Comput Stat
DOI 10.1007/s00180-015-0587-5

ORIGINAL PAPER

A sequential multiple change-point detection procedure via VIF regression

Xiaoping Shi[1] · Xiang-Sheng Wang[1] · Dongwei Wei[1] · Yuehua Wu[1]

Received: 6 March 2013 / Accepted: 12 May 2015

Abstract In this paper, we propose a procedure for detecting multiple change-points in a mean-shift model, where the number of change-points is allowed to increase with the sample size. A theoretic justification for our new method is also given. We first convert the change-point problem into a variable selection problem by partitioning the data sequence into several segments. Then, we apply a modified variance inflation factor regression algorithm to each segment in sequential order. When a segment that is suspected of containing a change-point is found, we use a weighted cumulative sum to test if there is indeed a change-point in this segment. The proposed procedure is implemented in an algorithm which, compared to two popular methods via simulation studies, demonstrates satisfactory performance in terms of accuracy, stability and computation time. Finally, we apply our new algorithm to analyze two real data examples.

Keywords CUSUM · Mean-shift model · Partition · Variable selection · Variance inflation factor regression algorithm

1 Introduction

Change-point problems can be found in many areas of science and engineering. Detecting change-points in a data sequence is of great importance. If there exists a change-point in a data sequence, it is not appropriate to make statistical inferences

The research was partially supported by Natural Sciences and Engineering Research Council of Canada.

✉ Yuehua Wu
wuyh@mathstat.yorku.ca

[1] Department of Mathematics and Statistics, York University, Toronto M3J 1P3, Canada

Published online: 03 June 2015

without considering its existence, because the results derived from such inferences may be misleading. The problem of detecting a single change-point has been studied extensively in the literature [see Csörgő and Horváth (1997) and Chen and Gupta (2012) among others]. However, in this data-rich era, many data sequences have a very large size, and thus it is not surprising that multiple change-points might occur in such a data sequence. It becomes desirable to find a fast and efficient method to detect the locations of these change-points. Recent literature in this area includes Harchaoui and Lévy-Leduc (2008, 2010), Killick et al. (2012), Jin et al. (2013) among others. In this paper, we will tackle the problem of multiple change-point detection in a mean-shift model given below

$$y_i = \sum_{r=0}^{b} \mu_r I_{\{k_r,\ldots,k_{r+1}-1\}}(i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $I_A(\cdot)$ denotes the indicator function of the set A; $1 < k_1 < \cdots < k_b < n$ are the unknown locations of b change-points satisfying $\lim_{n\to\infty} \min_r (k_r - k_{r-1})/n > 0$; $\mu_0, \ldots, \mu_b$ are the means such that $\mu_r \neq \mu_{r+1}$ for $0 \leq r \leq b-1$; and $\varepsilon_1, \ldots, \varepsilon_n$ are random errors with zero mean. Here, we have used the convention that $k_0 = 1$ and $k_{b+1} = n+1$. We denote the set of change-points by $\mathcal{K} = \{k_1, \ldots, k_b\}$.

Let us illustrate the application of multiple change-point detection by the following example. Consider the problem of recognizing a one-dimensional barcode that encodes 0123456789 in the top panel of Fig. 1 (http://barcode.tec-it.com/barcode-generator.aspx). When the image is converted into matrix form, all of the values in the matrix lie between 0 (black pixel) and 1 (white pixel). It is noted that all rows in this matrix are identical, and $\min_r (k_r - k_{r-1})$ in any row is 40. The barcode recognition problem here can be converted into a multiple change-point detection problem in a mean-shift model. Decontaminating the barcodes is equivalent to finding the set of change-points $\mathcal{K}$. To simulate the scanned input, we add two levels of noise to each element of the matrix. The resulting data are left-truncated at 0 and right-truncated at 1, which yields two barcodes, shown respectively in panels 2, 3 in Fig. 1.

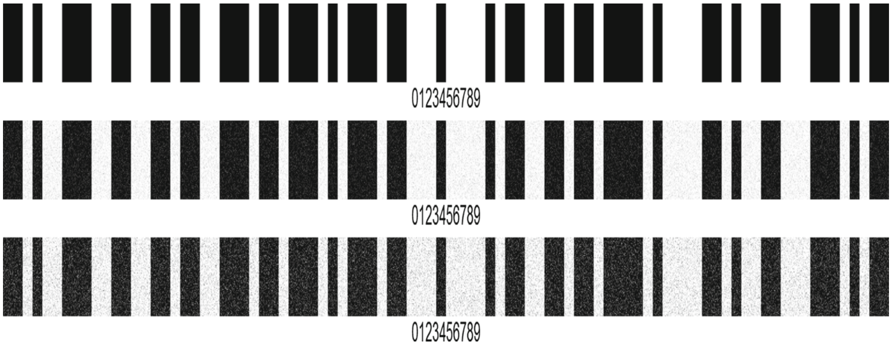

Fig. 1 *Top panel* the original barcode encoding 0123456789 without noise. *Middle panel* the original barcode contaminated by the added Gaussian noise with mean zero and $\sigma = 0.1$. *Bottom panel* the original barcode contaminated by the added Gaussian noise with mean zero and $\sigma = 0.2$

In addition to the barcode recognition problem, the detection of multiple change-points has many applications in areas such as genetic data analysis (see, e.g., Barry and Hartigan 1992, 1993; Erdman and Emerson 2007, 2008) and signal processing (see, e.g., Qu and Tu 2006).

There is a great need for efficient methods for detecting multiple change-points. Barry and Hartigan (1993) proposed a Bayesian analysis for change-point problems with the complexity of $O(n^3)$. This method was further improved by Erdman and Emerson (2008), who reduced the computation time to $O(n)$. While there are some other methods with computational complexity of order $O(n^2)$ (see e.g. Auger and Lawrence (1989), Jackson et al. (2005) and Rigaill (2010)), Scott and Knott (1974) proposed a faster binary segmentation algorithm with only $O(n \log n)$ computational complexity. The main feature of this algorithm is that it only considers a subset of the 2^{n-1} possible solutions (Killick and Eckley 2014).

The circular binary segmentation (CBS) and the Pruned Exact Linear Time (PELT) are two popular methods for detecting multiple change-points in a mean-shift model. CBS was proposed by Olshen et al. (2004) to detect change-points in the genomic data, and has been implemented in the R package **DNAcopy** (Seshan and Olshen 2015). PELT was proposed in Killick et al. (2012), and has also been implemented in the R package **changepoint** (Killick et al. 2014). The main idea behind PELT is to consider the data sequentially, and record the optimal segmentation at each step, for the data up to that step (Killick et al. 2012). The computation time of PELT is of order $O(n)$, but its R package **changepoint** is not stable when there are outliers in the data. Throughout this paper, we use CBS and PELT to stand for the R packages **DNAcopy** and **changepoint**, respectively.

It is noted that by properly segmenting a data sequence, the multiple change-point detection problem above can be equivalently expressed as a linear regression variable selection problem, with a large number of regression coefficients (see Harchaoui and Lévy-Leduc 2008; Jin et al. 2013 among others). Thus a modern variable selection method can be utilized to obtain a rough estimation of multiple change-points. Recently, Lin et al. (2011) proposed the variance inflation factor (VIF) regression algorithm for variable selection. This algorithm is much faster than many modern variable selection methods including LASSO and SCAD. In this paper, we modify the stagewise regression of the VIF regression algorithm, and perform the variable selection sequentially in segment order. Once the segment containing a possible change-point is flagged, we adopt a weighted cumulative sum to justify and locate the change-point in this segment. The proposed procedure is implemented by the algorithm VIFCP ("CP" stands for "change-point"). We would like to remark that our new algorithm allows the number of change-points to increase with the sample size, which makes our method applicable to various practical problems.

