Rapport de Stage de Recherche

Pour l'obtention du grade de

Master 2, Mathématiques et Applications (MM15 2)

UNIVERSITÉ PIERRE ET MARIE CURIE

Domaine de Recherche : **STATISTIQUE**

Présenté par

Mokhtar Zahdi ALAYA

Change-Points Detection with Total-Variation Penalization

Directeurs de Stage : Agathe GUILLOUX

Stéphane GAÏFFAS

Laboratoire de Statistique Théorique et Appliquée (LSTA)

Paris, Septembre 2012

Mokhtar Zahdi ALAYA

Change-Points Detection with Total-Variation Penalization

To my mother Sahara.

Acknowledgements

First and foremost, I owe innumerable thanks to my advisers Stéphane Gaïffas and Agathe Guilloux, for being great mentors, both professionally and personally. This report would never be possible without their continuous support over my training period. Many of their valuable and insightful suggestions not only encouraged me to constantly learn new things, but also taught me how to be an independent young researcher. I am in particular indebted to them for generously allowing me with enough freedom for exploring new research topics of my own interests.

I am deeply thankful for the support of my brother Mohamed and his family who made everything possible. Also I would like to give my special thanks to my mother Sahara whose love and unconditional encouragement enabled me to complete this work.

Paris, September 2012 Mokhtar Zahdi Alaya

Contents

Chapter I

Introduction

In this Chapter, we aim to give a very brief introduction to the high-dimensional problems that currently mathematicians, statisticians and data miners are trying to address. Rather than attempting to give an overview of this vast area, we will explain what is meant by high-dimensional data and then focus on some methods which have been introduced to deal with this sort of data. The approaches from these fields are often different from each other in the way of tackling high-dimensional data. However, there is one main point that reconcile theses scientific communities: something has to be done to reshape the classical approaches to better analyse high-dimensional data.

1 Challenges of High-Dimensional Modeling

In the current century, a mixture of expertise and the new technologies leads to the availability of massive amount of data. Our society invests massively in the collection and processing of data of all kinds; hyperspectral imagery, internet portals, financial tick by tick data, and DNA microarrays are just a few of the better-known sources, feeding data in torrential streams into scientific and business databases world-wide.
The trend today is towards more observations but even more larger number of variables. We are seeing examples where the data collected on individual observation are curves, or spectra, or images, or even movies, so that a single observations has dimensions in the thousands or billions, while there are only tens or hundreds of observations available to study. Classical methods cannot cope with this kind of explosive growth of the dimensionality of the observation matrix. Therefore high dimensional data analysis will be a very significant activity in the future, and completely new methods of high dimensional data analysis will be developed.

Over the last few decades, data, data management, and data processing have become ubiqui-

tous factors in modern life and work. Huge investments have been made in various data gathering and data processing mechanisms. The information technology industry is the fastest growing and most lucrative segment of the world economy, and much of the growth occurs in the development, management, and warehousing of streams of data for scientific,medical,engineering, and commercial purposes. Some recent examples include, Fan and Li (2006),:

— Biotech Data: the fantastic progress made in the last years in gathering data about the human genome have spread statistical concepts toward biological fields. This is actually just the opening round in a long series of developments. The genome is only indirectly related to protein function and protein function are only indirectly related to overall cell function. Over time, the focus is likely to switch from genomics to proteomics and beyond. In the process more and more massive databases will be compiled.

— Financial Data: over the last decade, high frequency financial data have become available; in the early to mid 1990s data on individual currency trades, became available, tracking individual transactions. After the recent economic crisis, statistical models for long and high dimension streams of data are required to better predict trembling situations.

— Consumer Financial Data: many transactions are made on the web; browsing, searching, purchasing are being recorded, correlated, compiled into databases, and sold and resold, as advertisers scramble to correlate consumer actions with pockets of demand for various goods and services.

Previous examples showed that we are in the era of massive automatic data collection, systematically obtaining many measurements, not knowing which ones will be relevant to the phenomenon of interest. Therefore, statisticians must face the problem of high dimensionality, reshaping the classical statistical thinking and data analysis.

2 High-Dimensioanl Data Analysis

Statistical estimation in high-dimensional situations, where the number of measured variables p is substantially larger than the sample size n, also known as, *large-p-small-n*, is fundamentally different from the estimation problems in the classical settings where we have *small-p-large-n*. Since high-dimensional datasets are not uncommon in modern real-world applications, such as gene expression microarray data and functional. In many real-world problems the number of covariates is very large and often statisticians have to tackle the challenge of treating data in which the number of variables p is much larger than the number of observations n, i.e

when $n \ll p$, or sometimes $p = p_n$ grows with n in the asymptotic analysis, possibly very fast, so that $n \ll p_n$ for n tends to infinity. Such high-dimensional settings with their many new scientific problems create great opportunities and significant challenges for the development of new techniques in statistics. From a classical statistical point of view, many algorithms for solving the problem of dimensional reduction and feature extraction have been conceived in order to obtain parsimonious models that are desirable as they provide simple and interpretable relations among scientific variables in addition to reducing forecasting errors. But in high-dimensional systems, we work with large size problems (from on the order of $50 - 100$ up to thousands of variables) and the space of all possible subset of variables is of the order of 2^p. Treating exhaustively all the possible subsets of models is not realistic because the study of all the sub-models is a NP-hard problem with computational time increasing exponentially with the dimensionality. Moreover, high dimensional real problems often involve costly experimentations and new techniques are needed to reduce the number of the experimental trials though guaranteeing satisfactory results. The expensive experimental and computational costs make traditional statistical procedures infeasible for high-dimensional data analysis. Generally speaking, learning salient information from relatively a few samples when many more variables are present is not possible without knowing special structures in the data.

