MULTIPLE CHANGE-POINT ESTIMATION WITH A TOTAL VARIATION PENALTY

Z. HARCHAOUI AND C. LÉVY-LEDUC

Abstract. We propose a new approach for dealing with the estimation of the location of change-points in one-dimensional piecewise constant signals observed in white noise. Our approach consists in reframing this task in a variable selection context. We use a penalized least-square criterion with a ℓ_1-type penalty for this purpose. We explain how to implement this method in practice by using the LARS/LASSO algorithm. We then prove that, in an appropriate asymptotic framework, this method provides consistent estimators of the change-points with an almost optimal rate. We finally provide an improved practical version of this method by combining it with a reduced version of the dynamic programming algorithm and we successfully compare it with classical methods.

1. Introduction

Retrospective Multiple Change-point Estimation consists in partitioning a nonstationary series of observations into several contiguous stationary segments of variable durations, see Brodsky and Darkhovsky (1993, 2000). It is particularly appropriate for analyzing *a posteriori* time series in which the quantity driving the behavior of the time series jumps from one level to another different level at random instants called change-points. Such a task, also known as temporal signal segmentation in signal processing, arises in many applications, ranging from EEG to speech processing and network intrusion detection (Basseville and Nikiforov, 1993; Ruanaidh and Fitzgerald, 1996).

As argued by both Carlstein et al. (1994) and Brodsky and Darkhovsky (2000), in most cases detecting changes of a time-evolving statistical quantity may be reduced to the detection of changes in the mean of a new sequence derived from the initial one. Thus, we

Date: April 27, 2010.

Key words and phrases. Change-point estimation, LARS, LASSO, ℓ_1-type penalty.

are interested in the estimation of the change-point locations $t_k^\star$ in the following model:

$$Y_t = \mu_k^\star + \varepsilon_t, \quad t_{k-1}^\star \leq t \leq t_k^\star - 1, \quad k = 1, \dots, K^\star + 1, \quad t = 1, \dots, n, \tag{1}$$

with the convention $t_0^\star = 1$ and $t_{K^\star+1}^\star = n + 1$ and where the $\{\varepsilon_t\}_{0 \leq t \leq n}$ are i.i.d zero-mean random variables, having a sub-Gaussian distribution.

This problem has recently received much attention on the theoretical side, both in a nonasymptotic and in an asymptotic setting by (Massart, 2004) and (Yao and Au, 1989; Lavielle and Moulines, 2000; Boysen et al., 2009) respectively. From a practical point of view, the standard approach for estimating the change-point locations is based on least-square fitting, performed *via* a dynamic programming algorithm (DP), coupled with an informational criterion such as the Schwarz criterion (Yao and Au, 1989) for choosing the unknown number of change-points. Indeed, for a given number of change-points K, the dynamic programming algorithm, proposed by Fisher (1958) and Bellman (1961), takes advantage of the intrinsic additive nature of the least-square objective to recursively compute the optimal change-points locations with a complexity of $O(Kn^2)$ in time. Then selecting the number of change-points is usually performed thanks to a Schwarz-like penalty $\lambda_n K$, where λ_n is often calibrated on data (Lavielle and Moulines, 2000; Lavielle, 2005), or a penalty $K(a + b\log(n/K))$ as in (Massart, 2004; Lebarbier, 2005), where a and b are data-driven as well. We should also mention that an abundant literature tackles both change-point estimation and model selection issues from a Bayesian point of view, see Ruanaidh and Fitzgerald (1996), Fearnhead (2006) and references therein; we shall not adopt such a point of view in this work.

While optimal from a maximum likelihood point of view in the case of Gaussian noise, the application of the standard least-square approach, called LS in the remainder, is seriously harmed by a quadratic time-complexity in the total duration of the series of observations in its exact implementation. Yet approximate dynamic programming procedures were devised in other contexts, such as for Dynamic Time Warping or the Viterbi algorithm (Kolesnikov and Fränti, 2003; Gales and Young, 2008). Moreover, as pointed in Hawkins (2001), a computationally efficient dynamic programming algorithm for change-point estimation may be devised when a prior assumption of order-structure between the segments is satisfied and therefore consists in restricting the change-point locations search to a pre-specified set. Yet, designing a computationally efficient dynamic programming algorithm for change-point estimation under general assumptions is still an open problem.

Therefore, an alternative formulation might be profitable from a computational point of view, while keeping comparable performance when compared to the least-square method. A natural way to lower the time-complexity of a ℓ_0-penalized least-square problem is to relax the ℓ_0-penalty to an ℓ_1-penalty. This strategy has proved to be appropriate in other statistical problems such as sparse PCA, sparse LDA, see d'Aspremont et al. (2008), and Moghaddam et al. (2006). Hence, it boils down to estimating the change-point locations by solving

$$\underset{u \in \mathbb{R}^n}{\text{Minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}(Y_i - u_i)^2 + \lambda_n \sum_{i=1}^{n-1} |u_{i+1} - u_i| \ , \tag{2}$$

and recovering the change-point locations from the jumps in the $\{\hat{u}_i\}_{i=1,\ldots,n}$ minimizing the criterion in Eq. (2). This alternative formulation yields a subquadratic time-complexity in the length of the sequence of observations, and still remains asymptotically consistent in terms of change-point estimation. Note that Tibshirani and Wang (2008) introduced the "fused lasso", which corresponds to a two-step procedure where the first step is a least-square change-point estimation with a total-variation penalty and the second is a thresholding one to discard small jumps from the zero-mean, a method specifically designed for spatial smoothing and hot spot detection in CGH data.

This article is organized as follows. In Section 2, we describe how Eq. (2) is related to the the well-known Least Absolute Shrinkage eStimatOr (LASSO) in least-square regression of Tibshirani (1996), usually used for efficient variable selection. We show that it turns out to be also useful for change-point estimation as well when used with a particular design matrix. We take advantage of this relationship to devise a subquadratic change-point estimation algorithm, called LS-TV for Least-Square with Total Variation penalty. In Section 3, we give theoretical results concerning the estimation of the underlying piecewise constant function and the estimation of the change-point locations. More precisely, we provide rates of convergence for the underlying piecewise constant function and for the change-point instants and we show that we can attain almost optimal rates of convergence in both cases. In Section 4, we run numerical experiments to assess the empirical behavior of LS-TV, and propose an enhanced version LS-TV* with better empirical performance.

2. Methodology

In this section, we describe the least-square change-point estimation with a total variation penalty LS-TV. In Section 2.1, we show how to recast the multiple change-point estimation problem into a particular variable selection problem. Then in Section 2.2, we describe a LAR-based implementation of LS-TV, and derive its time-complexity. The theoretical properties of LS-TV are given in Section 3.

2.1. **From change-point estimation to variable selection.**

The multiple change-point estimation problem may be relaxed into a LASSO-type problem using appropriate auxiliary variables.

Recall the multiple change-point model (Yao and Au, 1989):

$$Y_t = u_t^\star + \varepsilon_t \ , \quad t = 1, \dots, n, \tag{3}$$

where $u_t^\star = \mu_k^\star$ for $t_{k-1}^\star \leq t \leq t_k^\star - 1$, $k = 1, \dots, K^\star + 1$. We shall always assume in the remainder of this section that the true number of change-points $K^\star$ is known. The issue of dealing with an unknown number of change-points will be addressed later in Sections 3 and 4.

The least-square estimation method LS, which may also be viewed as the maximum-likelihood approach in the case of Gaussian white noise, solves the following problem:

$$\begin{cases} \text{Minimize}_{u \in \mathbb{R}^n} & \frac{1}{n} \sum_{i=1}^n (Y_i - u_i)^2 \\ \text{subject to} & \sum_{i=1}^{n-1} \mathbf{1}\{u_{i+1} - u_i\} = K^\star \ . \end{cases} \tag{4}$$

We propose here to relax the above ℓ_0 constraint into an ℓ_1 constraint on the magnitude of the jumps as follows:

$$\begin{cases} \text{Minimize}_{u \in \mathbb{R}^n} & \frac{1}{n} \sum_{i=1}^{n} (Y_i - u_i)^2 \\ \text{subject to} & \sum_{i=1}^{n-1} |u_{i+1} - u_i| \leq K^\star J^\star_{\max} \ , \end{cases} \tag{5}$$

where $J^\star_{\max} = \max_{1 \leq k \leq K^\star} |u^\star_{k+1} - u^\star_k|$. This alternative setting was previously elusively mentioned several times, in e.g. Mammen and Van De Geer (1997) and Boysen et al. (2009).

In order to further understand the behavior of the solution $(\hat{u}_1, \ldots, \hat{u}_n)$ of this criterion, let us denote by X_n the $n \times n$ lower triangular matrix with nonzero elements equal to one.

Then, by straightforward algebra, the problem in Eq. (5) may be rewritten as:

$$\begin{cases} \text{Minimize}_{\beta \in \mathbb{R}^n} & \frac{1}{n} (Y_i - (X_n \beta)_i)^2 \\ \text{subject to} & \sum_{i=1}^{n} |\beta_i| \leq K^\star J^\star_{\max} \ . \end{cases} \tag{6}$$

The underpinning insight is the sparsity-enforcing property of the ℓ_1-constraint, which is expected to give a sparse vector $\hat{\beta}^n$, whose non-zero components would match with change-points locations.

A major feature of Eq. (6) is that it exactly corresponds to the well-known Least Absolute Shrinkage eStimatOr (LASSO) in least-square regression of Tibshirani (1996), used for efficient variable selection. However, as far as we know, neither thorough practical implementation nor theoretical grounding has been given so far to support such an approach for change-point estimation. Actually, the corresponding minimization can be solved by using the LAR/LASSO algorithm described in Efron et al. (2004) and Hesterberg et al. (2008).

2.2. **Implementation with Least-Angle Regression.** In this Section, we detail the process of the Least-Angle Regression (LAR) algorithm of (Efron et al., 2004). For the sake of generality, we shall describe here this algorithm when we look for $K_{\max}$ change-points, $K_{\max}$ being a known upper bound on the true number of change-points. When implemented with care, we get a time-complexity in $O(n \log(n))$ of the LAR/LASSO algorithm in the particular case of our model. This substantial reduction of the computational complexity has to be contrasted with the complexity $O(K_{\max} n^2)$ of DP. We use in this section standard notation given for instance in Cormen et al. (2001).

The process is described in Table 1, the different notations involved being explained in the following in the description of each step of the algorithm. It essentially involves four steps, each of them being solved in sub-quadratic time-complexity with respect to the number of observations n. Suppose we have performed $k-1$ iterations in the main loop of the algorithm, then the current set of estimated change-points, that is the *active set* in the variable selection framework, is $\widehat{\mathcal{T}}_{n,k-1} = \{\hat{t}_1, \dots, \hat{t}_{k-1}\}$ and the current set of estimated segment levels is $\{\hat{u}_1(k-1), \dots, \hat{u}_n(k-1)\}$. We are now describing the computational requirements of the k-th iteration of the algorithm.

First, we look for the next change-point $\hat{t}_k$ to add to $\widehat{\mathcal{T}}_{n,k-1}$ yielding the largest discrepancy with the true signal. This requires, given $\{\hat{u}_1(k-1), \dots, \hat{u}_n(k-1)\}$, the computation of the n cumulative sums $\{\sum_{i=j}^{n} \hat{u}_i(k-1)\}_{j=1,\dots,n}$. These cumulative sums may actually be computed in $O(n)$ operations in time, using the simple recursion $\sum_{i=j}^{n} \hat{u}_i(k-1) = \sum_{i=j+1}^{n} \hat{u}_i(k-1) + \hat{u}_j(k-1)$. Besides, to be included in the current set of change-point estimates ("active set"), we need to locate the new change-point estimate with regard to the other change-point estimates, which is formally equivalent to sort the set of observations. Therefore, the "change-point addition"-step in Table 1 has a $O(n + n\log(n))$ time-complexity as long as k is smaller than $K_{\max}$.

