Noname manuscript No.
(will be inserted by the editor)

Degree Selection Methods for Curve Approximation via Bernstein Polynomials

Juliana Freitas de Mello e Silva · Sujit Kumar Ghosh · Vinícius Diniz Mayrink

Received: date / Accepted: date

Abstract It is well-known that Bernstein Polynomials (BPs) provide a suitable basis of functions that can be used to uniformly approximate any continuous function based on observed noisy samples from that unknown function. Yet, the primary challenge that still persists is the selection of an 'optimal' degree of the BPs, solely based on the observed data. In the absence of noises, the asymptotic theory suggests larger degree leads to better approximation but with noises (thus reducing biases), however, a larger degree also leads to high-dimensional parameter estimation and so a balance of classic bias-variance trade-off is warranted. The primary goal of this work is obtain minimum possible degree of the approximating BPs using probabilistic methods that are robust to various shapes of the unknown continuous function. In addition to providing theoretical guidance, the paper also provides extensive numerical illustrations to study the problem of obtaining minimal degree of BPs in approximating arbitrary continuous functions.

Keywords Bayesian inference · Bernstein basis · degree elevation · MCMC · model selection · optimal degree.

1 Introduction

Bernstein Polynomials (BP) were developed by Sergei Natanovich Bernstein in 1912 when he was proving a demonstration to a special case of Weiertrass' Theorem (2; 14). The main utility of the BP lies on approximating any smooth curve/function. The usage of polynomials to approximate functions

Juliana Freitas de Mello e Silva, Vinícius Diniz Mayrink
Departamento de Estatística, Universidade Federal de Minas Gerais
E-mail: julianafms27@gmail.com

Sujit Kumar Ghosh
Department of Statistics, North Carolina State University

has analytic advantages since they are easily written in the form of summation. Admittedly, polynomials are also known to be notorious as interpolating functions due to the well known *Runge's phenomenon*, but our problem is not about interpolation of function at equally spaced nodes but rather statistical estimation of unknown function based on noisy data. Moreover, among the class of polynomial basis functions, it has been shown that BPs enjoy an optimality property in preserving shapes as explained and demonstrated by (5). The BP approximation of a continuous function $f(x)$ defined on $[0,1]$ with degree/order m is given by

$$B_m(x;f) = \sum_{k=0}^{m} f\left(\frac{k}{m}\right) b_{k,m}(x), \tag{1}$$

where $b_{k,m}(x) = \binom{m}{k} x^k (1-x)^{m-k}$ for $k = 0, 1, \ldots, m$ are the basis functions. It has been demonstrated by Bernstein that $\max_{x\in[0,1]} |B_m(x;f) - f(x)| \longrightarrow 0$, as $m \longrightarrow \infty$ (see (2)). Notice that the interval $[0,1]$ can be replaced by any arbitrary interval $[a,b]$ by a linear transformation and the approximation remains valid on any closed interval.

For statistical estimation of the regression function $f(x)$ based on a model $y_i = f(x_i) + \epsilon_i$ for $i = 1, \ldots, n$, we assume ϵ_i's are independent noises with mean zero and bounded variance. Defining $\beta_k = f(k/m)$ for $k = 0, 1, \ldots, m$ as unknown parameters in Equation (1) we can use multiple linear regression framework to obtain estimates of the parameter vector $\boldsymbol{\beta}_m = (\beta_0, \ldots, \beta_m)$ where the entries of the $n \times (m+1)$ design matrix $\boldsymbol{Z}$ is given by $z_{i,k+1} = b_{k,m}(x_i)$ for $k = 0, 1, \ldots, m$ and $i = 1, 2, \ldots, n$. The primary focus is this work is the selection m that needs to be determined by the observed data as well.

In the statistical literature, there are abundant use of BPs for not only regression functions but also for estimating density, distribution and hazard functions among others, especially when it is desired to maintain a given shape of the functions (e.g., monotonicity, convexity etc.): some works make use of BP in the context of density estimation (20; 16; 17; 1). The BP can be applied to approximate density as well as intensity estimation function in a Poisson Process model (13). It can also be used as kernel estimator (4). Moreover, BP can be used aiming at variable selection, as in (9), and in shape-restriction problems like in (6) and (21).

The choice of the degree of the BP represents a great challenge due to the fact that it has an important role in the approximation performance. This characteristic was verified both in literature (16; 15; 22) and in practice. If the degree m is chosen too small, the approximated curve may have a large bias. On the other hand, if this m is chosen too large, it will inflate the variance of the parameter $\boldsymbol{\beta}$. Thus, the choice of m is equivalent to that of a bandwidth parameter (for density estimation) or penalty parameter (for regression function estimation). Nonetheless, the convergence of the BP is only guaranteed

when the degree $m \longrightarrow \infty$ (2; 14). Then, in a general framework, an "optimal" choice for the degree would be the *minimum* value that manages to approximate well and that accommodates the important features of the target function. There have been many results (see e.g., (15), (21), (24), (23), and (3)) on the "optimal" asymptotic choice of $m = O(n^{\alpha})$ for some $\alpha \in (0, 1)$, however such choices though of theoretical value, are not useful in practice. E.g., for the optimal rate $\alpha > 0$, if we set $m = Cn^{\alpha}$, the constant $C > 0$ usually depends on the unknown function f itself making is practically not useful. Thus, a data-dependent automatic choice of m is needed.

The data dependent selection m for the BPs have been also explored in the literature (e.g., (9) proposed the use of information criteria on choosing m). More recently, (24) claimed that their method seemed to be robust concerning the degree of the BP; but, perhaps, this result may be more related to the simplicity and smoothness of the function they were targeting than associated to the robustness of the method itself. Other novels attempts includes works like (16), (17), (7), (6) and (8) who have considered the degree of the BP as random quantity to be estimated within a Bayesian framework.

We propose a solution that consists on establishing a routine to choose the minimum suitable degree for the BP. This instruction is probabilistic method based on a previous knowledge of a possibly existing turning point on the target function. Here, we mention in anticipation that our method can still be used if one does not know this feature in the function. Our findings point out that the degree of the BP is more associated to how close a change in the behavior of the target function is to the boundaries of the domain, than it is related to sample sizes. We also derived two criteria to serve as stopping rule for the degree selection. This stopping rule is based on the degree elevation property of BPs (see (10)).

To the best of our knowledge, only one work (12) proposed an estimator for the degree of the BP and studied its performance. The mentioned proposal is different from ours. Our method seems to be more attractive in the sense that it is based on probabilities and properties of the BP. In addition, an entire distribution of the minimum degree can be studied. Also, Guan's method is based on the frequentist approach. On the other hand, the application of our method requires a previous knowledge of where the target function will change behavior to establish a minimum degree, and a posterior sample of the vector of coefficients to choose an optimal degree.

This paper is organized as follows: Section 2 explains how to use the BP to approximate curves. In addition, we introduce and discuss our proposed methods for the degree selection. Then, in Section 3 we present a simulation study to show the good performance of our proposals. In turn, in Section 4, we bring a short analysis of a real data set to illustrate how our methods can be used. At last, in Section 5 we discuss our conclusions on this matter.

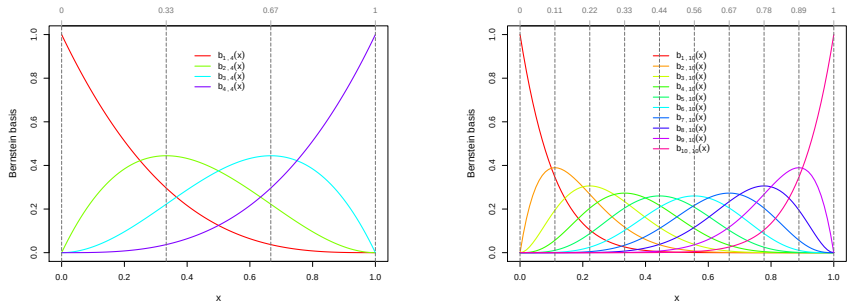
2 Bernstein Polynomials to approximate curves

