Robust semi-parametric multiple change-point detection

Jean-Marc Bardet, Charlotte Dion

Robust semi-parametric multiple change-point detection

Jean-Marc Bardet[(1)] and Charlotte Dion[(1),(2)]

[(1)] SAMM, Université Paris 1, 90 rue de Tolbiac, 75013 Paris
jean-marc.bardet@univ-paris1.fr
[(2)] LPSM, Sorbonne Université, 4 Place Jussieu, 75005 Paris
UMR CNRS 98001
charlotte.dion@upmc.fr

Abstract

This paper is dedicated to define two new multiple change-points detectors in the case of an unknown number of changes in the mean of a signal corrupted by additive noise. Both these methods are based on the Least-Absolute Value (LAV) criterion. Such criterion is well known for improving the robustness of the procedure, especially in the case of outliers or heavy-tailed distributions. The first method is inspired by model selection theory and leads to a data-driven estimator. The second one has a computational interest and it is based on total variation type penalty. These strategies are compared to standard alternatives in a numerical study.

Keywords. Change-points detection, Least-Absolute Value criterion.
AMS classification 62F86, 62F35, 62J05.

1 Introduction

The detection of change-points in time series has been studied a lot [8] and is very useful in many fields. One can cite finance [34], also genetic [25, 20], medical applications [14, 21], climatology [6] or agriculture [16].

In the sequel we consider a particular case of offline parametric multiple change-points detection. The framework is the following. Let $(y_1, \cdots, y_n)$ be an observed trajectory of a multiple mean process defined by:

$$y_t = \theta_k^* + \varepsilon_t, \quad t \in \{t_{k-1}^* + 1, \ldots, t_k^*\}, \quad k = 1, \ldots, D_m^*, \tag{1.1}$$

where D_m^* is the number of regimes (therefore $D_m^* - 1$ is the number of changes), the abrupt change instants are $0 = t_0^* < t_1^* < \cdots < t_{D_m^*-1}^* < t_{D_m^*}^* = n$ and $(\theta_1^*, \ldots, \theta_{D_m^*}) \in \mathbb{R}^{D_m^*}$ are the means of the process. The $(\varepsilon_k)_{k\in\mathbb{N}}$ is a white noise, *i.e.* a sequence of independent and identically distributed random variables (i.i.d.r.v.) with a positive continuous density, f_ε, at the neighborhood of zero. In addition we assume

$$\text{median}(\varepsilon_0) = 0 \iff F_{\varepsilon_0}^{-1}(1/2) = 0 \quad \mathbb{E}(\varepsilon_0) = 0. \tag{1.2}$$

If the median of the noise is non-zero, then the model is translated but the following stays. The model (1.1) can equivalently be defined with a functional formula:

$$y_t = s^*(t) + \varepsilon_t, \quad \text{for } t \in \{1, \ldots, n\} \quad \text{with} \quad s^* = \sum_{k=1}^{D_m^*} \theta_k^* \mathbf{1}_{I_{m^*}^*(k)} \tag{1.3}$$

where $I_{m^*}^*(k) = \{t_{k-1}^* + 1, \ldots, t_k^*\}$ for $k = 1, \ldots, D_m^*$ and s^* is the mean value function that is a piece-wise constant function. The goal of the paper is the detection (and the location) of change-points. This is a semi-parametric estimation problem since the distribution of ε is not supposed to

be known and the original signal is assumed to be piece-wise constant. Thus, this problem of change-points detection corresponds to the construction of estimators of the parameters $(D_m^*, (\theta_k^*)_k, (t_k^*)_k)$.

In the present work we consider a global approach by penalized contrast minimization, which means that all the change-points are detected simultaneously. As a consequence, the minimization is first realized for the contrast and for each number of change-points $D_m - 1$, leading to a collection of estimators $((\widehat{\theta}_k)_{1\leq k\leq D_m}, (\widehat{t}_k)_{1\leq k\leq D_m-1})$. The larger the number of changes $D_m - 1$ the smaller the contrast, therefore the penalty term provides a bias-variance trade-off that determines the optimal number of changes $\widehat{D}_m - 1$.

In this framework, estimators based on the Least-Squares (LS) contrast has received the most attention. It corresponds to the likelihood for the Gaussian framework, but it is also frequently used when the error term is not specified to be Gaussian. In [5], an estimation method based on LS criterion together with a testing method to find the good dimension are proposed. [27] derived the consistency and the rate of convergence of the change-points estimate from the LS criterion, in the situation where the number of changes is known. Then [26] proposed a penalized method to estimate the number of change-points which is unknown and their locations. [29] have developed a nonparametric strategy based on model selection tools to estimate D_m^*. This method is based on Gaussian model selection theory [13, 12]. [2] focuses on a cross-validation selection method and [1] is placed in the context of heteroscedastic data. Other nonparametric approaches exist, for example [23] proposed kernel estimators, also see [17] in a different framework. [32] investigate the fused LASSO procedure and showed some difficulties. Recently, [9] has built confidence sets in the special case of shape restricted regression.

But these methods based on classical LS regression rely on estimators that use means of sequences, and these estimators are not always ideal. Hence, estimators based on medians can be more relevant since they are significantly less sensitive to extreme values and therefore to outliers [as explained in 22]. This leads to consider the Least Absolute Value (LAV) criterion. The LAV criterion corresponds to the likelihood criterion for the Laplace framework, *i.e.* when the distribution of ε_t is a Laplace one. To the best of our knowledge, [4] is the most important contribution in this context: the consistency of the change-points estimators has been proved as well as the consistency of an estimator of D_m^* from a penalized LAV criterion (see below). If the variance of the noise is infinite the LAV criterion performed better than the LS criterion in the asymptotic context $n \to \infty$. This criterion is then often preferred [see for example 18]. Another paper [3] followed to the particular case of one shift. Note also that [35] chose a method based on LAV and total-variation for ℓ^1-trend filtering.

Our goal is to propose new LAV criteria to estimate a piece-wise constant signal. We first use a theoretical method based on model selection theory, inspired by [29], which produces an upper bound for the ℓ_1-risk. It is based on the following selection criterion: $\Gamma(m) = C(m, y) + \kappa \,\mathrm{pen}_n(m)$ where m represents a model that is a partition of n points with D_m regimes. The penalty function should depend only on n and the dimension D_m of the model m. The parameter κ is a trade-off parameter. Guided by the model selection literature this procedure leads to a non-asymptotic strategy. The choice of the penalty function is based on [7] and calibrated from a numerical study similarly to [29]. Finally, in order to obtain a totally data-driven procedure, we use the heuristic slope approach introduced in [2]. This provides an estimator $\widehat{\kappa}$ of the parameter κ and the estimator $\widehat{D}_m^{\mathrm{New}}$ of the number of regimes D_m^*.
Then we also propose a second method based on a Total-Variation approach (TV). It consists on a LAV deviation added to the penalty which depends on the difference of successive parameters, inducing that for a given segmentation the optimal parameters are not the minimizer of the LAV deviation. Hence the total variation approach does not provide a minimization of the ℓ^1-risk. But it gives an interesting estimator $\widehat{D}_m^{\mathrm{TV}}$ of the number of regimes D_m^*.

