Bayesian Analysis (2014) **9**, Number 4, pp. 859–882

Adaptive Priors Based on Splines with Random Knots

Eduard Belitser * and Paulo Serra †

Abstract. Splines are useful building blocks when constructing priors on nonparametric models indexed by functions. Recently it has been established in the literature that hierarchical adaptive priors based on splines with a random number of equally spaced knots and random coefficients in the B-spline basis corresponding to those knots lead, under some conditions, to optimal posterior contraction rates, over certain smoothness functional classes. In this paper we extend these results for when the location of the knots is also endowed with a prior. This has already been a common practice in Markov chain Monte Carlo applications, but a theoretical basis in terms of adaptive contraction rates was missing. Under some mild assumptions, we establish a result that provides sufficient conditions for adaptive contraction rates in a range of models, over certain functional classes of smoothness up to the order of the splines that are used. We also present some numerical results illustrating how such a prior adapts to inhomogeneous variability (smoothness) of the function in the context of nonparametric regression.

Keywords: Adaptive prior, Bayesian non-parametric, optimal contraction rate, spline, random knots.

1 Introduction

The Bayesian approach in statistics has become quite popular in recent years as an alternative to classical *frequentist* methods. The main appeal of the Bayesian methodology is its conceptual simplicity: given a model for the observed data $X \sim P_f$, $f \in \mathcal{F}$, some space of functions, put a prior on the unknown parameter f and draw inferences based on the resulting posterior $\Pi(f|X)$. Knowledge about the model under study can also be incorporated into the inference procedure via the prior. However, some seemingly "correct" priors can lead to unreasonable posteriors, especially in nonparametric models. It is therefore desirable to place ourselves in a setting where it is possible to assess the quality of the resulting posterior from some objective point of view. This gave rise to the development of the notion of contraction rate (cf. Ghosal et al. 2000), a Bayesian analog of a convergence rate: data is assumed to come from a fixed probability measure $P_0 = P_{f_0}$ for a "true" $f_0 \in \mathcal{F}$; the contraction rate is then the smallest radius such that the posterior mass in a ball (with respect to an appropriate distance) of probability measures around P_0 converges to 1 in P_0-probability as some information index such as a sample size goes to infinity.

Some general results about posterior contraction rates establish sufficient conditions

*VU University Amsterdam e.n.belitser@vu.nl
†Georg-August-Universität Göttingen pdeandr@uni-goettingen.de

 DOI:10.1214/14-BA879

on prior distributions such that the resulting posteriors attain a certain contraction rate. In this spirit, when studying specific priors, some authors now choose to present their results in the form of say *meta-theorems* which claim that sufficient conditions (such as the ones in Ghosal et al. 2000) required to attain a certain range of contraction rates hold for their choice of prior; cf. de Jonge and van Zanten (2012), Shen and Ghosal (2012), van der Vaart and van Zanten (2008) and further references therein. We adopt this practice here as well.

In the case where f_0 is a function from some functional space of smoothness α, the posterior contraction rate is typically compared to the convergence rate of the minimax risk (called optimal rate) over that space in the estimation problem. For example, if we observe a sample of size n and want to estimate a univariate α-smooth function, the typical optimal rate (e.g., density or regression function from the Hölder class) is of order $n^{-\alpha/(2\alpha+1)}$, possibly up to a logarithmic factor depending on the risk function. If the smoothness parameter α is unknown, and one wants to build estimators which attain the optimal rate corresponding to α but do not depend explicitly on α, one speaks of an adaptation problem. In a Bayesian context, the adaptation problem consists in finding a prior which leads to the optimal posterior contraction rate (usually up to a logarithmic factor) for any α-smooth function of interest and does not depend on the smoothness parameter α. Such priors are called rate adaptive. There is a growing number of papers, where this problem has been studied in different settings; cf. de Jonge and van Zanten (2012), Shen and Ghosal (2012), van der Vaart and van Zanten (2008), van der Vaart and van Zanten (2009) and Belitser and Ghosal (2003) among others.

Splines, in particular, can be used when constructing adaptive priors. A spline (cf. de Boor 1978) is a piecewise polynomial function designed to have a certain level of smoothness which is referred to as its order. Splines are easy to store, differentiate, integrate and evaluate on a computer, and are extensively used in practice for constructing good, parsimonious approximations of smooth functions. The points at which the different polynomial pieces of a spline connect are called knots. If an order (read: maximal polynomial degree) and a set of knots is fixed, then the space of all splines with that order and those knots forms a linear space which admits a basis of so called B-splines. Any spline of a fixed order is consequently characterized by a set of knots and its coordinates in the B-splines basis corresponding to those knots. Randomly generating a number of knots and, given those, generating random coordinates in the corresponding B-spline basis with equally spaced knots results in a random spline whose law can be used as a prior. If, given the number of knots, the coordinates in the corresponding B-spline basis are chosen to be independent and normally distributed, then the resulting spline has a conditionally Gaussian law and was studied by de Jonge and van Zanten (2012) by using Reproducing Kernel Hilbert Space techniques. Shen and Ghosal (2012) propose a more general, random series prior: the coefficients in the series are not necessarily independent or Gaussian and a basis other than the B-spline basis can also be used.

The case where the locations of the knots are also random is not covered by the results of either de Jonge and van Zanten (2012) or Shen and Ghosal (2012). However when practitioners put a prior on the number of knots they almost invariably also

put a prior on the locations of the knots (e.g., Denison et al. 1998, Di Matteo et al. 2001, Sharef et al. 2010) – a Poisson process is a popular choice. Their motivation for allowing arbitrarily located knots seems to be twofold. Firstly, this is attractive from the implementation point of view: designing reversible jump Markov chain Monte Carlo (MCMC) samplers is much simpler if any collection of knots is allowed since new knots can be inserted at arbitrary positions causing only localized changes in the spline. Secondly, the resulting posterior based on the prior with random locations of the knots is expected to be more adaptive with respect to inhomogeneous smoothness of the function of interest: the function may not have a fixed level of smoothness throughout its support, it may consist of rough and smooth pieces. To sustain an adequate level of accuracy over the whole support, more knots are needed in rough pieces and less in smooth ones. Therefore, to make it at least possible for the resulting posterior to pick up eventual spatial features of the function, the prior has to be flexible enough to model random locations of the knots.

In this paper, we extend the results of de Jonge and van Zanten (2012), and those of Shen and Ghosal (2012) with respect to the prior with random knots: we add one more level to the hierarchical spline prior by putting a prior on the location of the knots of the spline as well, making, in fact, the basis functions also random. Under some mild assumptions on the proposed hierarchical spline prior, we establish our main result for the proposed prior, providing sufficient conditions for adaptive, optimal contraction rates of the resulting posterior in a range of models (among others: density estimation, nonparametric regression, binary regression, Poisson regression, and classification). In doing so, we provide a theoretical basis for the common practice of using randomly located knots in spline based priors. Another interesting feature of a prior with random knot locations is that it leads to the posterior of the knots vector which provides (some sort of empirical Bayes) inference on the variability (smoothness inhomogeneity) of the underlying function. We present some numerical results illustrating how such a prior adapts to inhomogeneous variability (smoothness) of the function in the context of nonparametric regression.

2 Notation and preliminaries on splines

First we introduce some notation. For $d \in \mathbb{N}$ and $1 \le p < \infty$ denote by $\|\boldsymbol{x}\|_p = \big(\sum_{i=1}^d |x_i|^p\big)^{1/p}$ the l_p-norm of $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and by $\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,d} |x_i|$. For $1 \le p < \infty$ let the L_p-norm of a function f on [0,1] be $\|f\|_p = \big(\int_0^1 |f(x)|^p \, dx\big)^{1/p}$ and $\|f\|_\infty = \sup_{x\in[0,1]} |f(x)|$.

We use $\lesssim$ (respectively $\gtrsim$) to denote smaller (respectively greater) or equal up to a constant, the symbols $a \vee b$ and $a \wedge b$ stand for $\max\{a, b\}$ and $\min\{a, b\}$ respectively. The covering number $N(\epsilon, S, d)$ of a subset S of a metric space with balls of size ϵ is the smallest number of balls (with respect to distance d) of radius ϵ needed to cover S.

Now we provide some preliminaries on splines, which can be found, for example, in Schumaker (2007). A function is called a spline of order $q \in \mathbb{N}$, with respect to a

certain partition of its support, if it is $q-2$ times continuously differentiable and when restricted to each interval in this partition, coincides with a polynomial of degree at most $q-1$. Consider $q \in \mathbb{N}$, $q \geq 2$, which will be fixed throughout the remainder of this text. For any $j \in \mathbb{N}$, such that $j \geq q$ let $\mathcal{K}_j = \{(k_1, \ldots, k_{j-q}) \in (0,1)^{j-q} : 0 < k_1 < \cdots < k_{j-q} < 1\}$. We will refer to a vector $\boldsymbol{k} = \boldsymbol{k}_j \in \mathcal{K}_j$ as a set of inner knots; the index j in $\boldsymbol{k}_j$ will sometimes be used to emphasize the dependence on j. A vector $\boldsymbol{k} \in \mathcal{K}_j$ will be said to induce the partition $\big\{[k_0, k_1), [k_1, k_2), \ldots, [k_{j-q}, k_{j-q+1}]\big\}$, with $k_0 = 0$ and $k_{j-q+1} = 1$. For any $\boldsymbol{k} \in \mathcal{K}_j$ we will call $M(\boldsymbol{k}) = \max_{i=1}^{j-q+1} |k_i - k_{i-1}|$ the mesh size of the partition induced by $\boldsymbol{k}$ and $m(\boldsymbol{k}) = \min_{i=1}^{j-q+1} |k_i - k_{i-1}|$ the sparseness of the partition induced by $\boldsymbol{k}$. For a $\boldsymbol{k} \in \mathcal{K}_j$, denote by $\mathcal{S}^{\boldsymbol{k}} = \mathcal{S}_q^{\boldsymbol{k}}$ the linear space of splines of order q on $[0,1]$ with simple knots $\boldsymbol{k}$ (see the definition of knot multiplicity in Schumaker (2007)). This space has dimension j and admits a basis of so called B-splines $\{B_1^{\boldsymbol{k}}, \ldots, B_j^{\boldsymbol{k}}\}$. The construction of $\{B_1^{\boldsymbol{k}}, \ldots, B_j^{\boldsymbol{k}}\}$ involves the knots $k_{-q+1}, \ldots, k_{-1}, k_0, k_1, \ldots, k_{j-q}, k_{j-q+1}, k_{j-q+2}, \ldots, k_j$, with arbitrary extra knots $k_{-q+1} \leq \cdots \leq k_{-1} \leq k_0 = 0$ and $1 = k_{j-q+1} \leq k_{j-q+2} \leq \cdots \leq k_j$. Usually one takes $k_{-q+1} = \cdots = k_{-1} = k_0 = 0$ and $1 = k_{j-q+1} = \cdots = k_j$, and we adopt this choice here as well. These basis functions are nonnegative: $B_i^{\boldsymbol{k}}(x) \geq 0$, for all $x \in [0,1]$. Besides, they have local support and form a partition of unity:

$$B_i^{\boldsymbol{k}}(x) = 0 \ \text{ for } x \notin [k_{-q+i}, k_i], \quad \sum_{i=1}^{j} B_i^{\boldsymbol{k}}(x) = 1 \ \text{ for all } x \in [0,1]. \tag{1}$$

To refer explicitly to the coordinates $\boldsymbol{a} = (a_1, \ldots, a_j) \in \mathbb{R}^j$ of a spline in a specific B-spline basis with inner knots $\boldsymbol{k}$, we write $s_{\boldsymbol{a},\boldsymbol{k}}(x) = \sum_{i=1}^{j} a_i B_i^{\boldsymbol{k}}(x)$, $x \in [0,1]$. Since $\sum_{i=1}^{j} B_i^{\boldsymbol{k}}(x) = 1$, it is easy to see that for any $s_{\boldsymbol{a},\boldsymbol{k}}, s_{\boldsymbol{b},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}}$

$$\|s_{\boldsymbol{a},\boldsymbol{k}} - s_{\boldsymbol{b},\boldsymbol{k}}\|_2 \leq \|s_{\boldsymbol{a},\boldsymbol{k}} - s_{\boldsymbol{b},\boldsymbol{k}}\|_\infty \leq \|\boldsymbol{a} - \boldsymbol{b}\|_\infty \leq \|\boldsymbol{a} - \boldsymbol{b}\|_2. \tag{2}$$

