Robust semi-parametric multiple change-point detection

Jean-Marc Bardet, Charlotte Dion

Robust semi-parametric multiple change-points detection

Jean-Marc Bardet[a], Charlotte Dion[b,*]

[a] *SAMM, Université Paris 1, 90 rue de Tolbiac, 75013 Paris*
[b] *LPSM, Sorbonne Université, 4 Place Jussieu, 75005 Paris UMR CNRS 98001*

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 is an algorithm based on total variation type penalty. These strategies are numerically studied on Monte-Carlo experiments.

Keywords: Change-points detection, Least-Absolute Value criterion
2010 MSC: 62F86, 62F35, 62J05

1. Introduction

In the sequel we consider a particular case of off-line 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:

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

where $M^* \in \{1, 2, \ldots, n\}$ is the number of regimes and therefore $M^* - 1$ is the number of changes, $0 = t_0^* < t_1^* < \ldots < t_{M^*}^* < t_{M^*}^* = n$ and $(\theta_1^*, \ldots, \theta_{M^*}^*) \in \mathbb{R}^{M^*}$ are respectively the abrupt change instants and the means (satisfying also $\theta_i^* \neq \theta_{i+1}^*$, $i = 1, \ldots, M^* - 1$ for insuring thant changes occur) of the process. We assume that $(\varepsilon_k)_{k\in\mathbb{N}}$ 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 neighbourhood of zero. In addition we assume

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

If the median of the noise is non-zero, then the model is translated but the following stays. The model (1) can equivalently be defined with a functional formula:

$$y_t = s^*(t) + \varepsilon_t, \quad \text{for } t \in \{1, \ldots, n\} \quad \text{with} \quad s^* = \sum_{k=1}^{M^*} \theta_k^* \mathbb{1}_{I^*(k)} \tag{3}$$

where the intervals are $I^*(k) = \{t_{k-1}^* + 1, \ldots, t_k^*\}$ for $k = 1, \ldots, M^*$ and s^* is the mean value function that is a piece-wise constant function.

The goal of the paper is the estimation of the number of changes $(M^* - 1)$ and 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 $(M^*, (\theta_k^*)_{1\leq k\leq M^*}, (t_k^*)_{1\leq k\leq M^*-1})$.

*Corresponding author
Email addresses: jean-marc.bardet@univ-paris1.fr (Jean-Marc Bardet), charlotte.dion@upmc.fr (Charlotte Dion)

1.1. State of the art

The detection of change-points in time series has been widely studied (see for example the general book [1]) and is very useful in many fields. One can cite finance ([2]), medical applications ([3, 4]), climatology ([5]) or agriculture ([6]). It is a real challenge in genetic, and several change-point detection methods have been designed to deal with special kinds of genomic data, as [7, 8, 9] and [10].

In this framework, estimators based on the Least-Squares (LS) contrast have 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 [11], an estimation method based on LS criterion together with a testing method to find the good dimension are proposed. [12] derives 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. Once a collection of estimators is available, the model (or dimension) selection is crucial.

Then [13] proposes a penalized method to estimate the number of change-points which is unknown and their locations. A nonparametric strategy based on model selection tools to estimate M^* is developed in [14]. This method is based on Gaussian model selection theory ([15, 16]). [17] focuses on a cross-validation selection method and [18] is placed in the context of heteroscedastic data. Other nonparametric approaches exist, for example [19] proposes kernel estimators, also see [10] in a different framework. [20] investigates the fused LASSO procedure and showed some difficulties. Recently, [21] has built confidence sets in the special case of shape restricted regression.

But it is well known that 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. It 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, [23] 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 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 [24]). Another paper [25] has followed to the particular case of one shift. Note also that [26] chose a method based on LAV and total-variation for ℓ^1-trend filtering.

Our goal is to propose new penalized LAV criteria to estimate a piece-wise constant signal.

1.2. Main contribution

In the present work to estimate the signal s^* we consider global approaches by penalized contrast minimization. In particular, the change-points, which are of interest here, are detected simultaneously.

Naturally, a first approach is based on model selection theory. Inspired by [14] one can ideally want to produce an upper bound for the ℓ_1-risk. The selection criterion has the form $\Gamma(\mathbf{m}) = C(\mathbf{m}, y) + \kappa \,\mathrm{pen}(\mathbf{m})$ where $\mathbf{m}$ represents a model that is a partition of n points with $|\mathbf{m}| = \mathrm{Card}(\mathbf{m})$ regimes. The penalty function should depend only on n and the dimension $|\mathbf{m}|$ of the model $\mathbf{m}$. The parameter κ is a trade-off parameter. In practice, the minimization is first realized on the contrast and for each possible dimension or number of regimes M leading to a collection of estimators $((\widehat{\theta}_k)_{1\leq k\leq M}, (\widehat{t}_k)_{1\leq k\leq M-1})$, one by possible dimension. The larger the number of changes the smaller the contrast, therefore the penalty term provides a bias-variance trade-off that determines the optimal number of regimes $\widehat{M}$. Guided by the model selection literature this procedure leads to a non-asymptotic strategy. The choice of the penalty function is based on [27] and calibrated from a numerical study similarly to [14]. Finally, in order to obtain a totally data-driven procedure, we use the heuristic slope approach introduced in [17]. This provides an estimator $\widehat{\kappa}$ of the parameter κ and the estimator $\widehat{M}^{\mathrm{New}}$ of the true number of regimes M^*. This method is theoretically different from the classical LAV penalization methods (Bai's penalty or BIC). Indeed, in the presentation and in the theoretical part, it is presented as model selection, meaning not just a dimension selection. In other word, the collection of model is much larger and models are not nested. This is why the induced

penalty is much complicated. Nevertheless, in practice, this procedure rely also on the dynamic algorithm to select one model by dimension first according to the criterion, just as Bai, BIC procedure or [14]'s.

Then we also propose a second method based on a Total-Variation approach (TV). It consists on a LAV deviation and an additional term which penalizes the first order difference of θ, inducing that for a given segmentation the optimal parameters are not the minimizer of the LAV deviation. This time, the penalty term does not only depend on the number of regimes. Since its introduction in the field of image processing, the total variation proved to be a valuable regularizer (see *e.g.* [28] for ℓ^1 trend filtering, we drive the reader attention on the fact that in these articles the LS loss is chosen). This approach does not provide a minimization of the ℓ^1-risk. But it provides interesting different estimators $\widehat{s}^{TV}$ and $\widehat{M}^{\mathrm{TV}}$ of the signal s^* and the number of regimes M^*.

We realized Monte-Carlo experiments to compare both these new approaches, and to challenge them with previous estimator from literature. The results are convincing for the new penalized LAV deviation estimator with respect to other LAV, LS, Huber loss criteria (which combines both approaches). 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.

1.3. Organisation of the paper

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 $((\theta_k^*)_{1\leq k\leq M^*}, (t_k^*)_{1\leq k\leq M^*-1})$, which are the value on each segment and the change-points from the observed trajectory $(y_1, \ldots, y_n)$ defined in (1), when the number of regime is fixed.

2.1. Notations

Here are the notations used in the sequel. The set of all the partitions of $\{1, \ldots, n\}$ is denoted $\mathcal{M}_n$. Then, for $\mathbf{m} \in \mathcal{M}_n$, its length is $|m| = \mathrm{Card}(\mathbf{m})$, where $\mathbf{m} = \{t_1, \ldots, t_{|m|-1}\}$ with $I_{\mathbf{m}}(k) = \{t_{k-1}+1, \ldots, t_k\}$ for $k = 1, \ldots, |m|$, with $t_0 = 0$ and $t_{|m|} = n$. The true model induced by (3) is denoted $\mathbf{m}^*$ with $M^* = |\mathbf{m}^*|$ and $(I_{\mathbf{m}^*}(k))_{1\leq k\leq M^*}$ are the true segments.