We realized Monte-Carlo experiments to compare both these new approaches as well as other classical ones. The results are extremely convincing for the new penalized LAV deviation estimator

$\widehat{D}_m^{\text{New}}$ with respect to other LAV or LS criteria. One can especially note the important gain in terms of robustness for this estimator as well as the TV procedure compared to the more usual least squares criterion. We also applied our new criteria to genomic data and the new LAV procedure leads to interesting conclusions.

In Section 2 the LAV deviation risk is introduced. Section 3 is devoted to a presentation of classical and new penalized LAV criteria. In Section 4 the total variation estimator is introduced and the algorithm point of view is detailed. Finally the numerical results for the new estimators are presented in Section 5, and testify of the accuracy and the robustness of the data-driven penalized LAV criterion.

2 The Least Absolute Value Deviation and its implementation

We begin with the construction of the estimators of parameters $(D_m^*, (\theta_k^*)_{1\leq k\leq D_m^*+1}, (t_k^*)_{1\leq k\leq D_m^*})$, which are respectively the number of regimes, the value on each segment and the change-points from the observed trajectory $(y_1, \ldots, y_n)$ defined in (1.1).

2.1 Notations

Here are the notations used in the following. The set of all the partitions of $\{1, \ldots, n\}$ is denoted $\mathcal{M}_n$. Then, for $m \in \mathcal{M}_n$, its length is $D_m = \text{Card}(m)$, where $m = \{I_m(1), I_m(2), \ldots, I_m(D_m)\}$ with $I_m(k) = \{t_{k-1}+1, \ldots, t_k\}$ for $k = 1, \ldots, D_m$, with $t_0 = 0$ and $t_{D_m} = n$. The true model induced by (1.3) is denoted m^* with $D_m^* = \text{Card}(m^*)$ and $(I_{m^*}(k))_{1\leq k\leq D_m^*}$ are the true segments.

The set of segmentations of $\mathcal{M}_n$ with $M \in \mathbb{N}^*$ points is

$$\mathcal{A}_{n,M} := \{\mathbf{t} = (t_1, \ldots, t_M), t_0 = 0 < t_1 < \ldots < t_M = n\}. \tag{2.1}$$

As a consequence, for $m \in \mathcal{M}_n$ we also have $m \in \mathcal{A}_{n,D_m}$. For $m \in \mathcal{M}_n$, $\mathcal{S}_m$ is the linear subspace of the piece-wise constant functions on m, *i.e.*,

$$\mathcal{S}_m := \left\{ \sum_{k=1}^{D_m} u_k \mathbf{I}_{I_m(k)}, \quad (u_k)_{1\leq k\leq D_m} \in \mathbb{R}^{D_m} \right\}. \tag{2.2}$$

For M a fixed integer number in $\{0, 1, \ldots, n\}$, we also define the subspace of piece-wise constant functions with M shifts, *i.e.*,

$$\mathcal{S}_M := \left\{ \sum_{k=1}^{M} u_k \mathbf{I}_{\{t_{k-1}+1,\ldots,t_k\}}, \quad (u_k)_{1\leq k\leq M+1} \in \mathbb{R}^{M+1}, \ (t_1, \ldots, t_M) \in \mathcal{A}_{n,M} \right\}. \tag{2.3}$$

Finally, for $M_{\max}$ a fixed integer number in $\{0, 1, \ldots, n\}$, define the subspace of piece-wise constant functions with less than $M_{\max}$ shifts by

$$\mathcal{S}^{M_{\max}} := \bigcup_{0\leq M\leq M_{\max}} \mathcal{S}_M. \tag{2.4}$$

2.2 Least Absolute Deviation criterion

For $(y_1, \ldots, y_n) \in \mathbb{R}^n$, we define the Least Absolute Value (LAV) distance or Least Absolute Deviation by

$$\widehat{\gamma}_n(u) = \frac{1}{n} \sum_{t=1}^{n} |y_t - u_t| =: \|y - u\|_{1,n}, \quad \text{for} \quad u = (u_t)_{1\leq t\leq n} \in \mathbb{R}^n. \tag{2.5}$$

For $m \in \mathcal{M}_n$, we define the LAV-contrast estimator $\widehat{s}_m$ of s^* defined in (1.3), by

$$\widehat{s}_m = \underset{s \in \mathcal{S}_m}{\operatorname{argmin}}\, \widehat{\gamma}_n(s) = \sum_{k=1}^{D_m} \widehat{\theta}_k^m \mathbf{1}_{I_m(k)}, \tag{2.6}$$

where $\widehat{\theta}_k^m$ is an empirical median defined by $\widehat{\theta}_k^m := \text{median}\{y_{t_{k-1}+1}, \ldots, y_{t_k}\}$. When the number of changes $M \in \mathbb{N}$ is specified and not the specific segmentation, we deduce an estimator of s^* by

$$\widehat{s}_M = \underset{s \in \mathcal{S}_M}{\operatorname{argmin}}\, \widehat{\gamma}_n(s) = \underset{\boldsymbol{\theta} \in \mathbb{R}^{M+1}}{\operatorname{argmin}} \underset{\boldsymbol{t} \in \mathbb{R}^M}{\operatorname{argmin}} \frac{1}{n} \sum_{k=1}^{M} \sum_{t=t_{k-1}+1}^{t_k} |y_t - \theta_k|. \tag{2.7}$$

Note that $\widehat{s}_M = \sum_{k=1}^{M} \sum_{t=\widehat{t}_{k-1}^M+1}^{\widehat{t}_k^M} \widehat{\theta}_k^M$ where $\widehat{\theta}_k^M = \text{median}\left\{y_{\widehat{t}_{k-1}^M+1}, \ldots, y_{\widehat{t}_k^M}\right\}$. In [4] was established the following asymptotic result.

Proposition 2.1 ([4]). *For model* (1.1) *with assumptions* (1.2), *if* $t_k^* - t_{k-1}^* \geq n^{3/4}$, *if there exists* $c > 0$ *such as* $\left|\theta_k^* - \theta_{k-1}^*\right| \geq c$, *then* $\widehat{t}_k^{M^*} \xrightarrow[n\to\infty]{\mathcal{P}} t_k^*$ *for any* $k \in \{1, \ldots, M^* - 1\}$.

This consistency result motivates the study of LAV-contrast estimators.