To alleviate the ill-posed problem, it is natural to restrict our attention to subsets of all solutions with certain special structures or properties and meanwhile to incorporate the regularization ideas into estimation. Crucially, one has to assume in this setting that the data have *sparse structure*, meaning that most of the variables are irrelevant for accurate prediction. The task is hence to filter-out the relevant subset of variables. While high dimensionality of a data set is evident from the start, it is usually not easy to verify structural sparseness.
Sparsity is one commonly hypothesized condition and it seems to be realistic for many real-world applications. There has been a surge in statistical literature, which is the LASSO.

The LASSO, proposed by Tibshirani (1996), is an acronym for Least Absolute Shrinkage and Selection Operator. Among the main reasons why it has become very popular for high-dimensional estimation problems are its statistical accuracy for prediction and variable selection coupled with its computational feasibility.
The LASSO opens a new door to variable selection by using the ℓ_1-penalty in the model fitting criterion. Due to the nature of the ℓ_1-penalty, the LASSO performs continuous shrinkage and variable selection simultaneously. Thus the LASSO possesses the nice properties of both the ℓ_2-penalization (ridge) and best-subset selection. It is forcefully argued that the automatic

feature selection property makes the LASSO a better choice than the ℓ_2-penalization in high dimensional problems, especially when there are lots of redundant noise features although the ℓ_2 regularization has been widely used in various learning problems such as smoothing splines. An ℓ_1 method called basis pursuit was also used in signal processing Chen, Donoho and Saunders (2001). There are many theoretical work to prove the superiority of the ℓ_1-penalization in sparse settings. It is also shown that the ℓ_1-approach is able to discover the "right" sparse representation of the model under certain conditions (ref.).

3 Report Outlines

Now, we outline the structure of the rest of this report.

In Chapter 2, we address to present the ordinary regression methods of the linear models, more specifically, we present the least squares estimation and the ridge estimation. We further define the LASSO estimator and we study some of its theoretical properties. By the end of this chapter, we devote our study to a classical efficient algorithm, namely, least angle regression (LARS, Efron et al. (2004)) which is a great conceptual tool for understanding the behaviour of LASSO solutions.

In Chapter 3, we are based in our study to the article Harchaoui and Levy-Leduc (2010). The authors deal with the estimation of change-points in one- dimensional piecewise constant signals observed in white noise. Their approach consists in reframing the task in a variable selection context. For this purpose, they use a penalized least square criterion with a ℓ_1-type penalty.

Chapter II

LASSO-Type Estimator

The LASSO was proposed as a technique for linear regression. Linear regression is itself a specific technique of regression and this focus on techniques for computing this operator. This introductory chapter precises this hierarchy of problem with their settings, motivations and notations. Particular attention is given to the LASSO itself and algorithms for solving it, handling the ℓ_1-norm, and a generalized definition of the LASSO. We discuss in this chapter some fundamental methodological and computational aspects which addresses some bias problems of the LASSO. The methodological steps are supported by describing various theoretical results which will be fully developed.

1 Linear Regression Model

In this chapter, we consider the problem of estimating the coefficient vector in a linear regression model, defined as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}. \tag{II.1}$$

Or equivalently

$$\mathbf{Y} = \sum_{j=1}^{p} \beta_j^\star \mathbf{X}_j + \boldsymbol{\varepsilon}, \tag{II.2}$$

where we use the following notations:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_p \end{pmatrix},$$

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \text{ and } \boldsymbol{\beta}^\star = \begin{pmatrix} \beta_1^\star \\ \vdots \\ \beta_p^\star \end{pmatrix}.$$

Here $\mathbf{X}$ is the $n \times p$ **design** matrix which can either be non-stochastic or random. It is selected by the experimenter to determine its relationship to the observation. As per convention, rows of $\mathbf{X}$ represent the p-dimensional observations and columns of $\mathbf{X}$ represent the predictors. $\mathbf{Y}$ is the **observation** vector and the outcome of a statistical experiment. The coefficients Y_i are also called the endogenous variables, response variables, measured variables, or dependent variables. $\boldsymbol{\beta}^\star$ is the **target** coefficient vector to be estimated. The statistical estimation focuses on it. It represents the variables of interest. The entries of $\boldsymbol{\beta}^\star$ are the regression coefficients. We regard $\boldsymbol{\varepsilon}$ as a column vector, and use $\boldsymbol{\varepsilon}^\top$ to denote its conjugate transpose. The **noise** measurement error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_n)^\top$ captures all others factors which influence the observation. Depending on the model, $\boldsymbol{\varepsilon}$ is assumed to be *iid* according to a known distribution. Here we do not have to generally assume that the error possesses a finite second moment σ^2 for each component. This corresponds to a situation where one observes some real variables (here *variable* is taken in its physical sense, not the probabilistic one) $\mathbf{X}_1$, ..., $\mathbf{X}_p$ and $\mathbf{Y}$ at n different times or under n different circumstances. This results in n groups of values of those variables ($\mathbf{X}_1$, ..., $\mathbf{X}_p$, $\mathbf{Y}_i$) for $i \in \{1, ..., n\}$ each group corresponding to a time of observation or a particular experiment. We denote by $\mathbf{Y} = (\mathbf{Y}_i)_{1 \leq i \leq n}$ and ($\mathbf{X}_1$, ..., $\mathbf{X}_p$) the corresponding vectors. In this setting the main assumption is that the variable of interest $\mathbf{Y}$ is a linear (but otherwise unknown) function of the explanatory variables $\boldsymbol{X_1}, ..., \boldsymbol{X_p}$ plus some random perturbation. Classically, we are interested in estimation of the parameters $\beta_j^\star$ or equivalently $\mathbf{X}\boldsymbol{\beta}^\star$. As a particular case, we present an elementary but important statistical model, the Gaussian linear model. Gaussian linear regression is a statistical framework in which, the vector of noise had been distributed according to a zero-mean Gaussian distribution. It reads as

