

Bayesian piecewise linear regression via product partition models

R. H. Loschi^a, V. D. Mayrink^a, F. R. B. Cruz^{a,1}, and
V. L. Aguiar^b

^a *Department of Statistics, Federal University of Minas Gerais,
31270-901 - Belo Horizonte - MG, Brazil*

^b *Pontifical Catholic University of Minas Gerais,
55588-000 - Arcos - MG, Brazil*

Abstract

In the usual piecewise regression models, we should arbitrate beforehand the number of segments (or blocks) and to define the observations in each segment. In this paper, we propose a Bayesian approach for the segmented regression model that considers as random variables both the number of segments and the instants when the changes occur, thus removing the ad hoc nature of some segmented regression models. Our proposal is based on the product partition model (PPM). We also consider a modification in the usual piecewise regression model, in which the blocks of observations are defined by the most probable partition provided by the Bayesian model. Then, we apply the least squares method to estimate the parameters. The proposed methods are compared for simulated data sets and used in the estimation of the relationship between the Brazilian Industrial Production Index (BIP) and the Brazilian Employment Index (BEI). We conclude that both methods explain reasonably well the relationship between these indexes.

Key words: Change point identification, least squares estimates, product estimates, Yao's cohesions.

1 Introduction

In many studies, the random variables of interest may sometimes be connected by linear relations whose parameters are subject to change in time. In such

¹ Corresponding author. Prof. Frederico Cruz. E-mail fcruz@est.ufmg.br. Phone: (+55 31) 3499 5929. Fax: (+55 31) 3499 5924.

situations, in order to estimate the parameters in the usual way, firstly it is necessary to define the number of changes, then to estimate the positions of the change points. Being the data sequentially observed and assuming that only contiguous blocks are possible, by identifying the position of the changes, we are also identifying the observations within each block. Least squares, maximum likelihood, and Bayes estimators are the procedures usually suggested to obtain the parameters in each block. This kind of problem is very common in the economic settings as we, for instance, notice for the Brazilian Industrial Production Index (BIP) and the Brazilian Employment Index (BEI), from January, 1985, to April, 2001, presented in Figure 1.



Fig. 1. Brazilian Industrial Production Index (BIP) and Employment Index (BEI).

There are many proposals in the literature to analyze piecewise regression models. However, most of these papers refers either to the one-change-point problem (Quant, 1958; McGee & Carleton, 1970; Ferreira, 1975; Holbert, 1982; Hocking, 1996; Bai, 1997; Diniz et al., 2003), or to a case that does not fix previously the number of change points but has to perform sequential tests of hypothesis to define them (Bai & Perron, 1998). In spite of avoiding the *ad hoc* choice of the number of changes in the latter, the type I error cannot be quantified.

Notice that another possible way to treat segmented regression problems that has been successful and well accepted within the Bayesian community is through splines (see Wahba, 1983; Dias, 1999; Holmes & Mallick, 2001, for example). However, for the sake of simplicity and because we sense that the idea of bringing the PPM to piecewise regression with unknown number of segments might be a useful tool and a basis for future research on more elaborated approaches, we shall not deal with splines in here.

In this paper, we propose a full Bayesian approach for segmented regression model with multiple segments that considers as random variables both the number of segments and the instants when the changes occur. Our proposal is based on the product partition model (PPM) introduced by Barry & Hartigan (1993). Notice that the approach introduced here applies concepts and ideas

developed by the authors in previous papers (Loschi & Cruz, 2002, 2005). Thus, one of the main contributions of this paper is the inclusion of covariates in the mean of the normal distribution leading to regression models.

Additionally, by means of an implementation developed in C++, available from the authors upon request, the proposed Bayesian method is compared with the usual piecewise regression model, estimated by the least squares method. We also modify the usual piecewise regression model by considering as blocks of observations the most probable partition provided by the Bayesian model. In order to evaluate the performances, we apply both methods to simulated data sets, and to illustrate the methods, we estimate the relationship between the Brazilian Industrial Production Index (BIP) and the Brazilian Employment Index (BEI). From Figure 1 it is noticeable that a simple regression model would not be appropriate to model these two indices. A possible way is by linear regression models that obey different regimes.

The paper is organized as follows. In Section 2 we introduce a full Bayesian approach for analyzing piecewise linear regression models and present the algorithm for change point detection. Section 3 presents a comparison between the proposed model and the least squares method for piecewise regression model in simulated data sets. In Section 4, the proposed methods are applied to analyze the Brazilian Industrial Production Index (BIP) and the Brazilian Employment Index (BEI) data sets. Finally, in Section 5, some final conclusions and remarks close the paper.

2 Statistical model

In this section we present a full Bayesian approach for cluster analysis in linear regression model. We also adapt the algorithm proposed by Loschi & Cruz (2005), for cluster detection, to the piecewise linear regression model.

2.1 Cluster for linear regression model

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a sequence of observations. Assume that, for all $k = 1, \dots, n$, the vector (X_k, Y_k) obeys the usual regression model specification

$$Y_k = \alpha_k + \psi_k X_k + e_k, \quad (1)$$

in which $\alpha_k \in \mathbb{R}$ denotes the intercept and $\psi_k \in \mathbb{R}$ denotes the slope at instant k . Assume also independent errors $e_k \sim \mathcal{N}(0, \sigma_k^2)$, $k = 1, \dots, n$. Consequently,

we have that, conditional on X_k, α_k, ψ_k , and σ_k^2 , the variables Y_1, \dots, Y_n are independent such that

$$Y_k | X_k, \alpha_k, \psi_k, \sigma_k^2 \stackrel{\text{ind}}{\sim} \mathcal{N}(\alpha_k + \psi_k X_k, \sigma_k^2), \forall k = 1, \dots, n. \quad (2)$$

Notice that to make the regression more robust the Student-t distribution could be used but conjugacy could be lost and more general and less efficient sampling algorithms could be necessary, compromising the computational efficiency of the methodology.

Let us consider the vector $\theta = (\theta_1, \dots, \theta_n)$, in which $\theta_k = (\alpha_k, \psi_k, \sigma_k^2)$, that can be rewritten as follows

$$\theta = \sum_{j=1}^n \left(\theta_{[i_{j-1}i_j]} \mathbf{1}\{i_{j-1} < 1 \leq i_j\}, \dots, \theta_{[i_{j-1}i_j]} \mathbf{1}\{i_{j-1} < n \leq i_j\} \right), \quad (3)$$

in which $\{i_0, \dots, i_b\}$ is the value of the random partition ρ that denotes the instants in which the changes occur and satisfies the condition $0 = i_0 < \dots < i_b = n$, $B = b$ denotes the number of blocks in ρ , $\mathbf{1}\{A\}$ is the indicator function of event A , and $\theta_{[ij]} = (\alpha_{[ij]}, \psi_{[ij]}, \sigma_{[ij]}^2)$ denotes the common value for θ_k , $i < k \leq j$.

