type 2 distribution. The PC prior resulting from this new scaling is a Gumbel type 2 as in (3) with $\theta = -\log(\alpha)\sqrt{d^{-1}(U)}$. In the following, $\pi_{PC}(d)$ will denote the *induced PC prior for degrees of freedom*, with $U \in (r, K)$ and $\alpha \in (0, 1)$ the parameters specifying the distribution.

4.2 Invariance under design

While the PC prior for τ_β in (3) does not take into account any information regarding the adopted design, the induced PC prior for d does. Indeed, different designs return different mappings d (see the left panel in Figure 2), which implies a desirable property: the PC prior $\pi_{PC}(d)$ is *invariant under design*; the term invariant here applies to the interpretation of the PC prior in terms of degrees of freedom, not to the density. Figure 3 illustrates this property. The density of $\pi_{PC}(d)$, with parameters $(U = 5, \alpha = 0.01)$, is displayed both in the d scale (left panel) and $log(\tau_\beta)$ scale (right panel), for different designs. By changing the design, the range of d also changes and the density $\pi_{PC}(d)$ adapts accordingly (Figure 3 left). However, even if $\pi_{PC}(d)$ materializes differently in different designs, the probability mass assigned to $d > U$ is always α. In the $\log(\tau_\beta)$ scale, the location of the PC prior is shifted between different designs (Figure 3 right). In our opinion, this shows well how difficult it is to define priors for degrees of freedom in the original scale: in this case, the user would have to figure out the correct location of $\pi_{PC}(\log(\tau_\beta))$ and shift it according to the adopted design.

The PC prior $\pi_{PC}(d)$ plotted in the left panel of Figure 3 has been obtained numerically. Let $\pi_{PC}(\tau'_\beta)$ be the PC prior (3) evaluated at some predefined $\tau'_\beta > 0$ (a convenient approach is to take $\log(\tau'_\beta)$ on a regular grid inside $(-20, 20)$) and $d' = d(\tau'_\beta)$ be the associated degrees of freedom computed by (2). The induced PC prior evaluated at d' is $\pi_{PC}(d^{-1}(d'))\left|\frac{\partial d^{-1}(d')}{\partial d'}\right|$.

4.3 Behaviour near the base model

According to Simpson et al. (2014), the prior $\pi(\delta)$, where δ is the distance from the base model, is said to overfit, or to force overfitting, if $\pi(0) = 0$. Theorem 1 in Simpson et al. (2014) states that, if $\pi(\tau_\beta)$ is an absolutely continuous prior for the IGMRF precision τ_β, with $E(\tau_\beta) < \infty$, this prior overfits. The commonly used Gamma(a, b), $a, b > 0$, with $a/b < \infty$ falls in this class of overfitting priors.

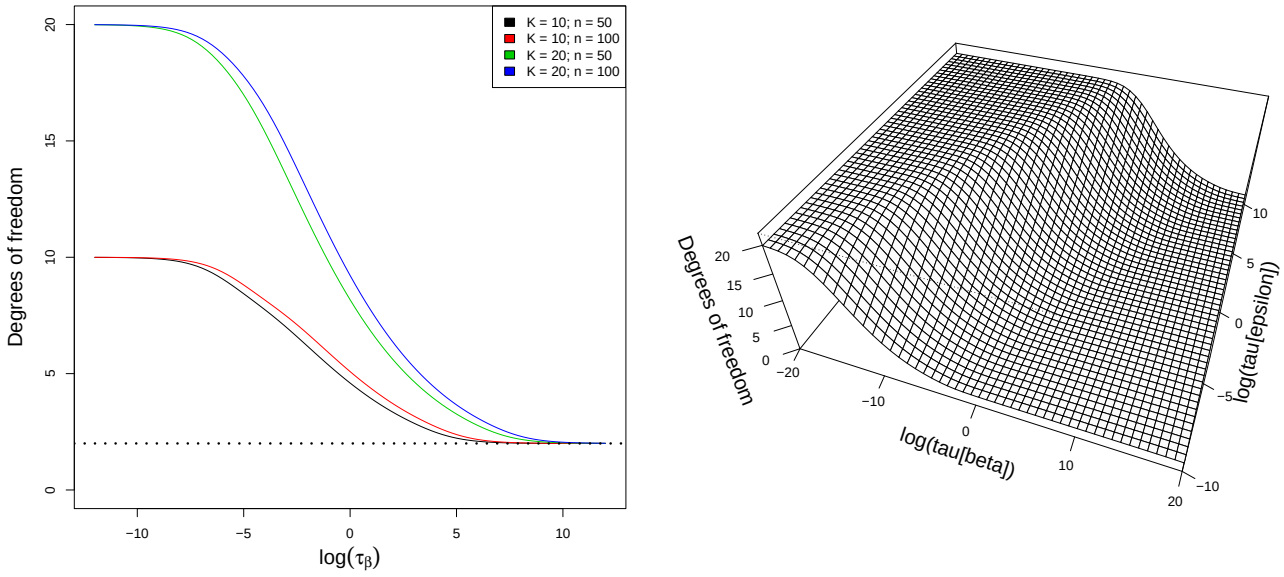

Figure 2: Mapping the degrees of freedom. The plot on the left shows the mapping d in the $\log(\tau_\beta)$ scale, conditional on $\tau_\epsilon = 1$, for four designs (choices of K and n). The dotted horizontal line at $d = 2$ indicates the base model (assuming an IGMRF of order 2 on the spline coefficients). The plot on the right shows d as a function of both τ_ϵ and τ_β, for the specific design $\{K = 20, n = 50\}$.

PC priors avoid over-fitting by construction: by applying the first three principles outlined in section 3 the mode of a PC prior is always at the base model. The fourth principle essentially allows the user to specify the penalty for increasing distances from the base model. The following result describes the behaviour near the base model for both the PC prior for degrees of freedom and the distribution of the degrees of freedom induced by a Gamma(a, b) on τ_β, which we denote as $\pi_G(d)$. (In the result below, we consider $\theta = -\log(\alpha)\sqrt{d^{-1}(U)}$, where d is the mapping given a positive and finite τ_ϵ.)

Result. Let r be the dimension of the null space of $\boldsymbol{R}$ in (1), then $\pi_{PC}(d) \to \infty$ as $d \to r$, for $\theta > 0$, and $\pi_G(d) \to 0$ as $d \to r$, for $a, b > 0$.

The proof is given in appendix A. The density $\pi_{PC}(d)$ goes to infinity as approaching the base model, avoiding over-fitting. Instead, the Gamma-induced $\pi_G(d)$ does not prevent over-fitting as it repulses the base model. In Figure 4, the Gamma-induced priors with $a = 1, b = 5 \cdot 10^{-4}$ (left panel) and $a = 10^{-3}, b = 10^{-3}$ (right panel) are displayed under four different designs. These two different priors have different interpretations in terms of degrees of freedom: the first favours over-smoothing while the latter favours over-fitting. For both choices of a and b, the base model is repulsed at a different rate according to design. Indeed, for a and b fixed, the density $\pi_G(d)$ clearly changes with design: in general, Gamma priors are not invariant under design as a

