A breakpoint detection in the mean model with heterogeneous variance on fixed time intervals

Olivier Bock · Xavier Collilieux · François Guillamon · Emilie Lebarbier · Claire Pascal

Received: 17 July 2018 / Accepted: 26 January 2019 / Published online: 3 May 2019

Abstract
This work is motivated by an application for the homogenization of global navigation satellite system (GNSS)-derived integrated water vapour series. Indeed, these series are affected by abrupt changes due to equipment changes or environmental effects. The detection and correction of the series from these changes are a crucial step before any use for climate studies. In addition to these abrupt changes, it has been observed in the series a non-stationary of the variability. We propose in this paper a new segmentation model that is a breakpoint detection in the mean model of a Gaussian process with heterogeneous variance on known time intervals. In this segmentation case, the dynamic programming algorithm used classically to infer the breakpoints cannot be applied anymore. We propose a procedure in two steps: we first estimate robustly the variances and then apply the classical inference by plugging these estimators. The performance of our proposed procedure is assessed through simulation experiments. An application to real GNSS data is presented.

Keywords Segmentation · Robust estimator of the variance parameter · GNSS time series

1 Introduction

Breakpoint detection aims at detecting abrupt changes, called breakpoints, in the distribution of a signal. Such problems arise in many fields, such as genomics (Braun et al. 2000; Picard et al. 2005; Cleynen et al. 2014; Lévy-Leduc et al. 2014), medical (Lavielle 2005), econometrics (Lai et al. 2005a; Lavielle 1999; Bai and Perron 2003), geodesy (Gazeaux et al. 2013, 2015) or climate (Caussinus and Mestre 2004; Mestre et al. 2013; Lue et al. 2010). This massive number of applications results in the abundant literature on this subject. The motivation of our work comes from the analysis of Global Navigation Satellite System (GNSS)-derived integrated water vapour (IWV) series. Such series are obtained from the analysis of GNSS measurements which are continuously recorded at permanent GNSS stations in a global observing network (www.igs.org). The IWV quantity is an essential climate variable which is of significant importance for monitoring climate change (Trenberth et al. 2005; Parracho et al. 2018). However, the GNSS IWV series have been affected by inhomogeneities due to equipment changes, changes in processing procedure and/or changes in electromagnetic properties of the environment at the measurement site (Vey et al. 2009; Ning et al. 2016). These inhomogeneities usually produce an abrupt change in the mean IWV data which may disturb the detection of the underlying natural climate variability and trends.

Up to now, no automatic method exists to determine the position of breakpoints in IWV series. The most commonly used approach in climatology is the relative abrupt change detection which compares the candidate series to one or several reference series (e.g. from nearby stations) which are assumed to contain nearly the same climate signal (Lindau
and Venema 2013; Caussinus and Mestre 2004). In the case of our application, the stations in the global GNSS network which have long records are usually too far from each other to remove completely the climate signal in the differences. Instead, we extract the reference time series for each station from a gridded global atmospheric model reanalysis. The resulting difference time series (GNSS minus reanalysis) can to a good degree be modelled by an identically and independently distributed (IID) Gaussian series (more details about the data and discussion of this assumption are provided in Sect. 5).

The statistical purpose consists thus in detecting the times at which the mean changes in the process. Many approaches have been proposed in the literature about this problem. Among them, we focus on segmentation methods. More precisely, the model of interest will be the following: the signal is supposed to be a realization of an independent Gaussian process whose parameters are affected by an unknown number of changes at unknown times. Two models can be considered, according to the characteristics of the signal that are affected by the changes: it can be either the mean of the signal only (usually called the homoscedastic model) or both the mean and the variance (usually called the heteroscedastic model), as proposed by Picard et al. (2005) in a genomic application field or by Gazeaux et al. (2015) in a geodesic application for the analysis of GPS GNSS coordinates series.

However, in the GNSS minus reanalysis series, it has been observed a non-stationarity of the variance due to increased variability of IWV in summer (see Sect. 5). Inspection of the annual variation in the series shows that a monthly sampling of the variance will be adequate. Consequently, the two above models will fail.

In this paper, we thus propose a new segmentation model that is a breakpoint detection in the mean of a Gaussian process model for which the variance is supposed to be piecewise stationary on fixed (and known) intervals. For the GNSS application, the latter intervals will correspond to the months.

It is now well known in segmentation framework that segmentation raises algorithmic issues due to the discrete nature of the breakpoint parameters. Indeed, the inference of these parameters requires to visit the whole segmentation space, which is prohibitive in terms of computational time when the visit is performed in a naïve way. The Dynamic Programming (DP) algorithm (introduced by Bellman 1954 and used for the first time in segmentation by Auger and Lawrence 1989) is the only efficient algorithm that retrieves the exact solution (i.e. the optimal segmentation according to the log-likelihood or least square contrasts for example) in a faster way. However, this algorithm can only be used if the quantity to be optimized is segment additive (see for example Bai and Perron 2003 or Lavielle 2005). In other words, a sufficient condition to satisfy this assumption is the fact that the segments are not linked both in terms of observations (i.e. independence) and parameters (i.e. no common parameters). In our case, the both stationary time intervals of the means and the variances do not coincide. Two problems will appear: first, the estimators of these two parameters will be linked, and then, we have no hope that DP can be applied. In order to circumvent this problem and retain the use of DP, we consider the same inference strategy as in Chakar et al. (2015) or Cleynen and Robin (2014) which consists in a two-step procedure: we first estimate the ‘nuisance’ parameters (here the variances) and then we apply the classical inference procedure by plugging these estimators.