The rest of this paper is organized as follows. In Sect. 2, the proposed procedure VIFCP is presented in detail and its theoretical justification is provided. In Sect. 3, we run simulation studies to examine the proposed procedure and to compare its performance with CBS and PELT. In Sect. 4, we give two real data examples. We conclude the paper in Sect. 5. The proof of the theoretical justification of the proposed procedure of Sect. 2 is provided in an "Appendix". The following notation is used throughout the rest of this paper. Let $\{c_n\}$ be a sequence of nonnegative numbers and

$\{d_n\}$ be a sequence of positive numbers. If the sequence $\{c_n/d_n\}$ is bounded, it is denoted as $c_n = O(d_n)$. If $c_n/d_n \to 0$ as $n \to \infty$, it is denoted as $c_n = o(d_n)$. If $c_n/d_n \to 1$ as $n \to \infty$, it is denoted as $c_n \sim d_n$. If a sequence of random variables $\{\xi_n\}$ tends to 0 in probability, it is denoted as $\xi_n = o_p(1)$. The symbol $\xrightarrow{d}$ denotes convergence in distribution. For convenience, we denote the $m \times 1$ vectors $(1, \ldots, 1)^T$ and $(0, \ldots, 0)^T$ by $\mathbf{1}_m$ and $\mathbf{0}_m$, respectively, and write $\boldsymbol{\ell}_{m_1,m_2} = (\mathbf{0}_{m_1}^T, \mathbf{1}_{m_2}^T)^T$. In addition, I_m stands for an $m \times m$ identity matrix (the subscript m may be suppressed if there is no confusion), $\|\cdot\|$ stands for the Euclidean norm, $\lfloor c \rfloor$ the largest integer less than or equal to a real number c, and $\Phi(\cdot)$ the cumulative distribution function of the standard normal random variable.

2 The VIFCP procedure and its theoretical justification

To establish a connection between the multiple change-point detection and variable selection, we follow the ideas of Harchaoui and Lévy-Leduc (2008) and Jin et al. (2013) to reformulate the model (1) as follows:

$$\boldsymbol{y}_n = \sum_{r=0}^{b} \gamma_r \boldsymbol{\ell}_{k_r-1,n-(k_r-1)} + \boldsymbol{\varepsilon}_n, \tag{2}$$

where $\boldsymbol{y}_n = (y_1, \ldots, y_n)^T$ is a column vector of n observations, γ_r with $r = 1, \ldots, b$ are the differences between two successive means $\mu_r - \mu_{r-1}$, and $\gamma_0 = \mu_0$, and $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Thus we can consider detecting multiple change-points for model (1) as carrying out variable selection for model (2). It is noted that this variable selection problem is different from the traditional one, since $\mathcal{K}$ is unknown in model (2). Nevertheless, the problem can be solved by applying the multiple change-point detection procedure as given below. The main idea of our new procedure is to divide the data sequence into smaller segments and sample each segment in sequential order. If no change-point is detected in a segment, the next segment is added to the collective pool of other segments that have been labeled as such. If this segment exhibits potential for containing a change-point, it is flagged and a weighted cumulative sum (CUSUM) is applied to test if there is a change-point in this segment.

2.1 Modified VIF regression algorithm and its justification

We first introduce an artificial partition $\mathcal{Q} = \{q_1, \ldots, q_a\}$ which divides the set $\{1, \ldots, n\}$ into $a + 1$ segments, where $l = \lfloor n/(a+1) \rfloor$ is the length of each segment excluding the first one. We set $q_s = n - (a + 1 - s)l$ for each $s = 1, \ldots, a$. Without loss of generality, we may assume that n is a multiple of $a + 1$, and hence $q_s = sl$ with $l = n/(a+1)$ being the length of all segments. By convention, we also set $q_0 = 0$.

Note that each artificial segment contains at most one change-point by the setup of model (1) and Assumption A1 below.

To reflect the artificial partition in model (2), we rewrite it as

$$\boldsymbol{y}_n = \sum_{s=0}^{a} \beta_s \boldsymbol{\ell}_{q_s,n-q_s} - \boldsymbol{\eta}_n + \boldsymbol{\varepsilon}_n. \tag{3}$$

The regression coefficients β_s (with $s = 1, \ldots, a$) are zeros, except when the artificial segment $[q_s + 1, q_{s+1}]$ contains a change-point, say k_r, and in this case, $\beta_s = \gamma_r$. By convention, we set $\beta_0 = \gamma_0$. The error vector $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is defined in the same way as in (2). Thus, we have correction vector $\boldsymbol{\eta}_n = \sum_{s=0}^{a} \beta_s \boldsymbol{\tau}_n(q_s)$ with $\boldsymbol{\tau}_n(q_s)$ being the zero vector $\mathbf{0}_n$ if $\beta_s = 0$, that is, no change-point exists in the segment $[q_s + 1, q_{s+1}]$, and

$$\boldsymbol{\tau}_n(q_s) = \boldsymbol{\ell}_{k_r-1,n-(k_r-1)} - \boldsymbol{\ell}_{q_s,n-q_s} = (\mathbf{0}^T_{q_s}, \mathbf{1}^T_{k_r-1-q_s}, \mathbf{0}^T_{n-(k_r-1)})^T$$

if $\beta_s = \gamma_r$, i.e., the rth change-point $k_r \in [q_s + 1, q_{s+1}]$. By convention, $\boldsymbol{\tau}_n(q_0) = \mathbf{0}_n$. It is readily seen that $\boldsymbol{\eta}_n$ is a sparse vector, because the change-points are sparse and the length of each artificial segment is comparably small. We would like to remark that if the artificial partition has exactly n segments, then model (2) reduces to the one studied by Harchaoui and Lévy-Leduc (2008). An illustration of the artificial partition is plotted in Fig. 2, where $n = 10$, $\boldsymbol{\varepsilon}_{10} = \mathbf{0}_{10}$, $b = 2$, $k_1 = 4$ and $k_2 = 7$. The model (2) is

$$\boldsymbol{y}_{10} = \gamma_0 \mathbf{1}_{10} + \gamma_1 \boldsymbol{\ell}_{3,7} + \gamma_2 \boldsymbol{\ell}_{6,4}.$$

Given an artificial partition $\mathcal{Q} = (2, 4, 6, 8)$, namely, $a = 4$, $l = 2$ and $q_s = 2s$, this model can be re-expressed as follows:

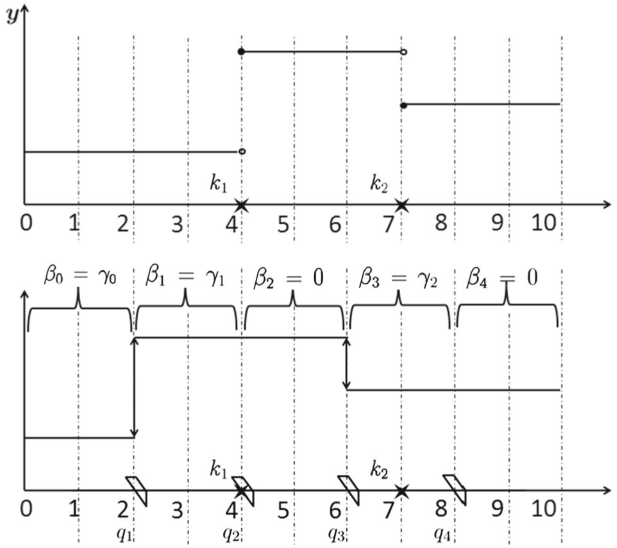

Fig. 2 The *upper plot* is the observations $\boldsymbol{y}$ of size 10 without random errors; the *one below* is the symbolic illustration of a parametric transformation (without the correction vector) by an artificial partition. Here the signs 'star' and 'diagonal stripe' represent locations of change-points and segments, respectively

$$\boldsymbol{y}_{10} = \beta_0 \mathbf{1}_{10} + \beta_1 \boldsymbol{\ell}_{2,8} + \beta_2 \boldsymbol{\ell}_{4,6} + \beta_3 \boldsymbol{\ell}_{6,4} + \beta_4 \boldsymbol{\ell}_{8,2} - \boldsymbol{\eta}_{10},$$

where $\beta_0 = \gamma_0$, $\beta_1 = \gamma_1$, $\beta_2 = 0$, $\beta_3 = \gamma_2$, $\beta_4 = 0$, and the correction vector $\boldsymbol{\eta}_{10} = (0, 0, \gamma_1, 0, 0, 0, 0, 0, 0, 0)^T$, which is symbolically illustrated in Fig. 2.