2.3 Dynamic programming

From a computing point of view, the dynamic programming algorithm is classically used to compute recursively the optimal paths, meaning, the collection $(\widehat{s}_M)$ for a given finite collection of $M \in \{0, 1, \ldots, M_{\max}\}$ [see 10]. It is based on the computation of the optimal cost $\widehat{C}_M(s,t)$ in M segments included in $\{s, s+1, \ldots, t\}$ for $s, t \in \mathbb{N}$, given by:

$$\widehat{C}_M(s,t) := \min_{t_0 = s < t_1 < t_2 < \cdots < t_{M-1} < t_M = t} \quad \min_{(\theta_k)_{1 \leq k \leq M} \in \mathbb{R}^M} \quad \frac{1}{n} \sum_{k=1}^{M} \sum_{j=t_{k-1}+1}^{t_k} \left|y_j - \theta_k\right|. \tag{2.8}$$

It is a two parts algorithm: the first part computes recursively the cost of the optimal segmentation with M changes for ℓ data for $1 \leq M \leq M_{\max}$ and $1 \leq \ell \leq n$; the second part is called *backtracking* and is used to find the optimal segmentation for each dimension [see the details in 19]. More formally we have:

Box 1 : Dynamic programming

1. Compute, in an iterative way, for $M = 1, \ldots, M_{\max}$ and $t \in \{M, \ldots, n\}$,

$$\widehat{C}_{M+1}(1,t) = \min_{s \in \{M+1, M+2, \ldots, t\}} \left\{\widehat{C}_M(1,s) + \widehat{C}_1(s+1,t)\right\}$$

2. Backtracking.

Finally, for each $M = 1, \ldots, M_{\max} + 1$, we obtain $C_M(1,n)$ and the change-points $t_1 < t_2 < \cdots < t_{M-1}$ that minimize $\widehat{C}_M(1,n)$. The time-consuming cost is $O(D_m N^2)$ instead of $O(\binom{N-1}{D_m})$ without the dynamic programming [see *e.g.* 24]. This algorithm gives finally one estimator by dimension, optimal for $\widehat{\gamma}_n$.

3 Estimation of the number of abrupt changes from penalized LAV criterion

3.1 Two first known dimension selection criteria

The previous cost allows to compute a robust estimator of $\big((t_k^*)_{1\leq k\leq M-1},(\theta_k^*)_{1\leq k\leq M}\big)$ for each $M\in\{1,\dots,M_{\max}\}$. But it is clear that $\widehat{C}_{M+1}(1,n)\leq\widehat{C}_M(1,n)$ and therefore other procedures are required for estimating D_m^* (meaning choosing M).

In the sequel we first assume that $D_m^*\leq M_{\max}$. Then a usual method for estimating D_m^* is to penalize the cost $\widehat{C}_M(s,t)$. This can be classically done using the following general criterion

$$\widehat{D}_m=\underset{1\leq M\leq M_{\max}}{\operatorname{argmin}}\Big\{f(\widehat{C}_M(1,n))+\kappa_n\operatorname{pen}(M)\Big\}=\underset{1\leq M\leq M_{\max}}{\operatorname{argmin}}\Big\{f(\widehat{\gamma}_n(\widehat{s}_M))+\kappa_n\operatorname{pen}(M)\Big\},\quad(3.1)$$

with an increasing function f, a sequence of penalization parameters $(\kappa_n)_n\in(0,\infty)^{\mathbb{N}}$ and a penalty function $k\in\mathbb{N}\mapsto\operatorname{pen}_n(k)$, which is also an increasing function.

For example [4] proposes to select $\widehat{D}_m^{\mathrm{BAI}}$ defined by

$$\widehat{D}_m^{\mathrm{BAI}}=\underset{1\leq M\leq M_{\max}}{\operatorname{argmin}}\Big\{\log\big(\widehat{\gamma}_n(\widehat{s}_M)\big)+\frac{\sqrt{n}}{n}\,M\Big\}.$$

The author choose to use $\kappa_n=\frac{\sqrt{n}}{n}$ for insuring the consistency of $\widehat{D}_m^{\mathrm{BAI}}$ to D_m^*. But $\sqrt{n}$ could also be replaced by any increasing sequence with infinite limit and bounded by $n\mapsto\sqrt{n}$.
Then, with $\kappa_n=\frac{\log(n)}{n}$, we could also consider the classical BIC penalty defined by:

$$\widehat{D}_m^{\mathrm{BIC}}=\underset{1\leq M\leq M_{\max}}{\operatorname{argmin}}\Big\{\log\big(\widehat{\gamma}_n(\widehat{s}_M)\big)+\frac{\log(n)}{n}\,M\Big\}.$$

Note that there is no heuristic justification for using this criterion because the usual Laplace approximation is no longer valid for non-differentiable functions [see 30, 28].

3.2 A new data-driven oracle penalization

In order to define a new data-driven penalized least absolute values estimator we first follow some non asymptotic results on model selection developed in [11, 7]. Hence the following general estimator can be considered

$$\widehat{m}:=\underset{m\in\mathcal{M}_n,u\in\mathcal{S}_m}{\operatorname{argmin}}\{\widehat{\gamma}_n(u)+\operatorname{pen}_n(D_m)\}\quad(3.2)$$

where $\operatorname{pen}_n(D_m)$ only depends on D_m and n. The differences with the previous methods are that the minimization is done (theoretically) on all the models and the penalization function $\operatorname{pen}_n(D_m)$ is not necessary a linear function of D_m.

First, let us recall the general result of [7] (Theorem 8 and 11) [see also 11, for details on LAV].

Theorem 3.1 (Barron Birgé Massart (1999))**.** *Assume that there exist $\Sigma>0$ and a family of weights $(L_m)_{m\in\mathcal{M}_n}$, such that $L_m\geq 1$ and $\sum_{m\in\mathcal{M}_n}\exp(-L_m\,D_m)\leq\Sigma$ for any $n\in\mathbb{N}$. With $K>0$, define also the penalty function $pen_n(\cdot)$ such as*

$$\operatorname{pen}_n(D_m)\geq K\,(L_m+\mathcal{L}_m)\frac{D_m}{n},\quad\textit{where}\quad\mathcal{L}_m=\log\left[c\Big(1+c'\Big(\frac{D_m}{n}\Big)^{1/2}\Big)\right]+1$$

with $c,c'>0$. Then there exist C, $C'\in(0,\infty)$ depending on σ and $\kappa>0$ such as $\widetilde{m}$ defined in (3.2) *satisfies*

$$\mathrm{E}[d(s^*,\widehat{s}_{\widehat{m}})^2]\leq\kappa\inf_{m\in\mathcal{M}_n}\Big\{d(s^*,\mathcal{S}_m)^2+C\operatorname{pen}_n(D_m)\Big\}+C'\,\frac{\Sigma}{n}.$$

In the present framework it is classical to choose variable weights, depending only on the dimension of the model $L_m = L_{D_m}$. Using the same computation done in [29] we obtain

$$\Sigma = \sum_{m \in \mathcal{M}_n} e^{-L_m D_m} \leq \sum_{d=1}^{n} e^{-d(L_d - 1 - \log(n/d))}.$$

Then, with $\theta > 0$, we can choose, using a counting argument,

$$L_d = 1 + \theta + \log(n/d) \quad \text{for any } d \in \mathbb{N}.$$

We deduce from this result the following bound for the considered risk.