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(O, \sigma^2 Id_n),$$

where $\mathcal{N}_n$ is the n-multivariate Gaussian distribution, $Id_n \in \mathbb{R}^{n \times n}$ is the identity matrix, and σ is the standard deviation. In this case, the random vector $\boldsymbol{\varepsilon}$ is called a a Gaussian white noise.

1.1 Least Squares Estimator and Ridge Estimator

We present two popular methods to estimate the parameter $\boldsymbol{\beta}^\star$, the least squares estimator and the ridge estimator.

Least Squares Estimator

The usually method for estimating the parameter $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ is the least squares. It consists in the search of a value $\hat{\beta}$ of the parameter which minimizes the the residual sum of squares (RSS):

$$\sum_{i=1}^{n} (y_i - x_i\hat{\beta})^2 = \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - x_i\beta)^2.$$

One can write this minimization problem in a matrix form as following:

$$\|\mathbf{Y} - \mathbf{X}\hat{\beta}\|_n^2 = \min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_n^2, \tag{II.3}$$

where $\|\cdot|_2$ is the standard ℓ_2-norm given by $\|x\|_2^2 = \frac{1}{n}\sum_{i=1}^m x_i^2$, for all $x \in \mathbb{R}^m$. It is clear that there is always a solution $\hat{\beta}$ of the minimization problem II.3, namely, least squares estimator (LSE) of $\boldsymbol{\beta}^\star$ which will be noted as β^{ls}. We write

$$\hat{\beta}_{ls} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i\beta)^2 = \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_n^2.$$

If the design matrix $\mathbf{X}^\top\mathbf{X}$ is invertible then the least squares estimator has an unique solution, defined by

$$\hat{\beta}_{ls} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}. \tag{II.4}$$

It is well known that ordinary least squares often does poorly in both prediction and interpretation. Penalization techniques have been proposed to improve ordinary least squares. For example, ridge regression Hoerl and Kennard (1988), minimizes RSS subject to a bound on the ℓ_2-norm of the coefficients. As a continuous shrinkage method, ridge regression achieves its better prediction performance through a bias-variance trade-off.

Ridge Estimator

Note that the basic requirement for the Least squares estimation of a linear regression is $\mathbf{X}^\top\mathbf{X}^{-1}$ exists. There are two reasons that the inverse does not exits. First, $n \ll p$ and collinearity between the explanatory variables. The technique of ridge regression is one of the most popular and best performing alternatives to the ordinary least squares methods. A simple way to guarantee the invertibility is adding a diagonal matrix to $\mathbf{X}^\top\mathbf{X}$, i.e. $\mathbf{X}^\top\mathbf{X}+\lambda\ \mathbf{I}_p$ where $\mathbf{I}_p$ is a

$p \times p$ identity matrix. The ridge regression estimator is then

$$\hat{\beta}_r(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}. \tag{II.5}$$

where $\lambda > 0$ is a parameter needs to be chosen. The motivation of ridge regression is very simple, but it has good performance. Another way to understand it is that we dont expect an estimator with too large $\boldsymbol{\beta}^\star$. Thus, we penalize the value of $\boldsymbol{\beta}^\star$. Recall the least square estimation is to minimize

To penalize the value of $\boldsymbol{\beta}^\star$, we can consider estimate $\boldsymbol{\beta}^\star$ by minimizing

$$\hat{\beta}_r(\lambda) \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \|\beta\|_2^2 = \arg\min_{\beta \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}\beta\|_n^2 + \lambda \|\beta\|_2^2. \tag{II.6}$$

It is not difficult to prove that to solution of β to the above problem is

$$\hat{\beta}_r(\lambda) = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p)^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Note that with larger λ, the penalty on β tends to be stronger; the solution of $\boldsymbol{\beta}^\star$ will be smaller.

Variable Selection

The parameter $\boldsymbol{\beta}^\star = (\beta_1^\star, ..., \beta_p^\star)^\top$ shows the weight of the explanatory variables $\mathbf{X}_1$, ..., $\mathbf{X}_p$ over the response $\mathbf{Y}$. When the number of the explanatory variables is very important, an objectif would be evaluated the contribution of each variable and eliminated the non-pertinent variables. This typical approach gives an interpretable estimators. In this context, the least squares and ridge estimator are not efficient. It is useful to consider some competent methods to select the subset of the explanatory variables, affording an almost complete representation of the response variable $\mathbf{Y}$. Therefore, diverse strategies are been proposed for achieving the determination of the pertinent variables. Some classical approach is *Subset Selection*. Let B_k a subset of explanatory variables of size k which reduces the maximum of RSS (ref.).