As mentioned previously, we will adopt the VIF regression algorithm (Lin et al. 2011) because it is an extremely fast algorithm for variable selection with satisfactory accuracy. It consists of two steps: the search step and the evaluation step. The search step takes advantage of sparsity (i.e., the nonzero regression coefficients are sparse in the set of all regression coefficients). The evaluation step is similar to that of a variation of a stepwise regression,*forward stagewise regression*, which evaluates variables using only marginal correlations. A typical forward stagewise regression can be used to test the following alternative model:

$$\boldsymbol{y} = \sum_{j=0}^{m} \beta_j \boldsymbol{x}_j + \beta_{\text{new}} \boldsymbol{x}_{\text{new}} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_m$ are linearly independent predictors and $\boldsymbol{x}_{\text{new}}$ is a new predictor. Let $X = (\boldsymbol{x}_0, \ldots, \boldsymbol{x}_k)$. Now, we define $\boldsymbol{r}_y = \boldsymbol{y} - X(X^T X)^{-1} X^T \boldsymbol{y}$ and $\boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}} - X(X^T X)^{-1} X^T \boldsymbol{x}_{\text{new}}$ to be the residuals of $\boldsymbol{y}$ and $\boldsymbol{x}_{\text{new}}$, respectively. The least squares estimation of β_{new} is given by

$$\hat{\beta}_{\text{new}} = \boldsymbol{r}_{\text{new}}^T \boldsymbol{r}_y / \boldsymbol{r}_{\text{new}}^T \boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_y / \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}}^T [I - X(X^T X)^{-1} X^T] \boldsymbol{y} / \rho^2, \tag{4}$$

where

$$\rho^2 = \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_{\text{new}} = \boldsymbol{x}_{\text{new}}^T [I - X(X^T X)^{-1} X^T] \boldsymbol{x}_{\text{new}}. \tag{5}$$

Since $I - X(X^T X)^{-1} X^T$ is an idempotent symmetric matrix (a fact which will be used frequently throughout this paper), one can derive that the variance of $\hat{\beta}_{\text{new}}$ is $\rho^{-2}\sigma^2$. Lin et al. (2011) suggested constructing the t-statistic $\hat{t} = \hat{\beta}_{\text{new}} \rho / \hat{\sigma} = \boldsymbol{x}_{\text{new}}^T \boldsymbol{r}_y / (\hat{\sigma} \rho)$, where $\hat{\sigma} = \|\boldsymbol{r}_y\| / \sqrt{(n-k-2)}$, the corresponding root-mean-square error (RMSE) of the residual $\boldsymbol{r}_y$. If $\Phi(|\hat{t}|) > 1 - \alpha/2$ for significance level α, then the new predictor $\boldsymbol{x}_{\text{new}}$ is added to the model. This is the key to the algorithm given in Lin et al. (2011).

We remark that the VIF regression algorithm cannot be directly applied to our variable selection problem, because any two successive vectors $\boldsymbol{\ell}_{q_s, n-q_s}$ and $\boldsymbol{\ell}_{q_{s+1}, n-q_{s+1}}$ differ only by $o\left(n^{2/3}\right)$ number of elements under Assumption A1 below, and hence are asymptotically correlated. However, to overcome these obstacles, we can modify the stagewise regression of the VIF regression algorithm as follows.

Suppose the predictors $\boldsymbol{x}_{1,i}, \ldots, \boldsymbol{x}_{m,i}$ have been selected based on the first il rows of $\boldsymbol{y}_n$. Here $\boldsymbol{x}_{r,i} = \boldsymbol{\ell}_{s_r l, (i-s_r) l}$ and $s_1 < \cdots < s_m < i$. We now check whether $\boldsymbol{x}_{\text{new}}^{(i+1)} = \boldsymbol{\ell}_{il,l}$ should be included as a new predictor via the following model

$$\boldsymbol{y}^{(i+1)} = \sum_{j=0}^{m} \beta_{j,i+1} \boldsymbol{x}_{j,i+1} + \beta_{\text{new}}^{(i+1)} \boldsymbol{x}_{\text{new}}^{(i+1)} - \boldsymbol{\eta}^{(i+1)} + \boldsymbol{\varepsilon}^{(i+1)}, \tag{6}$$

where $\boldsymbol{y}^{(i+1)} = \boldsymbol{y}_{(i+1)l}$ contains the first $(i+1)l$ rows of $\boldsymbol{y}_n$ and $\boldsymbol{x}_{0,i+1} = \mathbf{1}_{(i+1)l}$. The error vector $\boldsymbol{\varepsilon}^{(i+1)}$ and correction vector $\boldsymbol{\eta}^{(i+1)}$ are the first $(i+1)l$ rows truncated from the original vectors $\boldsymbol{\varepsilon}_n$ and $\boldsymbol{\eta}_n$, respectively. Let $X^{(i+1)} = (\boldsymbol{x}_{0,i+1}, \ldots, \boldsymbol{x}_{m,i+1})$. $\beta_{\text{new}}^{(i+1)}$ is estimated by

$$\hat{\beta}_{\text{new}}^{(i+1)} = \rho_{i+1}^{-2} \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \left\{ I - X^{(i+1)} \left[\left(X^{(i+1)}\right)^T X^{(i+1)} \right]^{-1} \left(X^{(i+1)}\right)^T \right\} \boldsymbol{y}^{(i+1)}, \tag{7}$$

where

$$\rho_{i+1}^2 = \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \left\{ I - X^{(i+1)} \left[\left(X^{(i+1)}\right)^T X^{(i+1)} \right]^{-1} \left(X^{(i+1)}\right)^T \right\} \boldsymbol{x}_{\text{new}}^{(i+1)}. \tag{8}$$

Applying (6) and (8) to (7) gives

$$\hat{\beta}_{\text{new}}^{(i+1)} = \beta_{\text{new}}^{(i+1)} + \rho_{i+1}^{-2} \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \left\{ I - X^{(i+1)} \left[\left(X^{(i+1)}\right)^T X^{(i+1)} \right]^{-1} \left(X^{(i+1)}\right)^T \right\} \times \left(\boldsymbol{\varepsilon}^{(i+1)} - \boldsymbol{\eta}^{(i+1)}\right). \tag{9}$$

Following Lin et al. (2011), let

$$\hat{t}_{i+1} = \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \boldsymbol{r}^{(i+1)} / (\hat{\sigma}_{i+1} \rho_{i+1}), \tag{10}$$

where $\boldsymbol{r}^{(i+1)} = [I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1} (X^{(i+1)})^T] \boldsymbol{y}^{(i+1)}$ is the residual and $\hat{\sigma}_{i+1} = \|\boldsymbol{r}^{(i+1)}\| / \sqrt{(i+1)l - m - 2}$ is the corresponding RMSE. If $\Phi(|\hat{t}_{i+1}|) > 1 - \alpha/2$, we put $\boldsymbol{x}_{m+1,i+1} = \boldsymbol{x}_{\text{new}}^{(i+1)}$ and $s_{m+1} = i + 1$, and repeat the above process with i and m replaced respectively by $i+1$ and $m+1$. Otherwise, we repeat the above process by replacing i by $i+1$.

Before giving a theoretical justification of the modified VIF regression algorithm, we make the following two assumptions.

A1. Assume that $l \to \infty$ and $bl^{3/2} \ll n$ as $n \to \infty$.

A2. Assume that the errors $\{\varepsilon_i\}$ in model (1) are independent and identically distributed (iid) zero-mean random variables with variance σ^2. Furthermore, $E|\varepsilon_i|^{2+\nu} < \infty$ for some positive constant $\nu > 0$.