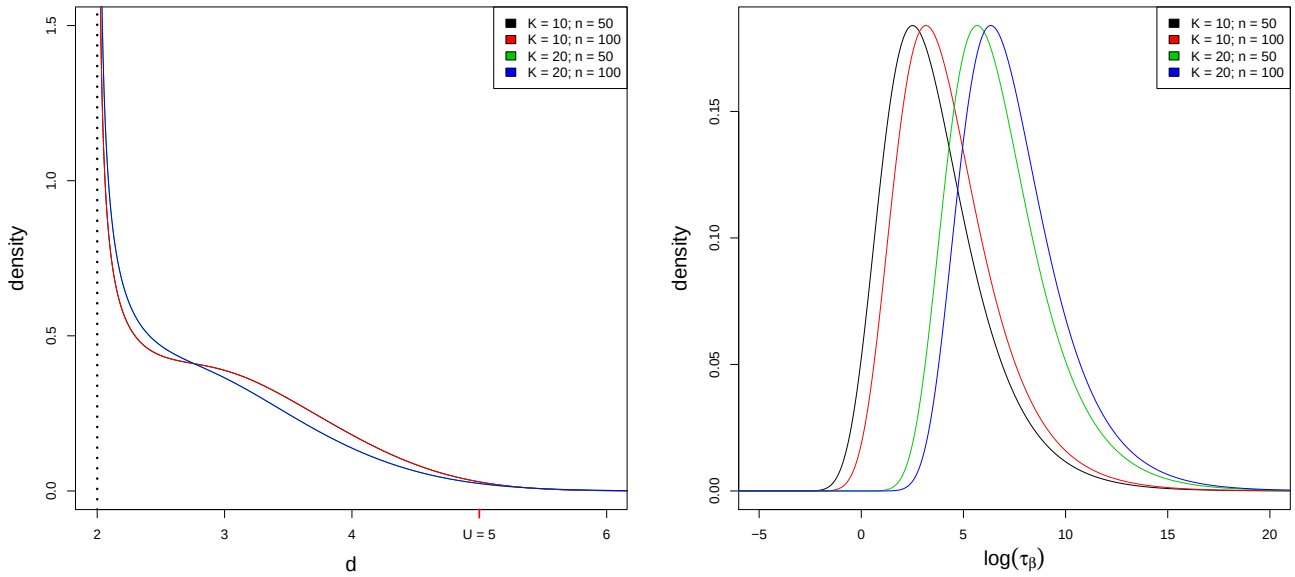

Figure 3: Invariance under design. Both panels show the PC prior density $\pi_{PC}(d)$, for $(U = 5, \alpha = 0.01)$ and $\tau_\epsilon = 1$, in two different scales: d (left panel) and $\log(\tau_\beta)$ (right panel). In the left panel, the dotted vertical line at $d = 2$ indicates the base model (assuming an IGMRF of order 2 on $\boldsymbol{\beta}$), while the red tick indicates the upper bound for degrees of freedom. Even if the PC prior density changes for varying K and n, the probability assigned to $d > 5$ is always 0.01. In this particular example, the density change between $n = 50$ and $n = 100$ is evident in the $\log(\tau_\beta)$ scale (right panel), but not in the d scale (left panel), where the black and red lines, as well as the green and blue lines, appear superimposed (however, they are not the same because the eigenvalues of $\boldsymbol{R}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}$ are different).

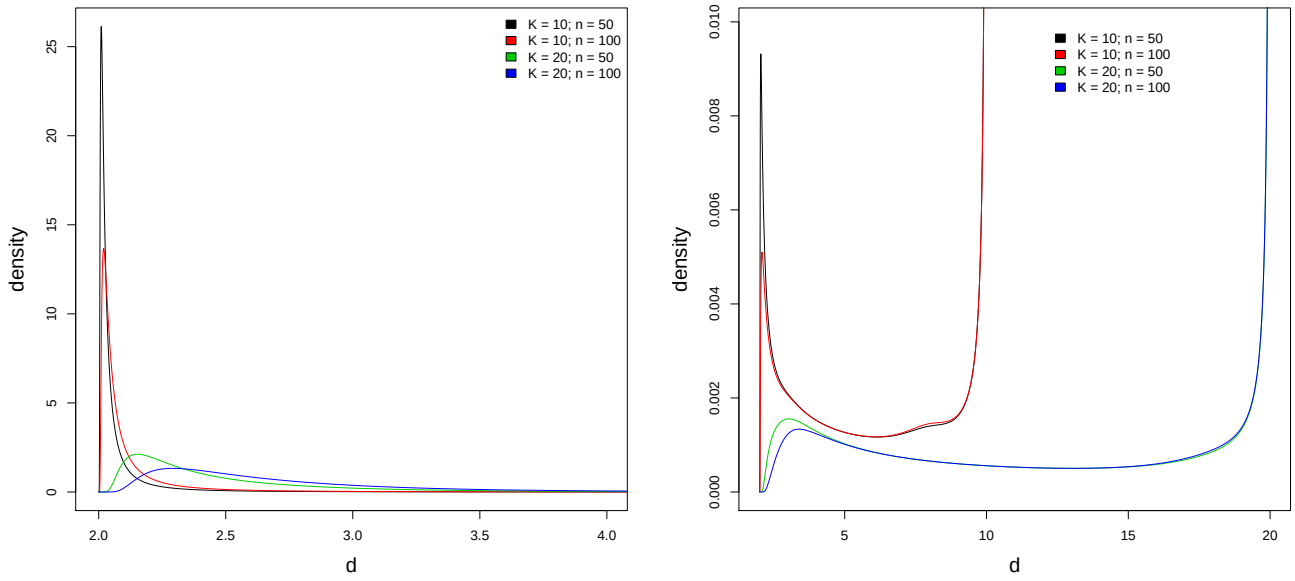

Figure 4: The distribution of the number of effective degrees of freedom induced by a Gamma prior with $a = 1, b = 5e-4$ (left panel) and $a = 10^{-3}, b = 10^{-3}$ (right panel), under four different designs. The base model is at $d = 2$, assuming an IGMRF prior of order 2 on $\boldsymbol{\beta}$.

consequence of the scaling issue discussed in section 2.1.

5 P-splines with a joint prior on $(\tau_\beta, \tau_\epsilon)$

So far we have worked under the assumption of known noise precision τ_ϵ. The number of effective degrees of freedom is a function of τ_β, which scales differently according to the level of noise present in the data (see the right panel in Figure 2). Knowing τ_ϵ is then crucial to scale the PC prior correctly, in order to guarantee the upper bound for degrees of freedom specified by the user.

The noise precision is typically unknown in applications. One could estimate τ_ϵ from the data and then specify the PC prior for d conditional on this estimate; this strategy has been proposed by Fong et al. (2010) to define Gamma-induced priors for degrees of freedom. We, instead, adopt a fully Bayesian model,

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \tau_\epsilon &\sim N(\boldsymbol{B}\boldsymbol{\beta}, \tau_\epsilon^{-1}\boldsymbol{I}) && (4)\\
\boldsymbol{\beta}|\tau_\beta &\sim N(0, \tau_\beta^{-1}\boldsymbol{R}^{-1}) && (5)\\
\tau_\beta|\tau_\epsilon &\sim \text{Gumbel}(1/2, \theta(\tau_\epsilon)) && (6)\\
\tau_\epsilon &\sim \pi(\tau_\epsilon) \propto 1/\tau_\epsilon, && (7)
\end{aligned}$$

where (6) and (7) specify a joint prior $\pi(\tau_\beta, \tau_\epsilon) = \pi_{PC}(\tau_\beta|\tau_\epsilon)\pi(\tau_\epsilon)$. The scaling parameter in (6) is given by $\theta(\tau_\epsilon) = -\log(\alpha)\sqrt{d^{-1}(U|\tau_\epsilon)}$, where $d(\cdot|\tau_\epsilon)$ is the mapping conditional on the random noise precision τ_ϵ. We use the improper $\pi(\tau_\epsilon) \propto 1/\tau_\epsilon$ since the data usually contain sufficient information with respect to τ_ϵ. The joint prior in (6) and (7) corresponds to the induced PC prior for d conditional on a random τ_ϵ, with $U \in (r, K)$ and $\alpha \in (0, 1)$ the parameters specifying the distribution.

We developed a Markov chain Monte Carlo (MCMC) algorithm to fit model (4) to (7); see pseudo-code reported in Appendix B algorithm 1. The algorithm includes a Metropolis-Hasting step to jointly update $(\tau_\epsilon, \tau_\beta, \boldsymbol{\beta})$. In our experience, block updating ensures good mixing and fast convergence of the proposed MCMC algorithms; see Rue and Held (2005) for details on block updating in hierarchical models with GMRF components. A brief description on how algorithm 1 works follows. At iteration j, both precision parameters (here denoted simply as τ) are sampled from the proposal distribution adopted in Knorr-Held and Rue (2002): $q(\tau^*|\tau^{(j-1)}) = t\tau^{(j-1)}$, where τ^* and $\tau^{(j-1)}$ are, respectively, the proposed and current values at iteration j, t is random with density $\pi(t) \propto 1 + 1/t$, for $t \in [1/T, T]$ and $T > 1$ is a tuning parameter; in our experience, setting T approximately equal to 1.5 works well in most applications. The proposal $q(\cdot)$ has the advantage that the ratio $q(\tau^*|\tau^{(j-1)})/q(\tau^{(j-1)}|\tau^*)$ equals one; for more details on this see Rue and Held (2005) ch. 4.2. Given τ_ϵ^* and τ_β^*, we draw $\boldsymbol{\beta}^*$ from the full conditional $\pi(\boldsymbol{\beta}|\tau_\epsilon^*, \tau_\beta^*, \boldsymbol{y})$. Within this scheme, the acceptance probability for $(\tau_\epsilon^*, \tau_\beta^*, \boldsymbol{\beta}^*)$ simplifies to (dropping the superscript $(j-1)$ to ease the notation)