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

$$\mathcal{A}_M \quad := \{\mathbf{t} = (t_1, \ldots, t_{M-1}), t_0 = 0 < t_1 < \ldots < t_M = n\} \tag{4}$$

(we classically use bold letters for vectors). For M a fixed integer number in $\{0, 1, \ldots, n\}$, the subspace of piece-wise constant functions with M shifts is denoted $\mathcal{S}_M$, and for $M_{\max} \leq n$ a fixed integer,the subspace of piece-wise constant functions with less than $M_{\max}$ shifts by

$$\mathcal{S}^{M_{\max}} := \bigcup_{0\leq M\leq M_{\max}} \mathcal{S}_M. \tag{5}$$

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

$$\mathcal{S}_{\mathbf{m}} := \left\{ \sum_{k=1}^{|\mathbf{m}|} u_k \mathbf{I}_{I_{\mathbf{m}}(k)}, \quad (u_k)_{1\leq k\leq |\mathbf{m}|} \in \mathbb{R}^{|\mathbf{m}|} \right\}. \tag{6}$$

2.2. Least Absolute Deviation criterion

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

$$\widehat{\gamma}(\mathbf{u}) = \frac{1}{n} \sum_{t=1}^{n} |y_t - u_t| =: \|y - u\|_{1,n}, \quad \text{for} \quad \mathbf{u} = (u_t)_{1 \le t \le n} \in \mathbb{R}^n. \tag{7}$$

When the number of changes $M \in \mathbb{N}$ is specified, the LAV estimator of s^* is given by

$$\widehat{s}_M = \underset{s \in \mathcal{S}_M}{\operatorname{argmin}}\, \widehat{\gamma}(s) = \underset{\boldsymbol{\theta} \in \mathbb{R}^{M+1}}{\operatorname{argmin}}\, \underset{\boldsymbol{t} \in \mathbb{R}^M}{\operatorname{argmin}}\, \frac{1}{n} \sum_{k=1}^{M} \sum_{t=t_{k-1}+1}^{t_k} |y_t - \theta_k|. \tag{8}$$

and

$$\widehat{s}_M = \sum_{k=1}^{M} \widehat{\theta}_k^M \mathbb{1}_{\{t = \widehat{t}_{k-1}^M + 1, \dots, \widehat{t}_k^M\}}$$

where $\widehat{\theta}_k^M = \text{median}\left\{ y_{\widehat{t}_{k-1}^M + 1}, \dots, y_{\widehat{t}_k^M} \right\}$. In [23] was established the following asymptotic result.

Proposition 2.1 ([23])**.** *For model* (1) *with assumptions* (2)*, if* $t_k^* - t_{k-1}^* \ge n^{3/4}$*, if there exists* $c > 0$ *such as* $\left|\theta_k^* - \theta_{k-1}^*\right| \ge c$*, then* $\widehat{t}_k^{M^*} \xrightarrow[n \to \infty]{\mathcal{P}} t_k^*$ *for any* $k \in \{1, \dots, M^* - 1\}$*.*

This consistency result motivates the study of LAV-contrast estimators. Nevertheless, the huge size of the models with M regimes makes the solution non computable. To solve it, we classically use the dynamic programming algorithm.

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 $M \in \{0, 1, \dots, M_{\max}\}$ (see [29]). It is based on the computation of the optimal cost $\widehat{C}_M(s,t)$ in M segments included in $\{u, u+1, \dots, v\}$ for $u, v \in \{1, \dots, n\}$, given by:

$$\widehat{C}_M(u,v) := \min_{t_0 = u < t_1 < t_2 < \cdots < t_{M-1} < t_M = v} \quad \min_{(\theta_k)_{1 \le k \le M} \in \mathbb{R}^M} \quad \frac{1}{n} \sum_{k=1}^{M} \sum_{j=t_{k-1}+1}^{t_k} \left|y_j - \theta_k\right|. \tag{9}$$

Note that this problem is computable here because the cost matrix $\min_{\theta} \sum_{j=s}^{t} \left|y_j - \theta\right| = \sum_{j=s}^{t} \left|y_j - \text{median}(y_{s:t})\right|$, $0 \le s < t \le n$. First compute $\widehat{C}_1[u,v]_{u,v \in \{1,\dots,n\}}$. Then it is a two parts algorithm: the first part computes recursively the cost of the optimal segmentation with M changes for ℓ data for $1 \le M \le M_{\max}$ and $1 \le \ell \le n$; the second part is called *backtracking* and is used to find the optimal segmentation for each dimension (see the details in [30]). More formally we have:

Input: $\widehat{C}_1$
Initialization $\widehat{C}_2, \dots \widehat{C}_{M_{\max}} = 0$;
for $M = 1, \dots, M_{\max}$ **do**
  for $v \in \{M, \dots, n\}$ **do**

$$\widehat{C}_{M+1}[1,v] = \min_{u \in \{M+1, M+2, \dots, v\}} \left\{ \widehat{C}_M[1,u] + \widehat{C}_1[u+1,v] \right\}$$

  end
end
Output: $(\widehat{C}_M[1,v])_{M \in \{1,\dots,M_{\max}\},\ v \in \{M,\dots,n\}}$

Algorithm 1: Dynamic Programming

And then for each cost we follow the path witch gives this minimal cost and obtain one optimal segmentation by dimension.
Finally, for each $M = 1, \dots, M_{\max}$, we obtain $\widehat{C}_M(1,n)$ and the change-points $t_1 < t_2 < \cdots < t_{M-1}$ that minimize $\widehat{C}_M(1,n)$. The time-consuming cost is $O(Mn^2)$ instead of $O(\binom{n-1}{M})$ without the dynamic programming (see *e.g.* [31]). This algorithm gives finally one estimator by dimension, optimal for $\widehat{\gamma}_n$.

In Section 3, we describe penalized LAV criteria. First we remind the reader some existing ones and secondly we propose a new criterion and provide some theoretical justifications.

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 $\widehat{C}_{M+1}(1,n) \leq \widehat{C}_M(1,n)$, implying that a minimization of $M \to C_M(1,n)$ leads to the choice $M_{\max}$ with probability 1 and therefore other procedures are required for estimating M^* (meaning choosing M).

Assuming that $M^* \leq M_{\max}$, a usual method for estimating M^* is to penalize the cost $\widehat{C}_M(1,n)$. This can be classically done using the following general criterion

$$\widehat{M} = \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \Big\{ f(\widehat{C}_M(1,n)) + \kappa_n \operatorname{pen}(M) \Big\} = \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \Big\{ f(\widehat{\gamma}_n(\widehat{s}_M)) + \kappa_n \operatorname{pen}(M) \Big\}, \tag{10}$$

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

For example [23] proposes to select $\widehat{M}^{\mathrm{BAI}}$ defined by

$$\begin{aligned}
\widehat{M}^{\mathrm{BAI}} &= \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \Big\{ \log\big(\widehat{\gamma}(\widehat{s}_M)\big) + \frac{\sqrt{n}}{n} M \Big\} \\
&= \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \Big\{ \underset{\boldsymbol{t}\in\mathcal{A}_M}{\operatorname{argmin}} \Big(\log\Big(\frac{1}{n} \sum_{k=1}^{M} \sum_{t=t_{k-1}+1}^{t_k} \big| y_t - \operatorname{median}\big\{ y_{t_{k-1}+1}, \dots, y_{t_k} \big\} \big| \Big) + \frac{\sqrt{n}}{n} M \Big) \Big\}.
\end{aligned}$$