Remark 1 Assumption A1 allows b to go to infinity in the order of $n/M(n)$, where $M(n) \to \infty$ as $n \to \infty$. Assumption A2 is a very basic assumption that is necessary for establishing asymptotic normality of the estimators of βs.

Remark 2 The choice of l should follow the rule that there is no more than one change-point in one partition. Under this condition, we can expect a more accurate estimate of a change-point with larger l.

The following theorem shows that under the assumptions A1–A2, the modified stagewise regression is warranted. Its proof is given in the "Appendix".

Theorem 1 *If the assumptions A1–A2 are satisfied, then as $n \to \infty$,*

$$\rho_{i+1}^{-1}\left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T\left\{I - X^{(i+1)}\left[\left(X^{(i+1)}\right)^T X^{(i+1)}\right]^{-1}\left(X^{(i+1)}\right)^T\right\}\boldsymbol{\varepsilon}^{(i+1)} \xrightarrow{d} N\left(0, \sigma^2\right),$$

$$\left[\left(X^{(i+1)}\right)^T X^{(i+1)}\right]^{-1} = O(1/n), \quad \rho_{i+1}^2/l \to 1, \tag{11}$$

where $l = n/(a+1)$ is the length of each artificial segment. Note that l is large but $l^{3/2}/n$ is small by Assumption A1. Furthermore, the following statements hold true:

(a) If the null hypothesis is accepted, i.e., $\beta_{\text{new}}^{(i+1)} = 0$, then the scaled estimate $\rho_{i+1}\hat{\beta}_{\text{new}}^{(i+1)}$ converges to $N(0, \sigma^2)$ in distribution as $n \to \infty$.
(b) If the alternative hypothesis is accepted, i.e., $\beta_{\text{new}}^{(i+1)} \neq 0$, then
  (i) $\hat{\beta}_{\text{new}}^{(i+1)} = \beta_{\text{new}}^{(i+1)}[1-\rho_{i+1}^{-2}(k_m - il)] + o_p(1)$, where k_m denotes the change-point in the artificial segment $[1+il, (i+1)l]$.
  (ii) Moreover, if the change-point k_m lies in the artificial segment $[1+(i-1)l, il]$ (i.e., the change-point was not detected during the previous search), then $\hat{\beta}_{\text{new}}^{(i)} = \beta_{\text{new}}^{(i)} + o_p(1)$.

2.2 A CUSUM test and its justification

By Theorem 1(b)(i), we may conclude that a change-point exists in the artificial time segment $[il+1, (i+1)l]$ if $\hat{\beta}_{\text{new}}^{(i+1)} \neq 0$. The precise location of the change-point, however, is unknown because the formula (7) does not fully reflect the information contained in the correction vector η_n. To locate a change-point in the artificial segment $[il+1, (i+1)l]$, one may conduct a test for a single change-point over this segment, which, jointly with Theorem 1(b)(ii) suggests that the test only needs to be carried out over the segment $[1+(i-1)l, il+\lfloor l/2 \rfloor]$.

Consider a univariate sequence $\{Z_i\}$ for $i = 1, \ldots, n$ with variance σ^2. We intend to test the null hypothesis

$$H_0 : E(Z_1) = \cdots = E(Z_n)$$

versus the alternative hypothesis

$$H_a : E(Z_1) = \cdots = E(Z_{k^*}) \neq E(Z_{k^*+1}) = \cdots = E(Z_n)$$

for some $k^* \in (1, n)$. The change-point k^* is unknown, and both k^*/n and $1 - k^*/n$ are assumed to be bounded away from zero as $n \to \infty$. Many single change-point detection methods in the literature can be used to solve this problem. Here, we apply the following CUSUM

$$U_k = C_k/w_k \tag{12}$$

to perform the test, where

$$C_k = \left(\frac{n}{k(n-k)}\right)^{1/2}\left(\sum_{i=1}^{k} Z_i - \frac{k}{n}\sum_{i=1}^{n} Z_i\right), \tag{13}$$

and

$$w_k = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\left(Z_i - \frac{1}{k}\sum_{j=1}^{k} Z_j\right)^2 + \frac{1}{n}\sum_{i=k+1}^{n}\left(Z_i - \frac{1}{n-k}\sum_{j=k+1}^{n} Z_j\right)^2}. \tag{14}$$

If $B(\log n)\max_{1\le k<n}|U_k| \le -\log(-\frac{1}{2}\log(1-\alpha)) + D(\log n)$, where $B(x) = (2\log x)^{1/2}$ and $D(x) = 2\log x + (1/2)\log\log x - (1/2)\log\pi$, then there is no change-point, otherwise the change-point exists and is estimated by $\hat{k} = \arg\max|U_k|$. It is noted that the above CUSUM is also related to the quasi-likelihood ratio test statistic. See Csörgő and Horváth (1997) (Eq. 1.4.25) for details.

2.3 The algorithm

The proposed method is implemented by the algorithm VIFCP below. Here, we provide the pseudocode for the VIFCP algorithm:

```
 1. INPUT $\boldsymbol{y}_n$ and $l$.
 2. INITIALIZATION, $a = n/l - 1$, $w = 0.05$, $dw = 0.05$, flag $= 0$, $\widehat{\mathcal{K}} = \emptyset$, $i = 1$,
    $j = 1$.
 3. LOOP{
 4.     SET $\alpha = w/(1 + i - \text{flag})$.
 5.     OBTAIN statistic $\hat{t}_{i+1}$ by (10).
 6.     IF $2\Phi(|\hat{t}_{i+1}|) > 2 - \alpha$
 7.         Test for a change-point $k^*$ in $[(i-1)l, il + \lfloor l/2 \rfloor]$ using the CUSUM.
 8.         IF the test is significant, obtain $\hat{k}_j$.
 9.             $\widehat{\mathcal{K}} \leftarrow \widehat{\mathcal{K}} \cup \{\hat{k}_j\}$, flag$\leftarrow i$, $w \leftarrow w + dw$, $j = j + 1$.
10.         ELSE $w \leftarrow w - \alpha/(1-\alpha)$.
11.         END IF
12.     ELSE $w \leftarrow w - \alpha/(1-\alpha)$.
13.     END IF
14.     UPDATE $i \leftarrow i + 1$.
15. }UNTIL $i \ge a + 1$ or $w \le 0$.
16. RETURN $\widehat{\mathcal{K}}$.
```

The values w, dw and α represent the wealth, payout and significance level, respectively. The details are given in Lin et al. (2011). From the third line to the 15th line, we use a loop to find all change-points. In the fifth line, we calculate the statistic $\hat{t}_{i+1}$ using the formula (10); this is the first key part of our algorithm. If the test is significant, then there may exist a change-point in the artificial segment $[il+1, (i+1)l]$. The 7th line is the second key part of our algorithm, where we apply the algorithm CUSUM defined in (12) to locate the change-point k^* in the interval $[(i-1)l, il + \lfloor l/2 \rfloor]$. Here we set the significance level of CUSUM to be 0.05. After the loop, we obtain $\widehat{\mathcal{K}}$, the estimates of multiple change-points. We remark that this algorithm has been implemented in the R package **VIFCP** (Shi et al. 2015).

We use the example in Sect. 2 to provide a more thorough explanation of our algorithm. The true change-points are located at 4 and 7. As the sample size is 10 and $l = 2$, firstly we set $i = 1$ and will find there is no change-point in the interval [0, 3]; after setting $i = 2$, we will find a change-point in the interval [2, 5] at 4. If we set $i = 3$, we will not detect any changes in the interval [4, 7]. The change-point at 7 will be detected when we set $i = 4$ in the interval [6, 9]. No change-point will be detected in [8, 10] upon setting $i = 5$.