$$\begin{aligned} a &= \min\left(1, \frac{\pi(\tau_\epsilon^*, \tau_\beta^*, \boldsymbol{\beta}^*|\boldsymbol{y})}{\pi(\tau_\epsilon, \tau_\beta, \boldsymbol{\beta}|\boldsymbol{y})} \frac{\pi(\boldsymbol{\beta}|\tau_\epsilon, \tau_\beta, \boldsymbol{y})}{\pi(\boldsymbol{\beta}^*|\tau_\epsilon^*, \tau_\beta^*, \boldsymbol{y})} \frac{q(\tau_\epsilon|\tau_\epsilon^*)}{q(\tau_\epsilon^*|\tau_\epsilon)} \frac{q(\tau_\beta|\tau_\beta^*)}{q(\tau_\beta^*|\tau_\beta)}\right) \\ &= \min\left(1, \frac{\pi(\tau_\epsilon^*, \tau_\beta^*|\boldsymbol{y})}{\pi(\tau_\epsilon, \tau_\beta|\boldsymbol{y})}\right), \end{aligned} \tag{8}$$

where $\pi(\tau_\epsilon^*, \tau_\beta^*|\boldsymbol{y}) \propto \pi(\boldsymbol{y}|\boldsymbol{\beta}^*, \tau_\epsilon^*)\pi(\boldsymbol{\beta}^*|\tau_\beta^*)\pi_{PC}(\tau_\beta^*|\tau_\epsilon^*)\pi(\tau_\epsilon^*)/\pi(\boldsymbol{\beta}^*|\tau_\epsilon^*, \tau_\beta^*, \boldsymbol{y})$. Computing the acceptance probability in (8) only requires the marginal for the precision parameters $(\tau_\epsilon, \tau_\beta)$ evaluated at the proposed and current values. Note that, rescaling $\pi_{PC}(\tau_\beta|\tau_\epsilon)$ according to the proposed τ_ϵ^* is needed before accepting/rejecting $(\tau_\epsilon^*, \tau_\beta^*, \boldsymbol{\beta}^*)$; this implies re-evaluating the inverse mapping $d^{-1}(\cdot|\tau_\epsilon)$, given τ_ϵ^*, and recomputing $\theta(\tau_\epsilon^*)$ at each iteration.

5.1 Additive P-splines

We now focus on an additive P-spline modelling framework, where the linear predictor is the sum of a number of smooth functions. Let $\boldsymbol{y}$ be a Gaussian response and $\boldsymbol{x}_j$, $j = 1, ..., J$, be a

set of J continuous covariates, the model is

$$\begin{aligned} \boldsymbol{y} &= \sum_{j=1}^{J} f_j(\boldsymbol{x}_j) + \boldsymbol{\epsilon} \quad ; \quad \boldsymbol{\epsilon} \sim N(0, \tau_\epsilon^{-1}) \\ f_j(\boldsymbol{x}_j) &= \boldsymbol{B}_j \boldsymbol{\beta}_j, \\ \boldsymbol{\beta_j} | \tau_{\beta_j} &\sim N(0, \tau_{\beta_j}^{-1} \boldsymbol{R}^{-1}) \end{aligned}$$

where $\boldsymbol{B}_j$ is the n x K_j B-spline basis matrix and $\boldsymbol{\beta}_j$ the vector of spline coefficients associated to the smooth function f_j; with no loss of generality, we consider the same number of knots $\forall j$, yielding $K = K_j$, $j = 1, ..., J$. We assume the joint prior $\prod_{j=1}^{J} \pi_{PC}(\tau_{\beta_j}|\tau_\epsilon)\pi(\tau_\epsilon)$, where $\pi(\tau_\epsilon) \propto 1/\tau_\epsilon$ and $\pi_{PC}(\tau_{\beta_j}|\tau_\epsilon) = \text{Gumbel}(1/2, \theta_j(\tau_\epsilon))$, $j = 1, ..., J$. The scaling parameter $\theta_j(\tau_\epsilon) = -\log(\alpha_j)\sqrt{d_j^{-1}(U_j|\tau_\epsilon)}$, where (U_j, α_j) are the parameters calibrating the induced PC prior for the degrees of freedom of f_j. The mapping for the degrees of freedom of f_j is given by $d_j(\tau_{\beta_j}|\tau_\epsilon) = \text{tr}\left\{\left(\boldsymbol{B_j}^\mathsf{T}\boldsymbol{B_j} + \frac{\tau_{\beta_j}}{\tau_\epsilon}\boldsymbol{R}\right)^{-1}\boldsymbol{B_j}^\mathsf{T}\boldsymbol{B_j}\right\}$.

Identifiability constraints are important in additive P-splines. The IGMRF prior on $\boldsymbol{\beta}_j$ controls deviations of the smooth term $\boldsymbol{B}_j\boldsymbol{\beta}_j$ from a polynomial base model; in the following, for simplicity, we consider an IGMRF of order 2 (which forces shrinkage towards a linear base model). All smooths $\boldsymbol{B}_j\boldsymbol{\beta}_j$ include the linear base model, thus they all compete to capture the mean of the data. To ensure identifiability we adopt the following re-parametrization,

$$\boldsymbol{y} = \mu + \sum_{j=1}^{J} \boldsymbol{x}_j\gamma_j + \sum_{j=1}^{J} \boldsymbol{B}_j\boldsymbol{\beta}_j^{ULC} + \boldsymbol{\epsilon}, \tag{9}$$

where μ is the intercept, γ_j is the slope coefficient for covariate $\boldsymbol{x}_j$ and $\boldsymbol{\beta}_j^{ULC}$ are the spline coefficients $\boldsymbol{\beta}_j$ under the two following linear constraints: $\left[\boldsymbol{c}^\mathsf{T}\boldsymbol{B}_j\right]\boldsymbol{\beta}_j = 0$ and $\left[\boldsymbol{l}^\mathsf{T}\boldsymbol{B}_j\right]\boldsymbol{\beta}_j = 0$, with *constant* vector $\boldsymbol{c} = \boldsymbol{1}_n$ and *line* vector $\boldsymbol{l} = [1, 2, ..., n]^\mathsf{T}$. In this way, the smooth term $\boldsymbol{B}_j\boldsymbol{\beta}_j^{ULC}$ captures residual variations from the linear base model $\mu + \gamma_j\boldsymbol{x}_j$. In other words, the constrained model (9) allows each smooth component to be identified, by separating the linear and flexible terms which coexist in each $\boldsymbol{B}_j\boldsymbol{\beta}_j$. Model (9) can be expressed in compact form as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{B}\boldsymbol{\beta}^{ULC} + \boldsymbol{\epsilon}, \tag{10}$$

where $\boldsymbol{B} = [\boldsymbol{B}_1 : ... : \boldsymbol{B}_j]$ is the n x (KJ) joint basis matrix, $\boldsymbol{\beta}^{ULC}$ is the joint vector of spline coefficients subject to the linear constraints, $\boldsymbol{X} = [\boldsymbol{1}_n : \boldsymbol{x}_1 : ... : \boldsymbol{x}_J]$ is the n x $(J+1)$ matrix of covariates with an additional column of ones for the intercept term and $\boldsymbol{\gamma}$ is the vector of fixed effects. We assume $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \tau_\gamma\boldsymbol{I}_{J+1})$ with a small precision, e.g. $\tau_\gamma = 10^{-4}$, as a prior for the

fixed effects. Other covariates can be added to $\boldsymbol{X}$ in model (10), if we assume them to have a simple linear effect.