Second, we have to compute the descent direction, which involves the multiplication of the inverse of a $(k \times k)$-matrix by a k-long vector. Indeed, X_k is a matrix which consists of the columns of X indexed by the elements of $\widehat{\mathcal{T}}_{n,k}$ and $\mathbf{1}_k$ denotes a vector of dimension k with each component equal to one. Given the current set of change-points $\widehat{\mathcal{T}}_{n,k}$, the inverse may be computed in $O(k^2)$ operations, since the entries of the inverse matrix of size $(k \times k)$ are available in close-form beforehand, see (35) in the Appendix. Then, the multiplication of the $(k \times k)$-inverse by $\mathbf{1}_k$ is computed in $O(k^2)$ operations. If $k < K_{\max}$, then the time-complexity of "descent direction computation"-step is upper-bounded by $O(K_{\max}^2)$.

```
LS-TV with LAR/LASSO

Initialization, $k = 0$.
    (a) Set $\widehat{\mathcal{T}}_{n,0} = \emptyset$.
    (b) Set $\hat{u}_i(0) = 0$, for all $i = 1, \ldots, n$.

While $k < K_{\mathsf{max}}$
    (a) Change-point addition:
        Find $\hat{t}_k$ such that
            $\hat{t}_k = \underset{t \in \{1,\ldots,n\} \setminus \widehat{\mathcal{T}}_{n,k-1}}{\mathrm{Argmax}} \left| \sum_{i=t}^{n} Y_i - \sum_{i=t}^{n} \hat{u}_i(k-1) \right| .$
    (b) Descent direction computation:
        Compute
            $\mathbf{w}_k = (X_k^T X_k)^{-1} \mathbf{1}_k .$
    (c) Descent step search:
        Search for $\hat{\gamma}$ such that
            $\hat{\gamma} = \underset{t \in \{1,\ldots,n\} \setminus \widehat{\mathcal{T}}_{n,k}}{\mathrm{Min}} \left( \frac{\sum_{i=t}^{n} Y_i - \sum_{i=t}^{n} \hat{u}_i(k)}{1 - \sum_{i=t}^{n} w_{k,i}}, \frac{\sum_{i=t}^{n} Y_i + \sum_{i=t}^{n} \hat{u}_i(k)}{1 + \sum_{i=t}^{n} w_{k,i}} \right) .$
    (d) Zero-crossing check:
        If
            $\hat{\gamma} > \tilde{\gamma} \overset{\mathrm{def}}{=} \min_{j} \quad (\alpha_j w_{k,j})^{-1} \left( \sum_{i=j}^{n} \hat{u}_i(k) \right) ,$
        then, decrease $\hat{\gamma}$ down to $\hat{\gamma} = \tilde{\gamma}$, and remove $\tilde{t}$ from $\widehat{\mathcal{T}}_{n,k}$, where
            $\tilde{t} \overset{\mathrm{def}}{=} \underset{j}{\mathrm{Argmin}} \quad (\alpha_j w_{k,j})^{-1} \left( \sum_{i=j}^{n} \hat{u}_i(k) \right) .$
```

Table 1. Description of the adaptation of LAR/LASSO algorithm for solving the LS-TV problem.

Third, we search for the descent step. For similar reasons as for the first step, the "descent step search"-step may be performed in linear-time $O(n)$ time-complexity. Indeed, again, this step involves the computation of n cumulative sums, which may be computed recursively.

Fourth, we check the zero-crossing of the coefficients to exactly track the regularization path of the LASSO. In this step, $\alpha_j = \text{sign}(\hat{u}_{j+1}(k) - \hat{u}_j(k))$. Again, all computations involved in this step hinge on cumulative sums as previously in the first step, and therefore may be performed in $O(n)$ time-complexity. Note that the maximum number of iterations N needed in practice to decrease $\hat{\gamma}$ to a small enough value to satisfy $\hat{\gamma} = \tilde{\gamma}$ is unknown in general, and no theoretically grounded upper-bound on N was provided in the literature so far. In practice, we set $N < K_{\max}$ in our implementation, and we never encountered any numerical issue which demanded a different (larger) setting of N. Hence, the "zero-crossing"-step has at most $O(K_{\max}\ n)$ time-complexity.

Thus, the implementation of LS-TV based upon the LAR/LASSO algorithm runs in $O(K_{\max}^3 + K_{\max} n \log n)$ in time.

3. Theoretical results

In this section, we give some theoretical results providing justification on the relevance of LS-TV for multiple change-point estimation. First, in Section 3.1, we prove that LS-TV is consistent in terms of estimation of the underlying signal. Second, in Section 3.2, we show that LS-TV is also consistent in terms of change-point estimation.

The main point of both Section 3.1 and Section 3.2 is the following. While the equivalence of LS-TV to a particular LASSO problem is fruitful from a computational point of view, it turns out to be less relevant for theoretical analysis. To get optimal results for LS-TV both in terms of means and change-points estimation, the original formulation (2) is more useful than the LASSO formulation.

3.1. **Estimation of the means.** We consider here the multiple changes in the mean problem as described in (1). Our purpose is to estimate the unknown means $\mu_1^\star, \dots, \mu_{K^\star+1}^\star$ together with the change-points from observations $Y_1, \dots, Y_n$.

Let us first work with the LASSO formulation to establish the consistency in terms of means estimation. The model (1) can be rewritten as:

$$Y^n = X_n \beta^n + \varepsilon^n \ , \tag{7}$$

where $Y^n = (Y_1, \dots, Y_n)'$ is the $n \times 1$ vector of observations, X_n the $n \times n$ lower triangular matrix with nonzero elements equal to one and $\varepsilon^n = (\varepsilon_1^n, \dots, \varepsilon_n^n)'$ is a zero mean random vector such that the ε_j^n's are i.i.d random variables with finite variance equal to σ^2. As for β^n, it is a $n \times 1$ vector having all its components equal to zero except those corresponding to the change-points instants.

Let us denote by $\mathcal{A}$ the set of non-zero components of β^n and by $\bar{\mathcal{A}}$ its complementary set defined as follows:

$$\mathcal{A} = \{k, \ \beta_k^n \neq 0\} \text{ and } \bar{\mathcal{A}} = \{1, \dots, n\} \backslash \mathcal{A} \ . \tag{8}$$

With the reformulation (7), the evaluation of the means estimation rate amounts to finding the rate of convergence of $\|X_n(\hat{\beta}^n(\lambda_n) - \beta^n)\|_n$ to zero, $\hat{\beta}^n(\lambda_n)$ satisfying:

$$\hat{\beta}^n(\lambda_n) = (\hat{\beta}_1(\lambda_n), \dots, \hat{\beta}_n(\lambda_n))' = \operatorname*{Arg\,min}_{\beta \in \mathbb{R}^n} \left\{ \|Y^n - X_n \beta\|_n^2 + \lambda_n \|\beta\|_1 \right\} \ , \tag{9}$$

where $\|u\|_n$ and $\|u\|_1$ are defined for a vector $u = (u_1, \dots, u_n) \in \mathbb{R}^n$ by $\|u\|_n = n^{-1} \sum_{j=1}^n u_i^2$ and $\|u\|_1 = \sum_{j=1}^n |u_j|$ respectively. Hence, within this framework, we are able to prove the following result regarding the consistency in means estimation of LS-TV.

Proposition 1. *Consider $Y_1, \dots, Y_n$ a set of observations following the model described in (7). Assume that the ε_j^n's are centered i.i.d Gaussian random variables with variance $\sigma^2 > 0$. Assume also that there exists β_{max} such that for all k in $\mathcal{A}$, $|\beta_k^n| \leq \beta_{max}$, the set $\mathcal{A}$ being defined in* (8). *Then, for all $n \geq 1$ and $C > 2\sqrt{2}$, we obtain that with a probability* *larger than* $1 - n^{1-C^2/8}$*, if* $\lambda_n = C\sigma\sqrt{\log n/n}$*,*

$$\|X_n(\hat{\beta}^n(\lambda_n) - \beta^n)\|_n \leq (2C\sigma\beta_{max}K^\star)^{1/2}\left(\frac{\log n}{n}\right)^{1/4} .$$

The proof, which follows similar lines as Bickel et al. (2009), is postponed to Section 7. Note that in Proposition 1, where no upper bound on the number of change-points is assumed to be known, we do not attain the known (parametric) optimal rate which is of order $1/\sqrt{n}$ derived by Yao and Au (1989) where an upper bound for the number of change-points is available. But, as we shall see in Proposition 2, the rate of Proposition 1 can be improved if the model and the criterion are rewritten in a different way and if an upper bound for the number of change-points is available.

Indeed, let us now work in the standard formulation of LS-TV instead of its LASSO counterpart, and write model (1) as:

$$Y_t = u_t^\star + \varepsilon_t \ , \quad t = 1, \dots, n \ , \tag{10}$$

where $u_t^\star = \mu_k^\star$ for $t_{k-1}^\star \leq t \leq t_k^\star - 1$, $k = 1, \dots, K^\star + 1$ and estimate the vector $(u_1^\star, \dots, u_n^\star)$ by using a criterion based on a total variation penalty as in Mammen and Van De Geer (1997):

$$\hat{u}(\lambda_n) = (\hat{u}_1(\lambda_n), \dots, \hat{u}_n(\lambda_n)) = \operatorname*{Arg\,min}_{u \in \mathbb{R}^n} \left\{ \|Y^n - u\|_n^2 + \lambda_n \sum_{i=1}^{n-1} |u_{i+1} - u_i| \right\} . \tag{11}$$

The following Proposition gives the rate of convergence of $\hat{u}(\lambda_n)$ when an upper bound for the number of change-points is known and equal to $K_{\max}$.

Proposition 2. *Consider* $Y_1, \dots, Y_n$ *a set of observations following the model described in* (10) *where the* ε_t*'s are zero-mean i.i.d Gaussian random variables with a variance* $\sigma^2 > 0$*. Assume also that* $\hat{u}$ *defined in* (11) *belongs to a set of dimension at most* $K_{max} - 1$*. Then, for all* $n \geq 1$*,* A *in* $(0, 1)$ *and* $B > 0$*, if* $\lambda_n = \sigma(A\sqrt{B}/2)(K_{max}\log n)^{1/2}n^{-3/2} - \sigma(2K_{max} + 1)^{1/2}n^{-3/2}$*,*

$$\mathbb{P}\left(\|\hat{u} - u^\star\|_n \geq \sigma(BK_{max}\log n/n)^{1/2}\right) \leq K_{max}\, n^{\{1-B(1-A)^2/8\}K_{max}} . \tag{12}$$

The proof of this Proposition is postponed to Section 7. The rate of convergence that we obtain for the estimation of the means is almost optimal up to a logarithmic factor since the optimal rate derived by Yao and Au (1989) is $O(n^{-1/2})$.

Let us now study the consistency in terms of change-point estimation, which is more of interest in this paper. Again, we shall see that the LASSO formulation is less relevant than the standard formulation for establishing the change-point estimation consistency.

3.2. **Estimation of the change-point locations.** In this section, we aim at estimating the change-point locations from the observations $(Y_1, \ldots, Y_n)$ satisfying model (7). The change-point estimates that we propose to study are obtained from the $\hat{\beta}_i(\lambda_n)$'s satisfying the criterion (9) as follows. Let us define the set of active variables by:

$$\hat{\mathcal{A}}(\lambda_n) = \left\{ i \in \{1, \ldots, n\},\ \hat{\beta}_i(\lambda_n) \neq 0 \right\} . \tag{13}$$

Then, we define the change-point estimates by $\hat{t}_i(\lambda_n)$ satisfying:

$$\hat{\mathcal{A}}(\lambda_n) = \left\{ \hat{t}_1(\lambda_n), \ldots, \hat{t}_{|\hat{\mathcal{A}}(\lambda_n)|}(\lambda_n) \right\}, \quad \text{where} \quad \hat{t}_1(\lambda_n) < \cdots < \hat{t}_{|\hat{\mathcal{A}}(\lambda_n)|}(\lambda_n) , \tag{14}$$

$|\hat{\mathcal{A}}(\lambda_n)|$ denoting the cardinal of the set $\hat{\mathcal{A}}(\lambda_n)$.

Discussion and related works. With such a reformulation of the change-point in the mean problem, the change-point estimates can be seen as Lasso-type estimates in a sparse framework. But, many classical assumptions under which the asymptotic properties of the Lasso estimates have been studied are not satisfied.

For instance, the *irrepresentable condition* as defined in Meinshausen and Yu (2009) (P. 5) which ensures *sign consistency* defined in Zhao and Yu (2006) is not satisfied in the change-point in the mean problem. More precisely, *sign consistency* ensures that $\mathbb{P}(\text{sign}(\hat{\beta}^n(\lambda_n) = \text{sign}(\beta^n))$ tends to one as n tends to infinity and the *irrepresentable condition* is a condition on the covariance matrix C^n defined by

$$C^n = n^{-1} X_n' X_n ,$$

which requires that the following inequality holds element-wise:

$$\left|C^n_{\bar{\mathcal{A}}\mathcal{A}}(C^n_{\mathcal{A}\mathcal{A}})^{-1}\mathrm{sign}(\beta^n_{\mathcal{A}})\right| < 1 \; , \tag{15}$$

where C^n_{IJ} is a sub-matrix of C^n obtained by keeping rows with index in the set I and columns with index in J. The vector $\beta^n_{\mathcal{A}}$ is defined by $\beta^n_{\mathcal{A}} = (\beta^n_k)_{k\in\mathcal{A}}$ and sign denotes a function mapping positive entries of a vector to 1, negative entries to -1 and null entries to zero. In our case, there exists at least one component i_0 such that

$$(|C^n_{\bar{\mathcal{A}}\mathcal{A}}(C^n_{\mathcal{A}\mathcal{A}})^{-1}\mathrm{sign}(\beta^n_{\mathcal{A}})|)_{i_0} = 1$$

This can be proved by computing explicitly the matrices $C^n_{\bar{\mathcal{A}}\mathcal{A}}$ and $(C^n_{\mathcal{A}\mathcal{A}})^{-1}$, see the Appendix for further details. In terms of change-point estimation, it means, as already known, see for example Yao and Au (1989) or Lavielle and Moulines (2000), that we cannot have a perfect estimation of the change-points.