Now, we study the computational complexity of the algorithm VIFCP. Under Assumption A1, the computation time for the variable selection is of order $O(n^2/l)$, and the computation time for performing all the single change-point tests is of order $O(bl)$. Hence the complexity of the algorithm VIFCP is $O(n^2/l + bl)$. It is noted that for finite b, the complexity of the algorithm VIFCP can be as low as $O(n^{4/3}M(n))$ ($M(n)$ is defined in Remark 1), while for $b = o(n)$, the complexity is $o(n^2)$.

3 Simulation studies

In this section, we present three simulation studies. A Dell server (two E5520 Xeon Processors, two 2.26 GHz 8 M Caches, 16 GB Memory) is used to perform the simulation studies. We will compare the performance of the algorithm VIFCP with CBS and PELT in terms of the accuracy of successfully detecting each true change-point, the accuracy of successfully detecting all true change-points under the condition that the number of true change-points is correctly estimated, and efficiency as determined by the elapsed running time in seconds (ERT).

In the simulation studies, we consider the following three model settings:

$S1$: $y_i = \sum_{r=0}^{5} \mu_r I_{\{k_r,\ldots,k_{r+1}-1\}}(i) + \varepsilon_i, \quad i = 1, \ldots, 2000$, where

1. $\{k_0, k_1, k_2, k_3, k_4, k_5, k_6 - 1\} = \{1, 324, 620, 1102, 1386, 1610, 2000\}$,
2. $(\mu_0, \mu_1, \mu_2, \mu_3, \mu_4, \mu_5) = (0, 0.3, 0.7, 0.2, -0.2, 0.3)$,
3. ε_i, $1 \le i \le n$, are iid $\sim N(0, \sigma^2)$,
4. $\sigma = 0.2, 0.3$ and 0.4.

Simulated data for $S1$ with different value of σ, are plotted in Fig. 3.

$S2$: This setting is the same as the setting $S1$ with only the following exception: in each simulation, we randomly select five locations between 1 and n, and then add five to each value at these locations. The values at these five locations are considered as outliers.

$S3$: This setting is the same as the setting $S1$ with only the following exception: in each simulation, we randomly select ten locations between 1 and n, and then add five to each value at these locations. The values at these ten locations are considered as outliers.

For each of the model settings $S1$, $S2$, and $S3$, we first generate a data sequence, and then apply all three methods PELT, CBS, and VIFCP, to detect multiple change-points

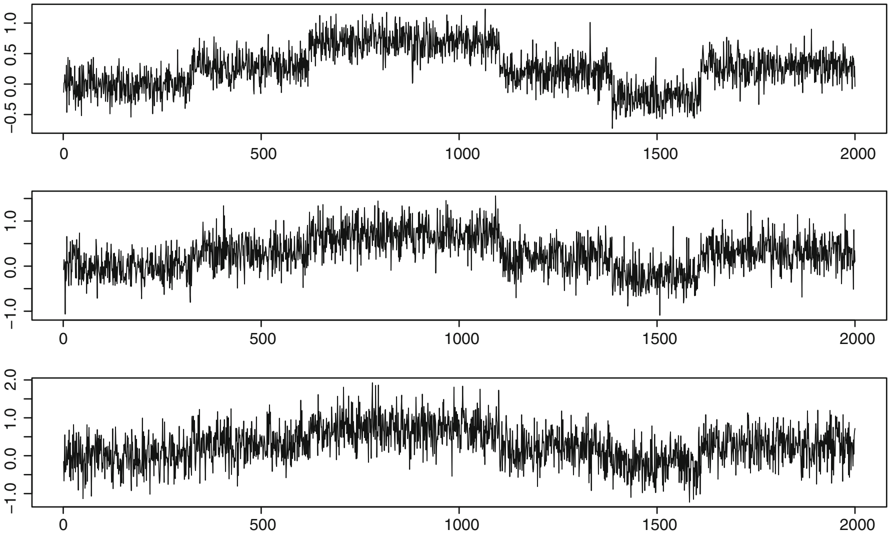

Fig. 3 Simulated data for $S1$ with $\sigma = 0.2, 0.3$ and 0.4 (from the *top* to *bottom panels*)

in the dataset. We denote the set of estimated change points in the mth simulation by $\mathcal{K}^{(m)}$ and define

$$A_{k_i}^{(m)} = \mathcal{K}^{(m)} \cap [k_i - 5, k_i + 5], \tag{15}$$

where $i = 1, \ldots, 5$. Actually, $A_{k_i}^{(m)}$ is the set of estimated change-points, derived from the mth simulation, lying in the neighborhood of the ith true one. Moreover, we define the function $J(x)$ as

$$J(A) = \begin{cases} 0 & \text{if } A = \emptyset \\ 1 & \text{otherwise} \end{cases} \tag{16}$$

Furthermore, we define $B^{(m)} = 1$ if $J(A_{k_i}^{(m)}) = 1$ for all $i = 1, \ldots, 5$ and the size of $\mathcal{K}^{(m)}$ is exactly 5; $B^{(m)} = 0$, otherwise. Note that $B^{(m)} = 1$ if and only if the mth simulation is successful in the sense that it detects exactly five change-points and all of these five estimated change-points are close to the corresponding exact change-points. With the aid of $B^{(m)}$, we define

$$ALLCP = \sum_{m=1}^{M} B^{(m)}/M,$$

the successful simulations as a percentage of all simulations. Here, M is the number of simulations, and in the current paper, we choose $M = 1000$.

The simulation results are reported respectively in Tables 1, 2 and 3.

We observe from Tables 1, 2 and 3 that VIFCP, PELT and CBS have similar performances in accuracy for $S1$. As for $S2$, that differs from $S1$ by having five outliers,

Table 1 Simulation results of PELT, CBS, and VIFCP based on 1000 simulations for three different noise levels ($\sigma = 0.2$, $\sigma = 0.3$, and $\sigma = 0.4$) of scenario 1

Method	PELT			CBS			VIFCP					
l							100			80		
σ	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
$\sum_{m=1}^{1000} J\left(A_{k_1}^{(m)}\right)$[a]	957	831	674	957	830	676	947	811	583	947	795	508
$\sum_{m=1}^{1000} J\left(A_{k_2}^{(m)}\right)$	997	954	838	946	841	733	994	936	819	986	923	702
$\sum_{m=1}^{1000} J\left(A_{k_3}^{(m)}\right)$	999	971	921	999	969	904	999	968	907	998	968	863
$\sum_{m=1}^{1000} J\left(A_{k_4}^{(m)}\right)$	985	931	834	988	921	812	980	924	797	979	930	797
$\sum_{m=1}^{1000} J\left(A_{k_5}^{(m)}\right)$	1000	982	915	997	975	905	994	972	904	997	977	899
cpnumber.R	1000	998	994	872	854	814	980	980	863	940	948	630
ALLCP (%)	93.8	69.9	41.0	89.2	59.5	32.3	91.3	65.4	34.2	90.6	65.6	33.0
ERT.S	9.312	9.502	9.269	201.196	202.099	205.639	0.422	0.503	0.420	0.402	0.417	0.376

cpnumber.R stands for the number of simulations in which the true number of change-points is correctly estimated. ALLCP denotes the percentage of simulations in which exactly five change-points are estimated, and all these five estimated change-points are close to the corresponding exact change-points; namely, $\{|\hat{k}_r - k_r| \leq 5\}$ for $r = 1, \ldots, 5$. ERT.S means the total running time in seconds

[a] $A_{k_i}^{(m)}$ is defined in (15) and $J(A)$ is given in (16)

Table 2 Simulation results of PELT, CBS, and VIFCP based on 1000 simulations for three different noise levels ($\sigma = 0.2$, $\sigma = 0.3$, and $\sigma = 0.4$) of scenario 2