Splines have good approximation properties for sufficiently smooth functions provided they are defined on a partition with appropriately small mesh size. We say that a function f on $[0,1]$ belongs to a generic smoothness class $\mathcal{F}_\alpha$, $\alpha > 0$, if f is Lipschitz, i.e., $f \in \mathcal{L}(\kappa_\alpha, L_\alpha) = \{f : |f(x_1) - f(x_2)| \leq L_\alpha |x_1 - x_2|^{\kappa_\alpha}, x_1, x_2 \in [0,1]\}$ for some $\kappa_\alpha, L_\alpha > 0$, and for any set of inner knots $\boldsymbol{k}$ there exists a spline $s_{\boldsymbol{a},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}}$ such that for some bounded C_f

$$\|f - s_{\boldsymbol{a},\boldsymbol{k}}\|_\infty \leq C_f M^\alpha(\boldsymbol{k}). \tag{3}$$

A leading example of a smoothness class $\mathcal{F}_\alpha$ is the Hölder space $\mathcal{H}_\alpha = \mathcal{H}_\alpha(L, [0,1])$, $0 < \alpha \leq q$, which is the collection of all functions f that have bounded derivatives up to order $\alpha_0 = \lfloor \alpha \rfloor = \max\{z \in \mathbb{Z} : z < \alpha\}$ and such that the α_0-th derivative satisfies the Hölder condition $|f^{(\alpha_0)}(x) - f^{(\alpha_0)}(y)| \leq L|x-y|^{\alpha - \alpha_0}$, for $L > 0$ and $x, y \in [0,1]$. In this case, a well-known spline approximation result (cf. de Boor 1978) states that (3) holds with $C_f = C_q \|f^{(\alpha)}\|_\infty$ for some constant C_q depending only on q. Other examples of smoothness classes for which the approximation property (3) holds include α-times continuously differentiable functions, Sobolev and Besov spaces; cf. Theorems 6.21, 6.25 and 6.31 in Schumaker (2007).

3 Main Result

We begin by describing a hierarchical prior on $\mathcal{S} = \mathcal{S}_q = \cup_{j=q}^{\infty} \cup_{\boldsymbol{k} \in \mathcal{K}_j} \mathcal{S}_q^{\boldsymbol{k}}$: first draw a number $J \in \mathbb{N}$, $J \geq q$; then, given J, generate independently $(J-q)$ inner knots $\boldsymbol{K}_J \in \mathcal{K}_j$ and also independently, J B-spline coefficients $\boldsymbol{\theta} \in \mathbb{R}^J$. Our prior on $\mathcal{S}$ will be the law of the random spline $s_{\boldsymbol{\theta},\boldsymbol{K}_J}$. We impose the following conditions on this prior. For $c_1, c_2 > 0$, $0 \leq t_1, t_2 \leq 1$ and all sufficiently large j,

$$\mathbb{P}(J > j) \lesssim \exp\big(-c_1 j \log^{t_1} j\big), \tag{4}$$

$$\mathbb{P}(J = j) \gtrsim \exp\big(-c_2 j \log^{t_2} j\big). \tag{5}$$

For some $\tau \geq 1$, $c_3 > 0$, $0 \leq t_3 \leq 1$, and all $j \geq q$,

$$\mathbb{P}\big(m(\boldsymbol{K}_j) < \delta(j) | J = j\big) = 0, \tag{6}$$

$$\mathbb{P}\big(M(\boldsymbol{K}_j) \leq \tau/j | J = j\big) \gtrsim \exp\big(-c_3 j \log^{t_3} j\big), \tag{7}$$

where $\delta(i)$ is a positive, strictly decreasing function on $\mathbb{N}$. Without loss of generality assume that $\delta(i) \leq 1$, $i \in \mathbb{N}$. For each $j \geq q$, the conditional distribution of $\boldsymbol{\theta} \in \mathbb{R}^j$ satisfies the following condition: for any $M > 0$ there exists $c_0 = c_0(M)$ such that

$$\mathbb{P}\big(\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_\infty \leq \epsilon | J = j\big) \gtrsim \exp\big(-c_0 j \log(1/\epsilon)\big) \tag{8}$$

for all $\epsilon > 0$ and all $\boldsymbol{\theta}_0 \in \mathbb{R}^j$ such that $\|\boldsymbol{\theta}_0\|_\infty \leq M$.

For examples of particular choices on the components of our hierarchical prior which verify these conditions we refer the reader to Section 5.

Denote $\mathcal{C}^j(M) = [-M, M]^j$. The following theorem is our main result.

Theorem 1. *Let $\|f_0\|_\infty < M$ and $f_0 \in \mathcal{F}_\alpha$ so that (3) holds with C_{f_0}. Let $\epsilon_n, \bar{\epsilon}_n$ be two positive sequences such that $\epsilon_n \geq \bar{\epsilon}_n$, $\epsilon_n \to 0$ as $n \to \infty$ and $n\bar{\epsilon}_n^2 > 1$. Assume that there exist sequences $J_n, \bar{J}_n > q$, $M_n \geq 1$ and a constant $c_M \geq c_1$ satisfying:*

$$J_n \log\Big[\frac{J_n M_n}{\epsilon_n \delta(J_n)}\Big] \lesssim n\epsilon_n^2, \tag{9}$$

$$n\bar{\epsilon}_n^2 \leq J_n \log^{t_1} J_n, \quad \mathbb{P}\big(\boldsymbol{\theta} \notin \mathcal{C}^j(M_n) | J = j\big) \lesssim \exp(-c_M n\bar{\epsilon}_n^2), \ q \leq j \leq J_n, \tag{10}$$

$$\Big[\frac{\bar{\epsilon}_n}{\tau^\alpha C_{f_0}}\Big]^{-1/\alpha} \leq \bar{J}_n, \quad \log^{t_2 \vee t_3} \bar{J}_n \lesssim \log\frac{1}{\bar{\epsilon}_n}. \tag{11}$$

Let $\mathcal{S}_n = \cup_{j=q}^{J_n} \cup_{\boldsymbol{k} \in \mathcal{K}_j^{\delta(j)}} \big\{s_{\boldsymbol{\theta},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}} : \|\boldsymbol{\theta}\|_\infty \leq M_n\big\}$, where $\mathcal{K}_j^\delta = \{\boldsymbol{k} \in \mathcal{K}_j : m(\boldsymbol{k}) \geq \delta\}$. Then it holds that

$$\log N(\epsilon_n, \mathcal{S}_n, \|\cdot\|_2) \lesssim n\epsilon_n^2, \tag{12}$$

$$\mathbb{P}\big(s_{\boldsymbol{\theta},\boldsymbol{K}_J} \notin \mathcal{S}_n\big) \lesssim \exp\big\{-c_1 n\bar{\epsilon}_n^2\big\}, \tag{13}$$

$$\mathbb{P}\big(\|s_{\boldsymbol{\theta},\boldsymbol{K}_J} - f_0\|_\infty \leq 2\bar{\epsilon}_n\big) \gtrsim \exp\big\{-(c_0(M) + c_2 + c_3)\bar{J}_n \log(1/\bar{\epsilon}_n)\big\}. \tag{14}$$

Proof. First we establish (12). Let $L_n(j) = 4M_n j(q+1)(\delta(j))^{-(q+1)}$ and $j > q$. Let $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{m_1}\}$ be an $\epsilon_n/2$-net of the set $\{\boldsymbol{\theta} \in \mathbb{R}^j : \|\boldsymbol{\theta}\|_\infty \le M_n\}$ and let $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_{m_2}\}$ be an $\epsilon_n/(2L_n(j))$-net of $\mathcal{K}_j^{\delta(j)} \subseteq \{\boldsymbol{x} \in \mathbb{R}^{j-q} : \boldsymbol{x} \in (0,1)^{j-q}\}$, both with respect to the $\|\cdot\|_\infty$-norm. Then, by using (2) and Lemma 2 (Lemma 2 is applicable since $\epsilon_n/(2L_n(j)) \le 2/(q-1)$ for sufficiently large n), $\{s_{\boldsymbol{\theta}_k, \boldsymbol{x}_l}, k = 1, \dots, m_1,\ l = 1, \dots m_2\}$ forms an ϵ_n-net of $\cup_{\boldsymbol{k} \in \mathcal{K}_j^{\delta(j)}} \{s_{\boldsymbol{\theta},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}} : \|\boldsymbol{\theta}\|_\infty \le M_n\}$ with respect to the $\|\cdot\|_\infty$-norm. By using this fact, we obtain for sufficiently large n that

$$
\begin{aligned}
&N\big(\epsilon_n, \mathcal{S}_n, \|\cdot\|_2\big) \le N\big(\epsilon_n, \mathcal{S}_n, \|\cdot\|_\infty\big) \\
&\le \sum_{j=q}^{J_n} N\Big(\epsilon_n, \cup_{\boldsymbol{k} \in \mathcal{K}_j^{\delta(j)}} \{s_{\boldsymbol{\theta},\boldsymbol{k}} \in \mathcal{S}_q^{\boldsymbol{k}} : \|\boldsymbol{\theta}\|_\infty \le M_n\}, \|\cdot\|_\infty\Big) \\
&\le \sum_{j=q}^{J_n} \Big[N\Big(\frac{\epsilon_n}{2}, \{\boldsymbol{\theta} \in \mathbb{R}^j : \|\boldsymbol{\theta}\|_\infty \le M_n\}, \|\cdot\|_\infty\Big) N\Big(\frac{\epsilon_n}{2L_n(j)}, (0,1)^{j-q}, \|\cdot\|_\infty\Big)\Big] \\
&\le J_n \Big[\frac{2M_n}{\epsilon_n}\Big]^{J_n} \Big[\frac{2L_n(J_n)}{\epsilon_n}\Big]^{J_n - q} \le J_n \Big(\frac{16(q+1)M_n^2 J_n}{\epsilon_n^2 (\delta(J_n))^{q+1}}\Big)^{J_n}.
\end{aligned}
$$

The last relation and (9) imply (12):

$$
\log N\big(\epsilon_n, \mathcal{S}_n, \|\cdot\|_2\big) \lesssim J_n \log\Big[\frac{J_n M_n}{\epsilon_n \delta(J_n)}\Big] \lesssim n\epsilon_n^2.
$$

Now we check (13). From the definition of $\mathcal{S}_n$, the relations (4), (6) and (10), it follows that

$$
\begin{aligned}
&\mathbb{P}\big(s_{\boldsymbol{\theta},\boldsymbol{K}_J} \notin \mathcal{S}_n\big) \le \mathbb{P}\Big(\{J > J_n\} \cup \big[\{q \le J \le J_n\} \cap \big(\{m(\boldsymbol{K}_j) < \delta(j)\} \cup \{\boldsymbol{\theta} \notin \mathcal{C}^j(M_n)\}\big)\big]\Big) \\
&\le \mathbb{P}\big(J > J_n\big) + \sum_{j=q}^{J_n} \mathbb{P}\big(J = j\big)\Big(\mathbb{P}\big(m(\boldsymbol{K}_j) < \delta(j) | J = j\big) + \mathbb{P}\big(\boldsymbol{\theta} \notin \mathcal{C}^j(M_n) | J = j\big)\Big) \\
&\lesssim \exp\big\{-c_1 J_n \log^{t_1} J_n\big\} + 0 + \exp\big\{-c_M n\bar{\epsilon}_n^2\big\} \lesssim \exp\big\{-c_1 n\bar{\epsilon}_n^2\big\}.
\end{aligned}
$$

It remains to prove (14). First note that, by using (3) and (11), for all $j \ge \bar{J}_n$ and for all sets of knots $\boldsymbol{k}_j \in \mathcal{K}_j$ such that $M(\boldsymbol{k}_j) \le \tau/j$, there exists a spline $s_{\boldsymbol{\theta}_0,\boldsymbol{k}_j} \in \mathcal{S}_q^{\boldsymbol{k}_j}$ (of course, $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(\boldsymbol{k}_j) = \boldsymbol{\theta}_0(\boldsymbol{k}_j, f_0)$) such that

$$
\|f_0 - s_{\boldsymbol{\theta}_0,\boldsymbol{k}_j}\|_\infty \le C_{f_0} M^\alpha(\boldsymbol{k}_j) \le C_{f_0} \tau^\alpha \bar{J}_n^{-\alpha} \le \bar{\epsilon}_n. \tag{15}
$$

Since $\|f_0\|_\infty < M$, there exists an $\varepsilon > 0$ such that the spline $s_{\boldsymbol{\theta}_0,\boldsymbol{k}_j}$ from (15) satisfies $\|s_{\boldsymbol{\theta}_0,\boldsymbol{k}_j}\|_\infty \le M - \varepsilon$ for sufficiently large n. Besides, $\bar{J}_n$ must grow with n in view of (11). Then, according to Lemma 3, there exists a $\delta = \delta(\mathcal{F}_\alpha, \varepsilon)$ such that, for sufficiently large n, $\|\boldsymbol{\theta}_0(\boldsymbol{k}_j)\|_\infty \le M$ for all sets of knots $\boldsymbol{k}_j \in \mathcal{K}_j$ such that $M(\boldsymbol{k}_j) \le \tau/\bar{J}_n \le \delta$ and $j \ge \bar{J}_n$.