Note that, Meinshausen and Yu (2009) brought to light some less restrictive conditions than the irrepresentable condition on the matrix C^n under which the Lasso estimates can be proved to be consistent in the ℓ_2-norm sense. The main assumption consists in assuming a m_n-*incoherent design* which means:

$$\liminf_{n\to\infty} \phi_{\min}(m_n) > 0 \; , \text{ where } \phi_{\min}(m) = \min_{\beta:\|\beta\|_{\ell_0}\leq m} \frac{\beta' C^n \beta}{\beta'\beta} \; , \tag{16}$$

with $m_n = s_n \log n$, s_n being the sparsity of the model that is the number of non-zero coefficients. In other words, a design is called m_n-*incoherent* if the minimal eigenvalue of a collection of m_n variables is bounded from below by a positive constant. In our setting, if the distance between two consecutive indices of non null coefficients is equal to one, then for all $n \geq 1$

$$\phi_{\min}(m_n) \leq 1/n \; ,$$

this making the condition (16) not satisfied in our case. A justification of this statement is given in the Appendix.

These particularities of the change-point in the mean model prevent us from using the techniques recently devised to study the asymptotic properties of the Lasso estimates in a general regression framework. However, the consistency of the $\hat{t}_i(\lambda_n)$ defined in (14) is established in Proposition 5.

Let us now detail the assumptions under which our theoretical results are established. Define

$$I^\star_{\min} = \min_{1\leq k\leq K^\star} |t^\star_{k+1} - t^\star_k|, \quad J^\star_{\min} = \min_{1\leq k\leq K^\star} |\mu^\star_{k+1} - \mu^\star_k|, \quad J^\star_{\max} = \max_{1\leq k\leq K^\star} |\mu^\star_{k+1} - \mu^\star_k|,$$

which are respectively the minimum interval length, the minimum and maximum jump sizes. From now on, we shall work under the following assumptions

(**A1**) The $\varepsilon_1, \ldots, \varepsilon_n$ are iid zero-mean random variables with $\mathrm{Var}[\varepsilon_1] = \sigma^2$ satisfying: there exists a positive constant β such that for all $\nu \in \mathbb{R}$, $\mathbb{E}\{\exp(\nu\varepsilon_1)\} \leq \exp(\beta\nu^2)$.

(**A2**) The sequence $\{\delta_n\}_{n\geq 1}$ is a non increasing and positive sequence tending to zero as n tends to infinity and satisfying $n\delta_n(J^\star_{\min})^2/\log(n) \to \infty$.

(**A3**) The change-points $t^\star_1, \ldots, t^\star_{K^\star}$ satisfy $I^\star_{\min} \geq n\delta_n$, for all $n \geq 1$.

(**A4**) The sequence of regularization parameters $\{\lambda_n\}_{n\geq 1}$ is such that $(n\delta_n J^\star_{\min})^{-1} n\lambda_n \to 0$, as n tends to infinity.

We first state a Lemma arising from the Karush-Kuhn-Tucker conditions of the optimization problem stated in (9) which will be useful in the proof of the consistency of our procedure.

Lemma 3. *Consider $Y_1, \ldots, Y_n$ a set of observations following the model described in (10). Then, $(\hat{t}_1(\lambda_n), \ldots, \hat{t}_n(\lambda_n))$ defined by (14) and $(\hat{u}_1(\lambda_n), \ldots, \hat{u}_n(\lambda_n))$ defined by: $\hat{u}_i(\lambda_n) = (X_n\hat{\beta}^n(\lambda_n))_i$, where X_n is a $n \times n$ lower triangular matrix with nonzero elements equal to one and the $(\hat{\beta}_i(\lambda_n))_{1\leq i\leq n}$ are obtained in (9), satisfy:*

$$\sum_{i=\hat{t}_\ell(\lambda_n)}^{n} Y_i - \sum_{i=\hat{t}_\ell(\lambda_n)}^{n} \hat{u}_i = \frac{n\lambda_n}{2}\hat{\alpha}_\ell \;, \quad \textit{for all } \ell = 1, \ldots, |\hat{\mathcal{A}}(\lambda_n)| \;, \tag{17}$$

and

$$\left|\sum_{i=j}^{n} Y_i - \sum_{i=j}^{n} \hat{u}_i\right| \leq \frac{n\lambda_n}{2} \;, \quad \textit{for all } j = 1, \ldots, n \;, \tag{18}$$

using the convention: $\hat{\alpha}_\ell = +1$, *if* $\hat{u}_{\hat{t}_\ell(\lambda_n)} > \hat{u}_{\hat{t}_\ell(\lambda_n)-1}$ *and* $\hat{\alpha}_\ell = -1$, *otherwise. The vector* $(\hat{u}_1(\lambda_n), \dots, \hat{u}_n(\lambda_n))$ *has the following additional property:*

$$\hat{u}_t(\lambda_n) = \hat{\mu}_k \ , \quad for \quad \hat{t}_{k-1}(\lambda_n) \le t \le \hat{t}_k(\lambda_n) - 1, \ k = 1, \dots, |\hat{\mathcal{A}}(\lambda_n)| + 1 \ , \tag{19}$$

where $|\mathcal{A}(\lambda_n)|$ *denotes the cardinal of the set* $\mathcal{A}(\lambda_n)$ *defined in (14).*

The proof of Lemma 3 is given in Section 7. Then, we state a Lemma which allows us to control the supremum of the average of the noise and which will also be useful for proving the consistency of our estimation criterion.

Lemma 4. *Let* $(\varepsilon_i)_{1\le i\le n}$ *be a sequence of random variables satisfying Assumption (A1). If* $\{v_n\}_{n\ge1}$ *and* $\{x_n\}_{n\ge1}$ *are two positive sequences such that* $v_n x_n^2/\log(n) \to \infty$, *then*

$$\mathbb{P}\left(\max_{\substack{1\le r_n<s_n\le n \\ |r_n-s_n|\ge v_n}} |(s_n-r_n)^{-1} \sum_{i=r_n}^{s_n-1} \varepsilon_i| \ge x_n \right) \to 0 \ , \ \ as\ n\to\infty \ .$$

The proof of Lemma 4 is postponed to Section 7.

Proposition 5. *Let* $Y_1, \dots, Y_n$ *be a set of observations satisfying model (1) then under Assumptions (A1)–(A4), the change-points estimators* $\{\hat{t}_1(\lambda_n), \dots, \hat{t}_{|\hat{\mathcal{A}}(\lambda_n)|}(\lambda_n)\}_{n\ge1}$ *defined by* (14), *satisfy, if* $|\hat{\mathcal{A}}(\lambda_n)| = K^\star$ *with probability tending to one:*

$$\mathbb{P}\left(\max_{1\le k\le K^\star} |\hat{t}_k - t_k^\star| \le n\delta_n \right) \to 1, \ \ as\ n\to\infty \ . \tag{20}$$

The proof of Proposition 5 is given in Section 7.

Under the assumptions of Proposition 5, the $\hat{\tau}_k$'s defined for all $k \in \{1, \dots, K^\star\}$ by $\hat{t}_k = [n\hat{\tau}_k]$ are consistent estimators of the $\tau_k^\star$'s defined by $t_k^\star = [n\tau_k^\star]$, for all $k \in \{1, \dots, K^\star\}$ with the rate δ_n.

Note that with $\delta_n = (\log n)^2/n$, $J_{\min}^\star \ge (\log n)^{1/4}$, $\lambda_n = \sqrt{\log n/n}$ or $\lambda_n = \sqrt{\log n}/n^{3/2}$, the assumptions (A2)–(A4) are satisfied leading thus to a rate of order $(\log n)^2/n$ for the estimation of the $\hat{\tau}_k$. With this choice of parameters, we obtain an almost optimal rate for the estimation of the $\tau_k^\star$ (up to a logarithmic factor) since the optimal rate is of order 1/n according to Yao and Au (1989).

This result has also to be compared with the work by Lavielle and Moulines (2000). They also obtained a rate in 1/n using a least-square approach in the case where the (ε_t) are not necessarily independent random variables but with more restrictive assumptions than ours on $I^\star_{\min}$ and $J^\star_{\min}$. Indeed, it is assumed in Theorem 7 of Lavielle and Moulines (2000), that $\min_{1\leq k\leq K^\star}|\tau^\star_{k+1}-\tau^\star_k| = \Delta^\star_\tau$ where $\Delta^\star_\tau$ is a positive constant and that $J^\star_{\min}$ is a positive constant.

In Proposition 5, the number of estimated change-points is assumed to be equal to the true number of change-points. Since this information is not in general available, we propose to evaluate the distance between the set $\widehat{\mathcal{T}}_{n,K}=\{\hat{t}_1,\dots,\hat{t}_K\}$ of K estimated change-points and the set of the true change-points $\mathcal{T}^\star_n=\{t^\star_1,\dots,t^\star_{K^\star}\}$ by using as in Boysen et al. (2009) the two quantities $\mathcal{E}(\widehat{\mathcal{T}}_{n,K}\|\mathcal{T}^\star_n)$ and $\mathcal{E}(\mathcal{T}^\star_n\|\widehat{\mathcal{T}}_{n,K})$, where $\mathcal{E}(\cdot\|\cdot)$ is defined for two sets A and B by

$$\mathcal{E}(A\|B)=\sup_{b\in B}\inf_{a\in A}|a-b| \ . \tag{21}$$

Note that we recover the Hausdorff distance between the sets A and B with

$$\Delta(A,B)=\sup\{\mathcal{E}(A\|B);\mathcal{E}(B\|A)\} \ .$$

Obviously, when $K=K^\star$, Proposition 5 implies that, under the same assumptions, $\mathcal{E}(\widehat{\mathcal{T}}_{n,K^\star}\|\mathcal{T}^\star_n)\leq n\delta_n$ and $\mathcal{E}(\mathcal{T}^\star_n\|\widehat{\mathcal{T}}_{n,K^\star})\leq n\delta_n$ with probability tending to one as n tends to infinity. In the case where $K>K^\star$, we prove in Proposition 6 that $\mathcal{E}(\widehat{\mathcal{T}}_{n,K}\|\mathcal{T}^\star_n)\leq n\delta_n$ with probability tending to one as n tends to infinity.

Proposition 6. *Let $Y_1,\dots,Y_n$ be a set of observations satisfying model (1) then under Assumptions (A1), (A3), (A4) and if $n\delta_n {J^\star_{min}}^2/\log(n^3/\lambda_n^2)\to\infty$, the change-points estimators $\{\hat{t}_1(\lambda_n),\dots,\hat{t}_{|\hat{\mathcal{A}}(\lambda_n)|}(\lambda_n)\}_{n\geq1}$ defined by* (14)*, satisfy, if $|\hat{\mathcal{A}}(\lambda_n)|\geq K^\star$ with probability tending to one:*

$$\mathbb{P}\left(\mathcal{E}(\widehat{\mathcal{T}}_{n,|\hat{\mathcal{A}}(\lambda_n)|}\|\mathcal{T}^\star_n)\leq n\delta_n\right)\to 1, \ \textit{as } n\to\infty \ . \tag{22}$$

Note that with $\delta_n=(\log n)^2/n$, $J^\star_{\min}\geq(\log n)^{1/4}$, $\lambda_n=\sqrt{\log n/n}$ or $\lambda_n=\sqrt{\log n}/n^{3/2}$, the assumptions (A3), (A4) and $n\delta_n {J^\star_{\min}}^2/\log(n^3/\lambda_n^2)\to\infty$ of Proposition 6 are fulfilled.

Now, we shall investigate the empirical behavior of LS-TV on simulated data. In the remainder, we focus on the so-called Blocks dataset introduced in Donoho and Johnstone (1995) which contains $K^\star = 11$ change-points. One may indeed consider the Blocks dataset as a typically difficult dataset for multiple change-point estimation, since both segment levels and segment lengths are highly heterogeneous.

4. Experimental results

4.1. **Specified number of change-points.** The Blocks dataset introduced in the paper (Donoho and Johnstone, 1995, page 1201, Table 1) was subsampled down to 1000 points as depicted in Figure 1, and corrupted with Gaussian white noise at three different levels: **low-noise** when $\sigma = 0.05$, **medium-noise** when $\sigma = 0.10$, and **high-noise** when $\sigma = 0.50$.

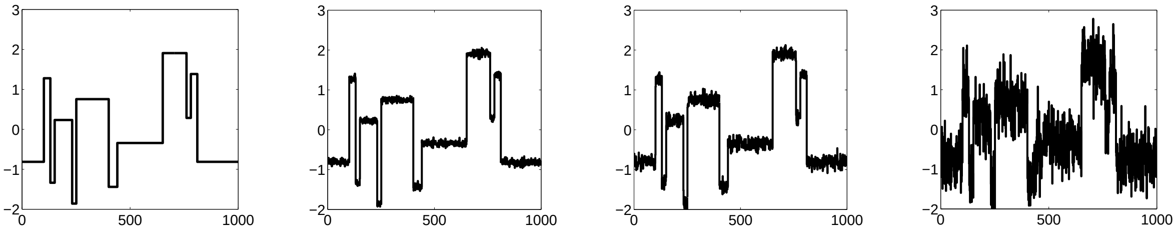

Figure 1. The Blocks dataset, subsampled to 1000 observations, and rescaled to mean zero and variance one, displayed without noise (on the far left), and with respectively low-noise, medium-noise and high-noise (from left to right).

