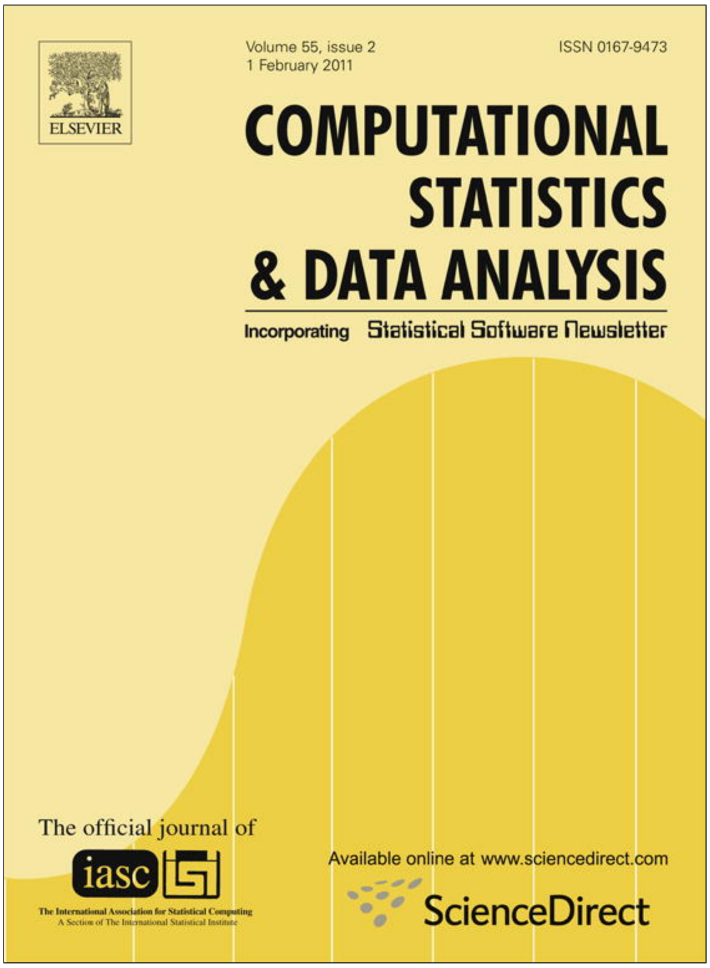
ELSEVIER
Volume 55, issue 2
1 February 2011
ISSN 0167-9473
COMPUTATIONAL STATISTICS & DATA ANALYSIS
Incorporating Statistical Software Newsletter
The official journal of
iasc
The International Association for Statistical Computing
A Section of The International Statistical Institute
Available online at www.sciencedirect.com
ScienceDirect

Computational Statistics and Data Analysis 55 (2011) 1160–1170

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Joint segmentation of multivariate Gaussian processes using mixed linear models

F. Picard[a,b,e], E. Lebarbier[c,*], E. Budinskà[d], S. Robin[c]

[a] UMR CNRS-8071/INRA-1152/Université d'Évry, Évry, France
[b] UMR CNRS 5558 Université Lyon 1 Claude Bernard, Université de Lyon, Laboratoire de Biométrie et Biologie Evolutive, Lyon, France
[c] AgroParisTech/INRA MIA 518, Paris, France
[d] Institute of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic
[e] BAMBOO Project, INRIA Rhône-Alpes, Saint-Martin, France

ARTICLE INFO

Article history:
Received 18 February 2010
Received in revised form 16 September 2010
Accepted 17 September 2010
Available online 1 October 2010

Keywords:
Segmentation
Mixed linear model
Multivariate Gaussian process
Dynamic programming
EM algorithm

ABSTRACT

The joint segmentation of multiple series is considered. A mixed linear model is used to account for both covariates and correlations between signals. An estimation algorithm based on EM which involves a new dynamic programming strategy for the segmentation step is proposed. The computational efficiency of this procedure is shown and its performance is assessed through simulation experiments. Applications are presented in the field of climatic data analysis.

Introduction

Many application fields in statistics provide signals which are in the form of non-stationary time series. To simplify the interpretation of such signals, segmentation models are often used to identify intervals in which the signal is homogeneous. To do this, the data are modelled by a random process whose parameters are subject to abrupt changes at unknown coordinates. In this work, we consider the *off-line* multiple-change-point problem, for which the segmentation is uncovered once the data have been observed. When considering single time series, the objective is to identify the number and the position of the change points, as well as the values of the parameter within intervals of homogeneity. Many strategies exist in this framework, and intensive research has been conducted to develop efficient segmentation algorithms. The literature is abundant in this field, and the reader is referred to Lavielle and Lebarbier (2001), Bai and Perron (2003), Lebarbier (2005), Fearnhead (2006) and references therein.

As the problem has shifted from single-change-point to multiple-change-point detection, attention has now focused on the detection of multiple changes in multiple time series. Now the purpose is to detect and characterize structure in two or more related series. In the following we focus on *joint* segmentation, to be distinguished from *simultaneous* segmentation. In *joint* segmentation, each series has its own segmentation which is achieved jointly with others, whereas for *simultaneous*

[*] Corresponding address: AgroParisTech/INRA MIA 518, 16 rue Claude Bernard, F - 75 231 Paris Cedex 05, France. Tel.: +33 0 1 44 08 16 72; fax: +33 0 1 44 08 16 66.
E-mail addresses: franck.picard@univ-lyon1.fr (F. Picard), lebarbie@agroparistech.fr, Lebarbier@agroparistech.fr (E. Lebarbier), budinska@cba.muni.cz (E. Budinskà), robin@agroparistech.fr (S. Robin).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.09.015

segmentation changes are common among series. The main motivation for joint segmentation arises when multiple series share some characteristics that cannot be modelled by the segmentation part. An illustration is provided in climatology (Caussinus and Mestre, 2004). Climate series are recorded at different stations and are subject to specific abrupt changes due to climate or instrumental effects for instance (Caussinus and Mestre, 2004). In addition to local abrupt changes, the same climate series can share some global trend that does not show any break. Another example is provided by the joint segmentation of wind speed and direction (Dobigeon and Tourneret, 2007).

A first strategy for handling multiple series is to derive a segmentation model for independent series using the linear model framework, with pure or partial structural changes (Bai and Perron, 2003; Caussinus and Mestre, 2004). In this work, we propose a generalization of those models to mixed linear models. This strategy allows us to model complex correlation structures, and as an example we show how it can be used to link time series at every instant. We develop an algorithm for estimating the parameters by using the maximum likelihood. Optimization is a challenging task when dealing with mixed models and with classical segmentation models. Parameters are estimated using the Expectation/Conditional Maximization algorithm (Meng and Rubin, 1993). ECM is an instance of the traditional EM algorithm (Dempster et al., 1977), which replaces a complicated M-step of EM with several computationally simpler CM-steps. This algorithm can be used in this context as linear mixed models can be put in the more general framework of models with incomplete data (van Dyk, 2000).