The author choose to use $\kappa_n = \frac{\sqrt{n}}{n}$ for insuring the consistency of $\widehat{M}^{\mathrm{BAI}}$ to M^* in our framework. 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:

$$\begin{aligned}
\widehat{M}^{\mathrm{BIC}} &= \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \Big\{ \log\big(\widehat{\gamma}(\widehat{s}_M)\big) + \frac{\log n}{n} M \Big\} \\
&= \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \Big\{ \underset{\boldsymbol{t}\in\mathcal{A}_M}{\operatorname{argmin}} \Big(\log\Big(\frac{1}{n} \sum_{k=1}^{M} \sum_{t=t_{k-1}+1}^{t_k} \big| y_t - \operatorname{median}\big\{ y_{t_{k-1}+1}, \dots, y_{t_k} \big\} \big| \Big) + \frac{\log n}{n} M \Big) \Big\}
\end{aligned}$$

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 [32, 33]). Moreover, since for $n \geq 1$ we have $\log n < \sqrt{n}$, the penalty term in $\widehat{M}^{\mathrm{BAI}}$ is always larger than the one in $\widehat{M}^{\mathrm{BIC}}$: we deduce that:

$$\widehat{M}^{\mathrm{BAI}} \leq \widehat{M}^{\mathrm{BIC}}. \tag{11}$$

This is the only possible comparison that can be made between the different criteria proposed in this article.

3.2. A data-driven oracle penalization

The point of view adopted in this paragraph is slightly different. The segmentation (or model) $\mathbf{m} \in \mathcal{M}_n$ is now fixed. Then we define the LAV-contrast estimator $\widehat{s}_{\mathbf{m}}$ of s^* defined in (3), is

$$\widehat{s}_{\mathbf{m}} = \underset{s \in \mathcal{S}_{\mathbf{m}}}{\operatorname{argmin}}\, \widehat{\gamma}(s) = \sum_{k=1}^{|\mathbf{m}|} \widehat{\theta}_k^{\mathbf{m}} \mathbb{1}_{I_{\mathbf{m}}(k)}, \tag{12}$$

where $\widehat{\theta}_k^{\mathbf{m}}$ is an empirical median defined by $\widehat{\theta}_k^{\mathbf{m}} := \text{median}\{y_{t_{k-1}+1}, \ldots, y_{t_k}\}$. This time the minimisation is done only on the $\boldsymbol{\theta}$ parameter as soon as the segmentation (thus the change-points) is fixed.

In order to define data-driven penalized least absolute values estimator we first follow some non asymptotic results on model selection developed in [34, 27]. Hence the following general estimator can be considered

$$\widehat{\mathbf{m}} := \underset{\mathbf{m} \in \mathcal{M}_n}{\operatorname{argmin}} \Big\{ \underset{s \in \mathcal{S}_{\mathbf{m}}}{\operatorname{argmin}} \left\{ \widehat{\gamma}(s) + \text{pen}(|\mathbf{m}|) \right\} \Big\} \tag{13}$$

where $\text{pen}(|\mathbf{m}|)$ only depends on $|\mathbf{m}|$ and n. The differences with the previous methods are that the minimization is done (theoretically) on all the models and the penalization function $\text{pen}(|\mathbf{m}|)$ is not necessary a linear function of $|\mathbf{m}|$ (since the model are no more nested).

First, let us recall the general result of [27] (Theorem 8 and 11) in the fixed design setting (see also [34] for details on LAV).

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

$$\text{pen}(|\mathbf{m}|) \geq K\,(L_{\mathbf{m}} + \mathcal{L}_{\mathbf{m}}) \frac{|\mathbf{m}|}{n}, \quad \textit{where} \quad \mathcal{L}_{\mathbf{m}} = \log\left[c\Big(1 + c'\Big(\frac{|\mathbf{m}|}{n}\Big)^{1/2}\Big)\right] + 1$$

with $c, c' > 0$. Then there exist C, $C' \in (0, \infty)$ depending on σ and $\kappa > 0$ such as $\widehat{\mathbf{m}}$ defined in (13) *satisfies*

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

This result is very strong for the LAV-contrast estimator, and the control of its quadratic risk. Furthermore, it is not a Gaussian framework result.

Then, in the present context, it is classical to choose variable weights, depending only on the dimension of the model $L_{\mathbf{m}} = L_{|\mathbf{m}|}$. Using the same computation done in [14] we obtain

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

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 $\mathbf{m} \in \mathcal{M}_n$*

$$\text{pen}(|\mathbf{m}|) := \sigma^2\,\frac{|\mathbf{m}|}{n}\left(c_1\,\log\left(\frac{n}{|\mathbf{m}|}\right) + c_2 \right) \tag{14}$$

there exist two positive constants $C(c_1, c_2)$, $C'(c_1, c_2)$ such that $\widehat{\mathbf{m}}$ defined in (13) *satisfies*

$$\mathbb{E}[\|s^* - \widehat{s}_{\widehat{\mathbf{m}}}\|_{1,n}] \leq \kappa \inf_{\mathbf{m} \in \mathcal{M}_n} \left\{ d(s^*, \mathcal{S}_{\mathbf{m}})^2 + C(c_1, c_2)\,\text{pen}(|\mathbf{m}|) \right\}^{1/2} + C'(c_1, c_2)\sqrt{\frac{\Sigma}{n}}. \tag{15}$$

This result is obtained from the usual inequality $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}|X_i|\right] \leq \mathbb{E}[\|X\|_2^2]^{1/2}$ and with the relationship $d^2(s^*, \mathcal{S}_{\mathbf{m}}) = \dfrac{1}{n}\displaystyle\sum_{t=1}^{n}|s^*(t) - s^*_{\mathbf{m}}(t)|^2 = \|s^* - s^*_{\mathbf{m}}\|^2_{2,n}$ where $s^*_{\mathbf{m}}$ 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 (14) because of the complexity of the collection of models, meaning the huge dimension. 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 [14]. 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*

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

which satisfies

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

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

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

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 (14) of the penalization. To be data-driven procedure, we propose a procedure to estimate this quantity from the data. We chose to estimate this constant using slope heuristic method introduced in ([16, 17]). It consists on computing the graph $\left(\frac{M}{n}\left(\log\left(\frac{n}{M}\right) + 2\right), \gamma(\widehat{s}_M)\right)$ for $1 \leq M \leq M_{\max}$. On such graph one can see an abrupt change of regime for M going from 1 to M^*, and a linear decrease for $M > M^*$. Using a classical off-line change detection for linear models, the slope κ of the linear part of the graph can be estimated by $\widehat{\kappa}$. The main idea of the slope heuristic procedure is to consider the new estimator of the number of regimes M^* by

$$\begin{aligned}\widehat{M}^{\text{New}} &= \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \left\{\widehat{\gamma}(\widehat{s}_M) - 2\widehat{\kappa}\,\frac{M}{n}\left(\log\left(\frac{n}{M}\right) + 2\right)\right\} \qquad (16)\\ &= \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}}\left\{\underset{\boldsymbol{t}\in\mathcal{A}_M}{\operatorname{argmin}}\left(\frac{1}{n}\sum_{k=1}^{M}\sum_{t=t_{k-1}+1}^{t_k}\left|y_t - \text{median}\left\{y_{t_{k-1}+1}, \ldots, y_{t_k}\right\}\right| - 2\widehat{\kappa}\,\frac{M}{n}\big(\log\left(\frac{n}{M}\right) + 2\big)\right)\right\}\end{aligned}$$

Hence we obtain a new data-driven estimator $\widehat{\mathbf{m}}$ of the true segmentation and thus an estimator $\widehat{M}^{\text{New}}$ of the number of abrupt changes for the LAV-contrast estimator.