To assess our large-sample consistency result which stated that $n^{-1}\mathcal{E}(\widehat{\mathcal{T}}_{n,K^\star}\|\mathcal{T}^\star) = o_P(1)$, as n tends to infinity, we ran Monte-Carlo simulations to investigate the empirical performance of LS-TV in terms of $n^{-1}\mathcal{E}(\widehat{\mathcal{T}}_{n,K^\star}\|\mathcal{T}^\star) = n^{-1}\max_{k=1,\dots,K^\star}|\hat{t}_k - t_k^\star|$ in the three different noise settings. For each noise setting, we generated 100 replications of the Blocks dataset corrupted with Gaussian white noise. The results are displayed in Table 2. In all noise conditions, the large-sample change-point estimation consistency of LS-TV is confirmed. In high-noise conditions, even for medium-scale samples, that is for $n = 1000$, the change-point detection ability of LS-TV remains satisfactory. For large-scale samples, that is for $n = 5000$, the performance continue improving both on average and standard deviation.

	Low-noise	Medium-noise	High-noise
n=1000	0.0200 ± 0.0068	0.0200 ± 0.0098	0.0230 ± 0.0185
n=5000	0.0127 ± 0.0059	0.0127 ± 0.0082	0.0127 ± 0.0169

TABLE 2. Performance in terms of $\mathcal{E}(\widehat{\mathcal{T}}_{n,K^\star}\|\mathcal{T}^\star)$ of LS-TV on the Blocks dataset corrupted with low-noise ($\sigma = 0.05$), medium-noise ($\sigma = 0.10$), and high-noise ($\sigma = 0.50$). The values after ± correspond to the standard deviations.

Since, in general, the number of change-points is unknown, we shall investigate in the next section the impact of misspecifying the number of change-points. For this purpose, we study the evolution of both $\{\mathcal{E}(\widehat{\mathcal{T}}_{n,K}\|\mathcal{T}^\star), \mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}_{n,K})\}$, as $K = 1, \ldots, 3K^\star$ in the three different noise settings.

4.2. **Unspecified number of change-points.**

4.2.1. *Performance of* LS-TV. We consider here the performance of LS-TV on the Blocks dataset corrupted with three different levels of noise, when the true number of change-points is unknown. For each noise level, we generated 100 replications of corrupted versions of the Blocks dataset. For each noise replication, we measured both $\mathcal{E}(\widehat{\mathcal{T}}_{n,K}\|\mathcal{T}^\star)$ and $\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}_{n,K})$ for all $K = 1, \ldots, 3K^\star$. We display in Figure 2 the results averaged over all replications for both errors. Also, note that the optimal trade-off between the two types of error is reached almost exactly at the true number of change-points $K = K^\star$.

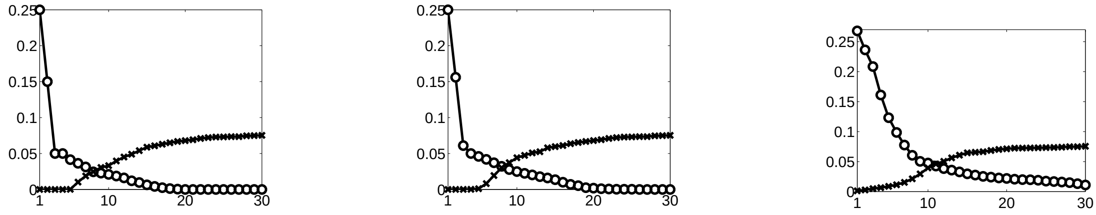

FIGURE 2. The evolution of the two types of error as $K = 1, \ldots, 3K^\star$, that is $\{\mathcal{E}(\widehat{\mathcal{T}}^{\text{LS-TV}}_{n,K}\|\mathcal{T}^\star)\}_{K=1,\ldots,3K^\star}$ ("o") and $\{\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}^{\text{LS-TV}}_{n,K})\}_{K=1,\ldots,3K^\star}$ ("×"), in different noise settings (low, medium, high noise from left to right).

4.2.2. *Comparison with the standard least-square* (LS) *approach.* Let us now compare the performance of LS-TV with the performance of the standard least-square estimation of multiple change-points theoretically studied by Yao and Au (1989). The latter criterion provides a number of K change-points for the model (1) by

$$\underset{t_1<\dots<t_K}{\text{Minimize}} \sum_{k=1}^{K} \sum_{i=t_{k-1}+1}^{t_k} (Y_i - \bar{\mu}_k)^2, \text{ where } \bar{\mu}_k \overset{\text{def}}{=} (t_k - t_{k-1})^{-1} \sum_{i=t_{k-1}+1}^{t_k} Y_i \,.$$

A computationally efficient way of solving this minimization is based on a dynamic programming algorithm (DP), originally proposed by Fisher (1958); Bellman (1961) and described in (Kay, 1993, Chapter 12). While a naive approach would require a $O(2^n)$ time-complexity, DP has a time-complexity of $O(K\,n^2)$ if we look for at most K change-points within the signal. For a fair comparison, we used exactly the same settings for both methods LS-TV and LS.

From Table 4 displayed in Section 5, we can see that LS-TV reaches satisfactory performance, in terms of both types of errors, in all noise settings as well as LS. It is worthwhile to emphasize that, while LS has a $O(Kn^2)$ time-complexity when implemented with the DP algorithm, our method LS-TV has only $O(Kn\log(n))$ time-complexity. We can also remark that the variance of $\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}^{\text{LS-TV}}_{n,K})$ is larger than the variance of $\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}^{\text{LS}}_{n,K})$. It is then interesting to remedy this issue, without harming the sub-quadratic time-complexity of LS-TV.

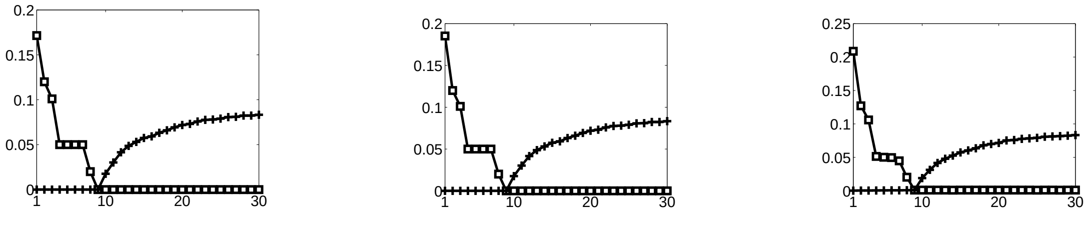

FIGURE 3. The evolution of the two types of error as $K = 1,\dots,3K^\star$, that is $\{\mathcal{E}(\widehat{\mathcal{T}}^{\text{LS-TV}*}_{n,K}\|\mathcal{T}^\star)\}_{K=1,\dots,3K^\star}$ displayed with squares and $\{\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}^{\text{LS-TV}*}_{n,K})\}_{K=1,\dots,3K^\star}$ ("+"), in different noise settings (low, medium and high noise from left to right).

In the next section, we show how LS-TV may be further enhanced, both in mean and variance in both types of errors, when combined with an additional step based on a reduced-search dynamic programming algorithm.

5. An enhanced version of LS-TV: LS-TV*

We now propose an enhanced version of LS-TV, called LS-TV*, which combines two steps. First, we run LS-TV with $K = K_{\max}$ larger than $K^\star$, and get a set of change-point estimates $\widehat{\mathcal{T}}_{n,K_{\max}}$. Second, we run a reduced version of DP searching $L < K_{\max}$ change-points over the set $\widehat{\mathcal{T}}_{n,K_{\max}}$, instead of $\{1, \dots, n\}$ as in the raw DP algorithm, which finally yields a new set of change-point estimates $\mathcal{S}_{n,L} \subsetneq \widehat{\mathcal{T}}_{n,K_{\max}}$.

From Table 4, we observe that for $K = 30$ the error $\mathcal{E}(\mathcal{T}^\star \| \widehat{\mathcal{T}}^{\mathrm{LS}}_{n,K})$ becomes larger than $\mathcal{E}(\mathcal{T}^\star \| \widehat{\mathcal{T}}^{\mathrm{LS\text{-}TV}}_{n,K})$ in all noise settings. This suggests that one type of error made by LS-TV stabilizes in the **over-segmentation** regime, that is when $K \gg K^\star$, whereas the same type of error made by LS still increases. Therefore, one might think of running LS-TV to look for an *a priori* much larger set of change-points than the true number of change-points, that is to look for $K_{\max} \gg K^\star$ change-points with $K_{\max} \ll n$, and then propose a way of selecting the best change-point estimates within the large set of change-point estimates obtained by LS-TV.

We suggest running a dynamic programming algorithm to perform this post-selection. More precisely, we aim at minimizing, for each K in $\{1, \dots, K_{\max}\}$:

$$\underset{\substack{t_1 < \cdots < t_K \\ \text{s.t } t_1, \dots, t_K \in \widehat{\mathcal{T}}_{n,K_{\max}}}}{\text{Minimize}} \quad \sum_{k=1}^{K} \sum_{i=t_{k-1}+1}^{t_k} (Y_i - \bar{\mu}_k)^2, \text{ where } \bar{\mu}_k \overset{\text{def}}{=} (t_k - t_{k-1})^{-1} \sum_{i=t_{k-1}+1}^{t_k} Y_i \,. \tag{23}$$

The above algorithm, subsequently called rDP, outputs for each $K = 1, \dots, K_{\max}$ a new set of change-point estimates $\mathcal{S}_{n,K} \subsetneq \widehat{\mathcal{T}}_{n,K_{\max}}$. We call LS-TV* the method which combines LS-TV with a post-selection based on rDP.

First, we investigate how LS-TV* improves on LS-TV in terms of error variance. The settings are the same as previously. We observe in Table 4 that the post-selection step indeed consistently reduces the variance of both errors obtained by LS-TV.

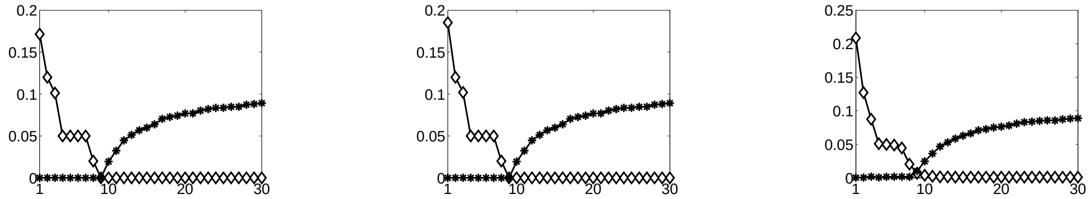

FIGURE 4. The evolution of the two types of error as $K = 1, \ldots, 3K^\star$, that is $\{\mathcal{E}(\widehat{\mathcal{T}}^{\text{LS-TV*}}_{n,K} \| \mathcal{T}^\star)\}_{K=1,\ldots,3K^\star}$ ("$\Diamond$") and $\{\mathcal{E}(\mathcal{T}^\star \| \widehat{\mathcal{T}}^{\text{LS-TV*}}_{n,K})\}_{K=1,\ldots,3K^\star}$ ("$*$"), in different noise settings (low, medium and high noise from left to right).

Second, we check whether LS-TV* improves, on average, the performance of LS-TV. As Figures 2, 4 and Table 4 show, not only LS-TV* yields much lower error rates than LS-TV in both types of errors, but LS-TV* also obtains similar error rates when compared to LS. Since the overall time-complexity of LS-TV* is $O(K_{\max}^3 + K_{\max} n \log(n))$, and the overall time-complexity of LS is $O(K_{\max} n^2)$, then, as long as $K^\star < K_{\max} \ll n$, LS-TV* obtains the same performance results as LS at a much lower computational cost. In order to give an idea to the reader of the actual computation times of LS-TV*and LS, we give in Table 3 the computation times of both methods when they are applied to the Blocks dataset for several values of n and $K_{\max}$.

$(n,K_{\max})$	(100,5)	(500,15)	(1000,30)
LS	0.021s	0.466s	2.464s
LS-TV*	0.005s	0.119s	0.689s

TABLE 3. Computation times in seconds of LS and LS-TV* for several values of n and $K_{\max}$.

Note that Figure 4 gives an appealing intuitive understanding of the statistical behavior of multiple change-point estimation methods. While the first type of error $\mathcal{E}(\widehat{\mathcal{T}} \| \mathcal{T}^\star)$ may be interpreted as the maximum error in the change-point location from estimated change-points to true change-points, the second type of error $\mathcal{E}(\mathcal{T}^\star \| \widehat{\mathcal{T}})$ may be interpreted as the maximum error in the change-point location from true change-points to estimated change-points. As the number of estimated change-points increases, the first type of error $\mathcal{E}(\widehat{\mathcal{T}}\|\mathcal{T}^\star)$ decreases while the second type of error $\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}})$ increases. Finally, $\mathcal{E}(\widehat{\mathcal{T}}\|\mathcal{T}^\star)$ quantifies the over-segmentation error while $\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}})$ quantifies the under-segmentation error.