We wrote a block updating MCMC algorithm implementing the joint prior in the model described above; pseudo-code is given in Appendix B algorithm 2. This includes Metropolis-Hasting steps to jointly update the blocks $(\tau_\epsilon, \boldsymbol{\gamma})$ and $(\tau_{\beta_j}, \boldsymbol{\beta}_j^{ULC})$, for each $j = 1, ..., J$; in our experience, this scheme gives good mixing (and convergence) properties. Algorithm 2 presents two main changes with respect to algorithm 1. First, rescaling the conditional PC prior $\pi_{PC}(\tau_{\beta_j}|\tau_\epsilon)$ is no longer necessary at each iteration; $\theta(\tau_\epsilon^*)$ must be recomputed only when $(\tau_\epsilon^*, \boldsymbol{\gamma}^*)$ is accepted. Second, the spline coefficients are sampled under linear constraints. To do this we use the algorithm proposed in Rue and Held (2005) ch. 2, which samples first the unconstrained coefficients and then *corrects* them for the constraints. To compute the acceptance probability for the candidate $(\tau_{\beta_j}^*, \boldsymbol{\beta}_j^{*ULC})$ in step (14) of algorithm 2 Appendix B, the full conditional density $\pi(\boldsymbol{\beta}^*|\tau_\beta^*, \boldsymbol{y})$ is evaluated at the constrained $\boldsymbol{\beta}_j^{*ULC}$; for computational details see Rue and Held (2005), formula (2.31).

6 Simulation study

The scaling parameter θ of the PC prior (3) can be tuned in several ways. Our proposal is to select θ through assumptions on the number of effective degrees of freedom, $d = h(\tau_\beta)$, of the function $f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\beta}$. This approach seems intuitive in the Gaussian case where quantity d relates immediately to the degree of an equivalent polynomial (which an expert user might have prior information about). Moreover, literature on smoothing often refers to degrees of freedom as a way to summarize model complexity. In a recent paper, Klein and Kneib (2015) specify θ through assumptions on the *scale*, or range of variation, of $f(\boldsymbol{x}) = \boldsymbol{B}\boldsymbol{\beta}$ and denote this PC prior as *scale-dependent prior* (SD prior). Precisely, θ is derived numerically by requiring that $Pr(|f(x)| \leq c) \geq 1 - \alpha$ for each x in the covariate domain, where $\alpha \in (0, 1)$ and c indicates an upper bound for the scale of f. Both scaling approaches lead to a PC prior which is invariant under design, in the sense that the computation of θ accounts for the adopted B-spline design $\boldsymbol{B}$. However, the two PC priors differ regarding conditioning on the noise variance. Our PC prior, $\pi_{PC}(d)$, is a joint prior $\pi(\tau_\beta|\tau_\epsilon) = \pi_{PC}(\tau_\beta|\tau_\epsilon)\pi(\tau_\epsilon)$, while the SD prior is defined unconditionally on τ_ϵ as the scale of f does not depend on τ_ϵ.

In this section we present a simulation study which investigates further the relevance of degrees of freedom for designing priors for Gaussian P-splines. The objective of our study is to assess the behaviour of our joint prior in scenarios with different noise levels, and to compare this with two alternative priors which are defined unconditionally on τ_ϵ, namely the conjugate Gamma prior and the SD prior. Therefore, our simulation study does not aim to generically

assess the behaviour of PC priors compared to standard priors. For an extended simulation study evaluating the performance of PC priors (in particular, SD priors) compared to several alternative hyper-priors for variance parameters, in both Gaussian and non Gaussian contexts, the reader is referred to Klein and Kneib (2015). Furthermore, our simulation is restricted to the Gaussian case.

We consider two different models regarding the shape of the true $f(x)$:

- $f_1(x) = \sin(x); \qquad x \in (-1, 1)$

- $f_2(x) = \cos(x); \qquad x \in (0, 2\pi)$

Model $f_1(x)$ is close to the base model (almost a linear effect; this is the same model considered in Klein and Kneib (2015)), while $f_2(x)$ is a one cycle sinusoidal curve (highly non linear effect). In both scenarios, data are simulated as $\boldsymbol{y} \sim N(f(\boldsymbol{x}), \tau_\epsilon^{-1}\boldsymbol{I})$, where covariate $\boldsymbol{x} = \{x_1, .., x_n\}$ takes values on a regular grid. We assume a standard P-spline model with one covariate, $\boldsymbol{y} \sim N(\boldsymbol{B}\boldsymbol{\beta}, \tau_\epsilon^{-1}\boldsymbol{I})$, $\boldsymbol{\beta} \sim N(0, \tau_\beta^{-1}\boldsymbol{R}^{-1})$, where $\boldsymbol{R}$ is the structure of an IGMRF of order 2, and $\boldsymbol{B}$ contains K cubic B-splines evaluated at $\boldsymbol{x}$.

Different scenarios are generated by setting: $n = \{20, 50\}$ (small and moderate sample sizes), $K = \{20, 30\}$, $\tau_\epsilon = \{0.25, 1, 5\}$ (high, moderate and low noise). We aim to assess the model fit obtained by the following priors:

- conjugate Gamma prior on τ_β, with two specifications widely used in applications: Gamma$(0.001, 0.001)$ and Gamma$(1, 5e-04)$.

- our joint prior $\pi_{PC}(\tau_\beta|\tau_\epsilon)\pi(\tau_\epsilon)$, inducing a PC prior for degrees of freedom, $\pi_{PC}(d)$, with parameters U and α. We set $\alpha = 0.01$ and various upper bounds $U = \{2, 3, 5, 7, 10\}$. Note that we specify $U = 2$ only to check consistency of results in the limit case where any deviation from the base model is strongly penalized (in applications, this is not a sensible choice as it forces the fit towards a linear trend; for more details see the the joint prior in action with simulated data in the supplemental material).

- SD prior on τ_β (Klein and Kneib, 2015), with $\alpha = 0.01$ and three specifications for the scale of f, $c = \{1.5, 2, 3\}$. Note that, since both f_1 and f_2 vary within $(-1, 1)$, $c = 1.5$ seems the most sensible choice as an upper bound for the scale of both functions, resulting in a sufficiently flexible prior, while $c = \{2, 3\}$ leads to an even more flexible prior. However, from Table 1 we see that the degrees of freedom implied by an SD prior with parameter c strongly depend on the noise present in the data (and to some extent on the adopted

B-spline design	High noise ($\tau_\epsilon = 0.25$)			Moderate noise ($\tau_\epsilon = 1$)			Low noise ($\tau_\epsilon = 5$)		
	$c = 1.5$	$c = 2$	$c = 3$	$c = 1.5$	$c = 2$	$c = 3$	$c = 1.5$	$c = 2$	$c = 3$
$n = 20$; $K = 20$	2.72	2.99	3.41	3.41	3.78	4.36	4.54	5.08	5.89
$n = 50$; $K = 20$	3.11	3.44	3.95	3.95	4.40	5.10	5.31	5.95	6.90
$n = 20$; $K = 30$	2.70	2.95	3.44	3.39	3.75	4.43	4.54	5.06	6.04
$n = 50$; $K = 30$	3.09	3.40	4.00	3.93	4.37	5.20	5.34	5.98	7.16

Table 1: Implied degrees of freedom, d, for the SD prior. The entries in the table refer to the upper bound, U, for d, obtained by assuming an SD prior with parameters c and $\alpha = 0.01$, in the different simulation scenarios. The computation of U involves the use of the `sdPrior` `R` package (Klein, 2015); for more details see the supplemental material.

design): for instance, $c = 1.5$ implies an upper bound for d around 2.72 in the high noise case (for the design $n = 20$, $K = 20$) which results into a very restrictive prior in fact.

6.1 Results

In each scenario, goodness of fit was assessed for each of 1000 simulated datasets by the mean squared error (MSE) of $\hat{f}(\boldsymbol{x}) = \boldsymbol{B}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the posterior mean, as $MSE = n^{-1}\sum_{i=1}^{n}(\hat{f}(x_i) - f(x_i))^2$. The posterior for $\boldsymbol{\beta}$ was computed using INLA (Rue et al., 2009) for the Gamma and SD priors, and using MCMC algorithm 1 for the joint prior. For the sake of comparison between the three classes of priors, we assume $\pi(\tau_\epsilon) = \text{Gamma}(1, 5e-04)$ throughout the simulation study (hence, our joint prior is $\pi_{PC}(\tau_\beta|\tau_\epsilon)\text{Gamma}(1, 5e-04)$).