Introduce the events: $E_1^j = \{M(\boldsymbol{K}_j) \le \tau/j\}$, $E_2^j = \{\|f_0 - s_{\boldsymbol{\theta}_0(\boldsymbol{K}_j),\boldsymbol{K}_j}\|_\infty \le \bar\epsilon_n\}$, $E_3^j = \{\|\boldsymbol{\theta}_0(\boldsymbol{K}_j) - \boldsymbol{\theta}\|_\infty \le \bar\epsilon_n\}$, $E_4^j = \{\|f_0 - s_{\boldsymbol{\theta},\boldsymbol{K}_j}\|_\infty \le 2\bar\epsilon_n\}$ and $E_5^j = \{\|\boldsymbol{\theta}_0(\boldsymbol{K}_j)\|_\infty \le M\}$. Using the argument from the previous paragraph, the triangle inequality, (2) and (15), we obtain that

$$E_1^{\bar J_n} \subseteq E_2^{\bar J_n}, \quad E_1^{\bar J_n} \subseteq E_5^{\bar J_n}, \quad E_2^j \cap E_3^j \subseteq E_4^j, \quad j \ge q. \tag{16}$$

Combining (5), (7), (8), (11) and (16), we prove (14):

$$\begin{aligned}
&\mathbb{P}\big(\|s_{\boldsymbol{\theta},\boldsymbol{K}_J} - f_0\|_\infty \le 2\bar\epsilon_n\big) = \mathbb{P}(E_4^J) \ge \mathbb{P}(J=\bar J_n)\mathbb{P}\big(E_4^{\bar J_n}|J=\bar J_n\big)\\
&\ge \mathbb{P}(J=\bar J_n)\mathbb{P}\big(E_2^{\bar J_n}\cap E_3^{\bar J_n}|J=\bar J_n\big)\\
&\ge \mathbb{P}(J=\bar J_n)\mathbb{P}\big(E_1^{\bar J_n}\cap E_3^{\bar J_n}\cap E_5^{\bar J_n}|J=\bar J_n\big)\\
&= \mathbb{P}(J=\bar J_n)\mathbb{E}\big[P\big(E_1^{\bar J_n}\cap E_3^{\bar J_n}\cap E_5^{\bar J_n}|J=\bar J_n,\boldsymbol{K}_{\bar J_n}\big)\big]\\
&= \mathbb{P}(J=\bar J_n)\mathbb{E}\big[\mathbb{I}\{\boldsymbol{K}_{\bar J_n}\in E_1^{\bar J_n}\cap E_5^{\bar J_n}\}\mathbb{P}\big(E_3^{\bar J_n}|J=\bar J_n,\boldsymbol{K}_{\bar J_n}\big)\big]\\
&\ge \mathbb{P}(J=\bar J_n)\mathbb{P}\big(E_1^{\bar J_n}|J=\bar J_n\big)\inf_{\|\boldsymbol{\theta}_0\|_\infty\le M}\mathbb{P}\big(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_\infty\le\bar\epsilon_n|J=\bar J_n\big)\\
&\gtrsim \exp\big(-(c_2+c_3)\bar J_n\log^{t_2\vee t_3}\bar J_n\big)\exp\big(-c_0(M)\bar J_n\log(1/\bar\epsilon_n)\big)\\
&\gtrsim \exp\big(-(c_0(M)+c_2+c_3)\bar J_n\log(1/\bar\epsilon_n)\big).
\end{aligned}$$

□

Remark 1. *Condition (6) is used in the proof of Theorem 1 exclusively to enforce* $\sum_{j=q}^J \mathbb{P}\big(J=j\big)\mathbb{P}\big(m(\boldsymbol{K}_j)<\delta(j)|J=j\big)$ *to be zero, when proving (13). Inspection of the proof shows however that, instead of condition (6), it would suffice to require this sum to be upper-bounded by a multiple of* $\exp\big\{-c_1 n\bar\epsilon_n^2\big\}$. *Although this would be a weaker requirement, typically the sequence* $\bar\epsilon_n$ *will depend on the unknown smoothness* α. *To avoid the dependence on* $\bar\epsilon_n$, *a slightly stronger condition (based on the fact that* $n\bar\epsilon_n^2$ *is of a smaller order than* n *as* $n\to\infty$*) can be proposed. Namely, if condition (6) is replaced by*

$$\sum_{j=q}^{J_n}\mathbb{P}\big(J=j\big)\mathbb{P}\big(m(\boldsymbol{K}_j)<\delta(j)|J=j\big)\le c_5\exp(-c_4 n), \tag{6'}$$

for some $c_4, c_5 > 0$ *and a function* $\delta(\cdot)$ *as in (6), then the conclusions of Theorem 1 remain valid as long as* J_n *is a sequence satisfying (9) and (10); cf. Section 5 for a comparison of (6) and (6').*

Remark 2. *If the range of the underlying curve* f_0 *is contained in some known interval* $[a,b]\subset\mathbb{R}$, *then, according to Lemma 3 and the proof of property (14), the prior on* $\boldsymbol{\theta}\in\mathbb{R}^j$ *can be chosen to be supported on, say,* $[a-1,b+1]^j$ *so that (8) has to hold only for* $\boldsymbol{\theta}_0\in[a-1,b+1]^j$. *Condition (10) will trivially be satisfied for* $M_n > (1-a)\vee(b+1)$.

Remark 3. *If (26) is assumed instead of (7), the proof of (14) can then be simplified a lot, as in this case one can condition on the event* $\{\boldsymbol{K}_{\bar J_n}=\bar{\boldsymbol{k}}_{\bar J_n}\}$ *so that* $\boldsymbol{\theta}_0=\boldsymbol{\theta}_0(\bar{\boldsymbol{k}}_{\bar J_n})$ *becomes fixed and* $\mathbb{P}(E_1^J|J=\bar J_n,\boldsymbol{K}_{\bar J_n}=\bar{\boldsymbol{k}}_{\bar J_n})=1$.

4 Implications of the main result

We clarify now the relevance of our result. Consider a family of models $\mathcal{P} = \{\mathbb{P}_f : f \in \mathcal{F}_{\mathcal{A}}\}$, $\mathcal{F}_{\mathcal{A}} = \cup_{\alpha\in\mathcal{A}}\mathcal{F}_\alpha$, with densities p_f with respect to some common dominating measure. Assume that we observe a sample $\boldsymbol{X}^{(n)} = (X_1, \ldots, X_n) \sim p_{f_0}^{(n)}$, $X_i \overset{ind}{\sim} p_{f_0}$, $f_0 \in \mathcal{F}_\alpha$ for some unknown smoothness $\alpha \in \mathcal{A}$. The Bayesian approach consists of putting a prior measure Π on $\mathcal{F} \subseteq \mathcal{F}_{\mathcal{A}}$ which, together with the likelihood $p_f^{(n)}$, leads to the posterior distribution $\Pi(\cdot|\boldsymbol{X}^{(n)})$ via Bayes' formula:

$$\Pi\big(A|\boldsymbol{X}^{(n)}\big) = \frac{\int_A p_f^{(n)}(\boldsymbol{X}^{(n)})\, d\Pi(f)}{\int_{\mathcal{F}} p_f^{(n)}(\boldsymbol{X}^{(n)})\, d\Pi(f)}$$

for a measurable $A \subseteq \mathcal{F}$. The asymptotic behavior of the posterior distribution can be studied from the point of view of the probability measure $\mathbb{P}_0 = \mathbb{P}_{f_0}$; see Ghosal et al. (2000).

For two densities p_f and p_g with $f, g \in \mathcal{F}_{\mathcal{A}}$, define the (squared) Hellinger metric $h^2(p_f, p_g) = 2\big(1 - \mathbb{E}_g\sqrt{p_f(X)/p_g(X)}\big)$, Kullback-Leibler divergence $K(p_f, p_g) = -\mathbb{E}_g \log\big(p_f(X)/p_g(X)\big)$ and the Csiszár f-divergence $V(p_f, p_g) = \mathbb{E}_g \log^2\big(p_f(X)/p_g(X)\big)$. Define also the ball $B(\epsilon_n, f_0) = \big\{f \in \mathcal{F} : K(f, f_0) \le \epsilon^2, V(f, f_0) \le \epsilon^2\big\}$.

The following theorem is a version of Theorem 2.1 from Ghosal and van der Vaart (2001) which makes a statement about the asymptotic behavior of a posterior measure.

Theorem 2 (Theorem 2.1 of Ghosal et al. 2000)**.** *Let Π_n be a sequence of priors on $\mathcal{F}$. Suppose that for two positive sequences $\kappa_n \ge \bar{\kappa}_n$ such that $n\bar{\kappa}_n^2 \to \infty$ and $\kappa_n \to 0$ as $n \to \infty$, sets $\mathcal{F}_n \subseteq \mathcal{F}$ and constants $b_1, b_2, b_3, b_4 > 0$, the following conditions hold:*

$$\log N\big(\kappa_n, \mathcal{F}_n, h\big) \le b_1 n\kappa_n^2, \tag{17}$$

$$\Pi_n(\mathcal{F}\backslash\mathcal{F}_n) \le b_2 e^{-(b_3+4)n\bar{\kappa}_n^2}, \tag{18}$$

$$\Pi_n(B(\bar{\kappa}_n, f_0)) \ge b_4 e^{-b_3 n\bar{\kappa}_n^2}. \tag{19}$$

Then, for large enough $M > 0$, $\Pi_n\big(f \in \mathcal{F} : h(p_f, p_{f_0}) \ge M\kappa_n|\boldsymbol{X}^{(n)}\big) \to 0$ as $n \to \infty$ in $\mathbb{P}_{f_0}$-probability.

The conditions of this theorem require the existence of a *sieve* $\mathcal{F}_n$ with small entropy (17) which contains most of the prior mass (18) and with enough prior mass around the parameter f_0 which indexes the "true" underlying measure of the data. Assume now that the models in $\mathcal{P}$ are such that for d^2 being h^2, K or V, $d^2(p_f, p_{f_0}) \lesssim \|f - f_0\|_2^2$. If in addition one can prove that in the considered model $h(p_f, p_{f_0}) \gtrsim \|f - f_0\|_2$, then Theorem 2 delivers a contraction rate ϵ_n with respect to the L_2-distance as well. Some examples of models for which the above relations between norms can be established are, among others, white noise, density estimation, non-parametric regression, binary regression, Poisson regression and classification; cf. Ghosal et al. (2000), de Jonge and van Zanten (2012), Shen and Ghosal (2012). We should note here that it requires a fair

piece of effort to implement this idea for many concrete models, only for the white noise model are the above relations between norms straightforward. Once these relations between norms are established, one can apply our meta-theorem (Theorem 1) to obtain an adaptive contraction rate which essentially verifies (17)–(19) for our spline-based prior. We summarize this in the following theorem.