As presented here LS-TV* does not include a model selection part. A thorough practical version of LS-TV* should incorporate a data-driven way of choosing the optimal number of change-points $\hat{K}$, and hence the optimal set of change-point estimates $\mathcal{S}_{n,\widehat{K}}$. This issue is left for future research.

6. Conclusion and prospects

The standard least-square estimation approach LS suffers from an overwhelming time-complexity for performing change-point estimation in long time series of observations. We showed, both theoretically and practically, that an alternative solution to the multiple change-point estimation problem, solved by a least-square fitting with a total variation penalty LS-TV, allowed us to get a lower time-complexity while keeping competitive performance in terms of change-point estimation, even in high-noise settings.

We see several future research directions for this work. In the last section of the paper, we proposed an enhanced version of LS-TV called LS-TV*, with better empirical performance and similar time-complexity. We would like to provide thorough theoretical support to this method, which would involve a statistical analysis of the two steps LS-TV and reduced DP (rDP). Besides, since a lot of real datasets include a non-negligible proportion of outliers, we would like to derive a robust version of both LS-TV and LS-TV*, and establish the corresponding theoretical results.

7. Proofs

Proof of Proposition 1. By definition of $\hat{\beta}^n(\lambda_n)$ given by (9), we have

$$\|Y^n - X_n\hat{\beta}^n(\lambda_n)\|_n^2 + \lambda_n\|\hat{\beta}^n(\lambda_n)\|_1 \leq \|Y^n - X_n\beta^n\|_n^2 + \lambda_n\|\beta^n\|_1 \; .$$

	$K=1$			$K=11$			$K=20$			$K=30$		
	Low	Medium	High	Low	Medium	High	Low	Medium	High	Low	Medium	High
LS	.169	.169	.169	.000	.000	.001	.000	.000	.001	.000	.000	.001
	(.014)	(.033)	(.041)	(.000)	(.000)	(.001)	(.000)	(.000)	(.001)	(.000)	(.000)	(.001)
	.000	.000	.000	.023	.023	.027	.072	.072	.072	.086	.086	.086
	(.000)	(.000)	(.001)	(.025)	(.025)	(.025)	(.018)	(.018)	(.017)	(.014)	(.014)	(.014)
LS-TV	.250	.250	.250	.020	.020	.040	.000	.000	.020	.000	.000	.019
	(.000)	(.000)	(.042)	(.006)	(.006)	(.009)	(.000)	(.000)	(.007)	(.000)	(.000)	(.009)
	.000	.000	.001	.034	.041	.042	.071	.071	.075	.081	.081	.081
	(.000)	(.000)	(.002)	(.030)	(.031)	(.028)	(.025)	(.025)	(.022)	(.020)	(.020)	(.020)
LS-TV*	.169	.169	.169	.000	.000	.001	.000	.000	.001	.000	.000	.001
	(.014)	(.033)	(.041)	(.000)	(.000)	(.005)	(.000)	(.000)	(.001)	(.000)	(.000)	(.001)
	.000	.000	.000	.029	.029	.033	.071	.071	.072	.082	.082	.082
	(.000)	(.000)	(.001)	(.023)	(.023)	(.024)	(.015)	(.015)	(.014)	(.013)	(.013)	(.013)

TABLE 4. Performance in terms of $\mathcal{E}(\widehat{\mathcal{T}}_{n,K}\|\mathcal{T}^\star)$ and $\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}_{n,K})$ for different values of K of LS, LS-TV and LS-TV* on the Blocks dataset corrupted with low-noise ($\sigma = 0.05$), medium-noise ($\sigma = 0.10$) and high-noise ($\sigma = 0.50$). For each method, the first and second lines correspond to the mean and standard deviation of $\mathcal{E}(\widehat{\mathcal{T}}_{n,K}\|\mathcal{T}^\star)$ respectively and the third and fourth lines correspond to the mean and standard deviation of $\mathcal{E}(\mathcal{T}^\star\|\widehat{\mathcal{T}}_{n,K})$ respectively. $K_{\max}$ was set to 30 in all experiments.

Using (7), we get

$$\|X_n(\beta^n - \hat{\beta}^n(\lambda_n))\|_n^2 + \frac{2}{n}(\beta^n - \hat{\beta}^n(\lambda_n))' X_n' \varepsilon^n + \lambda_n \sum_{k=1}^{n} |\hat{\beta}_k(\lambda_n)| \leq \lambda_n \sum_{k=1}^{n} |\beta_k^n| \,.$$

Thus,

$$\|X_n(\beta^n - \hat{\beta}^n(\lambda_n))\|_n^2 \leq \frac{2}{n}(\hat{\beta}^n(\lambda_n) - \beta^n)' X_n' \varepsilon^n + \lambda_n \sum_{j \in \mathcal{A}} (|\beta_j^n| - |\hat{\beta}_j(\lambda_n)|) - \lambda_n \sum_{j \in \bar{\mathcal{A}}} |\hat{\beta}_j(\lambda_n)| \,.$$

Observe that

$$\frac{2}{n}(\hat{\beta}^n(\lambda_n) - \beta^n)'X_n'\varepsilon^n = 2\sum_{j=1}^{n}(\hat{\beta}_j(\lambda_n) - \beta_j^n)\left(\frac{1}{n}\sum_{i=j}^{n}\varepsilon_i^n\right) .$$

Let us define the event $E = \bigcap_{j=1}^n \left\{n^{-1}\left|\sum_{i=j}^n \varepsilon_i^n\right| \leq \lambda_n/2\right\}$. Then, given that the ε_i^n's are iid zero-mean Gaussian random variables with variance σ^2, we obtain that

$$\mathbb{P}(\bar{E}) \leq \sum_{j=1}^{n}\mathbb{P}\left(n^{-1}\left|\sum_{i=j}^{n}\varepsilon_i^n\right| > \lambda_n/2\right) \leq \sum_{j=1}^{n}\exp\left(-\frac{n^2\lambda_n^2}{8\sigma^2(n-j+1)}\right) .$$

Thus, if $\lambda_n = C\sigma\sqrt{\log n/n}$,

$$\mathbb{P}(\bar{E}) \leq n^{1-C^2/8} .$$

With a probability larger than $1 - n^{1-C^2/8}$, we get

$$\|X_n(\beta^n - \hat{\beta}^n(\lambda_n))\|_n^2 \leq \lambda_n\sum_{j=1}^{n}|\hat{\beta}_j(\lambda_n) - \beta_j^n| + \lambda_n\sum_{j\in\mathcal{A}}(|\beta_j^n| - |\hat{\beta}_j(\lambda_n)|) - \lambda_n\sum_{j\in\bar{\mathcal{A}}}|\hat{\beta}_j(\lambda_n)| ,$$

where $\mathcal{A}$ and $\bar{\mathcal{A}}$ are defined in (8). Given that

$$\sum_{j=1}^{n}|\hat{\beta}_j(\lambda_n) - \beta_j^n| = \sum_{j\in\mathcal{A}}|\hat{\beta}_j(\lambda_n) - \beta_j^n| + \sum_{j\in\bar{\mathcal{A}}}|\hat{\beta}_j(\lambda_n)| ,$$

we obtain that, with a probability larger than $1 - n^{1-C^2/8}$,

$$\|X_n(\beta^n - \hat{\beta}^n(\lambda_n))\|_n^2 \leq 2\lambda_n\sum_{j\in\mathcal{A}}|\beta_j^n| = 2C\sigma\sqrt{\frac{\log n}{n}}\sum_{j\in\mathcal{A}}|\beta_j^n| \leq 2C\sigma\beta_{max}K^\star\sqrt{\frac{\log n}{n}} .$$

□

Proof of Proposition 2. For notational simplicity, we shall remove the dependence of $\hat{u}$ in λ_n. By definition of $\hat{u}$ as a minimizer of the criterion (11), we get:

$$\|Y^n - \hat{u}\|_n^2 + \lambda_n\sum_{i=1}^{n-1}|\hat{u}_{i+1} - \hat{u}_i| \leq \|Y^n - u^\star\|_n^2 + \lambda_n\sum_{i=1}^{n-1}|u_{i+1}^\star - u_i^\star| .$$

Using model (10), the previous inequality can be rewritten as follows:

$$\|\hat{u} - u^\star\|_n^2 \leq \lambda_n\left(\sum_{i=1}^{n-1}|u_{i+1}^\star - u_i^\star| - \sum_{i=1}^{n-1}|\hat{u}_{i+1} - \hat{u}_i|\right) + \frac{2}{n}\sum_{i=1}^{n}\varepsilon_i(\hat{u}_i - u_i^\star) .$$

Using the Cauchy-Schwarz inequality, we obtain

$$\|\hat{u}-u^\star\|_n^2 \leq 2n\lambda_n\|\hat{u}-u^\star\|_n + \frac{2}{n}\sum_{i=1}^n \varepsilon_i(\hat{u}_i-u_i^\star)\ .$$

Thus, defining $G(.)$ for v in $\mathbb{R}^n$, by: $G(v) = (\sum_{i=1}^n \varepsilon_i(v_i-u_i^\star))/(\sigma\sqrt{n}\|v-u^\star\|_n)$, we get:

$$\|\hat{u}-u^\star\|_n^2 \leq 2n\lambda_n\|\hat{u}-u^\star\|_n + \frac{2\sigma}{\sqrt{n}}\|\hat{u}-u^\star\|_n G(\hat{u})\ .$$

Let $\{S_K\}_{1\leq K\leq K_{\max}}$ be the collection of linear spaces to which $\hat{u}$ may belong, S_K denoting a space of dimension K. Then, given that the number of sets of dimension K is bounded by n^K, we obtain

$$\begin{aligned}\mathbb{P}(\|\hat{u}-u^\star\|_n \geq \alpha_n) &\leq \mathbb{P}(n\lambda_n + \sigma n^{-1/2}G(\hat{u}) \geq \alpha_n/2)\\ &\leq \sum_{K=1}^{K_{\max}} n^K\ \mathbb{P}(\sup_{v\in S_K} G(v) \geq n^{1/2}\sigma^{-1}\alpha_n/2 - n^{3/2}\sigma^{-1}\lambda_n)\ . \quad (24)\end{aligned}$$

Using that, $\mathrm{Var}(G(v)) = 1$, for all v in $\mathbb{R}^n$, we obtain by using an inequality due to Cirel'son, Ibragimov and Sudakov in the same way as in the proof of Theorem 1 in Birgé and Massart (2001), that for all $\beta > 0$,

$$\mathbb{P}(\sup_{v\in S_K} G(v) \geq \mathbb{E}[\sup_{v\in S_K} G(v)] + \beta) \leq \exp(-\beta^2/2)\ . \quad (25)$$

Let us now find an upper bound for $\mathbb{E}[\sup_{v\in S_K} G(v)]$. Denoting by W the D-dimensional space to which $v-u^\star$ belongs and some orthogonal basis $\psi_1,\dots,\psi_D$ of W, we obtain

$$\sup_{v\in S_K} G(v) \leq \sup_{w\in W}\frac{\sum_{i=1}^n \varepsilon_i w_i}{\sigma\sqrt{n}\|w\|_n} = \sup_{\alpha\in\mathbb{R}^D}\frac{\sum_{i=1}^n \varepsilon_i(\sum_{j=1}^D \alpha_j\psi_{j,i})}{\sigma\sqrt{n}\|\sum_{j=1}^D \alpha_j\psi_j\|_n} = \sup_{\alpha\in\mathbb{R}^D}\frac{\sum_{i=1}^n \varepsilon_i(\sum_{j=1}^D \alpha_j\psi_{j,i})}{\sigma\sqrt{n}(\sum_{j=1}^D \alpha_j^2)^{1/2}}\ .$$

Using the Cauchy-Schwarz inequality, we derive

$$\sup_{v\in S_K} G(v) \leq \sup_{\alpha\in\mathbb{R}^D}\frac{\sum_{j=1}^D \alpha_j(\sum_{i=1}^n \varepsilon_i\psi_{j,i})}{\sigma\sqrt{n}(\sum_{j=1}^D \alpha_j^2)^{1/2}} \leq (\sigma^2 n)^{-1/2}\left\{\sum_{j=1}^D\left(\sum_{i=1}^n \varepsilon_i\psi_{j,i}\right)^2\right\}^{1/2}\ .$$