The BP approximation for a function $f(t)$ with degree $m-1$ is given by

$$B_{m-1}(t;f) = \sum_{k=1}^{m} f\left(\frac{k-1}{m-1}\right)\binom{m-1}{k-1}\left(\frac{t}{T_{max}}\right)^{k-1}\left(1-\frac{t}{T_{max}}\right)^{m-k}$$
$$= \sum_{k=1}^{m} \xi_{k,m-1} b_{k,m-1}\left(\frac{t}{T_{max}}\right), \qquad (2)$$

where T_{max} is such that $t/T_{max} \in (0,1)$ and $\boldsymbol{\xi}_{m-1} = (\xi_{1,m-1}, \xi_{2,m-1}, \ldots, \xi_{m,m-1})$ is the vector of coefficients. Each component of this vector represents the target function $f(.)$ in a point of the domain, i.e., $\xi_{k,m-1} = f((k-1)/(m-1))$, for $k = 1, 2, \ldots, m$. Note by Equation (2) that each coefficient is related to a Bernstein basis $b_{k,m-1}(.)$, $k = 1, 2, \ldots, m$. With this same equation we can notice that the approximation provided by the BP is a linear combination of these two vectors. Since Bernstein basis can be regarded to as weights, it means that each coefficient $\xi_{k,m-1}$ will have more weight on sub-intervals of the domain that takes place around t such that $b'_{k,m-1}\left(\frac{t}{T_{max}}\right) = 0 \Leftrightarrow \frac{t}{T_{max}} = \frac{k-1}{m-1} \Leftrightarrow t = \frac{k-1}{m-1}T_{max}$, $k = 1, 2, \ldots, m$.

In Figure 1, each colored line represent a Bernstein basis, of a BP with degree $m-1$, for $t/T_{max} \in (0,1)$ and the vertical dotted gray lines represent at which point t each basis reaches its maximum value, according to the relationship above.



(a) $m = 4$. (b) $m = 10$.

Fig. 1: Illustration of the vector of Bernstein basis, of a BP with degree $m-1$, for $m = 4$ and $m = 10$ and the representation of the time where these functions reach their maximum.

In what follows, one of our interests lies on studying the increasing and decreasing behavior of the target function f being approximated. Thus, we

must evaluate the derivative of Bernstein approximation for this function. The derivative of $B_{m-1}(t;f)$ is

$$\begin{aligned} B'_{m-1}(t;f) &= \sum_{k=1}^{m} \xi_{k,m-1} \binom{m-1}{k-1} \left[\frac{(k-1)}{T_{max}} \left(\frac{t}{T_{max}}\right)^{k-2} \left(1-\frac{t}{T_{max}}\right)^{m-k} - \right. \\ &\quad \left. \frac{(m-k)}{T_{max}} \left(\frac{t}{T_{max}}\right)^{k-1} \left(1-\frac{t}{T_{max}}\right)^{m-k-1} \right] \\ &= \frac{(m-1)}{T_{max}} \sum_{k=1}^{m-1} (\xi_{k+1,m-1} - \xi_{k,m-1}) \binom{m-2}{k-1} \left(\frac{t}{T_{max}}\right)^{k-1} \left(1-\frac{t}{T_{max}}\right)^{m-k-1} \\ &= \frac{(m-1)}{T_{max}} \sum_{k=1}^{m-1} (\xi_{k+1,m-1} - \xi_{k,m-1}) b_{k,m-2}\left(\frac{t}{T_{max}}\right). \end{aligned} \tag{3}$$

It is well-known that if $B'_{m-1}(t;f) < 0$, the curve $B_{m-1}(t;f)$ presents a decreasing behavior and if $B'_{m-1}(t;f) > 0$, then $B_{m-1}(t;f)$ will be an increasing function on t. Also, it is clear that $(m-1)/T_{max} \geq 0$ and $b_{k,m-2}(t/T_{max}) \geq 0$, for $k = 1, 2, \ldots, m-1$. As a conclusion, the only term in Equation (3) that controls the increasing/decreasing behavior of this approximation is the vector of coefficients $\boldsymbol{\xi}_{m-1}$. This result is coherent considering that each of the coefficients represents the function f in a specific point of the domain. Therefore, the vector of Bernstein basis does not affect the form of the approximation directly.

Then, if the difference $(\xi_{k,m-1} - \xi_{k-1,m-1})$ has an opposite sign than $(\xi_{k+1,m-1} - \xi_{k,m-1})$ or, equivalently, if $(\xi_{k,m-1} - \xi_{k-1,m-1})(\xi_{k+1,m-1} - \xi_{k,m-1}) < 0$, for $k = 2, 3, \ldots, m-2$, we can assure that the approximated function f has a change of behavior. This behavior is such that it begins at $\xi_{k-1,m-1} \approx f\left(\frac{k-2}{m-1} T_{max}\right)$, it happens around $\xi_{k,m-1} \approx f\left(\frac{k-1}{m-1} T_{max}\right)$, and it finally ends at $\xi_{k+1,m-1} \approx f\left(\frac{k}{m-1} T_{max}\right)$. Therefore, we need three coefficients to be able to inform this feature. Thus, both k and m impact this result and this method is capable of detecting a change for a time $t \geq 1/(m-1)$ and $t \leq (m-2)/(m-1)$. In other words, we will be able to capture this change only on an interval (a, b) such that $\frac{1}{m-1} < a < b < \frac{m-2}{m-1} \Leftrightarrow m > \max\left(\frac{1}{a} + 1, \frac{2-b}{1-b}\right)$.

From now on, we will focus on the interval $(0, 1)$, because any scale can be transformed into this range. Thus, consider two random variables U_1 and U_2 defined on $(0, 1)$ that will form an interval $(u_{(1)}, u_{(2)}) \subset (0, 1)$, where $u_{(1)} = \min(U_1, U_2)$ and $u_{(2)} = \max(U_1, U_2)$. Assume a random variable $M = \left\lceil \max\left(\frac{1}{U_{(1)}} + 1, \frac{2-U_{(2)}}{1-U_{(2)}}\right) \right\rceil$, which refers to the minimum degree that is necessary to capture the change point in the interval $(u_{(1)}, u_{(2)})$. We have that

$$\mathbb{P}(M = m|(U_1, U_2)) = \mathbb{P}\left(\left\lceil \max\left(\frac{1}{U_{(1)}} + 1, \frac{2 - U_{(2)}}{1 - U_{(2)}}\right)\right\rceil = m\right) \tag{4}$$
$$= \left[\mathbb{P}\left(U_1 \leq \frac{m-2}{m-1}\right) - \mathbb{P}\left(U_1 < \frac{1}{m-1}\right)\right]^2 -$$
$$\left[\mathbb{P}\left(U_1 \leq \frac{m-3}{m-2}\right) - \mathbb{P}\left(U_1 < \frac{1}{m-2}\right)\right]^2,$$

for $m = 4, 5, \ldots$ (more details about this result can be found in Section A.1.1 of the Appendix). The cumulative distribution function of M conditioned on (U_1, U_2) is

$$F_{M|(U_1,U_2)}(m) = \sum_{l=4}^{m} \left\{\left[\mathbb{P}\left(U_1 \leq \frac{l-2}{l-1}\right) - \mathbb{P}\left(U_1 < \frac{1}{l-1}\right)\right]^2 -\right.$$
$$\left.\left[\mathbb{P}\left(U_1 \leq \frac{l-3}{l-2}\right) - \mathbb{P}\left(U_1 < \frac{1}{l-2}\right)\right]^2\right\}$$
$$= \left[\mathbb{P}\left(U_1 \leq \frac{m-2}{m-1}\right) - \mathbb{P}\left(U_1 < \frac{1}{m-1}\right)\right]^2. \tag{5}$$

Once we have derived the cumulative distribution function of M, it becomes trivial to calculate quantiles of the minimum degree. Table 1 shows some quantiles considering that U_1 and U_2 follow a $Beta(\theta_1, \theta_2)$ distribution. The values of θ_1 and θ_2 were chosen so that we could have high densities near 0 and/or 1 as well as densities concentrated in the middle of this interval.

Table 1: Quantiles of the minimum degree that is necessary to model a change in the approximated curve on an interval $(U_{(1)}, U_{(2)})$; greatest m such that $\mathbb{P}(M \leq m|(U_1, U_2)) \leq p$.

Distribution of U_1 and U_2	$p = 0.5$	$p = 0.9$	$p = 0.95$	$p = 0.99$	$p = 0.999$
$Beta(1, 1)$	7	39	79	399	3999
$Beta(3, 1)$	10	58	118	598	5998
$Beta(2, 150)$	141	418	618	1453	4709
$Beta(3, 3)$	4	7	9	16	34
$Beta(50, 800)$	19	22	23	25	29
$Beta(200, 1300)$	8	9	9	9	10

As an example, suppose that a change is expected on the interval $(u_{(1)}, u_{(2)})$ such that both U_1 and U_2 follow a $Beta(3, 1)$. Then, if $m = 10$ we have a 0.5 probability that the change will be represented in the approximated curve. If we want to be more accurate about this result, we should consider $m \geq 118$, which is a very high degree.