In Section 4, we develop a criterion based on Total-Variation penalty form. This penalty is adapted in the present context and leads to another way to estimate the change-points and the regime parameters together with the number of regimes.

4. A Total Variation criterion

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

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

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 $(\theta_k)_k$. 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 estimated parameters θ_k are different to the one obtained from the minimization of the LAV deviation only, according to the total variation penalty term. This criterion differs from the classical ℓ^1-trend filtering. First we use divergence, we introduce the knowledge of the number of regimes in the criterion, as a result, the minimization problem is different an we propose an algorithm to approximate the solution. This procedure produces in one algorithm an estimation for the change-points, the parameters of the regimes and the number of it.

For each λ, which is the tuning parameter of the TV penalization, one would like to minimize $\xi_\lambda^M(\boldsymbol{\theta},\mathbf{t})$ in $\boldsymbol{\theta}\in\mathbb{R}^M$ and $\mathbf{t}\in\mathcal{A}_M$. But, the cost matrix 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 $\widehat{\mathbf{m}} = (\widehat{t}_k)_k$ for each dimension M in the collection. Secondly, the following minimization problem can be solved:

$$\big(\widehat{M}^\lambda, (\theta_k^\lambda)_{1\leq k\leq \widehat{M}^\lambda}\big) = \underset{1\leq M\leq M_{\max}}{\operatorname{argmin}} \ \underset{(\theta_k)_{1\leq k\leq M}\in\mathbb{R}^M}{\operatorname{argmin}} \ \xi_\lambda^M\big(\boldsymbol{\theta},(\widehat{t}_1,\ldots,\widehat{t}_{M-1})\big) \tag{17}$$

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 [35] (see *e.g.*[36, 37]). The ADMM algorithm rewrites the minimization problem over $\boldsymbol{\theta}$ as an equality constraint optimization problem where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_M)$ is split in two parts $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$. It is based on the formulation:

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

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:

Input: A, λ, ρ, M
Initialization: $\boldsymbol{\theta}$, $\boldsymbol{\alpha} = \mathbf{0}$, $\boldsymbol{\gamma} = \mathbf{0}$, ;
for $\ell = 1, \ldots, \ell_{\text{conv}}$ **do**

$$\begin{aligned}
\boldsymbol{\theta}^{(\ell+1)} &= \underset{\boldsymbol{\theta}\in\mathbb{R}^M}{\operatorname{argmin}} \left\{ f(\boldsymbol{\theta}) + \frac{\rho}{2}\|A\boldsymbol{\theta} - \boldsymbol{\gamma}^{(\ell)} + \rho^{-1}\boldsymbol{\alpha}^{(\ell)}\|^2 \right\} \\
\boldsymbol{\gamma}^{(\ell+1)} &= \underset{\boldsymbol{\gamma}\in\mathbb{R}^{M-1}}{\operatorname{argmin}} \left\{ \lambda g(\boldsymbol{\gamma}) + \frac{\rho}{2}\|A\boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\gamma} + \rho^{-1}\boldsymbol{\alpha}^{(\ell)}\|^2 \right\} \\
\boldsymbol{\alpha}^{(\ell+1)} &= \boldsymbol{\alpha}^{(\ell)} + \rho(A\boldsymbol{\theta}^{(\ell+1)} - \boldsymbol{\gamma}^{(\ell+1)})
\end{aligned}$$

end
Output: $\widehat{\boldsymbol{\theta}}^{\lambda} \in \mathbb{R}^M$

Algorithm 2: ADMM algorithm

The parameter ℓ_{conv} is the number of iterations used until convergence. The next step is the minimization over M given in (17) and we obtain the optimal dimension (in the sense of this TV criterion) denoted $\widehat{M}^{\text{TV}}$.

In the previous, 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 [3] in the regression case and the consistency of the change-points estimator is established when the number of regressors tends to infinity with fixed n (see also [20]). In this classical ℓ^1-trend filtering context, the consistency holds only if the sign of the parameters θ_k are all different.

As it has been said before, the approach developed here is slightly different because we impose the length of the vector $\boldsymbol{\theta}$, 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. The consistency is beyond the scope of this work and should be the subject of future works.

In Section 5, we illustrate the several presented strategies, together with classical Least-Squares one and a Huber-loss criterion. Finally, we provide two examples of results for our estimators on real data sets.

5. Numerical illustrations

In the section, we first provide details on the Monte-Carlo experiments allowing to compare the different criteria as well as the numerical implementation of the different methods. Then two real dataset are studied using the new criteria.

5.1. Presentation of the Monte-Carlo experiments

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

- Gaussian noise, denoted $\mathcal{N}$, centred with variance σ^2.
- 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

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

  where ϕ_{μ,σ^2} is the density of a Gaussian random variable with mean μ and variance σ^2, $\mu = \frac{qp}{\sqrt{q^2p+\sigma^2}}$ and $\gamma^2 = \frac{\sigma^4}{q^2p+\sigma^2}$. The distribution of this noise contains 3 modes and can mimic the presence of outliers. In the sequel we use $p = 1/10$ and $q = 10$.

We considered several scenarios for the distribution of changes:

1. the example given in [23] with $M^* = 4$ regimes with parameters $\boldsymbol{\theta}^* = (1, 3, 1, -1)$, with $t_i^* = [in/4]$ for $i = 1, 2$ and 3;
2. the case $M^* = 7$ with $\theta^* = (1, 3, 1, -1, 1, -3, -1)$ with $t_i^* = [in/7]$ for $i = 1, \ldots, 6$;
3. finally the case of randomized values of M^*, $(\theta_i^*)_i$ and $(t_i^*)_i$ that allows for a greater universality of numerical results, has been studied. More precisely, the parameters are simulated according to the following scheme:
   - the number of change-points $M^* - 1$ is simulated from a binomial distribution with parameters $(6, 0.5)$,
   - 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.

Figure 1 illustrates the good behavior of estimator $\widehat{s}_{\widehat{M}^{\mathrm{New}}}$ estimating s^*(in dotted black) with $n = 1200$, a Student noise, $M^* = 5$ and parameters $\boldsymbol{\theta}^* = (0, 1, 3, 2, 1)$ (with equal width for each regime). The observed signal y (orange crosses) is very corrupted by the noise and nevertheless the estimator is very close to the true underlying signal.

The procedures are evaluated in two terms:

- with the empirical score in % , *i.e.* the frequencies of estimation of the true values of 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 $\widehat{M}^{\mathrm{Bai}}$, $\widehat{M}^{\mathrm{BIC}}$ and $\widehat{M}^{\mathrm{New}}$, detailed in Section 3. Indeed, this paragraph is a continuation of the numerical study of [23]. 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 $\widehat{M}^{\mathrm{New}}$, $\widehat{M}^{\mathrm{Bai}}$ and $\widehat{M}^{\mathrm{BIC}}$. These experiments exhibit the consistency of these three LAV criteria when n increases. Moreover, the larger σ^2 (or 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 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 $\widehat{M}^{\mathrm{Bai}}$ provides excellent results. This is not surprising since it has been defined in [23] in an

asymptotic framework, On the contrary, in "difficult" conditions, meaning that M^* and the variance σ^2 are large, and differences $|\theta^*_{i+1} - \theta^*_i|$ and $t^*_{i+1} - t^*_i$ are small, then $\widehat{M}^{\text{New}}$ and $\widehat{M}^{\text{BIC}}$ provide much better results than $\widehat{M}^{\text{Bai}}$.