By the concavity of the square root function and by using that $D \leq K_{\max} + K^\star + 1 \leq 2K_{\max}+1$, we get

$$\mathbb{E}[\sup_{v\in S_K} G(v)] \leq (2K_{\max}+1)^{1/2}\ . \quad (26)$$

Using (24), (25) and (26) with $\beta = n^{1/2}\sigma^{-1}\alpha_n/2 - n^{3/2}\sigma^{-1}\lambda_n - (2K_{\max}+1)^{1/2}$, we get

$$\mathbb{P}(\|\hat{u}-u^\star\|_n \geq \alpha_n) \leq K_{\max}\exp\left\{K_{\max}\log n - \frac{1}{2}\left(\frac{n^{1/2}\alpha_n}{2\sigma} - n^{3/2}\sigma^{-1}\lambda_n - (2K_{\max}+1)^{1/2}\right)^2\right\} ,$$

which is valid only if $\beta = n^{1/2}\sigma^{-1}\alpha_n/2 - n^{3/2}\sigma^{-1}\lambda_n - (2K_{\max}+1)^{1/2}$ is positive. Thus, writing for a constant A in $(0,1)$,

$$n^{3/2}\sigma^{-1}\lambda_n + (2K_{\max}+1)^{1/2} = An^{1/2}\sigma^{-1}\alpha_n/2 ,$$

gives

$$\mathbb{P}(\|\hat{u}-u^\star\|_n \geq \alpha_n) \leq K_{\max}\ \exp\left\{K_{\max}\log n - \frac{(1-A)^2}{8}\frac{n\alpha_n^2}{\sigma^2}\right\} .$$

Thus, if $\alpha_n = (B\sigma^2 K_{\max}\log n/n)^{1/2}$, we obtain the expected result. □

Proof of Lemma 3. A necessary and sufficient condition for a vector $\hat{\beta}$ in $\mathbb{R}^n$ to minimize Φ defined by: $\Phi(\beta) = \sum_{i=1}^n (Y_i - (X_n\beta)_i)^2 + n\lambda_n \sum_{i=1}^n |\beta_i|$, is that the zero vector in $\mathbb{R}^n$ belongs to the subdifferential of Φ at point $\hat{\beta}$ that is:

$$\begin{aligned}(X_n^T(Y^n - X_n\hat{\beta}))_j &= \frac{n\lambda_n}{2}\text{sign}(\hat{\beta}_j),\ \text{if } \hat{\beta}_j \neq 0 ,\\ |(X_n^T(Y^n - X_n\hat{\beta}))_j| &\leq \frac{n\lambda_n}{2},\ \text{if } \hat{\beta}_j = 0 .\end{aligned}$$

Using that $(X_n^T Y^n)_j = \sum_{k=j}^n Y_k$ and that $(X_n^T\hat{u})_j = \sum_{k=j}^n \hat{u}_k$, since X_n is a $n\times n$ lower triangular matrix having all its non zero elements equal to one, we obtain the expected result. □

In the remainder, for any sequence of random variables, say, $Z_1,\dots,Z_n$, we shall use the following notation

$$Z(r;s) \overset{\text{def}}{=} \sum_{i=r}^{s} Z_i,\quad \text{for any } 1 \leq r < s \leq n. \tag{27}$$

Proof of Lemma 4. Using the notation introduced in (27), we obtain

$$\mathbb{P}\left(\max_{\substack{1\leq r_n<s_n\leq n\\ |r_n-s_n|\geq v_n}}\left|\frac{\varepsilon(r_n;s_n-1)}{s_n-r_n}\right| \geq x_n\right) \leq \sum_{\substack{1\leq r_n<s_n\leq n\\ |r_n-s_n|\geq v_n}} \mathbb{P}\left(\left|\frac{\varepsilon(r_n;s_n-1)}{s_n-r_n}\right| \geq x_n\right) .$$

Using Assumption (A1), we get that for all $\eta > 0$,

$$\begin{aligned}\mathbb{P}\left(\frac{\varepsilon(r_n; s_n-1)}{s_n - r_n} \geq x_n\right) &\leq \exp[-\eta(s_n-r_n)x_n][\mathbb{E}\{\exp(\eta\varepsilon)\}]^{(s_n-r_n)} \\ &\leq \exp[-\eta(s_n-r_n)x_n + \beta\eta^2(s_n-r_n)] \;.\end{aligned}$$

Since the sharpest bound holds for $\eta = x_n/2\beta$, we get

$$\mathbb{P}\left(\frac{\varepsilon(r_n; s_n-1)}{s_n - r_n} \geq x_n\right) \leq \exp[-x_n^2(s_n-r_n)/4\beta] \;.$$

Since the same bound is valid when ε_i is replaced by $-\varepsilon_i$, we get that

$$\mathbb{P}\left(\left|\frac{\varepsilon(r_n; s_n-1)}{s_n - r_n}\right| \geq x_n\right) \leq 2\exp[-x_n^2(s_n-r_n)/4\beta] \;.$$

Hence, we obtain that

$$\mathbb{P}\left(\max_{\substack{1\leq r_n < s_n \leq n \\ |r_n - s_n| \geq v_n}} \left|\frac{\varepsilon(r_n; s_n-1)}{s_n - r_n}\right| \geq x_n\right) \leq 2n^2\exp[-v_n x_n^2/4\beta] \;,$$

which completes the proof. □

Proof of Proposition 5. In this proof, we shall use the notation introduced in (27). Since $\mathbb{P}(\max_{1\leq k\leq K^\star}|\hat{t}_k - t_k^\star| > n\delta_n) \leq \sum_{k=1}^{K^\star}\mathbb{P}(|\hat{t}_k - t_k^\star| > n\delta_n)$, it suffices to prove that for all $k = 1,\dots,K^\star$, $\mathbb{P}(A_{n,k}) \to 0$, where $A_{n,k} = \{|\hat{t}_k - t_k^\star| \geq n\delta_n\}$. Defining the set C_n by

$$C_n = \left\{\max_{0\leq k\leq K^\star}|\hat{t}_k - t_k^\star| < I_{\min}^\star/2\right\} \;, \tag{28}$$

it is enough to prove that $\mathbb{P}(A_{n,k}\cap C_n) \to 0$, and that $\mathbb{P}(A_{n,k}\cap\overline{C}_n) \to 0$. Let us first prove the first statement. Note that (28) implies that

$$t_{k-1}^\star < \hat{t}_k < t_{k+1}^\star, \text{ for all } k \in \{1,\dots,K^\star\} \;.$$

Let us first consider the case where $\hat{t}_k \leq t_k^\star$. Applying (18) in Lemma 3 with $j = t_k^\star$ and (17) in Lemma 3 with $\ell = k$ gives respectively

$$\left|\sum_{i=t_k^\star}^{n} Y_i - \sum_{i=t_k^\star}^{n}\hat{u}_i\right| \leq n\lambda_n/2 \quad \text{and} \quad \sum_{i=\hat{t}_k}^{n} Y_i - \sum_{i=\hat{t}_k}^{n}\hat{u}_i = n\hat{\alpha}_k\lambda_n/2 \;.$$

This yields, using (19) in Lemma 3, that the event $C_{n,k}$ defined as follows, occurs with probability one:

$$C_{n,k} = \left\{ |(\hat{t}_k - t_k^\star)(\mu_{k+1}^\star - \mu_k^\star) + (\hat{t}_k - t_k^\star)(\hat{\mu}_{k+1} - \mu_{k+1}^\star) + \varepsilon(\hat{t}_k; t_k^\star - 1) \leq n\lambda_n \right\} \ .$$

Using that $\mathbb{P}(A_{n,k} \cap C_n) = \mathbb{P}(A_{n,k} \cap C_{n,k} \cap C_n)$, we get

$$\begin{aligned}\mathbb{P}(A_{n,k} \cap C_n) &\leq \mathbb{P}\left(n\lambda_n/n\delta_n \geq |\mu_{k+1}^\star - \mu_k^\star|/3\right) + \mathbb{P}\left(\{|\hat{\mu}_{k+1} - \mu_{k+1}^\star| \geq |\mu_{k+1}^\star - \mu_k^\star|/3\} \cap C_n\right) \\ &\quad + \mathbb{P}\left(\left\{\left|\frac{\varepsilon(\hat{t}_k; t_k^\star - 1)}{t_k^\star - \hat{t}_k}\right| \geq |\mu_{k+1}^\star - \mu_k^\star|/3\right\} \cap A_{n,k}\right) \\ &\overset{\text{def}}{=} \mathbb{P}(A_{n,k,1}) + \mathbb{P}(A_{n,k,2}) + \mathbb{P}(A_{n,k,3}) \ .\end{aligned}$$

By Assumption (A4), $n\lambda_n/(n\delta_n J_{\min}^\star) < 1/3$, for n large enough, leading to $\mathbb{P}(A_{n,k,1}) \to 0$. By Lemma 4 with $x_n = J_{\min}^\star/3$, $v_n = n\delta_n$ and Assumption (A2), $\mathbb{P}(A_{n,k,3}) \to 0$. Let us now address $\mathbb{P}(A_{n,k,2})$. Using (18) in Lemma 3 with $j = (t_k^\star + t_{k+1}^\star)/2$ and with $j = t_k^\star$, and using the triangle inequality, we get

$$\left| \sum_{i=t_k^\star}^{(t_k^\star + t_{k+1}^\star)/2 \ -1} Y_i - \sum_{i=t_k^\star}^{(t_k^\star + t_{k+1}^\star)/2 \ -1} \hat{u}_i \right| \leq n\lambda_n \ .$$

Since we are in the event C_n and $\hat{t}_k \leq t_k^\star$, $\hat{u}_i \equiv \hat{\mu}_{k+1}$ within the interval $[t_k^\star, (t_k^\star + t_{k+1}^\star)/2 \ -1]$, which gives $|(t_{k+1}^\star - t_k^\star)\left(\mu_{k+1}^\star - \hat{\mu}_{k+1}\right)/2 + \varepsilon(t_k^\star; (t_k^\star + t_{k+1}^\star)/2 \ -1)| \leq n\lambda_n$. This implies that

$$(t_{k+1}^\star - t_k^\star)\left|\mu_{k+1}^\star - \hat{\mu}_{k+1}\right|/2 \leq n\lambda_n + \left|\varepsilon(t_k^\star; (t_k^\star + t_{k+1}^\star)/2 \ -1)\right| \ .$$

Therefore, we may upper bound $\mathbb{P}(A_{n,k,2})$ as follows

$$\begin{aligned}&\mathbb{P}\left(\{|\hat{\mu}_{k+1} - \mu_{k+1}^\star| \geq |\mu_{k+1}^\star - \mu_k^\star|/3\} \cap C_n\right) \\ &\leq \mathbb{P}\left(n\lambda_n \geq (t_{k+1}^\star - t_k^\star)|\mu_{k+1}^\star - \mu_k^\star|/12\right) + \mathbb{P}\left(\left|\frac{\varepsilon(t_k^\star; (t_k^\star + t_{k+1}^\star)/2 \ -1)}{t_{k+1}^\star - t_k^\star}\right| \geq |\mu_{k+1}^\star - \mu_k^\star|/6\right) \ ,\end{aligned}$$

which is arbitrarily small if $n\lambda_n < I_{\min}^\star \cdot J_{\min}^\star/12$ for n large enough, and, by Lemma 4, if $I_{\min}^\star (J_{\min}^\star)^2/\log(n) \to \infty$, as n tends to infinity. The last two conditions hold thanks to Assumptions (A2), (A3) and (A4). Since the proof in the case $\hat{t}_k \geq t_k^\star$ follows from similar reasoning, we have proved that $\mathbb{P}(A_{n,k} \cap C_n) \to 0$, as n tends to infinity.