Proposition 3.2. *For any given positive real numbers c_1 and c_2, define for any $m \in \mathcal{M}_n$*

$$\mathrm{pen}_n(D_m) := \sigma^2 \frac{D_m}{n} \left(c_1 \log\left(\frac{n}{D_m}\right) + c_2 \right) \tag{3.3}$$

there exist two positive constants $C(c_1, c_2)$, $C'(c_1, c_2)$ such that $\widehat{m}$ defined in (3.2) satisfies

$$\mathrm{E}[\|s^* - \widehat{s}_{\widehat{m}}\|_{1,n}] \leq \kappa \inf_{m \in \mathcal{M}_n} \left\{ d(s^*, \mathcal{S}_m)^2 + C(c_1, c_2)\, \mathrm{pen}_n(D_m) \right\}^{1/2} + C'(c_1, c_2) \sqrt{\frac{\Sigma}{n}}. \tag{3.4}$$

This result is obtained from the usual inequality $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} |X_i|\right] \leq \mathrm{E}[\|X\|_2^2]^{1/2}$ and with the relationship $d^2(s^*, \mathcal{S}_m) = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n} |s^*(t) - s_m^*(t)|^2 = \|s^* - s_m^*\|_{2,n}^2$ where s_m^* is the orthogonal projection of s^*, this segmentation is obtained with the mean of s^* taken on each segment (and not the median). The logarithm appears in (3.3) because of the complexity of the collection of models, meaning the huge dimension (there are $\binom{n}{M}$ possible segmentations of length M). This result is a non-asymptotic one. Besides, the choice of constants c_1 and c_2 could be done through an extensive simulation study as it was done in [29]. After those Monte-Carlo experiments, we have chosen $c_1 = 1$ and $c_2 = 2$.

Remark 3.3. *This results could be improved. Indeed, the result comes from the bound obtained for the quadratic loss (the ℓ^2-distance). The main difficulty with the LAV criterion, which differs from the LS criterion is that the theoretical loss function that the literature encourages to consider is*

$$\begin{aligned} \ell(s^*, u) &:= \mathrm{E}\big[\widehat{\gamma}_n(u) - \widehat{\gamma}_n(s^*)\big] = \mathrm{E}\big[\|y - u\|_{1,n}\big] - \mathrm{E}\big[\|y - s^*\|_{1,n}\big] \\ &= \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}\big[|s^*(t) - u(t) + \varepsilon_t| - |\varepsilon_t|\big] \end{aligned}$$

which satisfies

$$\|s^* - u\|_{1,n} - \frac{2}{n}\sum_{t=1}^{n} \mathrm{E}[|\varepsilon_t|] \leq \ell(s^*, u) \leq \|s^* - u\|_{1,n}$$

but not with the equality. In the LS case, we have that

$$\ell(s^*, u) := \mathrm{E}\big[\widehat{\gamma}_n(u) - \widehat{\gamma}_n(s^*)\big] = \mathrm{E}\big[\|y - u\|_{2,n}^2\big] - \mathrm{E}\big[\|y - s^*\|_{2,n}^2\big] = \mathrm{E}\big[\|u - s^*\|_{2,n}^2\big].$$

This makes the issue more challenging and could be the subject of further works.

However the unknown constant σ^2 is still present in the definition (3.3) of the penalization. We chose to estimate this constant using heuristic slope method introduced in [12, 2]. It consists on computing the graph $\left(\frac{M}{n}\left(c_1 \log\left(\frac{n}{M}\right) + c_2\right), \gamma_n(\widehat{s}_M)\right)$ for $1 \leq M \leq M_{\max}$. On such graph one can see an abrupt change of regime for M going from 1 to D_m^*, and a linear decrease for $M > D_m^*$. Using a classical offline change detection for linear models, the slope α of the linear part of the graph can be estimated by $\widehat{\alpha}$. The main idea of the slope heuristic is to consider the new estimator of the number of regimes D_m^* by

$$\widehat{D}_m^{\mathrm{NEW}} := \operatorname*{argmin}_{1 \leq M \leq M_{\max}} \left\{ \widehat{\gamma}_n(\widehat{s}_M) - 2\widehat{\alpha}\, \frac{M}{n} \left(c_1 \log\left(\frac{n}{M}\right) + c_2 \right) \right\}. \tag{3.5}$$

Hence we obtain a new data-driven estimator of the number of abrupt changes.

4 A Total Variation criterion

The developed criterion is called a convex Total Variation (TV) criterion (or sometimes fused lasso), since for any $\lambda > 0$, it is defined by

$$\begin{aligned}\widehat{s} &= \underset{1\leq M\leq M_{\max},(\theta_k)_{1\leq k\leq M}\in\mathbb{R}^M,(t_1,\ldots,t_M)\in\mathcal{A}_{n,M}}{\operatorname{argmin}} \left\{\sum_{t=1}^{n}\left|y_t-\sum_{k=1}^{M}\theta_k\mathbb{1}_{t_{k-1}+1\leq t\leq t_k}\right|+\lambda\sum_{k=2}^{M}\left|\theta_k-\theta_{k-1}\right|\right\}\\ &=: \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}}\ \underset{(\theta_k)_{1\leq k\leq M}\in\mathbb{R}^M,(t_1,\ldots,t_M)\in\mathcal{A}_{n,M}}{\operatorname{argmin}} \xi_\lambda^M((\theta_k)_{1\leq k\leq M},(t_1,\ldots,t_M)).\end{aligned}$$

The total variation allows to measure the variability of the sequence of $(\theta_k)_k$. The second term in the right hand side of the sum is the ℓ_1-norm of the first-difference sequence of $(\theta_k)_k$ and it can be seen as a convex approximation of the number of changes and it should tend towards a reduction of it.

For a fixed dimension, the segmentation is the same than the one obtained with the classical LAV criterion. Then, the estimated parameters are different according to the total variation penalty term, and finally the dimension parameter is chosen.

For each λ, which is the tuning parameter of the TV penalization, one would like to solve $\underset{(\theta_k)_{1\leq k\leq M}\in\mathbb{R}^M,(t_1,\ldots,t_M)\in\mathcal{A}_{n,M}}{\operatorname{argmin}} \xi_\lambda^M((\theta_k)_{1\leq k\leq M},(t_1,\ldots,t_M))$ using the dynamic programming (see Box 1). But, the cost matrix $\widehat{C}$ depending on λ cannot be explicit this time, and this would notably improves the complexity of such a method.

An alternative solution is to compute first, for each M, the segmentation minimizing the least absolute value criterion with the dynamic programming and obtain the vector $(\widehat{t}_k)_k$ for each dimension M in the collection. Secondly, the following minimization problem can be solved:

$$\left(\widehat{D}_m^\lambda,(\theta_k^\lambda)_{1\leq k\leq\widehat{D}_m^\lambda}\right) = \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}}\ \underset{(\theta_k)_{1\leq k\leq M}\in\mathbb{R}^M}{\operatorname{argmin}} \xi_\lambda^M\left((\theta_k)_{1\leq k\leq M},(\widehat{t}_1,\ldots,\widehat{t}_M)\right)$$

For any $\lambda > 0$ and $1 \leq M \leq M_{\max}$, a numerical approximation of the solution $(\theta_k^\lambda)_{1\leq k\leq M}$ can be done using the Alternating Direction Method of Multipliers (ADMM). The principle of the algorithm and its convergence are given in [15] [see 36, 31, for examples]. The ADMM algorithm rewrites the minimization problem over θ as an equality constraint optimization problem where $\theta = (\theta_1,\ldots,\theta_M)$ is split in two parts θ and γ. It is based on the formulation:

$$\begin{aligned}\underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}} \left\{\sum_{t=1}^{n}\left|y_t-\sum_{k=1}^{M}\theta_k\mathbb{1}_{\{\widehat{t}_{k-1}+1\leq t\leq\widehat{t}_k\}}\right|+\lambda\sum_{k=2}^{M}\left|\theta_k-\theta_{k-1}\right|\right\} &= \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}}\left\{f(\theta)+\lambda\, g(A\,\theta)\right\}\\ &= \underset{\substack{\theta\in\mathbb{R}^M,\gamma\in\mathbb{R}^{M-1}\\ A\,\theta=\gamma}}{\operatorname{argmin}}\left\{f(\theta)+\lambda\, g(\gamma)\right\}\end{aligned}$$

with

$$A=\begin{pmatrix}-1 & 1 & 0 & \ldots\\ 0 & -1 & 1 & \ldots\\ \vdots & \vdots & \vdots & \vdots\\ \ldots & 0 & -1 & 1\end{pmatrix}\in\mathcal{M}_{(M-1,M)}(\mathbb{R}) \quad\text{and}\quad g(x_1,\ldots,x_M)=\sum_{k=1}^{M}|x_k|.$$