Generally speaking, estimator $\widehat{M}^{\text{New}}$ 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 $\widehat{M}^{\text{New}}$often provides the minimal risk even when $\widehat{M}^{\text{Bai}}$ obtains the best empirical score (see typically the case $M^* = 4$).

5.3. Comparison with total variation, least-squares and Huber criteria

In the sequel the new data-driven estimator $\widehat{M}^{\text{New}}$ is compared to three other criteria: TV, LS, Huber's.

5.3.1. TV criterion

We implement the estimator $\widehat{M}^{\text{TV}}$ 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. It is a challenge during the computation phase, to choose the best λ parameter. We choose to select it with a BIC criteria. Nevertheless, Figure 2 illustrates the behaviour of the selected dimension with respect to λ, when $n = 300$, $M^* = 3$ and $\sigma = 1$. We notice that the number of detected change-points decreases quickly when λ increases and converges to the true value 3. As remarked in [28], choose λ of order n^c with $c = 0.7$ (between $1/2$ and 1) seems also justified.

Note that the asymptotic properties of the estimator $\widehat{M}^{\text{TV}}$ have still not be studied. Nevertheless, Table 8 exhibits its convergence when n increases. Note that we do not exhibit the ℓ^1-risk scores of $\widehat{s}_{\widehat{M}^{\text{TV}}}$ since the TV criterion is only devoted to select the number of change-points and not to minimize the risk.

For this estimator it is interesting to study the empirical ℓ_1 risk of estimator $\widehat{\boldsymbol{\theta}}^{\text{TV}}$ of $\boldsymbol{\theta}^*$. The results are presented in Table 7 for a Student noise distribution in the fixed case $\boldsymbol{\theta}^* = (1, 3, 1, -1)$. The risks tends to 0 with n.

5.3.2. LS criterion

The presentation in analogue with the one on Section 3.2. For a given segmentation $\mathbf{m} \in \mathcal{M}_n$, the LS estimator is

$$\widehat{s}^{\text{LS}}_{\mathbf{m}} = \underset{u \in \mathcal{S}_m}{\text{argmin}}\, \widehat{\gamma}^{\text{LS}}(u) = \sum_{k=1}^{|\mathbf{m}|} \overline{y}_k \mathbb{1}_{I_k}, \quad \overline{y}_k = \frac{1}{n_k} \sum_{t=t_{k-1}+1}^{t_k} y_t,\ n_k = \text{Card}\{t \in I_{\mathbf{m}}(k)\}. \tag{18}$$

Again, the model selection literature under the Gaussian noise assumption (this time), leads the author of [14] to propose the following estimator:

$$\widehat{\mathbf{m}}^{\text{LS}} = \underset{\mathbf{m} \in \mathcal{M}_n}{\text{argmin}}\, \{\widehat{\gamma}^{\text{LS}}(\widehat{s}^{\text{LS}}_{\mathbf{m}}) + \text{pen}^{\text{LS}}(|\mathbf{m}|)\}$$

with $\text{pen}^{\text{LS}}(M) = \frac{M}{n}\sigma^2 \left(2 \log\left(\frac{n}{M}\right) + 5\right)$. This is the same penalty function as pen of Section 3.2 with different constants. Nevertheless, the final estimator is different, because the criterion differs.

As previously, the problem comes down to find one estimator by dimension, optimized for the LS criterion. They are computed using the Dynamic Programming again. To compute the estimator called $\widehat{M}^{\text{LS}}$ we use again the slope heuristic following the same process as for $\widehat{M}^{\text{New}}$. Finally we used

$$\widehat{M}^{\text{LS}} = \underset{1 \le M \le M_{\max}}{\text{argmin}} \Big\{ \underset{\boldsymbol{t} \in \mathcal{A}_M}{\text{argmin}} \Big(\frac{1}{n} \sum_{k=1}^{M} \sum_{t=t_{k-1}+1}^{t_k} \Big(y_t - \frac{1}{t_k - t_{k-1}} \sum_{i=t_{k-1}+1}^{t_k} y_i \Big)^2 - 2\widehat{\kappa}\, \frac{M}{n} \big(2 \log\Big(\frac{n}{M}\Big) + 5 \big) \Big) \Big\}.$$

5.3.3. Huber criterion

The standard Huber loss, motivated by the non-Gaussian case, combines LS and LAV losses for more robustness in location parameter estimation. It can also be considered to provide another estimation of the number of changes. First define the classical Huber function (see [38]):

$$\psi(x) := x^2 \, \mathbb{1}_{|x|\leq k} + k\left(2|x| - k\right) \mathbb{1}_{|x|>k}, \tag{19}$$

where $k > 0$. As it was suggested in [38], we choose $k = 1.345$. For a given segmentation $\mathbf{m} \in \mathcal{M}_n$, the Huber estimator is

$$\widehat{s}^{\text{Hub}}_{\mathbf{m}} \quad = \quad \underset{u\in\mathcal{S}_{\mathbf{m}}}{\operatorname{argmin}} \sum_{t=1}^{n} \psi(y_t - u_t). \tag{20}$$

As for the LS estimator, we use a Huber criterion with a penalty constant $\widehat{\kappa}$ estimated through the slope heuristic method, *i.e.*

$$\widehat{\mathbf{m}}^{\text{Hub}} = \underset{\mathbf{m}\in\mathcal{M}_n}{\operatorname{argmin}} \left\{ \sum_{t=1}^{n} \psi(y_t - \widehat{s}^{\text{Hub}}_{\mathbf{m}}(t)) - 2\widehat{\kappa} \, \frac{|\mathbf{m}|}{n} \left(2\log\left(\frac{n}{|\mathbf{m}|}\right) + 5 \right) \right\}.$$

5.3.4. Results of comparisons

The empirical scores are given in Table 8 where we also consider the previous version of randomized values of M^*, $\boldsymbol{\theta}^*$ and $\mathbf{t}^*$. It appears that $\widehat{M}^{\text{New}}$ provides the most accurate estimations, except for Gaussian time series for which the classical LS criterion (and also Huber 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 $\widehat{M}^{\text{New}}$ 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 3 shows the two estimators $\widehat{s}_{\widehat{\mathbf{m}}}$ (red) $\widehat{s}^{\text{LS}}_{\widehat{\mathbf{m}}^{\text{LS}}}$ (blue) together with the true signal s^* in dotted black line when the noise is a Student noise and $\sigma = 2$. On this example, the estimator based on LS criterion detects two artificial change-points during the first regime. This is due to the large variance of the data. On the contrary, the estimator based on LAV is very close to the real signal. 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. As we could imagine from its definition, the Huber criterion provides an interesting trade-off between LAV and LS criteria, and is really convincing especially when $\sigma = 2$ and for Gaussian or Student distributions.

Finally, Table 8 shows that $\widehat{M}^{\text{TV}}$ provides an interesting alternative except for Gaussian processes. But it is quite always less efficient than $\widehat{M}^{\text{New}}$. However, this criterion is built on the same principle than the LASSO criterion and we can suspect that it could especially be useful when M^* is really large and not negligible with respect to the data length n. Also, when σ increases, the TV estimator is advantageous compared to the New one.

5.4. Application to genomic data

In this paragraph, we apply the new criteria on a real-life data set, which consists on normalized copy-number logratios of data array CGH study set of Corriel institute taken from the package `DNAcopy` of V. Seshan and A. Olshen, see also [39] (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 strategies on some part on the data. The results are presented in Figure 4: on the left the true data are plotted together with estimator $\widehat{s}_{\widehat{M}^{\text{New}}}$ in red, and on the right the same graph with $\widehat{s}^{\text{TV}}$ in orange (and the LS estimator $\widehat{s}_{\widehat{M}^{\text{LS}}}$ is barely equal to $\widehat{s}_{\widehat{M}^{\text{New}}}$ 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.