Another strategy for the variable selection is the *thresholding*. In this case, we use a preliminary estimator (e.g. the LSE when $p \leq n$), which we exploit it to exclude some variables from the study. A variable will be selected only when the estimation of the corresponding regressor coefficient, obtained by the preliminary estimator, exceeds some threshold defined by the statistician. As an example, we can consider the *soft thresholding* and the *hard thresholding*

(ref.)

To reduce the number of explanatory variables, diverse tests based on the LSE are been proposed for testing the relevance of each variable X_j. For all $j \in \{1, ..., p\}$, these procedures test under the null hypothesis $\beta_j^\star = 0$ and the alternative hypothesis $\beta_j^\star \neq 0$. Frequently, when the noise is gaussian, someone uses the Student test or Fisher test.

1.2 Penalized Least Squares and Sparsity

Let A an arbitrary set, we note by $|A|$ the cardinal of A. For the study of the method of variable selection, it is convenient to define the sparsity set as the following:

Definition 1.1 *Let the model defined by II.2. One can define the* ***support set*** *associated to the vector* $\boldsymbol{\beta}^\star$ *by*

$$S^\star = S^\star(\beta^\star) := \{j \in \{1, ..., p\} : \beta_j^\star \neq 0\}. \tag{II.7}$$

Thereafter, we call that the vector $\boldsymbol{\beta}^\star$ *has the* ***sparsity assumption*** *if the quantity* $|S^\star| \ll p$.

The construction of interpretable estimators is an important issue. Some of them are obtained from the ℓ_0-penalization such that the Information Criterion C_p of Mallows, Akaike Information Criterion (AIK) or the Bayesian Information Criterion (BIC). These criterions select from a collection of size D estimators of $\beta^\star$,

$$\hat{F} = \{\hat{\beta}_1, ..., \hat{\beta}_D\},$$

whose has the good estimation of $\mathbf{X}\boldsymbol{\beta}^\star$ and the good estimation to the set of the pertinent variables $S^\star$ defined in II.8. Clearly, one can understand the important of the choice of this family $\hat{F}$. Moreover, these criterions are constructed from the penality $\lambda\|\beta\|_0$ which interferes the ℓ_0-norm of the vector β, defined by

$$\|\beta\|_0 := \sum_{j=1}^{p} \mathbb{1}_{\{\beta_j \neq 0\}},$$

$\mathbb{1}_{\{\cdot\}}$ denotes the indicator function

Unfortunately, the ℓ_0-minimization problems are known to be NP-hard in general, so that the existence of polynomial-time algorithms is highly unlikely. This challenge motivates the use of computationally tractable approximations or relaxations to ℓ_0 minimization.In particular, a

great deal of research over the past decade has studied the use of the ℓ_1-norm as a computationally tractable surrogate to the ℓ_0-norm. The LASSO for linear models is the core example to develop the methodology for ℓ_1-penalization in high-dimensional settings. Moreover, it is a penalized least squares method imposing a ℓ_1-penalty on the regression coefficients. Due to the nature of the ℓ_1-penalty, the LASSO does both continuous shrinkage and automatic variable selection simultaneously.

2 LASSO Estimator

2.1 Definition

Definition 2.1 *The LASSO estimator of $\boldsymbol{\beta}^\star \in \mathbb{R}^p$ is defined as*

$$\hat{\beta}_{lasso} = \hat{\beta}_{lasso}(\lambda) := arg\min_{\beta\in\mathbb{R}^p} \Big\{ \frac{1}{2} \| \boldsymbol{Y} - \boldsymbol{X}\beta \|_n^2 + \lambda \|\beta\|_1 \Big\}, \tag{II.8}$$

where the $\|\beta\|_1 := \sum_{j=1}^{p} |\beta_j|$ is the ℓ_1-norm.

The parameter λ can be depended to the number of observation n, i.e. $\lambda \equiv \lambda_n$. Also, $\lambda \geq 0$ is a shrinkage tuning parameter. A larger λ yields a sparser linear sub-model whereas a smaller λ corresponds to a less-sparse one. In extreme cases, $\lambda = 0$ gives the unregularized model and $\lambda = \infty$ produces the null model consisting of no predictor.

Equivalently, the convex program *II*.8 can be reformulated as the ℓ_1-constrained quadratic problem as following:

$$\begin{cases} \min_{\beta\in\mathbb{R}^p} \Big\{ \frac{1}{2} \|\mathbf{Y} - \mathbf{X}\beta\|_n^2 \Big\} \\ s.t. \ \|\beta\|_1 \leq t \end{cases} \tag{II.9}$$

for some $t > 0$. If t is greater than or equal to the ℓ_1-norm of the ordinary least squares estimator, then that estimator is, of course, unchanged by the LASSO. For smaller values of t, the LASSO shrinks the estimated coefficient vector towards the origin (in the ℓ_1 sense), typically setting some of the coefficients equal to zero. Thus, the LASSO combines characteristics of ridge regression and subset selection and promises to be a useful tool for variable selection.
Problems *II*.8 and *II*.9 are equivalent; that is, for a given λ, $0 < \lambda < \infty$, there exists a $t > 0$ such that the two problems share the same solution, and vice versa. Optimization problems like *II*.9 are usually referred to as constrained regression problems while *II*.8 would be called

a penalized regression.