This leads to consider the following algorithm:

Box2 : ADMM algorithm

1. Initialization.
2. $\theta^{(\ell+1)} = \underset{\theta\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ f(\theta) + \frac{\rho}{2}\|A\theta - \gamma^{(\ell)} + \rho^{-1}\alpha^{(\ell)}\|^2 \right\}$
3. $\gamma^{(\ell+1)} = \underset{\gamma\in\mathbb{R}^{M-1}}{\operatorname{argmin}} \left\{ \lambda g(\gamma) + \frac{\rho}{2}\|A\theta^{(\ell+1)} - \gamma + \rho^{-1}\alpha^{(\ell)}\|^2 \right\}$
4. $\alpha^{(\ell+1)} = \alpha^{(\ell)} + \rho(A\theta^{(\ell+1)} - \gamma^{(\ell+1)})$.

Here, ρ is the augmented Lagrangian parameter, and the ADMM consists in applying the previous steps. In practice, $\rho = 1$. We remark also that the choice of the tuning parameter λ is crucial. In practice λ could be selected using the BIC criterion. This penalty has been used in [14] in the regression case and the consistency of the change-points estimator is established when the number of regressor tends to infinity with fixed n (see also [32]).

Nevertheless the approach developed here is slightly different because we impose the length of the vector θ, imposing the number of change-points (and this reduces the number of possible models as $M_{\max} << n$). Then a minimization with respect to the number of stages M is done.

5 Numerical illustrations

In the section we provide first details about Monte-Carlo experiments allowing to compare the different criteria as well as the numerical implementation of the different methods. Then a real-life dataset is studied using the new criteria.

5.1 Presentation of the Monte-Carlo experiments

We led a large simulation study, investigating different kind of signals, from different distributions of noise (ε) and different lengths (n). We choose $n \in \{50, 200, 500\}$, $D_{M_{\max}} = 40$. Signals are simulated randomly (the change-points, the parameters values) and the resulting estimators are compared with the oracle estimator (available on simulations). In the following in order to illustrate our purpose we choose four different distributions of the noise with the same variance σ^2:

- Gaussian noise, denoted $\mathcal{N}$, with density $\phi_{(0,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\Big(-\frac{1}{2\sigma^2}\, x^2 \Big)$;
- Laplace noise, denoted $\mathcal{L}$, with density $f(x) = \frac{1}{2\sqrt{2}}\sigma \exp(-|x|/\sigma)$;
- Normalized Student noise, denoted $\mathcal{S}$, with 3 degrees of freedom, *i.e.* $\sqrt{3\sigma^2}\, t(3)$, where $t(3)$ is the classical Student distribution with 3 degrees of freedom;
- Mixture of Gaussian noises, denoted M$\mathcal{N}$, defined by

$$f_\varepsilon(x) = (1-p)\,\phi_{(0,\gamma^2)}(x) + \frac{p}{2}\,\phi_{(-\mu,\gamma^2)}(x) + \frac{p}{2}\,\phi_{(\mu,\gamma^2)}(x),$$

  with $\mu = \frac{mp}{\sqrt{m^2p+\sigma^2}}$ and $\gamma^2 = \frac{\sigma^4}{m^2p+\sigma^2}$. The distribution of this noise contains 3 modes and can mimic the presence of outliers. In the sequel we use $p = 1/10$ and $m = 10$.

We investigate and illustrate first the example given in [4] with $D_m^* = 4$ change-points and 4 regimes with parameters $\theta^* = (1, 3, 1, -1)$. Then we consider the case $D_m^* = 7$ with $\theta^* = (1, 3, 1, -1, 1, -3, -1)$ and finally the case of randomized values of D_m^*, $(\theta_i^*)_i$ and $(t_i^*)_i$ has been studied. More precisely, the parameters are simulated according to the following scheme:

- the number of changes $D_m^* - 1$ is simulated from a binomial distribution with parameters $(6, 0.5)$,

	$n=50$			$n=200$			$n=500$		
$\sigma=1$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	**65.4**	57.2	57.7	87.8	**99.8**	70.0	92.1	**100**	74.9
$\mathcal{L}$	67.8	**84.1**	69.5	90.5	**100**	85.4	95.7	**100**	90.6
$\mathcal{S}$	67.9	**96.1**	70.2	88.7	**100**	83.4	94.1	**100**	83.4
M$\mathcal{N}$	47.2	**92.3**	72.9	38.6	**99.6**	48.6	48.1	**100**	49.8
$\sigma=2$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	19.1	1.3	**22.1**	63.5	2.9	**64.7**	**92.5**	31.8	77.3
$\mathcal{L}$	32.6	3.9	**34.7**	**88.4**	34.2	84.4	**96.1**	95.9	90.8
$\mathcal{S}$	19.7	1.3	**22.2**	**64.5**	2.6	64.4	**92.3**	33.0	77.0
M$\mathcal{N}$	41.1	7.9	**43.2**	**81.2**	63.0	77.9	91.7	**99.7**	85.8

Table 1: Empirical score in % with 10000 repetitions and $D_m^* = 4$

- the change-points are uniformly distributed $\mathcal{U}_{[\lfloor\sqrt{N}/2, N-\rfloor\sqrt{N}/2]}$, under the constraint that the difference between two successive times must be at least $\sqrt{N}/4$,
- the parameters values θ_i^* for each regime are Gaussian $\mathcal{N}(0,1)$ under the constraint to be separated by 1 at least.

The procedures are evaluated in two terms:

- with the empirical score in % , *i.e.* the frequencies of estimation of the true values of D_m^*;
- with the empirical ℓ_1-risk.

Both values illustrate a question: what is more important between minimizing the distance between the estimated and the true signals, and finding the "true" number of break-points ?

5.2 Comparisons of criteria based on penalized LAV deviations

In this subsection, we investigate the estimators $\widehat{D}_m^{\text{Bai}}$, $\widehat{D}_m^{\text{BIC}}$ and $\widehat{D}_m^{\text{New}}$, detailed in Section 3. The results are given in Tables 1-6.

General purposes can be first deduced from the results of Monte-Carlo experiments for the three penalized LAV criteria $\widehat{D}_m^{\text{New}}$, $\widehat{D}_m^{\text{Bai}}$ and $\widehat{D}_m^{\text{BIC}}$.

These experiments exhibit the consistency of these three LAV criteria when n increases. Moreover, the larger σ^2 (or D_m^*) the smaller the empirical score and the larger the empirical ℓ_1-risk. Let us note that the Gaussian noise, which has the flattest distribution tail, gives the least accurate results.