Theorem 3. *Let Π be the spline prior described in Section 3. Consider a family of models $\mathcal{P} = \{\mathbb{P}_f : f \in \mathcal{F}_{\mathcal{A}}\}$, $\mathcal{F}_{\mathcal{A}} = \cup_{\alpha\in\mathcal{A}}\mathcal{F}_\alpha$, with densities p_f with respect to some common dominating measure. Assume also that the models in $\mathcal{P}$ are such that for d^2 being h^2, K or V, $d^2(p_f, p_{f_0}) \lesssim \|f - f_0\|_2^2$. Take an i.i.d. sample $\boldsymbol{X}^{(n)} = (X_1, \dots, X_n)$, $X_i \sim p_{f_0}$, $f_0 \in \mathcal{F}_\alpha$, $\|f_0\|_\infty < M$, for some unknown smoothness $\alpha \in \mathcal{A}$, $\alpha \le q$. Consider a prior Π which verifies (4) through (8) for certain constants c_1, c_2, c_3, t_1, t_2 and t_3. Assume that at least one of the two conditions, $\alpha > 1$ or $t_2 \wedge t_3 < 1$, is fulfilled.*

Then, for large enough $C > 0$, $\Pi\big(f \in \mathcal{F} : h(p_f, p_{f_0}) \ge C r_n | \boldsymbol{X}^{(n)}\big) \to 0$ as $n \to \infty$ in $\mathbb{P}_0$-probability for $r_n = n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$. If $h(p_f, p_{f_0}) \gtrsim \|f - f_0\|_2$ then in the previous statement the Hellinger distance may be replaced by the L_2 distance and the statement remains valid.

Proof. We have that for some constant $\rho > 0$ and $\mathcal{F} = \mathcal{S}$, $\mathcal{F}_n = \mathcal{S}_n$,

$$N\big(\kappa_n, \mathcal{F}_n, h\big) \le N(\kappa_n/\rho, \mathcal{F}_n, \|\cdot\|_2), \tag{20}$$

$$\Pi(\mathcal{F}\backslash\mathcal{F}_n) = \mathbb{P}\big(s_{\boldsymbol{\theta},\boldsymbol{K}_J} \notin \mathcal{F}_n\big), \tag{21}$$

$$\Pi(B(\bar\kappa_n, f_0)) \ge \mathbb{P}\big(\|s_{\boldsymbol{\theta},\boldsymbol{K}_J} - f_0\|_\infty \le \bar\kappa_n/\rho\big). \tag{22}$$

The first inequality follows from the fact that by assumption $h(p_f, p_g) \le \rho\|f - g\|_2$ and so a κ/ρ-cover of $\mathcal{F}_n$ according to $\|\cdot\|_2$ induces a κ-cover of $\mathcal{F}_n$ according to h. Then, since for d^2 being K or V, $d^2(p_f, p_{f_0}) \le \rho\|f - f_0\|_2^2$, we have $B(\bar\kappa_n, f_0) \supset \big\{f \in \mathcal{F} : \|f - f_0\|_2 \le \bar\kappa/\rho\big\}$ and the last inequality follows.

By assumption $f_0 \in \mathcal{F}_\alpha$ satisfies the conditions of Theorem 1; assume (3) holds for some C_{f_0}. Consider then a prior that satisfies (4)–(8). Let us present a choice of quantities M_n, $\delta(j)$, J_n, $\bar J_n$, ϵ_n and $\bar\epsilon_n$ which meet conditions (9)–(11). First of all, sequence M_n can be taken as a polynomial in n (for instance, for normal or exponential conditional priors for $\boldsymbol{\theta} \in \mathbb{R}^j$ in (10)) and $1/\delta(j)$ as a polynomial in j. Next, note that there is no $\bar J_n$ that satisfies (11) if both $\alpha \le 1$ and $t_2 \wedge t_3 = 1$ hold. If either $\alpha > 1$ or $t_2 \wedge t_3 < 1$, then the best possible choices are $\bar J_n = \bar J_n(C_1) = \tau C_{f_0}^{1/\alpha}(\bar\epsilon_n(C_1))^{-1/\alpha}$ so that the first inequality of (11) is satisfied, $\bar\epsilon_n = \bar\epsilon_n(C_1) = C_1(\log n/n)^{\alpha/(2\alpha+1)}$ for sufficiently large $C_1 \ge 1$ so that the second inequality of (11) is satisfied, $J_n = C_2 n^{1/(2\alpha+1)}(\log n)^{2\alpha/(2\alpha+1)-t_1}$ for sufficiently large C_2 (any $C_2 \ge C_1^2$ will do) so that the first inequality of (10) is satisfied, and finally,

$$\epsilon_n = C_3 n^{-\alpha/(2\alpha+1)}(\log n)^{\alpha/(2\alpha+1)+(1-t_1)/2}$$

for sufficiently large C_3 so that (9) is satisfied. Since these quantities satisfy (9)–(11), Theorem 1 implies conditions (12)–(14) for the quantities defined above. Besides, we take constants C_1, C_2, C_3 so big that (13) and (14) also hold for $\bar\epsilon_n(\sqrt{C_1})$ and $\bar J_n(\sqrt{C_1})$.

Now, take $\kappa_n = 2\rho\epsilon_n$ and $\bar{\kappa}_n = 2\rho\bar{\epsilon}_n(\sqrt{C_1})$. Then it follows from (12) and (20) that

$$N\big(\kappa_n, \mathcal{F}_n, h\big) \le N(\kappa_n/\rho, \mathcal{F}_n, \|\cdot\|_2) = N(\epsilon_n, \mathcal{F}_n, \|\cdot\|_2) \lesssim n\epsilon_n^2 \lesssim n\kappa_n^2. \tag{23}$$

Next, using (21) and (13) for $\bar{\epsilon}_n(C_1)$ and $\bar{J}_n(C_1)$, we obtain that

$$\begin{aligned}\Pi(\mathcal{F}\backslash\mathcal{F}_n) &= \mathbb{P}\big(s_{\boldsymbol{\theta},\boldsymbol{K}_J} \notin \mathcal{F}_n\big) \lesssim \exp\big\{-c_1 n\bar{\epsilon}_n^2(C_1)\big\} \\ &= \exp\big\{-c_1(2\rho)^{-2}C_1 n\bar{\kappa}_n^2\big\} \le \exp\big\{-5n\bar{\kappa}_n^2\big\}\end{aligned} \tag{24}$$

for sufficiently large C_1. Denote $K = (c_0(M) + c_2 + c_3)\tau C_{f_0}^{1/\alpha}(2\rho)^{-2}\alpha/(2\alpha+1)$, then

$$(c_0(M) + c_2 + c_3)\bar{J}_n(\sqrt{C_1})\log(1/\bar{\epsilon}_n(\sqrt{C_1})) = KC_1^{-(1+1/\alpha)}n\bar{\kappa}_n^2(1+o(1))$$

as $n \to \infty$. The last relation, (22) and (14) for $\bar{\epsilon}_n(\sqrt{C_1})$ and $\bar{J}_n(\sqrt{C_1})$ imply that

$$\begin{aligned}\Pi(B(\bar{\kappa}_n, f_0)) &\ge \mathbb{P}\big(\|s_{\boldsymbol{\theta},\boldsymbol{K}_J} - f_0\|_\infty \le 2\bar{\epsilon}_n(\sqrt{C_1})\big) \\ &\gtrsim \exp\big\{-(c_0(M) + c_2 + c_3)\bar{J}_n(\sqrt{C_1})\log(1/\bar{\epsilon}_n(\sqrt{C_1}))\big\} \\ &\gtrsim \exp\big\{-KC_1^{-(1+1/\alpha)}n\bar{\kappa}_n^2\big\} \ge \exp\big\{-n\bar{\kappa}_n^2\big\}\end{aligned} \tag{25}$$

for sufficiently large C_1. Thus, for sufficiently large C_1 (and C_2, C_3), relations (17)–(19) follow from (23), (24) and (25) respectively.

Finally, applying Theorem 2 (since (17)–(19) are fulfilled), we conclude that the contraction rate of the resulting posterior is at most ϵ_n, which appears to be optimal (up to a logarithmic factor) in a minimax sense over the Hölder class $\mathcal{H}_\alpha$ (also over α-smooth Sobolev class). □

Remark 4. *A priori, it may be unknown whether $\alpha > 1$ or not, or it may be simply known that $\alpha \le 1$. We can however always ensure the condition $t_2 \wedge t_3 < 1$ by an appropriate choice of prior. For example, we take a geometric prior on J so that $t_2 = 0$ and a prior on $\boldsymbol{K}_j$ such that (26) (which implies (7)) holds with, say, $t_3 = 0$.*

5 Examples of Priors

We give now examples of particular choices for the several components of our hierarchical prior which verify conditions (4)–(8), (6') and the second relation in (10).

As for the prior on the number of basis functions, assumptions (4) and (5) hold for the geometric, Poisson and negative binomial distributions; cf. Shen and Ghosal (2012) (assumption (5) is slightly different from the corresponding assumption (B1) in Shen and Ghosal 2012). Assumption (8), in turn, will trivially hold if we assume, for example, the coordinates of $\boldsymbol{\theta} \in \mathbb{R}^j$ to be (conditionally on $J = j$) independent and identically distributed according to a density ϕ uniformly bounded away from zero on the interval $[-M, M]$. On the other hand, the prior distribution on $\boldsymbol{\theta} \in \mathbb{R}^j$ (conditionally on $J = j$) should have sufficiently light tails so that the second requirement in (10) holds for a sequence M_n that converges to infinity as $n \to \infty$ not faster than polynomially in n.

It can easily be checked for normal and Gamma densities ϕ. Let us consider standard normal ϕ. As $q \le j \le J_n$ and taking $M_n = n$, we immediately derive the required relation:

$$\mathbb{P}\big(\boldsymbol{\theta} \notin \mathcal{C}^j(M_n) | J = j\big) \le j\mathbb{P}(|\theta_1| \ge M_n | J = j) \le \frac{J_n 2\exp(-M_n^2/2)}{\sqrt{2\pi}M_n} \le \exp(-c_M n\bar{\epsilon}_n^2).$$

There is an ample choice of priors on $\boldsymbol{K}_J$, given $J = j$, which satisfy condition (6). First note that this condition enforces the prior on the location of the knots, for each $J = j$, to be such that, with probability 1, adjacent knots are at least $\delta(j)$ apart. The function $1/\delta(j)$ can be taken as a polynomial in j of high degree which makes the requirement less restrictive. If a certain sequence ϵ_n verifies the conditions of Theorem 1, then an increase in the exponent of $1/\delta(j)$ can be accommodated by making ϵ_n larger by a multiplicative factor (cf. condition (9)).

A simple choice for the prior on $\boldsymbol{K}_J$, given $J = j$, is to pick $(j - q)$ knots uniformly at random, without replacement, on a uniform $\delta(j)$-sparse grid. This construction is possible if δ is chosen in such a way that $\lfloor 1/\delta(j) \rfloor > j - q$ for all j. Another way is to generate the $(j - q)$ inner knots in $\boldsymbol{K}_j$ sequentially in the following way: add a knot K_1 uniformly at random on the interval $[\delta(j), 1-\delta(j)]$, then a knot K_2 uniformly at random on the interval $[\delta(j), 1-\delta(j)]\backslash(K_1-\delta(j), K_1+\delta(j))$ and so on. Finally, take the ordered $\boldsymbol{K}_j = (K_{(1)}, \ldots, K_{(j-q)})$. This construction is always possible if $1/\delta(j) > 2(j - q)$. If J is Poisson distributed, these points are simply distributed like a homogeneous Poisson process, conditioned to have all points at least $\delta(J)$ apart. Clearly, condition (6) is satisfied for these two constructions since all prior mass is concentrated on partitions with sparseness larger than $\delta(j)$.