Under a few assumptions, which are detailed in the sequel, the solution of this problem is unique. We denote it by$\hat{\beta}_{lasso} \equiv \hat{\beta}_{lasso}(\lambda)$ and define the regularization path $\mathbf{P}$ as the set of all solutions for all positive values of λ

$$\mathbf{P} := \{\hat{\beta}_{lasso}(\lambda) : \lambda > 0\}. \tag{II.10}$$

The following proposition presents classical optimality and uniqueness conditions for the Lasso solution, which are useful to characterize $\mathbf{P}$:

2.2 Convex Optimality and Uniqueness

We begin with some basic observations about the LASSO problem *II*.8. First, the minimum in the Lasso is always achieved by at least one vector . This fact follows from the Weierstrass theorem, because in its ℓ_1-constrained form *II*.9, the minimization is over a compact set, and the objective function is continuous. Second, although the problem is always convex, it is not always strictly convex, so that the optimum can fail to be unique. Indeed, a little calculation shows that the Hessian of the quadratic component of the objective is the $p \times p$ $\frac{X^\top X}{n}$ matrix , which is positive de?nite but not strictly so whenever . Nonetheless, as stated below in the Lemma 1, strict dual feasibility conditions are sufficient to ensure uniqueness, even under high-dimensional scaling $n \ll p$.

The objective function is not always differentiable, since the ℓ_1-norm is a piecewise linear function. However, the optima of the Lasso *II*.8 can be characterized by a zero subgradient condition. A vector is $w \in \mathbb{R}^p$ a subgradient for the ℓ_1-norm evaluated at $\beta \in \mathbb{R}^p$, written as $w \in \partial\|\beta\|_1$, if its elements satisfy the relations

$$\begin{cases} w_j = sign(\beta_j), & if\ \beta_j \neq 0 \\ w_j \in [-1, +1], & otherwise \end{cases} \tag{II.11}$$

For any subset $A \in \{1, ..., p\}$, let X_A be the $n \times |A|$ matrix formed by by concatenating the columns $\{X_j : j \in A\}$ indexed by A With these definitions, we state the following.

Lemma 2.1 ***(Karush Kuhn Tucker(KKT) Optimality Conditions)***

A vector $\hat{\beta} \in \mathbb{R}^p$ *is a solution of II.8 if and only if for all* $j \in \{1, ..., p\}$,

$$\begin{cases} \boldsymbol{X}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) = \lambda sign(\hat{\beta}_j), & if\ \hat{\beta}_j \neq 0 \\ |\boldsymbol{X}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})| \leq \lambda, & otherwise. \end{cases} \tag{II.12}$$

Define

$$\hat{S} := \{j \in \{1, ..., p\} : |\boldsymbol{X}_j^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta})| = \lambda\}.$$

Assuming the matrix $\boldsymbol{X}_{\hat{S}}$ *to be full rank, the solution is unique and we have*

$$\hat{\beta} = (\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{X}_{\hat{S}})^{-1} (\boldsymbol{X}_{\hat{S}}^\top \boldsymbol{Y} - z_{\hat{S}}), \tag{II.13}$$

where $z_{\hat{S}} = sign(\boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}))$ *is in* $\{-1; 0; +1\}^p$, *and the notation* $u_{\hat{S}}$ *for a vector* u *denotes the vector of size* $|\hat{S}|$ *recording the entries of* u *indexed by* $\hat{S}$.

Proof. The propertie *II.12* can be obtained by considering subgradient optimality conditions. These can be written as $0 \in \{(-X^\top (Y - X\hat{\beta} + \lambda w : w \in \partial \|\hat{\beta}\|))\}$. The equalities in *II.12* define a linear system that has a unique solution given by *II.13* when $\mathrm{X}\hat{S}$ is full rank.
Let us now show the uniqueness of the Lasso solution. Consider another solution $\hat{\beta}'$ and choose a scalar α in $(0, 1)$. By convexity, $\hat{\beta}^\alpha := \alpha\hat{\beta} + (1-\alpha)\hat{\beta}'$ is also a solution. for all $j \subseteq \hat{S}$, we have

$$|X_j^\top (Y - X\hat{\beta}^\alpha)| \leq \alpha |X_j^\top (Y - X\hat{\beta})| + (1-\alpha)|X_j^\top (Y - X\hat{\beta}')| < \lambda.$$

Combining this inequality with the conditions *II.12* we necessarily have $\hat{\beta}^\alpha_{\hat{S}^c} = \hat{\beta}_{\hat{S}^c} = 0$, and the vector $\hat{\beta}^\alpha_{\hat{S}^c}$ is also a solution of the following reduced problem:

$$\min_{\widetilde{\beta} \in \mathbb{R}^{|\hat{S}|}} \{\frac{1}{2}\|Y - X\widetilde{\beta}\|_n^2 + \lambda \|\widetilde{\beta}\|_1\}.$$

When $\mathrm{X}_{\hat{S}}$ is full rank, the Hessian $\mathrm{X}_{\hat{S}}^\top \mathrm{X}_{\hat{S}}$ is positive definite and this reduced problem is strictly convex. Thus, it admits a unique solution $\hat{\beta}^\alpha_{\hat{S}} = \hat{\beta}_{\hat{S}}$ It is then easy to conclude that $\hat{\beta}_{\hat{S}} = \hat{\beta}^\alpha_{\hat{S}} = \hat{\beta}'_{\hat{S}}$. ∎

Lemma 2.2 ***(Piecewise Linearity of the Path)****. Assume that for any* $\lambda > 0$ *and solution of II.8 the matrix* $X_{\hat{S}}$ *defined in Lemma 2.1 is full-rank. Then, the regularization path* $\boldsymbol{P} := \{\hat{\beta}_{lasso}(\lambda) : \lambda > 0\}$ *is well defined, unique and continuous piecewise linear.*