Among the CM-steps, one is dedicated to the estimation of the change-point location. Dynamic Programming (DP) is an efficient algorithm for change-point positioning (Bai and Perron, 2003; Caussinus and Mestre, 2004; Picard et al., 2005). The interest of Dynamic Programming is that it allows the exact computation of the maximum likelihood estimates with a reduction in the algorithmic complexity from $\mathcal{O}(N^K)$ to $\mathcal{O}(KN^2)$. However, even if the computational load is reduced when using DP, this method cannot be used directly for joint segmentations, since N (the total number of points) becomes too large. Heuristics have been proposed: Bai and Perron (2003) propose adding constraints on the length of segments and Caussinus and Mestre (2004) propose a prior selection of possible breaks. These two strategies do not lead to the optimal segmentation solution. In this work, we overcome this computational issue and propose a new procedure which gives the optimal segmentation. This procedure is based on a two-step dynamic programming algorithm. It can be used when dealing with linear models with or without random effects.

In addition to modelling and computational issues, joint segmentation raises a model selection issue which is the choice of the total number of segments. Choosing the number of segments has been widely investigated in the univariate segmentation case using Bayesian (Caussinus and Mestre, 2004; Zhang and Siegmund, 2007) or adaptive strategies (Lavielle, 2005; Lebarbier, 2005). In this article we study the generalization of four model selection criteria to the multivariate case, and we assess their performance using simulations.

This work is organized as follows. In Section 1 we present a statistical model for joint segmentation whose parameters can be estimated by using the maximum likelihood, using an algorithm described in Section 2. We propose a computational trick for using Dynamic Programming in the joint segmentation case (Section 3) and model selection criteria are developed in Section 4. Sections 5 and 6 are dedicated to simulations and to a real-data example on climate series.

1. Linear models for joint segmentation

We consider M time series with n_m observations each, and we denote as $N = \sum_m n_m$ the total number of observations. We denote by t the position of the signal and by $n_{\max}$ the maximum number of points in a series, $t \in \{1, \dots, n_{\max}\}$. We observe Y_{mt}, the signal of series m at position t. We suppose that part of the mean of the process $\{Y_{mt}\}_t$ is subject to $K_m - 1$ abrupt changes at breakpoints $\{t_k^m\}$ for series m (with the conventions $t_0^m = 0$ and $t_{K_m}^m = n_{\max}$) and is constant between two breakpoints within the interval $I_k^m =]t_{k-1}^m, t_k^m]$. In the following we denote by $K = \sum_m^M K_m$ the total number of segments across series which is fixed in this section. Following Bai and Perron (2003) we consider the following linear model:

$$\forall t \in I_k^m, \quad Y_{mt} = \mu_{mk} + \mathbf{x}_{mt}\boldsymbol{\theta} + E_{mt},$$

with $\mathbf{x}_{mt}$ a $[1 \times p]$ vector of covariates, $\boldsymbol{\theta}$ the corresponding parameter which is not subject to changes, whereas μ_{mk} is the one which is subject to changes. When $p = 0$ we obtain a pure structural change model. E_{mt} stands for the noise.

To use the matricial formulation of linear models, we introduce the $[N \times K]$-incidence matrix of breakpoints denoted by $\mathbf{T} = \text{Bloc}\,[\mathbf{T}_m]$ with $\mathbf{T}_m = \text{Bloc}\left[\mathbb{1}_{n_{K_m}^m}\right]$ of size $[n_m \times K_m]$, and with $n_k^m = t_k^m - t_{k-1}^m$ being the length of segment k for series m. We also introduce the notation $\boldsymbol{\mu} = [\mu_{mk}]$ which corresponds to the fixed effects subject to changes (of size $[K \times 1]$). Using the matricial formulation of linear models, we have

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\theta} + \mathbf{E},$$

where $\mathbf{Y}$ ($[N \times 1]$) stands for the observed data, and where $\mathbf{T}$, $\mathbf{X}$ are incidence matrices of breakpoints and the constant parameter with respective size $[N \times K]$, $[N \times p]$. Note that unlike in classical linear models, the incidence matrix $\mathbf{T}$ is unknown and should be estimated. $\mathbf{E}$ is centered Gaussian with covariance matrix $\mathbf{R}$.

Here we consider a more general model by introducing $\mathbf{U}$ ($[n_{\max} \times 1]$), the vector of $n_{\max}$ random effects with incidence matrix $\mathbf{Z}$ ($[N \times n_{\max}]$). We suppose that $\mathbf{U}$ is centered Gaussian with covariance matrix $\mathbf{G}$. $\mathbf{U}$ and $\mathbf{E}$ are supposed independent.

In the harvest example the random effect U_t would correspond to an 'annual' effect and any systematic climatic effect should be modeled as a fixed effect $\mathbf{X}\boldsymbol{\theta}$. Overall, the mixed linear model that we consider is

$$\mathbf{Y} = \mathbf{T}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\mathbf{U} + \mathbf{E},$$

and thus $\mathbf{Y}$ has a mean $\mathbf{T}\boldsymbol{\mu} + \mathbf{X}\boldsymbol{\theta}$ and variance $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$.
The covariance matrix $\mathbf{R}$ is supposed to be diagonal. This condition is required to allow the use of the DP algorithm (used to obtain the segmentation parameters; see the following section). The identifiability of the model is ensured when $\begin{bmatrix}\mathbf{T} & \mathbf{X}\end{bmatrix}$ is a full rank matrix whatever the form of the covariance matrix $\mathbf{G}$.

2. Parameter estimation using the ECM algorithm

We propose to estimate the parameters of the model by using the maximum likelihood. In the following, we denote by $\phi = (\boldsymbol{\mu}, \boldsymbol{\theta}, \mathbf{G}, \mathbf{R}, \mathbf{T})$ the set of parameters to be estimated. The use of the EM algorithm Dempster et al. (1977) is now well established in the context of parameter estimation for mixed linear models van Dyk (2000), as these models can be put in the general framework of models with incomplete data. In this case, random effects $\mathbf{U}$ constitute the unobserved data, and the use of EM lies in the breakdown of the complete-data log-likelihood such that $\log \mathcal{L}(\mathbf{Y}, \mathbf{U}; \phi) = \log \mathcal{L}(\mathbf{Y}|\mathbf{U}; \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{\mu}, \mathbf{R}) + \log \mathcal{L}(\mathbf{U}; \mathbf{G})$. We denote by $\mathbb{E}_{\phi}\{\cdot\}$ the expectation operator using ϕ as the parameter value and by $\mathbb{V}_{\phi}\{\cdot\}$ the corresponding variance. The conditional expectation $Q(\phi; \phi^{(h)})$ of $\log \mathcal{L}(\mathbf{Y}, \mathbf{U}; \phi)$ given $\mathbf{Y}$ is also a sum of two terms $Q_0(\phi; \phi^{(h)})$ and $Q_1(\phi; \phi^{(h)})$:

$$\begin{aligned}
-2Q_0(\phi; \phi^{(h)}) &= -2\mathbb{E}_{\phi^{(h)}}\{\log \mathcal{L}(\mathbf{Y}|\mathbf{U}; \boldsymbol{\theta}, \mathbf{T}, \boldsymbol{\mu}, \mathbf{R})|\mathbf{Y}\} \\
&= N\log(2\pi) + \log|\mathbf{R}| + \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta} - \mathbf{T}\boldsymbol{\mu} - \mathbf{Z}\widehat{\mathbf{U}}^{(h)}\|^2_{\mathbf{R}^{-1}} + \mathrm{Tr}\left(\mathbf{R}^{-1}\mathbf{Z}\mathbf{W}^{(h)}\mathbf{Z}'\right), \\
-2Q_1(\phi; \phi^{(h)}) &= -2\mathbb{E}_{\phi^{(h)}}\{\log \mathcal{L}(\mathbf{U}; \mathbf{G})|\mathbf{Y}\} \\
&= q\log(2\pi) + \log|\mathbf{G}| + \widehat{\mathbf{U}}^{(h)\prime}\mathbf{G}^{-1}\widehat{\mathbf{U}}^{(h)} + \mathrm{Tr}\left(\mathbf{G}^{-1}\mathbf{W}^{(h)}\right),
\end{aligned}$$

where $\widehat{\mathbf{U}}^{(h)} = \mathbb{E}_{\phi^{(h)}}\{\mathbf{U}|\mathbf{Y}\}$ stands for the best linear unbiased predictor (BLUP) of the random effects $\mathbf{U}$, $\mathrm{Tr}(A)$ for the trace of matrix A, $|A|$ for its determinant and where $\mathbf{W}^{(h)} = \mathbb{V}_{\phi^{(h)}}\{\mathbf{U}|\mathbf{Y}\}$.

2.1. E-step

This step consists in the calculation of $Q(\phi; \phi^{(h)})$ which only requires the calculation of $\widehat{\mathbf{U}}$ and $\mathbf{W}$. The BLUP is such that $\widehat{\mathbf{U}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta} - \mathbf{T}\boldsymbol{\mu})$, and we use Henderson's trick which avoids the inversion of $\mathbf{V}$. So we get at iteration $(h+1)$

$$\begin{cases}
\widehat{\mathbf{U}}^{(h+1)} = \mathbf{W}^{(h)}\mathbf{Z}'\mathbf{R}^{(h)-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}^{(h)} - \mathbf{T}^{(h)}\boldsymbol{\mu}^{(h)}\right), \\
\mathbf{W}^{(h+1)} = \left(\mathbf{Z}'\mathbf{R}^{(h)-1}\mathbf{Z} + \mathbf{G}^{(h)-1}\right)^{-1}.
\end{cases}$$

2.2. CM-steps

The principle of the ECM algorithm is to break down the maximization of $Q(\phi; \phi^{(h)})$ with respect to ϕ (the global M-step) into simpler CM-steps which focus on one parameter, the others being fixed. The convergence properties of ECM are provided in Meng and Rubin (1993).

Estimation of $\boldsymbol{\theta}$. The update of $\boldsymbol{\theta}$ is done with the classical least-squares estimator

$$\mathbf{X}'\mathbf{R}^{(h)-1}\mathbf{X}\boldsymbol{\theta}^{(h+1)} = \mathbf{X}'\mathbf{R}^{(h)-1}(\mathbf{Y} - \mathbf{T}^{(h)}\boldsymbol{\mu}^{(h)} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)}).$$

Estimation of variance components. We get the estimates $\mathbf{G}^{(h+1)}$ and $\mathbf{R}^{(h+1)}$:

$$\mathbf{G}^{(h+1)} = \arg\max_{\mathbf{G}} Q_1(\phi; \boldsymbol{\theta}^{(h+1)}, \mathbf{T}^{(h)}, \boldsymbol{\mu}^{(h)}, \mathbf{G}^{(h)}, \mathbf{R}^{(h)})$$

and

$$\mathbf{R}^{(h+1)} = \arg\max_{\mathbf{R}} Q_0(\phi; \boldsymbol{\theta}^{(h+1)}, \mathbf{T}^{(h)}, \boldsymbol{\mu}^{(h)}, \mathbf{G}^{(h+1)}, \mathbf{R}^{(h)}).$$

Estimation of segmentation parameters. This step is done such that

$$\left\{\mathbf{T}^{(h+1)}, \boldsymbol{\mu}^{(h+1)}\right\} = \arg\max_{\mathbf{T},\boldsymbol{\mu}} Q_0\left(\phi; (\boldsymbol{\theta}^{(h+1)}, \mathbf{G}^{(h+1)}, \mathbf{R}^{(h+1)})\right),$$

and the computation of this particular CM-step is equivalent to the minimization of the residual sum of squares:

$$\begin{aligned}
\mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu}) &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}^{(h+1)} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)} - \mathbf{T}\boldsymbol{\mu}\|^2_{\mathbf{R}^{(h+1)-1}}, \\
&= \|\widetilde{\mathbf{Y}} - \mathbf{T}\boldsymbol{\mu}\|^2_{\mathbf{R}^{(h+1)-1}},
\end{aligned}$$

where $\widetilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\theta}^{(h+1)} - \mathbf{Z}\widehat{\mathbf{U}}^{(h+1)}$. This minimization should be done with respect to the constraint $\sum_m K_m = K$. At this stage the problem is to obtain the best segmentation of the new series $\widetilde{\mathbf{Y}}$ into K segments. We provide a computational solution for performing this step in the next section.

3. Using dynamic programming for joint segmentation

In this section we propose to reduce the computational load of traditional Dynamic Programming (DP) when segmenting multiple time series $\mathbf{Y}$. DP is used to solve the following minimization problem:

$$\{\hat{\mathbf{T}}, \hat{\boldsymbol{\mu}}\} = \arg\min_{\{\mathbf{T},\boldsymbol{\mu}\}} \mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu}),$$

with $\mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu})$ the residual sum of squares of a segmentation model with K segments. When dealing with multiple time series, this minimization must be done under an additional constraint compared with traditional individual segmentations, namely $\sum_m K_m = K$. With homoskedastic noise, the residual sum of squares is

$$\begin{aligned}\mathrm{RSS}_K(\boldsymbol{\mu}, \mathbf{T}) = \|\mathbf{Y} - \mathbf{T}\boldsymbol{\mu}\|^2 &= \sum_{m=1}^{M}\sum_{k=1}^{K_m} \mathrm{RSS}_k^m(\boldsymbol{\mu}_m, \mathbf{T}_m),\\ &= \sum_{m=1}^{M}\sum_{k=1}^{K_m}\sum_{t\in I_k^m}(y_{mt} - \mu_{km})^2. \end{aligned} \qquad (1)$$

Note that, in our problem, this is applied to $\mathbf{Y} = \widetilde{\mathbf{Y}}$. In a heteroskedastic noise case, the corresponding weighted sum of squares is considered.

