Nonparametric adaptive Bayesian regression with tractable normalizing constants under qualitative assumptions

Khader Khadraoui

Laval University, Department of Mathematics and statistics, Québec city G1V 0A6, Canada

Abstract

Shape constrained regression models are useful for analyzing data with specific shape responses, such as (monotone) dose-response curves, the (concave) utility functions of a risk averse decision maker, the (increasing) growth curves of children's height through time, that are particularly common in medicine, economic and epidemiological studies. This paper proposes a new adaptive Bayesian approach towards constructing prior distributions, with known normalizing constants, which enables us to take into account combinations of shape constraints and to localize each shape constraint on a given interval. Our strategy enables us to compute the simulation from the posterior distribution using a reversible jump Metropolis-Hastings scheme. The major advantages of the proposal are its flexibility achieved by adjusting local shape restrictions to detect better the high and low variability regions of the data and to facilitate the control of the regression function shape when there is no data at all in some regions. We give asymptotic results that show that our Bayesian method provide consistent function estimator from the adaptive prior. The performance of our method is investigated through a simulation study with small samples. An analysis of two real data sets are presented to illustrate the new approach.

Keywords: Bayesian regression, Localized shape constraints, Free-knot B-spline, Reversible jump MCMC scheme.

1. Introduction

The problem of estimating functions under qualitative assumptions has a long history dating back at least to [15]. Regression under shape and smoothness constraints is of considerable interest both for theoretical and practical reasons. From a theoretical point of view, shape constraints and smoothing usually reduce the variance of the estimators. From a practical point of view, there are many situations in which it is necessary to take into account some prior knowledge on the shape of the regression function. The extensive literature on shape constrained problems has partly been motivated by specific applications but also by the fact that has features that are shared with non-parametric function estimation.

As smoothing splines are defined as minimizers of a penalized sum of squares, they provide a natural theoretical framework to incorporate shape constraints simply by restricting the minimization over the set of constrained estimates [32, 23, 22]. Nevertheless, minimizing over a restricted set may lead to numerical difficulties in practice. Another strategy for monotone or convex regressions is to decompose the regression function into a tailored basis of splines as in [27] and [24]. The constraints are then taken into account simply by restricting the coefficients to be nonnegative. In the two papers just cited, the knots of the basis spline functions are fixed as it is usually the case in the frequentist literature on shape restricted regression. Regression splines with fixed knots also appear in the Bayesian literature on shape restricted regression. In these cases, the number of knots is typically large. A roughness penalty can be introduced from the prior distribution in order to avoid overfitting as in [6] and [1]. In [26] and Shively & Sager [31, Section 3], indicator variables are introduced in the model for selecting a subset of the basis functions. Usually, the constraint is included by restricting the spline coefficients to some subset.

Another way of selecting the number and the position of the knots is to use the Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm introduced by [14]. Recall that a RJMCMC algorithm is a Metropolis-Hastings algorithm whose acceptance probability involves the computations of the ratio of priors and the ratio of likelihoods for two sets of parameters whose dimension may differ. In the unconstrained case, [8] and [10] get rid of the coefficients of the basis functions and run a RJMCMC algorithm for sampling from the posterior marginal distribution of knots only while [21] use a RJMCMC algorithm with all the parameters. [36] follows the former strategy but in the monotone constrained case. Contrary to unconstrained situation, the conditional posterior distribution of the data given the knots cannot be computed analytically and the likelihood ratio is not exactly known. The latter strategy [21] can also induce numerical difficulties in the computation of the prior ratio. In some convenient situations, the constraint on the coefficients is simple and the prior distribution is completely known. This is the case in [25] where the coefficients are constrained to be nonnegative. Then, a Gamma distribution for each coefficient fulfills the constraint and the prior ratio can easily be computed even for coefficients vectors with different dimensions. The same happens in [5] but with a different model. However, in these two papers, the support of the basis functions is the whole interval of the data. Then, such models do not enable us to localize the constraint or to consider a combination of constraints like being increasing on $[1, 2]$ and concave on $[3, 4]$. Such combinations of constraints have been undertaken by using B-splines with fixed knots in [2]. Free-knot B-splines for monotone regression are used in [17] and [18]. With B-splines, the constraint on coefficients is typically not as simple as forcing the coefficients to be nonnegative. Thus, it is usually included in the prior by means of truncated distributions like the truncated Gaussian distributions and the prior density is only known up to the normalizing constant which does change with the dimension of the coefficients vector. Thus, simplification of the unknown normalizing constants in the prior ratio cannot be done and the prior ratio remains typically unknown. This feature was already pointed out by [14] and it is apparently missed by [18]. Note that this problem can be avoided by running the algorithm in the unconstrained case and retaining only the samples for which the constraint is fulfilled as in [17]. Nevertheless, this strategy may behave poorly if the shape constraint is not sufficiently supported by the data.

In the present paper a new spline framework is introduced for estimating function under localized shape restrictions with unknown knot vector. This framework is focused on the construction of an adaptive prior that extends to estimating functions with all or at a least large classes of shape restrictions as well as combination of several constraints. There are two main goals of the present paper. The first is, as the dimension of the model parameters is unknown, to introduce a new construction of a truncated normal prior distribution with known normalizing constant for usual shape constraints. The idea is this prior enables us to compute the prior ratio of the acceptance probability of the RJMCMC algorithm used for sampling from the posterior distribution. The second goal is, when the regression functions are uniformly bounded and the design points are random, to prove almost sure consistency of the posterior probabilities of Kullback-Leibler and Hellinger neighborhoods of the joint density of the response and design point.

The paper is organized as follows. Section 2 introduces and discusses the concepts of B-spline and the control polygon. In Section 3, we introduce the model and the associated prior. Section 4 is devoted to the reversible jump Markov chain Monte Carlo sampling scheme to implement the method. Section 5 discusses the asymptotic properties of the posterior distribution. Section 6 presents simulation results to show the small sample properties across examples of functions as well as applications based on real world data sets. Auxiliary results and proofs (except for Lemma 2 and Theorem 2) are gathered in Section 7. Section 8 contains a short discussion.

2. Notations and preliminaries

The intent in this section is to give a simple and direct development for B-splines via the recurrence relations. Let the integer k denote the order of the B-spline. Given a nondecreasing sequence of points $\{t_j \in \mathbb{R} | t_j \leq t_{j+1}\}$ called knots, the B-spline function of order 1 is given by: $B_{j1}(x) = 1$ if $t_j \leq x < t_{j+1}$ and $B_{j1}(x) = 0$ otherwise. From the first-order B-splines, we obtain higher-order B-splines by recurrence:

$$B_{jk}(x) = \omega_{jk}(x)\, B_{j,k-1}(x) + \Big(1 - \omega_{j+1,k}(x)\Big)\, B_{j+1,k-1}(x), \qquad (2.1)$$

with

$$\omega_{jk}(x) = \begin{cases} \frac{x-t_j}{t_{j+k-1}-t_j}, & \text{if } t_j < t_{j+k-1} \\ 0, & \text{otherwise.} \end{cases} \qquad (2.2)$$

From this, we infer that B_{jk} is a piecewise polynomial of degree $< k$ which vanishes outside the interval $[t_j, t_{j+k})$. In particular, B_{jk} is just the zero function in case $t_j = t_{j+k}$. Also, by induction, B_{jk} is positive on the open interval (t_j, t_{j+k}), since both ω_{jk} and $(1 - \omega_{j+1,k})$ are positive there. We refer the reader to de Boor [7] for a thorough presentation of the many other properties of B-splines. A spline of order k with knot sequence $t = (t_j)$ is, by definition, a linear combination of the B-splines B_{ik} associated with that knot sequence. The control of the spline shape suggests consideration of the control polygon;

Definition 1. *The control polygon associated with the representation $\sum_j \beta_j B_{jk}$, for $\beta_j \in \mathbb{R}$, is the broken line or piecewise linear function that interpolates the vertices (control points) P_j defined by*

$$P_j = (t_j^*, \beta_j), \tag{2.3}$$

where

$$t_j^* = (t_{j+1} + \cdots + t_{j+k-1})/(k-1). \tag{2.4}$$

This control polygon will be denoted by $C_{\beta,t}$. From the definition 1, the control polygon is a 2-order spline given by:

$$C_{\beta,t}(x) = \sum_j \mathbf{1}_{[t_j^*, t_{j+1}^*)} \Big\{ \Big(1 - \frac{x - t_j^*}{t_{j+1}^* - t_j^*}\Big)\beta_i + \Big(\frac{x - t_j^*}{t_{j+1}^* - t_j^*}\Big)\beta_{j+1} \Big\},$$

where $\mathbf{1}_A$ denotes the indicator function of a set A. The curve obtained is continuous but not differentiable.

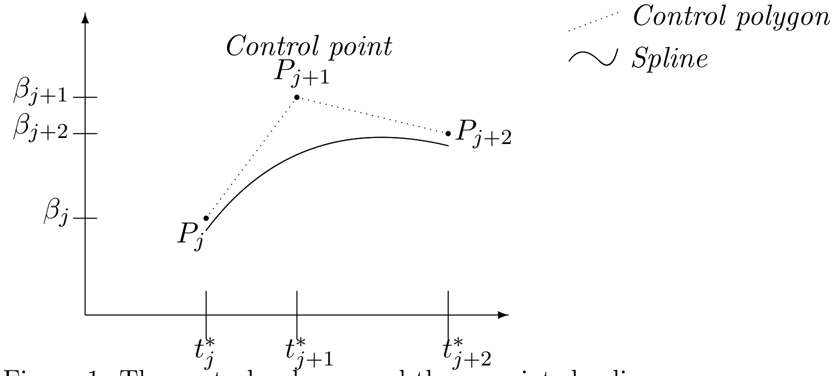

Figure 1: The control polygon and the associated spline.

Since the B-spline order, k, will not change in the remainder of the paper, we will usually suppress it and write B_j instead of B_{jk}, ω_j instead of ω_{jk}, etc.

3. A prior distribution with known normalizing constant

Consider the usual regression model with n independent observations $(x_i, y_i) \in [a_0, b_0] \times \mathbb{R}$, $i = 1, \ldots, n$:

$$y_i | x_i, f, \sigma^2 \sim N(f(x_i), \sigma^2). \tag{3.1}$$

Denote by $\mathcal{S}$ the set of functions with a given constraint on $[a, b] \subseteq [a_0, b_0]$ and assume it is known that f is smooth and belongs to $\mathcal{S}$. From this prior information, we construct a prior distribution on (f, σ^2) as follows. We set:

$$f = \sum_{j=1}^{m} \beta_j B_j, \tag{3.2}$$

where $B_1, \ldots, B_m$ is the B-spline basis of order k associated with a nondecreasing sequence of knots $t = (t_1, \ldots, t_{m+k})$ such that $t_j < t_{j+k}$ for all j and $t_k = a_0$ and $t_{m+1} = b_0$. Clearly, (3.1) and (3.2) reduce to

$$y|\beta, \sigma^2, t \sim N_n(B\beta, \sigma^2 I_n), \tag{3.3}$$

where $y = (y_1, \ldots, y_n)'$, $\beta = (\beta_1, \ldots, \beta_m)'$, B is the $n \times m$ matrix whose entry (i, j) is $B_j(x_i)$ and I_n is the $n \times n$ identity matrix. By construction, f is a spline of order k with knot sequence t, i.e. a piecewise polynomial of degree $< k$ with breakpoints t_j which is $k - 1 - \#t_j$ times continuously differentiable at t_j where $\#t_j$ denotes the cardinal of $\{t_l : t_l = t_j\}$. We are now in position to include the localized shape constraint in the prior distribution. For ease of exposition this idea can be explained as follows: For all $x \in [a_0, b_0]$, denote by t_{j_x} the smallest knot greater than x. From now on, it will be convenient to assume that $t_j < t_{j+1}$ for all j as it enables us to write that $t_{j_x - 1} < x \leq t_{j_x}$ for all $x \in [a_0, b_0]$. From de Boor [7], it can be deduced that the shape of f on $[a, b]$ can be controlled by the shape of the finite sequence of points $P_{j_a - k}, \ldots, P_{j_b - 1}$, where $P_j = (t_j^*, \beta_j)$. Note that t_j^* is simply the arithmetic mean of the knots which are in the interior of the support of B_j and that $B_{j_a - k}, \ldots, B_{j_b - 1}$ are precisely the B-splines whose support intersects with $[a, b]$. More precisely, we have that:

- if $k > 2$ and if $P_{j_a - k}, \ldots, P_{j_b - 1}$ is increasing (decreasing), then the spline f is increasing (decreasing) on $[a, b]$,
- if $k > 3$ and if $P_{j_a - k}, \ldots, P_{j_b - 1}$ is convex (concave), then the spline f is convex (concave) on $[a, b]$,
- if $k > 3$ and if $P_{j_a - k}, \ldots, P_{j_b - 1}$ is unimodal, then the spline f is unimodal or monotone on $[a, b]$.

As the constraints can be localized on a given interval, it is straightforward with our approach to impose combinations of constraints on possibly different intervals. Contrary to most of the methods cited in Section 1, it is worth noting that the induced constraints on the coefficients β_j are very easy to derive for any order k.