The problem is thus reduced to the estimation of the variance parameter in a series with changes in the mean. Due to the presence of breakpoints in the series, the classical estimators for the variance will fail. Here, we follow the same idea as in Chakar et al. (2015) who proposed a robust estimator of the autocorrelation parameter for estimating breakpoints in the mean of an AR(1) process. Briefly speaking, instead of using the raw series, the idea is to work with the differentiated series that is then a zero-mean Gaussian process except at the position of the breakpoints. These points can be then seen as outliers, and a robust approach can be used to obtain a good estimator of the scale parameter, as Rousseeuw and Croux (1993) proposed. We adapt in particular this estimator to our case for which, using the results of Lévy-Leduc et al. (2011), we obtain asymptotic properties.

For the second step of the inference, if DP can be applied to obtain the best segmentation of the series in a given number of segments, the question arises of the choice of this number. This question has been widely investigated. In this paper, we propose to adapt the criteria proposed by Lavielle (2005), Lebarbier (2005) and Zhang and Siegmund (2007).

This paper is organized as follows: Sect. 2 presents the proposed segmentation model, describes the algorithmic issue for the inference and gives the outline of the proposed inference strategy. The details of this strategy are given in Sect. 3. More precisely, the robust estimator of the variance and the different model selection criteria for choosing the number of segments are given. A simulation study is performed in Sect. 4 and Sect. 5 is dedicated to an application of our method on GNSS-derived IWV series.

2 Model and inference issue

2.1 Model

We observe a series \( y = \{ y_t \}_{t=1,...,n} \) modelled by a Gaussian independent random process \( Y = \{ Y_t \}_{t=1,...,n} \) such that

- the mean of \( Y \) is affected by \( K - 1 \) abrupt changes at some unknown times in the GNSS application, called breakpoints, \( 0 = t_0 < t_1 < \cdots < t_{K-1} < t_K = n \) and is
constant between two breakpoints or within the interval \( I_k^{\text{mean}} = [t_{k-1} + 1, t_k] \), denoted segment, and

* the variance of \( Y \) is constant within each interval \( I_j^{\text{var}} \) and different from one to another. The time intervals \( I_j^{\text{var}} \) are fixed.

The model is thus the following:

\[
Y_t \sim \mathcal{N}(\mu_k, \sigma_j^2) \quad \forall t \in I_k^{\text{mean}} \cap I_j^{\text{var}},
\]

for \( k = 1, \ldots, K \) with \( K \) is the number of segments or intervals \( I_k^{\text{mean}} \) and for \( j = 1, \ldots, J \) with \( J \) is the number of intervals \( I_j^{\text{var}} \). This model assumes that the variance is not constant, compared to the homoscedastic model where \( Y_t \sim \mathcal{N}(\mu_k, \sigma_k^2) \). Compared to the heteroscedastic model where both the mean and the variance are subject to the same changes, i.e. \( Y_t \sim \mathcal{N}(\mu_k, \sigma_j^2) \), the intervals of stationary of the variance \( I_j^{\text{var}} \) do not coincide to the intervals \( I_k^{\text{mean}} \), the intervals \( I_j^{\text{var}} \) being fixed.

### 2.2 Segmentation inference: an algorithmic issue

Parameter inference in model (1) amounts to estimating the number of segments \( K \), the breakpoints \( \mathbf{t} = (t_k)_{k=1}^{K-1} \) and the distribution parameters, i.e. the means \( \mu = (\mu_k)_{k=1}^{K} \) and the variances \( \sigma^2 = (\sigma_j^2)_{j=1}^{J} \). To this end, we use a (penalized) maximum likelihood framework and proceed as classically in segmentation inference in three steps: (i) estimate the distribution parameters the breakpoints and their number being fixed, (ii) estimate the breakpoints for a fixed \( K \) and (iii) choose the number of segments \( K \).

The log-likelihood of model (1) is

\[
\log p(y; \mathbf{t}, \mathbf{\mu}, \sigma) = -\frac{n}{2} \log (2\pi) - \sum_{j=1}^{J} \frac{n_j}{2} \log (\sigma_j^2) - \frac{1}{2} SSw_{gK}(\mathbf{t}, \mathbf{\mu}, \sigma^2),
\]

where

\[
SSw_{gK}(\mathbf{t}, \mathbf{\mu}, \sigma^2) = \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{t \in I_k^{\text{mean}} \cap I_j^{\text{var}}} \frac{(y_t - \mu_k)^2}{\sigma_j^2},
\]

and \( n_j \) is the length of interval \( I_j^{\text{var}} \). Recall that in the segmentation framework, it is now well known that the step (ii) leads to a discrete optimization problem and that the only efficient algorithm that retrieves the solution (exact solution in a fast way) is the Dynamic Programming algorithm (DP). This algorithm can be applied under the constraint that the quantity to be optimized is additive with respect to the segments or intervals \( I_k^{\text{mean}} \) (see, for example, Bai and Perron 2003; Picard et al. 2005 or Lavielle 2005). Here the optimization problem for breakpoint estimation is