Proof. The existence/uniqueness of the regularization path was shown in Lemma 2.1. Let us define $\{\hat{z}(\lambda) := sign(\hat{\beta}(\lambda)) : \lambda > 0\}$ the set of sparsity patterns. Let us now consider

$\lambda_1 < \lambda_2$ such that $\hat{z}(\lambda_1) = \hat{z}(\lambda_2)$. For all θ in $[0,1]$, it is easy to see that the solution $\hat{\beta}^\theta := \alpha\hat{\beta}(\lambda_1) + (1-\theta)\hat{\beta}(\lambda_2)$ satisfies the optimality conditions of Lemma 2.1 for $\lambda = \theta\lambda_1 + (1-\theta)\lambda_2$, and that $\hat{\beta}(\theta\lambda_1 + (1-\theta)\lambda_2) = \hat{\beta}^\theta$.
This shows that whenever two solutions $\hat{\beta}(\lambda_1)$ and $\hat{\beta}(\lambda_2)$ have the same signs for $\lambda_1 \neq \lambda_2$, the regularization path between λ_1 and λ_2 is a linear segment. As an important consequence, the number of linear segments of the path is smaller than 3^p, the number of possible sparsity patterns in $\{-1,0,1\}^p$. The path P is therefore piecewise linear with a finite number of kinks. Moreover, since the function $\lambda \to \hat{\beta}(\lambda)$ is piecewise linear, it is piecewise continuous and has right and left limits for every $\lambda > 0$. It is easy to show that these limits satisfy the optimality conditions of the propretie *II*.12. By uniqueness of the LASSO solution, they are equal to $\hat{\beta}$ and the function is in fact continuous. ■

In the next section we discuss some theoretical properties of LASSO.

2.3 Theoretical Results of the LASSO: A Brief of View

We begin by some definitions. we assume in our regression setting that the vector β is sparse in the ℓ_0-sense and many coefficients of β are identically zero.The corresponding variables have thus no influence on the response variable and could be safely removed. The sparsity pattern of ß is understood to be the *sign* function of its entries,

$$sign(x) = \begin{cases} +1, & if x > 0 \\ 0, & if x = 0 \\ -1, & if x < 0 \end{cases}$$

The sparsity pattern of a vector might thus look like

$$sign(\beta) = (+1, -1, 0, 0, +1, +1, -1, +1, 0, 0, ...),$$

distinguishing whether variables have a positive, negative or no influence at all on the response variable. It is of interest whether the sparsity pattern of the LASSO estimator is a good approximation to the true sparsity pattern. If these sparsity patterns agree asymptotically, the estimator is said to be *sign consistent*.

Definition 2.2 ***(Sign Consistency)***
An estimator $\hat{\beta}$ is sign consistent if and only if

$$\mathbb{P}\Big(sign(\hat{\beta}) = sign(\beta^\star)\Big) \to 1, \qquad as\, n \to \infty.$$

Asymptotic properties of the LASSO estimator have been extensively studied and analyzed. In a seminal work (ref.), Knight and Fu, first derived the asymptotic distribution of the LASSO estimator and proved its estimation consistency under the shrinkage rate $\lambda_n = o(\sqrt{n})$ and $\lambda_n = o(n)$. More specifically, as long as errors are *iid* and possess a common finite second moment σ^2 , the $\sqrt{n}$ scaled LASSO estimator with a sequence of properly tuned shrinkage parameters $\{\lambda_n\}_{n\in N}$ has an asymptotic normal distribution with variance $\sigma^2 C^{-1}$, where $\frac{1}{n}X^\top$X $\to C$ and C is a positive definite matrix.
Zhao and Yu (2006) found a sufficient and necessary condition required on the design matrix for the LASSO estimator to be model selection consistent, i.e. the *irrepresentable condition*.

Definition 2.3 ***(Irrepresentable condition)***
Let $S^\star$ the support set of $\beta^\star$ also it is the set of the relevant variables and let $S^{\star c} = \{1, ..., p\} - \beta^\star$ be the set of noise variables. The sub matrix C_{UV} is understood as the matrix obtained from C by keeping rows with index in the set U and columns with index in V. The irrepresentable condition is fulfilled if

$$\|C_{S^{\star c}S^\star} {C_{S^\star S^\star}}^{-1} (sign(\beta_{S^\star})\|_{\ell_\infty} < 1.$$

These conditions are in general not easy to verify. Therefore, instead of requiring conditions on the design matrix for model selection consistency, there are also several variants of the original LASSO. For examples, the relaxed LASSO, Meinshausen (2007), uses two parameters to separately control the model shrinkage and selection; the adaptive LASSO, Zou (2006), leverages a simple adaptation procedure to shrink the irrelevant predictors to 0 while keeping the relevant ones properly estimated. Meinshausen and Yu (2009) suggested employ a two-stage hard thresholding rule, in the spirit of the Gauss-Dantzig selector, Candès and Tao (2007), to set very small coefficients to 0.