It is usual to set the prior distribution of (β, σ^2, t) by specifying the conditional distribution of β given (σ^2, t) and by taking σ^2 and t to be independent. Classically, we choose an inverse gamma prior for σ^2 and the prior for t is decomposed as follows. The exterior knots $t_1, \ldots, t_k$ and $t_{m+1}, \ldots, t_{m+k}$ are fixed and we denote by $\nu = m - k$ the number of interior knots. We use a Poisson distribution $\mathcal{P}(\lambda)$ for ν. We allow the ν knots $h = (t_{k+1}, \ldots, t_m)'$ to take positions on a larger, predetermined set of candidate positions, say $h^c = (\xi_1^c, \ldots, \xi_q^c)'$. These candidate knot specifications constitute a prior on knot placement. We denote by π_2 the joint distribution of (σ^2, t) induced by (3.4):

$$\begin{cases} \sigma^2 & \sim & IG(c,d), \\ h|\nu & \sim & \left[\binom{q}{\nu}\right]^{-1}, \\ \nu & \sim & \mathcal{P}(\lambda). \end{cases} \tag{3.4}$$

The general idea for deriving a prior density for β given (σ^2, t) in order that, firstly, the normalizing constant is known and, secondly, the constraint is fulfilled, is as follows. Let $\beta^1 = (\beta_{j_a-k}, \ldots, \beta_{j_b-1})'$ and $m_1 = j_b - j_a + k$. For most types of constraint, it is usually easy to find a function T^* from $\mathbb{R}^{m_1}$ to $\mathbb{R}^{m_*}$, with $m^* \leq m_1$, such that the shape condition on $P_{j_a-k}, \ldots, P_{j_b-1}$ is equivalent to an inequality of the form $T^*(\beta^1) \geq 0$. Such functions are given in special cases in the following. From T^*, we derive an invertible map T from $\mathbb{R}^{m_1}$ on $\mathbb{R}^{m_1}$ whose the m^* last coordinates are exactly those of T^*. Then we set $z = T(\beta^1)$ and choose a distribution for z with support $\mathbb{R}^*$ for the m^* last coordinates of z. By construction, the distribution of β^1 will necessarily fulfills the shape constraint and its density can easily be calculated. As the normalization constant is known for z, the normalization constant of $\beta^1 = T^{-1}z$ is also known. Note that the distribution of z may depend on σ^2 and t if needed. Usually, we can use Gamma distributions or independent normal distributions with variance σ^2 truncated to $\mathbb{R}_+$ for the coordinates of z. In this paper, we use independent normal distributions. Then, it remains to choose the conditional distribution of $\beta^0 = (\beta_1, \ldots, \beta_{j_a-k-1}, \beta_{j_b}, \ldots, \beta_m)'$ given (σ^2, t, β^1). We typically use a normal distribution for β^0. This strategy is illustrated for three usual constraints in the next section, namely the monotonicity, unimodality and convexity (concavity) constraints. Thus, the prior distribution can be summarized by:

$$\begin{aligned}\beta^0|\sigma^2, t, \beta^1 &\sim \pi_0,\\ \beta^1|\sigma^2, t &\sim \pi_1,\\ (\sigma^2, t) &\sim \pi_2.\end{aligned} \tag{3.5}$$

By noting that t and β are just deterministic functions of h and (β^0, z) respectively, another parametrization of the full Bayesian model is given by:

$$\begin{aligned}y|z, \beta^0, \sigma^2, h &\sim N_n(B\beta, \sigma^2 I_n),\\ \beta^0|z, \sigma^2, h &\sim \pi_0,\\ z|\sigma^2, \nu &\sim N^+_{m_1}(0, \tau^2\sigma^2 I_{m_1}),\\ \sigma^2 &\sim IG(c, d),\\ h|\nu &\sim \left[\binom{q}{\nu}\right]^{-1},\\ \nu &\sim \mathcal{P}(\lambda).\end{aligned} \tag{3.6}$$

By definition of j_a and j_b, we have that $j_b \geq j_a$, (actually, $j_a = j_b$ when there is no knots in $[a, b[$, which is allowed with our method). Thus, we always have that $m_1 \geq k$.

3.1. Monotone prior with known normalizing constant

Assume that the regression function has to be increasing on $[a, b]$ and take $k > 2$. The constraint is fulfilled as soon as $P_{j_a-k}, \ldots, P_{j_b-1}$ is increasing, i.e.

$$\beta_{j_a-k} \leq \beta_{j_a-k+1} \leq \cdots \leq \beta_{j_b-1}. \tag{3.7}$$

Clearly, (3.7) is equivalent to $T^*\beta^1 \geq 0$ where $\beta^1 = (\beta_{j_a-k}, \ldots, \beta_{j_b-1})'$, $m_1 = j_b - j_a + k$, T is the following $m_1 \times m_1$-matrix:

$$T = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0\\ -1 & +1 & 0 & \cdots & 0\\ 0 & -1 & +1 & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & 0\\ 0 & \cdots & 0 & -1 & +1 \end{pmatrix},$$

and T^* is the $(m_1 - 1) \times m_1$ matrix defined by removing the first row of T. Thus, we always have that $m_1 \geq k$ and T can always be defined since

$k > 2$. Denote by N^+ the univariate normal distribution truncated on $\mathbb{R}_+$. Take $z_1|\sigma^2 \sim N(0, \tau^2\sigma^2)$ and $z_i|\sigma^2 \sim N^+(0, \tau^2\sigma^2)$ for $i \in \{2, \dots, m_1\}$ and assume that $z_1, \dots, z_{m_1}$ are conditionally independent given σ^2. Let $T\beta^1 = z$ where $z = (z_1, \dots, z_{m_1})'$. Note that this last equality reduces to set $\beta_{j_a-k+1} = z_1,\ \beta_{j_a-k+2} = z_1 + z_2, \dots,\ \beta_{j_b-1} = z_1 + \dots + z_{m_1}$. By noting that $\int_0^\infty \exp(-z^2/2)\,\mathrm{d}z = \sqrt{\pi/2}$ and that $|Det(T)| = 1$, it is straightforward to deduce that the density, with respect to the Lebesgue measure on $\mathbb{R}^{m_1}$, of the conditional distribution of β^1 given (σ, ν) is given by:

$$\pi_1(\beta^1|\sigma^2, t) = (\pi/2)^{-m_1/2} \left(\sigma^2\tau^2\right)^{-m_1/2} \exp\left(-\frac{1}{2\sigma^2\tau^2}\beta^{1\prime}(T'T)\beta^1\right) \mathbf{1}_S(\beta^1),$$

where S denotes the set of β^1 such that (3.7) is fulfilled.

3.2. Unimodal prior with known normalizing constant

Assume that the regression function has to be unimodal on $[a, b]$ and take $k > 3$. The constraint is fulfilled as soon as $P_{j_a-k}, \dots, P_{j_b-1}$ is unimodal, i.e.

$$\beta_{j_a-k} \leq \beta_{j_a-k+1} \leq \dots \leq \beta_l \geq \beta_{l+1} \geq \dots \geq \beta_{j_b-1}, \tag{3.8}$$

for all $l \in \{j_a - k + 1, \dots, j_b - 2\}$. Clearly, (3.8) is equivalent to $T^*\beta^1 \geq 0$ where T is the following $m_1 \times m_1$-matrix:

$$T = \begin{pmatrix} +1 & 0 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ -1 & +1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & -1 & +1 & 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & +1 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & +1 & -1 \end{pmatrix}.$$

and T^* is the $(m_1 - 1) \times m_1$ defined by removing the first row of T. Note that $m_1 \geq k$ so that T can always be defined since $k > 3$. Denote by N^+ the univariate normal distribution truncated on $\mathbb{R}_+$. Take $z_1|\sigma^2 \sim N(0, \tau^2\sigma^2)$ and $z_i|\sigma^2 \sim N^+(0, \tau^2\sigma^2)$ for $i \in \{2, \dots, m_1\}$ and assume that $z_1, \dots, z_{m_1}$ are conditionally independent given σ^2. Let $T\beta^1 = z$ where $z = (z_1, \dots, z_{m_1})'$. Note that this last equality reduces to set $\beta_{j_a-k+1} = z_1,\ \beta_{j_a-k+2} = z_1 +$

$z_2, \ldots, \beta_{j_a-k+\ell} = z_1 + \cdots + z_\ell, \beta_{j_a-k+\ell+1} = z_1 + \cdots + z_\ell - z_{\ell+1}, \ldots,\ \beta_{j_b-1} = z_1 + \cdots + z_\ell - z_{\ell+1} - \cdots - z_{m_1}$. By noting that $\int_0^\infty \exp(-z^2/2)\,\mathrm{d}z = \sqrt{\pi/2}$ and that $|Det(T)| = 1$, it is straightforward to deduce that the density, with respect to the Lebesgue measure on $\mathbb{R}^{m_1}$, of the conditional distribution of β^1 given (σ, ν) is given by:

$$\pi_1(\beta^1|\sigma^2, t) = (\pi/2)^{-m_1/2} \left(\sigma^2\tau^2\right)^{-m_1/2} \ \exp\left(-\frac{1}{2\sigma^2\tau^2}\beta^{1\prime}(T'T)\beta^1\right) \mathbf{1}_S(\beta^1).$$

3.3. Concave prior with known normalizing constant

Assume that the regression function has to be concave on $[a, b]$ and take $k > 3$. The constraint is fulfilled as soon as $P_{j_a-k}, \ldots, P_{j_b-1}$ is concave. It is easy to check that this condition reduces to:

$$\mathit{\Delta}_{1,l+1}\ \beta_{l-1} - \mathit{\Delta}_{2,l}\ \beta_l + \mathit{\Delta}_{1,l}\ \beta_{l+1} \geq 0, \tag{3.9}$$

for all $l \in \{j_a - k + 1, \ldots, j_b - 2\}$ where $\mathit{\Delta}_{1,l} = t_l^* - t_{l-1}^*$ and $\mathit{\Delta}_{2,l} = t_{l+1}^* - t_{l-1}^*$. Clearly, (3.9) is equivalent to $T^*\beta^1 \geq 0$ where T is the following $m_1 \times m_1$-matrix:

$$T = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & +1 & 0 & 0 & \cdots & 0 \\ \mathit{\Delta}_{1,j_a-k+2} & -\mathit{\Delta}_{2,j_a-k+1} & \mathit{\Delta}_{1,j_a-k+1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \mathit{\Delta}_{1,j_b-1} & -\mathit{\Delta}_{2,j_b-2} & \mathit{\Delta}_{1,j_b-2} \end{pmatrix},$$

and T^* is the $(m_1 - 2) \times m_1$ defined by removing the two first rows of T. Note that $m_1 \geq k$ so that T can always be defined since $k > 3$. By noting that $|Det(T)| = |\prod_{l=j_a-k+1}^{j_b-2} \mathit{\Delta}_{1,l}| = |J_{\mathit{\Delta}}|$ and by a similar reasoning as in the monotone and unimodal shapes, it is straightforward to deduce that the density, with respect to the Lebesgue measure on $\mathbb{R}^{m_1}$, of the conditional distribution of β^1 given (σ, ν) is given by:

$$\pi_1(\beta^1|\sigma^2, t) = |J_{\mathit{\Delta}}|\,(\pi/2)^{-m_1/2} \left(\sigma^2\tau^2\right)^{-m_1/2} \ \exp\left(-\frac{1}{2\sigma^2\tau^2}\beta^{1\prime}(T'T)\beta^1\right) \mathbf{1}_S(\beta^1).$$

Until now in this paper, a little attention has been paid to the choice of the number of interior spline knots ν and their positions h. Both have a profound effect on the smoothness of the estimated regression function f. An advantage of the Bayesian approach is that the number and positions of knots may be treated as unknown parameters which permit the data to help determine the best choices for the knots. However, such additional flexibility making estimation via MCMC will become more complicated in comparison with the fixed knot setting because the total parameter dimension is not fixed a priori, but changes across successive MCMC iterations. Further details on the MCMC procedure are provided in Section 4.

4. Sampling from the posterior distribution

The scheme used to sample from the joint posterior distribution consists of a trans-dimensional Metropolis-Hastings algorithm [14]. In order to enable adaptive knot selection, we not only allow knots to move, but also permit changes in dimension, such as deleting a knot (death step) or adding a knot (birth step). Of course, the procedure is identical for spline coefficients. From now on, let $\nu^{(\kappa)}$ denote the number of interior spline knots $h^{(\kappa)}$, and $\beta^{(\kappa)}$ the spline coefficients vector of length $\nu^{(\kappa)} + k$, at iteration κ. Following [8], at each iteration we choose randomly whether to execute a birth, death, or move step. The probabilities b_ν, d_ν, m_ν of birth, death and move steps respectively are set to: $b_\nu = e \min\left\{1, \frac{p(\nu+1)}{p(\nu)}\right\}$, $d_\nu = e \min\left\{1, \frac{p(\nu-1)}{p(\nu)}\right\}$, $m_\nu = 1 - b_\nu - d_\nu$, where $e = 0.4$ as in [8]. These parameters are chosen so that

$$b_\nu p(\nu) = d_{\nu+1} p(\nu + 1). \tag{4.1}$$

In the sequel, we give details on the birth, death and move step.

