# A new model for Bayesian nonparametric regression

*Un nuovo modello di regressione bayesiana non parametrica*

Rebecca Graziani
Istituto Metodi Quantitativi
Università L.Bocconi
Rebecca.Graziani@uni-bocconi.it

**Riassunto:** Uno dei problemi principali che si affronta nell'analisi bayesiana di un modello di regressione non parametrica è rappresentato dalla difficoltà di assegnare una distribuzione iniziale alla funzione di regressione. In questo lavoro si prende in considerazione solo il caso della regressione non parametrica univariata e si propone di rappresentare la funzione di regressione come una combinazione lineare di polinomi definiti su intervalli di valori della variabile esplicativa. Si suppone altresì di non conoscere né il numero, né l'ordine dei polinomi, nè la posizione dei punti che delimitano gli intervalli in cui i polinomi sono definiti. Assegnando una distribuzione iniziale a tali quantità si riesce ad assegnare una distribuzione inziale all'intera funzione di regressione. Si propone come stimatore della funzione di regressione non nota il valore atteso della sua distribuzione finale, che viene approssimato facendo ricorso ad un algoritmo Reversible jump MCMC.
**Keywords:** Piecewise Polynomials; Splines; Bayesian Model Averaging, Reversible jump Markov chain Monte Carlo method.

## 1. The model

The basic nonparametric univariate regression model has the form

$$y_i = f(x_i) + \varepsilon_i \qquad i = 1, ..., n, \tag{1}$$

where the $y_i's$ are the observations on the response variable $Y$, the $x_i's$ are the observations on the covariate $X$ and $f(x_i)$ is the unknown regression function.

We suggest to model the unknown regression function as a piecewise polynomial, made up of pieces of unknown number, unknown location, unknown order and unknown orientation.

Then, the regression model in (**??**) can be written in the following way:

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j (x_i - \tau_j)_{R_j}^{q_j} + \varepsilon_i i = 1, ..., n \tag{2}$$

that is the regression function is modeled as a linear combination of polynomial splines, where for $j = 1, ..., k$:

- $R_j$ are binary variables, taking values 0, if the spline basis is left orientated and 1, if the spline basis is right orientated:

$$\begin{aligned} R_j &= 0 \Longrightarrow (x_i - \tau_j)_- = \min(0, (x_i - \tau_j)), \\ R_j &= 1 \Longrightarrow (x_i - \tau_j)_+ = \max(0, (x_i - \tau_j)). \end{aligned}$$

- the knot points $\tau_j$ can be located everywhere in the interval defined by the observations on the covariate $x_i$.

Let us denote by $M_k$ the model with $k$ given splines. Given $k$, the structure of the model $M_k$ is specified by the vector $\tau^{(k)}$ of the locations of the splines, by the vector $q^{(k)}$ of their orders and the vector $R^{(k)}$ of their orientations. Let us denote by $\vartheta_k$ the vector formed by $\tau^{(k)}$, $q^{(k)}$ and $R^{(k)}$, ($\vartheta_k = \left(\tau^{(k)}, q^{(k)}, R^{(k)}\right)$). Given $k$ and $\vartheta_k$, the model (**??**) can be viewed as a multivariate linear model with design matrix $\mathbf{X}^{(k)}$, whose columns contain the values of the splines basis for each of the observations on the covariate.

## 2. Priors specification

We suppose that the vector $\varepsilon$ of the error terms $\varepsilon_i$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$ ($\mathbf{I}$ is the identity $n$x$n$ matrix).

We assign a prior distribution to the unknown model parameters $k, \tau^{(k)}, q^{(k)}, R^{(k)}$, $\beta^{(k)}$, $\sigma^2$ in the following conditional way:

$$\begin{aligned} p\left(k, \tau^{(k)}, q^{(k)}, R^{(k)}, \beta^{(k)}, \sigma^2\right) &= p\left(k\right) p\left(\tau^{(k)}|k\right) p\left(q^{(k)}|k\right) p\left(R^{(k)}|k\right) \cdot \\ &\cdot p\left(\beta^{(k)}|k, \tau^{(k)}, q^{(k)}, R^{(k)}, \sigma^2\right) p\left(\sigma^2\right) \end{aligned}$$

where the priors on $\tau^{(k)}, q^{(k)}, R^{(k)}$ are independent conditionally on $k$ .

- We assume that the number of splines $k$ has a discrete uniform distribution on $\{1, 2, ..., n\}$ .
- We suppose that, conditionally on $k$, the locations of the knots $\tau_1, ..., \tau_k$ are *i.i.d* from a uniform distribution on the interval defined by the observations on the covariate $X$.
- We assume that conditionally on $k$ the orders of the splines $q_1, ..., q_k$ are *i.i.d* and we shall consider two kinds of distributions for the single components $q_j$, a Poisson prior and a Uniform prior on $(0, q_{\max})$ for a suitable choice of $q_{\max}$.
- The components of the vector $R^{(k)}$ are assumed, conditionally on $k$, *i.i.d* from a Bernoulli of parameter 1/2:

We choose to perform a complete Bayesian analysis of the model and to assign to the vector of coefficients $\beta^{(k)}$ and to the error variance $\sigma^2$ the conjugate priors.