\[
\min_{t \in \mathcal{M}_{K,n}} \min_{\mu \in \mathbb{R}^K} \min_{\sigma \in \mathbb{R}^J} -\log p(y; \mathbf{t}, \mathbf{\mu}, \sigma) = \min_{t \in \mathcal{M}_{K,n}} -\log p(y; \mathbf{t}, \hat{\mathbf{\mu}}, \hat{\sigma}),
\]

where \( \log p(y; \mathbf{t}, \mathbf{\mu}, \sigma) \) is given in (2) and \( \mathcal{M}_{K,n} = \{(t_1, \ldots, t_{K-1}) \in \mathbb{N}^{K-1}, 0 = t_0 < t_1 < \ldots, t_{K-1} < t_K = n \} \) is the set of all possible partitions in \( K \) segments of the grid \([1, n] \). The carriers of the mean parameters and the variance parameters are not the same, \( I_k^{\text{mean}} \) for \( \mu_k \) and \( I_j^{\text{var}} \) for \( \sigma_j^2 \). In this case, the only efficient algorithm (the DP algorithm) to infer the breakpoints cannot be used. Indeed, the estimators of the mean and the variance parameters are

\[
\hat{\mu}_k = \frac{\sum_{j=1}^{J} \sum_{t \in I_k^{\text{mean}} \cap I_j^{\text{var}}} y_t / \sigma_j^2}{\sum_{j=1}^{J} \sum_{t \in I_k^{\text{mean}} \cap I_j^{\text{var}}}},
\]

\[
\hat{\sigma}_j^2 = \frac{1}{n_j} \sum_{k=1}^{K} \sum_{t \in I_k^{\text{mean}} \cap I_j^{\text{var}}} (y_t - \hat{\mu}_k)^2,
\]

and thus, the quantity to be optimized \( -\log p(y; \mathbf{t}, \hat{\mathbf{\mu}}, \hat{\sigma}) \) will be not segment additive that is the required condition to use DP. In order to keep possible the use of this algorithm, we propose an other inference strategy described in the next section.

### 3 Inference procedure

We consider the same strategy as proposed by Chakar et al. (2015) or Cleynen and Robin (2014) which consists in

1. estimating the variance parameters (see Sect. 3.1), the estimators are denoted \( \hat{\sigma}_j^2 \),
2. using the classical inference with ‘known’ variances. In this case,

* the mean estimators are the same as (4) where \( \hat{\sigma}_j^2 \) is replaced by \( \hat{\sigma}_j^2 \),
* the optimization problem for breakpoint estimation is

\[
\min_{t \in \mathcal{M}_{K,n}} -\log p(y; \mathbf{t}, \hat{\mathbf{\mu}}, \hat{\sigma}) = \min_{t \in \mathcal{M}_{K,n}} \sum_{k=1}^{K} \sum_{j=1}^{J} \sum_{t \in I_k^{\text{mean}} \cap I_j^{\text{var}}} \frac{(y_t - \hat{\mu}_k)^2}{\hat{\sigma}_j^2}
\]

\[
= \min_{t \in \mathcal{M}_{K,n}} SSw_{gK}(\hat{\mathbf{t}}, \hat{\mathbf{\mu}}, \hat{\sigma}^2),
\]

\[
= SSw_{gK}(\hat{\mathbf{t}}, \hat{\mathbf{\mu}}, \hat{\sigma}^2),
\]
and DP applied. Note that these two steps are done, $K$ being fixed.

- the number of segments $K$ is chosen according to a model selection strategy which consists in maximizing a penalized log-likelihood (see Sect. 3.2).

### 3.1 A robust estimator of the variance parameters

For a sake of simplicity, let us consider that the variance of the process $Y$ is homogeneous, i.e. in model (1) we have $J = 1$, $\sigma_j^2 = \sigma^2$ whatever $j$ (i.e. $I_j^\text{car} = [1, n]$) and the purpose is to estimate $\sigma$. Since we need to estimate it in a series with breakpoints, classical estimators failed. The objective is to provide a robust (faced to the presence of breakpoints) estimator of $\sigma$. Following Chakar et al. (2015), the idea is to work on the differentiated series $\{X_t\} = \{\nabla Y_t\} = \{Y_t - Y_{t-1}\}$, where $\nabla$ stands for the differentiation operator. Indeed, the means of this novel series is equal to 0 except at the breakpoint positions (only $K - 1$ ($K \ll n$) differences are non-centred). These latter breakpoints can then be seen as outliers, and robust approaches can be used to estimate $\sigma$. Rousseeuw and Croux (1993) proposed a robust estimator of the scale parameter of an independent Gaussian stationary process $X$ that is proportional to the first quartile of the $n^2$ differences $\{X_i - X_j; 1 \leq i < j \leq n\}$, i.e.

$$Q_{CR,n}(X) = c_Q \{\{X_i - X_j; 1 \leq i < j \leq n\}\left(\frac{\sqrt{3}}{2}\right)^2\},$$

(6)

with

$$c_Q = \frac{1}{\sqrt{2\Phi^{-1}\left(\frac{3}{8}\right)}} \approx 2.2191,$$

(7)

to ensure the consistency of the estimator and where $\Phi$ denotes the cumulative distribution function of a standard Gaussian random variable. The asymptotic properties of this estimator have been studied by Lévy-Leduc et al. (2011) for Gaussian stationary processes with either short-range or long-range dependence.

Using this estimator, the robust estimator $\hat{\sigma}$ we proposed in our context is given in Proposition 1 for which asymptotic properties are obtained.