We now prove that $\mathbb{P}(A_{n,k} \cap \overline{C}_n) \to 0$. Recall that by definition of C_n given in (28), $\overline{C}_n = \left\{\max_{k\in\{1,\dots,K^\star\}} |\hat{t}_k - t_k^\star| \geq I_{\min}^\star/2\right\}$. We now split $\mathbb{P}(A_{n,k} \cap \overline{C}_n)$ into three terms:

$$\mathbb{P}(A_{n,k} \cap \overline{C}_n) = \mathbb{P}(A_{n,k} \cap D_n^{(l)}) + \mathbb{P}(A_{n,k} \cap D_n^{(m)}) + \mathbb{P}(A_{n,k} \cap D_n^{(r)}) \; .$$

where

$$\begin{aligned}
D_n^{(\ell)} &\overset{\text{def}}{=} \left\{\text{there exists}\, p \in \{1,\dots,K^\star\}\;,\; \hat{t}_p \leq t_{p-1}^\star\right\} \cap \overline{C}_n \;, \\
D_n^{(m)} &\overset{\text{def}}{=} \left\{\text{for all}\, k \in \{1,\dots,K^\star\}\;,\; t_{k-1}^\star < \hat{t}_k < t_{k+1}^\star\right\} \cap \overline{C}_n \;, \\
D_n^{(r)} &\overset{\text{def}}{=} \left\{\text{there exists}\, p \in \{1,\dots,K^\star\}\;,\; \hat{t}_p \geq t_{p+1}^\star\right\} \cap \overline{C}_n \;.
\end{aligned}$$

Let us first focus on $\mathbb{P}(A_{n,k} \cap D_n^{(m)})$ and consider the case where $\hat{t}_k \leq t_k^\star$, since the other case can be addressed in a similar way. Note that

$$\mathbb{P}(A_{n,k} \cap D_n^{(m)}) \leq \mathbb{P}(A_{n,k} \cap B_{k+1,k} \cap D_n^{(m)}) + \sum_{l=k+1}^{K^\star} \mathbb{P}(C_{l,l} \cap B_{l+1,l} \cap D_n^{(m)}) \;, \tag{29}$$

where $B_{p,q} = \{(\hat{t}_p - t_q^\star) \geq I_{\min}^\star/2\}$ with the convention $B_{K^\star+1,K^\star} = \{(n - t_{K^\star}^\star) \geq I_{\min}^\star/2\}$ and $C_{p,q} = \{(t_p^\star - \hat{t}_q) \geq I_{\min}^\star/2\}$. Let us now prove that the first term in the right-hand side of (29) tends to zero as n tends to infinity, the arguments for addressing the other terms being similar. Using (18) and (17) in Lemma 3 with $j = t_k^\star$ and $\ell = k$, on the one hand and (18) in Lemma 3 with $j = t_k^\star$ and (17) in Lemma 3 with $\ell = k+1$ on the other hand, we obtain respectively:

$$\left|\hat{t}_k - t_k^\star\right| |\hat{\mu}_{k+1} - \mu_k^\star| \leq n\lambda_n + \left|\varepsilon(\hat{t}_k; t_k^\star - 1)\right| \;\text{ and }\; \left|\hat{t}_{k+1} - t_k^\star\right| \left|\hat{\mu}_{k+1} - \mu_{k+1}^\star\right| \leq n\lambda_n + \left|\varepsilon(t_k^\star; \hat{t}_{k+1} - 1)\right| \;. \tag{30}$$

Defining E_n by:

$$E_n = \{|\mu_{k+1}^\star - \mu_k^\star| \leq \; n\lambda_n/(n\delta_n) + 2n\lambda_n/I_{\min}^\star + (t_k^\star - \hat{t}_k)^{-1}|\varepsilon(\hat{t}_k; t_k^\star - 1)| + (\hat{t}_{k+1} - t_k^\star)^{-1}|\varepsilon(t_k^\star; \hat{t}_{k+1} - 1)|\} \;,$$

we obtain

$$
\begin{aligned}
&\mathbb{P}(A_{n,k} \cap B_{k+1,k} \cap D_n^{(m)}) \le \mathbb{P}(E_n \cap \{(t_k^\star - \hat{t}_k) \ge n\delta_n\} \cap \{(\hat{t}_{k+1} - t_k^\star) \ge I_{\min}^\star/2\}) \\
&\le \mathbb{P}(n\lambda_n/(n\delta_n) \ge |\mu_{k+1}^\star - \mu_k^\star|/4) + \mathbb{P}(2n\lambda_n/I_{\min}^\star \ge |\mu_{k+1}^\star - \mu_k^\star|/4) \\
&+ \mathbb{P}(\{(t_k^\star - \hat{t}_k)^{-1}|\varepsilon(\hat{t}_k; t_k^\star - 1)| \ge |\mu_{k+1}^\star - \mu_k^\star|/4\} \cap \{(t_k^\star - \hat{t}_k) \ge n\delta_n\}) \\
&+ \mathbb{P}(\{(\hat{t}_{k+1} - t_k^\star)^{-1}|\varepsilon(t_k^\star; \hat{t}_{k+1} - 1)| \ge |\mu_{k+1}^\star - \mu_k^\star|/4\} \cap \{(\hat{t}_{k+1} - t_k^\star) \ge I_{\min}^\star/2\}) \ .
\end{aligned}
$$

By Assumptions (A2), (A3) and (A4), $\mathbb{P}(A_{n,k} \cap B_{k+1,k} \cap D_n^{(m)}) \to 0$, as n tends to infinity, which concludes that $\mathbb{P}(A_{n,k} \cap D_n^{(m)}) \to 0$.

Let us now focus on $\mathbb{P}(A_{n,k} \cap D_n^{(\ell)})$. The latter probability can be upper bounded by

$$
\begin{aligned}
\mathbb{P}(D_n^{(\ell)}) &\le \sum_{k=1}^{K^\star} 2^{k-1}\mathbb{P}(\max\{1 \le l \le K^\star,\ \hat{t}_l \le t_{l-1}^\star\} = k) \\
&\le 2^{K^\star-1} \sum_{k=1}^{K^\star-1} \sum_{m \ge k}^{K^\star-1} \mathbb{P}(\{t_m^\star - \hat{t}_m > I_{\min}^\star/2\} \cap \{\hat{t}_{m+1} - t_m^\star > I_{\min}^\star/2\}) \\
&+ 2^{K^\star-1}\mathbb{P}(\{t_{K^\star}^\star - \hat{t}_{K^\star} > I_{\min}^\star/2\}) \ . \qquad (31)
\end{aligned}
$$

Consider one term of the sum in the right-hand of (31). Using (30) with $k = m$, we get

$$
\begin{aligned}
&\mathbb{P}(\{t_m^\star - \hat{t}_m > I_{\min}^\star/2\} \cap \{\hat{t}_{m+1} - t_m^\star > I_{\min}^\star/2\}) \le \mathbb{P}(4n\lambda_n/I_{\min}^\star \ge |\mu_{m+1}^\star - \mu_m^\star|/3) \\
&+ \mathbb{P}(\{(t_m^\star - \hat{t}_m)^{-1}|\varepsilon(\hat{t}_m; t_m^\star - 1)| \ge |\mu_{m+1}^\star - \mu_m^\star|/3\} \cap \{(t_m^\star - \hat{t}_m) \ge I_{\min}^\star/2\}) \\
&+ \mathbb{P}(\{(\hat{t}_{m+1} - t_m^\star)^{-1}|\varepsilon(t_m^\star; \hat{t}_{m+1} - 1)| \ge |\mu_{m+1}^\star - \mu_m^\star|/3\} \cap \{(\hat{t}_{m+1} - t_m^\star) \ge I_{\min}^\star/2\}) \ .
\end{aligned}
$$

By Assumptions (A2), (A3) and (A4), $\mathbb{P}(\{t_m^\star - \hat{t}_m > I_{\min}^\star/2\} \cap \{\hat{t}_{m+1} - t_m^\star > I_{\min}^\star/2\}) \to 0$, as n tends to infinity. Let us now consider the last term in the right-hand side of (31). Using (30) with $k = K^\star$ leads to

$$
\begin{aligned}
&\mathbb{P}(\{t_{K^\star}^\star - \hat{t}_{K^\star} > I_{\min}^\star/2\}) \le \mathbb{P}(3n\lambda_n/I_{\min}^\star \ge |\mu_{K^\star+1}^\star - \mu_{K^\star}^\star|/3) \\
&+ \mathbb{P}(\{(t_{K^\star}^\star - \hat{t}_{K^\star})^{-1}|\varepsilon(\hat{t}_{K^\star}; t_{K^\star}^\star - 1)| \ge |\mu_{K^\star+1}^\star - \mu_{K^\star}^\star|/3\} \cap \{(t_{K^\star}^\star - \hat{t}_{K^\star}) \ge I_{\min}^\star/2\}) \\
&+ \mathbb{P}(\{(n - t_{K^\star}^\star + 1)^{-1}|\varepsilon(t_{K^\star}^\star; n)| \ge |\mu_{K^\star+1}^\star - \mu_{K^\star}^\star|/3\}) \ .
\end{aligned}
$$

By Assumptions (A2), (A3) and (A4), $\mathbb{P}(\{t^\star_{K^\star} - \hat{t}_{K^\star} > I^\star_{\min}/2\}) \to 0$, as n tends to infinity, which gives: $\mathbb{P}(D_n^{(\ell)}) \to 0$. In a similar way, we can prove that: $\mathbb{P}(D_n^{(r)}) \to 0$, as n tends to infinity which gives that $\mathbb{P}(A_{n,k} \cap \overline{C}_n) \to 0$ and concludes the proof. □

Proof of Proposition 6. In this proof, we shall use the notation introduced in (27). By Lemma 2 of Meinshausen and Yu (2009), we get that with probability tending to one

$$|\hat{\mathcal{A}}(\lambda_n)| \leq C\frac{n}{\lambda_n^2} , \tag{32}$$

where C is a positive constant equal to $\sigma^2 + K^{\star 2} J^{\star 2}_{\max}$. In order to prove that

$$\mathbb{P}(\{\mathcal{E}(\hat{\mathcal{T}}_{n,|\hat{\mathcal{A}}(\lambda_n)|} \| \mathcal{T}_n^\star) \geq n\delta_n\} \cap \{|\hat{\mathcal{A}}(\lambda_n)| \geq K^\star\}) \to 0, \text{ as } n \to \infty ,$$

it is enough to prove that

$$\mathbb{P}(\{\mathcal{E}(\hat{\mathcal{T}}_{n,|\hat{\mathcal{A}}(\lambda_n)|} \| \mathcal{T}_n^\star) \geq n\delta_n\} \cap \{K^\star \leq |\hat{\mathcal{A}}(\lambda_n)| \leq Cn/\lambda_n^2\}) \to 0, \text{ as } n \to \infty .$$

Note that

$$\begin{aligned}\mathbb{P}(\{\mathcal{E}(\hat{\mathcal{T}}_{n,|\hat{\mathcal{A}}(\lambda_n)|} \| \mathcal{T}_n^\star)\} &\cap \{K^\star \leq |\hat{\mathcal{A}}(\lambda_n)| \leq Cn/\lambda_n^2\}) \\ &\leq \mathbb{P}(\mathcal{E}(\hat{\mathcal{T}}_{n,K^\star} \| \mathcal{T}_n^\star) \geq n\delta_n) + \sum_{K > K^\star}^{Cn/\lambda_n^2} \mathbb{P}(\mathcal{E}(\hat{\mathcal{T}}_{n,K} \| \mathcal{T}_n^\star) \geq n\delta_n) . \end{aligned} \tag{33}$$

The first term of the right hand-side of (33) tends to zero as $n \to \infty$ since it is upper bounded by $\mathbb{P}(\max_{1 \leq k \leq K^\star} |\hat{t}_k - t^\star_k| > n\delta_n)$ which tends to zero by Proposition 5. Let us now focus on the second term on the right hand-side of (33). Note that

$$\begin{aligned}\sum_{K > K^\star}^{Cn/\lambda_n^2} \mathbb{P}(\mathcal{E}(\hat{\mathcal{T}}_{n,K} \| \mathcal{T}_n^\star) \geq n\delta_n) &\leq \sum_{K > K^\star}^{Cn/\lambda_n^2} \sum_{k=1}^{K^\star} \mathbb{P}(\forall 1 \leq l \leq K,\ |\hat{t}_l - t^\star_k| \geq n\delta_n) \\ &\stackrel{\text{def}}{=} \sum_{K > K^\star}^{Cn/\lambda_n^2} \sum_{k=1}^{K^\star} \mathbb{P}(E_{n,k,1}) + \mathbb{P}(E_{n,k,2}) + \mathbb{P}(E_{n,k,3}) ,\end{aligned}$$

where

$$
\begin{aligned}
&E_{n,k,1} = \{\forall 1 \le l \le K,\ |\hat{t}_l - t_k^\star| \ge n\delta_n \text{ and } \hat{t}_l < t_k^\star\},\\
&E_{n,k,2} = \{\forall 1 \le l \le K,\ |\hat{t}_l - t_k^\star| \ge n\delta_n \text{ and } \hat{t}_l > t_k^\star\},\\
&E_{n,k,3} = \{\exists 1 \le l \le K-1,\ |\hat{t}_l - t_k^\star| \ge n\delta_n,\ |\hat{t}_{l+1} - t_k^\star| \ge n\delta_n \text{ and } \hat{t}_l < t_k^\star < \hat{t}_{l+1}\}\ .
\end{aligned}
$$

Let us first upper bound $\mathbb{P}(E_{n,k,1})$. Remark that

$$\mathbb{P}(E_{n,k,1}) = \mathbb{P}(E_{n,k,1} \cap \{\hat{t}_K > t_{k-1}^\star\}) + \mathbb{P}(E_{n,k,1} \cap \{\hat{t}_K \le t_{k-1}^\star\}).$$

Applying (18) in Lemma 3 with $j = t_k^\star$ and (17) in Lemma 3 with $\ell = K$ in the case where $\hat{t}_K > t_{k-1}^\star$ gives with probability one

$$|(t_k^\star - \hat{t}_K)\{(\mu_k^\star - \mu_{k+1}^\star) + (\mu_{k+1}^\star - \hat{\mu}_{K+1})\} + \varepsilon(\hat{t}_K; t_k^\star - 1)| \le n\lambda_n\ .$$