The first example in Table 1 is an $Uniform(0,1)$ distribution; in this case we would have no prior information about where the function being approximated may change its behavior. We can see that $\mathbb{P}(M \leq 39|(U_1, U_2)) \approx 0.9$. In what follows, the next two examples $Beta(3,1)$ and $Beta(2,150)$ admit high density on the lower boundary. So, it is required a large minimum degree in order for the BP to be able to approximate this behavior. On the other hand, if the focus relies far from 0 and 1, m can be considerably lower, as in the last three rows of Table 1.

In practice, one can follow the subsequent step-by-step in order to choose the minimal degree m:

Step 1: In which part of the domain the curve will probably change behavior? If the domain is other than $(0,1)$, then proceed to the next step with a transformation on the scale;
Step 2: Find (θ_1, θ_2) such that $Beta(\theta_1, \theta_2)$ has high density on the region above;
Step 3: Find quantiles of minimum m by solving $\mathbb{P}(M \leq m|(U_1, U_2)) = p \Leftrightarrow F_{Beta}\left(\frac{m-2}{m-1}; \theta_1, \theta_2\right) - F_{Beta}\left(\frac{1}{m-1}; \theta_1, \theta_2\right) - \sqrt{p} = 0$.

The instructions above are a conclusion that leads to a probabilistic view of a suitable minimum degree for the BP. We highlight that all the discussion and results involving this minimum value is a contribution of this paper. Next, we focus on completing the guidance on choosing the degree by discussing a reasonable maximum value for m.

2.1 Degree elevation to achieve optimal m

The stopping rule we will discuss in this section was based on the degree elevation property, seen in the work of (10).

Property 1 (Degree elevation) This property refers to the relationship between the vector of Bernstein coefficients of degree m and $m + r$, $r = 1, 2, \ldots$. It allows us to determine coefficients of a BP with degree $m + r$ given the vector of coefficients of a BP with degree m. This relationship is

$$\tilde{\xi}_{k,m+r} = \sum_{j=\max(0,k-r)}^{\min(m,k)} \frac{\binom{r}{k-j}\binom{m}{j}}{\binom{m+r}{k}} \xi_{k,m},\ k = 1, 2, \ldots, m + r, \tag{6}$$

where $\tilde{\xi}_{k,m+r}$ represents the k-th coefficient of a BP with degree $m+r$ obtained via degree elevation and $\xi_{k,m}$ is the k-th coefficient from a BP with degree m. Note that: (i) the Binomial coefficient terms of the summation in Equation (6) come from a $HyperG(m + r, m, k)$ distribution and (ii) if $r = 1$, Equation (6) reduces to

$$\tilde{\xi}_{k,m+1} = \frac{k}{m+1}\xi_{k-1,m} + \left(1 - \frac{k}{m+1}\right)\xi_{k,m},\ k = 1, 2, \ldots, m \tag{7}$$

and $\tilde{\xi}_{0,m+1} = \xi_{0,m}$ and $\tilde{\xi}_{m+1,m+1} = \xi_{m,m}$.

It means that, theoretically, it is only required a vector of Bernstein coefficients of degree m and then all the other approximations with higher degrees would follow immediately. For this reason, at a first look it may seem dispensable to estimate the vector of coefficients $\boldsymbol{\xi}_{m+r}$, $r = 1, 2, \ldots$ with an entire procedure of estimation. That is, we could simply use Property 1 to instantly obtain this result. However, this behavior of similarity was not observed in practice.

Regarding this property, it is intuitive to expected that estimates of Bernstein coefficients obtained either via a direct way or using degree elevation should be at least similar. That is, $|\xi_{k,m} - \tilde{\xi}_{k,m}| \approx 0$ for all k. However, in practice, we observed that when m is small there can be a significant difference between estimates based on $\boldsymbol{\xi}_{m-1} = (\xi_{1,m-1}, \xi_{2,m-1}, \ldots, \xi_{m,m-1})^\top$ and $\boldsymbol{\xi}_m = (\xi_{1,m}, \xi_{2,m} \ldots, \xi_{m+1,m})^\top$. Consequently, this discrepancy leads to a difference between $\tilde{\boldsymbol{\xi}}_m$ and $\boldsymbol{\xi}_m$. Nevertheless, when m is large enough, the estimated curve stabilizes and the difference between these two vectors of coefficients gets considerably small. Based on this discussion, we came up with two criteria that can help us to provide an optimal degree for the BP. These two criteria are another contribution of this paper.

Criterion 1 - difference between coefficients

Under a Bayesian framework, let $D_{m-1}^{(s)} = \dfrac{1}{m}\displaystyle\sum_{k=1}^{m} |\xi_{k,m-1}^{(s)} - \tilde{\xi}_{k,m-1}^{(s)}|$, for $m \geq 5$ and $s = 1, 2, \ldots, S$. The quantity $D_{m-1}^{(s)}$ represents the difference between coefficients obtained via degree elevation ($\tilde{\xi}_{k,m-1}$) and direct estimation ($\xi_{k,m-1}$) based on the s-th posterior sampled values. If m is small, then it is expected that $\text{Median}(\mathbf{D}_{m-1}) > \text{Median}(\mathbf{D}_m)$, where $\mathbf{D}_m = (D_m^{(1)}, D_m^{(2)}, \ldots, D_m^{(S)})^\top$. However, if m is considered to be large enough there will be no significant difference between these two quantities, meaning that we have reached an optimal degree.

Hence, we will test the hypothesis $H_0 : \text{Median}(\mathbf{D}_{m-1}) = \text{Median}(\mathbf{D}_m)$ vs. $H_1 : \text{Median}(\mathbf{D}_{m-1}) > \text{Median}(\mathbf{D}_m)$ and we will increase m unity by unity until we no longer reject the null hypothesis. Thus, the optimal degree is given by $m_{opt} = \min\{m > 4 : \text{p-value} > 0.1\} + 1$. Here, $m \geq 5$ because our proposed method for the minimum degree is available for $m \geq 4$. Then, we will start the comparison of the direct estimation for $m = 5$ with the degree elevation method based on the posterior sample of a BP with $m = 4$.

Criterion 2 - difference between estimated curves

The estimation for the function of interest, as defined in Equation (2), can be rewritten as $B_{m-1}(t;f) = (\boldsymbol{\xi}_{m-1})^\top \mathbf{b}_{m-1}(t/T_{max})$, where $\mathbf{b}_{m-1}(t/T_{max}) = (b_{1,m-1}(t/T_{max}), b_{2,m-1}(t/T_{max}), \ldots, b_{m,m-1}(t/T_{max}))$ is the vector of Bernstein basis and $\boldsymbol{\xi}_{m-1}$ is the vector of coefficients. The second criterion is based on the distance between estimated curves based on both $\boldsymbol{\xi}_{m-1}$ and $\tilde{\boldsymbol{\xi}}_{m-1}$. This distance will be defined as

$$
\begin{aligned}
D_{m-1} &= \int_0^1 (B_{m-1}(t;f) - \tilde{B}_{m-1}(t;f))^2 d(t/T_{max}) \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m}(\xi_{k,m-1} - \tilde{\xi}_{k,m-1})(\xi_{l,m-1} - \tilde{\xi}_{l,m-1}) \\
&\quad \int_0^1 \left[b_{k,m-1}\left(\frac{t}{T_{max}}\right) b_{l,m-1}\left(\frac{t}{T_{max}}\right)\right] d(t/T_{max}) \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m}(\xi_{k,m-1} - \tilde{\xi}_{k,m-1})(\xi_{l,m-1} - \tilde{\xi}_{l,m-1})\binom{m-1}{k-1}\binom{m-1}{l-1} \\
&\quad \frac{\Gamma(k+l-1)\Gamma(2m-k-l+1)}{\Gamma(2m)} \\
&= (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1})^\top \mathrm{A} (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1}),
\end{aligned} \tag{8}
$$

where A is an $m \times m$ matrix such that $\mathrm{A} = [a_{kl}]$, for $k = 1, 2, \ldots, m$, $l = 1, 2, \ldots, m$ with $a_{kl} = \dfrac{1}{(2m-1)} HyperG(k-1; 2m-2, k+l-2, m-1)$. At last, $HyperG(k-1; 2m-2, k+l-2, m-1)$ is the probability of $k-1$ successes from an *Hypergeometric* distribution with parameters $(2m-2, k+l-2, m-1)$. See this result with details in the Appendix, Section A.1.2.