**Proposition 1** Let $\{Y_t\}$ and $\{E_t\}$, such that $Y_t = \mu_k + E_t$ if $t \in I_k = [t_{k-1} + 1, t_k]$ for $k = 1, \ldots, K$ where $\{E_t\}$ are i.i.d centred Gaussian with variance $\sigma^2$, and let further assume that $Y_0 \sim \mathcal{N}(\mu_1, \sigma^2)$. Let denote $X = \{X_t\}_{t=1,\ldots,n} = \{
abla Y_t\}_{t=1,\ldots,n} = \{Y_t - Y_{t-1}\}_{t=1,\ldots,n}$ and $\{V_t\}_{t=1,\ldots,n} = \{
abla E_t\}_{t=1,\ldots,n}$, where $\nabla$ stands for the differentiation operator. Let

$$\hat{\sigma} = Q_n(X) = \frac{Q_{CR,n}(X)}{\sqrt{2}},$$

(8)

where $Q_{CR,n}(X)$ is given in (6). Then, $Q_n$ satisfies the following Central Limit Theorem

$$\sqrt{n}(Q_n(X) - \sigma) \overset{d}{\rightarrow} \mathcal{N}(0, \sigma^2), \text{ as } n \rightarrow \infty,$$

where

$$\sigma^2 = \sigma^2 \mathbb{E}\left[IF^2\left(\frac{v_0}{\sqrt{2}\sigma}\right)\right] + 2\sigma^2 \sum_{h \geq 1} \mathbb{E}\left[IF(\frac{v_0}{\sqrt{2}\sigma})IF(\frac{v_h}{\sqrt{2}\sigma})\right],$$

$$IF(x) = c_Q \frac{1/4 - \Phi(x + 1/c_Q) + \Phi(x - 1/c_Q)}{\int_{-\infty}^{x} \Phi(y)\Phi(y + 1/c_Q)dy},$$

and where $\Phi$ and $\phi$ denote the cumulative distribution function and the probability distribution function of a standard Gaussian random variable, respectively.

The proof of Proposition 1 is given in ‘Appendix’.

Note that by working on the differentiated series $\{X_t\}_{t=1,\ldots,n}$, the independence is lost, but it is of short range and remark that $X_t$ is a Gaussian process with variance $2\sigma^2$ (that explained the normalization by $\sqrt{2}$ in (8)).

Come back to our segmentation model (1) and using Proposition 1, the proposed estimator of $\sigma_j$ is

$$\hat{\sigma}_j = Q_n(X^{(j)}),$$

(9)

where $X^{(j)} = \{X_t^{(j)}\}_{t=1,\ldots,n} = \{Y_t - Y_{t-1}\}_{t=1,\ldots,n}$. We note $\hat{\sigma} = (\hat{\sigma}_j)_j$.

### 3.2 Selecting the number of segments

In order to select the number of segments $K$, we consider three criteria proposed by Lavielle (2005), Lebarbier (2005) and Zhang and Siegmund (2007). We use these criteria forgetting the fact that $\hat{\sigma}_j^2$ has been estimated in a first step. The first two criteria are penalized contrast criteria which differ from the form of the penalty and depend on constants to be calibrated contrary to the last one. The penalty proposed by Lavielle (2005), denoted Lav, depends on the number of parameters in a model with dimension $K$ (i.e. a segmentation with $K$ segments) denoted $D_K$. It is defined as follows:

$$\text{Lav}(K) = SS_{W_K}(\hat{\theta}, \mu, \hat{\sigma}^2) + \beta D_K,$$

(10)

where $SS_{W_K}(t, \mu, \sigma^2)$ is the sum of squares given in (3). The constant $\beta$ is chosen using an adaptive method which involves a threshold $s$, taken in the simulation study and the applications to $s = 0.7$ as suggested by Lavielle (2005).
Fig. 1 Boxplots of $\tilde{\sigma}_1 - \sigma_1^*$ in black and $\tilde{\sigma}_2 - \sigma_2^*$ in grey for different values of $\sigma_2^*$ with $n = 200$ (left) and $n = 800$ (right). $x$-axis: $\sigma_2^*$.

Fig. 2 Results for $n = 200$ and for different values of $\sigma_2^*$, obtained for MFixedHetero (in which $\sigma_1^*$ and $\sigma_2^*$ are estimated) and the different model selection criteria: a boxplots of $\hat{K} - K^*$; b boxplots of the RMSE($\mu$); c boxplots of the distance $d_1$; d boxplots of the distance $d_2$. $x$-axis: $\sigma_2^*$. (Color figure online)

Applying the works of Birg and Massart (2001) in the segmentation context, Lebarbier (2005) proposed a more complex penalty in which, in addition to $D_K$, the number of possible segmentations with $K$ segments (that is $\binom{n-1}{K-1}$) is taken into account for. This criterion is denoted BM and is defined as follows:
Fig. 3 Similar to Fig. 2a when the true values of $\sigma_1^*$ and $\sigma_2^*$ are used
$x -$ axis : $\sigma_2^*$. (Color figure online)

$BM(K) = SS_{wgK}(\hat{t}, \hat{\mu}, \hat{\sigma}^2) + \alpha \left[ 5D_K + 2K \log \left( \frac{n}{K} \right) \right].$

(11)

This penalty also depends on a constant $\alpha$ which can be calibrated in practice using the slope heuristic method proposed in Arlot and Massart (2009). More precisely, there exist two algorithms based on this heuristic: the dimension jump algorithm and the data-driven slope estimation algorithm. We use for the simulations and the application the package R capushe (Baudry et al. 2011) and denote BM1 and BM2 the criteria BM where the constant is calibrated using these two algorithms, respectively. Note that these criteria have to be minimized.