4.1. Insertion of a knot and a coefficient (birth step)

In the birth move, a random candidate knot $\xi^c \in h^c$ is selected uniformly from the set of candidate locations h^c. The candidate ξ^c is chosen to be added to the current set of knots $(t_{k+1}^{(\kappa)}, \ldots, t_{m^{(\kappa)}}^{(\kappa)})'$, of length $\nu^{(\kappa)}$. Denote by ℓ the interval of the current knot set containing ξ^c, so that $t_\ell^{(\kappa)} < \xi^c < t_{\ell+1}^{(\kappa)}$. The new candidate knot set is then given by $(t_{k+1}^{(\kappa)}, \ldots, t_\ell^{(\kappa)}, \xi^c, t_{\ell+1}^{(\kappa)}, \ldots, t_{m^{(\kappa)}}^{(\kappa)})'$ of length $\nu^{(\kappa+1)} = \nu^{(\kappa)} + 1$. In the same spirit, the set of spline coefficients $\beta^{(\kappa)}$ of length $m^{(\kappa)}$ must be updated to a candidate set $\tilde{\beta}^{(\kappa+1)}$ of length $\tilde{m}^{(\kappa+1)} = m^{(\kappa)} + 1$. There are simple rules for nondestructively inserting a new knot into a B-spline function [7], but using these directly may violate the shape constraint. In addition, it would violate the reversibility and dimension-matching constraint between the birth and death moves. Since the birth move begins in a model of dimension $\nu^{(\kappa)}$ and its reverse begins at dimension $\nu^{(\kappa)}+1$, we need to generate an additional random number for the birth move. Intuitively, since removing a knot is a destructive procedure and may change the shape of the curve, we must during the birth move be able to generate the set of curves that would reduce to the original curve upon removal of the new knot. To do this, we compute the candidate spline coefficients $\tilde{\beta}^{(\kappa+1)}$ for inserting a knot ξ^c as follows:

$$\tilde{\beta}_j^{(\kappa+1)} = \begin{cases} \beta_j^{(\kappa)} & \text{if} \quad j \leq \ell+1, \\ \beta_{j-1}^{(\kappa)} & \text{if} \quad j > \ell+k, \\ \omega_j \beta_j^{(\kappa)} + (1-\omega_j)\beta_{j-1}^{(\kappa)} & \text{if} \quad \ell+1 < j < \ell+k, \\ u\beta_j^{(\kappa)} + (1-u)\beta_{j-1}^{(\kappa)} & \text{if} \quad j = \ell+k, \end{cases} \tag{4.2}$$

where $\omega_j = (\xi^c - t_{j-k}^{(\kappa)})/(t_{j-1}^{(\kappa)} - t_{j-k}^{(\kappa)})$ and $u \sim \mathcal{U}(0,1)$. These rules correspond to the deterministic rules in [7], except that the coefficient $\tilde{\beta}_{\ell+k}^{(\kappa+1)}$ is perturbed by a random amount. Of course, whatever random amount u is used, the resulting coefficient $\tilde{\beta}_{\ell+k}^{(\kappa+1)}$ must satisfy the shape constraint if $\tilde{\beta}_{\ell+k}^{(\kappa+1)} \in \tilde{\beta}^{1^{(\kappa+1)}}$. Let $\tilde{\sigma}^{2^{(\kappa+1)}} \sim IG(c,d)$. The likelihood ratio R_L is given by the ratios of the likelihoods for the spline parameters. The prior ratio R_π for the birth move is given by

$$R_\pi = \frac{\pi_0(\tilde{\beta}^{0^{(\kappa+1)}}|\tilde{\sigma}^{2^{(\kappa+1)}}, \tilde{t}^{(\kappa+1)}, \tilde{\beta}^{1^{(\kappa+1)}})}{\pi_0(\beta^{0^{(\kappa)}}|\sigma^{2^{(\kappa)}}, t^{(\kappa)}, \beta^{1^{(\kappa)}})} \times \frac{\pi_1(\tilde{\beta}^{1^{(\kappa+1)}}|\tilde{\sigma}^{2^{(\kappa+1)}}, \tilde{t}^{(\kappa+1)})}{\pi_1(\beta^{1^{(\kappa)}}|\sigma^{2^{(\kappa)}}, t^{(\kappa)})} \times \frac{p(\tilde{\sigma}^{2^{(\kappa+1)}})}{p(\sigma^{2^{(\kappa)}})} \frac{p(\nu^{(\kappa)}+1)}{p(\nu^{(\kappa)})} \frac{\nu^{(\kappa)}+1}{q-\nu^{(\kappa)}}, \tag{4.3}$$

and the transition ratio R_Q is given by

$$R_Q = \frac{d_{\nu^{(\kappa)}+1}/(\nu^{(\kappa)}+1)p(\tilde{\sigma}^{2^{(\kappa+1)}})}{b_{\nu^{(\kappa)}}/(q-\nu^{(\kappa)})p(\sigma^{2^{(\kappa)}})}. \tag{4.4}$$

Note that using (4.1), this implies that

$$R_\pi \times R_Q = \pi_0(\tilde{\beta}^{0^{(\kappa+1)}}|\cdot,\cdot,\cdot)\pi_1(\tilde{\beta}^{1^{(\kappa+1)}}|\cdot,\cdot)/\pi_0(\beta^{0^{(\kappa)}}|\cdot,\cdot,\cdot)\pi_1(\beta^{1^{(\kappa)}}|\cdot,\cdot). \tag{4.5}$$

Lastly, the Jacobian $|J|$ for the transformation from $(\beta^{(\kappa)}, u)$ to $(\tilde{\beta}^{(\kappa+1)})$ is

$$|J| = \Big|(\beta^{(\kappa)}_{\ell+k} - \beta^{(\kappa)}_{\ell+k-1}) \prod_{j=\ell+2}^{\ell+k-1} \omega_j\Big|. \tag{4.6}$$

Thus, the candidate spline parameters are accepted with probability

$$\rho_b = \min\{1, R_L \times R_\pi \times R_Q \times |J|\}. \tag{4.7}$$

4.2. Deletion of a knot and a coefficient (death step)

In the death step, a single knot $t^{(\kappa)}_\ell$ is chosen uniformly from the set of knots $h^{(\kappa)}$ to be removed. The candidate knot set for the next iteration is then

$$\tilde{h}^{(\kappa+1)} = (t^{(\kappa)}_{k+1}, \ldots, t^{(\kappa)}_{\ell-1}, t^{(\kappa)}_{\ell+1}, \ldots, t^{(\kappa)}_{m^{(\kappa)}}). \tag{4.8}$$

The spline parameters are correspondingly adjusted by the inverse of the transformation in (4.2), that is, by deleting the coefficient $\beta^{(\kappa)}_{\ell+k-1}$ and adjusting the remaining coefficients as

$$\tilde{\beta}^{(\kappa+1)}_j = \begin{cases} \beta^{(\kappa)}_j & \text{if} \quad j < \ell+1, \\ \beta^{(\kappa)}_{j+1} & \text{if} \quad j \geq \ell+k-1, \\ \frac{1}{\omega_j}\beta^{(\kappa)}_j + \frac{(1-\omega_j)}{\omega_j}\beta^{(\kappa)}_{j-1} & \text{if} \quad \ell+1 \leq j < \ell+k-1. \end{cases} \tag{4.9}$$

Because the birth and death moves are symmetrically defined, the likelihood ratio, prior ratio, transition ratio and Jacobian determinant are the inverses of those in ρ_b.

4.3. Knot and coefficient positions change (move step)

In the move step, a single knot position $t^{(\kappa)}_\ell$ to be moved is chosen uniformly from the set of interior knots, and changed to a random new candidate position located between its neighboring knots. That is, the candidate knot position $\tilde{t}^{(\kappa+1)}_\ell$ is selected uniformly from the set of candidate locations $\xi^c \in h^c$ such that $t^{(\kappa)}_{\ell-1} < \xi^c < t^{(\kappa)}_{\ell+1}$. One method to draw a candidate vector of coefficients is to perturb a coefficient, $\beta^{(\kappa)}_j$ with $j \sim \mathcal{U}\{1, \ldots, m^{(\kappa)}\}$, uniformly between its neighboring coefficients. If $\beta^{(\kappa)}_j \in \beta^{1^{(\kappa)}}$, the resulting coefficient $\tilde{\beta}_j^{(\kappa)}$ must satisfy the shape constraint. Since no dimension change is required, the new spline parameters are accepted with probability

$$\rho_m = R_L \times \frac{\pi_0(\tilde{\beta}^{0^{(\kappa+1)}}|\tilde{\sigma}^{2^{(\kappa+1)}}, \tilde{t}^{(\kappa+1)}, \tilde{\beta}^{1^{(\kappa+1)}})}{\pi_0(\beta^{0^{(\kappa)}}|\sigma^{2^{(\kappa)}}, t^{(\kappa)}, \beta^{1^{(\kappa)}})} \times \frac{\pi_1(\tilde{\beta}^{1^{(\kappa+1)}}|\tilde{\sigma}^{2^{(\kappa+1)}}, \tilde{t}^{(\kappa+1)})}{\pi_1(\beta^{1^{(\kappa)}}|\sigma^{2^{(\kappa)}}, t^{(\kappa)})}. \tag{4.10}$$

Since the prior on the knot positions is discrete uniform over the set of candidate knots, the prior probabilities for knots $h^{(\kappa)}$ and the candidate $\tilde{h}^{(\kappa+1)}$ are identical.

5. Asymptotic consistency of the posterior

In this section, we consider the Bayesian consistency for the shape constrained regression function from the B-spline prior. Consistency of the posterior distribution is an important large-sample property for the validation of a Bayesian method. Because posterior consistency may fail in an infinite-dimensional model, it seem that checking the consistency of any non-parametric Bayesian procedure is required. We study the behaviour of the posterior distribution for the Kullback-Leibler and the Hellinger divergences when $x_1, \ldots, x_n$ is an i.i.d. sequence having density h_x which has a support $[a_0, b_0]$. In the unconstrained Bayesian setting, [12] presented general results on the rates of convergence of the posterior measure and [13] generalized the results to case when the observations are not i.i.d. From the approach based on the two preceding seminal papers, the non-parametric B-splines estimator attains the minimax rate of convergence $n^{-\frac{\alpha}{2\alpha+1}}$ under the assumption that the true regression function belong to the Hölder space $C^{\alpha}[0,1]$ and $\alpha > 0$ could be fractional. The rate of convergence in our constrained case is beyond the scope of this paper and we will focus on the study of the consistency.

5.1. Weak consistency

Consider that σ^2 is assigned a prior distribution $\pi(\sigma^2)$ on $[0, \tau_\sigma]$, for some large constant τ_σ, and f has prior distribution $\mathit{\Pi}(f)$ a probability measure on the space of shape constrained continuous functions. For simplicity of notation and with no loss of generality, we assume that $[a_0, b_0] = [a, b]$. Let P_0 denote the true distribution of (x, y), let $\theta = (f, \sigma^2) \in (\mathcal{S}, [0, \tau_\sigma])$ and let $l_n(\theta)$ denote the likelihood function given by

$$l_n(\theta) = \prod_{i=1}^{n} h_y^{\theta}(y_i) h_x(x_i),$$

where $h_y^\theta(y)$ is a normal density with mean $f(x)$ and unknown variance σ^2. The posterior distribution for a set B is given by

$$\Pi_n(B) = \Pi(B|(x_1, y_1), \ldots, (x_n, y_n)) = \frac{\int_B l_n(\theta) d\Pi(f) \pi(\sigma^2) d\sigma^2}{\int l_n(\theta) d\Pi(f) \pi(\sigma^2) d\sigma^2}$$

and, for all $\epsilon > 0$, we shall consider

$$A_\epsilon = \{(f, \sigma^2) \in (\mathcal{S}, [0, \tau_\sigma]) : K(h_x h_y^{\theta_0}, h_x h_y^\theta) < \delta + \epsilon/2\}$$

where $\theta_0 = (f_0, \sigma_0^2)$ is the true value of θ, $K(\cdot, \cdot)$ denote the Kullback-Leibler divergence and

$$\delta = \inf_{\theta \in (\mathcal{S}, [0, \tau_\sigma])} \{K(h_x h_y^{\theta_0}, h_x h_y^\theta)\}. \tag{5.1}$$

Clearly, our aim is to show that, for all $\epsilon > 0$, $\Pi_n(A_{2\epsilon}) \to 1$ almost surely P_0^∞. For this, consider that the regression function f is bounded such that

$$-C \leq \min_{j=1,\ldots,m} \beta_j \leq \max_{j=1,\ldots,m} \beta_j \leq C \tag{5.2}$$

for some known constant $C \in (0, \infty)$ (depending on the data). To refer explicitly to the space $\mathcal{S}$ of a specific shape constrained continuous bounded functions on $[a_0, b_0]$, we write $\mathcal{S}^C$. In the same spirit, we write S^C the set of β such that the shape constraints and (5.2) are fulfilled. (The notation is slightly different from that in previous sections.) Let $D_\nu = S^C \times [a_0, b_0]^\nu$ and we assumed that $\pi(\nu) > 0$ for all integers $\nu = 1, 2, \ldots, q$; this assumption is necessary for proving consistency. In the following, with some abuse of notation, we will use Π as the prior distribution on (β, h, σ^2) as well as for the induced prior distribution on regression function f.