Here, we are interested in testing $H_0 : \text{Median}(\mathbf{D}_{m-1}) = \text{Median}(\mathbf{D}_m)$ vs. $H_1 : \text{Median}(\mathbf{D}_{m-1}) > \text{Median}(\mathbf{D}_m)$. Similarly to the first criterion, to reject the null hypothesis means that m is still low and, as a result, a better estimated curve can be obtained by increasing the degree by a unity. The optimal value of m is $m_{opt} = \min\{m > 4 : \text{p-value} > 0.1\} + 1$.

The difference between the proposed criteria is that the first one considers only BP coefficients, which evaluate the approximated curve difference in m points. The second one takes into account the entire approximation difference. In order to test these hypothesis, we can use Sign Test and/or Wilcoxon Rank Test (11). Next section focuses on the simulation study.

3 Simulation Study

The main goal of this study was to verify the performance of the proposed degree selection methods. We aimed at investigating if the theoretical results that we derived could be put into practice.

Suppose that the true curve of a certain phenomena evolves according to a function $f(t) = 10+10\sin(2\pi t)$, $t \in (0,1)$. We generated data in a longitudinal configuration merely for convenience. It is important to mention that it does not favor our degree selection method. We can use the proposed method in any curve approximation framework.

Here, the random variable referring to the observed values is $Y_i(t_{ij})$ for subjects $i = 1, 2, \ldots, n$ and measurements $j = 1, 2, \ldots, J_i$. Each observation is equal to the "true" and unobserved value $f(t_{ij})$ plus a measurement error:

$$Y_i(t_{ij}) = W_i(t_{ij}) + \epsilon(t_{ij}) = f(t_{ij}) + \epsilon(t_{ij}), \tag{9}$$

where, $W_i(t_{ij})$ represents the true value of the longitudinal variable. This true value behaves according to the function f. Next, $\epsilon(t_{ij})$ is an independent and identically normally distributed random error with mean 0 and variance σ^2_ϵ.

We generated 100 data sets with sample size $n = 50$ each. The number of measurements were uniformly distributed within the set $\{3, 4, \ldots, 10\}$. The first measurement time was always 0 (i.e., $t_{i1} = 0 \ \forall \ i$) and the measurement times for $i = 1, 2, \ldots, n$ and $j = 2, 3, \ldots, J_i$ were sampled from a $Beta(1, 3)$ distribution. This distribution has high density at the beginning of the interval $(0, 1)$, representing the idea that there are more measurements at initial times and fewer ones at final times. Thus, $\forall \ i$ we sampled $\mathbf{W}_i = (W_i(t_{i1}), W_i(t_{i2}), \ldots, W_i(t_{iJ_i}))^\top \sim Normal_{J_I}(\mathbf{f}_i, \Sigma_{J_i})$, where $\mathbf{f}_i = (f(t_{i1}), f_i(t_{i2}), \ldots, f_i(t_{iJ_i}))^\top$, $\Sigma_{J_i} = [\sigma_{jj'}]$, $\sigma_{jj'} = 6$ if $j \neq j'$ and $\sigma_{jj'} = 3^2$ if $j = j'$. At last, we used Equation (9) to obtain the observed values $\mathbf{y}_i = (y_i(t_{i1}), y_i(t_{i2}), \ldots, y_i(t_{iJ_i}))$, $\forall \ i$, with $\sigma_\epsilon = 1.5$.

Evaluating the true curve $f(t)$, we can verify that it changes behavior in two points, at $t = 1/4$ and $t = 3/4$. Therefore, considering that this information is known by an expert, the next step is to translate this knowledge to Beta distributions with high densities in intervals including these two points (see step by step in page 7). Then, we chose $Beta(11, 34)$, $Beta(117, 351)$ and $Beta(1172, 3515)$. The expected values for these distributions are 0.2444, 0.2500 and 0.2501 and the variances are 0.0040, 0.0004 and 0.00004.

In this example both turning points are equally spaced in relation to the boundaries. So, in this specific case, we can consider only one of them, because the indicated minimum value will be exactly the same. We chose the first one, $t = 1/4$. In a case where the points in which a change of behavior is expected are not equally spaced, one can apply the step by step for each one of them. Then, the final minimum degree would be the maximum of the indicated values. Another strategy would be to choose the point that is closer to the boundaries.

Table 2 shows the probability function of the random variable M given the three chosen Beta distributions for U_1 and U_2. We can see that the highest probability for all three cases is when $m = 6$. In addition, $\mathbb{P}(M \geq 6|(U_1, U_2)) \geq 0.5$ in all examples. Thus, a suitable minimum value for m in this example is 6. We will use the proposed methods to point out the optimal degree for the BP that best approximates $f(t)$.

Table 2: Probability function of M given different distributions for U_1 and U_2.

Distribution of	m											
U_1 and U_2	4	5	6	7	8	9	10	11	12	13	14	...
$Beta(11, 34)$	0.0078	0.1878	0.3612	0.2463	0.1148	0.0477	0.0195	0.0082	0.0035	0.0016	0.0008	...
$Beta(117, 351)$	0.0000	0.2430	0.7480	0.0090	0.0000	0.0000	0.000	0.0000	0.0000	0.0000	0.0000	...
$Beta(1172, 3515)$	0.0000	0.2511	0.7489	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	...

For each data set, we fitted the model

$$\begin{aligned} Y_i(t_{ij}) &= W_i(t_{ij}) + \epsilon_i(t_{ij}),\ i = 1, 2, \dots n,\ j = 1, 2, \dots, J_i \\ &= \sum_{l=1}^{m} f_i\left(\frac{l-1}{m-1}T_{max}\right)\binom{m-1}{l-1}\left(\frac{t_{ij}}{T_{max}}\right)^{l-1}\left(1-\frac{t_{ij}}{T_{max}}\right)^{m-l} + \epsilon_i(t_{ij}) \\ &= (\boldsymbol{\xi}_{i,m-1})^\top \mathbf{b}_{m-1}\left(\frac{t_{ij}}{T_{max}}\right) + \epsilon_i(t_{ij}), \end{aligned} \tag{10}$$

where $Y_i(t_{ij})$ is a random variable referring to the observed value of the longitudinal variable for the i-th subject and the j-th measurement time, $W_i(t_{ij})$ represents the true value of this variable, and $\epsilon_i(t_{ij})$ is the measurement error. Here, $\epsilon_i(t_{ij}) \sim Normal(0, \sigma^2_\epsilon)$. Suppose that the temporal evolution of the longitudinal variable is in accordance with a function f_i and that it will be modeled by a BP with degree $m-1$. Thus, $\boldsymbol{\xi}_{i,m-1} = (\xi_{i,1,m-1}, \xi_{i,2,m-1}, \dots, \xi_{i,m,m-1})^\top$ is the vector composed of m Bernstein coefficients, at subject-level, approximating the values of the function $f_i(.)$ for each subject i at time points $[(l-1)/(m-1)]T_{max}$, $l = 1, 2, \dots, m$. In addition, we considered that $\boldsymbol{\xi}_{i,m-1} \sim Normal_m(\boldsymbol{\mu}_\xi, \Sigma_\xi)$. Here, $\boldsymbol{\mu}_\xi$ represents the vector of unknown means and Σ_ξ is the $(m \times m)$ variance-covariance matrix. At last, $\mathbf{b}_{m-1}\left(\frac{t_{ij}}{T_{max}}\right) = \left(b_{1,m-1}\left(\frac{t_{ij}}{T_{max}}\right), b_{2,m-1}\left(\frac{t_{ij}}{T_{max}}\right), \dots, b_{m,m-1}\left(\frac{t_{ij}}{T_{max}}\right)\right)^\top$ is the vector of m Bernstein basis that will weight the information from the vector of coefficients. More details about how to use the BP on the longitudinal data context can be seen in (22).