The last criterion is a modified version of the classical BIC criterion (Schwarz 1978) adapted by Zhang and Siegmund (2007) to the segmentation in the mean with homogeneous variance framework and so-called mBIC. Two versions are derived depending on the knowledge or not of the variance. Here, we considered the one for which the variance is supposed to be known, denoted mBIC,

$mBIC(K) = -\frac{1}{2}SS_{wgK}(\hat{t}, \hat{\mu}, \hat{\sigma}^2) - \frac{1}{2} \sum_{k=1}^{K} \log (\hat{n}_k)$

$+ \left( \frac{3}{2} - K \right) \log (n),$

(12)

where $\hat{n}_k = \hat{t}_k - \hat{t}_{k-1}$ is the length of the $k$th segment of the best segmentation with $K$ segments (i.e. of $\hat{t}$). Note that this criterion has to be maximized.

Fig. 4 Frequencies of each possible breakpoint for MFixedHetero when the number of segments is selected with mBIC and the variances are estimated (left), when the number of segments is selected with mBIC and the variances are the true values (middle) and when the number of segments is true ($K = 7$) and the variances are estimated (right), with $n = 200$. The value of $\sigma_2^*$ is fixed to 0.1 (a), 0.5 (b) and 1.5 (c). The dotted lines correspond to the true breakpoint locations. $x$-axis: time
4 Simulation study

In order to assess the performance of our procedure, we conduct the simulation study described below. Note that we indicated the true parameters by $\star$.

4.1 Simulation design and quality criteria

We use a similar design as in Chakar et al. (2015) for the segmentation parameters (breakpoint locations and means) and mimic our motivation application in the sense that the series include several years and the variance time inter-
Fig. 7 Frequencies of each possible breakpoint for MHetero when the number of segments is selected with the criteria mBIC (left), Lav (middle) and BM2 (left), with $n = 200$. The value of $\sigma^*_2$ is fixed to 0.1 (a), 0.5 (b) and 1.5 (c). The dotted lines correspond to the true breakpoint locations, and the changes of variances are fixed at locations 25, 50, 75, 100, 125, 150, 175. x-axis: time

vals correspond to the months. We consider series of length $n \in \{200, 800\}$ with 4 years of $n/4$ points each and 2 time intervals for variance by year with standard deviation $\sigma^*_1$ and $\sigma^*_2$, respectively. $\sigma^*_1$ is fixed to 0.5, and $\sigma^*_2$ varies from 0.1 to 1.5 by step of 0.2. The series are affected by 6 breakpoints ($K^* = 7$, the true number of segments) located at positions $t^* = (27, 38, 88, 111, 150, 183)$ for $n = 200$ and $4 \times t^*$ for $n = 800$. The mean within each segment alternates between 0 and 1, starting with $\mu_1 = 0$. Each configuration is simulated 100 times.

Furthermore, we compare our model, called MFixed-Hetero, to the two classical segmentation models (the homoscedastic and the heteroscedastic segmentation models, see Picard et al. 2005). The good performance of these two segmentation methods has been shown in the comparison study of many segmentation methods proposed by Lai et al. (2005b) in a genomic application context. The corresponding models are the following:

* changes in the mean with homogeneous variance, called MHomo:
  $$Y_{t \text{ ind.}} \sim N(\mu_k, \sigma^2) \text{ if } t \in I_k = [t_{k-1} + 1, t_k] \text{ for } k = 1, \ldots, K,$$
* changes in both the mean and the variance, called MHetero:
  $$Y_{t \text{ ind.}} \sim N(\mu_k, \sigma^2_k) \text{ if } t \in I_k = [t_{k-1} + 1, t_k] \text{ for } k = 1, \ldots, K.$$

In order to evaluate the performance of our proposed method, we use the following criteria:

* the difference between the estimated standard deviation and the true one, $\hat{\sigma}_{\cdot,n} - \sigma^*_2$,
* the difference between the estimated number of segments and the true one, $\hat{K} - K^*$,
* the two components of the Hausdorff distance $d_1(t^*, \hat{t})$ and $d_2(t^*, \hat{t})$ where
  $$d_1(a, b) = \max_b \min_a |a - b|.$$
and \( d_2(a, b) = d_1(b, a) \), in order to study the quality of the estimation of the breakpoint locations. A perfect segmentation results in both null \( d_1 \) and \( d_2 \). Under-segmentation results in a small \( d_1 \) and a large \( d_2 \), meaning that the estimated breakpoints are correctly located, but we miss some breakpoints. Over-segmentation results in a large \( d_1 \) and a small \( d_2 \), meaning we have too many detections, but the true breakpoints are ‘correctly’ recovered.