Lemma 1. *Suppose that the constrained regression function f is bounded by a known constant so that (5.2) holds. Then, for all $\epsilon > 0$, $\Pi(A_\epsilon) > 0$.*

The Lemma 1 shows that the prior puts positive mass on all Kullback-Leibler neighbourhoods of the true distribution under the condition that the range of the underlying curve f is contained in some known interval $[-C, C] \subset \mathbb{R}$ which will trivially be satisfied in applications.

Theorem 1. *Let Π be the spline prior described in Section 3. Consider a family of models $\mathcal{P} = \{P_\theta : \theta \in (\mathcal{S}^C, [0, \tau_\sigma])\}$ with densities $h_x h_y^\theta$ with* *respect to some common dominating measure. In particular, assume that the regression function is bounded by a known constant so that (5.2) holds. Take an independent pairs $(x_1, y_1), \ldots, (x_n, y_n)$ and given x_i, suppose y_i is normally distributed with conditional mean $f(x_i)$ and conditional variance σ^2. Then*

$$\Pi(A_{2\epsilon}|(x_1, y_1), \ldots, (x_n, y_n)) \to 1 \qquad \textit{almost surely } P_0^\infty,$$

for all $\epsilon > 0$.

Thus, the preceding results provide regularity conditions ensuring that the posterior concentrates almost surely on the subset θ_0 of the space $(\mathcal{S}^C, [0, \tau_\sigma])$ on which the Kullback-Leibler divergence of the true distribution $h_x h_y^{\theta_0}$ against the model distribution $h_x h_y^{\theta}$ is minimal. If in addition one can prove that in the considered model $K(h_x h_y^{\theta_0}, h_x h_y^{\theta}) \gtrsim \|h_x h_y^{\theta_0} - h_x h_y^{\theta}\|_2$, then Theorem 1 delivers a consistency with respect to the L_2-distance as well. We used above $\gtrsim$ to denote greater or equal up to a constant. We should note here that it not requires a fair piece of effort to implement this idea for our white noise model and the above relation between norms straightforward. For sake of completeness, we summarize this in the following corollary. For reason of simplicity, we assume in the following result that the variance σ^2 is a constant.

Corollary 1. *In the previous Theorem 1 the Kullback-Leibler distance may be replaced by the L_2 distance and the statement remains valid, which means we have*

$$K(h_x h_y^{\theta_0}, h_x h_y^{\theta}) \gtrsim \int_{a_0}^{b_0} \big(f_0(x) - f(x)\big)^2 h_x(x) dx \tag{5.3}$$

and then

$$\Pi(B_{2\epsilon}|(x_1, y_1), \ldots, (x_n, y_n)) \to 1 \qquad \textit{almost surely } P_0^\infty,$$

for all $\epsilon > 0$ and

$$B_\epsilon = \Big\{ f \in \mathcal{S} : c \int_{a_0}^{b_0} \big(f_0(x) - f(x)\big)^2 h_x(x) dx \leq \delta + \epsilon/2 \Big\}$$

where the constant $c > 0$ dependent on C, σ^2 and f_0.

5.2. Strong (Hellinger) consistency

Having derived weak consistency of the shape constrained regression estimator, we now turn to the strong consistency problem. [29] had established that a prior which puts positive mass on all Kullback-Leibler neighbourhoods of the true distribution is weakly consistent. However, [9] had demonstrated that priors which puts positive mass on all Kullback-Leibler neighbourhoods of the true distributions are not necessarily weakly consistent. In the last two decades, Bayesian nonparametric methods have switched to studying and characterizing sufficient conditions for strong consistency. Now, to study the Hellinger consistency in our shape constrained procedure, we need a little more notation. Let, for all $\eta > 0$, consider

$$K_\eta = \{(f, \sigma^2) \in (\mathcal{S}^C, [0, \tau_\sigma]) : H(h_x h_y^{\theta_0}, h_x h_y^{\theta}) > \eta\}$$

where $H(\cdot, \cdot)$ is the Hellinger divergence, given by

$$H(h_x h_y^{\theta_0}, h_x h_y^{\theta}) = \sqrt{2}\Big(1 - \int \sqrt{h_x h_y^{\theta_0} h_x h_y^{\theta}} dx dy\Big)^{1/2} = \sqrt{2}\Big(1 - \int h_x \sqrt{h_y^{\theta_0} h_y^{\theta}} dx dy\Big)^{1/2}.$$

The idea is to follow [39] and [35] by using a technique that combine the consistency of the maximum likelihood estimate (MLE) and a Bayesian component. This technique is valid whenever the MLE exists. To tackle the problem of MLE consistency, we employ optimization techniques. Exactly, we use projection on closed convex set and the uniform Lipschitz property in the ℓ_∞-norm. This property yields a uniform sup-norm bound on variations of spline coefficients regardless of the number of interior knots ν. Moreover, this property leads to boundary consistency of the shape constrained B-spline estimator, an important feature that many other estimators in the literature (e.g., the Brunk's monotone estimator studied in [28] and the convex estimator studied in [16]) do not have.

Lemma 2. *For the model (3.1) and the class $\mathcal{S}^C$ of shape constrained regression functions, the maximum likelihood $\hat{\theta} = (\hat{f}, \hat{\sigma}^2)$ exists and is unique. Furthermore, under the conditions of Theorem 2 [37] and Theorem 3.1 [38], we have*

$$\forall \epsilon > 0, \qquad \frac{l_n(\hat{\theta})}{l_n(\theta_0)} \leq \exp(2n\epsilon), \qquad \textit{for all large } n. \tag{5.4}$$

Proof. Fix an arbitrary $\theta_0 \in \mathcal{S}^C \times [0, \tau_\sigma]$. We show that there exists a $\hat{\theta} \in \mathcal{S}^C \times [0, \tau_\sigma]$ for which $l_n(\hat{\theta}) \geq l_n(\theta_0)$. Maximizing the likelihood ($\max_{\theta \in \mathcal{S}^C \times [0,\tau_\sigma]} l_n(\theta)$) amounts to choosing $\hat{f} \in \mathcal{S}^C$ such that

$$\hat{f} = \arg \min_{f \in \mathcal{S}^C} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 \tag{5.5}$$

and considering the variance $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \{y_i - \hat{f}(x_i)\}^2$. Hence any function in $\mathcal{S}^C$ which passes through each of the observations $(x_1, y_1), \dots, (x_n, y_n)$ achieves the minimum. Indeed, let f_0 be a given monotone and continuous function. Then one can construct a piecewise polynomial function which has the same values at the observation points by taking $\mu_j = f_0(j\frac{b-a}{m} + a)$, for $j = 0, \dots, m$, and $\hat{f}(x) = \sum_{j=0}^{m} \mu_j B_j(x)$. It follows from Bernstein-Weierstrass approximation Theorem that $\hat{f}$ converges to f_0 uniformly. Now, we will show existence in case of concave (convex) constraint. Let f_0 be a given concave and continuous function. Taking $\bar{\mu}_0 = f_0'(a) + 1/m^3$, $\bar{\mu}_{m-1} = f_0'(b) - 1/m^3$, $\bar{\mu}_j = f_0'(j\frac{b-a}{m-1} + a)$, for $j = 0, \dots, m-2$ and $\hat{f}'(x) = \sum_{j=0}^{m-1} \bar{\mu}_j B_{j,k-1}(x)$. Using Bernstein-Weierstrass Theorem, we know $\hat{f}'$ converges to f_0' uniformly. Additionally, let $\mu_0 = f_0(a)m/(b-a)$, $\mu_j = \mu_0 + \bar{\mu}_0 + \dots + \bar{\mu}_{j-1}$ and $\hat{f}(x) = \frac{b-a}{m} \sum_{j=0}^{m} \mu_j B_{j,k}(x)$. Thus $\hat{f}(0) = f_0(0)$ and $\hat{f}$ converges uniformly to f_0. To conclude the existence argument, it is clear that $\mu_{j+2} - 2\mu_{j+1} + \mu_j = \bar{\mu}_{j+1} - \bar{\mu}_j \leq 0$ as the sequence $\bar{\mu}_j$ is decreasing. Immediately, we know $\mu_1 - \mu_0 > 0$, $\mu_m - \mu_{m-1} < 0$ and $\mu_{j+2} + \mu_j \leq 2\mu_{j+1}$. This same argument can be used to show existence under unimodality constraint and for this reason we omit it.

Now for the uniqueness, we use the projection Theorem on nonempty closed convex set. For any μ and α, we may consider the scalar product $\langle \mu, \alpha \rangle_{\mathbf{M}} = \mu' \mathbf{M} \alpha$ where $\mathbf{M} = B'B$ which is a positive definite matrix when $m < n$. Let consider $P\alpha = \arg\min_{\mu \in S^C} \|\mu - \alpha\|_{\mathbf{M}}$ where P is a projection operator and $\|\cdot\|$ to denote the Euclidean norm. For a monotone increasing spline $s = \sum_{j=1}^{m} \mu_j B_j$, we have the constraint

$$S^C = \{s : |\mu| \leq C, \mu_1 \leq \mu_2 \leq \dots \leq \mu_m\}.$$

Similarly, for a concavity constraint on $[a_0, b_0]$, we have

$$S^C = \cap_{j=3}^{m-1} \{s : |\mu| \leq C, \mu_1 < \mu_2, \mu_{m-1} > \mu_m, \mu_j - 2\mu_{j-1} + \mu_{j-2} \leq 0\},$$

and for a unimodality constraint, we have

$$S^C = \cup_{l=2}^{m-1}\{s : |\mu| \leq C, \mu_1 \leq \mu_2 \leq \cdots \leq \mu_l \geq \mu_{l+1} \geq \cdots \geq \mu_m\}.$$

Since S^C is a closed convex set, their exists a unique projection of α on the constrained set S^C.

We have already checked existence and uniqueness of the maximum likelihood under the shape constraints. We are now in position to prove (5.4). To establish that inequality (5.4) holds under concave (convex) constraint, we base our argument on [38]. Let $\alpha_n = \sum_{i=1}^n B_j^2(x_i)$ for $j = k, \ldots, \nu$. Without loss of generality, we assume that the knots are equally spaced where α_n is independent of j. Thanks to the structure of the design matrix B together with its column linear independence, we can easily show that $\varOmega = B'B/\alpha_n \in \mathbb{R}^m \times \mathbb{R}^m$ is a positive definite and $(2k-1)$-banded matrix. Now, let define the weighted data vector $\bar{y} = B'y/\alpha_n \in \mathbb{R}^m$. Formulating (5.5) via the above matrix notation leads to

$$\hat{\beta} \equiv \hat{\beta}(\bar{y}) = \arg\min_{\beta \in S^C} \frac{1}{2}\beta'\varOmega\beta - \beta'\bar{y}. \tag{5.6}$$

By using this formulation (5.6), [38] derived a closed form of $\hat{\beta}$ despite the combinatorial nature of the problem and showed that given an order k, there exists a positive constant $c_{\infty,k}$ (dependent on k only) such that

$$\|\hat{\beta}(v) - \hat{\beta}(w)\|_\infty \leq c_{\infty,k}\|v - w\|_\infty, \qquad \text{for any } \nu \text{ and all } v, w \in \mathbb{R}^m. \tag{5.7}$$

Under the monotone constraint, a similar result to (5.7) has been established by [37] for the order $k = 2$. Another similar result has been established by [30] for all order k in the case of the monotone smoothing spline regression. In light of the fact that $\hat{\beta}(\bar{y})$ satisfies the uniform Lipschitz property (5.7), we obtain

$$\|\hat{\beta}(\bar{y}) - \hat{\beta}(\mathbb{E}[\bar{y}])\|_\infty \lesssim \|\bar{y} - \mathbb{E}[\bar{y}]\|_\infty = O_{\mathbb{P}}\Big(\sqrt{\frac{\nu\log(\nu)}{n}}\Big),$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator and $a = O_{\mathbb{P}}(b)$ means that a/b is bounded in probability. Let $\bar{f}(x) = \sum_{j=1}^m \bar{\beta}B_j(x)$ where $\bar{\beta} = \hat{\beta}(\mathbb{E}[\bar{y}])$. By a Taylor series expansion, we can easily check that $\|\bar{f} - f_0\|_\infty = O(\nu^{-1})$. This equality and the triangle inequality yield

$$
\begin{aligned}
\|\hat{f}-f_0\|_\infty = \sup_{x\in[a,b]}|\hat{f}(x)-f_0(x)| &\le \|\hat{f}-\bar{f}\|_\infty + \|\bar{f}-f_0\|_\infty \\
&\le \|\hat{\beta}(\bar{y})-\hat{\beta}(\mathbb{E}[\bar{y}])\|_\infty + O(\nu^{-1}) \\
&\lesssim \|\bar{y}-\mathbb{E}[\bar{y}]\|_\infty + O(\nu^{-1}) \\
&= O_{\mathbb{P}}\Big(\sqrt{\frac{\nu\log(\nu)}{n}}\Big) + O(\nu^{-1}),
\end{aligned}
$$

which establishes the uniform consistency of $\hat{f}$; i.e., assume f_0 is bounded on $[a,b]$, if $\nu\to\infty$ and $n^{-1}\nu\log(\nu)\to 0$ as $n\to\infty$, then

$$
\forall\epsilon>0, \qquad \mathbb{P}\Big(\sup_{x\in[a,b]}|\hat{f}(x)-f_0(x)|\ge\epsilon\Big)\longrightarrow 0. \tag{5.8}
$$