Method	PELT			CBS			VIFCP					
l							100			80		
σ	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
$\sum_{m=1}^{1000} J\left(A_{k_1}^{(m)}\right)^{a}$	534	462	401	612	494	381	781	643	441	797	646	407
$\sum_{m=1}^{1000} J\left(A_{k_2}^{(m)}\right)$	766	730	651	894	776	697	910	833	742	889	775	600
$\sum_{m=1}^{1000} J\left(A_{k_3}^{(m)}\right)$	866	836	787	963	921	874	956	905	844	925	838	743
$\sum_{m=1}^{1000} J\left(A_{k_4}^{(m)}\right)$	729	713	624	943	872	780	885	794	669	884	795	683
$\sum_{m=1}^{1000} J\left(A_{k_5}^{(m)}\right)$	863	818	788	961	926	879	954	902	819	933	884	784
cpnumber.R	0	0	0	602	539	486	817	738	551	695	609	373
ALLCP (%)	0	0	0	69.1	45.3	27.8	64.6	43.2	28.3	67.9	45.3	22.5
ERT.S	7.005	6.940	6.982	117.529	118.142	128.291	0.411	0.423	0.451	0.425	0.374	0.358

cpnumber. R stands for the number of simulations in which the true number of change-points is correctly estimated. ALLCP denotes the percentage of simulations in which exactly five change-points are estimated, and all these five estimated change-points are close to the corresponding exact change-points; namely, $\{|\hat{k}_r - k_r| \leq 5\}$ for $r = 1, \ldots, 5$. ERT.S means the total running time in seconds

[a] $A_{k_i}^{(m)}$ is defined in (15) and $J(A)$ is given in (16)

Table 3 Simulation results of PELT, CBS, and VIFCP based on 1000 simulations for three different noise levels ($\sigma = 0.2$, $\sigma = 0.3$, and $\sigma = 0.4$) of scenario 3

Method	PELT			CBS			VIFCP					
l							100			80		
σ	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4	0.2	0.3	0.4
$\sum_{m=1}^{1000} J\left(A_{k_1}^{(m)}\right)$[a]	290	287	276	376	311	235	643	511	343	643	477	337
$\sum_{m=1}^{1000} J\left(A_{k_2}^{(m)}\right)$	563	534	473	824	717	631	850	729	609	757	630	474
$\sum_{m=1}^{1000} J\left(A_{k_3}^{(m)}\right)$	746	715	668	926	902	836	884	854	748	793	685	577
$\sum_{m=1}^{1000} J\left(A_{k_4}^{(m)}\right)$	514	550	470	875	828	711	755	684	518	752	707	553
$\sum_{m=1}^{1000} J\left(A_{k_5}^{(m)}\right)$	718	703	658	897	886	820	886	860	715	855	783	651
cpnumber.R	0	0	0	389	369	313	627	549	310	490	380	181
ALLCP	0	0	0	52.7	36.0	21.4	45.9	30.4	16.1	43.1	27.9	17.7
ERT.S	6.287	6.095	6.197	91.659	99.610	104.866	0.482	0.431	0.413	0.490	0.406	0.407

cpnumber.R stands for the number of simulations in which the true number of change-points is correctly estimated. ALLCP denotes the percentage of simulations in which exactly five change-points are estimated, and all these five estimated change-points are close to the corresponding exact change-points; namely, $\{|\hat{k}_r - k_r| \leq 5\}$ for $r = 1, \ldots, 5$. ERT.S means the total running time in seconds

[a] $A_{k_i}^{(m)}$ is defined in (15) and $J(A)$ is given in (16)

VIFCP and CBS have better accuracy in multiple change-point detection than PELT, and hence, are more stable. However, for $S3$, the performance of PELT decreases sharply with the increase of the number of outliers. Actually, PELT is very sensitive to the sudden change in observations, and detects outliers as change-points in both $S2$ and $S3$.

If we compare these three methods in terms of ERT.S, we find that VIFCP is much faster than CBS and PELT in all three simulation studies. To examine whether or not the results obtained by using VIFCP are sensitive to the choice of l, we have varied the value of l. It can be seen from the three tables that the results for $l = 80$ and $l = 100$ are similar.

4 Real data examples

In this section, we will analyze the following two real data examples.

4.1 Denoising a barcode

The original barcode was given in the top panel of Fig. 1. As explained in Sect. 1, all the values in the original image matrix range from 0 (black) to 1 (white). We now add Gaussian noises with mean 0, and standard deviation $\sigma = 0.1$ or 0.2, to each element of the original image matrix. Note that the resulting matrices may have elements smaller than 0 or larger than 1. To mimic an image matrix, we replace such elements by 0 or 1, i.e., we apply the transformation $xI_{[0,\,1]}(x) + I_{(1,\,\infty)}(x)$ to each element of the two noise-added matrices to make the noised grayscales range from 0 to 1. We name these two resulting matrices as Matrix 1 and Matrix 2, respectively.

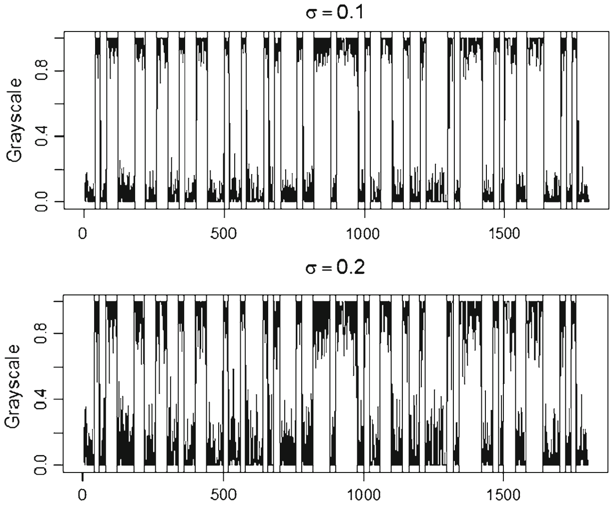

Fig. 4 The data produced by a scanner through reading the *first row* of contaminated barcode image with different noise levels ($\sigma = 0.1$ and $\sigma = 0.2$). The true change-points are marked by *vertical lines*

One realization of the first row of each of Matrices 1–2 is plotted in Fig. 4. The task is to reconstruct the original barcode, i.e., to find all the change-points marked by vertical lines (obtained from the original image in the top panel of Fig. 1). Here, the true number of change-points is 48. We choose $l = 20$ for applying the VIFCP algorithm. For both datasets, VIFCP correctly detected all change-points. In contrast, CBS and PELT failed to detect all change-points. For the case when $\sigma = 0.1$, PELT detected 17 change-points, while CBS correctly detected all of the change-points. When $\sigma = 0.2$, PELT still detected 17 change-points, but CBS failed to detect two change-points. Thus in terms of the multiple change-point detection accuracy, even though CBS and PELT failed to compete with VIFCP, CBS outperformed PELT in this example.

4.2 Genetic data

In this subsection, we consider a test using a genetic dataset involving 57 bladder tumor samples (Stransky et al. 2006); see web page http://microarrays.curie.fr/publications/oncologie_moleculaire/bladder_TCM/. The problem is to find changes in the DNA copy number of an individual using array comparative genomic hybridization (CGH).