It is also easy to see that condition (7) is verified for the knot vectors obtained from one of these two constructions. In fact, condition (7) is trivially fulfilled if, for some $0 \le t_3 < 1$,

$$\mathbb{P}(\boldsymbol{K}_j = \bar{\boldsymbol{k}}_j) \gtrsim \exp\big(-c_3 j \log^{t_3} j\big), \tag{26}$$

where $\bar{\boldsymbol{k}}_j \in \mathcal{K}_j$ is the set of $(j-q)$ equally spaced inner knots. This suggests a mechanism to assure that any prior which verifies (6) can be slightly modified to also verify (7): given $J = j$, generate a Bernoulli random variable X with success probability, say, $\exp(-c_3 j \log^{t_3} j)$; if $X = 1$, then take $\boldsymbol{K}_j = \bar{\boldsymbol{k}}_j$, otherwise pick the knots in $\boldsymbol{K}_j$ according to any procedure which verifies (6), for instance, one of the two procedures described above. The resulting prior will trivially satisfy both (6) and (7).

Condition (6) necessarily excludes some knot vectors from the support of the prior (and then also from the support of the posterior). It is therefore of interest to design a weaker alternative for condition (6). Condition (6') plays this role in that it allows priors on $\boldsymbol{K}$ which can have any set of knots of $[0, 1]$ in its support. Assuming condition (6') instead of (6) consequently allows us to put positive mass on any vector of simple knots in a straightforward way: generate a Bernoulli random variable with success probability $1 - c_5\exp(-c_4 n)$; if $X = 1$ take $\boldsymbol{K}_j = \bar{\boldsymbol{k}}_j$; if $X = 0$, then take an arbitrary $\boldsymbol{K}_j$ (for example, independent, uniformly distributed points on $[0, 1]$). If we take $1/\delta(j) = j$ and

$\tau \geq q$, then conditions (6') and (7) are verified. This procedure, although simple, does place little prior mass on knot vectors with inhomogeneous distributions.

An alternative, less degenerate prior which verifies (6') and (7) can be obtained in the following way. Given $J = j$, first generate a Bernoulli random variable X_1 with success probability $c_5 \exp(-c_4 n)$; if $X_1 = 1$ distribute the $(j - q)$ knots arbitrarily; if $X_1 = 0$, then generate another Bernoulli random variable X_2 with success probability $\exp(-j)$; if $X_2 = 1$, then take $(j - q)$ equally spaced knots $\bar{\boldsymbol{k}}_j$; if $X_2 = 0$, then place the knots in such a way that (6) is verified. This procedure allows good control of the prior on the knots while not excluding arbitrary knot vectors.

Note that the priors described above which verify (4)–(8) do not depend on the sample size n, as prescribed by the Bayesian paradigm. Condition (6') is a weaker requirement than condition (6), but it will introduce a dependence on the sample size n in the prior.

Remark 5. *The common practice, in applications, of endowing the location of the knots with a Poisson process prior results in a prior that does not verify assumption (6). Assumption (6'), however, will be satisfied if the prior is modified in such a way that a large enough prior mass is assigned to an equally spaced knot vector.*

6 Numerical Results

In this section we present some numerical results. By applying the reversible jump Markov chain Monte Carlo (RJMCMC) method introduced first by Green (1995), we compare the performance of two hierarchical priors in the nonparametric regression model. Both priors are based on splines, as described in Section 3, and they satisfy conditions (4)–(8). The first prior has equally spaced knots and the second has randomly located knots; we therefore refer to these priors as the fixed knots prior and the free knots prior. We also look into the possibility of using data driven priors on the knots based on some sort of two stage empirical Bayes procedure. We say that vector $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$ is ordered if $x_1 \leq \ldots \leq x_d$.

Consider $n = 1000$ observations $\boldsymbol{X}^{(n)} = \{(t_i, Y_i), i = 1, \ldots, n\}$ from the nonparametric regression model with regular design points $\boldsymbol{t}^{(n)} = (t_1, \ldots, t_n)$, $t_i = i/n$:

$$Y_i = f(t_i) + \xi_i, \quad i = 1, \ldots, n, \tag{M}$$

where the ξ_i's are independent standard Gaussian random variables. It is well known that the relations between appropriate norms required to apply Theorem 3 hold for the model (M). The regression function $f(\cdot)$ is taken to be the so called Doppler function

$$f(t) = 10\sqrt{t(1-t)} \sin\left(\frac{2\pi \cdot 1.05}{t + 0.05}\right), \quad t \in [0, 1], \tag{27}$$

which we plot in Figure 1 together with the observations from the model (M). This function is infinitely many times differentiable on $(0, 1)$ but has a high variability region in a small vicinity of zero.

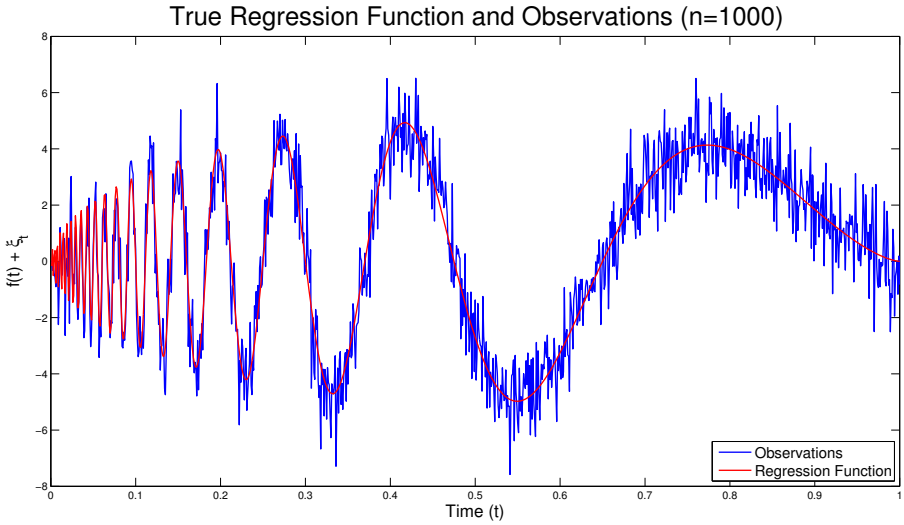

Figure 1: Simulated data from model (M) used in this section (in blue) and the true regression function (in red).

Now we describe the two priors, the fixed knots prior and the free knots prior. In both hierarchical priors we endow J with a (shifted, with support starting with $q \geq 2$) Poisson prior with mean ν and on each spline coefficient we put a uniform prior on $[-M, M]$. In the fixed knots prior, given $J = j$, the $j - q$ inner knots are taken to be equally spaced: $k_i = i/(j-q+1)$, $i = 1, \ldots, j-q$. In the free knots prior, given $J = j$, first generate $U_1, \ldots, U_{j-q}$, uniformly on $\big[0, 1-(j-q+1)\delta(j)\big)$ with $\delta(j) = 1/j^2$, and let $U_{(1)} \leq \ldots \leq U_{(j-q)}$. Next, take the knot vector $\boldsymbol{K}_j$ with entries $K_{i,j} = U_{(i)} + i\delta(j)$, $i = 1, \ldots, j-q$. We represent the fixed knots posterior density as $\bar{\pi}\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j \,|\, \boldsymbol{X}^{(n)}\big)$ and the free knots posterior as $\tilde{\pi}\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j \,|\, \boldsymbol{X}^{(n)}\big)$. We have

$$\bar{\pi}\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j \,|\, \boldsymbol{X}^{(n)}\big) \propto \varphi_n\big(\boldsymbol{X}^{(n)} - \boldsymbol{s}_{\boldsymbol{\theta}_j, \boldsymbol{k}_j}(\boldsymbol{t}^{(n)})\big)\, \nu^{j-q}\, (2M)^{-j},$$
$$\tilde{\pi}\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j \,|\, \boldsymbol{X}^{(n)}\big) \propto \varphi_n\big(\boldsymbol{X}^{(n)} - \boldsymbol{s}_{\boldsymbol{\theta}_j, \boldsymbol{k}_j}(\boldsymbol{t}^{(n)})\big)\, \nu^{j-q}\, (2M)^{-j}\, \big(1-(j-q+1)\delta(j)\big)^{j-q},$$

where φ_n is the density of n independent standard Gaussian random variables and $\boldsymbol{s}_{\boldsymbol{\theta}_j, \boldsymbol{k}_j}(\boldsymbol{t}^{(n)}) = (s_{\boldsymbol{\theta}_j, \boldsymbol{k}_j}(t_1), \ldots, s_{\boldsymbol{\theta}_j, \boldsymbol{k}_j}(t_n))$ represents the spline $s_{\boldsymbol{\theta}_j, \boldsymbol{k}_j}$ evaluated at the design points $\boldsymbol{t}^{(n)}$.

We implement RJMCMC procedures for these two priors to sample from the corresponding posteriors. A generic state of the sampler is a vector $\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j\big) \in \mathbb{N} \times \mathbb{R}^{j-q} \times \mathbb{R}^j$. To sample from the posterior corresponding to the fixed knots prior, we consider three types of moves: a) changing the coefficients of a spline, b) adding a knot and c) removing a knot. In addition to these moves, the sampler for the posterior corresponding to the free knots prior has an extra move: d) changing the location of the knots. These moves are attempted with probabilities p_a, p_b, p_c, p_d $(p_a + p_b + p_c + p_d = 1)$ respectively,

which are parameters of the sampler.

A move of type a) corresponds to jumping from the state $\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j\big)$ to a proposal $\big(j, \boldsymbol{k}_j, \boldsymbol{\vartheta}_j\big)$ where $\boldsymbol{\vartheta}_j = \boldsymbol{\theta}_j + \eta_a \boldsymbol{u}$ and $\boldsymbol{u}$ is a vector of j independent standard normal random variables. This move is attempted with probability p_a. Both η_a and p_a are parameters of the sampler. Moves of type d) correspond to jumping from the state $\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j\big)$ to a proposal $\big(j, \boldsymbol{\kappa}_j, \boldsymbol{\theta}_j\big)$ where $\boldsymbol{\kappa}_j$ is obtained from $\boldsymbol{k}_j$ by perturbing its i-th entry, with the index i chosen uniformly at random, and then ordering the resulting vector. The perturbation is $\kappa_{i,j} = k_{i,j} + \eta_d v$, with v a standard normal random variable. This move is attempted with probability p_d and again, both η_d and p_d are parameters of the sampler. The acceptance probabilities for moves of type a) and moves of type d) are given by, respectively,

$$\min\left(1,\, \frac{\pi\big(j, \boldsymbol{k}_j, \boldsymbol{\vartheta}_j \,|\, \boldsymbol{X}^{(n)}\big)}{\pi\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j \,|\, \boldsymbol{X}^{(n)}\big)}\right) \qquad \text{and} \qquad \min\left(1,\, \frac{\tilde{\pi}\big(j, \boldsymbol{\kappa}_j, \boldsymbol{\theta}_j \,|\, \boldsymbol{X}^{(n)}\big)}{\tilde{\pi}\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j \,|\, \boldsymbol{X}^{(n)}\big)}\right),$$

where π is either $\bar{\pi}$ or $\tilde{\pi}$.

Now we specify how proposals for moves of type b), where we add an extra knot to the current state of the chain $\big(j, \boldsymbol{k}_j, \boldsymbol{\theta}_j\big)$, are designed. The proposal will, for both priors, be a state $\big(j+1, \boldsymbol{\kappa}_{j+1}, \boldsymbol{\vartheta}_{j+1}\big)$. For the fixed knots prior we propose $\kappa_{i,j+1} = i/(j-q+2)$, for $i = 1, \ldots, j-q+1$. For the free knots prior, generate a new knot k uniformly on $(0,1)$ so that $k \in [k_{i-1,j}, k_{i,j}]$ (with $k_{0,j} = 0$ and $k_{j-q+1,j} = 1$) for some $i \in \{1, \ldots, j-q+1\}$, and propose $\boldsymbol{\kappa}_{j+1} = (k_{1,j}, \ldots, k_{i-1,j}, k, k_{i,j}, \ldots, k_{j-q,j})$.