Specifically, we deduce

$$
\begin{aligned}
\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i)-f_0(x_i))^2\ge\epsilon^2\Big) &\le \sum_{i=1}^{n}\mathbb{P}\Big(|\hat{f}(x_i)-f_0(x_i)|\ge\epsilon\Big) \\
&\le n\mathbb{P}\Big(\sup_{x\in[a,b]}|\hat{f}(x)-f_0(x)|\ge\epsilon\Big),
\end{aligned}
$$

and clearly we obtain

$$
\frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i)-f_0(x_i))^2\longrightarrow 0 \qquad \text{almost surely } P_0^\infty. \tag{5.9}
$$

Furthermore, by maximization of the likelihood ($l_n(\hat{\theta})\ge l_n(\theta_0)$) and the triangle inequality we can write

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{f}(x_i))^2 &\le \frac{1}{n}\sum_{i=1}^{n}(y_i-f_0(x_i))^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{f}(x_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i)-f_0(x_i))^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n}(y_i-f_0(x_i))^2 + \frac{1}{n}\sum_{i=1}^{n}(\hat{f}(x_i)-f_0(x_i))^2 \longrightarrow \sigma_0^2+0 \quad \text{a.s. } P_0^\infty.
\end{aligned}
$$

Hence, we conclude the proof by writing

$$\frac{1}{n}\log\Big(\frac{l_n(\hat{\theta})}{l_n(\theta_0)}\Big)$$

Theorem 2. *Let Π be the spline prior described in Section 3. Assume that the regression function is bounded and under the conditions of Lemma 2, for all sets K_η with $\eta > 0$, we have*

$$\Pi(K_\eta|(x_1,y_1),\ldots,(x_n,y_n)) \to 0 \qquad \textit{almost surely } P_0^\infty.$$

Proof. We can write the posterior distribution for a set K_η as

$$\Pi_n(K_\eta) = \frac{\int_{K_\eta}\frac{l_n(\theta)}{l_n(\theta_0)}d\Pi(f)\pi(\sigma^2)d\sigma^2}{\int\frac{l_n(\theta)}{l_n(\theta_0)}d\Pi(f)\pi(\sigma^2)d\sigma^2} \leq \frac{\Big(\frac{l_n(\hat{\theta})}{l_n(\theta_0)}\Big)^{1/2}\int_{K_\eta}\Big(\frac{l_n(\theta)}{l_n(\theta_0)}\Big)^{1/2}d\Pi(f)\pi(\sigma^2)d\sigma^2}{\int\frac{l_n(\theta)}{l_n(\theta_0)}d\Pi(f)\pi(\sigma^2)d\sigma^2}.$$

Taking expectation of the integral term in the numerator, we obtain

$$\begin{aligned}\mathbb{E}\Big(\int_{K_\eta}\Big(\frac{l_n(\theta)}{l_n(\theta_0)}\Big)^{1/2}d\Pi(f)\pi(\sigma^2)d\sigma^2\Big) &= \int_{K_\eta}\prod_{i=1}^{n}\Big\{\int\Big(\frac{h_y^{\theta}(y_i)}{h_y^{\theta_0}(y_i)}\Big)^{1/2}h_x(x_i)h_y^{\theta_0}(y_i) \\ &\qquad dx_i dy_i\Big\}d\Pi(f)\pi(\sigma^2)d\sigma^2 \\ &= \int_{K_\eta}\big(1-\frac{1}{2}H^2(h_xh_y^{\theta_0},h_xh_y^{\theta})\big)^n d\Pi(f)\pi(\sigma^2)d\sigma^2 \\ &\leq (1-\frac{\eta^2}{2})^n.\end{aligned}$$

Consequently, we can write

$$\begin{aligned}\mathbb{P}\Big(\int_{K_\eta}\Big(\frac{l_n(\theta)}{l_n(\theta_0)}\Big)^{1/2}d\Pi(f)\pi(\sigma^2)d\sigma^2 > \exp(-n\tilde{\eta})\Big) &< \exp(n\tilde{\eta})\mathbb{E}\Big(\int_{K_\eta}\Big(\frac{l_n(\theta)}{l_n(\theta_0)}\Big)^{1/2} \\ &\qquad \Pi(f)\pi(\sigma^2)d\sigma^2\Big) \\ &< \exp(n\tilde{\eta})(1-\frac{\eta^2}{2})^n \\ &< \exp\Big(n\tilde{\eta}-n\log(1-\frac{\eta^2}{2})^{-1}\Big).\end{aligned}$$

Hence, choosing $\tilde{\eta} < \log(1 - \frac{\eta^2}{2})^{-1}$, the Borel-Cantelli theorem gives that

$$\int_{K_\eta} \Big(\frac{l_n(\theta)}{l_n(\theta_0)}\Big)^{1/2} d\Pi(f)\pi(\sigma^2)d\sigma^2 < \exp(-n\tilde{\eta}), \qquad \text{a.s. for all large } n. \tag{5.10}$$

Thus, by combining (5.4) and (5.10) the numerator of the posterior can now be written, a.s. for all large n and for any $\tilde{\eta} < \log(1 - \frac{\eta^2}{2})^{-1}$, as

$$\forall \epsilon > 0, \qquad \int_{K_\eta} \frac{l_n(\theta)}{l_n(\theta_0)} d\Pi(f)\pi(\sigma^2)d\sigma^2 < \exp(-n(\tilde{\eta} - \epsilon)). \tag{5.11}$$

Now we turn to the denominator of the posterior. We proved that under the conditions of Lemma 1, we have $\Pi(A_\epsilon) > 0$ for all $\epsilon > 0$ which means that the prior puts positive mass on Kullback-Leibler neighborhood of the true distribution and consequently from [29] we have for any $\bar{\eta} > 0$

$$\int \frac{l_n(\theta)}{l_n(\theta_0)} d\Pi(f)\pi(\sigma^2)d\sigma^2 > \exp(-n\bar{\eta}), \qquad \text{a.s. for all large } n. \tag{5.12}$$

Finally, putting together (5.11), with obviously choose $\epsilon < \tilde{\eta}$, and (5.12) we obtain Bayesian consistency under our shape constrained prior and we conclude that $\Pi_n(K_\eta) \overset{\text{a.s.}}{\to} 0$. □

6. Numerical studies

We now explore the numerical performance of our methodology to provide some validation, strengthen the motivation for the use of localized shape constraints and better qualify the contribution. This section reports the results of a numerical experiment that was used to compare the small sample properties across examples of functions as well as applications based on real world data sets. We first apply the method to artificial data (Section 6.1) to illustrate that our approach performs well in comparison to the estimation with fixed-knot strategy and the unconstrained (naive) estimation. Secondly, two applications to food industry and Global Warming data are provided.

6.1. A simulation study

By applying the methodology described in this paper, we compare the performance of the free-knot localized shape estimation with the fixed-knot

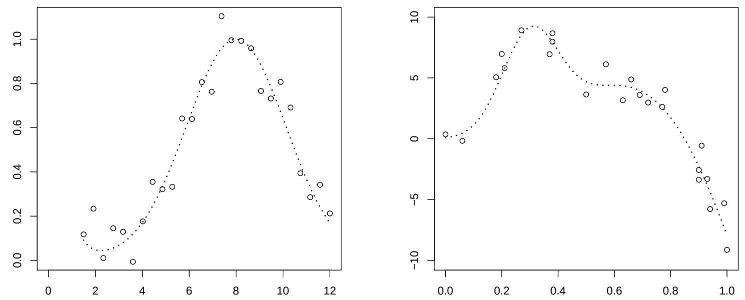

Figure 2: Figures shown simulated data (circle) from model (3.1) and the true regression functions (dotted line). The figure of f_1 on the left and f_2 on the right.

localized shape estimation [2] and the unconstrained estimation. The preceding three estimations are compared using the posterior mean $\widehat{f}$ and the posterior mode (MAP) $\tilde{f}$ of the regression function. We note that, although any posterior simulation of f necessarily fulfills the shape constraints, the posterior expectation $\widehat{f}$ may not satisfy the constraints (for instance, the mean of unimodal functions is not necessarily an unimodal function). In some cases, it might be more appropriate to use the posterior mode instead of the posterior expectation as an estimate. For sake of completeness, we reports the performance of $\widehat{f}$ and $\tilde{f}$ together across the L_1-norm, sup-norm, mean square error (MSE) of $|\widehat{f}-f|$ and $|\tilde{f}-f|$, as a function on $[a_0, b_0]$. Consider the following two regression examples: given a (possibly small) number of observation points, the aim is to find an estimate for f_1 and f_2. We generate data ($n = 26$) according to model (3.1) with a true regression function defined by $f_1(x) = \frac{1}{2(x^4+1)} + \exp\{(-(x-8)/3)^2\}$ on $[a_0, b_0] = [1.5, 12]$ with $\sigma = 0.1$. We assume that it is known that f_1, which we plot in Figure 2 (on the left) together with the observations from the model, is convex on $[1.5, 4.5]$, unimodal on $[4, 12]$ and continuously differentiable everywhere. In addition, we generate data ($n = 24$) according to model (3.1) with a true regression function defined by $f_2(x) = 15x^2 \sin(3.7x) + 2\psi_{0.3,0.1}(x)$ on $[a_0, b_0] = [0, 1]$ with $\sigma = 1$, where $\psi_{m,\sigma}$ denotes the density of the $N(m, \sigma^2)$ distribution. We assume that it is known that f_2, which we plot in Figure 2 (on the right) together with the observations from the model, is unimodal on $[0, 1]$ (first increasing and then decreasing), concave on $[0.55, 1]$ and twice differentiable.

Now we describe how we use the two free-knot shape constrained priors for the two regression functions f_1 and f_2 to start the RJMCMC samplers. In both hierarchical priors we endow ν with a Poisson prior with mean $\lambda = \{0.5, 1, 2\}$ for f_1 and with mean $\lambda = \{0.5, 2, 5\}$ for f_2. In the knots prior, given $\nu^{(0)} = 9$, the $\nu^{(0)}$ inner knots are taken to be equally spaced on (a_0, b_0). The construction of $B_1, \ldots, B_m$ involves an arbitrary extra knots $t_1 \leq \cdots \leq t_k = a_0$ and $b_0 = t_{m+1} \geq \cdots \geq t_{m+k}$. Usually one takes $t_1 = \cdots = t_k = a_0$ and $b_0 = t_{m+1} = \cdots = t_{m+k}$, and we adopt this choice here as well. We note here that in the predetermined set of candidate inner knots h we consider only simple knots (see the definition of knot multiplicity in Section 3). On spline coefficients we put a truncated Gaussian distribution if it are in the local interval of the constraint and a simple Gaussian distribution if not as explained in Section 3. We let both RJMCMC samplers run for the same number of iterations, starting from the state $\beta^{(0)} = (0.1, 0.047, 0.04, 0.51, 0.65, 0.7, 0.73, 0.88, 1.042, 1.008, 0.57, 0.0575)$ for f_1 which corresponds to a convex function on $[1.5, 4.5]$, unimodal on $[4, 12]$ and from $\beta^{(0)} = (3.29, 6.27, 7.83, 8.25, 8.66, 5.61, 4.95, 4.61, 4.48, -1.90, -6.72, -7.16)$ the starting state for f_2 which corresponds to a unimodal function on $[0, 1]$, concave on $[0.55, 1]$. We run the RJMCMC algorithm with $c = 60$, $d = 0.5$ and $\tau = 1$ for 10^5 updates, after a burn in period of 10^4 updates. we use the remaining 9×10^4 realizations of the Markov chain to obtain the posterior summaries. The above experiment is replicated 50 times; the averages of the resulting 50 L_1-norms, sup-norms and MSE of $|\widehat{f} - f|$ and $|\tilde{f} - f|$ are reported in Tables 1 and 2.

One remarkable aspect of the study is how our procedure performed to take into account combinations of shape constraints and to localize a shape constraint on a given interval. For different values of λ (see Figures 3 and 4) the posterior mean of ν is sensitive to the choice of the Poisson prior mean and in particular this sensitivity becomes significant unsurprisingly when the range of data $[a_0, b_0]$ is large. We remark again this sensitivity with the error of the estimation (see Tables 1 and 2). Usually we note that the error increases in function of λ for the three cases of L_1-norms, sup-norms and MSE but the difference between the error of the posterior mean and the posterior mode is not very significant. For function f_2 the error of estimation is greater than those of the function f_1. This feature can be explained easily by the complexity of the shape of f_2 with small number of observations. In particular, the value of $n = 24$ has been chosen sufficiently small to highlight the fact that few observations can give large

Table 1: Simulation study for f_1 defined in Section 6.1. All results are based on 50 runs.