The estimators can also be compared through their ability to estimate the true number of changes. Indeed, different configurations are presented. In "easy" conditions, meaning that D_m^* is small and also is the variance σ^2, and the differences $|\theta_{i+1}^* - \theta_i^*|$ and $t_{i+1}^* - t_i^*$ are large (and therefore large n), then the criterion $\widehat{D}_m^{\text{Bai}}$ provides excellent results. This is not surprising since it has been defined in [4] in an asymptotic framework,

On the contrary, in "difficult" conditions, meaning that D_m^* and the variance σ^2 are large, and differences $|\theta_{i+1}^* - \theta_i^*|$ and $t_{i+1}^* - t_i^*$ are small, then $\widehat{D}_m^{\text{New}}$ and $\widehat{D}_m^{\text{BIC}}$ provide much better results than $\widehat{D}_m^{\text{Bai}}$.

Generally speaking, estimator $\widehat{D}_m^{\text{New}}$ offers the best trade-off as it can be observed in the case of randomized choice of parameters.

Concerning the empirical ℓ_1-risk, the conclusions are almost the same, except that $\widehat{D}_m^{\text{New}}$often provides the minimal risk even when $\widehat{D}_m^{\text{Bai}}$ obtains the best empirical score (see typically the case $D_m^* = 4$).

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	**0.42**	0.52	0.42	**0.17**	0.17	0.19	**0.10**	0.11	0.11
$\mathcal{L}$	**0.28**	0.30	0.28	0.11	**0.10**	0.11	**0.06**	0.06	0.06
$\mathcal{S}$	0.37	0.49	**0.34**	**0.14**	0.14	0.14	**0.08**	0.08	0.08
M$\mathcal{N}$	0.46	0.61	**0.40**	**0.15**	0.15	0.15	**0.08**	0.08	0.09
$\sigma = 2$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	**0.93**	0.99	0.93	0.50	0.78	**0.47**	**0.24**	0.56	0.25
$\mathcal{L}$	**0.72**	0.88	0.73	**0.25**	0.54	0.25	**0.13**	0.15	0.14
$\mathcal{S}$	**0.93**	0.99	0.93	0.51	0.78	**0.48**	**0.23**	0.56	0.25
M$\mathcal{N}$	**0.62**	0.82	0.64	**0.22**	0.39	0.23	**0.12**	0.12	0.13

Table 2: Empirical ℓ_1-risk with 10000 repetitions and $D_m^* = 4$

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	25.5	6.0	**39.0**	**92.6**	71.6	64.4	95.8	**99.8**	69.5
$\mathcal{L}$	54.0	22.7	**56.8**	92.7	**97.9**	79.8	94.7	**100**	87.9
$\mathcal{S}$	**69.9**	42.4	63.2	90.6	**99.7**	79.3	92.3	**100**	86.4
M$\mathcal{N}$	54.0	22.7	56.7	92.7	97.9	79.7	94.7	100	87.5
$\sigma = 2$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	3.7	0	**4.8**	10.9	0	**25.7**	**75.3**	0	67.8
$\mathcal{L}$	6.5	0.0	**7.7**	**54.4**	0.0	61.9	**95.6**	10.3	87.9
$\mathcal{S}$	9.0	0.1	**11.4**	**71.3**	0.7	71.2	**94.2**	34.1	85.6
M$\mathcal{N}$	**11.5**	0.2	**11.5**	**69.9**	0.9	66.2	**93.2**	45.7	80.6

Table 3: Empirical score in % with 10000 repetitions with $D_m^* = 7$

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	0.70	0.86	**0.56**	**0.23**	0.34	0.25	0.14	**0.13**	0.15
$\mathcal{L}$	**0.39**	0.62	0.40	**0.15**	0.15	0.15	**0.08**	0.08	0.09
$\mathcal{S}$	0.37	0.49	**0.33**	**0.14**	0.14	0.14	**0.08**	0.08	0.08
M$\mathcal{N}$	0.46	0.62	**0.40**	**0.15**	**0.15**	**0.15**	**0.08**	**0.08**	0.09
$\sigma = 2$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	**1.11**	1.19	**1.11**	0.76	0.98	**0.70**	0.38	0.88	**0.36**
$\mathcal{L}$	0.97	1.16	**0.96**	0.45	0.88	**0.41**	**0.19**	0.52	0.20
$\mathcal{S}$	0.89	1.08	**0.87**	0.38	0.80	**0.36**	**0.19**	0.41	0.19
M$\mathcal{N}$	0.87	1.09	**0.86**	**0.34**	0.79	**0.34**	**0.17**	0.35	0.18

Table 4: Empirical ℓ_1-risk with 10000 repetitions and $D_m^* = 7$

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	16.9	24.4	**39.0**	66.5	42.7	**72.0**	83.5	49.0	**86.3**
$\mathcal{L}$	**44.6**	42	35.9	**79.0**	56.0	60.7	**92.1**	76.1	75.8
$\mathcal{S}$	**49.7**	31.0	35.0	**82.9**	53.8	38.7	**94.0**	60.2	31.9
M$\mathcal{N}$	**55.9**	40.5	29.9	**73.9**	69.3	55.52	**91.0**	75.5	66.2
$\sigma = 2$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	17.3	14.1	**20.5**	29.5	20.0	**36.6**	53.2	20.5	**55.2**
$\mathcal{L}$	**23.0**	20.3	19.4	38.9	**39.4**	30.04	**62.7**	55.6	43.9
$\mathcal{S}$	**25.5**	23.5	21.1	**45.7**	39.6	23.7	**67.1**	45.8	23.5
M$\mathcal{N}$	**25.9**	10.17	25.0	46.5	18.4	**50.3**	65.6	27.4	**71.6**

Table 5: Empirical score in % with 10000 repetitions with randomized D_m^*

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	**0.18**	0.21	0.19	**0.18**	0.24	0.19	**0.10**	0.14	0.11
$\mathcal{L}$	0.29	0.35	**0.28**	**0.11**	0.15	0.11	**0.06**	0.08	0.06
$\mathcal{S}$	0.26	0.31	**0.25**	**0.10**	0.13	0.11	**0.06**	0.07	0.06
M$\mathcal{N}$	0.17	0.18	**0.16**	**0.06**	**0.06**	0.08	**0.03**	0.04	0.05
$\sigma = 2$	New	Bai	BIC	New	Bai	BIC	New	Bai	BIC
$\mathcal{N}$	0.80	**0.78**	**0.78**	**0.44**	0.57	0.44	**0.26**	0.42	0.26
$\mathcal{L}$	0.62	0.66	**0.60**	0.29	0.43	**0.27**	0.15	0.28	**0.14**
$\mathcal{S}$	0.57	0.63	**0.56**	0.27	0.39	**0.26**	**0.14**	0.25	0.14
M$\mathcal{N}$	0.56	0.62	**0.55**	**0.25**	0.38	**0.25**	0.14	0.25	**0.13**