For moves of type b), it remains to describe how the coefficient vector $\boldsymbol{\vartheta}_{j+1}$ is generated in the proposal. Whatever the vector $\boldsymbol{\kappa}_{j+1}$ is, for the sake of comparing the priors, the procedure for proposing $\boldsymbol{\vartheta}_{j+1}$ is the same for both priors. To ease the notation, we abbreviate the current state and the proposed state as $\big(j, \boldsymbol{k}, \boldsymbol{\theta}\big)$ and $\big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\big)$, where both $\boldsymbol{\kappa}$ and $\boldsymbol{\vartheta}$ have one more element than $\boldsymbol{k}$ and $\boldsymbol{\theta}$, respectively. The coefficients $\boldsymbol{\vartheta}$ will be obtained via (perturbed) interpolation at $j+1$ points $\boldsymbol{t} = \boldsymbol{t}_{j+1} = (t_1, \ldots, t_{j+1})$. Of these $j+1$ points, $j-q+2$ points are taken to be the midpoints of the intervals comprised between the adjacent points of the vector $(0, \boldsymbol{\kappa}, 1) \in [0,1]^{j-q+3}$; the remaining $q-1$ points are the first $q-1$ elements from the list $0, 1, \kappa_1, \kappa_{j-q+1}, \kappa_2, \kappa_{j-q}, \kappa_3, \ldots$. The vector $\boldsymbol{t}$ is assumed to be ordered.

Consider now the $(j+1) \times (j+1)$ matrices $C = C_j(\boldsymbol{\kappa}, \boldsymbol{t})$ with (i,l)-entry $B_l^{\boldsymbol{\kappa}}(t_i)$ and the $(j+1) \times j$ matrices $D = D_j(\boldsymbol{k}, \boldsymbol{t})$ with (i,l)-entry $B_l^{\boldsymbol{k}}(t_i)$. One can show that for our choice of interpolation points C and D are of full column rank. For a matrix M denote by $\mathcal{L}(M)$ the linear space spanned by the columns of matrix M. Then $\mathcal{L}(C) = \mathbb{R}^{j+1}$, $\mathcal{L}(D) \subseteq \mathbb{R}^{j+1}$ with $\dim(\mathcal{L}(D)) = j$. Let $\boldsymbol{w} \in \mathcal{L}(D)^{\perp}$ (the orthogonal complement of $\mathcal{L}(D)$ so that $D^T \boldsymbol{w} = \boldsymbol{0}$) which is unique up to scaling. Clearly, the interpolation problem $\boldsymbol{s}_{\boldsymbol{\vartheta},\boldsymbol{\kappa}}(\boldsymbol{t}) = \boldsymbol{s}_{\boldsymbol{\theta},\boldsymbol{k}}(\boldsymbol{t})$ corresponds to the system of linear equations $C\boldsymbol{\vartheta} = D\boldsymbol{\theta}$. Because of the mismatch in the dimensions of $\boldsymbol{\theta}$ and of $\boldsymbol{\vartheta}$, this relation between $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$ is not a bijection. Indeed, $\boldsymbol{\theta} = (D^T D)^{-1} D^T C \boldsymbol{\vartheta}_\rho$ for all $\boldsymbol{\vartheta}_\rho = \boldsymbol{\vartheta}_\rho(\boldsymbol{\theta}) = C^{-1}(D\boldsymbol{\theta} + \rho \boldsymbol{w})$, $\rho \in \mathbb{R}$.

Assume that by default all vectors are column vectors. Our proposal for $\boldsymbol{\vartheta}$ is the

following linear function that matches the dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\vartheta}$:

$$\boldsymbol{\vartheta} = g(\boldsymbol{\theta}, \rho) = C^{-1}\big(D\,\boldsymbol{w}\big)\begin{pmatrix}\boldsymbol{\theta}\\ \eta\rho + \mathring{\rho}\end{pmatrix},$$

where ρ is a standard Gaussian random variable, $\eta > 0$ and $\mathring{\rho} = \mathring{\rho}(\boldsymbol{\theta})$ is taken to be

$$\mathring{\rho}(\boldsymbol{\theta}) = \arg\min_{\rho\in\mathbb{R}} \|s_{\boldsymbol{\vartheta}_\rho(\boldsymbol{\theta}),\boldsymbol{\kappa}} - s_{\boldsymbol{\theta},\boldsymbol{k}}\|^2 = \frac{\langle s_{\boldsymbol{\omega}_1,\boldsymbol{\kappa}},\, s_{\boldsymbol{\theta},\boldsymbol{k}} - s_{\boldsymbol{\omega}_2,\boldsymbol{\kappa}}\rangle}{\langle s_{\boldsymbol{\omega}_1,\boldsymbol{\kappa}},\, s_{\boldsymbol{\omega}_1,\boldsymbol{\kappa}}\rangle},$$

$\boldsymbol{\omega}_1 = C^{-1}\boldsymbol{w}$, $\boldsymbol{\omega}_2 = C^{-1}D\boldsymbol{\theta}$ and $\langle s_1, s_2\rangle$ represents the inner product $\int_0^1 s_1(t)s_2(t)dt$. The interpretation of $\mathring{\rho}$ is that our proposal for $\boldsymbol{s}_{\boldsymbol{\vartheta},\boldsymbol{\kappa}}$ is "centered" (cf. Brooks et al. 2003) around a good approximation $\boldsymbol{s}_{\boldsymbol{\vartheta}_{\mathring{\rho}},\boldsymbol{\kappa}}$ of the previous state $s_{\boldsymbol{\theta},\boldsymbol{k}}$. This central state $\boldsymbol{s}_{\boldsymbol{\vartheta}_{\mathring{\rho}},\boldsymbol{\kappa}}$ can be seen as an ideal interpolator.

It is straightforward to check that the Jacobian matrix of the mapping g is

$$J_g = J_g(\eta) = C^{-1}\Big[\big(D\,\eta\boldsymbol{w}\big) + \big(\boldsymbol{w}\ldots\boldsymbol{w}\big)\mathrm{diag}\big(\nabla_{\boldsymbol{\theta}}\mathring{\rho}(\boldsymbol{\theta}), \eta\big)\Big],$$

where $\mathrm{diag}(\boldsymbol{v})$ denotes a square matrix with the entries of $\boldsymbol{v}$ in its main diagonal and $\nabla_{\boldsymbol{\theta}}\mathring{\rho}(\boldsymbol{\theta})$ is the gradient of $\mathring{\rho}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Note that the determinant of this Jacobian does not depend on the gradient of $\mathring{\rho}$ and is given by

$$\det\big(J_g\big) = \frac{\eta\det\big[(D\,\boldsymbol{w})\big]}{\det\big(C\big)}.$$

We then take $\eta = \eta_b \det\big(C\big)/\det\big[(D\,\boldsymbol{w})\big]$, where η_b becomes a parameter of the sampler.

We propose moves of type b) and c) with probabilities $p_{b,j}$ and $p_{c,j}$, respectively, which depend on j, p_a and p_d $(0 < p_a + p_d < 1)$: $p_{b,j} = (1 - p_a - p_d)/2$, $p_{c,j} = (1 - p_a - p_d)/2$, $j \geq q$ and $p_{b,q-1} = (1 - p_a - p_d)$, $p_{c,q-1} = 0$; for the fixed knots prior take $p_d = 0$. These choices make sure if there are no inner knots in the current state, no knot is removed. For moves of type b), the acceptance probability of the proposed state $\big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\big)$ is as follows:

$$\min\left(1,\, \frac{\bar{\pi}\big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\,|\,\boldsymbol{X}^{(n)}\big)p_{c,j+1}}{\bar{\pi}\big(j, \boldsymbol{k}, \boldsymbol{\theta}\,|\,\boldsymbol{X}^{(n)}\big)p_{b,j}\varphi_1(\rho)}\eta_b\right) \quad \text{and}$$
$$\min\left(1,\, \frac{\tilde{\pi}\big(j+1, \boldsymbol{\kappa}, \boldsymbol{\vartheta}\,|\,\boldsymbol{X}^{(n)}\big)p_{c,j+1}(j-q+1)^{-1}}{\tilde{\pi}\big(j, \boldsymbol{k}, \boldsymbol{\theta}\,|\,\boldsymbol{X}^{(n)}\big)p_{b,j}\varphi_1(\rho)}\eta_b\right),$$

for the fixed knots prior and the free knots prior, respectively. Moves of type c) are simply the reverse move to a move of type b), so we omit the details. For this type of move, we remove one knot from the current state of the chain, uniformly at random, and recompute the spline coefficients via the inverse of the mapping g.

We let both MCMC samplers run for the same number of iterations, both starting from the state $\big(40, (1/37, 2/37, \ldots, 36/37), \mathbf{0}\big)$ which corresponds to a constant function

equal to zero with 38 equally spaced inner knots. We then collect 10.000 states from each chain. The results of the MCMC procedures are given in Figures 2 and 3. In both priors we use cubic splines ($q = 4$) and $n = 1000$. We take for the fixed knots prior $\nu = 40$, $M = 15$, $p_a = 0.5$, $p_d = 0$, $\eta_a = 1.12 \times 10^{-1}$, $\eta_b = 3 \times 10^{-2}$ and $\eta_d = 0$. For the free knots prior, we choose $\nu = 40$, $M = 15$, $p_a = 0.66$, $p_d = 0.33$, $\eta_a = 1.18 \times 10^{-1}$, $\eta_b = 6 \times 10^{-3}$ and $\eta_d = 4 \times 10^{-2}$.

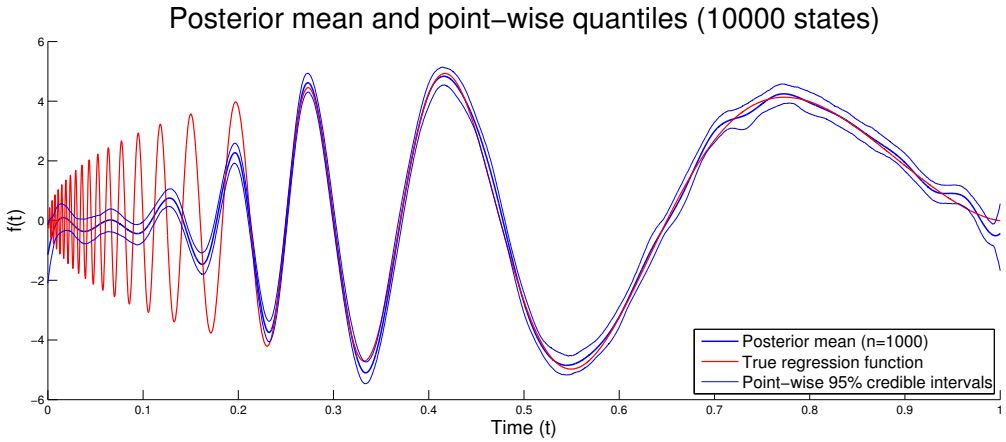

Figure 2: Results of the MCMC sampler for the fixed knots prior: posterior mean and respective 95% point-wise credible bands (in blue) and the true regression function (in red).

For both priors, we compute the proposed spline coefficients in the same way (described above), for the sake of comparing their performance. This is, however, not strictly necessary for the free knots prior. In this case, the insertion of a new knot has only a local effect on the spline: if all coefficients are kept the same, it is simple to propose a reasonable procedure for the new coefficient associated with the new added B-spline. In the case of the free knots prior, adding and removing knots from the current state of the chain can be made in a straightforward and computationally efficient way which does not involve recomputing all of the coefficients of the spline in the proposal.

As the simulations results in Figures 2 and 3 show, the free knots prior seems to outperform the fixed knots prior: the free knots posterior detects better the high and low variability regions of the regression function and facilitates the placement of more knots in the high variability region. In its turn, the fixed knots prior uses roughly 25% more knots to actually achieve a worse fit: 29 knots for the fixed knots posterior against about 23 knots for the free knots posterior. The fixed knots posterior fails to assign a number of knots that is compatible with the (inhomogeneous) structure of the true regression function f over the whole interval $[0, 1]$. As a consequence, the posterior seems to compromise on a number of knots which is clearly not sufficient for the high variability region close to zero (resulting in oversmoothing) and excessive for the low variability region close to 1 (undersmoothing the data).