Then the purpose is to optimize Eq. (1) under the constraint $\sum_m K_m = K$. An important property of the RSS is that it is additive according to the number of segments, which allows us to use Dynamic Programming at this step. The computational trick that we propose is based on the following breakdown:

$$\min_{\{\mathbf{T},\boldsymbol{\mu}\}} \mathrm{RSS}_K(\mathbf{T}, \boldsymbol{\mu}) = \min_{K_1+\cdots+K_M=K} \left\{ \sum_{m=1}^{M} \min_{\mathbf{T}_m,\boldsymbol{\mu}_m} \mathrm{RSS}_{K_m}^m(\mathbf{T}_m, \boldsymbol{\mu}_m) \right\}.$$

This optimization problem is a partitioning problem whose purpose is to cut the interval $[1, N]$ into K intervals structured according to series. Denoting by $J^m =]t_0^m, t_{K_m}^m]$ the time intervals for series m, the aim is to find the best partition of $[1, N]$ according to the RSS, such that

$$[1, N] = \bigcup_{m=1}^{M} J^m = \bigcup_{m=1}^{M}\bigcup_{k=1}^{K_m} I_k^m.$$

To account for this structure, and to get an efficient algorithm when N is large, we propose a two-stage dynamic programming.

Stage 1. This step gives an optimal solution for the first minimization step:

$$\forall m \in [1, M] \quad \{\hat{\mathbf{T}}_m, \hat{\boldsymbol{\mu}}_m\} = \min_{\mathbf{T}_m,\boldsymbol{\mu}_m} \mathrm{RSS}_{K_m}^m(\mathbf{T}_m, \boldsymbol{\mu}_m).$$

We denote by $\mathrm{RSS}_k^m(J^m)$ the minimal residual sum of squares when partitioning interval J^m of series m into k segments. This segmentation step is based on the calculation of $\mathrm{RSS}_1^m(]i, j])$ and on the recursive minimization

$$\begin{aligned}&\forall k \in [1 : K_m],\\ &\mathrm{RSS}_k^m(]t_1^m, j]) = \min_h \left\{ \mathrm{RSS}_{k-1}^m(]t_1^m, h]) + \mathrm{RSS}_1^m(]h, j]) \right\}.\end{aligned}$$

In this stage, each series is segmented for $k = 1, \ldots, K_m$ segments by using the classical DP algorithm.

Stage 2. The second step consists in solving

$$\min_{K_1+\cdots+K_M=K} \sum_{m=1}^{M} \mathrm{RSS}_{K_m}^m(\hat{\mathbf{T}}_m, \hat{\boldsymbol{\mu}}_m).$$

We denote by $\mathrm{RSS}_K(J^1, \ldots, J^m)$ the total sum of squares for a model with K segments spread over m series. The second step consists in the repartition of segments among M time series. This step is based on the calculation of $\mathrm{RSS}_k^m(J^m)$ which has been done in Step 1, and on the recursive minimization

$$\begin{aligned}&\forall m \in [1 : M],\\ &\mathrm{RSS}_K(J^1, \ldots, J^m) = \min_{k'+k''=K} \left\{ \mathrm{RSS}_{k'}(J^1, \ldots, J^{m-1}) + \mathrm{RSS}_{k''}^m(J^m) \right\}.\end{aligned}$$

Complexity time. The first stage corresponds to the segmentation of individual series into a given number of segments whose complexity is $\mathcal{O}\left(\sum_m n_m^2 K_m\right)$. The complexity of the second stage is $\mathcal{O}\left(K^2 \times M\right)$ which makes the overall complexity of order

$\mathcal{O}\left(\sum_m n_m^2 K_m + K^2 M\right)$. Using dynamic programming on the whole data set would result in a complexity of $(\mathcal{N}^2\mathcal{K})$. If all series had the same length $n_m = n$ (and thus $N = Mn$) and were segmented into $K_m = k$ segments each (thus $K = Mk$) and assuming that $k = \lambda n$ (with $\lambda \ll 1$), the two-stage dynamic programming algorithm would have a complexity of $\mathcal{O}(\lambda M n^2[n + \lambda M^2])$ whereas the overall one would have a complexity $\mathcal{O}(\lambda M^3 n^3)$.

Complexity space. The complexity in space of the first step is $\mathcal{O}\left(\sum_m n_m^2\right)$ and it is $\mathcal{O}\left(\sum_m K_m\right)$ for the second step. Consequently, the space complexity is of order $\mathcal{O}(Mn(1+\lambda))$ to be compared with $\mathcal{O}(M^2n^2)$ for the traditional dynamic programming.

4. Model selection

The procedure proposed above leads to the joint segmentation into K segments which is unknown and needs to be estimated. We propose different model selection strategies for selecting this number by maximizing a penalized log-likelihood criterion. The case of univariate series segmentation has been intensively studied in the literature and needs to be generalized to joint segmentation. Here we consider the generalization of four criteria Lavielle (2005), Lebarbier (2005), Caussinus and Mestre (2004), Zhang and Siegmund (2007). To our knowledge the criterion proposed by Caussinus and Mestre (2004) is the only one directly adapted to the multivariate case:

$$\mathrm{Crit}_{\mathrm{CM}}(K) = \log\left[1 - \frac{\sum_{m=1}^{M}\sum_{k=1}^{K_m} \hat{n}_k(\hat{\mu}^K_{mk} - \hat{\mu}^0_{mk})^2}{\sum_{m=1}^{M}\sum_{t=1}^{n_m}(Y_{mt} - \hat{\mu}^0_{mk}(t))^2}\right] + \frac{2K}{N-n}\log N,$$

with $\hat{\mu}^K_{mk}$ and $\hat{\mu}^0_{mk}$ being the estimated mean of the data that belong to the segment k of series m for the segmentation with K and 1 segments respectively, and $\hat{n}^m_k$ being the length of segment k in series m ($\hat{n}^m_k = \hat{t}^m_k - \hat{t}^m_{k-1} + 1$).

Then adaptive criteria depend on some constants to be calibrated Lavielle (2005), Lebarbier (2005). The penalty proposed by Lavielle (2005) depends on D_K, the number of parameters in a model with K segments. Thus its multivariate version is straightforward:

$$\mathrm{Crit}_{\mathrm{Lav}}(K) = -2\log\mathcal{L}_K(\mathbf{Y}; \widehat{\phi}) + \beta D_K,$$

where $\log\mathcal{L}_K(\mathbf{Y}; \widehat{\phi})$ stands for the log-likelihood calculated at its maximum. For example $D_K = K + 2$ for model (2). The constant β is chosen using an adaptive method which involves a threshold s. As Lavielle suggests $s \in [0.5, 0.75]$, we used $s = 0.7$ throughout the simulation study.

The previous penalty considers the complexity of one model via D_K (*i.e*. the number of parameters to be estimated in this model) but does not consider that there exist many K-dimensional models. Let us recall that in the segmentation setting, a model is defined by a segmentation. Then the number of K-dimensional models is the number of all the possible segmentations with K segments that is $\binom{N-1}{K-1}$. This is considered by Lebarbier (2005) and the general form of the penalty is $c_1 D_K + c_2\log\left(\binom{N}{K}\right)$ where c_1 and c_2 have been calibrated up to a constant. In the multivariate case we make a parallel between the segmentation of M series of respective lengths n_m into K segments, and the segmentation of one series with length N ($\sum_m n_m$) into K segments, M breakpoints being fixed. Thus for joint segmentation the number of possible choices for the $K-M$ breakpoint positions among $N-M$ positions is $\binom{N-M}{K-M}$. The generalization of this criterion to joint segmentation is then

$$\mathrm{Crit}_{\mathrm{Leb}}(K) = -2\log\mathcal{L}_K(\mathbf{Y}; \widehat{\phi}) + \alpha\left[5D_K + 2(K-M)\log\left(\frac{N-M}{K-M}\right)\right].$$

This penalty also depends on a constant α which can be calibrated in practice Birgé and Massart (2007).