Table 6: Empirical ℓ_1-risk , with 10000 repetitions and randomized D_m^*

	$n=50$			$n=200$			$n=500$		
$\sigma=1$	New	TV	LS	New	TV	LS	New	TV	LS
$\mathcal{N}$	16.9	24.4	**39.0**	66.5	42.7	**72.0**	83.5	48.8	**86.3**
$\mathcal{L}$	**44.6**	42.3	35.9	**79.0**	56.0	60.6	**92.1**	76.1	75.8
$\mathcal{S}$	**49.7**	31.2	35.0	**82.9**	53.8	38.7	**94.0**	60.2	31.8
M$\mathcal{N}$	**55.9**	40.5	29.9	**73.9**	69.3	55.5	**91.0**	75.5	66.2
$\sigma=2$	New	TV	LS	New	TV	LS	New	TV	LS
$\mathcal{N}$	17.3	14.1	**20.5**	29.5	20.0	**36.6**	53.2	20.5	**55.2**
$\mathcal{L}$	**23.0**	14.5	19.4	38.9	**39.4**	30.0	**62.8**	48.2	43.8
$\mathcal{S}$	**25.5**	23.5	21.1	**45.7**	39.6	23.7	**67.1**	45.8	23.5
M$\mathcal{N}$	**25.9**	14.8	18.5	**46.5**	28.9	31.1	**65.6**	38.0	47.3

Table 7: Empirical score with 10000 repetitions with randomized D_m^*

5.3 Comparison with total variation and least-squares criteria

In the sequel the new data-driven estimator $\widehat{D}_m^{\text{New}}$ is compared to two criteria, which are not based on penalized LAV deviation. The first one is provided from a classical approach, the LS offline detector presented in Introduction 1 and defined by

$$\widehat{\gamma}_n^{LS}(u) = \frac{1}{n}\sum_{t=1}^{n}(y_t - u(t))^2 = \|y-u\|_{2,n}^2. \tag{5.1}$$

For a given segmentation m, the LS estimator is

$$\widehat{s}_m^{LS} \quad = \quad \underset{u\in\mathcal{S}_m}{\operatorname{argmin}}\, \widehat{\gamma}_n^{LS}(u) = \sum_{k=1}^{D_m} \overline{y}_k \mathbf{1}_{I_k}, \quad \overline{y}_k = \frac{1}{n_k}\sum_{t=t_{k-1}+1}^{t_k} y_t,\ n_k = \text{Card}\{t\in I_m(k)\}. \tag{5.2}$$

The estimators are computed using the Dynamic Programming again, to obtain one model by dimension. Then, to select the best estimator in the collection the selected dimension is the one minimizing the sum of two terms: $\underset{1\le M\le M_{\max}}{\operatorname{argmin}} \{\widehat{\gamma}_n^{LS}(\widehat{s}_M^{LS}) + \text{pen}(M)\}$ with $\text{pen}^{LS}(M) = \frac{M}{n}\sigma^2\left(2\log\left(\frac{n}{M}\right)+5\right)$. according to [29], for a Gaussian noise. To compute the estimator called $\widehat{D}_m^{\text{LS}}$ we use again the slope heuristic following the same process as for $\widehat{D}_m^{\text{New}}$.

We also implement the estimator $\widehat{D}_m^{\text{TV}}$ obtained from the total variation (TV) criterion, described in Section 4. Different programming steps are followed: first the dynamic programming to get one segmentation by dimension, then the ADMM algorithm to optimize the minimization of the criterion on θ and finally the selection of the best dimension. Note that the asymptotic properties of the estimator $\widehat{D}_m^{\text{TV}}$ have still not be studied. Nevertheless, the following Table 7 exhibits its convergence when n increases. Note that we do not exhibit the ℓ^1-risk scores since the TV criterion is only devoted to select the number of change-points and not to minize the risk. The empirical scores are given in Table 7 where we also consider the previous version of randomized values of D_m^*, $(\theta_i)_i$ and $(t_i)_i$.

It appears first from Table 7 that $\widehat{D}_m^{\text{New}}$ provides the most accurate estimations, except for Gaussian time series for which the classical LS criterion is still the most interesting. This is not a surprise since in this Gaussian case the LS criterion can be derived from maximum likelihood estimation, while $\widehat{D}_m^{\text{New}}$ can also be derived but for Laplace distribution. And this confirms the well known robustness of LAV estimation with respect to the LS one. Figure 1 and Figure 2 illustrates this purpose about the robustness of criteria. In practice it is a common fact to observe pics values on real data set and instead of truncate them, the LAV criterion can deal with them without creating artificial new regimes.

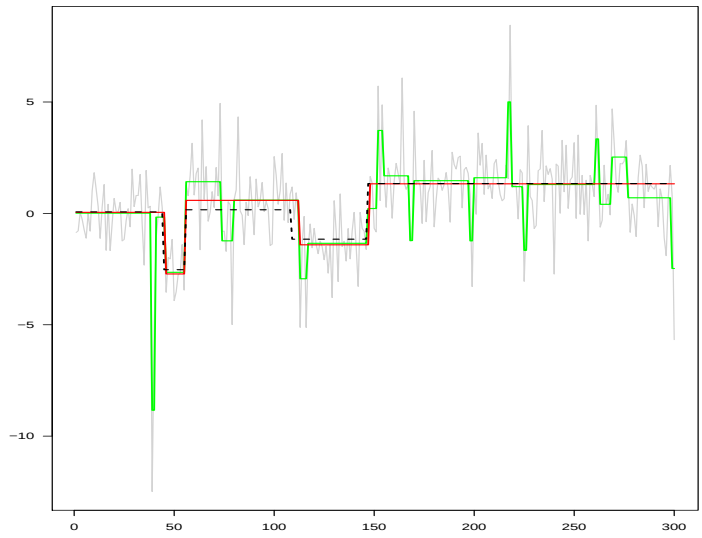

Figure 1: Comparison between LAV and LS criteria ($\widehat{s}_{\widetilde{m}}$ vs $\widehat{s}^{LS}_{\widehat{m}^{LS}}$): red line for LAV new estimator, green line for LS penalized estimator. Student case, $\sigma = 2$, signal y in light grey, true signal s^* il dotted black line.

Finally, $\widehat{D}_m^{\mathrm{TV}}$ provides an interesting alternative to LS criterion except for Gaussian processes. But it is quite always less efficient than $\widehat{D}_m^{\mathrm{New}}$. However, this criterion is built on the same principle than the LASSO criterion and we can suspect that it could especially be useful when K^* is really large and not negligible with respect to the data length n.