		Free-knot constrained		Unconstrained	Fixed-knot constrained		
λ	Criterion	$\lvert\widehat{f}-f\rvert$	$\lvert\tilde{f}-f\rvert$	$\lvert\widehat{f}-f\rvert$	n	$\lvert\widehat{f}-f\rvert$	$\lvert\tilde{f}-f\rvert$
$\lambda = 0.5$	Sup-norm	0.036179	0.041220	0.119243	$n = 26$	0.080620	0.087204
	L_1-norm	0.091103	0.116862	0.489892		0.256143	0.262035
	MSE	0.001697	0.002804	0.038385		0.012174	0.013043
$\lambda = 1$	Sup-norm	0.047979	0.051029	0.104684	$n = 50$	0.063345	0.076908
	L_1-norm	0.122699	0.134996	0.401337		0.180708	0.244450
	MSE	0.002640	0.003916	0.025264		0.007090	0.012556
$\lambda = 2$	Sup-norm	0.057773	0.057827	0.122528	$n = 100$	0.058089	0.067805
	L_1-norm	0.154615	0.171749	0.436041		0.168556	0.194949
	MSE	0.004726	0.005596	0.030887		0.006093	0.008098

errors for the unconstrained and even for a global constrained estimate (not given here) while the combination of the two local constraints estimation remains good on the whole interval (the estimate capture the sudden descent of the function f_2 on interval $[0.3, 0.5]$). The simulation results in Tables 1 and 2 indicate that the free-knot constrained method does considerably better than the free-knot unconstrained method for the two regression functions f_1 and f_2. On the basis of the numerical study evidence, it appears that the free-knot constrained estimator is a good robust choice since it is always competitive with the fixed-knot constrained estimator and does considerably better for functions that change direction sharply, such as the unimodal-concave function f_2. Clearly, the same pattern holds for different sample sizes ($n = 26, 50, 100$). In particular, free-knot approach under combination of several shape restrictions does better when the function being estimated changes direction sharply as the approach detects better the high and low variability regions of the data and facilitates the placement of more knots in the high variability region. In all cases, it is preferable to choose the free-knot constrained estimator that gives estimator of the regression function with better estimation properties and smaller error of estimation.

6.2. Application to acidification curves

The free-knot constrained regression that was developed in Section 3 can be very useful to estimate functions where there is no data at all in some regions. An example of such a situation is a recent study of acidification response curves (Figure 5) that model the pH decreasing due to a change in chemical treatment conditions: N_2 treatment; N_2H_2 treatment; O_2H_2 treatment and O_2 treatment. Independent of the data, it is known that acidification kinetics are monotonically decreasing function from a $\mathrm{pH}_{\max}$ value to

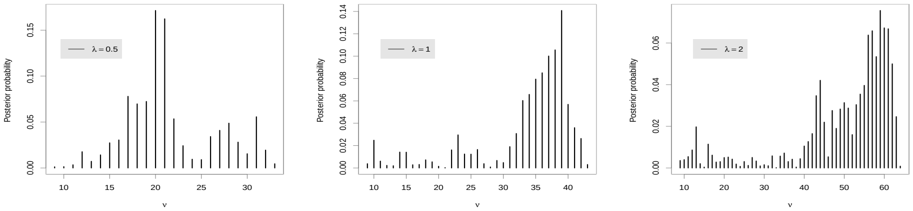

Figure 3: Figures shown the posterior summary of ν for different λ values in the prior of the regression function f_1.

Table 2: Simulation study for f_2 defined in Section 6.1. All results are based on 50 runs.

		Free-knot constrained		Unconstrained	Fixed-knot constrained		
λ	Criterion	$\lvert\widehat{f}-f\rvert$	$\lvert\tilde{f}-f\rvert$	$\lvert\widehat{f}-f\rvert$	n	$\lvert\widehat{f}-f\rvert$	$\lvert\tilde{f}-f\rvert$
$\lambda = 0.5$	Sup-norm	0.395172	0.412234	1.224587	$n = 24$	0.652693	0.690629
	L_1-norm	0.118227	0.123308	0.255491		0.134992	0.144507
	MSE	0.026817	0.028606	0.147344		0.036756	0.044142
$\lambda = 1$	Sup-norm	0.444417	0.410948	1.187223	$n = 50$	0.608905	0.577963
	L_1-norm	0.144337	0.141692	0.280565		0.151304	0.123109
	MSE	0.031038	0.030283	0.168651		0.053650	0.039749
$\lambda = 5$	Sup-norm	0.573937	0.583284	1.444670	$n = 100$	0.520936	0.610902
	L_1-norm	0.154615	0.171749	0.237899		0.129879	0.133811
	MSE	0.047265	0.045896	0.163769		0.033972	0.035829

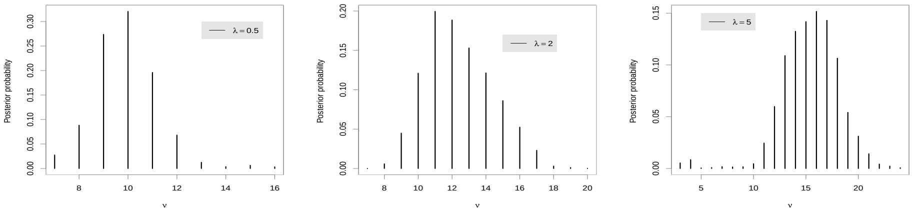

Figure 4: Figures shown the posterior summary of ν for different λ values in the prior of the regression function f_2.

a $\mathrm{pH}_{\min}$, i.e. the $\mathrm{pH}_{\min}$ is like an asymptote, but the beginning of the curve has not the shape of an asymptote, and the decrease has in fact already started from a unknown value to be estimated. As we can see in the Figure 5, the recorded acidification kinetics have an undesirable unimodal shape for x arround 12 and under O_2 treatment for x arround 10. Thus, acidification curves must be corrected to eliminate inconsistent points due to electrical interferences and it is an important issue to estimate what would be the true function if no inconsistent points have occurred. In this section we apply our methodology to reconstruct the acidification curves. For instance, the acidification kinetic under N_2H_2 treatment consists of 1428 measures of pH and extra information from specialists in milk acidification process suggests that any points (x_i, y_i) with $x_i \notin [7, 21]$ must be retained in the reconstruction. This data cleaning was required before any statistical inference because of some errors during the recording process. Therefore, the new data obtained after removing inconsistent points showed a completely absence of data in the region $[7, 21]$. Although specialists indicates that the acidification curve is monotonic they does not suggest a specific functional form. Hence, our nonparametric estimation approach can be appropriate by allowing the data to determine the relationship functional forms under the monotonicity restriction.

The four data sets have been analyzed with the hierarchical constrained model defined in Section 3, with the following setting for previously unspecified hyperparameters: $c = d = 0.01$ and $\tau = 1$. The prior on ν is taken as Poisson on the integers $1, 2, \ldots, \nu_{\max} = 20$. We compute four constrained estimates (Figure 6) of the acidification curves associated with the four data sets. The monotonicity constraint involves all the B-splines ($S = \{\beta \in \mathbb{R}^m, \beta_1 \geq \cdots \geq \beta_m\}$) as the constraint is located on the whole interval $[0, 23.8]$. Computations of $(\hat{\beta}, \hat{\sigma}, \hat{t}, \hat{m})$ are obtained using simulations from the posterior distribution obtained by running the RJMCMC sampler. Several initial values for $(\beta^0, \sigma^0, t^0, m^0)$ are used and give similar results. For each of the four data sets, we report results corresponding to 10^5 sweeps, following a burn-in period of 2×10^4 sweeps. In all the runs, the number of knots never exceeded 20; hence the chosen value of $\nu_{\max}$ was inconsequential.

6.3. Application to the Global Warming data

To confirm the advantages of the combinations of shape constraints to improve function estimation, we shall apply it to the Global Warming data

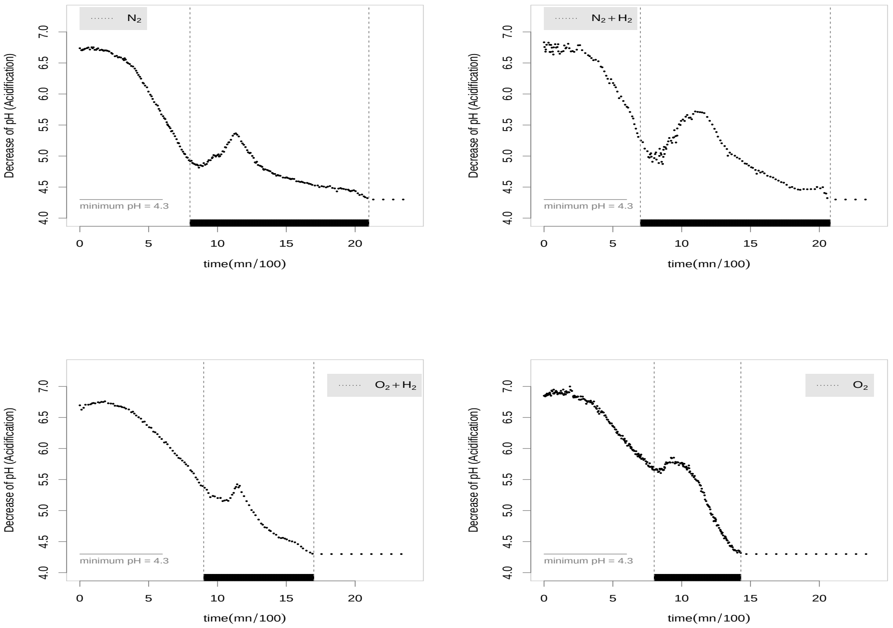

Figure 5: pH versus time (mn) rescaled by a factor of 10^{-2} under four chemical treatment conditions. Inconsistent data $\mathcal{D}_{\mathrm{N_2}} = \{(x_i, y_i), x_i \in [8, 21]\}$, $\mathcal{D}_{\mathrm{N_2H_2}} = \{(x_i, y_i), x_i \in [7, 21]\}$, $\mathcal{D}_{\mathrm{O_2H_2}} = \{(x_i, y_i), x_i \in [9, 17]\}$, $\mathcal{D}_{\mathrm{O_2}} = \{(x_i, y_i), x_i \in [8, 14.3]\}$ are indicated by the vertical lines and the horizontal stripes.

Table 3: Posterior distribution of ν for the four data sets.

Data set	λ	$p(\nu\|x, y)$				Proportion (%) of moves accepted	
						Birth-Death	Move
N_2H_2	4	$p(1) = 0.04044$	$p(2) = 0.08654$	$p(3) = 0.40436$	$p(4) = 0.01573$	$2.1 - 2.3$	6.41
		$p(5) = 0.00744$	$p(6) = 0.01498$	$p(7) = 0.03888$	$p(8) = 0.08159$		
		$p(9) = 0.09410$	$p(10) = 0.05180$	$p(11) = 0.06464$	$p(12) = 0.04002$		
		$p(13) = 0.04049$	$p(14) = 0.01139$	$p(15) = 0.00758$	$p(16) = 0.00002$		
		$\sum_{\nu \geq 17} p(\nu) = 0$					
N_2	3	$p(1) = 0.04781$	$p(2) = 0.16487$	$p(3) = 0.46782$	$p(4) = 0.0.08920$	$2.3 - 2.9$	7.64
		$p(5) = 0.14815$	$p(6) = 0.06110$	$p(7) = 0.00015$	$p(8) = 0.00725$		
		$p(9) = 0.00855$	$p(10) = 0.00395$	$p(11) = 0.00095$	$p(12) = 0.00020$		
		$\sum_{\nu \geq 13} p(\nu) = 0$					
O_2H_2	4	$p(1) = 0.06353$	$p(2) = 0.13521$	$p(3) = 0.17382$	$p(4) = 0.59182$	$2.2 - 2.2$	4.9
		$p(5) = 0.02679$	$p(6) = 0.00316$	$p(7) = 0.00300$	$p(8) = 0.00044$		
		$p(9) = 0.00038$	$p(10) = 0.00133$	$p(11) = 0.00028$	$p(12) = 0.00024$		
		$\sum_{\nu \geq 13} p(\nu) = 0$					
O_2	3	$p(1) = 0.10916$	$p(2) = 0.15623$	$p(3) = 0.12415$	$p(4) = 0.51046$	$1.13 - 1.14$	3.39
		$p(5) = 0.08631$	$p(6) = 0.00656$	$p(7) = 0.00182$	$p(8) = 0.00178$		
		$p(9) = 0.00264$	$p(10) = 0.00042$	$p(11) = 0.00005$	$p(12) = 0.00037$		
		$p(13) = 0.00005$	$\sum_{\nu \geq 14} p(\nu) = 0$				

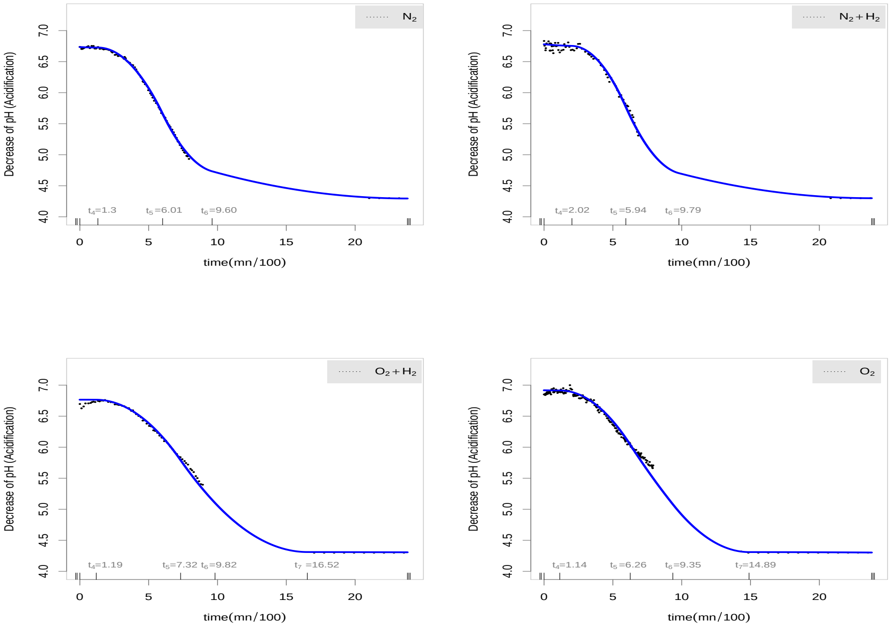

Figure 6: Constrained estimates for the four data without inconsistent points. The posterior means of knots are indicated by the vertical stripes.