Notice that a cluster structure follows the definition presented in (3) for θ . The prior distribution for θ can be constructed as follows. Firstly, elicit a prior distribution for the random partition ρ that discloses your opinion about the groups defined by each common parameter $\theta_{[ij]} = (\alpha_{[ij]}, \psi_{[ij]}, \sigma_{[ij]}^2)$, $i < j$, $i, j = 1, \dots, n$. After that, specify the prior distribution for θ , given that $\rho = \{i_0, i_1, \dots, i_{b-1}, i_b\}$, which is equivalent to assign a prior distribution for the common parameters $\theta_{[i_0i_1]}, \dots, \theta_{[i_{b-1}i_b]}$.

Assuming that the data are sequentially observed and that only contiguous blocks are possible, we construct the prior distribution for ρ as follows. Assume that p denotes the probability of a change to take place at any instant. Define the Markov chain $\{Z_k : k \in \mathbb{N}\}$ generated by the instants when the changes occurred, that is, Z_k is the instant when the k th change occurs and it is such that, $Z_0 = i_0$. For $k > 0$, Z_k assumes values in the set $\{Z_{k-1} + 1, \dots, n\}$, if $Z_{k-1} \neq n$, and Z_k is n , if $Z_{k-1} = n$. Consequently, the one step transition probability associated with $\{Z_k : k \in \mathbb{N}\}$ is given by

$$c_{[ij]} = \begin{cases} p(1-p)^{j-i-1}, & \text{if } j < n, \\ (1-p)^{j-i-1}, & \text{if } j = n. \end{cases} \quad (4)$$

Assume that $P(Z_0 = i_0|p) = 1$. Thus, the prior distribution of ρ given p is

$$\begin{aligned} P(\rho = \{i_0, \dots, i_b\}|p) &= P(Z_0 = i_0, Z_1 = i_1, \dots, Z_b = i_b|p) \\ &= P(Z_b = i_b|Z_{b-1} = i_{b-1}, \dots, Z_0 = i_0, p) \dots P(Z_0 = i_0|p) \\ &= p^b(1-p)^{(n-b)}, \end{aligned} \tag{5}$$

for all $b \in I = \{1, \dots, n\}$.

In order to assign the joint prior distribution for θ , given ρ and p , we assume that:

- the common independent parameters $\theta_{[i_0i_1]}, \dots, \theta_{[i_{b-1}i_b]}$ are independent from p ; and
- for all $i, j = 1, \dots, b$, $i < j$, parameters $\alpha_{[ij]}$ and $\psi_{[ij]}$ are independent, conditional on $\sigma_{[ij]}^2$. We also consider that they have the following prior distributions

$$\alpha_{[ij]}|\sigma_{2[ij]} \sim \mathcal{N}(M, \tau_0^2 \sigma_{[ij]}^2), \tag{6}$$

$$\psi_{[ij]}|\sigma_{2[ij]} \sim \mathcal{N}(m, \gamma_0^2 \sigma_{[ij]}^2), \tag{7}$$

$$\sigma_{[ij]}^2 \sim \mathcal{IG}(a/2, d/2), \tag{8}$$

in which a, d, τ_0^2 , and γ_0^2 are positive numbers, m and M are real values, and \mathcal{N} , \mathcal{IG} stand respectively for the normal and inverse-gamma distributions.

Observe that the model above is usually called product partition model (PPM), introduced in the general context by Hartigan (1990) and in its parametric version by Barry & Hartigan (1992). The model developed here is equivalent to the PPM, for sequentially observed data and Yao's prior cohesions (Yao, 1984).

The product estimates (or posterior means) for parameters $\theta_k = (\alpha_k, \psi_k, \sigma_k^2)$, for all $k = 1, \dots, n$, can be obtained by means of the respective expectations

$$\begin{aligned} E(\alpha_k | X_{[0n]}, Y_{[0n]}) &= \sum_{i=0}^{k-1} \sum_{j=k}^n r_{[ij]}^* M_{[ij]}^*, \\ E(\psi_k | X_{[0n]}, Y_{[0n]}) &= \sum_{i=0}^{k-1} \sum_{j=k}^n r_{[ij]}^* m_{[ij]}^*, \\ E(\sigma_k^2 | X_{[0n]}, Y_{[0n]}) &= \sum_{i=0}^{k-1} \sum_{j=k}^n r_{[ij]}^* \frac{a_{[ij]}^*}{d_{[ij]}^* - 2}, \end{aligned}$$

in which $r_{[ij]}^*$, called the posterior relevance of block $[ij]$, denotes the posterior

probability of a block $[ij]$ to be in the partition ρ . Additionally, we have that

$$\begin{aligned}
d_{[ij]}^* &= d + j - i, \\
M_{[ij]}^* &= \frac{G_{[ij]}(M + \tau_0^2 \sum_{k=i+1}^j Y_k) - \tau_0^2 \sum_{k=i+1}^j X_k (m + \gamma_0^2 \sum_{k=i+1}^j Y_k X_k)}{V_{[ij]y}}, \\
V_{[ij]y} &= G_{[ij]} + (j - i)\tau_0^2 G_{[ij]} - \tau_0^2 \gamma_0^2 \left(\sum_{k=i+1}^j X_k \right)^2, \\
G_{[ij]} &= \gamma_0^2 \sum_{k=i+1}^j X_k^2 + 1, \\
t_{[ij]} &= 1 + \tau_0^2 (j - i), \\
m_{[ij]}^* &= \frac{t_{[ij]}(m + \gamma_0^2 \sum_{k=i+1}^j Y_k X_k) - \gamma_0^2 \sum_{k=i+1}^j X_k (M + \tau_0^2 \sum_{k=i+1}^j Y_k)}{V_{[ij]y}}, \\
a_{[ij]}^* &= a + \frac{t_{[ij]}(m^2 \tau_0^2 + \tau_0^2 \gamma_0^2 \sum_{k=i+1}^j Y_k^2 + M^2 \gamma_0^2)}{\tau_0^2 \gamma_0^2 t_{[ij]}} - \frac{\tau_0^4 \gamma_0^2 (\sum_{k=i+1}^j Y_k)^2}{\tau_0^2 \gamma_0^2 t_{[ij]}} \\
&\quad - \frac{\left(t_{[ij]}(m + \tau_0^2 \gamma_0^2 \sum_{k=i+1}^j Y_k X_k) - \gamma_0^2 \sum_{k=i+1}^j X_k (\tau_0^2 \sum_{k=i+1}^j Y_k + M) \right)^2}{\gamma_0^2 t_{[ij]} (t_{[ij]} G_{[ij]} - \tau_0^2 \gamma_0^2 (\sum_{k=i+1}^j X_k)^2)} \\
&\quad - \frac{M^2 \gamma_0^2 + 2M \tau_0^2 \gamma_0^2 \sum_{k=i+1}^j Y_k}{\tau_0^2 \gamma_0^2 t_{[ij]}}.
\end{aligned}$$

In order to obtain the posterior relevance for each block and the posterior distributions for ρ , p , and B , we will apply the algorithm described in the next section.