We considered the orders of the BP within the set $\{4, 5, \dots, 16\}$. For the MCMC specification, we set a burn-in of 50,000, a lag of 20 and we saved 5,000 posterior values. The prior distributions were weakly informative: $\boldsymbol{\mu}_\xi \sim Normal_m(\mathbf{0}_m, (10)^2\mathbb{I}_m)$, $\Sigma_\xi^{-1} \sim Wishart(m+2, (1/m)\mathbb{I}_m)$, and $1/\sigma^2_\epsilon \sim Gamma(0.01, 0.01)$. Here, $\mathbf{0}_m$ represents a vector of length m in which all components are equal to 0, and $\mathbb{I}_m$ stands for an $m \times m$ identity matrix.

Our aims were (i) to compare the best degree for the BP suggested by both proposed criteria as well as by regular comparison measures (DIC, LPML, and WAIC); and (ii) given the optimal degree, to indicate when changes in the target curve occur. Moreover, since we know the true curve function $f(t)$, we

can use both proposed methods to obtain the optimal degree by comparing the estimates of $\boldsymbol{\xi}_{m-1} = (\xi_{1,m-1}, \xi_{2,m-1}, \ldots, \xi_{m,m-1})^\top$ and the true values that this vector of parameters represents, which are $f((l-1)/(m-1))$, for $l = 1, 2, \ldots, m$. It is worth emphasizing that this is not the main goal of our degree selection method. The comparison with the estimated vector of coefficients with its true value is merely for further understanding the role of the coefficients and the impact on the final results.

Table 3 shows results regarding the optimal degree for the BP. The contents in this table are the frequencies in which each degree was selected as the best one. This selection varied according to the criterion (1 or 2) and the non-parametric tests (Sign or Wilcoxon), as well as to the regular comparison measures. The left side of this table is based on Criterion 1, i.e., difference between coefficients. We compared the difference between *estimated BP coefficients versus BP coefficients obtained via degree elevation* as well as the difference between *estimated BP coefficients versus true BP coefficients*. In all these cases, both Sign and Wilcoxon tests indicated that $m_{opt} = 10$.

Table 3: Optimal degree for BP (m_{opt}) based on proposed criteria.

	Difference between coefficients				Difference between mean curves				DIC	LPML	WAIC
	degree elevation		true		degree elevation		true		-	-	-
m	Sign	Wilcoxon	Sign	Wilcoxon	Sign	Wilcoxon	Sign	Wilcoxon			
5	0	0	0	0	0	0	0	0	**19**	1	1
6	0	0	1	0	**56**	**52**	0	0	11	17	18
7	0	0	0	0	15	12	5	5	18	**29**	**26**
8	13	9	8	8	10	7	6	5	9	15	19
9	20	19	12	10	4	6	10	9	6	12	10
10	**45**	**44**	**35**	**37**	10	13	10	9	17	7	7
11	20	23	29	26	3	2	10	9	6	7	7
12	2	5	11	15	2	7	**18**	**18**	3	2	1
13	0	0	4	3	0	1	7	9	4	4	5
14	0	0	0	1	0	0	15	**18**	3	5	5
15	0	0	0	0	0	0	5	6	1	1	1
16	0	0	0	0	0	0	5	4	3	0	0
≥ 17	0	0	0	0	0	0	9	8	-	-	-

In turn, the second part of Table 3 shows results for the same comparisons described above, but based on the differences between the entire mean curves (Criterion 2). Thus, we have differences between *direct estimated mean curve versus mean curve via degree elevation* as well as the difference between *direct estimated mean curve versus true mean curve*. According to Criterion 2, the best degree is $m_{opt} = 6$. Nonetheless, when we compare the estimated results to the true mean curve, there was not a clear difference between $m = 12$ and $m = 14$. The fact that the results obtained based on estimation only and the ones that compare estimation and the true mean curve do not match, is not

completely unexpected. This seems to occur because while criterion 1 compares solely the coefficients, criterion 2 takes into account the basis information as well. Furthermore, in the real data scenario we do not know the true curve. For this reason we will choose the best degree for this criterion, $m_{opt} = 6$.

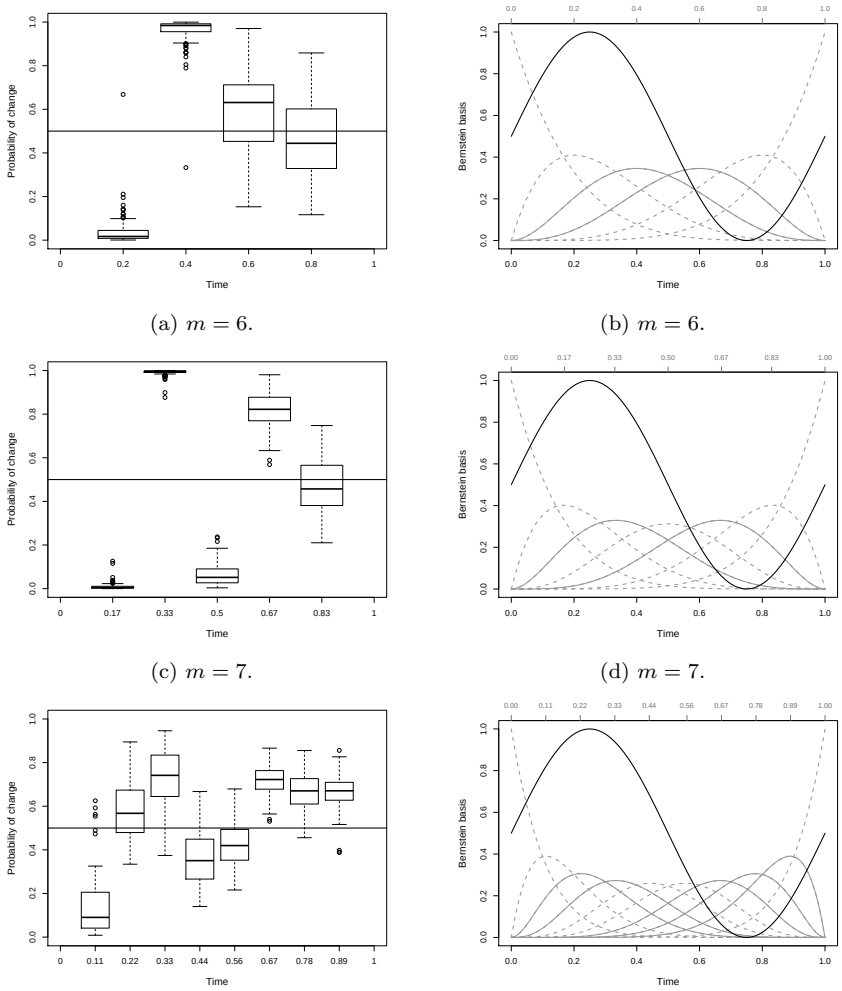
At last, according to DIC, the best model is the one considering $m = 5$. On the other hand, both LPML and WAIC selected $m = 7$ as the best degree to approximate the target function.

Based on the previous discussion, we evaluated the probability of observing a change in the behavior of the approximated mean curve for $m = 6$ (due to Criterion 2), 7 (according to LPML and WAIC) and 10 (pointed out by Criterion 1), see Figure 2. On the left panel of this figure we can observe boxplots based on 100 posterior probabilities of a change in the mean curve. These probabilities are defined as $\mathbb{P}\left((\mu_{\xi_l,m-1} - \mu_{\xi_{l-1},m-1})(\mu_{\xi_{l+1},m-1} - \mu_{\xi_l,m-1}) < 0|\text{Data}\right)$, for $l = 2, 3, \ldots, m-2$. Each value of the boxplots refers to the probability of a change of behavior in the approximated function in one of the $m-2$ time points available. The black horizontal line in these figures represent a cutting value of 0.5.

The right panel in Figure 2 shows each of the m Bernstein basis. The black straight line is the true curve f after it was reescaled to the $(0, 1)$ interval. We reescaled this function so that it could fit in the figures of the basis and, therefore, facilitate the comparisons. The gray straight lines represent the bases related to the coefficients that, in turn, represent at each time point the mean curve is expected to change its behavior. We determined that it is *likely* that the approximated function has a turning point if the median of the probabilities for the data sets was equal or above 0.5.