set [19]. This data set provides the annual temperature anomalies from 1850 to 2012, expressed in degrees Celsius (see Figure 7). Anomalies are departures from the temperatures average between 1961 and 1990. For reasons of simplicity, we assume that observations are supposed to be independent and identically distributed as in Alvarez & Dey [3], Wu et al. [40]. The methodology of the present paper enables us to test whether the temperature anomalies sequence is a combination of increasing and decreasing localized constraints (H_1) or simply an increasing shape (H_0). Note that a combination of localized shape constraints (given in Figure 7) has never been considered in the free-knot context. In practice, it can be of interest to test $H_0 : f \in \mathcal{S}_0$ versus $H_1 : f \in \mathcal{S}_1$ for the two particular constraints $\mathcal{S}_0$ and $\mathcal{S}_1$. Consider a prior as explained in Section 3 and denote by $\mathit{\Pi}_{\mathcal{S}_0}$ and $\mathit{\Pi}_{\mathcal{S}_1}$ the prior distributions conditioned on $\mathcal{S}_0$ and $\mathcal{S}_1$ respectively. Note that simulations from $\mathit{\Pi}_{\mathcal{S}_j}$, $j \in \{0, 1\}$, can be obtained by replacing the posterior distribution by $\mathit{\Pi}_{\mathcal{S}_j}$ in the RJMCMC sampler. Then, it is straightforward to see that the Bayes factor can be approximated by

$$B_{01} = \frac{\int_{\mathcal{S}_0 \times [0,\tau_\sigma]} l_n(\theta) d\mathit{\Pi}_{\mathcal{S}_0}(f)\pi(\sigma^2)d\sigma^2}{\int_{\mathcal{S}_1 \times [0,\tau_\sigma]} l_n(\theta) d\mathit{\Pi}_{\mathcal{S}_1}(f)\pi(\sigma^2)d\sigma^2} \simeq \frac{\sum_\kappa l_n\Big(\beta^\kappa_{(0)}, (\sigma^2_{(0)})^\kappa, t^\kappa_{(0)}, \nu^\kappa_{(0)}\Big)}{\sum_\kappa l_n\Big(\beta^\kappa_{(1)}, (\sigma^2_{(1)})^\kappa, t^\kappa_{(1)}, \nu^\kappa_{(1)}\Big)},$$

where $l_n(\beta, \sigma^2, t, \nu)$ denotes the conditional density of data $\{x_i, y_i\}_{i=1}^n$ given $(\beta, \sigma^2, t, \nu)$ and where $(\beta^\kappa_{(j)}, (\sigma^2_{(j)})^\kappa, t^\kappa_{(j)}, \nu^\kappa_{(j)})$ is simulated from $\mathit{\Pi}_{\mathcal{S}_j}$ by the RJMCMC sampler. However, this approach to compute the Bayes factor often does not lead to accurate results, especially when the support of the prior and the posterior do not match. For this reason we propose the use of the harmonic mean of the likelihood values to approximate the Bayes factor, see e.g. [11], by

$$B_{01} \simeq \frac{N_0 \sum_{\kappa=1}^{N_1} \Big\{ l_n\Big(\tilde\beta^\kappa_{(1)}, (\tilde\sigma^2_{(1)})^\kappa, \tilde t^\kappa_{(1)}, \tilde\nu^\kappa_{(1)}\Big)\Big\}^{-1}}{N_1 \sum_{\kappa=1}^{N_0} \Big\{ l_n\Big(\tilde\beta^\kappa_{(0)}, (\tilde\sigma^2_{(0)})^\kappa, \tilde t^\kappa_{(0)}, \tilde\nu^\kappa_{(0)}\Big)\Big\}^{-1}}, \tag{6.1}$$

where $(\tilde\beta^\kappa_{(j)}, (\tilde\sigma^2_{(j)})^\kappa, \tilde t^\kappa_{(j)}, \tilde\nu^\kappa_{(j)})$ is simulated from the posterior under H_j, $j = \{0, 1\}$, by the RJMCMC sampler and where N_j is the number of iterations after burn in. The computation of the Bayes factor (6.1) gives $B_{01} = 3.2658 \times 10^{-12}$; hence a strong evidence in favor of H_1. For both posteriors under H_0 and H_1, we run the reversible jump Metropolis-Hastings algorithm for

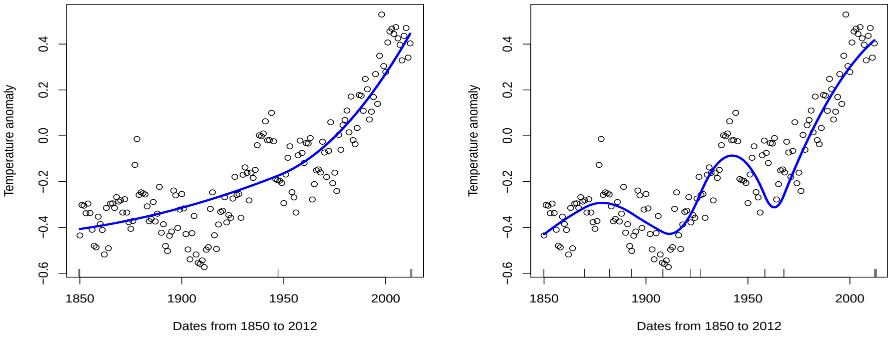

Figure 7: Figures shown the annual temperatures v.s. time, constrained estimate under H_0 and constrained estimate under H_1 for $\lambda = 0.5$. Posterior means of knots are indicated by the vertical stripes (under H_0, $\widehat{t} = (1849.5, 1849.9, 1850, 1947.2, 2012, 2012.5, 2012.9)$ and under H_1, $\widehat{t}$ =(1849.5,1849.9,1850 ,1869.863, 1882.176, 1892.944, 1908.273, 1921.671, 1926.586, 1958.322,1967.647,2012,2012.5,2012.9)).

10^5 iterations and check the convergence of the chain for different initial values. We also compute the posterior mean of $f(x)$ for every x in fine mesh over the whole interval $[1850, 2012]$ for each hypothesis H_j, $j \in \{0, 1\}$, for the sake of comparing their performance. As the simulations results in Figure 7 show, the free-knot prior under H_1 seems to outperform the free-knot monotone prior; The free-knot posterior under H_1 detects better the high and low variability regions of the data and facilitates the placement of more knots in the high variability region. In its turn, the free-knot posterior under H_1 succeeds to assign a number of knots that is compatible with the inhomogeneous structure of the data on the interval $[1850, 2012]$.

7. Auxiliary results and proofs

This section contains the proofs of some lemmas and theorems of the paper.

7.1. Proof of Lemma 1

Proof. Let $\gamma = (\beta, h, \sigma^2)$. We apply the dominated convergence theorem to show that $K(h_x h_y^{\theta_0}, h_x h_y^{\theta})$ is continuous as a function of γ. From (5.1) and

(5.2), there exists $\theta^* \in (\mathcal{S}^C, [0, \tau_\sigma])$, or in an equivalent form there exists $(\gamma^*, \nu^*) \in (D_\nu \times [0, \tau_\sigma], \{1, \dots, q\})$, such that

$$K(h_x h_y^{\theta_0}, h_x h_y^{\theta^*}) < \delta + \frac{\epsilon}{4}.$$

Now, because the divergence K is continuous at γ^*, there exists an open neighborhood $N_\epsilon \subset D_{\nu^*} \times [0, \tau_\sigma]$ of γ^* such that $K(h_x h_y^{\theta_0}, h_x h_y^{\theta^v}) < \delta + \epsilon/2$ for every $v \in N_\epsilon$. Now we turn to concluding the argument by writing

$$\mathit{\Pi}(A_\epsilon) \geq \mathit{\Pi}(N_\epsilon) \geq \pi(\nu^*) \int_{N_\epsilon} \mathit{\Pi}(\gamma) d\gamma > 0,$$

which completes the proof. □

7.2. Proof of Theorem 1

The proof is split into two mains lemmas. In addition to the Lemma 1, we consider the following Lemma 3

Lemma 3. *Let P_0^* denote the outer measure of P_0. The class*

$$\mathcal{G} = \Big\{\theta \in (\mathcal{S}^C, [0, \tau_\sigma]) : g = \log\Big(\frac{h_y^{\theta_0}}{h_y^{\theta}}\Big)\Big\}$$

is a Glivenko-Cantelli class, which means

$$\sup_{\theta \in (\mathcal{S}^C, [0, \tau_\sigma])} \Big| \frac{1}{n} \sum_{i=1}^{n} g\big(x_i, y_i\big) - K(h_x h_y^{\theta_0}, h_x h_y^{\theta}) \Big| \to 0, \qquad \textit{almost surely } P_0^{*\infty}.$$

Proof. It is straightforward that the function g decomposes as

$$g = \log(h_y^{\theta_0}) + \log(\sqrt{2\pi}\sigma) + \frac{1}{2\sigma^2}\big(y^2 - 2yf(x) + f(x)^2\big).$$

The aim is to prove that every one of the following classes

$$\mathcal{G}_1 = \Big\{\sigma^2 \in [0, \tau_\sigma] : \quad \log(h_y^{\theta_0}) + \log\Big(\sqrt{2\pi}\sigma \exp(\frac{y^2}{2\sigma^2})\Big)\Big\};$$
$$\mathcal{G}_2 = \Big\{\theta \in (\mathcal{S}^C, [0, \tau_\sigma]) : \quad -\frac{2yf(x)}{2\sigma^2}\Big\};$$
$$\mathcal{G}_3 = \Big\{\theta \in (\mathcal{S}^C, [0, \tau_\sigma]) : \quad \frac{f(x)^2}{2\sigma^2}\Big\},$$

is Glivenko-Cantelli. To show these, it suffices to show that these sets are Vapnik-Cervonenkis (VC) classes together with the fact that the envelope functions of $\mathcal{G}_1$, $\mathcal{G}_2$ and $\mathcal{G}_3$ are integrable. We recall that, for any $g \in \mathcal{G}$, G is an envelope function for $\mathcal{G}$ if $|g(x)| < G(x)$. First, as $\mathcal{G}_1$ consists of one element, thus obviously it is a VC class. The class $sp = \{\beta \in S^C : f = B\beta\}$ is by construction the subset of all piecewise polynomials fulfilled the shape constraints S^C. We recall that each function $B_j(x)$ is degree $(k-1)$ polynomials on each subinterval $x \in (t_j, t_{j+1})$ for $j = \{1, \ldots, m+k-1\}$ and the function $B\beta$ is a degree $(k-1)$ polynomial on the same subinterval. The degree $k-1$ is less than some number $(< k)$. For $m = \nu + k$, sp is the subset of linear combinations of a given, finite set of functions $B_1, \ldots, B_m$. These implies that sp is a VC class from van der Vaart [33, p.276]. Thus, from sp we deduce immediately that the class $\mathcal{G}_2$ is VC by lemma 2.6.18 cited in van der Vaart & Wellner [34, p.147]. Now, we consider the class $sp' = \{\beta \in S^C : f^2 = (B\beta)^2\}$. Then, straightforwardly the product $B\beta \times B\beta$ is a degree $2(k-1)$ polynomial on each subinterval $x \in (t_j, t_{j+1})$ for $j = \{1, \ldots, m+k-1\}$. The degree $2(k-1)$ is less than some number $(< 2k-1)$. A similar argument of the preceding shows also that the class $\mathcal{G}_3$ is VC. It is easy to verify that, for any $g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2, g_3 \in \mathcal{G}_3$, there is a measurable functions G_1, G_2 and G_3 such that

$$|g_1(x)| < G_1(x),\ |g_2(x)| < G_2(x),\ |g_3(x)| < G_3(x).$$

Finally, using the stability of the Glivenko-Cantelli property [34, p.125] it is clear that the class $\{\mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3\}$ is Glivenko-Cantelli. □

We now proceed to the proof of the Theorem 1.