\* the root mean square error of the mean signal: \( \text{RMSE}(\mu) = \sqrt{\frac{\sum (\hat{\mu}_t - \mu^*_t)^2}{n}} \)

### 4.2 Results

**Estimation of \( \sigma^*_1 \) and \( \sigma^*_2 \)** Figure 1 presents the proposed estimator for the two variances \( \sigma^*_1 \) and \( \sigma^*_2 \). We observe that it performs well to estimate the variances resulting in a similar performance in terms of segmentation estimation (see Figs. 2a and 3 for the selection of \( K \) and Fig. 4 for the locations of the breakpoints when the variances are estimated or fixed to the true values). We can also note that the accuracy of the variance estimations increases with the length of the series \( n \).

**Segmentation estimation for \( \text{MFixedHetero} \)** The two components of the Haussdorff distance \( (d_1 \) and \( d_2 \)) calculated on the obtained segmentations are plotted in Fig. 2c, d, respectively. These distances are also computed for the optimal segmentations with the true number of segments (on the same figures). The \( \text{RMSE}(\mu) \) is plotted in Fig. 2b. In addition, the histograms of breakpoint locations are given in Fig. 4 for three values of \( \sigma^*_2 \) when the number of segments is selected using \( \text{mBIC} \) or fixed to the true value and when the variances are estimated or not (the other criteria giving similar results).

First recall that in a segmentation in the mean context, it has been observed that when the noise is small, the detection problem is easy and the procedure detects the true breakpoints. However, when the problem gets difficult (large variance), the procedure tends to underestimate the number of segments in order to avoid the detection of false breakpoints (see for example Chakar et al. 2015). In our simulation design, among the six breakpoints, four belong to an interval with variance \( \sigma^*_2 \), the fourth one, \( t^*_4 \), belongs to an interval with variance \( \sigma^*_1 \) and the fifth one \( t^*_5 \) corresponds to a change both in the mean and in the variance. We thus observe that our procedure performs as expected and whatever the model selection criteria. First, the variance \( \sigma^*_2 \) does not alter the detection of the breakpoint \( t^*_1 \). When \( \sigma^*_2 \) is small, all the true breakpoints are recovered with a less of accuracy for \( t^*_2 \) and \( t^*_4 \), and when \( \sigma^*_2 \) becomes large, the procedure tends to underestimate the number of segments with estimated breakpoints that are close to the true ones \( (d_1 \) smaller compared to the segmentations with the true number of segments) and with a better precision of the estimator of the mean signal (RMSE(\( \mu \)) is lower than with the true number of segments).

For a very high value of \( \sigma^*_2 \), almost only \( t^*_4 \) is detected. We can also observe that our method performs as well as when the variances are known. Moreover, even if the different criteria for selecting the number of segments show a global same behaviour, there exist some slight differences: \( \text{BM1} \) fails when the detection problem is very easy due to the calibration heuristic; \( \text{BM2} \) tends to detect a little more number of segments compared to the other criteria when the variance is large.

**Comparison with models \( \text{MHomo and MHetero} \)** Figure 5 displays the boxplots of the number of segments selected by \( \text{Lav} \), \( \text{BM} \) and \( \text{mBIC} \) for models \( \text{MHomo} \) and \( \text{MHetero} \), and Figs. 6 and 7 give the histograms of the breakpoint locations for the different model selection criteria and three values of \( \sigma^*_2 \), obtained with \( \text{MHomo} \) and \( \text{MHetero} \), respectively. Note that the imposed changes of variance are located at the positions 25, 50, 75, 100, 125, 150, 175 and the true breakpoints at 27, 38, 88, 111, 150, 183.

Logically \( \text{MHomo} \), \( \text{MHetero} \) and \( \text{MFixedHetero} \) lead to close performances in terms of segmentation when the two variances are close, even if for \( \text{MHetero} \) we can observe an overestimation by \( \text{mBIC} \) (Fig. 7(b—left) or Fig. 5(right)) and a less of accuracy with \( \text{Lav} \) and \( \text{BM2} \). With model \( \text{MHetero} \), as expected, the changes in the variance are also detected. This explained the overestimation of the estimated number of segments. This is more marked with \( \text{mBIC} \). Model

| Series | Known changes                                      |
|--------|---------------------------------------------------|
| SYOG   | 1995-03-15 (RA)                                   |
|        | 1996-01-17 (R)                                    |
|        | 1999-12-24 (R)                                    |
|        | 2000-02-03 (R)                                    |
|        | 2002-01-26 (R)                                    |
|        | 2007-01-25 (R)                                    |
|        | 2008-03-31 (P)                                    |
|        | 2009-03-26 (P)                                    |
| ONSA   | 1999-02-01 (RAD)                                  |
|        | 1999-07-03 (R)                                    |
|        | 2003-08-19 (R)                                    |
|        | 2004-03-10 (R)                                    |
|        | 2007-11-01 (R)                                    |
|        | 2008-03-31 (P)                                    |
|        | 2008-05-15 (R)                                    |
|        | 2009-03-26 (P)                                    |

All changes correspond to a change of receiver (R), antenna (A), radome (D) or processing (P). RA and RAD indicate combined changes.

Table 1 Known changes in the two considered series
MHomo behaves similarly as model MFixedHetero, except when the variance is too large (Fig. 6c). In this latter case, MFixedHetero can be able to detect the fourth breakpoint $t^*_4$ contrary to MHomo for which the estimated standard deviation is larger than 0.5 in the corresponding interval (1.27 in average).