The criterion proposed by Zhang and Siegmund (2007) is a modified version of the classical BIC criterion. More precisely, the condition required for using the classical version is the differentiability of the likelihood with respect to the parameters. For segmentation models the breakpoints are discrete parameters which makes this condition not satisfied. A continuous-time version of the model has been proposed for deriving an efficient new BIC criterion Zhang and Siegmund (2007). The following proposition gives the generalization of this criterion in the multiple-series setting.

Proposition 1. *The modified BIC criterion proposed by Zhang and Siegmund (2007) for the selection of the number of breakpoints is generalized in the segmentation of M series context as follows:*

$$\mathrm{mBIC}_{\mathrm{JointSeg}}(K) = \left(\frac{N-K+1}{2}\right)\log\left[1 + \frac{SS_{\mathrm{bg}}(\hat{t})}{SS_{\mathrm{wg}}(\hat{t})}\right] + \log\left[\frac{\Gamma\left(\frac{N-K+1}{2}\right)}{\Gamma\left(\frac{N+1}{2}\right)}\right]$$
$$+ \frac{K}{2}\log(SS_{\mathrm{all}}) - \frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{k_m}\log\hat{n}^m_k + \left(\frac{1}{2} - (K-M)\right)\log(N),$$

where

$$SS_{\text{bg}} = \sum_{m=1}^{M}\sum_{k=1}^{k_m} \hat{n}_k^m (\bar{y}_{mk} - \bar{y})^2, \qquad SS_{\text{all}} = \sum_{m=1}^{M}\sum_{t=1}^{n_m} (y_m(t) - \bar{y})^2, \qquad SS_{\text{wg}} = SS_{\text{all}} - SS_{\text{bg}},$$

with $\hat{n}_k^m$ being the length of segment k in series m ($\hat{n}_k^m = \hat{t}_k^m - \hat{t}_{k-1}^m + 1$) and $\bar{y}_{mk} = \frac{1}{\hat{n}_k^m}\sum_{t=\hat{t}_{k-1}^m+1}^{\hat{t}_k^m} y_m(t)$.

Proof. The proof is only based on the adaptation of the prior distribution of the breakpoints denoted by $f(\tau)$ in Zhang (2006). Here $K - M$ breakpoints are spread on $[0, N]$ with uniform probabilities, so

$$f(\tau) = (K-M)!/N^{(K-M)}.$$

The rest of the derivation of the criterion follows the proof of Theorem 2.2 in Zhang (2006). □

Another possible form of this criterion with the likelihood term is

$$\text{mBIC}_{\text{JointSeg}}(K) = \left(\frac{N-K+1}{2}\right)\left[\frac{2}{N}\log \mathcal{L}_K(\mathbf{Y}; \widehat{\mathbf{T}\mu}) + 1 + \log(2\pi) - \log(N)\right] - \frac{1}{2}\sum_{m=1}^{M}\sum_{k=1}^{k_m}\log \hat{n}_k^m - (K-M)\log(N) + \log\left(\Gamma\left(\frac{N-K+1}{2}\right)\right).$$

$\log \mathcal{L}_K(\mathbf{Y}; \widehat{\mathbf{T}\mu})$ stands for the log-likelihood calculated at its maximum for a joint segmentation model with K segments. The penalty term depends not only on the number of breakpoints to be chosen but also on the length of the segments, favoring regularly spaced breakpoints.

Note that the dependency structure induced by the introduction of the random effect is not taken into account here. The simulations performed in the next section are used to demonstrate the quality of this approximation.

5. Simulation study

We study the performance of the proposed estimation procedure on synthetic data.

Model. We consider the following model:

$$\forall t \in I_k^m \quad Y_{mt} = \mu_{mk} + U_t + E_{mt}. \tag{2}$$

In this model no covariate is considered. The covariance matrices are set as $\mathbf{R} = \sigma_0^2\mathbf{I}$ and $\mathbf{G} = \sigma_u^2\mathbf{I}$. The introduction of random effect U_t is used to introduce correlation among series such that $\text{Cov}(Y_{mt}, Y_{m't}) = \sigma_u^2$. For this model, analytic formulas can be derived for the estimates which are given in the Appendix.

Simulation design. The length of the series is fixed at $n = 100$. For each series Y_m, the number of breakpoints $(K-1)$ is Poisson distributed with mean 2 and their positions are uniformly distributed. The mean value within each segment alternates between 0 and a value in $\{-2, -1, +1, +2\}$ with probability $\{0.1, 0.4, 0.4, 0.1\}$ respectively. We consider different numbers of series $M \in \{10, 50\}$, residual standard deviations $\sigma_0 \in \{0.1, 0.2, 0.5, 1\}$, and random effect standard deviations $\sigma_u \in \{0, 0.1, 0.2, 0.5, 1\}$. Each configuration is simulated 50 times. We first focus on the comparison of the four model selection criteria described in Section 4 for the choice of $\widehat{K}$. We then compare with the separate segmentation of each series.

Quality criteria. We consider criteria of two types for assessing the quality of the estimation. The first one is the root mean squared distance between the true mean and its estimate: $\text{RMSE}(\mu) = \left[(N)^{-1}\sum_m\sum_t(\widehat{\mu}_{mt} - \mu_{mt})^2\right]^{1/2}$. The second one is the proportion of erroneously detected breakpoints among detected breakpoints (false positive rate, FPR) and the proportion of undetected true breakpoints among true breakpoints (false negative rate, FNR). For each configuration we consider the average of these criteria over the 50 simulations.

Results. Figs. 1 and 2 represent the evolution of the criteria with respect to the different values of σ_0 (from easy to hard detection configurations) for $M = 50$ series (similar but less variable results are obtained for $M = 10$ series). The general conclusion is that our generalization of the mBIC criterion performs better for estimating the number of segments and so leads to better performance in terms of segmentation (Fig. 2).

The behavior of Crit_{Lav} is very erratic, selecting either too many segments or no breakpoint, particularly when the number of series is important. Then Crit_{CM} is the most sensitive to the random effect since the number of the segments that it selects is less than the true number and than other criteria. This trend is most marked for easy configurations. Both Crit_{CM} and Crit_{Lav} lead to bad performance in terms of segmentation (see $\text{RMSE}(\mu)$, Fig. 1). For easy configurations Crit_{ZS} selects segments whose number is close to the true one (whatever the variability of the random effect) whereas Crit_{Leb} tends to be more variable, in particular for many series (see Fig. 1). Consequently Crit_{ZS} leads to better segmentation estimates (see $\text{RMSE}(\mu)$, Fig. 1) for which the breakpoints are better positioned (smaller FPR and FNR; see Fig. 2). For noisy configurations fewer