5.5. Application to financial data

We further consider a publicly available financial data set. It is composed of the FTSE index between January 2004 and April 2013 and more precisely to its monthly estimations of the volatility (calculated from the empirical standard deviation taken over one month of FTSE log-returns). We are looking for the different possible disruptions with the different financial crises that occurred during these years in mind. Results are shown on Figure 5: we represent the New LAV estimator $\widehat{s}_{\widehat{M}^{\text{New}}}$ (red) as well as the more classical $\widehat{s}^{\text{LS}}$ (blue). Once again $\widehat{s}_{\widehat{M}^{\text{New}}}$ provides a very convincing result. The adaptive LS estimator detects 9 breaks (while there are 44 estimated breaks for the BIC-LS criterion!). But as it can be seen on Figure 5, 3 of these breaks correspond to peaks of distribution while only one of them (2009) is also detected by New LAV estimator (and during a larger period).

Besides, most of the estimated change dates obtained by $\widehat{s}_{\widehat{M}^{\text{New}}}$ are really meaningful:

1. $\widehat{t}_1$ is July 13, 2007. This corresponds to the beginning of the sub-prime crisis.
2. $\widehat{t}_2$ is September 10, 2008. this corresponds to Lehman Brothers bankruptcy (September 15, 2008).
3. $\widehat{t}_3$ is November 27, 2008: he American, British and European Union central banks proposed financial recovery plans.
4. $\widehat{t}_5$ and $\widehat{t}_6$ are July 29 and August 22, 2011: this occurs during the European financial crisis of summer 2011.

We have also computed LAV-Bai and LAV-BIC estimators that detected respectively 13 and 42 breaks, while TV criterion detected 43 breaks. In this case, New-LAV estimator provides clearly the most interesting segmentation.

Conclusion

The main contribution of this article concerns the construction of new robust estimators of the number of changes in the framework of multiple mean processes and their comparisons with other estimators. Let us recall the definition of the different estimators we considered:

$$\widehat{M}^{\text{BAI}} = \underset{1\le M\le M_{\max}}{\text{argmin}} \Big\{ \underset{\boldsymbol{t}\in\mathcal{A}_M}{\text{argmin}} \Big(\log\Big(\frac{1}{n}\sum_{k=1}^{M}\sum_{t=t_{k-1}+1}^{t_k} |y_t - \text{median}\Big\{y_{t_{k-1}+1},\dots,y_{t_k}\Big\}|\Big) + \frac{\sqrt{n}}{n} M\Big)\Big\}$$

$$\widehat{M}^{\text{BIC}} = \underset{1\le M\le M_{\max}}{\text{argmin}} \Big\{ \underset{\boldsymbol{t}\in\mathcal{A}_M}{\text{argmin}} \Big(\log\Big(\frac{1}{n}\sum_{k=1}^{M}\sum_{t=t_{k-1}+1}^{t_k} |y_t - \text{median}\Big\{y_{t_{k-1}+1},\dots,y_{t_k}\Big\}|\Big) + \frac{\log n}{n} M\Big)\Big\}$$

$$\widehat{M}^{\text{New}} = \underset{1\le M\le M_{\max}}{\text{argmin}} \Big\{ \underset{\boldsymbol{t}\in\mathcal{A}_M}{\text{argmin}} \Big(\frac{1}{n}\sum_{k=1}^{M}\sum_{t=t_{k-1}+1}^{t_k} |y_t - \text{median}\Big\{y_{t_{k-1}+1},\dots,y_{t_k}\Big\}| - 2\widehat{\kappa}\,\frac{M}{n}\big(\log\Big(\frac{n}{M}\Big)+2\big)\Big)\Big\}$$

$$\widehat{M}^{\text{TV}} = \underset{1\le M\le M_{\max}}{\text{argmin}} \Big\{ \underset{\boldsymbol{\theta}\in\mathbb{R}^M,\boldsymbol{t}\in\mathcal{A}_M}{\text{argmin}} \Big(\sum_{t=1}^{n}|y_t - \sum_{k=1}^{M}\theta_k \mathbb{1}_{t_{k-1}+1\le t\le t_k}| + \widehat{\lambda}\sum_{k=2}^{M}|\theta_k-\theta_{k-1}|\Big)\Big\}$$

$$\widehat{M}^{\text{LS}} = \underset{1\le M\le M_{\max}}{\text{argmin}} \Big\{ \underset{\boldsymbol{t}\in\mathcal{A}_M}{\text{argmin}} \Big(\frac{1}{n}\sum_{k=1}^{M}\sum_{t=t_{k-1}+1}^{t_k} \Big(y_t - \frac{1}{t_k-t_{k-1}}\sum_{i=t_{k-1}+1}^{t_k} y_i\Big)^2 - 2\widehat{\kappa}\,\frac{M}{n}\big(2\log\Big(\frac{n}{M}\Big)+5\big)\Big)\Big\}$$

$$\widehat{M}^{\text{Hub}} = \underset{1\le M\le M_{\max}}{\text{argmin}} \Big\{ \underset{\boldsymbol{\theta}\in\mathbb{R}^M,\boldsymbol{t}\in\mathcal{A}_M}{\text{argmin}} \Big(\sum_{k=1}^{M}\sum_{t=t_{k-1}+1}^{t_k} \psi(y_t-\theta_k) - 2\widehat{\kappa}\,\frac{M}{n}\big(2\log\Big(\frac{n}{M}\Big)+5\big)\Big)\Big\},$$

where $\widehat{\kappa}$ is obtained from a slope heuristic procedure and the function ψ is given in (19). There is only one general relation (11) between those estimators, *i.e.* $\widehat{M}^{\text{BAI}} \le \widehat{M}^{\text{BIC}}$. Otherwise, and this is verified by numerical simulations, the estimator $\widehat{M}^{\text{New}}$ has the best overall performance, particularly in terms of robustness.

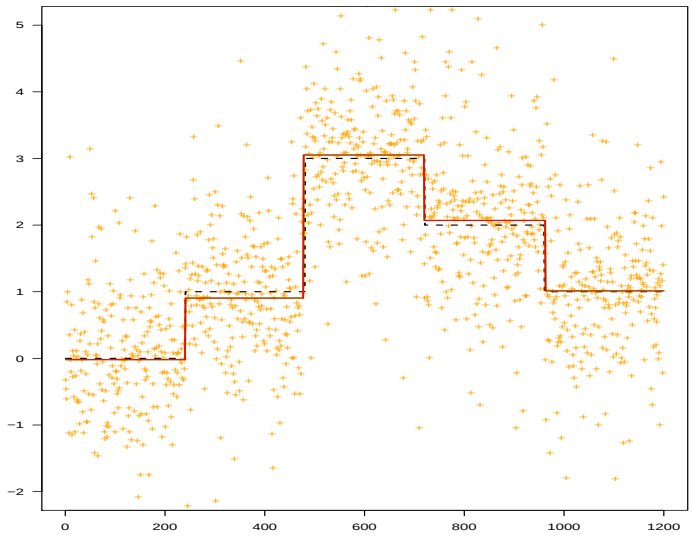

Figure 1: Example of estimator $\widehat{s}_{\widehat{M}^{\text{New}}}$ (red line) when $n = 1200$, $M^* = 5$, s^* the dotted black line and y orange points corrupted with Student noise.