Thus,

$$
\begin{aligned}
\mathbb{P}(E_{n,k,1} \cap \{\hat{t}_K > t_{k-1}^\star\}) &\le \mathbb{P}(n\lambda_n/(n\delta_n) \ge |\mu_k^\star - \mu_{k+1}^\star|/3) + \mathbb{P}(|\mu_{k+1}^\star - \hat{\mu}_{K+1}| \ge |\mu_k^\star - \mu_{k+1}^\star|/3)\\
&\quad + \mathbb{P}(\{|(t_k^\star - \hat{t}_K)^{-1}\varepsilon(\hat{t}_K; t_k^\star - 1)| \ge |\mu_k^\star - \mu_{k+1}^\star|/3\} \cap \{|t_k^\star - \hat{t}_K| \ge n\delta_n\})\\
&\overset{\text{def}}{=} \mathbb{P}(E_{n,k,1}^{(1)}) + \mathbb{P}(E_{n,k,1}^{(2)}) + \mathbb{P}(E_{n,k,1}^{(3)})\ .
\end{aligned}
$$

By Assumption (A4), $n\lambda_n/(n\delta_n J_{\min}^\star) < 1/3$, for n large enough, leading to $CnK^\star/\lambda_n^2\mathbb{P}(E_{n,k,1}^{(1)}) \to 0$. By Lemma 4 with $x_n = J_{\min}^\star/3$, $v_n = n\delta_n$ and using that $n\delta_n {J_{\min}^\star}^2/\log(n^3/\lambda_n^2) \to \infty$, $CnK^\star/\lambda_n^2\mathbb{P}(E_{n,k,1}^{(3)}) \to 0$. Let us now address $\mathbb{P}(E_{n,k,1}^{(2)})$. Using (18) in Lemma 3 with $j = t_k^\star$ and with $j = t_{k+1}^\star$, we get

$$(t_{k+1}^\star - t_k^\star)|\mu_{k+1}^\star - \hat{\mu}_{K+1}| \le n\lambda_n + |\varepsilon(t_k^\star; t_{k+1}^\star - 1)|\ .$$

Therefore, we may upper bound $\mathbb{P}(E_{n,k,1}^{(2)})$ as follows:

$$
\begin{aligned}
&\mathbb{P}(|\mu_{k+1}^\star - \hat{\mu}_{K+1}| \ge |\mu_k^\star - \mu_{k+1}^\star|/3)\\
&\quad \le \mathbb{P}(n\lambda_n \ge (t_{k+1}^\star - t_k^\star)|\mu_k^\star - \mu_{k+1}^\star|/6) + \mathbb{P}(|(t_{k+1}^\star - t_k^\star)^{-1}\varepsilon(t_k^\star; t_{k+1}^\star - 1)| \ge |\mu_k^\star - \mu_{k+1}^\star|/6)\ .
\end{aligned}
$$

By using Assumptions (A2), (A3) and $n\delta_n {J^\star_{\min}}^2/\log(n^3/\lambda_n^2) \to \infty$, we conclude as previously that $CnK^\star/\lambda_n^2 \mathbb{P}(E^{(2)}_{n,k,1}) \to 0$. The same arguments can be used for addressing $\mathbb{P}(E_{n,k,1} \cap \{\hat{t}_K \leq t^\star_{k-1}\})$. We can address in the same way the term $\mathbb{P}(E_{n,k,2})$.

Let us now focus on $\mathbb{P}(E_{n,k,3})$. Note that $\mathbb{P}(E_{n,k,3})$ can be split into four terms as follows:

$$\mathbb{P}(E_{n,k,3}) = \mathbb{P}(E^{(1)}_{n,k,3}) + \mathbb{P}(E^{(2)}_{n,k,3}) + \mathbb{P}(E^{(3)}_{n,k,3}) + \mathbb{P}(E^{(4)}_{n,k,3}) \ ,$$

where

$$\begin{aligned}
E^{(1)}_{n,k,3} &= E_{n,k,3} \cap \{t^\star_{k-1} < \hat{t}_l < \hat{t}_{l+1} < t^\star_{k+1}\} \ ,\\
E^{(2)}_{n,k,3} &= E_{n,k,3} \cap \{t^\star_{k-1} < \hat{t}_l < t^\star_{k+1}, \hat{t}_{l+1} \geq t^\star_{k+1}\} \ ,\\
E^{(3)}_{n,k,3} &= E_{n,k,3} \cap \{\hat{t}_l \leq t^\star_{k-1}, t^\star_{k-1} < \hat{t}_{l+1} < t^\star_{k+1}\} \ ,\\
E^{(4)}_{n,k,3} &= E_{n,k,3} \cap \{\hat{t}_l \leq t^\star_{k-1}, t^\star_{k+1} \leq \hat{t}_{l+1}\} \ .
\end{aligned}$$

As for addressing $\mathbb{P}(E_{n,k,1} \cap \{\hat{t}_K > t^\star_{k-1}\})$, we have to use Lemma 3 two times. For $\mathbb{P}(E^{(1)}_{n,k,3})$, we first use (18) and (17) in Lemma 3 with $j = t^\star_k$ and $\ell = l$ respectively. Second, we use (18) and (17) in Lemma 3 with $j = t^\star_k$ and $\ell = l+1$ respectively. For $\mathbb{P}(E^{(2)}_{n,k,3})$, we first use Lemma 3 with $j = t^\star_k$ and $\ell = l$. Second, we use Lemma 3 with $j = t^\star_k$ and $j = t^\star_{k+1}$. For $\mathbb{P}(E^{(3)}_{n,k,3})$, we first use Lemma 3 with $j = t^\star_{k-1}$ and $j = t^\star_k$. Second, we use Lemma 3 with $j = t^\star_k$ and $\ell = l+1$. Finally, for $\mathbb{P}(E^{(4)}_{n,k,3})$, we first use Lemma 3 with $j = t^\star_{k-1}$ and $j = t^\star_k$. Second, we use Lemma 3 with $j = t^\star_k$ and $j = t^\star_{k+1}$. □

Appendix

Discussion about Condition (15). Let us compute the different matrices arising in (15). The matrix $C^n_{\mathcal{A}\mathcal{A}}$ is a $K^\star \times K^\star$ matrix defined by:

$$nC^n_{\mathcal{A}\mathcal{A}} = \begin{pmatrix}
n - t^\star_1 + 1 & n - t^\star_2 + 1 & n - t^\star_3 + 1 & \dots & n - t^\star_{K^\star} + 1\\
n - t^\star_2 + 1 & n - t^\star_2 + 1 & n - t^\star_3 + 1 & \dots & n - t^\star_{K^\star} + 1\\
n - t^\star_3 + 1 & n - t^\star_3 + 1 & n - t^\star_3 + 1 & \dots & n - t^\star_{K^\star} + 1\\
\vdots & \vdots & \vdots & & \vdots\\
n - t^\star_{K^\star} + 1 & n - t^\star_{K^\star} + 1 & n - t^\star_{K^\star} + 1 & \dots & n - t^\star_{K^\star} + 1
\end{pmatrix} . \tag{34}$$

As for $(C^n_{\mathcal{A}\mathcal{A}})^{-1}$, it is a $K^\star \times K^\star$ symmetric tridiagonal matrix satisfying:

$$n^{-1}(C^n_{\mathcal{A}\mathcal{A}})^{-1} \tag{35}$$
$$= \begin{pmatrix} d_{2,1} & -d_{2,1} & 0 & 0 & \dots & \dots & 0 \\ -d_{2,1} & d_{2,1}+d_{3,2} & -d_{3,2} & 0 & \dots & \dots & 0 \\ 0 & -d_{3,2} & d_{3,2}+d_{4,3} & -d_{4,3} & 0 & \dots & 0 \\ 0 & 0 & \ddots & \ddots & \ddots & & \vdots \\ 0 & 0 & \dots & 0 & 0 & -d_{K^\star,K^\star-1} & d_{K^\star+1,K^\star}+d_{K^\star,K^\star-1} \end{pmatrix},$$

where $d_{k,l} = (t^\star_k - t^\star_l)^{-1}$, for $1 \leq k, l \leq K^\star$ and $d_{K^\star+1,K^\star} = (n - t^\star_{K^\star} + 1)^{-1}$.

Since $a_{1,1} = 1$ where $A = (a_{i,j})_{1\leq i\leq n-K^\star, 1\leq j\leq K^\star} = C^n_{\bar{\mathcal{A}}\mathcal{A}}(C^n_{\mathcal{A}\mathcal{A}})^{-1}$ and $a_{1,j} = 0$, for all $2 \leq j \leq K^\star$, the irrepresentable condition (15) is clearly not satisfied.

Discussion about Condition (16). Let $\mathcal{M} = \{t_1, \dots, t_m\}$ be a set of indices of cardinal m. Using (35), one can see that, as soon as $\mathcal{M}$ is such that $t_j - t_i = 1$ for all i and j such that $j - i = 1$, $n^{-1}(C^n_{\mathcal{A}\mathcal{A}})^{-1}$ is a tridiagonal matrix with diagonal terms equal to 2 except the first one which is equal to 1 and extra diagonal terms equal to -1. Such a matrix is symmetric and positive since all the determinants of its sub-matrices are equal to 1. Thus, the maximal eigenvalue of $(C^n_{\mathcal{A}\mathcal{A}})^{-1}$ is larger than n implying that the minimal eigenvalue of $C^n_{\mathcal{A}\mathcal{A}}$ is smaller than $1/n$. Hence, Condition (16) is not fulfilled.

References

Basseville, M. and N. Nikiforov (1993). *The detection of abrupt changes.* Information and System sciences series. Prentice-Hall.

Bellman, R. (1961). On the approximation of curves by line segments using dynamic programming. *Communications of the ACM 4*(6), 284–294.

Bickel, P., Y. Ritov, and A. Tsybakov (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics 37*(4), 1705–1732.

Birgé, L. and P. Massart (2001). Gaussian model selection. *Journal of the European Mathematical Society 3*, 203–268.

Boysen, L., A. Kempe, A. Munk, V. Liebscher, and O. Wittich (2009). Consistencies and rates of convergence of jump penalized least squares estimators. *Annals of Statistics 37*(1), 157–183.

Brodsky, B. and B. Darkhovsky (1993). *Nonparametric methods in change-point problems.* the Netherlands: Kluwer Academic Publishers.

Brodsky, B. and B. Darkhovsky (2000). *Non-parametric statistical diagnosis: problems and methods.* Kluwer Academic Publishers.

Carlstein, E., H. Müller, and D. Siegmund (Eds.) (1994). *Change-point Problems*, Number 23 in IMS Monograph. Institute of Mathematical Statistics, Hayward, CA.

Cormen, T. H., C. E. Leiserson, R. L. Rivest, and C. Stein (2001). *Introduction to Algorithms.* MIT Press.

d'Aspremont, A., F. Bach, and L. El Ghaoui (2008). Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research 9*, 1269–1294.

Donoho, D. and I. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association 90*(432), 1200–1224.

Efron, B., T. Hastie, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics 32*, 407–499.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and Computing 16*, 203–213.

Fisher, W. D. (1958). On grouping for maximum homogeneity. *Journal of the American Statistical Society 53*, 789–798.

Gales, M. and S. Young (2008). The application of hidden markov models in speech recognition. *Foundations and Trends in Information Retrieval 1*(3), 195–304.

Hawkins, D. M. (2001). Fitting multiple change-point models to data. *Computational Statistics & Data Analysis 37*, 323–341.

Hesterberg, T. C., N. H. Choi, L. Meier, and C. Fraley (2008). Least angle and L1 penalized regression: A review. *Statistics Surveys 2*, 61–93.

Kay, S. M. (1993). *Fundamentals of statistical signal processing: detection theory.* Prentice-Hall, Inc.

Kolesnikov, A. and P. Fränti (2003). Reduced-search dynamic programming for approximation of polygonal curves. *Pattern Recogn. Lett. 24*(14), 2243–2254.

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

Lavielle, M. and E. Moulines (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis 21*(1), 33–59.

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

Mammen, E. and S. Van De Geer (1997). Locally adaptive regression splines. *Annals of Statistics 25*(1), 387–413.

Massart, P. (2004). A non asymptotic theory for model selection. In *Proceedings of Mathematical Foundations of Stattistical Learning, Stockholm*, pp. 309–324.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics 37*(1), 246–270.

Moghaddam, B., Y. Weiss, and S. Avidan (2006). Generalized spectral bounds for sparse lda. In *ICML*.

Ruanaidh, J. and W. Fitzgerald (1996). *Numerical Bayesian Methods Applied to Signal Processing*. Statistics and Computing. Springer.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*(1), 267–288.

Tibshirani, R. and P. Wang (2008). Saptial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics 9*(1), 18–29.

Yao, Y. and S. T. Au (1989). Least-squares estimation of a step function. *Sankhya: The Indian Journal of Statistics, Series A, 51*(3), 370–381.

Zhao, P. and B. Yu (2006). On model selection consistency of lasso. *Journal of Machine Learning Research 7*, 2541–2563.

CNRS/LTCI/TelecomParisTech - 46, rue Barrault, 75634 Paris Cédex 13, France.
E-mail address: zaid.harchaoui@telecom-paristech.fr

CNRS/LTCI/TelecomParisTech - 46, rue Barrault, 75634 Paris Cédex 13, France.
E-mail address: celine.levy-leduc@telecom-paristech.fr