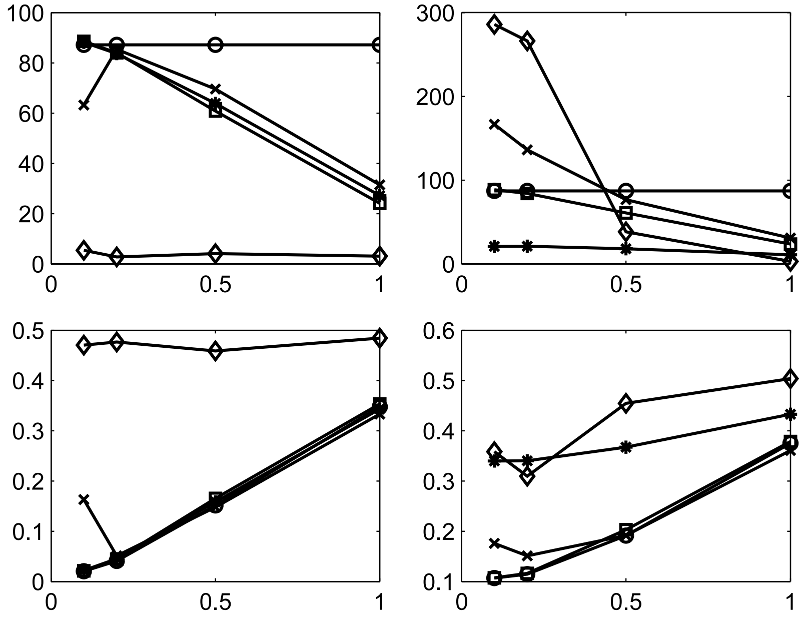

Fig. 1. Model selection procedure comparison according to the number of segments (top) and RMSE(μ) (bottom) with respect to σ_0 (on the x-axis) for $M = 50$ series. Left panel: $\sigma_u = 0.1$, right panel: $\sigma_u = 1$. ○: true number of segments, □: ZS, ×: Leb, ◇: Lav, ⋆: CM.

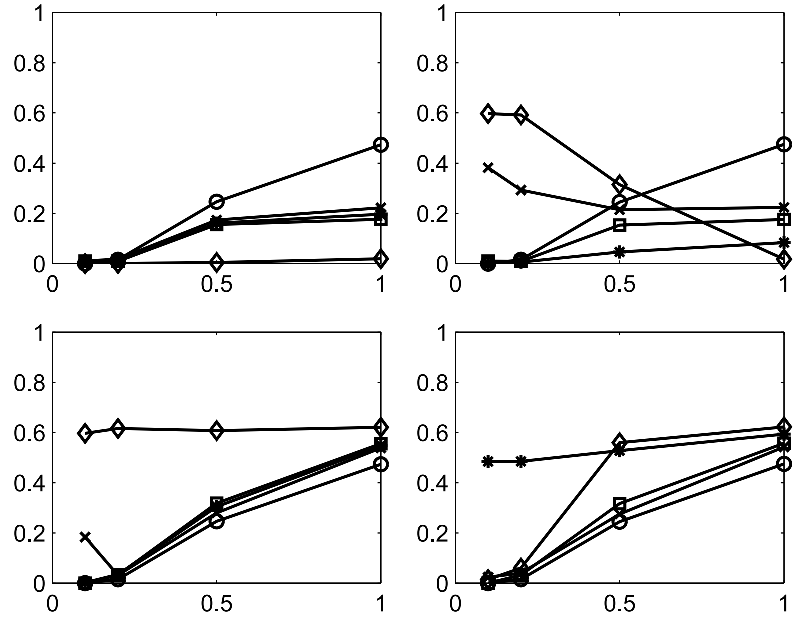

Fig. 2. Breakpoint detection performance according to the FPR (top) and FNR (bottom) with respect to σ_0 (on the x-axis) for $M = 50$ series. Left panel: $\sigma_u = 0.1$, right panel: $\sigma_u = 1$. ○: true number of segments, □: ZS, ×: Leb, ◇: Lav, ⋆: CM.

segments are selected by the different model selection procedures as compared with the true number, which was expected. Indeed in these situations small jumps in the mean are more difficult to detect and it may be preferable to ignore some of them. This is illustrated by the results obtained with the true number which slightly decreases the FNR but strongly increases the FPR compared to the results obtained with the Crit_{ZS} criterion (see Fig. 2). This means that the positioning of the breakpoints remains inaccurate in the presence of noise even when their true number is known. Best estimation qualities are again observed with Crit_{ZS} which also leads to accurate estimations of the variances σ_0 and σ_u (see Table 1 and Table 2). This criterion is then used in the application study.

Fig. 3 represents the evolution of the criteria as a function of σ_u for $M = 10$ and $M = 50$ series, and for a fixed $\sigma_0 = 0.2$. For both the joint and independent segmentation, the number of segments is selected by using the Crit_{ZS} criterion. As expected, independent segmentation does not improve when the number of series increases whereas joint segmentation does benefit from this increasing, especially in terms of breakpoint positioning (smaller FPR and FNR) and prediction of **U** (not shown). This is all the more marked as the variability of the random effect is large. In the absence of a random effect, the two strategies have similar performance. They still compare well when σ_u is very small ($\sigma_u < 0.2$). However for large variability of the random effect, the performance of the independent segmentation is dramatically degraded. Indeed in this case, too many

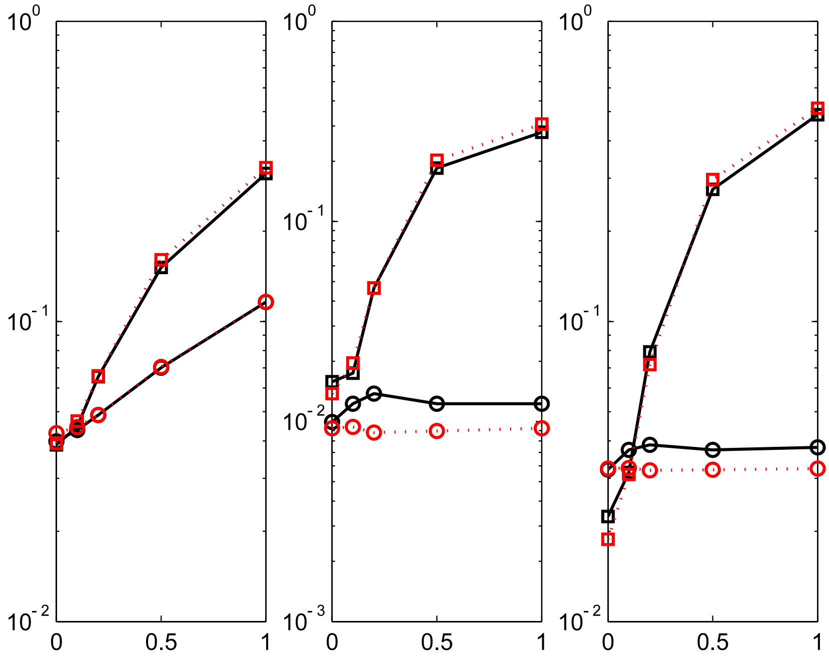

Fig. 3. RMSE(μ), FPR and FNR respectively (on a logarithm scale) with respect to σ_u (on the x-axis) for $M = 10$ series (solid line) and $M = 50$ series (dotted line). ○: joint segmentation, □: independent segmentation.