Figure 5 reports $\log(MSE)$ for small sample size $n = 20$ (which is when the hyper-prior is expected to be most influential on the posterior) and $K = 20$. We do not see much change for increasing n and K, thus results for other scenarios are only reported in the supplemental material. Our main findings are:

1. The Gamma priors are generally outperformed by the two PC priors (both for degrees of freedom and scale), unless the data are very informative about the true model (e.g. in model f_2 with low noise, see bottom-right panel in Figure 5). As expected, the Gamma$(10^{-3}, 10^{-3})$ overfits when model f_1 is the true one (see left panels in Figure 5), while the Gamma$(1, 5e-04)$ performs poorly in scenario f_2 (especially with high noise, see top-right panel in Figure 5).

2. For sensible choices of the upper bound U the joint prior performs better than, or at least as good as, the SD prior. The main difference is noticed in scenario f_2 with high noise (top-right panel in Figure 5) and scenario f_1 with low noise (bottom-left panel in Figure 5). In the former, setting the joint prior with $U = \{5, 7, 10\}$ outperforms most SD prior specifications; in particular, the SD prior with $c = \{1.5, 2\}$ achieves poor performance

because the implied upper bound for d is too small (i.e. below 3, see Table 1) for this case, where the true effect is highly non linear and the data are very noisy. The SD prior with $c = 3$ gives similar performance because it implies a larger upper bound for degrees of freedom. Instead, in scenario f_1 with low noise, SD priors are outperformed by the joint prior with $U = \{2, 3\}$, because any choice $c = \{1.5, 2, 3\}$ implies an upper bound for d clearly larger than needed (i.e. above 4, see Table 1) in this case, where data are very informative and the true effect is close to the base model.

3. In cases where the choice of the upper bound U is inappropriate to describe the complexity of the true function f, the joint prior performs poorly. For instance, when the true curve is close to the base model (i.e. scenario f_1), the joint prior with $U \geq 5$ is outperformed by both the SD priors and the Gamma$(1, 5e-04)$ (see left panels in Figure 5); in other words, when data are generated under a linear model, a joint prior with large upper bound U will result in a bad choice, as it allows *far more* degrees of freedom than needed. For the same reasons, when the true curve is highly non linear (i.e. scenario f_2), the joint prior with $U = \{2, 3\}$ generally achieves poor performance (see the right panels in Figure 5), because it assigns almost all weight to the base model or close to it.

In summary, the simulation study showed that both PC priors (i.e. the SD prior and our joint prior) provide potentially better performance than the standard Gamma priors, for sensible choices of the scaling parameter θ. None of these PC priors shows uniformly better performance than the other one. However, we think that the particular scaling approach used to tune θ plays an important role. Scaling the PC prior in terms of degrees of freedom automatically accounts for the level of noise present in the data. We were able to show via simulation that our joint prior has the potential to perform well in general, i.e. both in high and low noise cases, provided that the elicited U is an appropriate upper bound for the degrees of freedom of the function f underlying the data. Therefore, we conclude that the number of effective degrees of freedom is a relevant quantity for building PC priors for smoothing Gaussian data, more suitable than other transformations of τ_β that do not depend on τ_ϵ.

7 Application

We demonstrate the use of the joint prior within an additive P-spline framework for modelling nitrate concentration in river *Oglio*, Lombardia region, Italy. A total of $n = 576$ observations of $NO3^-$ concentration were collected during 2010-2012 by taking one sample at each season (spring, summer, autumn and winter) in 48 gauging stations located along the river catchment. The response variable is $\log(\texttt{NO3}^-_{ij})$, measured at station $i = 1, ..., 48$ and season $j = 1, ..., 4$. Covariate

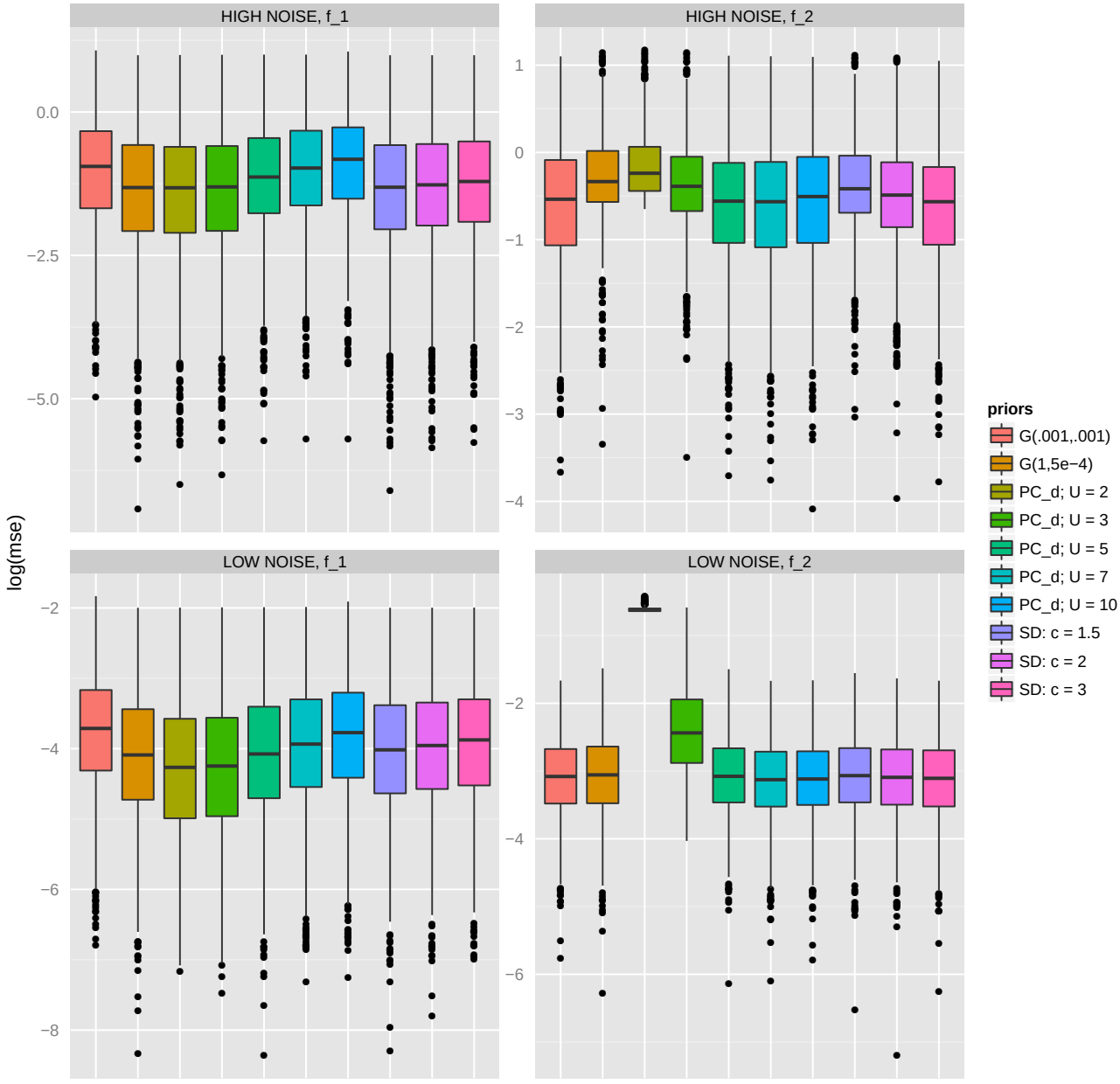

Figure 5: Simulation results: $\log(MSE)$ for f_1 (left panels) and f_2 (right panels), in presence of high noise ($\tau_\epsilon = 0.25$, top panels) and low noise ($\tau_\epsilon = 5$, bottom panels), sample size $n = 20$, $K = 20$. In the legend on the right, label "G" indicates the Gamma prior; "PC_d" indicates our PC prior for degrees of freedom (joint prior), with $\alpha = 0.01$ and $U = \{2, 3, 5, 7, 10\}$; "SD" denotes scale dependent prior with $\alpha = 0.01$ and $c = \{1.5, 2, 3\}$.

$\texttt{stream}_i$ is the distance from each station i to the *Iseo* lake (i.e., the river source) measured in km along the stream network; the first station in proximity to the lake is at $\texttt{stream} \approx 0$, while the last station downstream the river is at $\texttt{stream} \approx 150$. The goal is to understand *river enrichment* in terms of nitrates, by studying the behaviour of $\log(\texttt{NO3}^-)$ as the stream distance increases. A substantial amount of information comes from previous studies, see Delconte et al. (2014); Bartoli et al. (2012) and references therein, suggesting that river enrichment in terms of nitrates may vary non linearly as the stream distance increases. Due to different characteristics in terms of groundwater interactions and irrigation practices, the river catchment can be divided into upstream, middle and downstream reach. Different processes are expected within the three reaches and between seasons, hence the enrichment curve may show different shapes in the three river segments and seasons.