2.2 Algorithm for cluster analysis

Let U_i be an auxiliary random quantity reflecting whether or not a change point occurs at time i , that is

$$U_i = \begin{cases} 1, & \text{if } \theta_i = \theta_{i+1}, \\ 0, & \text{if } \theta_i \neq \theta_{i+1}, \end{cases}$$

for $i = 1, \dots, n - 1$. Notice that each vector $\mathbf{U} = (U_1, \dots, U_{n-1})$ immediately leads to a corresponding partition ρ . The s th partition $\mathbf{U}_s = (U_{1,s}, \dots, U_{(n-1),s})$

is generated by using Gibbs sampling through the following ratio

$$R_r = \frac{f_{[xy]}(Y_{[xy]}|X_{[xy]}) \int_0^1 p^{b-2}(1-p)^{n-b+1} d\pi(p)}{f_{[xr]}(Y_{[xr]}|X_{[xr]}) f_{[ry]}(Y_{[ry]}|X_{[ry]}) \int_0^1 p^{b-1}(1-p)^{n-b} d\pi(p)}, \quad (9)$$

for $r = 1, \dots, n-1$, in which x denotes the last change point before r and y denotes the next change point following r . For the linear regression model, the joint density $f_{[ij]}(Y_{[ij]}|X_{[ij]})$ is the $(j-i)$ -variate Student- t distribution with density function

$$f_{[ij]}(Y_{[ij]}|X_{[ij]}) = \frac{a^{d/2} \Gamma\left(\frac{d^*_{[ij]}}{2}\right)}{\pi^{(j-i)/2} \Gamma\left(\frac{d}{2}\right) (V^*_{[ij]y})^{1/2}} (a^*_{[ij]y})^{-d^*_{[ij]}/2},$$

in which $d^*_{[ij]}$, $V^*_{[ij]y}$, $t_{[ij]}$, $G_{[ij]}$, $M^*_{[ij]}$ and $m^*_{[ij]}$ are as defined earlier and

$$a^*_{[ij]y} = a + \frac{M^2}{\tau_0^2} + \frac{m^2}{\gamma_0^2} + \sum_{k=i+1}^j Y_k^2 - 2M^*_{[ij]} m^*_{[ij]} \sum_{k=i+1}^j X_k - \frac{(M^*_{[ij]})^2 t_{[ij]}}{\tau_0^2} - \frac{(m^*_{[ij]})^2 G_{[ij]}}{\gamma_0^2}.$$

Thus, assuming that $p \sim \mathcal{B}(\eta, \beta)$, that is, a beta prior distribution, each sample of the posterior distribution of p may be generated from

$$p_s \sim \mathcal{B}\left[(\eta + b_s - 1), (n + \beta - b_s)\right],$$

for $s \geq 1$, in which b_s is the number of blocks of the s th vector \mathbf{U}_s , obtained by $B_s = 1 + \sum_{i=1}^{n-1} (1 - U_{i,s})$.

Consequently, the posterior distribution of B (or alternatively, the number of change points, $B-1$) and the posterior relevance of block $[ij]$ may be estimated, respectively, by

$$\hat{P}(B = b | X_{[0n]}, Y_{[0n]}) = \frac{\sum_{s=1}^T \mathbf{1}\{B_s = b\}}{T},$$

$$\hat{r}^*_{[ij]} = \frac{M}{T},$$

in which $\mathbf{1}\{B\}$ denotes the indicator function of event B , M is the number of vectors \mathbf{U}_s for which it is observed that $U_{i,s} = 0$, $U_{(i+1),s} = \dots = U_{(j-1),s} = 1$,

$U_{j,s} = 0$, and T is the total number of vectors \mathbf{U}_s generated. For further details, the interested reader is encouraged to check in Loschi & Cruz (2005).

3 Comparing the methods by simulated data sets

In order to evaluate the performance of the proposed piecewise linear regression model, we consider in this section several generated data sets and compare the product estimates with the least squares estimates (Hocking, 1996). The least squares estimates are computed considering the most probable block indicated by the Bayesian model. The block information is included in the model through dummy variables. We consider R^2 as the fit-of-model measure. In the least squares method, R^2 is computed twice, that is, assuming and not assuming interactions among the variables. In the Bayesian model we do not consider interactions among the variables. The results for the least squares method may be obtained by using any common statistical software. For the Bayesian model, we developed an implementation in C++, a powerful and flexible programming language, available from the authors upon request.

We generated series of size 200 with only one change point, at the instant 101. We also considered that the change could occur in one, two, or in all three parameters of the regression model.

For the Gibbs sampling scheme 11,000 samples of 0-1 values were generated, with the same dimension of the time series, starting from a sequence of zeros. Because convergence was reached before the 1,000th step, the initial 1,000 iterations were discharged for *burn-in*. In order to avoid correlation among vectors a lag of 10 was selected. Notice that the inclusion of covariates slow down the convergence in comparison with the “free of covariate” case (see Loschi & Cruz, 2005).

Because the scenarios considered here presented only one change point, a beta prior distribution with most of its mass in small values of p was considered. We assumed that $p \sim \mathcal{B}(5, 50)$, which means that the prior mean of p is 0.091 and its standard deviation, 0.039. As a consequence, the mean number of change points in the sequence, $B - 1$, is 18.1. That is, in the prior evaluation we chose to overestimate the number of changes to see how the Bayesian method performs.

In all scenarios, to be described shortly, the independent variable was generated from a standard normal distribution. We also considered a low informative prior distribution for the variance σ^2 .

3.1 Analysis of scenario #1

In scenario #1, we assumed a data set with one change taking place only in the slope ψ_k . That is, we supposed a regression system obeying two separate regimes in which the dependent variables were generated as follows

$$\begin{aligned} Y_k | x_k, \alpha_k, \psi_k, \sigma_k^2 &\sim \mathcal{N}(1 - x_k, 0.1), \quad k = 1, \dots, 100, \\ Y_k | x_k, \alpha_k, \psi_k, \sigma_k^2 &\sim \mathcal{N}(1 + x_k, 0.1), \quad k = 101, \dots, 200. \end{aligned}$$

The data sets are presented in Figure 2, jointly with the fitted models. As prior specifications for the common parameters, we assumed the distributions $\sigma_{[ij]}^2 \sim \mathcal{IG}(0.201/2; 0.01/2)$, $\alpha_{[ij]} | \sigma_{[ij]}^2 \sim \mathcal{N}(1.0; \sigma_{[ij]}^2)$, and $\psi_{[ij]} | \sigma_{[ij]}^2 \sim \mathcal{N}(0.0; \sigma_{[ij]}^2)$. Notice that it means that under the 0-1 loss function the prior estimates for σ_k^2 , α_k , and ψ_k are 0.1, 1.0, and 0.0, respectively. In the posterior evaluation, we found

$$\begin{aligned} \hat{P}(\rho = \{0, 100, 200\} | Y_{[0,200]}, X_{[0,200]}) &= 0.997, \\ \hat{P}(B = 2 | Y_{[0,200]}, X_{[0,200]}) &= 0.997, \end{aligned}$$

that is, the posterior most probable partition indicated a change at instant 101. The posterior mode of the number of change points was 1.