Since the ground breaking work of Candès and Tao (2007) which provided non-asymptotic upper bounds on the ℓ_2- estimation loss of the Dantzig selector with large probability, parallel ℓ_2 error bounds were found for the LASSO estimator by Meinshausen and Yu (2009) under the incoherent design condition and by Bickel, Ritov, and Tsybakov (2009) under the restricted eigenvalue condition. In a previous work of Candès and Tao (2007), they showed that minimizing the ell_1-norm of the coefficient vector subject to the linear system constraint can exactly recover the sparse patterns, provided the restricted isometry condition holds and the support of the noise vector is not too large Candès and Tao (2005).
Cai, Xu, and Zhang (2009) tightened all previous error bounds for noiseless, bounded error and Gaussian noise cases. These bounds are nearly optimal in the sense that they achieve within a

logarithmic factor the least squares errors as if the true model were known (oracle property). Wainwright (2006) derived a set of sharp constraints on the dimensionality, sparsity of the model and the number of observations for the Lasso to correctly recover the true sparsity pattern. The ℓ_∞ convergence rate of the LASSO estimator was obtained by Lounici (2008). Other bounds for the sparsity oracle inequalities of the Lasso can be found in Bunea, Tsybakov, Wegkamp (2007).

Despite those appealing properties of the Lasso estimator and the advocacy of using the LASSO, the LASSO estimate is not guaranteed to provide a satisfactory estimation and detection performance, at least in some application scenarios. For instance, when the data are corrupted by some outliers or the noise is extremely heavy-tailed, the variance of the LASSO estimator can be quite large, usually become unacceptably large, even when the sample size approaches infinity, Knight and Fu (2000). Asymptotic analysis, Knight and Fu (2000), and non-asymptotic error bounds on the estimation loss, Bickel, Ritov, and Tsybakov (2009), both suggest that the performance of the LASSO linearly deteriorates with the increment of the noise power. A similar observation can sometimes be noted when the dimensionality of the linear model is very high while the data size is much smaller.

3 Least Angle Regression (LARS)

Least Angle Regression is a promising technique for variable selection applications, offering a nice alternative to stepwise regression. It provides an explanation for the similar behavior of LASSO (ℓ_1-penalized regression) and forward stagewise regression, and provides a fast implementation of both. The idea has caught on rapidly, and sparked a great deal of research interest. We write LAR for least angle regression, and LARS to include LAR as well as LASSO or forward stagewise as implemented by least-angle methods .In the sequel, we give the algorithm of Least Angle Regression . The LARS algorithm was proposed (and named) by Efron et al. (2004), though essentially the same idea appeared earlier in the works of Osborne et al. (2000).

3.1 Description of the Algorithm

The algorithm begins at $\lambda = \infty$, where the lasso solution is trivially $0 \in \mathbb{R}^p$. Then,as the parameter λ decreases, it computes a solution path $\hat{\beta}_{lars}(\lambda)$ that is piecewise linear and continuous as a function of λ. Each knot in this path corresponds to an iteration of the algorithm, in which the path s linear trajectory is altered in order to satisfy the KKT optimality

conditions.
The LARS algorithm recursively calculates a sequence of breakpoints $\infty = \lambda_0 > \lambda_1 > \lambda_2 >$ ůůů $= 0$ with $\hat{\beta}(\lambda)$ linear for each interval $\lambda_{k+1} \leq \lambda \leq \lambda_k$. The *active set* $\hat{S}$ of the coefficients changes, the incactive coefficients stay fixed at zero. Define the residual vector and correlations

$$R(\lambda) := Y - X\hat{\beta}(\lambda) \qquad \textit{and} \qquad C_j(\lambda) := X_j^\top R(\lambda).$$

To get a true correlation we would have to divide by $\|R(\lambda)\|$, which would complicate the constraints.
The algorithm will ensure that

$$\begin{cases} C_j(\lambda) = +\lambda & \text{if } \hat{\beta}_j(\lambda) > 0 \quad (\text{constraint } \oplus) \\ C_j(\lambda) = -\lambda & \text{if } \hat{\beta}_j(\lambda) > 0 \quad (\text{constraint } \ominus) \\ |C_j(\lambda)| < \lambda & \text{if } \hat{\beta}_j(\lambda) > 0 \quad (\text{constraint } \odot) \end{cases}$$

That is, for the minimizing $\hat{\beta}(\lambda)$ each $(\lambda, C_j(\lambda))$ needs to stay inside the region $\mathcal{R} := \{(\lambda, c) \in \mathbb{R}^+ \times \mathbb{R} : |c| \leq \lambda\}$, moving along the top boundary ($c = +\lambda$) when $\hat{\beta}_j(\lambda) > 0$ (constraint $\oplus$) along the lower boundary ($c = -\lambda$) when $\hat{\beta}_j(\lambda) < 0$ (constraint $\ominus$), and being any where in $\mathcal{R}$ when $\hat{\beta}_j(\lambda) = 0$ (constraint $\odot$).

3.2 The Algorithm

The solution $\hat{\beta}(\lambda)$ is to be constructed in a sequence of steps, starting with large λ and working towards $\lambda = 0$.

Step 1:
Start with $\hat{S}_0 = \emptyset$ and $\hat{\beta} = 0 \in \mathbb{R}^p$. Define $\lambda_1 = \max_{1 \leq j \leq p} |X_j^\top Y|$. Constraint $\odot$ is satisfied on $[\lambda_1, \infty)$. For $\lambda \geq \lambda_1$ take $\hat{\beta}(\lambda) = 0$, so that $|C_j(\lambda)| < \lambda_1$. Constraint $\odot$ would be violated if we kept $\hat{\beta}(\lambda)$ equal to zero for $\lambda < \lambda_1$; the $\hat{\beta}(\lambda)$ must move away from zero as λ decreases below λ_1.