In order to investigate possible seasonal effects on river enrichment, we adopt the model: $\log(\texttt{NO3}^-_{ij}) = \mu + \gamma_j + f_j(\texttt{stream}_i) + \epsilon_{ij}$, $i = 1, ..., 48$, $j = 1, ..., 4$, where μ is the overall intercept, γ_j is the season-specific intercept and $f_j(\texttt{stream}_i)$ is the season-specific smooth function of the stream distance, modelled with a P-spline with joint prior on $(\tau_{\beta_j}, \tau_\epsilon)$ as described in section 5,

$$\begin{aligned} f_j(\texttt{stream}) &= \boldsymbol{B}_j\boldsymbol{\beta}_j \qquad\qquad j = 1, ..., 4 \\ \boldsymbol{\beta_j}|\tau_{\beta_j} &\sim N(0, \tau_{\beta_j}^{-1}\boldsymbol{R}^{-1}) \\ \tau_{\beta_j}|\tau_\epsilon &\sim \text{Gumbel}(1/2, \theta_j(\tau_\epsilon)) \\ \tau_\epsilon &\sim \pi(\tau_\epsilon) \propto 1/\tau_\epsilon. \end{aligned} \tag{11}$$

We assume an IGMRF of order 2 with precision τ_{β_j} on each $\boldsymbol{\beta}_j$. Based on our prior information we specify an upper bound $U = 8$ for the PC prior (11), for all $j = 1, ..., 4$, assuming that each f_j is much more flexible than linear (i.e. $d > 2$) and assigning 2 additional degrees of freedom to each of the three river segments, to capture possibly different enrichment behaviours. We believe this prior is flexible enough to describe the possible smooth change between the upstream middle and downstream behaviours. We set $\alpha = 0.01$, saying that it is 1% likely that f_j is more flexible than 8 degrees of freedom, $j = 1, ..., 4$.

To construct the matrices $\boldsymbol{X}$ and $\boldsymbol{B}$ of model (10), we need to create suitable dummy vectors of length n: $\texttt{dummy}_j$, $j = 2, ..., 4$, taking value 1 when the observation is from season j and 0 elsewhere (these are associated to the season-specific intercepts); $\texttt{stream} * \texttt{dummy}_j$, $j = 1, ..., 4$, taking the actual stream distance when the observation is from season j and 0 elsewhere (these are associated to the season-specific slopes). The n x 8 fixed effect design matrix is $\boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{D}, \boldsymbol{S}]$, with $\boldsymbol{D} = [\texttt{dummy}_2, ..., \texttt{dummy}_4]$ and $\boldsymbol{S} = [\texttt{stream} * \texttt{dummy}_1, ..., \texttt{stream} * \texttt{dummy}_4]$. Each basis $\boldsymbol{B}_j$ contains $K = 30$ cubic B-splines, evaluated on the season-specific stream distances $\texttt{stream} * \texttt{dummy}_j$. The n x $(4K)$ B-spline design matrix is $\boldsymbol{B} = [\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3, \boldsymbol{B}_4]$. To separate the season-specific slopes

and intercepts from the season-specific smooth variation captured by the B-splines, suitable linear constraints must be applied to each f_j, as discussed in section 5.1. We fit this model via MCMC based on the pseudo-code reported in Appendix B algorithm 2; in particular, we use a two blocks updating scheme, which ensured good mixing: first block is $(\tau_\epsilon, \boldsymbol{\gamma})$ and the other one contains all the four sets of spline coefficients and associated precisions, $(\tau_{\beta_1}, ..., \tau_{\beta_4}, \boldsymbol{\beta}_1^{ULC}, ..., \boldsymbol{\beta}_4^{ULC})$. During MCMC iterations, the PC prior for each precision τ_{β_j} needs to be re-scaled according to the current τ_ϵ, but this can be done at a negligible computational cost. The R code is provided as supplemental material.

The results displayed in Figure 6 reveal different river enrichment curves ($\hat{f}_j(\texttt{stream})$) in different seasons: a distinctive pattern is observed in summer, where the fitted curve shows a fast increase upstream ($\texttt{stream} < 50\ km$) and tendency to decrease downstream the river ($\texttt{stream} > 80\ km$). This pattern supports the argument given in Delconte et al. (2014) about an upstream reach (from 0 to 25 km) where nitrates are stable, reflecting the chemistry of the lake; a middle reach (from 25 to 80 km) showing an increase of $NO3^-$ concentration, probably due to groundwater inputs as replacement of river water abstracted for irrigation; a downstream reach (from 80 to 150 km) where nitrates should remain constant or even decrease, mainly due to the dilution of river water with $NO3^-$ deprived inputs.

8 Discussion

PC priors are defined to penalize complexity with respect to a given base model, the magnitude of the penalty being elicited by the user using an intuitive scaling approach. The scaling tool allows the user to derive the PC prior on an interpretable scale, different from the scale of the original parameter, provided that the link between the two is known. We took advantage of this nice feature and derived PC priors for the number of effective degrees of freedom d of a P-spline model for Gaussian data.

For non Gaussian responses, the idea presented in this paper follows straightforwardly by assuming the definition of the degrees of freedom of a generalized P-spline model (Hastie and Tibshirani, 1990), $d = \text{tr}\left(\boldsymbol{B}^\intercal \boldsymbol{W} \boldsymbol{B} + \frac{\tau_\beta}{\tau_\epsilon}\boldsymbol{R}\right)^{-1} \boldsymbol{B}^\intercal \boldsymbol{W} \boldsymbol{B}$, where $\boldsymbol{W}$ is a diagonal matrix, with entries depending on the linear predictor of the model (i.e. on $\boldsymbol{B}\boldsymbol{\beta}$) and the adopted link function. MCMC methods similar to those proposed in this paper can then be developed being careful about implementing the mapping $d(\tau_\beta|\tau_\epsilon, \boldsymbol{W})$, which is conditional on $\boldsymbol{W}$ in the generalized case (note that for most distributions in the exponential family τ_ϵ is known, e.g. for Poisson we have $\tau_\epsilon = 1$). For Poisson and Binomial responses a convenient approach is to use auxiliary variable methods (Frhwirth-Schnatter and Frhwirth, 2007; Frühwirth-Schnatter et al., 2008) and work with an equivalent *augmented* P-spline model for Gaussian (pseudo) data. The PC prior for d