Table 1
Empirical confidence interval of σ_u. Top: $\sigma_u = 0.1$, bottom: $\sigma_u = 1$.

Crit	$M = 10$		$M = 50$	
	$\sigma_0 = 0.1$	$\sigma_0 = 1$	$\sigma_0 = 0.1$	$\sigma_0 = 1$
Ktrue	0.099 ± 0.001	0.19 ± 0.1	0.099 ± 0.0004	0.115 ± 0.04
Leb	0.102 ± 0.122	0.206 ± 0.117	0.1 ± 0.255	0.12 ± 0.059
ZS	0.099 ± 0.001	0.207 ± 0.113	0.099 ± 0.0004	0.122 ± 0.056
Lav	0.099 ± 0.001	0.2 ± 0.223	0.1 ± 0.168	0.129 ± 0.127
CM	0.099 ± 0.001	0.21 ± 0.119	0.099 ± 0.0004	0.121 ± 0.053
Crit	$M = 10$		$M = 50$	
	$\sigma_0 = 0.1$	$\sigma_0 = 1$	$\sigma_0 = 0.1$	$\sigma_0 = 1$
Ktrue	0.996 ± 0.009	0.98 ± 0.103	0.998 ± 0.0004	0.995 ± 0.039
Leb	0.989 ± 0.217	0.995 ± 0.128	0.986 ± 0.122	0.997 ± 0.06
ZS	0.997 ± 0.001	0.995 ± 0.129	0.998 ± 0.0004	0.998 ± 0.0579
Lav	0.886 ± 0.002	0.948 ± 0.134	0.92 ± 0.046	0.998 ± 0.128
CM	0.999 ± 0.073	0.999 ± 0.130	0.998 ± 0.045	0.998 ± 0.073

Table 2
Empirical confidence interval of σ_0. Top: $\sigma_0 = 0.1$, bottom: $\sigma_0 = 1$.

Crit	$M = 10$		$M = 50$	
	$\sigma_u = 0.1$	$\sigma_u = 1$	$\sigma_u = 0.1$	$\sigma_u = 1$
Ktrue	0.099 ± 0.0036	0.102 ± 0.3335	0.0987 ± 0.0036	0.099 ± 0.3308
Leb	0.156 ± 0.0081	0.209 ± 0.3428	0.296 ± 0.007	0.145 ± 0.3624
ZS	0.099 ± 0.0036	0.099 ± 0.33	0.099 ± 0.0036	0.099 ± 0.3308
Lav	0.098 ± 0.0036	0.095 ± 0.2664	0.497 ± 0.0063	0.109 ± 0.2697
CM	0.098 ± 0.004	0.307 ± 0.33	0.098 ± 0.004	0.33 ± 0.34
Crit	$M = 10$		$M = 50$	
	$\sigma_u = 0.1$	$\sigma_u = 1$	$\sigma_u = 0.1$	$\sigma_u = 1$
Ktrue	0.951 ± 0.02	0.959 ± 0.351	0.959 ± 0.0084	0.995 ± 0.36
Leb	1.003 ± 0.025	1.017 ± 0.358	1.013 ± 0.0088	0.997 ± 0.356
ZS	1.004 ± 0.025	1.017 ± 0.359	1.026 ± 0.0088	0.998 ± 0.358
Lav	0.959 ± 0.03	1.014 ± 0.343	1.027 ± 0.0109	1.101 ± 0.359
CM	1.003 ± 0.023	1.046 ± 0.35	1.02 ± 0.009	1.06 ± 0.36

breakpoints are selected for capturing this effect, leading to a bad quality of the positioning of the breakpoints. To conclude, joint segmentation should be preferred as it provides more reliable results.

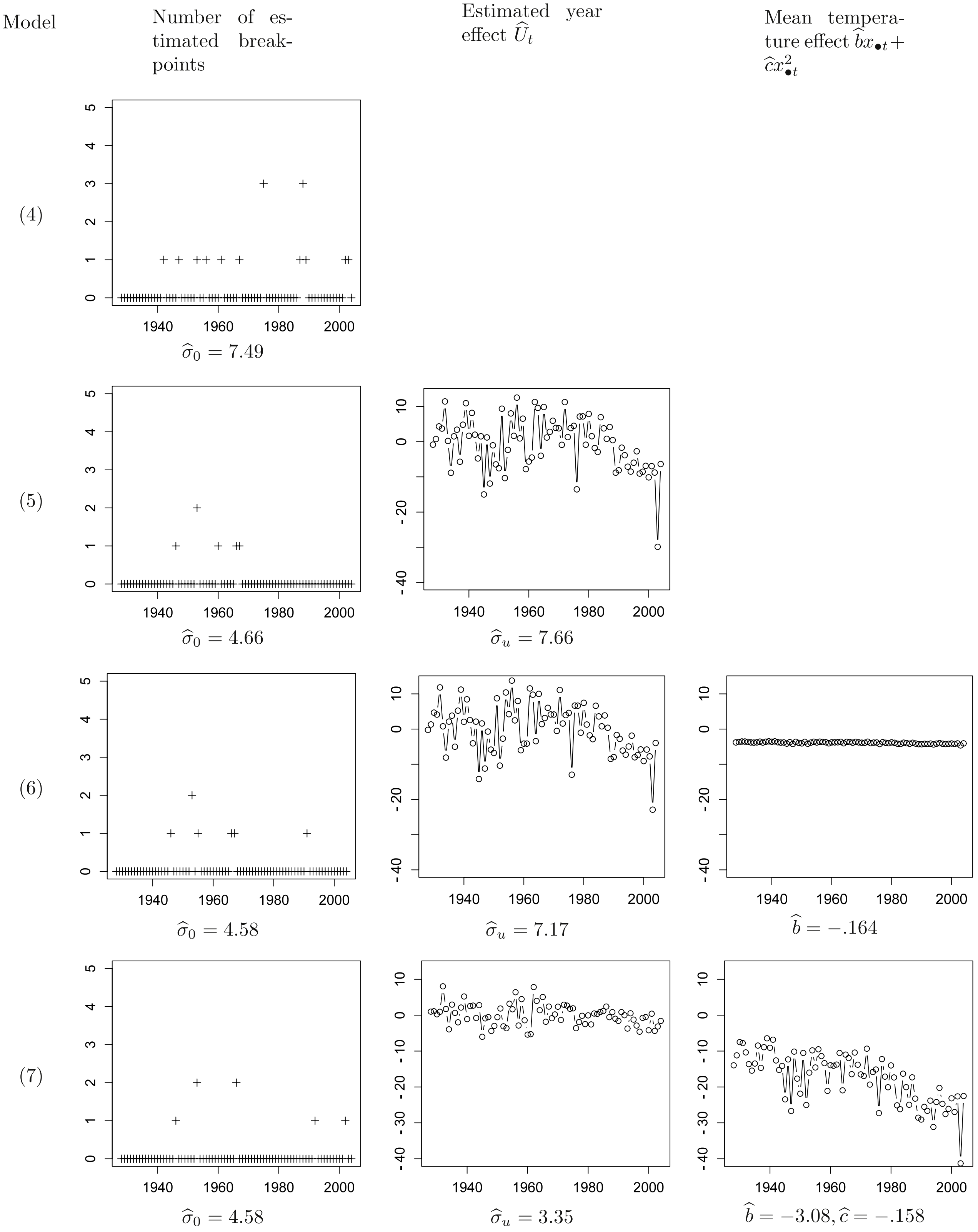

Fig. 4. Estimated effects considered in models (4), (5), (6), (7).