Figure 3 presents the product estimates, the least squares estimates, and the real value for α_k , ψ_k , and σ_k^2 , plotted together. In order to compute the least squares estimates, we used the two blocks pointed out by the posterior most probable partition.

It is noticeable from Figure 3 that the posterior estimates and the least squares estimates presented similar behavior identifying a change at instant 101, in all parameters. However, we also observe that the least squares estimates for the variance were closer to the real value than the posterior estimates. For the slope ψ_k (the only parameter that really experiences any change), we noticed that the estimates were very close to the real values. However, the intercept α_k was underestimated by both methods. Figure 2 presents the adjust by both methods. The fitted models presented $R^2 = 90.1\%$ and $R^2 = 1.6\%$, respectively by the Bayesian approach and least squares estimators, when any iteration between the independent and the dummy variables was considered. A considerable improvement in the least squares method was reached when the iteration was considered (its R^2 raised to 91.5%). However, an advantage of the Bayesian approach is that we can also measure the posterior uncertainty of the fitted model because we do have the posterior distributions of B and ρ .

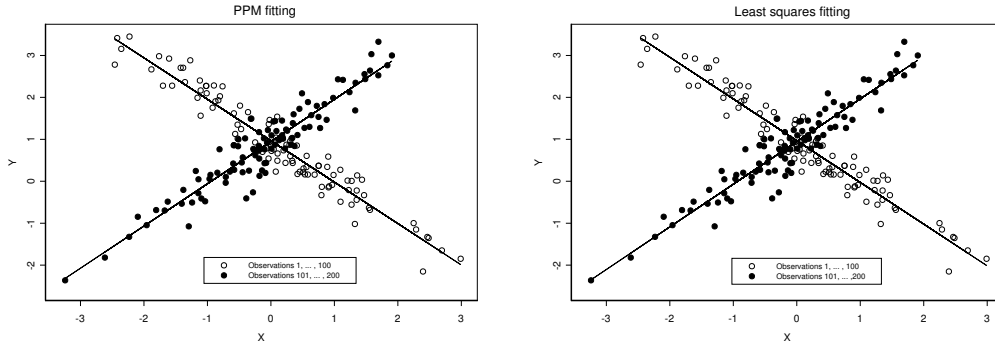


Fig. 2. Data set and fitted models for scenario #1.

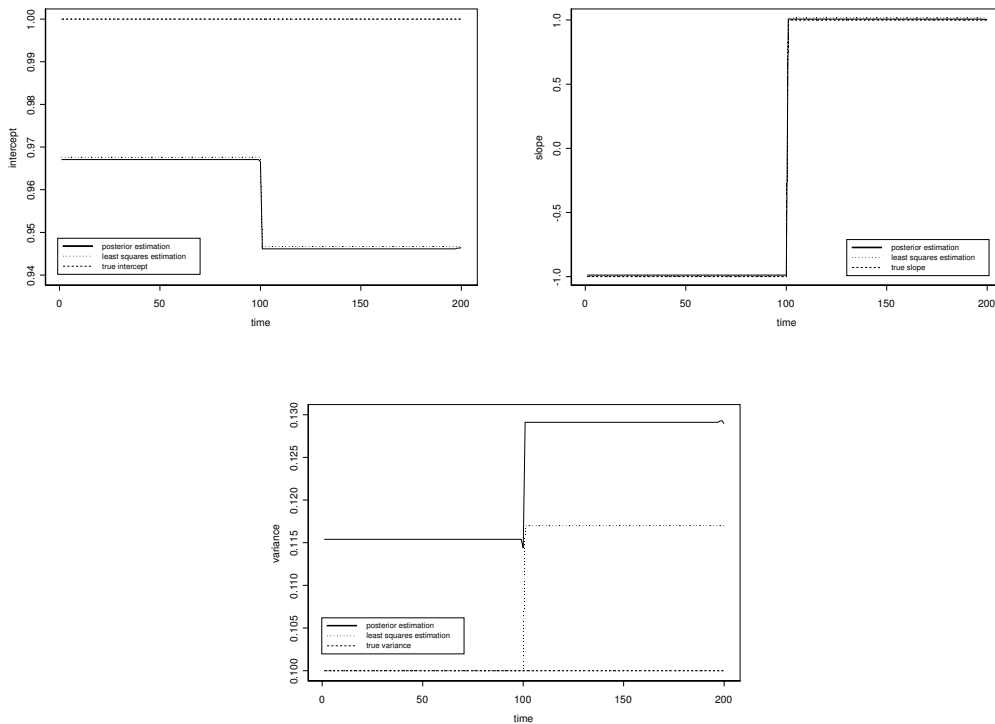


Fig. 3. Estimates for scenario #1.

3.2 Analysis of scenario #2

For scenario #2, we assumed a data set with one change occurring only in the intercept α_k . That is, we supposed that the values from the dependent variable in the regression model were generated by

$$Y_k | x_k, \alpha_k, \psi_k, \sigma_k^2 \sim \mathcal{N}(-2 + x_k, 0.1), \quad k = 1, \dots, 100,$$

$$Y_k | x_k, \alpha_k, \psi_k, \sigma_k^2 \sim \mathcal{N}(1 + x_k, 0.1), \quad k = 101, \dots, 200.$$

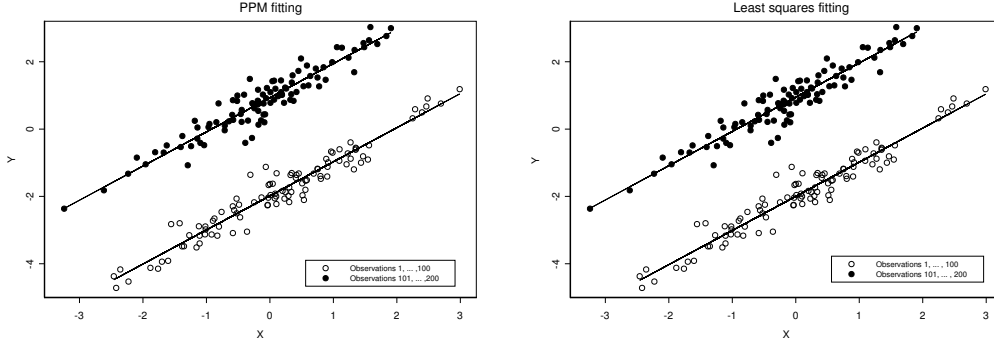


Fig. 4. Data set and fitted models for scenario #2.