Proof. Let us consider

$$Q_n = \frac{1}{n} \sum_{i=1}^{n} g\big(x_i, y_i\big).$$

From Lemma 3, for any $0 < d_\epsilon < \epsilon/6$ and for n sufficiently large, we have

$$K(h_x h_y^{\theta_0}, h_x h_y^{\theta}) - d_\epsilon < Q_n < K(h_x h_y^{\theta_0}, h_x h_y^{\theta}) + d_\epsilon$$

where $\theta \in (\mathcal{S}^C, [0, \tau_\sigma])$. In particular, for $\theta \in A_\epsilon$ we have

$$Q_n < \delta + \frac{\epsilon}{2} + d_\epsilon$$

and for $\theta \in A_{\epsilon+6d_\epsilon}$ we have

$$Q_n > \delta + \frac{\epsilon}{2} + 2d_\epsilon.$$

Let denote $A^c_{\epsilon+6d_\epsilon}$ the complementary of $A_{\epsilon+6d_\epsilon}$. It is clear that $A_\epsilon \subset A^c_{\epsilon+6d_\epsilon}$. It follows from these that

$$\begin{aligned}
\frac{\mathit{\Pi}_n(A_{\epsilon+6d_\epsilon})}{\mathit{\Pi}_n(A^c_{\epsilon+6d_\epsilon})} &\leq \frac{\mathit{\Pi}_n(A_{\epsilon+6d_\epsilon})}{\mathit{\Pi}_n(A_\epsilon)} \\
&= \frac{\int_{A_{\epsilon+6d_\epsilon}} l_n(\theta) d\mathit{\Pi}(f)\pi(\sigma^2)d\sigma^2}{\int_{A_\epsilon} l_n(\theta) d\mathit{\Pi}(f)\pi(\sigma^2)d\sigma^2} \\
&= \frac{\int_{A_{\epsilon+6d_\epsilon}} \prod_{i=1}^n h_y^\theta(y_i) d\mathit{\Pi}(f)\pi(\sigma^2)d\sigma^2}{\int_{A_\epsilon} \prod_{i=1}^n h_y^\theta(y_i) d\mathit{\Pi}(f)\pi(\sigma^2)d\sigma^2} \\
&= \frac{\int_{A_{\epsilon+6d_\epsilon}} \exp\left(-nQ_n\right) d\mathit{\Pi}(f)\pi(\sigma^2)d\sigma^2}{\int_{A_\epsilon} \exp\left(-nQ_n\right) d\mathit{\Pi}(f)\pi(\sigma^2)d\sigma^2} \\
&\leq \frac{\exp\{-n(\delta + \frac{\epsilon}{2} + 2d_\epsilon)\}\mathit{\Pi}(A_{\epsilon+6d_\epsilon})}{\exp\{-n(\delta + \frac{\epsilon}{2} + d_\epsilon)\}\mathit{\Pi}(A_\epsilon)} = \exp\left(-nd_\epsilon\right)\frac{\mathit{\Pi}(A_{\epsilon+6d_\epsilon})}{\mathit{\Pi}(A_\epsilon)}.
\end{aligned}$$

Now, we know from Lemma 1 that, for all $\epsilon > 0$, $\mathit{\Pi}(A_\epsilon) > 0$. Thus, it follows the following convergence

$$\exp\left(-nd_\epsilon\right)\frac{\mathit{\Pi}(A_{\epsilon+6d_\epsilon})}{\mathit{\Pi}(A_\epsilon)} \to 0 \qquad \text{almost surely } P_0^\infty. \tag{7.1}$$

We have already shown that the latter (7.1) converges to 0 almost surely P_0^∞ for all $\epsilon > 0$. Therefore, $\mathit{\Pi}_n(A_{2\epsilon}) \to 1$ almost surely P_0^∞. This completes the argument. □

7.3. Proof of Corollary 1

Proof. We note that it suffices to prove the following relation between norms

$$K(h_x h_y^{\theta_0}, h_x h_y^\theta) \gtrsim \int_{a_0}^{b_0} \big(f_0(x) - f(x)\big)^2 h_x(x) dx. \tag{7.2}$$

We first write

$$\begin{aligned}
K(h_x h_y^{\theta_0}, h_x h_y^\theta) &= \int_{a_0}^{b_0} \int_{-\infty}^{\infty} h_x(x) h_y^{\theta_0}(y) \log\Big(\frac{h_y^{\theta_0}(y)}{h_y^\theta(y)}\Big) dy dx \\
&\geq \int_{a_0}^{b_0} \Big(\text{const} \int_{-\infty}^{\infty} |h_y^{\theta_0}(y) - h_y^\theta(y)| dy\Big)^2 h_x(x) dx \qquad (7.3)
\end{aligned}$$

where the inequality (7.3) is a result of Kemperman [20, p.2174] and the constant const= $\frac{1}{\sqrt{2}}$ is the best possible. Now we shall be interested in the squared total variation norm term that appears in (7.3). It is easily seen that the only remaining problem is to establish lower-bound on $\|h_y^{\theta_0}(y)-h_y^{\theta}(y)\|_{TV}$ in terms of L_2-distance of the regression functions f_0 and f. To show that, let consider $\mathcal{L}(x,y)$ that lies between $y-f_0(x)$ and $y-f(x)$ and note

$$\begin{aligned}
\int_{\mathbb{R}} |h_y^{\theta_0}(y)-h_y^{\theta}(y)|dy &= (2\pi\sigma^2)^{-\frac{1}{2}}\int_{\mathbb{R}}\Big|\exp\{-\frac{(y-f_0(x))^2}{2\sigma^2}\}-\exp\{-\frac{(y-f(x))^2}{2\sigma^2}\}\Big|dy \\
&= (2\pi\sigma^2)^{-\frac{1}{2}}\int_{\mathbb{R}}\exp\{-\frac{(\mathcal{L}(x,y))^2}{2\sigma^2}\}\Big|\frac{(y-f_0(x))^2}{2\sigma^2}-\frac{(y-f(x))^2}{2\sigma^2}\Big|dy \\
&= (8\pi\sigma^6)^{-\frac{1}{2}}\Big|f_0(x)-f(x)\Big|\int_{\mathbb{R}}\exp\{-\frac{(\mathcal{L}(x,y))^2}{2\sigma^2}\}\Big|2y-f_0(x)-f(x)\Big|dy \\
&\geq (8\pi\sigma^6)^{-\frac{1}{2}}\Big|f_0(x)-f(x)\Big|\int_{-1}^{1}\exp\{-\frac{(\mathcal{L}(x,y))^2}{2\sigma^2}\}\Big|2y-f_0(x)-f(x)\Big|dy \\
&\geq C_0\Big|f_0(x)-f(x)\Big|\int_{-1}^{1}\Big|2y-f_0(x)-f(x)\Big|dy \\
&\geq C_1\Big|f_0(x)-f(x)\Big|, \qquad (7.4)
\end{aligned}$$

where (7.4) is obtained by application of the mean-value theorem and the constant C_1 depends only on f, f_0 and σ^2. Therefore we have, by combining (7.4) and (7.3), that

$$K(h_x h_y^{\theta_0}, h_x h_y^{\theta}) \gtrsim \int_{a_0}^{b_0}\big(f_0(x)-f(x)\big)^2 h_x(x)dx.$$

as we wanted to prove. □

8. Discussion

In this paper, a Bayesian regression with free-knot B-splines under localized shape restrictions is provided. On the one hand, our approach enables us to take into account combinations of shape constraints and to localize each shape constraint on a given interval. On the other hand, our main contribution is in providing a framework for constructing prior distributions, with known normalizing constants, which enables us to compute the simulation from the posterior distribution using a reversible jump Metropolis-Hastings scheme. Various tools from constrained optimization and asymptotic analysis are exploited to establish Bayesian weak and strong consistency. These techniques can be extended to handle convergence rates. Hence, the results developed in this paper open a door to other complex Bayesian nonparametric problems subject to shape constraints. However, an important question addressed partially in this paper is whether it is appropriate to impose a specific shape restrictions. The problem of testing for shape constraints by Bayes factor suggests some improvements. Clearly, a theoretical study to characterize the asymptotic properties of the Bayes factor for comparing between different combinations of shape constraints could be interesting to consider.

References

[1] Abraham C. (2012). Bayesian regression under combinations of constraints. *Journal of Statistical Planning and Inference*, **142**, 2672–2687.

[2] Abraham C. & Khadraoui K. (2015). Bayesian regression with b-splines under combinations of shape constraints and smoothness properties. *Statistica Neerlandica*, **69**, 150–170.

[3] Alvarez E. & Dey D. (2009). Bayesian isotonic changepoint analysis. *Annals of the Institute of Statistical Mathematics*, **61**, 355–370.

[4] Alvarez E. & Yohai V. (2012). M-estimators for isotonic regression. *Journal of Statistical Planning and Inference*, **142**, 2351–2368.

[5] Bornkam B. & Ickstadt K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics*, **65**, 198–205.

[6] Brezger A. & Steiner W. (2008). Monotonic regression based on bayesian p-splines: An application to estimating price response functions from store-level scanner data. *Journal of Business and Economic Statistics*, **26**, 91–104.

[7] de Boor C. (2001). *A practical guide to splines*. Springer-Verlag, New-York, NY.

[8] Denison D., Mallick B. & Smith A. (1998). Automatic bayesian curve fitting. *Journal of the Royal Statistical Society (B)*, **60**, 333–350.

[9] Diaconis P. & Freedman D. (1986). On the consistency of bayes estimates. *Annals of statistics*, **14**, 1–26.

[10] DiMatteo I., Genovese C. & Kass R. (2001). Bayesian curve-fitting with free-knot splines. *Journal of the Royal Statistical Society (B)*, **88**, 1055–1071.

[11] Gelfand A. & Dey D. (1994). Bayesian model choice: asymptotics and exact calculations. *Journal of Royal Statistical Society B*, **56**, 571–578.

[12] Ghosal S., Ghosh J. & van der Vaart A. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, **28**, 500–531.

[13] Ghosal S. & van der Vaart A. (2007). Convergence rates of posterior distribution for noniid observations. *The Annals of Statistics*, **35**, 192–223.

[14] Green P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711–732.

[15] Grenander U. (1956). On the theory of mortality measurement. *Scandinavian Actuarial Journal*, **1956**, 125–153.

[16] Groeneboom P., Jongbloed G. & Wellner J. (2001). Estimation of a convex function: characterizations and asymptotic theory. *Annals of statistics*, **29**, 1653–1698.

[17] Holmes C. & Heard N. (2003). Generalized monotonic regression using random change points. *Statistics in Medecine*, **22**, 623–638.

[18] Johnson M. (2007). Modeling dichotomous item response with free-knot splines. *Computational Statistics and Data Analysis*, **51**, 4187–4192.

[19] Jones P., Parker D., Osborn T. & Briffa K. (2013). Global and hemispheric temperature anomaliesland and marine instrumental records. *In Trends: A Compendium of Data on Global Change*, p. doi: 10.3334/CDIAC/cli.002.

[20] Kemperman J. (1969). On the optimum rate of transmitting information. *The Annals of Mathematical Statistics*, **40**, 2156–2177.

[21] Lindstrom M. (2002). Bayesian estimation of free-knot splines using reversible jump. *Computational Statistics and Data Analysis*, **41**, 255–269.

[22] Mammen E., Marron J., Turlach B. & Wand M. (2001). A general projection framework for constrained smoothing. *Statistical Science*, **16**, 232–248.

[23] Mammen E. & Thomas-Agnan C. (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics*, **26**, 239–252.

[24] Meyer M. (2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics*, **2**, 1013–1033.

[25] Meyer M., Hackstadt A. & Hoeting J. (2011). Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *Journal of Nonparametric Statistics*, **23**, 867–884.

[26] Neelon B. & Dunson D. (2004). Bayesian isotonic regression and trend analysis. *Biometrics*, **60**, 398–406.

[27] Ramsay J. (1988). Monotone regression splines in action. *Statistical Science*, **3**, 425–461.

[28] Robertson T. & Wright F. (1975). Consistency in generalized isotonic regression. *Annals of statistics*, **3**, 350–362.

[29] Schwartz L. (1965). On bayes procedures. *Z. Wahrsch. Verw. Gebiete*, **4**, 10–26.

[30] Shen J. & Wang X. (2011). Estimation of monotone functions via p-splines: A constrained dynamical optimization approach. *SIAM Journal of Control and Optimization*, **49**, 646–671.

[31] Shively T. & Sager T. (2009). A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society (B)*, **71**, 159–175.

[32] Turlach B. (2005). Shape constrained smoothing using smoothing splines. *Computational Statistics*, **20**, 81–103.

[33] van der Vaart A. (1998). *Asymptotic statistics.* Cambridge University Press, Cambridge.

[34] van der Vaart A. & Wellner J. (1996). *Weak convergence and empirical processes*. Springer, New York.

[35] Walker S. & Hjort N. (2001). On bayesian consistency. *Journal of Royal Statistical Society B*, **63**, 811–821.

[36] Wang X. (2008). Bayesian free-knot monotone cubic spline regression. *Journal of Computational and Graphical Statistics*, **17**, 373–387.

[37] Wang X. & Shen J. (2010). A class of grouped brunk estimators and penalized spline estimators for monotone regression. *Biometrika*, **97**, 585–601.

[38] Wang X. & Shen J. (2013). Uniform convergence and rate adaptive estimation of convex functions via constrained optimization. *SIAM Journal of Control and Optimization*, **51**, 2753–2787.

[39] Wasserman L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of Royal Statistical Society B*, **62**, 159–180.

[40] Wu W., Woodroofe M. & Mentz G. (2001). Isotonic regression: another look at the changepoint problem. *Biometrika*, **88**, 793–804.

[41] Zhao O. & Woodroofe M. (2012). Estimating a monotone trend. *Statistica Sinica*, **22**, 359–378.