5 Application to GNSS-derived integrated water vapour series

**Context and data description** The GNSS-derived IWV series are subject to small abrupt changes, due, for example, to equipment changes that are difficult to distinguish from the natural climate variation in the measured IWV signal (Ning et al. 2016). In this work, we analyse the differences between IWV series from GNSS and the ECMWF reanalysis, ERA-Interim (Dee et al. 2011) referred to as IWV series. The ERA-Interim reanalysis is produced from the assimilation of many observational data with 6-hourly sampling and provides a very good approximation of the true state of the atmosphere. The underlying climate variations are thus expected to be completely removed in difference time series IWV which contains then essentially the changes in the mean IWV due to inhomogeneities in the GNSS data and noise contributed by both data sources. Inhomogeneities in the reanalysis due to changes in the assimilation data have also been evoked by some authors (Ning et al. 2016; Schroeder et al. 2016). The main source of error in the IWV series is, however, due representativeness differences which is highly site dependent (Parracho et al. 2018). The representativeness differences arise from the fact that the reanalysis provides a gridded representation of the physical variables (spatial averages over areas of 0.75° by 0.75°), whereas the GNSS data are nearly point observations. As a result, it is commonly observed that the IWV series exhibit larger scatter during the summer period when IWV is large. The noise component of our time series is thus not stationary. This effect is explicitly modelled in our segmentation method as described in Sect. 2. At some sites, it may happen that the representativeness differences produce a small residual seasonal variation in the mean and time-correlated errors. These cases set a limit to the proposed method which assumes the series can be modelled by a IID Gaussian process. These features will be taken into account in future versions of the method. For now, we exclude the sites where the data do not reasonably follow the proposed model. In this application, we consider the IWV series of two stations SYOG (Syowa, Antarctica) and ONSA (Onsala, Sweden) contributing to the International GNSS Service (IGS) network of continuously operating reference stations (www.igs.org). The IWV data retrieved from these GNSS measurements are described in Parracho et al. (2018). In the present work, the IWV data series are used with daily time sampling. The equipment changes are available from the so-called IGS sitelogs and are given in Table 1.

**Model (1) for this application** For these series, the variance time intervals correspond to the different months, i.e.

\[ j = \text{month}, \quad J = 12, \quad \text{and each interval } I_{\text{month}}^{\text{var}} \text{ is the union of several intervals among the considered years,} \]
Fig. 9 Results for the series ONSA. a The estimated breakpoints with $\hat{K} = 2$ (left: Lav and BM2) and $\hat{K} = 15$ (right: BM1). b The estimated mean with $\hat{K} = 74$ (left: mBIC) and the estimated standard deviation for each month (right). c The estimated breakpoints obtained with model MHomo (left: $\hat{K} = 5$ with Lav) and with model MHetero (right: $\hat{K} = 14$ with Lav). Solid lines (in red): the estimated breakpoints and the fitted expectation. Dashed lines (in black): known equipment changes (see Table 11). (Color figure online)

$$I_{\text{var month}} = \bigcup_{\text{year}} I_{\text{var year month}},$$

where $I_{\text{var year month}}$ is the time-interval of the month ‘month’ of the year ‘year’.

\[ \sigma_j^2 = \sigma_{\text{month}}^2 \]

is estimated by $Q_n(x_{\text{month}})$ with

$$x_{\text{month}} = \left\{ (y_t - y_{t-1})_{\text{year date(t-1) and date(t)\in month}} \right\}_{\text{year}},$$

i.e. the differentiated series of the considered month of all the years where $Q_n$ is defined by (8).

Results The results for the series ONSA are given in Fig. 9. The criteria select different number of segments: $\hat{K} = 2$ for Lav and BM2, $\hat{K} = 15$ for BM1 and $\hat{K} = 76$ for mBIC. The big abrupt change at date 1999-02-04 is always detected
and is associated with a change in receiver, antenna and radome. When \( \hat{K} = 15 \), only one break corresponds to a known change and two others are close. Contrary to the previous series, the estimated monthly variances are different (higher in summer) resulting in a different segmentation for models MHomo and MHetero (see Fig. 9c where the criterion Lav is considered). The breakpoint at date 1999-02-04 is detected with the both. However, with our method we observe an overestimation of the number of breakpoints and the estimated breakpoints are clearly not linked to known equipment changes. Note that all equipment changes do not impact the time series (Ning et al. 2016).

For the SYOG series, three criteria select four breakpoints, while mBIC selects 81. Figure 8a shows that the four breakpoints coincide well with known equipment changes. Two of the coincide exactly (2008-03-31 and 2009-03-26), and two are close to known changes (1999-12-16 and 2007-02-15). This segmentation is also obtained by both the models MHomo and MHetero with BM2. This can be explained by the fact that the monthly variances are quite similar (see the estimated standard deviation of each month as shown in Fig. 8 (lower)).

For both series, we observe an overestimation of the number of segments when using the mBIC criterion. By looking to the estimated means (Figs. 8 (middle), 9b—left), this overestimation links to the detection of outliers and seems to an overestimation of the number of breakpoints and the estimated breakpoints are clearly not linked to known equipment changes.