6. Application to harvest date series

We are interested here in the detection of changes in grape harvest dates across several French regions. The aim is to detect changes in terms of agricultural practices that may affect the harvest date in specific stations. As harvesting is done earlier in hot years, we typically want to distinguish station-specific results from those due to variations of the climate. The harvest dates Y_{mt} and the mean temperature x_{mt} from April to August were recorded for $M = 10$ French stations for

$t = 1928 \cdots 2004$. The climatic data have been provided by O. Mestre from Météo—France and are detailed in Caussinus and Mestre (2004). The harvest data have been provided by I. Chuine and I. Garcia from CEFE-CNRS, and V. Daux, P. Yiou and N. Viovy from LSCE-CEA-CNRS. For all analysis, the number of segments was estimated using the mBIC model selection criterion described in Section 4.

Year effect. To distinguish station-specific breakpoints from changes due to some general year effect, we first compared the pure segmentation model

$$Y_{mt} = \mu_{mk} + E_{mt}, \tag{3}$$

with the mixed model including a random year effect

$$Y_{mt} = \mu_{mk} + U_t + E_{mt}. \tag{4}$$

Throughout this study, we set $\mathbf{R} = \sigma_0^2 \mathbf{I}$ and $\mathbf{G} = \sigma_u^2 \mathbf{I}$. Fig. 4 displays the total number of breakpoints observed each year across stations. Five breakpoints were found with model (3) in 1988 (± 1 year). A slight negative linearity was observed in the harvest dates from this year until the end, showing that, in recent years, grapes were harvested earlier every year. In addition, several breakpoints were found near 1976 and 2003, each corresponding to a very hot summer. As shown in Fig. 4, both this trend and year-specific effects were captured by the random year effect in model (4). This results in a dramatic decrease in the number of breakpoints and in the estimated residual standard deviation, which was found smaller than the estimated random effect standard deviation.

Modeling climate effects. As the year effect seemed to be related to climatic trends or events, we introduced the mean temperature x_{mt} as a covariate:

$$Y_{mt} = \mu_{mk} + b x_{mt} + E_{mt}. \tag{5}$$

As shown in Fig. 4 the effect of this covariate was found to be very small ($\hat{b} = -0.164$) and the year effect estimated with model (5) was very similar to that of model (4). A possible explanation is that moderate temperature variations do not affect the harvest dates, whereas exceptional variations do. We then added a quadratic term to account for the specific effect of large variations:

$$Y_{mt} = \mu_{mk} + b x_{mt} + c x_{mt}^2 + E_{mt}. \tag{6}$$

This later climatic effect modeling turned out to provide a good description of what was first captured by the random year effect. Both the linear trend and the exceptionally hot years were captured by the climatic effect. This results in a dramatic reduction of the estimated random effect variance.

Model (6) allowed us to detect station-specific breakpoints. It also provided insights into the climate variation underlying the random year effect. As an illustration, the breakpoint found in 1953 occurs in two neighboring stations and where the same grape variety is grown.

Conclusion

In this work we developed a linear mixed model for the joint segmentation of Gaussian series by taking into account correlations that could exist across series. The random part of the model could be considerably enriched. For instance, when studying climate series, stations are spread among different regions, and it is likely that spatial correlations exist between series. Such structure could be integrated in the random part of the model. Our second contribution lies in the computational framework that we propose. This two-stage dynamic programming schema constitutes an answer to the computational limitations faced by previous studies. We also provide some directions for the development of model selection criteria for selecting the number of segments in a joint segmentation model. The criteria that we propose work well in practice, but new theoretical developments are needed to integrate the random effect from a theoretical point of view.

Appendix

At the iteration $(h+1)$ of the EM algorithm, we get the BLUP and the variance estimates respectively:

$$\hat{U}_t^{(h+1)} = \frac{1}{M + \lambda^{(h)}} \sum_{m=1}^{M} (Y_{mt} - \mu_{mk}^{(h)}), \quad \text{if } t \text{ belong to segment } I_k^m,$$

$$\sigma_u^{(h+1)\,2} = \frac{1}{n_{\max}} \left(\sum_{t=1}^{n_{\max}} \hat{U}_t^{(h+1)2} + \frac{\sigma_0^{(h)\,2}}{n_{+t} + \lambda^{(h)}} \right)$$

and

$$\sigma_0^{(h+1)\,2} = \frac{1}{N} \left(\sum_{m,t} \hat{E}_{mt}^{(h+1)2} + \sigma_0^{(h)\,2} \left(n_{\max} + \sum_{t=1}^{n_{\max}} \frac{\lambda^{(h)}}{n_{+t} + \lambda^{(h)}} \right) \right),$$

where n_{+t} is the number of patients with position t being recorded, $\hat{E}^{(h+1)} = Y_{mt} - \mu_{mk}^{(h)} - \hat{U}_t^{(h+1)}$ if t belongs to segment I_k^m and $\lambda^{(h)} = {\sigma_0^{(h)}}^2 / {\sigma_u^{(h)}}^2$.

References

Bai, J., Perron, P., 2003. Computation and analysis of multiple structural change models. Journal of Applied Economics 18, 1–22.

Birgé, L., Massart, P., 2007. Minimal penalties for Gaussian model selection. Probability Theory and Related Fields 138, 33–73.

Caussinus, H., Mestre, O., 2004. Detection and correction of artificial shifts in climate series. Journal of the Royal Statistical Society: Series C 53 (3), 405–425.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B 39, 1–38.

Dobigeon, N., Tourneret, J.-Y., 2007. Joint segmentation of wind speed and direction using a hierarchical model. Computational Statistics and Data Analysis 51, 5603–5621.

Fearnhead, P., 2006. Exact and efficient Bayesian inference for multiple change-point problems. Statistics and Computing 16, 203–213.

Lavielle, M., 2005. Using penalized contrasts for the change-point problem. Signal Processing 85 (8), 1501–1510.

Lavielle, M., Lebarbier, E., 2001. An application of MCMC methods for the multiple change-points problem. Signal Processing 81, 39–53.

Lebarbier, E., 2005. Detecting multiple change-points in the mean of Gaussian process by model selection. Signal Processing 85, 717–736.

Meng, X.-L., Rubin, D., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80 (2), 267–278.

Picard, F., Robin, S., Lavielle, M., Vaisse, C., Daudin, J.-J., 2005. A statistical approach for array CGH data analysis. BMC Bioinformatics 6 (27), 1.

van Dyk, D., 2000. Fitting mixed-effects models using efficient EM-type algorithms. Journal of Computational and Graphical Statistics 9, 78–98.

Zhang, N.R., 2006. Change-point detection and sequence alignment: statistical problems of genomics, Ph.D. Thesis, Stanford University.

Zhang, N.R., Siegmund, D.O., 2007. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. Biometrics 63 (1), 22–32.