In order to perform multiple change-point detection, we first deal with missing values in the dataset. Following Matteson and James (2013), we remove all series that had more than 7 % of values missing, which left genome samples of 42 individuals

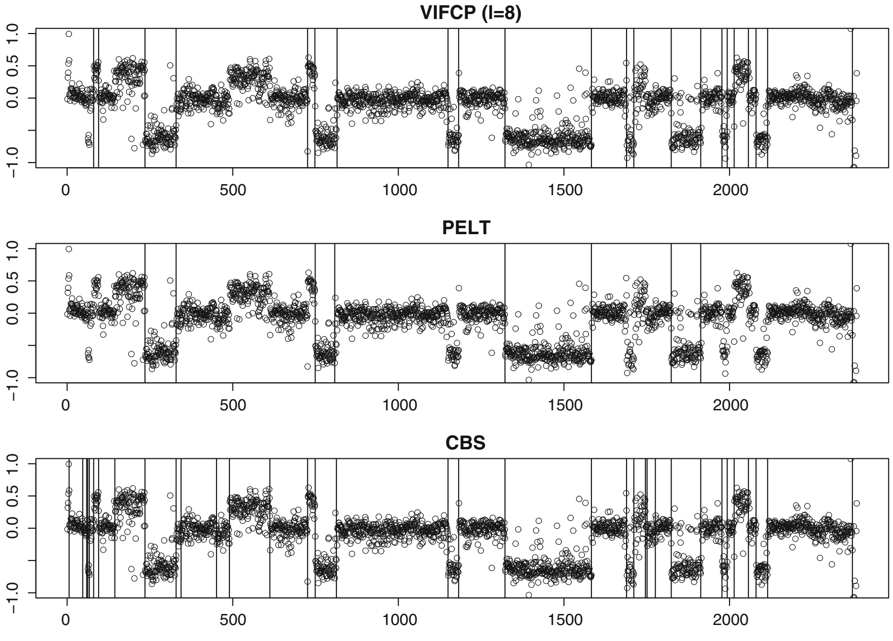

Fig. 5 The normalized relative aCGH signal for the 11th individual with a bladder tumor. The change-points detected by VIFCP with $l = 8$, PELT and CBS are indicated by the *vertical lines*

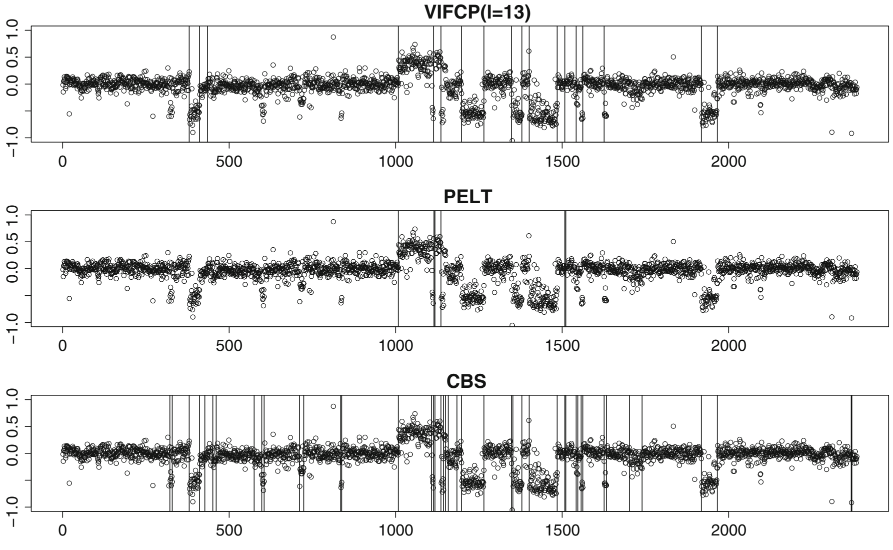

Fig. 6 The normalized relative aCGH signal for the 13th individual with a bladder tumor. The change-points detected by VIFCP with $l = 13$, PELT and CBS are indicated by the *vertical lines*

for analysis. As in Matteson and James (2013), we also normalize the data so that the modal ratio is zero on a logarithmic scale. For each missing value, we find the 3 nearest neighbors using a Euclidean metric, and infer the missing value by averaging the values of its neighbors. As an illustration, we randomly choose two individuals' array CGH dataset for analysis. Here we choose the 11th and 13th individuals.

The choice of l is critical in the real data analysis. We will give a criterion for choosing l later in Sect. 5 [see the formula (17) for more details]. We first limited the range of l to $\{5, 6, \ldots, 30\}$ before applying this criterion. The application of the criterion returned $l = 8$ and $l = 13$, respectively, for the $11th$ and $13th$ patient, as the optimal value of l.

For individual 11, VIFCP with $l = 8$ detects 22 change-points, while PELT and CBS claim respectively 9 and 35 change-points, which are displayed in Fig. 5. From this figure, we observe that both VIFCP and CBS perform better than PELT. PELT fails to detect some change-points. As a matter of fact, neither VIFCP or CBS is perfect in detecting the change-points. There are three potential change-points around 150, 500 and 600 but VIFCP fails to detect them. As for CBS, the minimum distance between two successive change-points is 3, and in addition, the distance between adjacent change-points in each of two pairs is 5. Thus CBS may overestimate the number of change-points.

For individual 13, VIFCP with $l = 13$ detects 18 change-points, while PELT and CBS report 6 and 44 change-points, respectively. The result is shown in Fig. 6. We conclude that CBS and VIFCP perform better than PELT because PELT fails to detect some potential change-points. Moreover, VIFCP may fail to claim some change-points while CBS obviously overestimates the number of change-points.

5 Discussion

In this paper, we propose a procedure, as well as its theoretical justification, for detecting multiple change-points in the mean-shift model, where the number of change-points is allowed to increase with the sample size. We first convert a change-point detection problem into a variable selection problem by partitioning the data sequence. This allows us to apply a modified variance inflation factor regression algorithm to perform the variable selection sequentially in segment order. Once the segment containing a possible change-point is flagged, a weighted CUSUM algorithm is applied to test if there is a change-point in this segment. This procedure is implemented in the algorithm, named VIFCP. Simulation studies demonstrate that VIFCP, when compared with two popular algorithms, CBS and PELT, has a satisfactory performance in accuracy and computation time. It is also shown in the barcode example that VIFCP is better than CBS and PELT in terms of detection accuracy of multiple change-points. In the second real-data analysis, VIFCP and CBS outperform PELT from the point-of-view of estimating change-point locations.

In the simulation studies, segment length l is set to be 100, 80 for $n = 2000$. In the barcode example, l is set as 20, to account for the barcode design. The choice of l is a very important issue. We may make the optimal choice of l by applying a Bayesian information criterion as follows:

$$l_{\text{opt}} = \arg\min_{l}\{\log(n)(DF_l + 1) + n\log(RSS_l/n)\}, \tag{17}$$

where DF is the number of estimated change-points and $RSS_l = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

In this paper, it can be seen that the proposed procedure for a mean-shift model can be extended to detect multiple change-points in other types of regression models, including generalized linear models. The algorithm for implementing such a procedure is also feasible, requiring only a straightforward extension of VIFCP.

Acknowledgments The authors would like to thank the associate editor and two anonymous reviewers for the critical comments and constructive suggestions which have led to the improvement of this paper. The authors would also like to thank Professor Trueman MacHenry for polishing the paper.

Appendix: Proof of Theorem 1

Since ε_i, $i = 1, 2, \ldots$, are iid zero-mean variables with variance σ^2, it follows from the definition of ρ_{i+1} in (8) and the idempotence of $I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T$ that the variance of $\rho_{i+1}^{-1}(\boldsymbol{x}_{\text{new}}^{(i+1)})^T\{I - X^{(i+1)}[(X^{(i+1)})^T X^{(i+1)}]^{-1}(X^{(i+1)})^T\}\boldsymbol{\varepsilon}^{(i+1)}$ is still σ^2. By the central limit theorem, we obtain that

$$\rho_{i+1}^{-1}\left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T\left\{I - X^{(i+1)}\left[\left(X^{(i+1)}\right)^T X^{(i+1)}\right]^{-1}\left(X^{(i+1)}\right)^T\right\}\boldsymbol{\varepsilon}^{(i+1)} \xrightarrow{d} N\left(0, \sigma^2\right).$$

Note that $(X^{(i+1)})^T X^{(i+1)}$ can be expressed as $(U^{(i+1)})^T \Lambda^{(i+1)} U^{(i+1)}$, where $U^{(i+1)}$ is the lower triangular matrix of order $k+1$ whose nonzero entries are all 1's, and $\Lambda^{(i+1)}$ is a diagonal matrix with diagonal entries being $k_1-k_0,\ k_2-k_1,\ \ldots,\ k_m-k_{m-1},\ 1+(i+1)l-k_m$. Since the change-points are well-separated, i.e., $k_r-k_{r-1}=O(n)$, $(\Lambda^{(i+1)})^{-1}$ is of order $O(1/n)$, we have that $[(X^{(i+1)})^T X^{(i+1)}]^{-1}$ is also of order $O(1/n)$.