Consider $m = 6$ (Figures 2a and 2b); this value was selected as the optimal degree according to Criterion 2. We can see that the posterior probabilities of having a change point around $t = 0.2$ are all below 0.5. That is, results show that is *extremely unlikely* that a change of course (increasing/decreasing) starts in $t = 0$, happens around $t = 0.2$ and ends until $t = 0.4$. The second boxplot is entirely above the straight line that marks the value of 0.5; thus, it is *extremely likely* that a change in the course of the approximated function will occur around $t = 0.4$. In turn, the median of the third boxplot is above 0.5, therefore we can consider that there will be another change of course around $t = 0.6$ and, at last, the fourth boxplot indicates that no change in the course of the target function is likely to occur in $t = 0.8$. Figure 2b shows all $m = 6$ Bernstein basis and the *straight* gray lines are those related to the coefficients that detected a change in the mean curve. We can see that the BP detects the changes of courses even with a small degree. However, this degree was too low to allow the estimation to detect this change at an accurate time point.

The same analysis was done considering $m = 7$. We can see in Figure 2c that the results indicate that changes are likely to occur in the intervals $t \in (0.17, 0.5)$ and $t \in (0.5, 0.83)$. The highest probabilities are on $t = 0.33$ and $t = 0.67$. In Figure 2d we can see the basis related to the changes in the mean curve.



(a) $m = 6$. (b) $m = 6$.

(c) $m = 7$. (d) $m = 7$.

(e) $m = 10$. (f) $m = 10$.

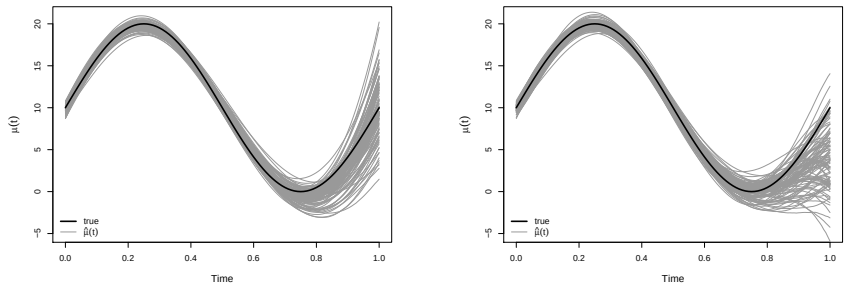
Fig. 2: Posterior probability of a change on the behavior of the mean curve (left panel) and basis related to these changes (right panel), for $m = 6, 7$ and 10.

In Figure 2e we can see results when we consider $m = 10$. These results indicate that a change with respect to the increasing/decreasing behavior will occur around $t \in (0.11, 0.44)$ with higher probability centered on $t = 0.33$. Another change of behavior is detected on $t \in (0.56, 1)$, being most likely to happen around $t = 0.67$. An interesting aspect to point out is that, since there were fewer observations at final times, there is more uncertainty in these estimates. The implication of this characteristic is that it leads to a wider interval pointing to where the change of course might occur. Figure 2f shows

the basis functions that are related to the changes. We can see that there are two basis related to the first change and three basis related to the second one. We have more basis related to the second change precisely due to less data at this part.

Turning our attention to the results of the regular comparison measures, it seems that they somewhat agreed with Criterion 2 and pointed out to simpler models. Nonetheless, on the contrary of DIC, LPML, and WAIC, our proposed method is specific for the BP approximation. Thus, in a sense, it can lead to a more appropriate choice of the degree.

Figure 3 shows the median of the estimated posterior mean curves for each of the 100 data sets (i.e., median of estimated mean function based on all 5000 posterior values) and $m = \{6, 10\}$. The true mean curve is the straight black line. In a general aspect, performance based on both degrees pointed out by our methods presented similar results. Then, we must consider the trade-off between a higher degree and the concept of parsimony. A higher degree leads to a more accurate indication of time when the change of behavior happens, due to the fact that the proposed method detects these changes around a time t such that $t = (l-1)/(m-1)$. However, regarding the overall mean curve the lower degree ($m = 6$) is enough to provide good estimates. In addition, the variability of estimates at the end of follow up is larger due to fewer data.



(a) $m = 6$. (b) $m = 10$.

Fig. 3: Overall median of the posterior mean curves along with the true curve, for $m = \{6, 10\}$.

This simulation study indicates that the two methods we propose for the degree selection presented adequate results. Comparing the outcomes of our degree selection criteria, both BP with degrees $m = 6$ and $m = 10$ were able to approximate well the target function. On one hand, estimates based on the lower degree presented lower variability even at the final times - this region is where we had less data. On the other hand, the greater is the degree of the BP, the more accurate it will point out the time points in which we have a change of behavior. That is, when $m = 6$ the possible times points evaluated were $t \in \{0.20, 0.40, 0.60, 0.80\}$; in turn, if $m = 10$, this approximation has the set $t \in \{0.11, 0.22, \ldots, 0.78, 0.89\}$ to point out the the change of course in the

target function. Thus, in the cases where the methods we proposed are not in accordance with each other, one should prioritize or (i) the final result of the estimation of the target function; or (ii) the will of being able to tell when the approximate curve will change behavior. In the first case, one should choose the lower indicated degree; on the other hand, if the main goal is (ii) then the choice of the optimal degree should be the maximum value indicated by our proposed criteria. Next section shows a succinct analysis of a real data set.

4 Berkeley Growth Study

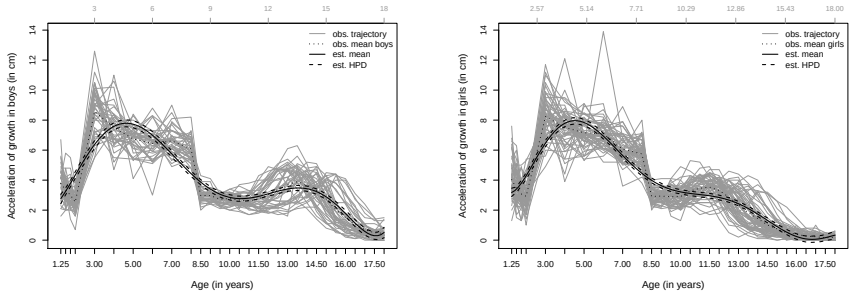
We will briefly analyze data from a Berkeley Growth Study (18; 19) as a form of illustrating our probabilistic method for the minimum degree, as well as both propositions to reach the optimal degree of the BP. The mentioned data set - available in the `fda R` package - is composed of 39 boys and 54 girls that were accompanied longitudinally along eighteen years. Along these years each subject had her/his height evaluated in 31 pre-specified and unequally spaced ages. It is well-known that, in regular conditions, height presents an increasing behavior for infants. For this reason, this curve do not present ups and downs in the mean curve. Then, a form of dealing with this data is to consider the response as the rate of growth. This rate is merely the value, in cm, of growth for each new measurement. So, after treating and transforming this data, we had 30 measurements for each infant.

The first step is to find out the minimum degree that enable us to accommodate important features of the target function. Using basic and non-expert knowledge of growth in infants, we expected that the mean curve for boys had an increasing behavior until the age of 13 or 14 years old. For girls, we assumed that growth will decelerate around 11 or 12 years old. Based on these assumptions, we followed the step by step described on Section 2. We found that the minimum degree necessary so that the estimated mean curve would include our assumptions were 5 girls and 6 for boys. In order to follow our proposed routine to find the minimum m, we used Beta(20, 7) for the mean curve of the boys and Beta(5, 3) for the girls. It resulted in $\mathbb{P}(M \leq 6|(U_1, U_2)) = 0.5586$ for boys and $\mathbb{P}(M \leq 5|(U_1, U_2)) = 0.5528$ for girls.

Then, we fit the model described by Equation (10). However, for this application we considered that boys and girls could have different mean curves. So, the vector of random effects had different distributions, they were: $\boldsymbol{\xi}_{i,m} \sim Normal(\boldsymbol{\mu}_m^{(b)}, \Sigma_m)$ for boys and $\boldsymbol{\xi}_{i,m} \sim Normal(\boldsymbol{\mu}_m^{(g)}, \Sigma_m)$ for girls. We fit the model with $m \in \{5, 6, \ldots, 30\}$, with the same degree for each mean curve and $T_{max} = 18$. Proceeding with the analysis, we used our degree selection method taking to account that it was possible that there could be a different optimal m for boys and girls. We also evaluated the same regular comparison measures aforementioned. It is noteworthy that these latter measures will point out to the best fitted model with the mean curves having the same degree. In contrast to that, with our degree selection method, we can verify the optimal degree using the posterior sample of $\boldsymbol{\mu}_m^{(b)}$ and $\boldsymbol{\mu}_m^{(g)}$ separately.