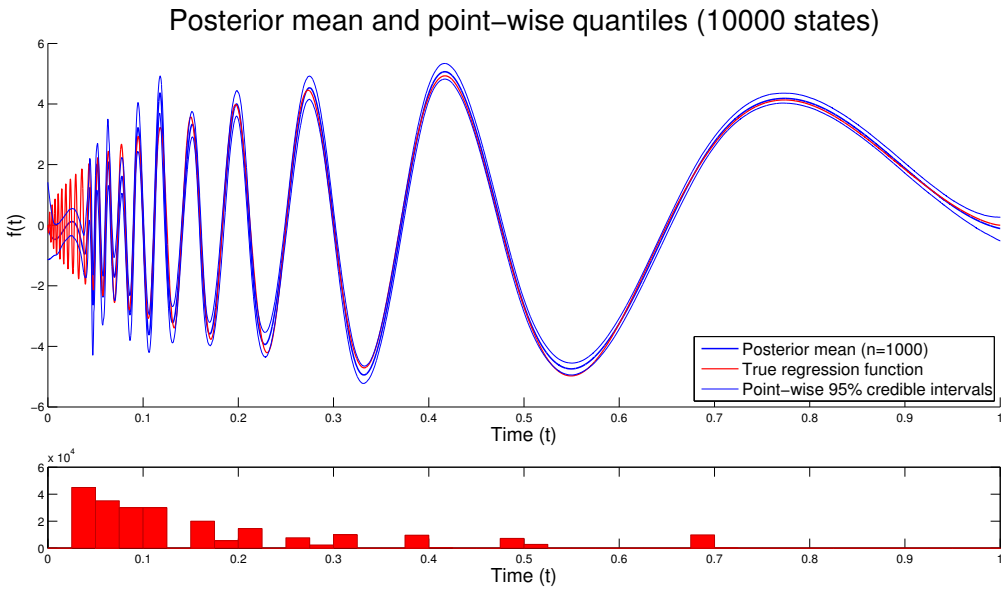

Figure 3: Results of the MCMC sampler for the free knots hierarchical prior. Above: posterior mean and respective 95% point-wise credible bands (in blue) and the true regression function (in red). Below: histogram of all the knots in all the sampled states.

Bayesian analysis based on the free knots prior has the advantage of providing relevant information about how the posterior chooses to place the knots. The bottom display of Figure 3 clearly shows a concentration of knots close to 0. This concentration, accompanied with the wider credible bands in the top display, suggests that the regression function is more variable ("volatile") in this region. Interestingly, the free knots posterior puts no mass on a sufficiently small vicinity of zero, which is sensible as well: clearly, it is impossible to reasonably estimate the Doppler function in a small vicinity of zero for a finite sample size n. This posterior distribution on knots locations can be used to make an inference on the variability (smoothness inhomogeneity, volatility) of the underlying function and to try and improve estimation procedures.

In fact, this leads to the following data-driven, empirical procedure for selecting a more appropriate prior on the location of the knots: sample $j - q$ knots from the empirical knot distribution presented in the bottom display of Figure 3 instead of our original prior on knots. We re-ran the MCMC procedure using such a prior on the knots. Actually, since some regions of the support had no knots, we constructed a new empirical (Bayes) prior for drawing one knot by mixing the distribution presented in the histogram in Figure 3 with a uniform distribution on $[0, 1]$, with a small mixture weight. This was done to assign a positive (although small) mass to the knot locations over the entire support of the regression function. The results are given in Figure 4. This data driven prior, at least in our numerical study, does not seem to improve significantly upon the

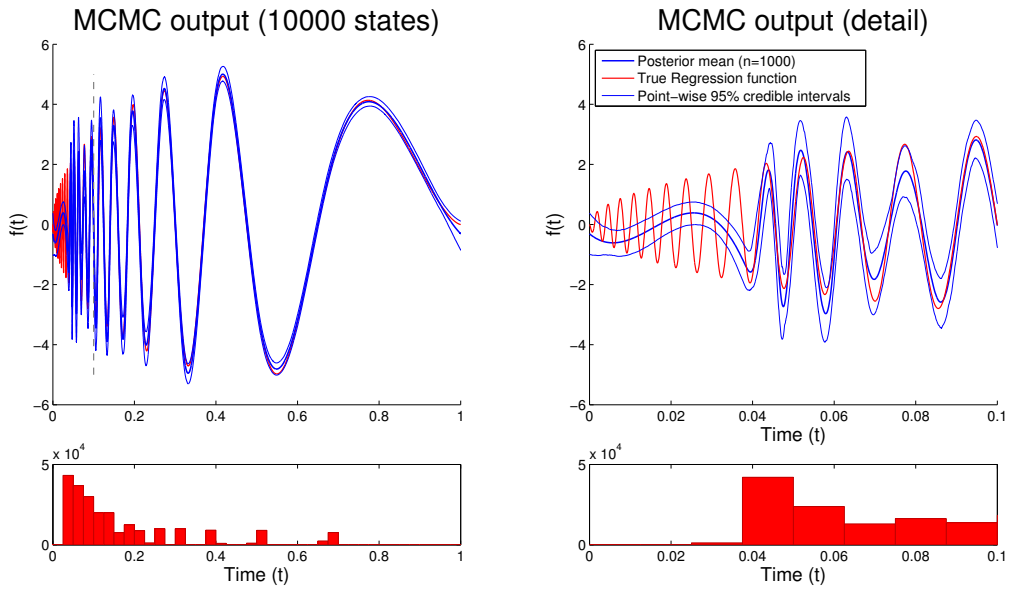

Figure 4: Results of the MCMC sampler for the free knots prior with a data driven prior on the knots locations. On the right side we display the same figures as on the left side, but on the interval $[0, 0.1]$. Above: posterior mean and respective 95% point-wise credible bands (in blue) and the true regression function for comparison (in red). Below: histogram of all the knots in all the sampled states.

free knots prior. This might simply mean that the free knots prior is already managing the location of the spline knots adequately, and reinforcing this via a data driven prior does not give an extra advantage in the inference procedure. Note that Theorem 1 may still be applied to such a data driven prior so that the resulting posterior retains (at least) the same theoretical properties as the free knots posterior.

Remark 6. *To summarize the above discussion, one can obtain the two stage sampler via the following procedure: a) split the dataset into two sets of observations; b) at the first stage, run the MCMC procedure on half of the data to obtain the posterior for the knots and use this to construct an empirical distribution for the knots; c) construct a new prior, using the empirical distribution of the knots obtained from the first sampler, possibly mixed (with a small mixture weight) with some "regular" prior distribution on knots (for example, the prior on knots from the first stage); d) at the second stage, run the MCMC sampler on the remaining data with the prior described in c).*

7 Concluding Remarks

Practitioners have already been using random knots when applying Bayesian methods in spline regression models, but a theoretical basis in terms of adaptive contraction rates was missing. In this paper, we provided theoretical justification of the random knots techniques for a rather general setting, which covers many known specific situations. Next, we specified certain conditions on the prior of the knots so that the resulting posterior is of the optimal rate, and at the same time it allows to model inhomogeneous variability of the curve adequately. Simulation results demonstrate that it is possible to design reversible jump MCMC algorithms to sample from the posterior that results from using the priors we consider in the paper. The procedure delivers results in a reasonable time horizon. Sampling from the posterior for the locations of knots might also provide useful information about the local regularity of the curve being estimated.

Empirical Bayes procedures are also quite popular among practitioners and, as suggested by a reviewer, it may provide an alternative approach to knot selection. One could treat the number and the locations of the knots as parameters of the prior distribution, and try to find their estimators by using, for example, a marginal likelihood for the knots which can be obtained by integrating out all other parameters of the model. However, as far as we know, the empirical Bayes approach for choosing the knots has never been treated theoretically in the literature and perhaps for good reason – a proper theoretical treatment of this approach in at least some generality seems to be a very complicated problem at the moment. The state of the art is not advanced enough in this area, one needs to develop an appropriate framework. It seems also that stringent conditions on the likelihood in the model are needed. In fact, theoretical treatment of empirical Bayes procedures in terms of optimal posterior contraction rates is already very challenging in the simplest nonparametric model (Gaussian white noise model), when the hyper-parameter is just one dimensional, as is demonstrated by Szabó et al. 2013. In our case, the hyper-parameter, the locations of the knots, would even have growing dimension.

As to the computational aspect of empirical Bayes methods for selecting the knots, clearly such methods can bypass sampling algorithms. This does not mean that the issue of computational complexity is resolved. Loosely speaking, the more flexibility (in our case, choosing the knots) we want to model, the more complexity we will have to deal with, either full Bayes or empirical Bayes. An empirical Bayes procedure invariably involves solving some optimization problem – typically, the maximization of the marginal likelihood with respect to the knots. The marginal likelihood must be computable, which can be achieved by using either approximations (e.g., Laplace approximation) or conjugate priors. This limits the range of models and priors for which the empirical Bayes approach can be applied. Another computational difficulty is that the marginal likelihood cannot be assured to be a convex function of the knots. The parameter space over which this problem has to be solved is of growing dimension, whereas solving a non-convex optimization problem is in general not computationally feasible even in low dimensions. It seems that in some situations an empirical Bayes approach can lead to computationally attractive procedures, further research in this area is needed.

8 Technical results

In this section we collect some technical results. Lemmas 1 and 2 are needed to bound the entropy number of the sieves $\mathcal{S}_n$ in Theorem 1. Lemma 3 claims in essence that if some bounds on the range of the function f_0 are known, then this knowledge can be incorporated into the prior on the coefficients $\boldsymbol{\theta}$.

Theorem 4.26 of Schumaker (2007) claims that if all the inner knots of a B-spline are simple, then the B-spline is continuous, uniformly over its support, with respect to its knots. In Lemma 2 we establish a slightly stronger result (a Lipschitz-type property): if we take two splines with the same coefficients in their respective B-spline basis, then the L_∞ distance between the splines can be bounded by a multiple of the l_∞ distance between the two sets of knots, as long as the sets of knots are sufficiently sparse. First, we present a preliminary lemma. Denote the $(r+1)$-th order divided difference of a function h over the points $t_1 \le \cdots \le t_{r+1}$ as $[t_1, \ldots, t_{r+1}]h = ([t_2, \ldots, t_{r+1}]h - [t_1, \ldots, t_r]h)/(t_{r+1} - t_1)$, with $[t_i]h = h(t_i)$. If $t_1 = \cdots = t_{r+1}$, then define $[t_1, \ldots, t_{r+1}]h = h^{(r)}(t_1)/r!$ for a function h with enough derivatives at t_1.

Lemma 1. *Let $i \in \{1, \ldots, r\}$, $r \ge 2$, $(k_1, \ldots, k_{r+1}) \in (0,1)^{r+1}$. Assume $k_{v+1} - k_v \ge \delta > 0$ for $v = 0, \ldots, i-1, i+1, \ldots, r$ and $k_{i+1} - k_i = 0$. For fixed $x \in [0,1]$ take the function $h(y) = (x-y)_+^{q-1}$ with $y \in [0,1]$ and $q \ge 2$. Then the divided difference $\left|[k_1, \ldots, k_{r+1}]h\right| \le 4/\delta^r$ for $x \ne k_i$ and any $\delta \le 2/(q-1)$.*

Proof. Notice that $|h'(y)| = (q-1)(x-y)_+^{q-2} \le (q-1) \le 2/\delta$ for $x \ne y$, as $q \ge 2$ and $\delta \le \frac{2}{q-1}$. Next, if $v = i-1$, $\left|[k_{v+1}, k_{v+2}]h\right| = |h'(k_{v+1})| \le 1/\delta$; if $v \ne i-1$, $\left|[k_{v+1}, k_{v+2}]h\right| = |h(k_{v+2}) - h(k_{v+1})|/|k_{v+2} - k_{v+1}| \le 2/\delta$. We conclude $\left|[k_{v+1}, k_{v+2}]h\right| \le 2/\delta$ as long as $x \ne k_i$.