- Hence, we assign to $\beta^{(k)}$ conditionally on $\left(k, \tau^{(k)}, q^{(k)}, R^{(k)}, \sigma^2\right)$ a multivariate normal distribution with $\mathbf{0}$ mean and covariance matrix $v\sigma^2\mathbf{I}$ ($\mathbf{I}$ being the identity $(k+1)$x$(k+1)$ matrix) and to $\sigma^2$ an Inverse-gamma distribution with parameters $a$ and $d$.
- Finally, we complete the analysis by assigning a prior distribution to the hyperparameter $v$.

## 3. The estimation of the regression function

All inferences about the unknown parameters $k, \tau^{(k)}, q^{(k)}, R^{(k)}, \beta^{(k)}$ and $\sigma^2$ are based on their joint posterior distribution, $p\left(k, \tau^{(k)}, q^{(k)}, R^{(k)}, \beta^{(k)}, \sigma^2 | \mathbf{y}\right)$.

The Bayes estimate of the regression function, assuming a quadratic loss function, is given by its posterior expectation $E\left(f | \mathbf{y}\right)$ which can be written as follows:

$$E\left(f\left(X\right)|\mathbf{y}\right) = \sum_{k} E\left(f\left(x|M_k\right)\right) p\left(M_k|\mathbf{y}\right)$$

where $M_k$ is the model with $k$ splines and parameters $\vartheta_k = \left(\tau^{(k)}, q^{(k)}, R^{(k)}\right)$, $\beta^{(k)}$ and $\sigma^2$ so that $p\left(M_k|\mathbf{y}\right)$ is the posterior distribution $p\left(k, \tau^{(k)}, q^{(k)}, R^{(k)}, \beta^{(k)}, \sigma^2 | \mathbf{y}\right)$. That is the estimate of the regression function is given by a weighted average of the Bayes estimates of the regression function obtained under each model $M_k$ with weights given by their posterior distributions, in accordance with the principle of the Bayesian Model Averaging.

The expression of the posterior expectation is analytically intractable. For this reason we suggest to approximate it by resorting to Markov Chain Monte Carlo simulation. In the following section, we describe in detail the algorithm used for this approximation.

## 3.1 A reversible jump algorithm

In the approximation of $E\left(f\left(X\right)|\mathbf{y}\right)$ we face an additional problem due to the fact that we know neither the parameters of the model, nor the dimension of the model itself. Therefore, following **?** we construct a reversible jump algorithm to approximate the joint posterior distribution of the parameters.

We should design a strategy allowing to move from a model to another in order to explore the whole support of the posterior distribution. The strategy lays on two main elements:

- the identification of the possible movements;
- the computation of the acceptance probability for each move

*The movements*

The exploration of the class of models $\mathcal{M} = \{M_0, ..., M_k, ..\}$ can be achieved by performing the following three move types:

- *Addition of a spline*. This movement implies a change in the dimension of the model, since a new explanatory variable (the new spline) is introduced.
- *Deletion of a spline.*
- *Change of the position of a spline or change in the order of a spline.* Both this move types cause no change in the dimension of the model. Each one is chosen with probability 1/2.

  We assign to each of the above move types the same probability to be chosen, that is 1/3.

*The acceptance probability*

Following Green, the probability $\alpha$ to accept each proposed move is given by

$$\alpha = \min(1, \text{likelihood ratio x prior ratio x proposal ratio})$$

Due to our choice of the proposal probabilities some of the elements of above expression cancel, so that we obtain $\alpha = \min(1, \text{Bayes factor})$

The models proposed at each iteration of the simulation process are evaluated on the basis of the Bayes factor against the current ones. No computational problems arise for the evaluation of their Bayes factor, since the models compared are multivariate normal linear models, with known design matrix. Moreover, the Bayes factor contains a natural penalty against over complicated models, that in our case preserve from overfitting.

## 4. Discussion of the model and a comparison with the previous models

Our work can be considered a generalization of the one due to **?**. In fact, like **?** we choose to model the regression function as a piecewise polynomial, made up of an unknown number of splines located at unknown positions. Again like **?** the problem of the computation of the posterior distributions of the unknown parameters is addressed by using Green's the reversible jumps algorithm.

Unlike as in **?** we do not suppose that the order of the splines is known and fixed and that the splines are all right orientated. We consider the orders of the splines and their orientations unknown as well as their number and locations, giving in this way higher flexibility to our model. The risk inherent in such an high flexibility is to over fit the data.

The model is tested with two simulated data sets, that are two of the test curves used in **?**, the so-called 'Blocks' and 'Heavisine' data sets. As in **?**, we resort to these data sets to test the performance of our methodology in the fit of unsmooth curves. The model works well: in both cases the shape of the true underlying function is entirely caught.

We compare the results found with the ones that can be obtained when we consider splines right orientated and with a fixed order equal to 1 (linear splines) or 2 (quadratic splines) as in **?**. Our model seems to have a better performance, leading to smaller mean squared errors.