Regarding the estimated mean curve of the rate of growth in boys, both criteria and both tests used pointed out to $m_{opt} = 7$. In turn, considering the mean curve for the girls, we found that $m_{opt} = 6$ according to criterion 1 and both tests; nonetheless, according to criterion 2 and both tests evaluated, $m_{opt} = 8$. Since in this application we are indeed interested in a more accurate information about the interval of when the rate of growth will change its course, it is more convenient to proceed the analysis with $m_{opt} = 8$ for girls. In addition, DIC selected the simplest model (in accordance with the simulation study), which considers $m = 6$, as the best fitted model. On the other hand, both LPML and WAIC pointed out to $m = 30$. An advantage of our method in this example is possibility of verifying, in a simple manner, different optimal degrees for the mean curves.

Figure 4 shows results for the estimated mean curve for both boys (Panel 4a) and girls (Panel 4b) based on their optimal degrees ($m_{opt} = 7$ and $m_{opt} = 8$, respectively). The gray lines represent the observed trajectories of each infant. In turn, the black dotted line is the observed mean curve. The median of the mean curves estimated by a BP with the optimal degree selected by our method is the black solid line; the HPD interval is the dashed black line.



(a) Boys. (b) Girls.

Fig. 4: Observed trajectory of boys and girls, observed mean curves, estimated mean curves and HPD intervals.

Observing this figure we can see that the optimal degrees selected by our method seems to present a good performance as it fits the main characteristics of the rate of growth. As discussed before, with a higher degree we are able to encompass even more characteristics of the target function. We would also have more accurate estimates of the ages in which the rate of height increases and/or decreases. However, according to our method and taking into account the concept of parsimony, $m = 6$ for boys and $m = 8$ for girls give enough accuracy. An extra comment is that we can verify that the rate of growth is indeed different for boys and girls.

Given the approximation of the mean curve using BP we have the estimated probability of change of behavior, see Table 4. For the boys these probabilities are very high for all age points available. It means that it is very likely that

the mean curve will change the behavior in the ages of 3, 6, 9, 12 and 15 years old (yrs). If we look back at Panel 4a, we can see that we have a peak at the age of 3 yrs. Then, around 6 yrs, growth decelerates and starts to increase again until 9 yrs. Next, when they have around 12 yrs, the behavior of rate of growth increases one last time, until approximately 15 yrs.

Table 4: Posterior probability of turning points in the overall mean curve. Results according to the optimal degrees for boys and girls separately.

Mean curve for boys ($m_{opt} = 7$).							
Age	0.00	3.00	6.00	9.00	12.00	15.00	18.00
p	-	0.9540	0.9974	0.9996	0.9996	0.9942	-

Mean curve for girls ($m_{opt} = 8$).								
Age	0.00	2.57	5.14	7.71	10.29	12.86	15.43	18.00
p	-	0.9920	0.9992	0.9970	0.9908	0.8876	0.2724	-

In turn, via Table 4 and Panel 4b, we can see that the rate of growth in girls presents a similar behavior to that of the boys until the age of 11 yrs, approximately. It is worth highlighting that, since we are presenting estimations with a higher degree, we will automatically have more accurate information about these changes. Thus, it is very likely that the growth will have an increasing behavior until the age of 2.57 yrs. Next, we can observe a small (but very likely to occur) deceleration until the age of 5.14 yrs, that turns into an increasing behavior until girls reach the age of 7.71 ft. After that, we can expect that the rate of growth will decrease until 10.29 yrs. From now on the difference of growth between boys and girls becomes clearer. After the age of approximately 11 yrs, the rate in girls starts to increase again until they are 12.86 yrs. After this age, the rate to growth reaches to a final stage of decreasing.

At last, it is worth highlighting that the estimated median mean curve is composed of a linear combination of the vector of coefficients (that leads to the probabilities evaluated above) and the vector of Bernstein basis. The next section brings an overall discussion of the methods we have proposed in this paper.

5 Discussion

Our main proposal in this paper was to elaborate methods that could serve as a complete guidance on the choice of the degree when using the BP to approximate unknown functions. This guidance is composed of a probabilistic based method to point out a minimum degree, along with two approaches that indicate the optimal one. The main motivation for the necessity of a guidance is that, theoretically, we reach convergence when $m \longrightarrow +\infty$. However, in

practice, a finite value must be considered. Moreover, it is well reported that this choice plays an important role on the estimation via BP.

The role of having a distribution of the minimum degree, i. e. the random variable M, is to give a starting point. In case this starting value is too large for a specific application, it informs that some of the important characteristics may not be well represented in the approximation. In other words, with the distribution of M we can also anticipate possible limitations of the approximation via BP. In turn, both methods that indicate the optimal degree suggest that a higher degree does not bring relevant information in the estimation. Thus, invoking the principle of parsimony, a higher degree would be superfluous.

We conducted a simulation study to verify the capacity of the proposed degree selection methods. According to the results, they have a satisfactory performance. Besides that, our proposal is a robust method, since its usage only needs an indication of a turning point in the function being approximated. Nonetheless, it is worth reinforcing that it can still be used even if such information is not available.

At last, we mention that several works have fixed the degree of the BP based on sample sizes. However, it seems that this degree is more associated to the point in which the target function changes its trajectory, than it is related to sample sizes. In addition to that, our methods are entirely based on probabilities and properties of the BP itself. For future works, we intend on building a package in order to help users to put our methods into practice.

Acknowledgements The first author thanks the Brazilian agency Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for the funding in the form of two scholarships: as a doctorate student and as a student in an exchange program.

Conflict of interest

The authors declare that they have no conflict of interest.

A Appendix

A.1 Details of calculations

This section shows details of the main results described in this paper. The calculations are separated in subsections. The order of these subsections are in accordance with the organization of the paper.

A.1.1 Probability function of M

The random variable M represents the minimum degree that is needed to capture a change in an interval $(U_{(1)}, U_{(2)})$. This variable is defined as $M = \left\lceil \max\left(\frac{1}{U_{(1)}} + 1, \frac{2 - U_{(2)}}{1 - U_{(2)}}\right)\right\rceil$. So, we have that

$$\begin{aligned}\mathbb{P}(M=m|(U_1,U_2)) &= \mathbb{P}\left(\left\lceil \max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\right\rceil = m|(U_1,U_2)\right)\\ &= \mathbb{P}\left(m-1<\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m|(U_1,U_2)\right)\\ &= \mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m|(U_1,U_2)\right)-\\ &\quad \mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m-1|(U_1,U_2)\right).\end{aligned}$$

Now, note that

$$\begin{aligned}\left[\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m\right] &= \left[\frac{1}{U_{(1)}}+1\leq m,\frac{2-U_{(2)}}{1-U_{(2)}}\leq m\right]\\ &= \left[U_{(1)}\geq\frac{1}{m-1},U_{(2)}\leq\frac{m-2}{m-1}\right]\\ &= \left[\frac{1}{m-1}\leq U_1\leq\frac{m-2}{m-1},\frac{1}{m-1}\leq U_2\leq\frac{m-2}{m-1}\right].\end{aligned}$$

Consequently,

$$\left[\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m-1\right] = \left[\frac{1}{m-2}\leq U_1\leq\frac{m-3}{m-2},\frac{1}{m-2}\leq U_2\leq\frac{m-3}{m-2}\right].$$

In the case that U_1 and U_2 are equally distributed, we have that

$$\begin{aligned}\mathbb{P}(M=m|(U_1,U_2)) &= \mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m|(U_1,U_2)\right)-\\ &\quad \mathbb{P}\left(\max\left(\frac{1}{U_{(1)}}+1,\frac{2-U_{(2)}}{1-U_{(2)}}\right)\leq m-1|(U_1,U_2)\right)\\ &= \left[\mathbb{P}\left(U_1\leq\frac{m-2}{m-1}\right)-\mathbb{P}\left(U_1<\frac{1}{m-1}\right)\right]^2-\\ &\quad \left[\mathbb{P}\left(U_1\leq\frac{m-3}{m-2}\right)-\mathbb{P}\left(U_1<\frac{1}{m-2}\right)\right]^2.\end{aligned}$$