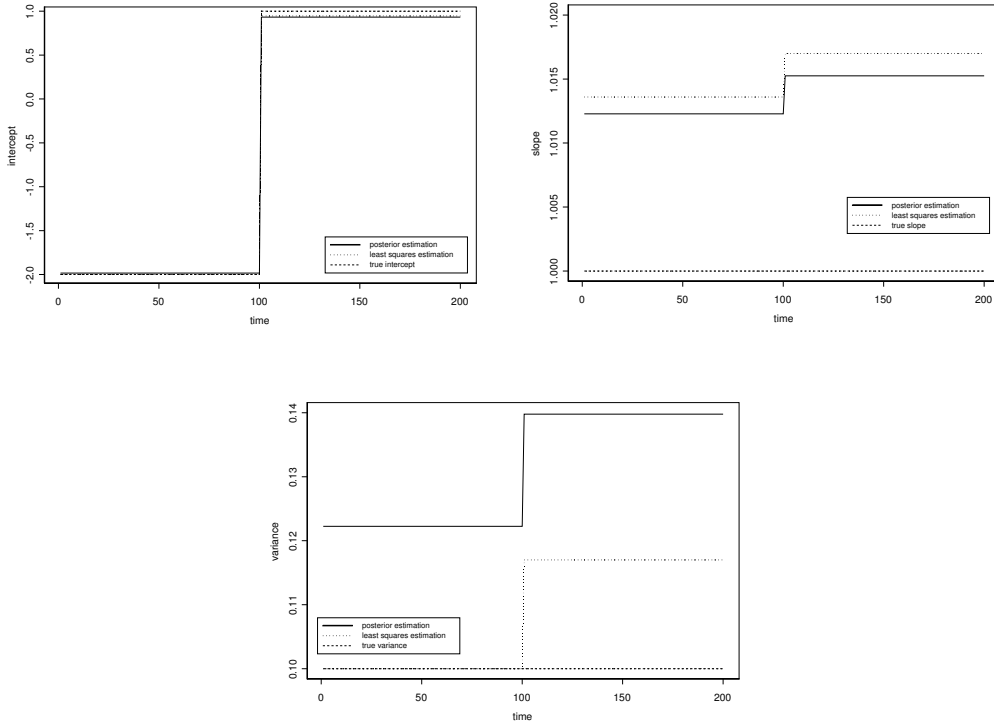


Fig. 5. Estimates for scenario #2.

The data sets and the fitted models are presented in Figure 4. As prior specifications for the common parameters, we assumed the distributions $\alpha_{[ij]}|\sigma_{[ij]}^2 \sim \mathcal{N}(-0.5; \sigma_{[ij]}^2)$ and $\psi_{[ij]}|\sigma_{[ij]}^2 \sim \mathcal{N}(1.0; \sigma_{[ij]}^2)$. The prior distribution for $\sigma_{[ij]}^2$ was the same as for scenario #1, $\sigma_{[ij]}^2 \sim \mathcal{IG}(0.201/2; 0.01/2)$. Notice that it follows from the 0-1 loss function that the prior estimates for σ_k^2 , α_k , and ψ_k are 0.1, -0.5 , and 1.0, respectively.

The posterior distribution of B indicated that the number of change points in the sequence was 1 with probability 1.0. We also noticed that the posterior most probable partition was $\rho = \{0, 100, 200\}$ with probability 1.0. Figure 5

presents the product estimates, the least squares estimates, and the real value for α_k , ψ_k , and σ_k^2 , $k = 1, \dots, n$, plotted together.

Similarly to what we have observed for scenario #1, both methods provided very close estimates to the real intercept α_k (the only parameter that really changed). As observed for scenario #1, we noticed that the least squares method provided the best estimates for the variance. However, we noticed for the intercept ψ_k that the posterior estimates were closer to the real values than the least squares estimates. In scenario #2 we observed that the fitted models presented $R^2 = 95.3\%$ and $R^2 = 96.6\%$, respectively by using the Bayesian approach and least squares estimators. For this scenario the iterations between the independent and dummy variables were not significant.

3.3 Analysis of scenario #3

The data now followed the two separate regression regimes bellow

$$Y_k|x_k, \alpha_k, \psi_k, \sigma_k^2 \sim \mathcal{N}(1 + x_k, 0.1), \quad k = 1, \dots, 100,$$

$$Y_k|x_k, \alpha_k, \psi_k, \sigma_k^2 \sim \mathcal{N}(1 + x_k, 0.5), \quad k = 101, \dots, 200.$$

The data sets are presented in Figure 6 jointly with the fitted models. As prior specifications for the common parameters, we assumed that $\sigma_{[ij]}^2 \sim \mathcal{IG}(0.603/2; 0.01/2)$ and, given $\sigma_{[ij]}^2$, we assumed that $\alpha_{[ij]}$ and $\psi_{[ij]}$ were identically distributed with normal distribution $\mathcal{N}(1.0; \sigma_{[ij]}^2)$. The prior modes for σ_k^2 , α_k , and ψ_k distributions are 0.3, 1.0, and 1.0, respectively.

As for scenario #2, the posterior distribution of B indicated that the number of change points in the sequence was one with probability 1.0. However, the posterior most probable partition for scenario #3 was $\rho = \{0, 97, 200\}$ with probability 58.8%.

As noticed for scenarios #1 and #2, both methods identified change points in parameters that really did not change at all. Notice also from Figure 7 that in this case the intercept was best estimated by the least squares method. The estimates for the slope and variance were very similar. The fitted models presented $R^2 = 81.2\%$ and $R^2 = 80.9\%$, respectively by the Bayesian approach and least squares estimators, explaining reasonably well the variability in the data. For scenario #3, we also observed that the iteration was not significant.

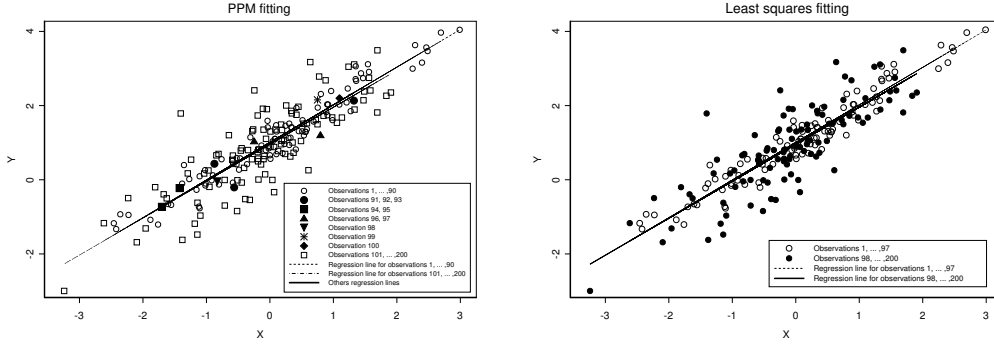


Fig. 6. Data set and fitted models for scenario #3.