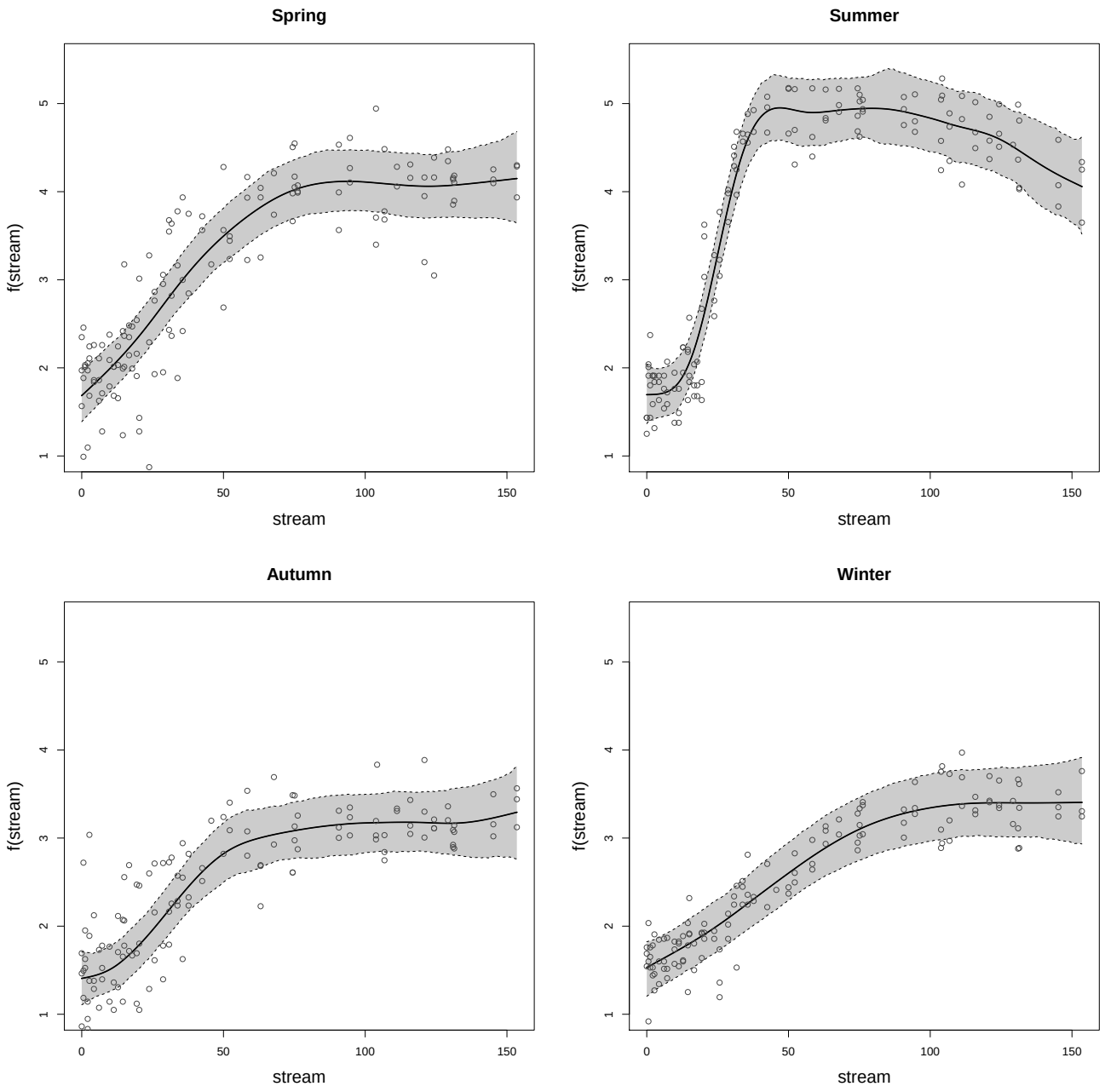

Figure 6: Estimated river enrichment curves ($\hat{f}_j(\texttt{stream})$, black line), for each season, and 95% credible bands (grey). The curve for summer shows clearly a distinctive pattern with respect to the other seasons. Partial residuals (Wood, 2006) are plotted in each panel (dots), indicating larger variability, than what assumed by the model, for the log $NO3^-$ concentrations observed in spring and autumn (with possible outliers in the upstream river segment in autumn).

can then be implemented in the same way as described in this paper, with algorithms 1 and 2 needing only the inclusion of a Gibbs-step to update the augmented parameters.

The potential advantages of using PC priors for degrees of freedom are twofold. First, they are *easy-to-elicit* by the user, who has to define two intuitive scaling parameters: U, an upper bound for d, and α, the prior mass assigned to $d > U$. This scaling tool can be handled flexibly. For instance, elicitation of the median M for the degrees of freedom results from fixing $\alpha = 0.5$. In this case, the PC prior density is bimodal: one mode is set at the base model (by definition) and another mode is set around M degrees of freedom. This bimodal behaviour is due to the attraction to the base model implicit in PC priors.

As a second advantage, these PC priors avoid overfitting and are invariant under design, which means that the parameters U and α do not need to be rescaled if the design changes. In other words, the PC prior is able to code into the model the prior knowledge on the complexity of the curve, or its degrees of freedom, in a design-adaptive way. The ability to adapt to design and avoid overfitting by construction, makes $\pi_{PC}(d)$ an appealing default choice in additive models where the latent structure includes several smooth functions (built on a basis of B-splines, e.g. P-splines) and other types of structures, such as individual random effects, spatial and spatio-temporal random effects.

Acknowledgements

Massimo Ventrucci is funded by a FIRB 2012 grant (project nr. RBFR12URQJ, title: Statistical modeling of environmental phenomena: pollution, meteorology, health and their interactions), for research projects of national interest provided by the Italian Ministry of Education, Universities and Research. The dataset used in section 7 was kindly provided by the *Consorzio dell'Oglio*, from the project "Experimental assessment of the environmental flow in the lower Oglio river". We thank Erica Racchetti and Alex Laini, Department of Life Science, University of Parma, for introducing us to the application in section 7 and for fruitful discussion on the ecological interpretation of results. Finally, we would like to thank the AE and two anonymous referees for their helpful comments.

References

Bartoli, M., Racchetti, E., Delconte, C. A., Sacchi, E., Soana, E., Laini, A., Longhi, D., and Viaroli, P. (2012). Nitrogen balance and fate in a heavily impacted watershed (Oglio River, Northern Italy): in quest of the missing sources and sinks. *Biogeosciences*, 9(1):361–373.

Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2000). BayesX software for Bayesian inference in structured additive regression models. Technical report. Available from `http://www.stat.uni-muenchen.de/~bayesx`.

Currie, I., Durbán, M., and Eilers, P. (2006). Generalized linear array models with applications to multidimensional smoothing. *J. R. Statist. Soc. B*, 68:259–280.

Delconte, C., Sacchi, E., Racchetti, E., Bartoli, M., Mas-Pla, J., and Re, V. (2014). Nitrogen inputs to a river course in a heavily impacted watershed: A combined hydrochemical and isotopic evaluation (Oglio River Basin, N Italy). *Science of The Total Environment*, 466467:924 – 938.

Eilers, P., Currie, I., and Durbán, M. (2006). Fast and compact smoothing on large multidimensional grids. *Computational Statistics & Data Analysis*, 5:61–76.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11:89–121.

Eilers, P. and Marx, B. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:637–653.

Fahrmeir, L. and Kneib, T. (2009). Propriety of posteriors in structured additive regression models: Theory and empirical evidence. *Journal of Statistical Planning and Inference*, 139:843–859.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *STATISTICA SINICA*, 14:715–745.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: models, methods and applications*. Springer-Verlag, Berlin.

Fong, Y., Rue, H., and Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412.

Frhwirth-Schnatter, S. and Frhwirth, R. (2007). Auxiliary mixture sampling with applications to logistic models. *Computational Statistics & Data Analysis*, 51(7):3509 – 3528.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2008). Improved auxiliary mixture sampling for hierarchical models of non-gaussian data. *Statistics and Computing*, 19(4):479–492.

Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154(1):85 – 100.

Frühwirth-Schnatter, S. and Wagner, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In *J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.)*, pages 165–200. Bayesian Statistics 9, Oxford.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Anal.*, 1(3):515–534.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag New York.

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.

Klein, N. (2015). *sdPrior: Scale-Dependent Hyperpriors in Structured Additive Distributional Regression*. R package version 0.3.

Klein, N. and Kneib, T. (2015). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for diasease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

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

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models using inte- grated nested laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71(2):319–392.

Ruppert, D., Wand, P., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Simpson, D. P., Rue, H., Martins, T. G., Riebler, A., and Sørbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *ArXiv e-prints*.

Srbye, S. H. and Rue, H. (2014). Scaling intrinsic gaussian markov random field priors in spatial modelling. *Spatial Statistics*, 8:39 – 51. Spatial Statistics Miami.

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics. Springer-Verlag New York.

Wood, S. (2006). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.

A Proofs

We show that $\pi_{PC}(d) \to \infty$ as $d \to r$, for $\theta > 0$. To simplify notation in the following we write τ instead of τ_β. Let us consider degrees of freedom d expressed as a function of τ, given a positive and finite τ_ϵ and assuming $v_1' < v_2' < ... < v_{K-r}'$ are the ordered positive eigenvalues of $\tau_\epsilon^{-1}\boldsymbol{R}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B})^{-1}$,

$$\begin{aligned} d(\tau) &= r + \sum_{k=1}^{K-r} \frac{1}{1+\tau v_k'} \\ &= r + \frac{1}{1+\tau v_1'} \cdot \left(1 + \frac{1/\tau + v_1'}{1/\tau + v_2'} + ... + \frac{1/\tau + v_1'}{1/\tau + v_{K-r}'}\right) \end{aligned} \tag{12}$$

When $\tau \to \infty$, the term inside the bracket on the right hand side of equation (12) is a constant, then d behaves like $r + \frac{1}{1+\tau v_1'}$. Since $d \to r$ if and only if $w \to 0$, $w = \frac{\sigma^2}{\sigma^2 + v_1'}$, it is sufficient to study the behaviour of $\pi_{PC}(w)$ when $w \to 0$. By a change of variable, $\pi_{PC}(w) = \theta \exp\left(-\theta \cdot const\sqrt{w}\right)\left|\frac{const}{2}\frac{1}{\sqrt{w}}\right|$, for positive σ (using the fact that $\pi_{PC}(\sigma)$ is exponential with rate θ; see Simpson et al. (2014)). When $w \to 0$, $\pi_{PC}(w) \propto \frac{1}{\sqrt{w}} = \infty$, for $\theta > 0$, which completes the proof.