At last, considering $U_1\sim Beta(\theta_1,\theta_2)$ and $U_2\sim Beta(\theta_1,\theta_2)$

$$\begin{aligned}\mathbb{P}(M=m|(U_1,U_2)) &= \left[F_{Beta}\left(\frac{m-2}{m-1};\theta_1,\theta_2\right)-F_{Beta}\left(\frac{1}{m-1};\theta_1,\theta_2\right)\right]^2-\\ &\quad \left[F_{Beta}\left(\frac{m-3}{m-2};\theta_1,\theta_2\right)-F_{Beta}\left(\frac{1}{m-2};\theta_1,\theta_2\right)\right]^2.\end{aligned}$$

A.1.2 Difference between two estimated curves

The difference between two approximated functions - one estimating the vector of coefficients directly ($B_{m-1}(t;f)$)) and the other one obtained by using the degree elevation property ($\tilde{B}_{m-1}(t;f)$), is given by

$$
\begin{aligned}
D_{m-1} &= \int_0^1 (B_{m-1}(t;f)) - \tilde{B}_{m-1}(t;f))^2 d(t/T_{max}) \\
&= \int_0^1 \left((\boldsymbol{\xi}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right) - (\tilde{\boldsymbol{\xi}}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right)\right)^2 d(t/T_{max}) \\
&= \int_0^1 \left[(\boldsymbol{\xi}_{m-1}^\top - \tilde{\boldsymbol{\xi}}_{m-1}^\top)\mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right)\right]^2 d(t/T_{max}) \\
&= \int_0^1 \left[(\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1})^\top \mathbf{b}_{m-1}\left(\frac{t}{T_{max}}\right)\right]^2 d(t/T_{max}) \\
&= \int_0^1 \left[\sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_{k,m-1} - \tilde{\xi}_{k,m-1})(\xi_{l,m-1} - \tilde{\xi}_{l,m-1}) b_{k,m-1}\left(\frac{t}{T_{max}}\right) b_{l,m-1}\left(\frac{t}{T_{max}}\right)\right] d(t/T_{max}) \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_{k,m-1} - \tilde{\xi}_{k,m-1})(\xi_{l,m-1} - \tilde{\xi}_{l,m-1}) \int_0^1 \left[b_{k,m-1}\left(\frac{t}{T_{max}}\right) b_{l,m-1}\left(\frac{t}{T_{max}}\right)\right] d(t/T_{max}) \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_{k,m-1} - \tilde{\xi}_{k,m-1})(\xi_{l,m-1} - \tilde{\xi}_{l,m-1}) \binom{m-1}{k-1}\binom{m-1}{l-1} \frac{\Gamma(k+l-1)\Gamma(2m-k-l+1)}{\Gamma(2m)} \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_{k,m-1} - \tilde{\xi}_{k,m-1})(\xi_{l,m-1} - \tilde{\xi}_{l,m-1}) \frac{1}{(2m-1)} \frac{\binom{k+l-2}{k-1}\binom{2m-k-l}{m-k}}{\binom{2m-2}{m-1}} \\
&= \sum_{k=1}^{m}\sum_{l=1}^{m} (\xi_{k,m-1} - \tilde{\xi}_{k,m-1})(\xi_{l,m-1} - \tilde{\xi}_{l,m-1}) \frac{1}{(2m-1)} HyperG(k-1; 2m-2, k+l-2, m-1) \\
&= (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1})^\top \mathrm{A} (\boldsymbol{\xi}_{m-1} - \tilde{\boldsymbol{\xi}}_{m-1}),
\end{aligned}
$$

where A is an $m \times m$ matrix. Each component a_{kl} of this matrix is given by $a_{kl} = \frac{1}{(2m-1)} HyperG(k-1; 2m-2, k+l-2, m-1)$, for $k = 1, 2, \ldots, m$ and $l = 1, 2, \ldots, m$. At last, $HyperG(k-1; 2m-2, k+l-2, m-1)$ represents the probability of $k-1$ successes from an *Hypergeometric* distribution with parameters $(2m-2, k+l-2, m-1)$.

References

1. Babu, G.J., Canty, A.J., Chaubey, Y.P.: Application of Bernstein polynomials for smooth estimation of a distribution and density function. Journal of Statistical Planning and Inference **105**(2), 377 – 392 (2002)
2. Bernstein, S.N.: Démonstration du théorème de Weiertrass fondée sur le calcul des probabilités. Kharkov Mathematical Society **13**(1-2) (1912)
3. Bertrand, A., Keilegom, I.V., Legrand, C.: Flexible parametric approach to classical measurement error variance estimation without auxiliary data. Biometrics **75**(1), 297–307 (2019)
4. Brown, B.M., Chen, S.X.: Beta-Bernstein smoothing for regression curves with compact support. Scandinavian Journal of Statistics **26**(1), 47–59 (1999)
5. Carnicer, J.M., Peña, J.M.: Shape preserving representations and optimality of the Bernstein basis. Advances in Computational Mathematics **1**(2), 173–196 (1993)
6. Chang, I.S., Chien, L.C., Hsiung, C.A., Wen, C.C., Wu, Y.J.: Shape restricted regression with random Bernstein polynomials. Lecture Notes-Monograph Series **54**, 187–202 (2007)
7. Chang, I.S., Hsiung, C.A., Yuh-Jennwu, Yang, C.C.: Bayesian survival analysis using Bernstein polynomials. Scandinavian Journal of Statistics **32**(3), 447–466 (2005)
8. Chen, Y., Hanson, T., Zhang, J.: Accelerated hazards model based on parametric families generalized with Bernstein polynomials. Biometrics **70**(1), 192–201 (2014)

9. Curtis, S.M., Ghosh, S.K.: A variable selection approach to monotonic regression with Bernstein polynomials. Journal of Applied Statistics **38**(5), 961–976 (2011)
10. Farouki, R.T.: The Bernstein polynomial basis: A centennial retrospective. Computer Aided Geometric Design **29**(6), 379 – 419 (2012)
11. Gibbons, J.D., Chakraborti, S.: Nonparametric Statistical Inference, 4th edn. Statistics: Textbooks & Monograph. Chapman and Hall/CRC (2003)
12. Guan, Z.: Efficient and robust density estimation using Bernstein type polynomials. Journal of Nonparametric Statistics **28**(2), 250–271 (2016)
13. Kottas, A.: Dirichlet process mixtures of Beta distributions, with applications to density and intensity estimation. In: In Proceedings of the Workshop on Learning with Nonparametric Bayesian Methods (2006)
14. Lorentz, G.G.: Bernstein Polynomials, *AMS Chelsea Publishing*, vol. 323. American Mathematical Society (1986)
15. Osman, M., Ghosh, S.K.: Nonparametric regression models for right-censored data using Bernstein polynomials. Computational Statistics & Data Analysis **56**(3), 559 – 573 (2012)
16. Petrone, S.: Bayesian density estimation using Bernstein polynomials. The Canadian Journal of Statistics / La Revue Canadienne de Statistique **27**(1), 105–126 (1999)
17. Petrone, S.: Random Bernstein polynomials. Scandinavian Journal of Statistics **26**(3), 373–393 (1999)
18. Ramsay, J., Silverman, B.W.: Functional Data Analysis, 2 edn. Springer Series in Statistics. Springer-Verlag, New York (2005)
19. Ramsay, J.O., Wickham, H., Graves, S., Hooker, G.: fda: Functional Data Analysis (2017). URL `https://CRAN.R-project.org/package=fda`. R package version 2.4.7
20. Vitale, R.A.: A Bernstein polynomial approach to density function estimation. In: M.L. Puri (ed.) Statistical Inference and Related Topics, pp. 87 – 99. Academic Press (1975)
21. Wang, J., Ghosh, S.K.: Shape restricted nonparametric regression with Bernstein polynomials. Computational Statistics & Data Analysis **56**(9), 2729 – 2741 (2012)
22. Wang, L., Ghosh, S.K.: Nonparametric models for longitudinal data using Bernstein polynomial sieve. Tech. rep., Department of Statistics, North Carolina State University (2013). URL `https://repository.lib.ncsu.edu/bitstream/handle/1840.4/8245/mimeo2651_Wang.pdf?sequence=1&isAllowed=y`
23. Zhou, H., Hanson, T.: A unified framework for fitting Bayesian semiparametric models to arbitrarily censored survival data, including spatially referenced data. Journal of the American Statistical Association **113**(522), 571–581 (2018)
24. Zhou, Q., Hu, T., Sun, J.: A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. Journal of the American Statistical Association **112**(518), 664–672 (2017)