Next, we prove that ρ_{i+1} defined in (8) is asymptotically equal to $\sqrt{l}$. Note that $\boldsymbol{x}_{\text{new}}^{(i+1)}=\boldsymbol{\ell}_{il,l}$ is the vector with only the last l elements being ones, and all other elements are zeros. It can be seen that $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T\boldsymbol{x}_{\text{new}}^{(i+1)}=l$ and $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T X^{(n+1)}=O(l)$. Therefore, as $n\to\infty$, it is readily seen from $[(X^{(i+1)})^T X^{(i+1)}]^{-1}=O(1/n)$ that

$$\begin{aligned}\rho_{i+1}^2 &= \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \boldsymbol{x}_{\text{new}}^{(i+1)} - \left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T \\ &\quad\times\left\{I - X^{(i+1)}\left[\left(X^{(i+1)}\right)^T X^{(i+1)}\right]^{-1}\left(X^{(i+1)}\right)^T\right\}\boldsymbol{x}_{\text{new}}^{(i+1)} \\ &= l - O\left(l^2/n\right)\sim l.\end{aligned}$$

Under the null hypothesis, there exists no change-point in the interval $[1+il,\ (i+1)l]$. It can be shown that the last l elements of the correction vector $\boldsymbol{\eta}^{(i+1)}$ are zeros, which implies that $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T\boldsymbol{\eta}^{(i+1)}=0$. Since $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T X^{(i+1)}=O(l)$, $(X^{(i+1)})^T\boldsymbol{\eta}^{(i+1)}=o_p(bl)$, $[(X^{(i+1)})^T X^{(i+1)}]^{-1}=O(1/n)$ and $\rho_{i+1}/\sqrt{l}\to 1$, by Assumption A1, it follows that

$$\rho_{i+1}^{-1}\left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T\left\{I - X^{(i+1)}\left[\left(X^{(i+1)}\right)^T X^{(i+1)}\right]^{-1}\left(X^{(i+1)}\right)^T\right\}\boldsymbol{\eta}^{(i+1)}=o(1).$$

In view of the fact that $\beta_{\text{new}}^{(i+1)}=0$, i.e., there is no change-point in $[1+il,\ (i+1)l]$, and $\rho_{i+1}\to\infty$, by (7) and (9), we obtain that

$$\rho_{i+1}\hat{\beta}_{\text{new}}^{(i+1)}\xrightarrow{d}N(0,\sigma^2).$$

This proves Theorem 1(a).

Under the alternative hypothesis, there exists a change-point, say k_m, in the segment $[1+il,\ (i+1)l]$. Moreover, k_m-il many of the last l elements of the correction vector $\boldsymbol{\eta}^{(i+1)}$ are equal to $\beta_{\text{new}}^{(i+1)}$, and $\beta_{\text{new}}^{(i+1)}\neq 0$, which implies $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T\boldsymbol{\eta}^{(i+1)}=\beta_{\text{new}}^{(i+1)}(k_m-il)$.

Moreover, we have

$$\rho_{i+1}^{-2}\left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T X^{(i+1)}\left[\left(X^{(i+1)}\right)^T X^{(i+1)}\right]^{-1}\left(X^{(i+1)}\right)^T\boldsymbol{\eta}^{(i+1)}=o_p(1)$$

from the Proof of Theorem 1(a). In view of (11), we obtain that

$$\rho_{i+1}^{-2}\left(\boldsymbol{x}_{\text{new}}^{(i+1)}\right)^T\left\{I-X^{(i+1)}\left[\left(X^{(i+1)}\right)^T X^{(i+1)}\right]^{-1}\left(X^{(i+1)}\right)^T\right\}\boldsymbol{\varepsilon}^{(i+1)}=o_p(1).$$

Applying these results to (7) yields

$$\hat{\beta}_{\text{new}}^{(i+1)}=\beta_{\text{new}}^{(i+1)}\left[1-\rho_{i+1}^{-2}(k_m-il)\right]+o_p(1).$$

Furthermore, if the change-point k_m is located in the artificial interval $[1+(i-1)l, il]$ (i.e., the change-point was previously undetected), then the correction vector $\boldsymbol{\eta}^{(i+1)}$ has zero components in the last l rows, which implies that $(\boldsymbol{x}_{\text{new}}^{(i+1)})^T\boldsymbol{\eta}^{(i+1)}=0$. A similar argument as above yields that $\hat{\beta}_{\text{new}}^{(i+1)}=\beta_{\text{new}}^{(i+1)}+o_p(1)$. This ends the proof of Theorem 1(b).

References

Auger I, Lawrence C (1989) Algorithms for the optimal identification of segment neighborhoods. Bull Math Biol 51:39–54

Barry D, Hartigan JA (1992) Product partition models for change-point problems. Ann Stat 20:260–279

Barry D, Hartigan JA (1993) A Bayesian analysis for change point problems. J Am Stat Assoc 35:309–319

Chen J, Gupta AK (2012) Parametric statistical change point analysis with applications to genetics medicine and finance, 2nd edn. Birkhäuser, Boston

Csörgő M, Horváth L (1997) Limit theorems in change-point analysis. Wiley, Chichester

Erdman C, Emerson JW (2007) bcp: an R package for performing a Bayesian analysis of change point problems. J Stat Softw 23:1–13

Erdman C, Emerson JW (2008) A fast Bayesian change point analysis for the segmentation of microarray data. Bioinformatics 24:2143–2148

Harchaoui Z, Lévy-Leduc C (2008) Catching change-points with Lasso. Adv Neural Inf Process Syst 20:617–624

Harchaoui Z, Lévy-Leduc C (2010) Multiple change-point estimation with a total variation penalty. J Am Stat Assoc 105:1480–1493

Jackson B, Sargle J, Barnes D, Arabhi S, Alt A, Gioumousis P, Gwin E, Sangtrakulcharoen P, Tan L, Tsai TT (2005) An algorithm for optimal partitioning of data on an interval. IEEE Signal Process Lett 12:105–108

Jin B, Shi X, Wu Y (2013) A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. Stat Comput 23:221–231

Killick R, Eckley IA (2014) changepoint: an R package for changepoint analysis. J Stat Softw 58(3):1–19

Killick R, Eckley IA, Haynes K (2014) changepoint: An R package for changepoint analysis. R package version 1(1):5

Killick R, Fearnhead P, Eckley IA (2012) Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc 107:1590–1598

Lin D, Foster DP, Ungar LH (2011) VIF regression: a fast regression algorithm for large data. J Am Stat Assoc 106:232–247

Matteson DS, James NA (2013) A nonparametric approach for multiple change point analysis of multivariate data. J Am Stat Assoc 109:334–345

Olshen A, Venkatraman E, Lucito R, Wigler M (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. Biostatistics 5:557–572

Qu L, Tu Y (2006) Change point estimation of bilevel functions. J Mod Appl Stat Methods 5:347–355

Rigaill G (2010) Pruned dynamic programming for optimal multiple change-point detection. Technical Report, arXiv:1004.0887v1

Scott AJ, Knott M (1974) A cluster analysis method for grouping means in the analysis of variance. Biometrics 30:507–512

Seshan VE, Olshen A (2015) DNAcopy: DNA copy number data analysis. R package version 1(40)

Shi X, Wang X, Wei W, Wu Y (2015) VIFCP: detecting change-points via VIFCP method. R package version 1.0

Stransky N, Vallot C, Reyal F, Bernard-Pierrot I, Diez de Medina SG, Segraves R, de Rycke Y, Elvin P, Cassidy A, Spraggon C, Graham A, Southgate J, Asselain B, Allory Y, Abbou CC, Albertson DG, Thiery J-P, Chopin DK, Pinkel D, Radvanyi F (2006) Regional copy number-independent deregulation of transcription in cancer. Nat Genet 38:1386–1396