## Appendix: Proof of Proposition 1

To proof the result of proposition 1, we use results obtained by Lévy-Leduc et al. (2011) and a trick used in Chakar et al. (2015). We denote by \( t^* = \{t_k^*\}_{k=1}^{K^*} \) the set of the true breakpoints. Let denote \( F \) the cumulative distribution function of \( \{|X_t|\}_{t \in \mathbb{Z}} \) and \( F^X \) the cumulative empirical distribution function of \( \{|X_t|\}_{t \in \mathbb{Z}} \) (same for \( F^\nu_n \)). As in Lévy-Leduc et al. (2011), we define the two mappings

\[
T_1 : F \mapsto \left\{ r \mapsto \int_{\mathbb{R}} \int_{\mathbb{R}} \mathbb{1}_{|x-y| \leq r} dF(x)dF(y) \right\},
\]

and

\[
T_2 : U \mapsto U^{-1}(1/4).
\]

If \( T_0 \circ T_1 \), then \( Q_{CR,n}(X) = c_Q T_0(F^X_n) \), and

\[
\sqrt{n} \left( Q_{CR,n}(X) - \sqrt{2}\sigma \right) = c_Q \sqrt{n} \left( T_0(F^X_n) - T_0(F) \right).
\]

Moreover, we have that

\[
\sqrt{n} \left( F^X_n(i) - F(i) \right) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\mathbb{1}_{|X_t| \leq |i|} - F(i)) = \frac{1}{\sqrt{n}} \sum_{t = 1}^{n} (\mathbb{1}_{|X_t| \leq |i|} - F(i)) = + \frac{1}{\sqrt{n}} \sum_{t = 1}^{n} (\mathbb{1}_{|X_t| \leq |i|} - F(i)).
\]

Since the series \( \{X_t\}_t = \{Y_t - Y_{t-1}\}_t \) is centred except at the breakpoint positions minus one, we obtain that

\[
\sqrt{n} \left( F^X_n(i) - F(i) \right) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} (\mathbb{1}_{|\nu_t| \leq |i|} - F(i)) + R_n(i),
\]

where \( R_n(i) \) tends uniformly to zero in probability. Thus, we get

\[
\sqrt{n} \left( Q_{CR,n}(X) - \sqrt{2}\sigma \right) = \sqrt{n} \left( Q_{CR,n}(\nu) - \sqrt{2}\sigma \right) + o_P(1).
\]

We apply Theorem 2 of Lévy-Leduc et al. (2011) on \( (\nu_t) \) with \( \gamma(0) = 2\sigma^2 \) and since \( \sum_{h \geq 1} |\gamma(h)| < \infty \), where \( \gamma(h) = \text{Cov}(\nu_t, \nu_{t+h}) \), to obtain the convergence of the first term in the right-hand side. That concludes the proof.

## References

Arlot, S., Massart, P.: Data-driven calibration of penalties for least-squares regression. J. Mach. Learn. Res. 10, 245–279 (2009); (electronic)

Auger, I., Lawrence, C.: Algorithms for the optimal identification of segments neighborhoods. Bull. Math. Biol. 66(6), 503515 (1954)

Bellman, R.: The theory of dynamic programming. Bull. Am. Math. Soc. 60(6), 503515 (1954)

Birg, L., Massart, P.: Gaussian model selection. J. Eur. Math. Soc. 3, 203–268 (2001)

Braun, J.V., Braun, R., Müller, H.G.: Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. Biometrika 87(2), 301–314 (2000)

Causin, H., Mestre, O.: Detection and correction of artificial shifts in climate series. Appl. Stat. 53, 405–425 (2004)

Chakar, S., Lebarbier, E., Levy-Leduc, C., Robin, S.: A robust approach for estimating change-points in the mean of an ar(1) process. Bernoulli to appear (2015); (arXiv:14031958)

Cleynen, A., Robin, S.: Comparing change-point location in independent series. Statist. Comput. 26(1–2), 263–276 (2016)
Cleynen, A., Dudoit, S., Robin, S.: Comparing segmentation methods for genome annotation based on rna-seq data. J. Agric. Biol. Environ. Stat. 19(1), 101–118 (2014)

Dee, D.P., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, dP, et al.: The era-interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137(656), 553–597 (2011)

Gazeaux, J., Williams, S., King, M., Bos, M., Dach, R., Deo, M., Moore, A.W., Webb, F.H.: Detecting offsets in GPS time series: first results from the detection of offsets in GPS experiment. J. Geophys. Res. Solid Earth 118, 2397–2407 (2013). https://doi.org/10.1002/jgrb.50152

Gazeaux, J., Lebarbier, E., Collilieux, X., Métivier, L.: Joint segmentation of multiple gps coordinate series. Journal de la Société Française de Statistique 156(4), 163–179 (2015)

Lai, T.L., Liu, H., Xing, H.: Autoregressive models with piecewise constant volatility and regression parameters. Stat. Sin. 15, 279–301 (2005a)

Lai, W., Johnson, M., Kucherlapati, R., Park, P.J.: Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics 21(19), 3763–3770 (2005b)

Lebarbier, E.: Detecting multiple change-points in the mean of Gaussian process by model selection. Signal Process. 85(8), 1501–1510 (2005)

Lévy-Leduc, C., Boistard, H., Moulines, E., Taqqu, M.S., Reisen, V.A.: Robust estimation of the scale and of the autocovariance function of Gaussian short-and long-range dependent processes. J. Time Ser. Anal. 32(2), 135–156 (2011)

Ley, R., Volz, M., Vervelde, K., van der Gaag, L., van der Maas, P.: The homogeneity and interpretation of air pollution and climate data: a bayesian approach. J. Appl. Meteorol. Clim. 55(7), 1633–1649 (2016)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.