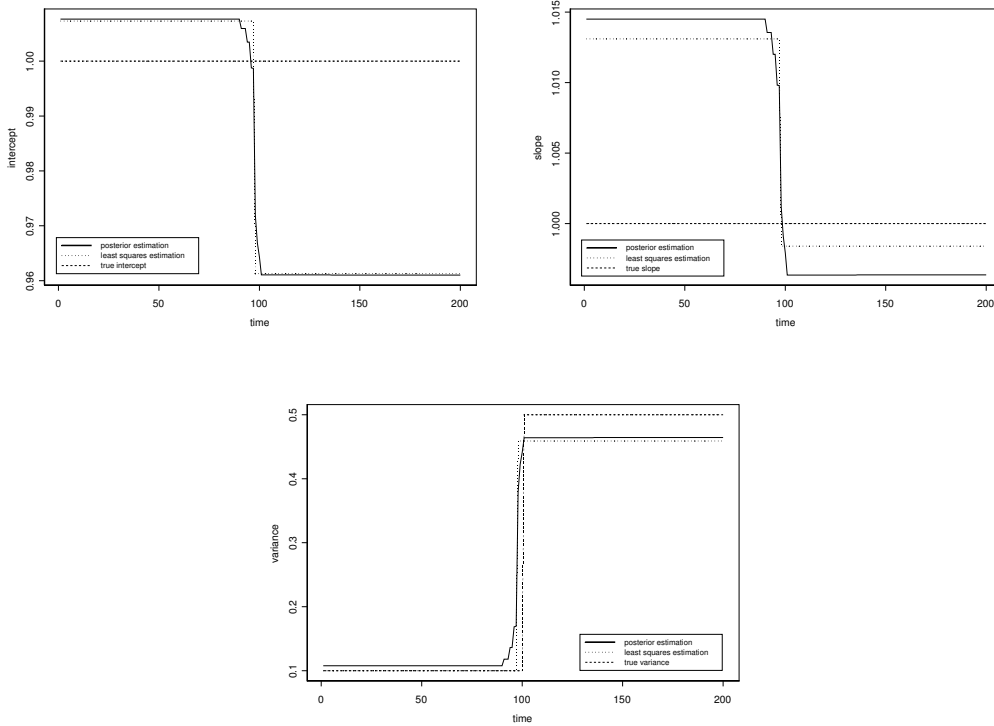


Fig. 7. Estimates for scenario #3.

3.4 Analysis of scenario #4

In scenario #4 we supposed that all parameters experienced one change at the instant 101. That is, we considered a data set that obeyed the following two separate regimes

$$Y_k | x_k, \alpha_k, \psi_k, \sigma_k^2 \sim \mathcal{N}(0 + x_k, 0.1), \quad k = 1, \dots, 100,$$

$$Y_k | x_k, \alpha_k, \psi_k, \sigma_k^2 \sim \mathcal{N}(5 - x_k, 0.5), \quad k = 101, \dots, 200.$$

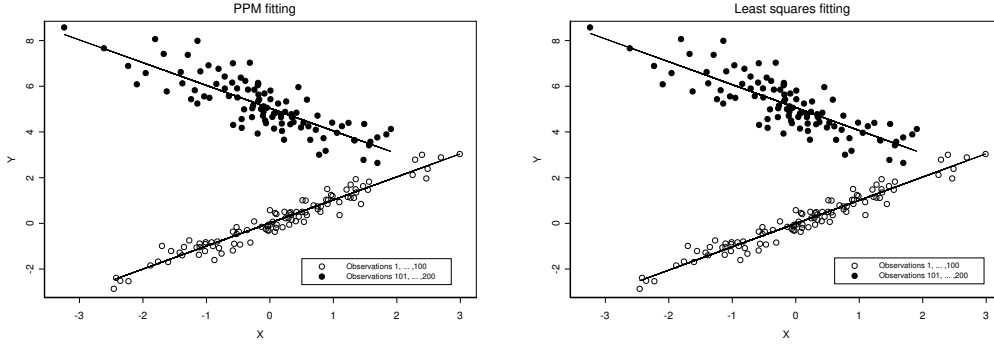


Fig. 8. Data set and fitted models for scenario #4.

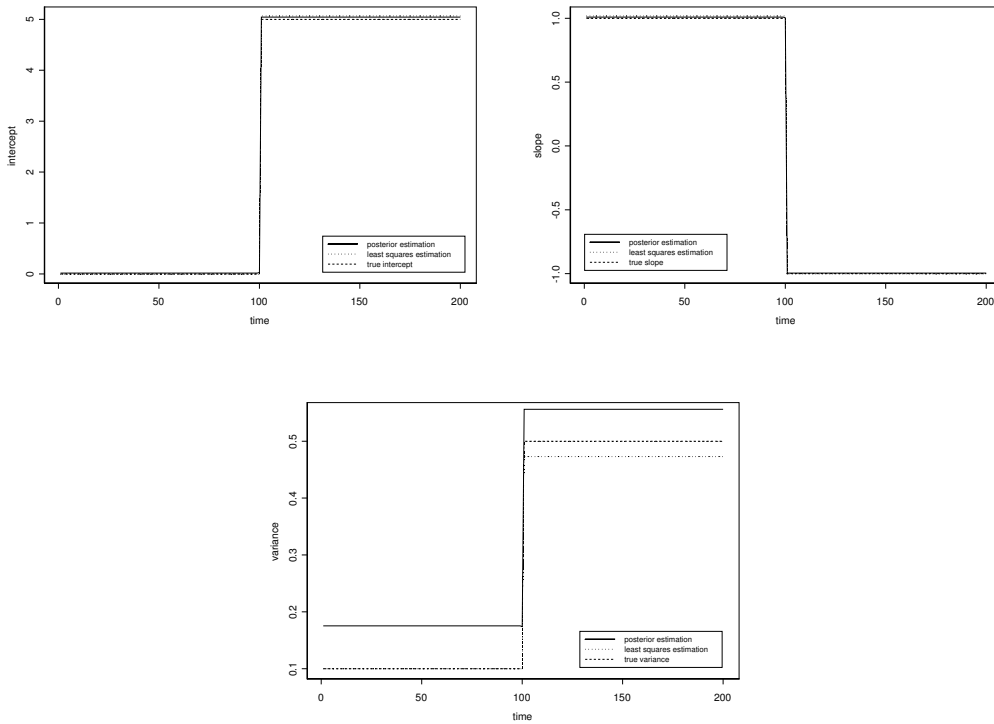


Fig. 9. Estimates for scenario #4.