We now prove that $\pi_G(d)$ goes to 0 as $d \to r$, for $a, b, > 0$. Using the same argument as in the proof above, it is sufficient to check that, for positive w, $\pi_G(w) \propto w^{-a-1}\exp(-b/w) \to 0$ as $w \to 0$, because $\lim_{w \to 0^+} \exp(-b/w) = 0$.

B MCMC

Algorithm 1 MCMC for fitting a P-spline model $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, assuming the joint prior $\pi_{PC}(\tau_\beta|\tau_\epsilon)\pi(\tau_\epsilon)$. (In step 5, N_C indicates the canonical parametrization of a GMRF, see Rue and Held (2005) chapter 2.2.)

1: Initialise $\tau_\epsilon^{(0)}$, $\tau_\beta^{(0)}$, $\boldsymbol{\beta}^{(0)}$
2: **for** $j = 1..., \text{n.iter}$ **do**
3: sample $\tau_\epsilon^* \sim q(\tau_\epsilon^*|\tau_\epsilon^{(j-1)})$
4: sample $\tau_\beta^* \sim q(\tau_\beta^*|\tau_\beta^{(j-1)})$
5: sample $\boldsymbol{\beta}^* \sim N_C(\tau_\epsilon^*\boldsymbol{B}^\mathsf{T}\boldsymbol{y}, \tau_\epsilon^*\boldsymbol{B}^\mathsf{T}\boldsymbol{B} + \tau_\beta^*\boldsymbol{R})$
6: Rescale $\pi_{PC}(\tau_\beta|\tau_\epsilon)$, according to the proposed τ_ϵ^*:

- evaluate the inverse mapping $d^{-1}(U|\tau_\epsilon^*)$ and compute $\theta(\tau_\epsilon^*)$, the rescaled PC prior is a Gumbel$(1/2, \theta(\tau_\epsilon^*))$

7: sample $r \sim U(0,1)$
8: **if** $r < \min\left(1, \frac{\pi(\tau_\epsilon^*, \tau_\beta^*|\boldsymbol{y})}{\pi(\tau_\epsilon^{(j-1)}, \tau_\beta^{(j-1)}|\boldsymbol{y})}\right)$ **then**
9: $\tau_\epsilon^{(j)} \leftarrow \tau_\epsilon^*$
10: $\tau_\beta^{(j)} \leftarrow \tau_\beta^*$
11: $\boldsymbol{\beta}^{(j)} \leftarrow \boldsymbol{\beta}^*$
12: $\theta(\tau_\epsilon^{(j)}) \leftarrow \theta(\tau_\epsilon^*)$
13: **end if**
{note: if $(\tau_\epsilon^*, \tau_\beta^*, \boldsymbol{\beta}^*)$ are not accepted, then all current parameters (those with superscript $(j-1)$) are repeated in the chain, i.e. $\tau_\epsilon^{(j)} = \tau_\epsilon^{(j-1)}$, $\tau_\beta^{(j)} = \tau_\beta^{(j-1)}$ and so on (we skip this here, to reduce the number of lines of code). The same reasoning holds for each acceptance step of algorithm 2 below.}
14: **end for**

Algorithm 2 MCMC for fitting an additive P-spline model under linear constraints, $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{B}\boldsymbol{\beta}^{ULC} + \boldsymbol{\epsilon}$, assuming the joint prior $\pi_{PC}(\tau_\beta|\tau_\epsilon)\pi(\tau_\epsilon)$. For simplicity, the pseudo-code considers only one covariate, hence, following notation of section 5.1: $J = 1$, $X = [\mathbf{1}, x_1]$, $\boldsymbol{B} = [\boldsymbol{B}_1]$. Extension to $J > 1$ only requires to repeat step 10 and steps 12:17 for each smooth $j = 1, ..., J$. Another option would be to jointly update all J sets of spline coefficients and associated precision in one block.

1: define 2 x K matrix $\boldsymbol{A} = \left[(\boldsymbol{c}^\mathsf{T}\boldsymbol{B})^\mathsf{T}, (l^\mathsf{T}\boldsymbol{B})^\mathsf{T}\right]^\mathsf{T}$ with 2 linear constraints (note: in our experience it is convenient to rescale $\boldsymbol{A}$, dividing each row by its maximum)
2: Initialise $\tau_\epsilon^{(0)}$, $\tau_\beta^{(0)}$, $\boldsymbol{\gamma}^{(0)}$, $\boldsymbol{\beta}^{(0)}$; τ_γ is fixed
3: **for** $j = 1..., \text{n.iter}$ **do**
4: sample $\tau_\epsilon^* \sim q(\tau_\epsilon^*|\tau_\epsilon^{(j-1)})$
5: sample $\boldsymbol{\gamma}^* \sim N_C(\tau_\epsilon^*\boldsymbol{X}^\mathsf{T}\tilde{\boldsymbol{y}}, \tau_\epsilon^*\boldsymbol{X}^\mathsf{T}\boldsymbol{X} + \tau_\gamma\boldsymbol{I}_{J+1})$, with $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{B}\boldsymbol{\beta}^{ULC(j-1)}$
6: sample $r \sim U(0, 1)$
7: **if** $r < \min\left(1, \frac{\pi(\tau_\epsilon^*|\boldsymbol{y})}{\pi(\tau_\epsilon^{(j-1)}|\boldsymbol{y})}\right)$ **then**
8: $\tau_\epsilon^{(j)} \leftarrow \tau_\epsilon^*$
9: $\boldsymbol{\gamma}^{(j)} \leftarrow \boldsymbol{\gamma}^*$
10: Rescale $\pi_{PC}(\tau_\beta|\tau_\epsilon)$, according to the current $\theta(\tau_\epsilon^{(j)})$:

- evaluate the inverse mapping $d^{-1}(U|\theta(\tau_\epsilon^{(j)})$ and compute $\theta(\tau_\epsilon^{(j)})$, the rescaled conditional PC prior is a Gumbel$(0.5, \theta(\tau_\epsilon^{(j)}))$

11: **end if**
12: sample $\tau_\beta^* \sim q(\tau_\beta^*|\tau_\beta^{(j-1)})$
13: sample $\boldsymbol{\beta}^*|\boldsymbol{A}\boldsymbol{\beta} = \mathbf{0}$, using algorithm 2.6 in Rue and Held (2005), steps are:

- sample $\boldsymbol{\beta}^* \sim N_C(\tau_\epsilon^{(j)}\boldsymbol{B}^\mathsf{T}\tilde{\boldsymbol{y}}, \tau_\epsilon^{(j)}\boldsymbol{B}^\mathsf{T}\boldsymbol{B} + \tau_\beta^*\boldsymbol{R})$, with $\tilde{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma}^{(j)}$
- $\boldsymbol{\beta}^{*ULC} = \boldsymbol{\beta}^* - \boldsymbol{Q}^{-1}\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{Q}^{-1}\boldsymbol{A}^\mathsf{T})^{-1}\boldsymbol{A}\boldsymbol{\beta}^*$, with $\boldsymbol{Q} = \tau_\epsilon^{(j)}\boldsymbol{B}^\mathsf{T}\boldsymbol{B} + \tau_\beta^*\boldsymbol{R}$

14: **if** $r < \min\left(1, \frac{\pi(\tau_\beta^*|\boldsymbol{y})}{\pi(\tau_\beta^{(j-1)}|\boldsymbol{y})}\right)$ **then**
15: $\tau_\beta^{(j)} \leftarrow \tau_\beta^*$
16: $\boldsymbol{\beta}^{ULC(j)} \leftarrow \boldsymbol{\beta}^{*ULC}$
17: **end if**
18: **end for**