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	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 $M^* = 4$.

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	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 $M^* = 4$.

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	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	0.0	**4.8**	10.9	0.0	**25.7**	**75.3**	0.0	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 $M^* = 7$.

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	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 $M^* = 7$.

	$n = 50$			$n = 200$			$n = 500$		
$\sigma = 1$	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**	42.0	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	55.5	**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**	30.0	**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 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 M^*.

		$n = 50$	$n = 200$	$n = 500$
$\sigma = 1$	New	0.73(0.28)	0.36(0.13)	0.22(0.086)
	TV	0.76(0.29)	0.37(0.14)	0.22(0.087)
$\sigma = 2$	New	2.42(2.08)	2.12(1.86)	0.45(0.161)
	TV	2.23(1.79)	2.08(0.66)	0.44(0.162)

Table 7: Empirical ℓ_1-risk of $\widehat{\boldsymbol{\theta}}^{\mathrm{TV}}$ with $\varepsilon \sim \mathcal{S}$.

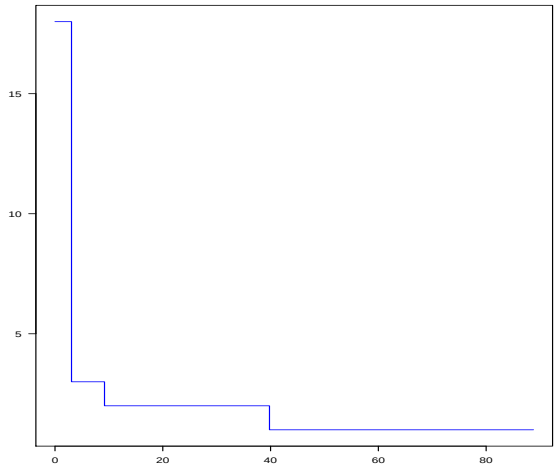

Figure 2: $\widehat{M}_{\lambda}^{\mathrm{TV}}$ as a function of λ when $M^* = 3$.

	$n = 50$				$n = 200$				$n = 500$			
$\sigma = 1$	New	TV	LS	Hub	New	TV	LS	Hub	New	TV	LS	Hub
$\mathcal{N}$	16.9	24.4	**39.0**	38.8	66.5	42.7	**72.0**	71.2	83.5	48.8	86.3	**87.1**
$\mathcal{L}$	**44.6**	42.3	35.9	42.9	**79.0**	56.0	60.6	73.9	**92.1**	76.1	75.8	90.5
$\mathcal{S}$	**49.7**	31.2	35.0	47.7	**82.9**	53.8	38.7	76.1	**94.0**	60.2	31.8	88.8
M$\mathcal{N}$	**55.9**	40.5	29.9	40.0	**73.9**	69.3	55.5	42.1	**91.0**	75.5	66.2	50.2
$\sigma = 2$	New	TV	LS	Hub	New	TV	LS	Hub	New	TV	LS	Hub
$\mathcal{N}$	17.3	14.1	**20.5**	19.7	29.5	20.0	**36.6**	34.6	53.2	20.5	**55.2**	48.7
$\mathcal{L}$	**23.0**	14.5	19.4	22.9	38.9	**39.4**	30.0	39.3	**62.8**	48.2	43.8	62.4
$\mathcal{S}$	25.5	23.5	21.1	**26.4**	45.7	39.6	23.7	**48.7**	67.1	45.8	23.5	**71.0**
M$\mathcal{N}$	25.9	14.8	18.5	**26.5**	**46.5**	28.9	31.1	40.5	**65.6**	38.0	47.3	63.8

Table 8: Empirical score with 10000 repetitions with randomized M^*: comparisons between several estimators of the number of changes: $\widehat{M}^{\text{New}}$, $\widehat{M}^{\text{TV}}$, $\widehat{M}^{\text{LS}}$ and $\widehat{M}^{\text{Hub}}$.

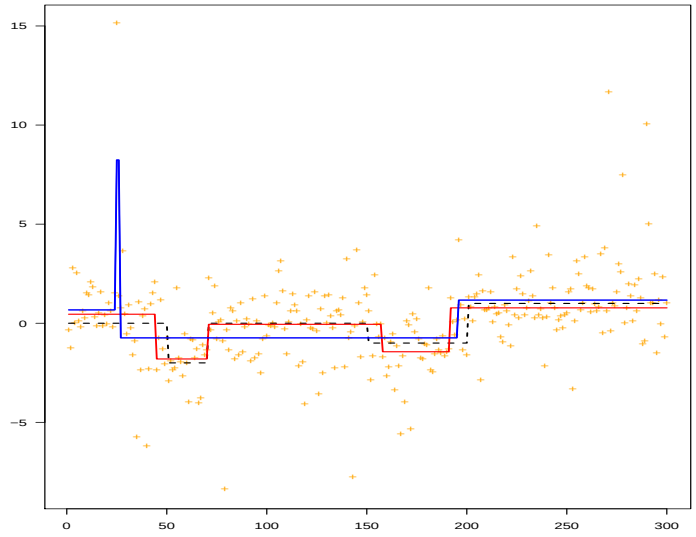

Figure 3: Comparison between LAV and LS criteria ($\widehat{s}_{\widehat{\mathbf{m}}}$ vs $\widehat{s}^{\text{LS}}_{\widehat{\mathbf{m}}^{\text{LS}}}$): red line for LAV new estimator, blue line for LS penalized estimator. True signal s^* in dotted black line, and y in orange corrupted with Student noise, $\sigma = 2$.

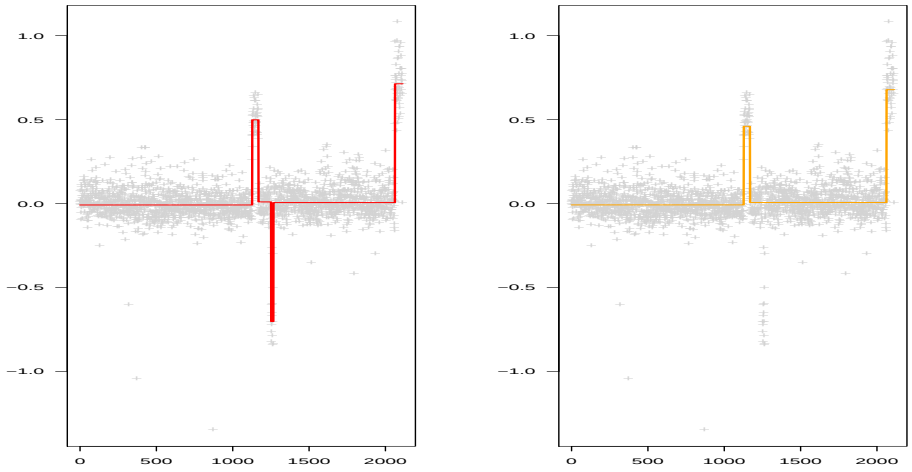

Figure 4: Grey: normalized copy-number logratio, left: red $\widehat{s}_{\widehat{M}^{\mathrm{New}}}$, right: orange $\widehat{s}^{\mathrm{TV}}$.