The data set are presented in Figure 8, jointly with the fitted models. As prior specifications we assumed that $\sigma_{[ij]}^2 \sim \mathcal{IG}(0.603/2; 0.01/2)$, and, given $\sigma_{[ij]}^2$, we assumed that $\alpha_{[ij]}$ and $\psi_{[ij]}$ had normal distributions with means equal to 2.5 and 0.0, respectively, and variance equal to $\sigma_{[ij]}^2$. The prior modes for the distributions of σ_k^2 , α_k , and ψ_k are 0.3, 2.5, and 0.0, respectively.

The posterior distribution of B indicated that the number of change points in the sequence was one with probability 99.9%. The posterior most probable partition for scenario #4 was $\rho = \{0, 100, 200\}$, with probability 99.7%.

Notice from Figure 9 that the estimates for α_k and ψ_k are very similar to the real parameters. Here, we also observe that the Bayesian method overestimate the variance. From Figure 8, it is noticeable that both models explain reasonable well the variability in the data. The fitted models presented $R^2 = 94.6\%$ and $R^2 = 81.6\%$, respectively by using the Bayesian approach and least squares estimators (without iteration). In this scenario the Bayesian method was the most efficient in explaining the variability of the dependent variable. An improvement in the least squares method was obtained by considering the iteration between the independent and dummy variables, for which we had $R^2 = 96.4\%$.

4 Application

In this section we apply the methodology developed to analyze the behavior throughout the time of the Brazilian Industrial Production Index (BIP) and the Brazilian Employment Index (BEI), recorded monthly, from January, 1985, to April, 2001. From Figure 1, it is noticeable that the variables seem to be linearly correlated within different blocks. Notice that for the sake of conciseness aspects concerning the time series structure, such as long memory, unit roots, and cointegration were not discussed explicitly in here but they must be considered before proceeding with the regression (Bai, 1997; Bai & Perron, 1998).

Because we do not have much prior information about the parameters and we expect few changes along the time in the relationship between BIP and BEI, the following prior distributions are considered reasonable

$$\begin{aligned}\sigma_{[ij]}^2 &\sim \mathcal{IG}(0.001/2; 0.001/2), \\ \alpha_{[ij]} | \sigma_{[ij]}^2 &\sim \mathcal{N}(0.0; \sigma_{[ij]}^2), \\ \psi_{[ij]} | \sigma_{[ij]}^2 &\sim \mathcal{N}(0.0; \sigma_{[ij]}^2), \\ p &\sim \mathcal{B}(5.0; 50.0).\end{aligned}$$

As a consequence, it follows that the prior expected number of change points is 17.4, with standard deviation equals to 8.3 changes. For the Gibbs sampling scheme we considered the same sample size, burn-in, and lag assumed earlier.

The posterior most probable partition identified that changes occurred at instants 102 (June, 1993) and 149 (May, 1997) with probability 1.0. It was also identified that $\hat{P}(B = 3 | Y_{[0,196]}, X_{[0,196]}) = 1.0$, which means that only two changes took place with probability 1.0 in the linear relationship between BIP and BEI.

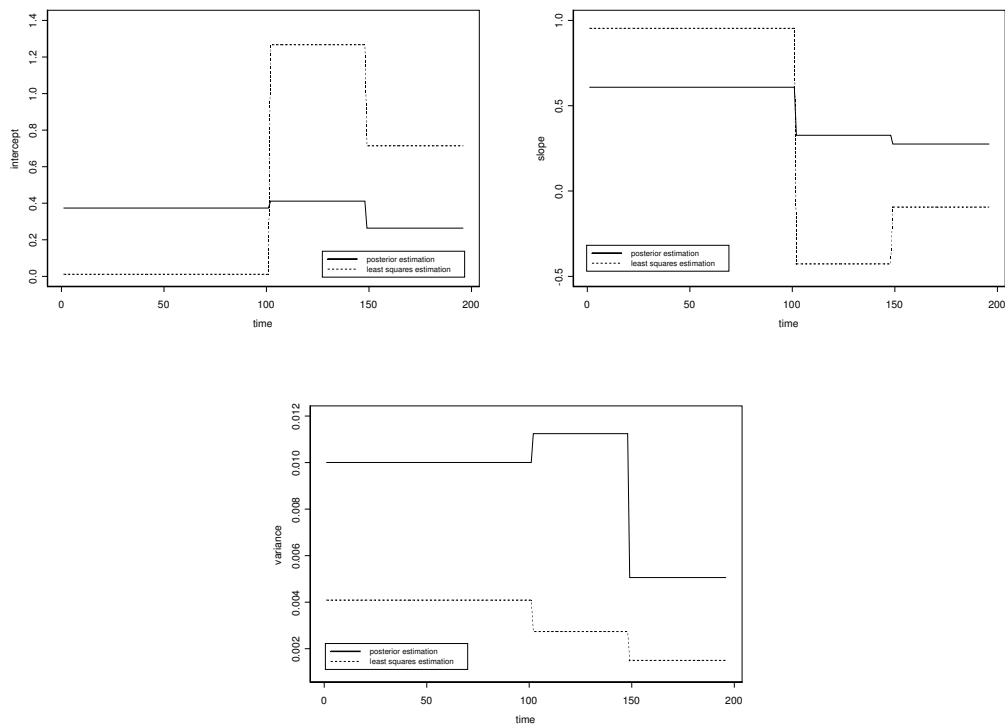


Fig. 10. Estimates for BIP *versus* BEI.

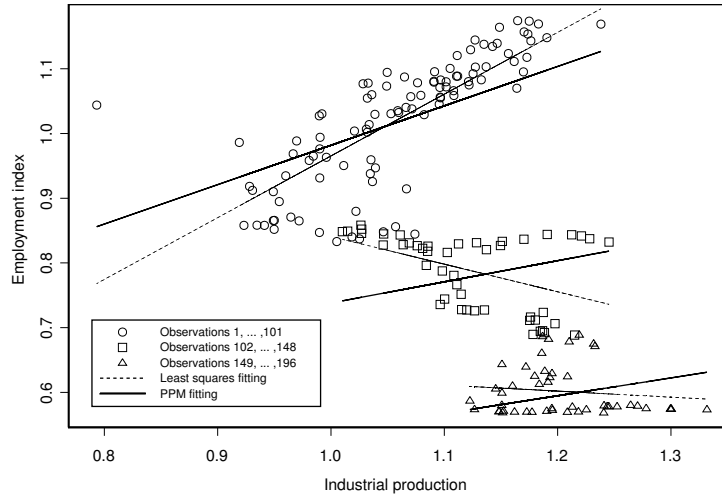


Fig. 11. Data set and fitted models for BEI *vs.* BIP.