5.4 Application to genomic data

We also apply new criteria on a real-life data set, which consists on normalized copy-number lograties of data array CGH study set of Corriel institute taken from the package `DNAcopy` of V. Seshan and A. Olshen, see also [33] (the authors have assembled arrays of around 2400 clones of DNA copy number across the human genome). These data and their analysis help to detect chromosomal copy number variations which could cause disorders. We apply the previous two new

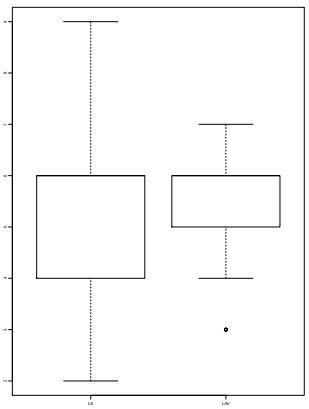

Figure 2: Boxplots of $\widehat{D}_m^{\mathrm{LS}}$ and $\widehat{D}_m^{\mathrm{New}}$ for Student case, $N = 200$, $D^* = 7$

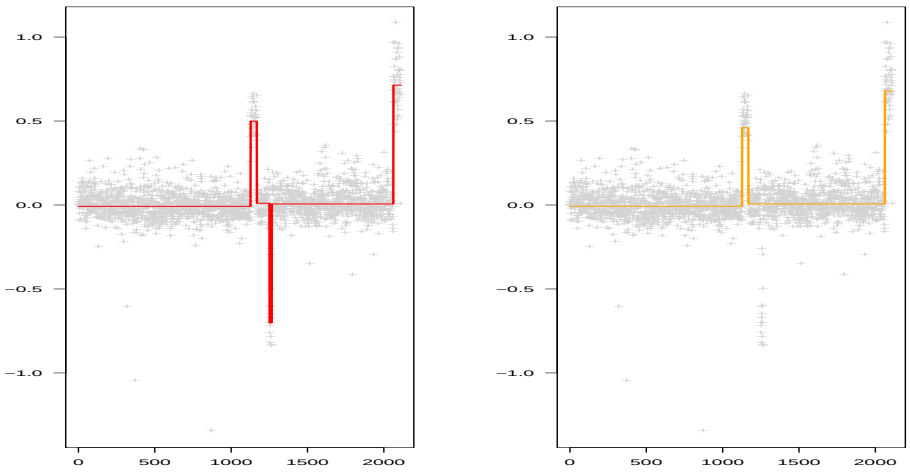

Figure 3: Grey: normalized copy-number logratio, left: red $\widehat{s}_{\widehat{D}_m^{\mathrm{New}}}$, right: orange $\widehat{s}_{\widehat{D}_m^{\mathrm{NTV}}}$

strategies on some part on the data. The results are presented in Figure 3: on the left the true data are plotted together with estimator $\widehat{s}_{\widehat{D}_m^{\mathrm{New}}}$ in red, and on the right the same graph with $\widehat{s}_{\widehat{D}_m^{\mathrm{TV}}}$ in orange (and the LS estimator $\widehat{s}_{\widehat{D}_m^{\mathrm{LS}}}$ is barely equal to $\widehat{s}_{\widehat{D}_m^{\mathrm{New}}}$ in this case). Here the decomposition of the signal obtained with the new LAV criterion procedure seems to fit well the data, nevertheless the TV estimator may have removed an artifact of the data (large variance) choosing only 3 change-points. However, the biological context and medical knowledge are required to interpret the results.

Acknowledgment

The authors are very grateful to the Foundation des Sciences Mathématiques de Paris (FSMP), which provided a grant for Charlotte Dion's postdoc. They also thank Pascal Massart (University Paris-Sud) for fruitful advise.

References

[1] S. Arlot and A. Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, 21(4):613–632, 2011.

[2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

[3] J. Bai. Least absolutes deviation estimation of a shift. *Econometric Theory*, 11:403–436, 1995.

[4] J. Bai. Estimation of multiple-regime regressions with least absolutes deviation. *Journal of Statistical Planning and Inference*, 74(1):103–134, 1998.

[5] J. Bai and P Perron. Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78, 1998.

[6] R.T Baillie and S.K. Chung. Modeling and forecasting from trend-stationary long memory models with applications to climatology. *Int. J. Forecasting*, 19:215–226, 2002.

[7] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3):301–413, 1999.

[8] M. Basseville and I. V Nikiforov. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

[9] Pierre C Bellec. Sharp oracle inequalities for least squares estimators in shape restricted regression. *arXiv preprint arXiv:1510.08029*, 2015.

[10] Richard Bellman. The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515, 1954.

[11] L. Birgé and P Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

[12] L. Birgé and P Massart. Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 138(1-2):33–73, 2006.

[13] Lucien Birgé and Pascal Massart. Gaussian model selection. *Journal of the European Mathematical Society*, 3(3):203–268, 2001.

[14] Kevin Bleakley and Jean-Philippe Vert. The group fused lasso for multiple change-point detection. *arXiv preprint arXiv:1106.4199*, 2011.

[15] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[16] V. Brault, C. Lévy-Leduc, A. Mathieu, and A. Jullien. Change-point estimation in the multivariate model taking into account the dependence: Application to the vegetative development of oilseed rape. *Journal of Agricultural, Biological and Environmental Statistics*, 2018.

[17] V. Brault, S. Ouadah, L. Sansonnet, and C Lévy-Leduc. Nonparametric multiple change-point estimation for analyzing large hi-c data matrices. *Journal of Multivariate Analysis*, 165:143 – 165, 2018.

[18] S. Chakar, E. Lebarbier, .C Lévy-Leduc, and S Robin. A robust approach for estimating change-points in the mean of an ar(1) process. *Bernoulli*, 23(2):1408–1447, 2017.

[19] S Charkar. *Segmentation de processus avec un bruit autoregressif*. PhD thesis, Université Paris Sud Orsay, 2015.

[20] Alice Cleynen, Michel Koskas, Emilie Lebarbier, Guillem Rigaill, and Stéphane Robin. Segmentor3isback: an r package for the fast and exact segmentation of seq-data. *Algorithms for Molecular Biology*, 9(1):1, 2014.

[21] C Denis. Classification in postural style based on stochastic process modeling. *The International Journal of Biostatistics*, 10(2):251—260, 2014.

[22] X D'Haultfoeuille and P Givord. La régression quantile en pratique. *Economie et Statistiques*, 471, 2014.

[23] D. Garreau and S. Arlot. Consistent change-point detection with kernels. *arXiv preprint arXiv:1612.04740*, 2016.

[24] S. Gey and E. Lebarbier. Using cart to detect multiple change points in the mean for large sample. 2008.

[25] S Greenland and M.P Longnecker. Methods for trend estimation from summarized dose-response data with applications to meta-analysis. *Amer. J. Epidemiology*, 11:1301–1309, 1992.

[26] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Processing*, 85(8):1501 – 1510, 2005.

[27] Marc Lavielle and Eric Moulines. Least-squares estimation of an unknown number of shifts in a time series. *Journal of time series analysis*, 21(1):33–59, 2000.

[28] E Lebarbier and T Mary-Huard. Une introduction au critère bic: fondements théoriques et interprétation. *Journal de la Société française de statistique*, 147(1):39–57, 2006.

[29] Émilie Lebarbier. Detecting multiple change-points in the mean of gaussian process by model selection. *Signal processing*, 85(4):717–736, 2005.

[30] A.E. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

[31] Aaditya Ramdas and Ryan J Tibshirani. Fast and flexible admm algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858, 2016.

[32] C. Rojas and B Wahlberg. On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*, 2014.

[33] A. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. Hindle, B. Huey, and K. Kimura. Assembly of microarrays for genome-wide measurement of dna copy number. *Nature genetics*, 29(3):263, 2001.

[34] R. S. Tsay. *Analysis of Financial Time Series*, volume 543. Wiley, 2005.

[35] B. Wahlberg, C. Rojas, and M. Annergren. On l 1 mean and variance filtering. In *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 1913–1916. IEEE, 2011.

[36] Y Zhu. An augmented ADMM algorithm with application to the generalized lasso problem. *Journal of Computational and Graphical Statistics*, 26(1):195–204, 2017.