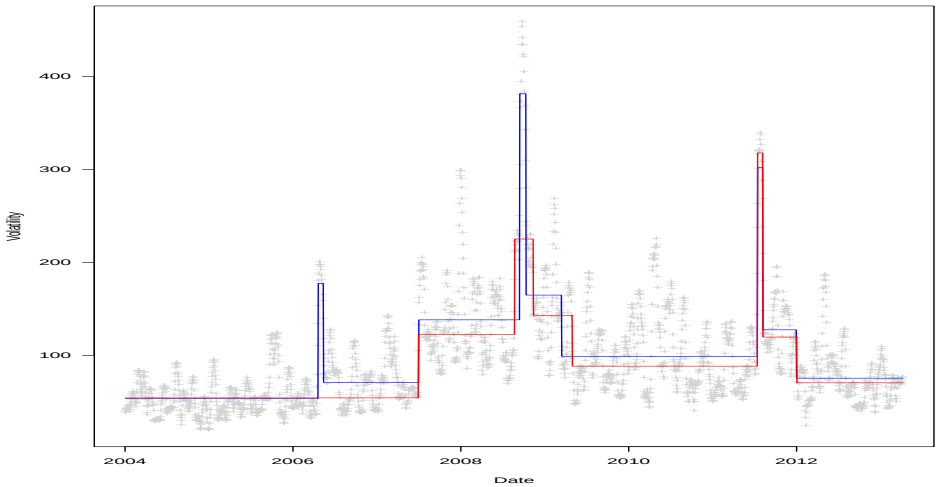

Figure 5: Black: estimated monthly volatility of FTSE index each day from January 2004 and April 2013. In red the new LAV estimator $\widehat{s}_{\widehat{M}^{\mathrm{New}}}$, and in blue, the LS estimator $\widehat{s}^{\mathrm{LS}}_{\widehat{\mathbf{m}}^{\mathrm{LS}}}$.

Acknowledgements

The authors are very grateful to the Foundation des Sciences Mathématiques de Paris (FSMP) for financial support. They also thank Pascal Massart (University Paris-Sud) for fruitful advise. Finally, many thanks to the Associate Editor and to the three referees for their careful reading of the paper and thoughtful critical comments.

References

[1] M. Basseville, I. V. Nikiforov, Detection of abrupt changes: theory and application, Vol. 104, Prentice Hall Englewood Cliffs, 1993.
[2] R. S. Tsay, Analysis of Financial Time Series, Vol. 543, Wiley, 2005.
[3] K. Bleakley, J.-P. Vert, The group fused lasso for multiple change-point detection, arXiv preprint arXiv:1106.4199.
[4] C. Denis, Classification in postural style based on stochastic process modeling, The International Journal of Biostatistics 10 (2) (2014) 251—260.
[5] R. Baillie, S. Chung, Modeling and forecasting from trend-stationary long memory models with applications to climatology, Int. J. Forecasting 19 (2002) 215–226.
[6] V. Brault, C. Lévy-Leduc, A. Mathieu, 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 23 (3) (2018) 374 – 389.
[7] S. Greenland, M. Longnecker, Methods for trend estimation from summarized dose-response data with applications to meta-analysis, Amer. J. Epidemiology 11 (1992) 1301–1309.
[8] A. Cleynen, M. Koskas, E. Lebarbier, G. Rigaill, S. Robin, Segmentor3isback: an r package for the fast and exact segmentation of seq-data, Algorithms for Molecular Biology 9 (1) (2014) 1.
[9] A. Cleynen, S. Dudoit, S. Robin, Comparing segmentation methods for genome annotation based on rna-seq data, Journal of Agricultural, Biological, and Environmental Statistics 19 (2013) 101–118.
[10] V. Brault, S. Ouadah, L. Sansonnet, C. Lévy-Leduc, Nonparametric multiple change-point estimation for analyzing large hi-c data matrices, Journal of Multivariate Analysis 165 (2018) 143 – 165.
[11] J. Bai, P. Perron, Estimating and testing linear models with multiple structural changes, Econometrica 66 (1) (1998) 47–78.
[12] M. Lavielle, E. Moulines, Least-squares estimation of an unknown number of shifts in a time series, Journal of time series analysis 21 (1) (2000) 33–59.
[13] M. Lavielle, Using penalized contrasts for the change-point problem, Signal Processing 85 (8) (2005) 1501 – 1510.
[14] É. Lebarbier, Detecting multiple change-points in the mean of gaussian process by model selection, Signal processing 85 (4) (2005) 717–736.
[15] L. Birgé, P. Massart, Gaussian model selection, Journal of the European Mathematical Society 3 (3) (2001) 203–268.
[16] L. Birgé, P. Massart, Minimal penalties for gaussian model selection., Probability Theory and Related Fields 138 (1-2) (2006) 33–73.
[17] S. Arlot, P. Massart, Data-driven calibration of penalties for least-squares regression, Journal of Machine Learning Research 10 (2009) 245–279.
[18] S. Arlot, A. Celisse, Segmentation of the mean of heteroscedastic data via cross-validation, Statistics and Computing 21 (4) (2011) 613–632.
[19] D. Garreau, S. Arlot, Consistent change-point detection with kernels, arXiv preprint arXiv:1612.04740.
[20] C. Rojas, B. Wahlberg, On change point detection using the fused lasso method, arXiv preprint arXiv:1401.5408.
[21] P. C. Bellec, Sharp oracle inequalities for least squares estimators in shape restricted regression, arXiv preprint arXiv:1510.08029.
[22] X. D'Haultfoeuille, P. Givord, La régression quantile en pratique, Economie et Statistiques 471.
[23] J. Bai, Estimation of multiple-regime regressions with least absolutes deviation, Journal of Statistical Planning and Inference 74 (1) (1998) 103–134.
[24] S. Chakar, E. Lebarbier, C. Lévy-Leduc, S. Robin, A robust approach for estimating change-points in the mean of an ar(1) process, Bernoulli 23 (2) (2017) 1408–1447.
[25] J. Bai, Least absolutes deviation estimation of a shift, Econometric Theory 11 (1995) 403–436.
[26] B. Wahlberg, C. Rojas, 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), IEEE, 2011, pp. 1913–1916.
[27] A. Barron, L. Birgé, P. Massart, Risk bounds for model selection via penalization, Probability theory and related fields 113 (3) (1999) 301–413.
[28] J. Ottersten, B. Wahlberg, C. R. Rojas, Accurate changing point detection for ℓ_1 mean filtering, IEEE Signal Processing Letters 23 (2) (2016) 297–301.
[29] R. Bellman, The theory of dynamic programming, Bulletin of the American Mathematical Society 60 (6) (1954) 503–515.
[30] S. Charkar, Segmentation de processus avec un bruit autoregressif, Ph.D. thesis, Université Paris Sud Orsay (2015).
[31] S. Gey, E. Lebarbier, Using cart to detect multiple change points in the mean for large sample, hal-00327146.
[32] A. Raftery, Bayesian model selection in social research, Sociological Methodology 25 (1995) 111–163.

[33] E. Lebarbier, T. Mary-Huard, Une introduction au critère bic: fondements théoriques et interprétation, Journal de la Société française de statistique 147 (1) (2006) 39–57.
[34] L. Birgé, P. Massart, Rates of convergence for minimum contrast estimators, Probability Theory and Related Fields 97 (1993) 113–150.
[35] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and Trends in Machine Learning 3 (1) (2011) 1–122.
[36] Y. Zhu, An augmented ADMM algorithm with application to the generalized lasso problem, Journal of Computational and Graphical Statistics 26 (1) (2017) 195–204.
[37] A. Ramdas, R. J. Tibshirani, Fast and flexible admm algorithms for trend filtering, Journal of Computational and Graphical Statistics 25 (3) (2016) 839–858.
[38] P. J. Huber, Robust estimation of a location parameter, The Annals of Mathematical statistics 35 (1) (1964) 73–101.
[39] A. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, G. Hamilton, A. Hindle, B. Huey, K. Kimura, Assembly of microarrays for genome-wide measurement of dna copy number, Nature genetics 29 (3) (2001) 263.