From Figure 10 we can notice that the product estimates and least squares estimates identified changes in all three parameters at the same time. However, we can perceive that the estimates were not only different but also presented different behavior. For instance, while the product estimates indicated that the intercept decreased after May, 1997, the least squares estimate indicated

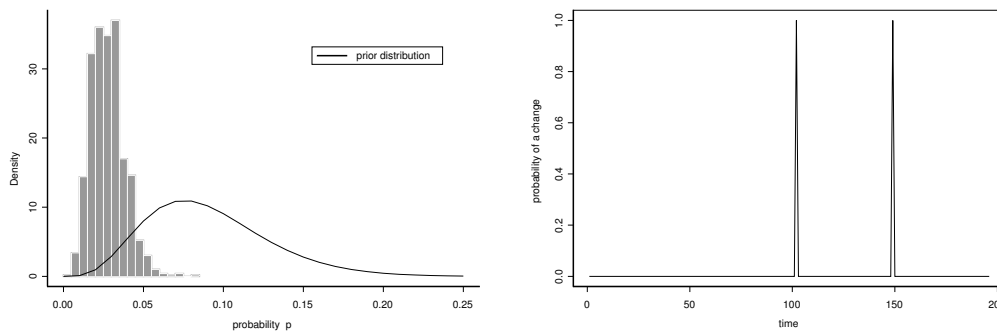


Fig. 12. Posterior distribution of p and probability of change for BEI *vs.* BIP.

that the intercept increased.

Figure 11 shows the fitted models by using the Bayesian and least squares approaches for which it was obtained $R^2 = 88.3\%$ and $R^2 = 87.2\%$, respectively. As observed for the simulated data sets, an improvement was reached for the least squares method by adding iteration between the dummy variables and the BIP, which raised R^2 to 91.2%.

By means of the Bayesian approach we have some other measurements that can help evaluating the model, which are the posterior distribution for the probability p of a change to take place at any month and the probability of each month to be a change point. It is noticeable from Figure 12 that the posterior distribution for p is more concentrated (standard deviation = 0.0104) and located at smaller values (posterior mean = 0.0277) than the prior distribution. We also noticed that all months but June, 1993, and May, 1997 (which are change points with probability 1.0) had probability zero of being change points.

5 Conclusions and Final Comments

We proposed a full Bayesian approach based on the product partition model (PPM) for piecewise linear regression. The proposed method considers multiple segments and models. Besides, it assumes as random both the number of segments and the instants when the changes occur. We compared the Bayesian model with a modified least squares approach whose blocks of observations were provided by the Bayesian posterior most probable partition. The methods were applied to analyze simulated data sets and to estimate the relationship between the Brazilian Industrial Production Index (BIP) and the Brazilian Employment Index (BEI) along the time. In all cases, low informative prior distribution was considered for the variances and the results were

quite satisfactory.

In general, the Bayesian approach provided the highest R^2 values, when no iteration was considered between the independent and dummy variables in the least squares approach. An improvement was obtained in the least squares models when the iterations were included in the model. Despite of this, the Bayesian approach has advantages over the least squares approach because it also may provide some other measures that permit a more complete evaluation of the model, such as the posterior distributions of the number of blocks, the instants when changes occurred, and the posterior probability of each instant to be a change point. For the simulated scenarios, it was noticed that the product estimates tend to overestimate the variance of the model. This could be a consequence of using low informative prior distribution for the parameter.

As a final remark, it has been confirmed here an issue related to the PPM, which is its inefficiency to precisely identifying change points when only one parameter changes its levels. Thus it is expected that heteroschedasticity, typically present in regression data, may compromise the quality of the inferences for the intercept and slope.

Acknowledgements

Rosângela Loschi acknowledges PRPq-UFMG (grant 4801-UFMG/RTR/FUNDO/PRPq/RECEM-DOCTORES/00) and CNPq (*Conselho Nacional de Desenvolvimento Científico e Tecnológico*) of the Ministry for Science and Technology of Brazil (grants 300325/2003-7 and 472066/2004-8), for a partial allowance to her research. Vinicius D. Mayrink had his research partially supported by CNPq, grant 110061/2002-0. The research of Frederico Cruz has been partially funded by the CNPq, grants 201046/1994-6, 301809/1996-8, 307702/2004-9, and 472066/2004-8, the FAPEMIG (*Fundação de Amparo à Pesquisa do Estado de Minas Gerais*), grants CEX-289/98 and CEX-855/98, and PRPq-UFMG, grant 4081-UFMG/ RTR/FUNDO/PRPq/99.

References

- Bai, J., 1997. Estimation of a change point in multiple regression models. *The Review of Economics & Statistics* 79 (4), 551–563.
- Bai, J. & Perron, P., 1998. Estimating and testing linear models with multiple structural changes. *Econometrica* 66 (1), 47–78.
- Barry, D. & Hartigan, J., 1992. Product partition models for change point problems. *The Annals of Statistics* 20 (1), 260–279.

- Barry, D. & Hartigan, J., 1993. A Bayesian analysis for change point problem. *Journal of the American Statistical Association* 88 (421), 309–319.
- Dias, R., 1999. Sequential adaptative non parametric regression via h-splines. *Communications in Statistics: Computation & Simulations* 28, 501–515.
- Diniz, C., Milan, L. A. & Mazuchelli, J., 2003. Bayesian inference of a linear segmented regression model. *Brazilian Journal of Probability & Statistics* 17, 1–16.
- Ferreira, P. E., 1975. A Bayesian analysis of a switching regression model: Known number of regimes. *Journal of the American Statistical Association* 70, 370–355.
- Hartigan, J., 1990. Product partitions models. *Communications in Statistics: Theory & Methods* 19, 2745–2756.
- Hocking, R. R., 1996. *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. John Wiley & Sons, New York.
- Holbert, D., 1982. A Bayesian analysis of a switching linear model. *Journal of Econometrics* 19, 77–87.
- Holmes, C. C. & Mallick, B. K., 2001. Bayesian regression with multivariate linear splines. *Journal of the Royal Statistical Society B* 63, 3–17.
- Loschi, R. H. & Cruz, F. R. B., 2002. An analysis of the influence of some prior specifications in the identification of change points via product partition model. *Computational Statistics & Data Analysis* 39 (4), 477–501.
- Loschi, R. H. & Cruz, F. R. B., 2005. Extension to the product partition model: Computing the probability of a change. *Computational Statistics & Data Analysis* 48 (2), 255–268.
- McGee, V. E. & Carleton, W. T., 1970. Piecewise regression. *Journal of the American Statistical Association* 65 (331), 1109–1124.
- Quant, R. E., 1958. The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association* 53, 873–880.
- Wahba, G., 1983. Bayesian ‘confidence intervals’ for the cross-validated smoothing spline. *Journal of the Royal Statistical Society B* 45, 133–150.
- Yao, Y., 1984. Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches. *The Annals of Statistics* 12 (4), 1434–1447.