For $j = 2, \ldots, r$, define $\gamma_j = \min_{v=1,\ldots,r+1-j} |k_{v+j} - k_v| \ge (j-1)\delta$. Now we make use of Theorem 2.56 from Schumaker (2007) and the previous bound:

$$\left|[k_1, \ldots, k_{r+1}]h\right| \le \sum_{v=0}^{r-1} \binom{r-1}{v} \frac{\left|[k_{v+1}, k_{v+2}]h\right|}{\gamma_2 \ldots \gamma_r} \le \frac{2^r}{(r-1)!\delta^r} \le \frac{4}{\delta^r}$$

holds for all $x \ne k_i$. This completes the proof of the Lemma. □

Recall that $s_{\boldsymbol{\theta},\boldsymbol{k}}(x)$, $x \in [0,1]$, is a spline of order $q \ge 2$ with the coordinates $\boldsymbol{\theta}$ in the B-spline basis and the inner knots vector $\boldsymbol{k}$.

Lemma 2. *Let $\boldsymbol{\theta} \in \mathbb{R}^j$ satisfy $\|\boldsymbol{\theta}\|_\infty \le M$ and let $\boldsymbol{k}, \boldsymbol{k}' \in \mathcal{K}_j^\delta = \{\boldsymbol{k} \in \mathcal{K}_j : m(\boldsymbol{k}) \ge \delta\}$. Then $\|s_{\boldsymbol{\theta},\boldsymbol{k}} - s_{\boldsymbol{\theta},\boldsymbol{k}'}\|_\infty \le L\|\boldsymbol{k} - \boldsymbol{k}'\|_\infty$, for $L = 4j(q+1)M\delta^{-(q+1)}$ and any $\delta \le 2/(q-1)$.*

Proof. Define $\boldsymbol{k}^l = (k_1^l, \ldots, k_{j-q}^l) = (k_1', \ldots, k_l', k_{l+1}, \ldots, k_{j-q})$ for $l = 0, \ldots, j-q$, such

that $\boldsymbol{k}^0 = \boldsymbol{k}$ and $\boldsymbol{k}^{j-q} = \boldsymbol{k}'$. By (1) and the triangle inequality, we get

$$\begin{aligned}\big\|s_{\boldsymbol{\theta},\boldsymbol{k}} - s_{\boldsymbol{\theta},\boldsymbol{k}'}\big\|_\infty &= \Big\|\sum_{i=1}^{j}\theta_i B_i^{\boldsymbol{k}^0} - \sum_{i=1}^{j}\theta_i B_i^{\boldsymbol{k}^{j-q}}\Big\|_\infty \le M\Big\|\sum_{i=1}^{j}(B_i^{\boldsymbol{k}^0} - B_i^{\boldsymbol{k}^{j-q}})\Big\|_\infty \\ &\le jM \max_{1\le i\le j}\big\|B_i^{\boldsymbol{k}^0} - B_i^{\boldsymbol{k}^{j-q}}\big\|_\infty \le jM \max_{1\le i\le j}\sum_{l=0}^{j-q-1}\big\|B_i^{\boldsymbol{k}^l} - B_i^{\boldsymbol{k}^{l+1}}\big\|_\infty \\ &\le (q+1)jM \max_{1\le i\le j}\max_{0\le l\le j-q-1}\big\|B_i^{\boldsymbol{k}^l} - B_i^{\boldsymbol{k}^{l+1}}\big\|_\infty.\end{aligned}$$

The last inequality in the above display follows from (1). Indeed, the inner knots of $B_i^{\boldsymbol{k}^l}$ and $B_i^{\boldsymbol{k}^{l+1}}$ differ only at the $(l+1)$-th entry. Therefore, according to (1), for each i there are at most $(q+1)$ nonzero terms $\big\|B_i^{\boldsymbol{k}^l} - B_i^{\boldsymbol{k}^{l+1}}\big\|_\infty$ in the last sum.

Theorem 4.27 of Schumaker (2007) gives explicit expressions for the derivative of a B-spline with respect to one of its knots. These expressions are in terms of the divided differences which satisfy the conditions of Lemma 1, so that combining this with Lemma 1 for $r = q+1$ (the maximal number of knots in the support of a B-spline) yields that this derivative is bounded in absolute value by $4\delta^{-(q+1)}$, except at $x = k_{l+1}^l$, where it is not defined. Then, as $\|\boldsymbol{k}^l - \boldsymbol{k}^{l+1}\|_\infty \le \|\boldsymbol{k} - \boldsymbol{k}'\|_\infty$, we obtain that, for $x \ne k_{l+1}^l$, $l = 0, \dots, j-q-1$,

$$\big|B_i^{\boldsymbol{k}^l}(x) - B_i^{\boldsymbol{k}^{l+1}}(x)\big| \le |k_{l+1}^{l+1} - k_{l+1}^l| \sup_{k_{l+1}^l \in (0,1)}\Big|\frac{\partial B_i^{\boldsymbol{k}^l}(x)}{\partial k_{l+1}^l}\Big| \le \frac{4\|\boldsymbol{k} - \boldsymbol{k}'\|_\infty}{\delta^{q+1}}.$$

Since splines are continuous for all $q > 1$, so is $s_{\boldsymbol{\theta},\boldsymbol{k}} - s_{\boldsymbol{\theta},\boldsymbol{k}'}$ and we conclude that the same bound must also hold for $x = k_{l+1}^l$. Combining the above two relations concludes the proof. □

The properties of B-splines allow to relate the range of the coefficients of the approximating spline to the range of the approximated function. The following lemma generalizes Lemma 1 of Shen and Ghosal (2012) for non-equally spaced knots.

Lemma 3. *Let $f \in \mathcal{F}_\alpha$ (so that (3) holds), $a < b$, $\varepsilon > 0$. Assume that $f(x) \in [a+\varepsilon, b-\varepsilon]$ for all $x \in [0,1]$. Then there exists a positive constant $\delta = \delta(\mathcal{F}_\alpha, \varepsilon)$ such that for any $\boldsymbol{k} \in \mathcal{K}_j$, $j \ge q$, such that $M(\boldsymbol{k}) \le \delta$, the coefficients $\boldsymbol{a}$ of the approximating spline $s_{\boldsymbol{a},\boldsymbol{k}}$ in (3) can be taken to be contained in (a,b).*

Proof. Fix q, j and inner knots $\boldsymbol{k}$, assume $I = [a,b]$, $a < b$ and $a + \varepsilon < f < b - \varepsilon$, for some $\varepsilon > 0$.

We use results from section 4.6 of Schumaker (2007) on dual basis of B-splines. If $B_1^{\boldsymbol{k}}, \dots, B_j^{\boldsymbol{k}}$ is the B-spline basis associated with the inner knots $\boldsymbol{k}$, then there exists a dual basis $\lambda_1, \dots, \lambda_j$ of linear functionals such that, for each $i, r = 1, \dots, j$, $\lambda_r B_i^{\boldsymbol{k}} = 1$ if $i = r$ and is 0 otherwise. As a consequence, we obtain that $\lambda_i s_{\boldsymbol{a},\boldsymbol{k}} = a_i$, and since $\sum_{i=1}^j B_i^{\boldsymbol{k}}(x) = 1$, it follows that $\lambda_i c = c$ for any constant c and all $i = 1, \dots, j$. This dual

basis is not necessarily unique and, according to Theorem 4.41 from Schumaker (2007), can be taken such that $|\lambda_i f| \leq C_1 \sup_{x\in I_i} |f(x)|$ where I_i represents the support of $B_i^{\boldsymbol{k}}$ and constant C_1 depends only on q. Each I_i consists of at most q adjacent intervals in the partition induced by $\boldsymbol{k}$ and thus the length of I_i is bounded by $qM(\boldsymbol{k})$.

Let $s_{\boldsymbol{a},\boldsymbol{k}}$ be such that (3) is fulfilled for f. Then for any constant c

$$\begin{aligned} |a_i - c| &= \left|\lambda_i s_{\boldsymbol{a},\boldsymbol{k}} - \lambda_i f + \lambda_i f - c\right| \leq \left|\lambda_i (s_{\boldsymbol{a},\boldsymbol{k}} - f)\right| + \left|\lambda_i (f - c)\right| \\ &\leq C_1 C_f M^{\alpha}(\boldsymbol{k}) + C_1 \sup_{x\in I_i} |f(x) - c|. \end{aligned}$$

Take $c = \inf_{x\in I_i} f(x)$ and recall that $f \in \mathcal{F}_\alpha \subseteq \mathcal{L}(\kappa_\alpha, L_\alpha)$. Using the Lipschitz property, we derive that $\sup_{x\in I_i} |f(x) - c| = \sup_{x\in I_i} f(x) - \inf_{x\in I_i} f(x) \leq L_\alpha (q\, M(\boldsymbol{k}))^{\kappa_\alpha}$ and therefore

$$|a_i - \inf_{x\in I_i} f(x)| \leq C_1 C_f M^{\alpha}(\boldsymbol{k}) + C_1 L_\alpha (q\, M(\boldsymbol{k}))^{\kappa_\alpha} \leq C_2 M^{\alpha\wedge\kappa_\alpha}(\boldsymbol{k}).$$

In the same way, if we take $c = \sup_{x\in I_i} f(x)$, we derive that $\sup_{x\in I_i} |f(x) - c| \leq L_\alpha (q\, M(\boldsymbol{k}))^{\kappa_\alpha}$ and thus $\left|a_i - \sup_{x\in I_i} f(x)\right| \leq C_2 M^{\alpha\wedge\kappa_\alpha}(\boldsymbol{k})$.

Now for $\delta = (\varepsilon/(2C_2))^{1/(\alpha\wedge\kappa_\alpha)}$ conclude that if $M(\boldsymbol{k}) \leq \delta$, then $a_i \geq \inf_{x\in I_i} f(x) - C_2 M^{\alpha\wedge\kappa_\alpha}(\boldsymbol{k}) \geq \inf_{x\in I_i} f(x) - \varepsilon/2 > a$. For the same choice of δ we have $a_i \leq \sup_{x\in I_i} f(x) + C_2 M^{\alpha\wedge\kappa_\alpha}(\boldsymbol{k}) \leq \sup_{x\in I_i} f(x) + \varepsilon/2 < b$.

□

References

Belitser, E. and Ghosal, S. (2003). "Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution." *Annals of Statistics*, 31: 536–559. 860

De Boor, C. (1978). *A practical guide to splines*. Springer-Verlag, New York. 860, 862

Brooks, S. P., Giudici, P., and Roberts, G.O. (2003). "Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions." *Journal of the Royal Statistical, Society Series B (Statistical Methodology)*, 65: 3–55. 873

De Jonge, R. and van Zanten, H. (2012). "Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors." *Electronic Journal of Statistics* 6: 1984–2001. 860, 861, 866

Denison, D., Mallick, B., and Smith, A. (1998). "Bayesian MARS." *Statistics and Computing*, 8: 337–346. 861

Di Matteo, I., Genovese, C., and Kass, R. (2001). "Bayesian curve-fitting with free-knot splines." *Biometrika*, 88: 1055–1071. 861

Ghosal, S., Ghosh, J., and van der Vaart, A. (2000). "Convergence rates of posterior distributions." *Annals of Statistics*, 28: 500–531. 859, 860, 866

Ghosal, S. and van der Vaart, A. (2001). "Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities." *Annals of Statistics*, 29: 1233–1263. 866

Green, P.J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination." *Biometrika*, 82: 711–732. 870

Schumaker, L. (2007). *Spline functions: basic theory*. John Wiley & Sons, New York. 861, 862, 878, 879, 880

Sharef, E., Strawderman, R., Ruppert, D., Cowen, M., and Halasyamani, L. (2010). "Bayesian adaptive B-spline estimation in proportional frailty models." *Electronic Journal of Statistics*, 4: 606–642. 861

Shen, W. and Ghosal, S. (2012). "MCMC-free adaptive Bayesian procedures using random series prior." *Preprint arXiv:1204.4238*. 860, 861, 866, 868, 879

Szabó, B., van der Vaart, A., and van Zanten, H. (2013). "Emprirical Bayes scaling of Gaussian priors in the white noise model." *Electronic Journal of Statistics*, 7: 991–1018. 877

Van der Vaart, A. and van Zanten, H. (2008). "Rates of contraction of posterior distributions based on Gaussian process priors." *Annals of Statistics*, 36: 1435–1463. 860

Van der Vaart, A. and van Zanten, H. (2009). "Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth." *Annals of Statistics*, 37: 2655–2675. 860