We must have $|C_j(\lambda_1)| = \lambda_1$ for at least one j. For convenience of exposition, suppose that $|C_1(\lambda_1)| = \lambda_1 > |C_j(\lambda_1)|$ for all $j \geq 2$. The active set becomes now $\hat{S} = 1$.

For $\lambda_2 \leq \lambda < \lambda_1$, with λ_2 to be specified soon, keep $\hat{\beta}_j = 0$ for $j = 2$ but let

$$\hat{\beta}_1(\lambda) = 0 + v_1(\lambda_1 - \lambda),$$

for some constant v_1. To maintain the equalities

$$\lambda = C_1(\lambda) = X_1^\top (Y - X_1\hat{\beta}_1(\lambda)) = C_1(\lambda_1) - X_1^\top X_1 v_1(\lambda_1 - \lambda) = \lambda_1 - v - 1(\lambda_1 - \lambda)$$

we need $v_1 = 1$. This choice also ensures that $\hat{\beta}_1(\lambda) > 0$ for a while, so that Constraint $\oplus$ is the relevant constraint for $\hat{\beta}_1(\lambda)$.
For $\lambda < \lambda_1$, with $v_1 = 1$ we have $R(\lambda) = Y - X_1(\lambda_1 - \lambda)$ and

$$C - j(\lambda) = C - j(\lambda - 1) - a - j(\lambda_1 - \lambda) \quad where \quad a_j := X_j X_1.$$

Notice that $|a_j| < 1$ unless $X_j = \pm X_1$. Also, as long asmax$_{j\leq 2} |C_j(\lambda)| \leq \lambda$ the other $\hat{\beta}_j(\lambda)$ is s still satisfy constraint $\odot$.
We need to end the first step at λ_2, the largest λ less than λ_1 for which $\max_{j\geq 2} |Cj(\lambda)| = \lambda$. Solve for $C_j(\lambda) = \pm\lambda$ for each fixed $j \leq 2$:

$$\lambda = \lambda_1 - (\lambda_1 - \lambda) = C - j(\lambda_1) - a - j(\lambda_1 - \lambda) - \lambda = -\lambda_1 + (\lambda_1 - \lambda) = C - j(\lambda_1) - a_j(\lambda_1 - \lambda)$$

if and only if

$$\lambda_1 - \lambda = (\lambda_1 - C_j(\lambda_1))/(1 - a - j)\lambda_1 - \lambda = (\lambda_1 + C_j(\lambda_1))/(1 + a_j).$$

Both right-hand sides are strictly positive. Thus $\lambda_2 = \lambda_1 - \delta\lambda$ where

$$\delta\lambda := \min_{j\geq 2} \left\{ \frac{\lambda_1 - C_j(\lambda_1))}{1 - a_j} \wedge \frac{\lambda_1 + C_j(\lambda_1))}{1 + a_j} \right\}.$$

Step 2:
We have $C_1(\lambda_2) = \lambda_2 = \max_{j\geq 2} |C_j(\lambda_2)|$, by construction. For convenience of exposition, suppose $|C_2(\lambda_2)| = \lambda_2 > |C_j(\lambda_2)|$ for all $j \geq 3$. The active set now becomes $\hat{S} = \{1, 2\}$.

For $\lambda_3 \leq \lambda < \lambda_2$ and a new v_1 and v_2, define

$$\hat{\beta}_1(\lambda) = \hat{\beta}_1(\lambda_2) + (\lambda_2 - \lambda)v_1$$

$$\hat{\beta}_2(\lambda) = 0 + (\lambda_2 - \lambda)v - 2$$

with all other $\hat{\beta}_j(\lambda)$ still zero. Write Z for $(X1, X2)$. The new C_j become

$$C_j(\lambda) = \mathbf{X}_j^\top \Big(\mathbf{Y} - \mathbf{X}_1 \hat{\beta}_1(\lambda) - \mathbf{X}_2 \hat{\beta}_2(\lambda) \Big) = C_j(\lambda_2) - (\lambda_2 - \lambda) \mathbf{X}_j^\top Z v',$$

where $v' = (v1, v2)$.

Let λ_3 be the largest λ less than λ_2 for which $\max_{j \geq 3} |C_j(\lambda)| = \lambda$.

General Step:

At each λ_k a new active set $\hat{S}_k$ is defined. During the kth step the parameter λ decreases from λ_k to λ_{k+1}. For all j in the active set $\hat{S}_k$, the coefficients $\hat{\beta}_j(\lambda)$ change linearly and the $C_j(\lambda)$ move along one of the boundaries of the feasible region: $C_j(\lambda) = \lambda$ if $\hat{\beta}_j(\lambda) > 0$ and $C_j(\lambda) = -\lambda$ if $\hat{\beta}_j(\lambda) < 0$. For each inactive j the coefficient $\hat{\beta}_j(\lambda) > 0$ remains zero throughout $[\lambda_{k+1}, \lambda_k]$. Step k ends when either an inactive $C_j(\lambda)$ hits a $\pm\lambda$ boundary or if an active $\hat{\beta}_j(\lambda)$ becomes zero: λ_{k+1} is defined as the largest λ less than λ_k for which either of these conditions holds:

— (i) $\max_{j \notin \hat{S}_k} |C_j(\lambda)| = \lambda$. In that case add the new $j \in \hat{S}_k^c$ for which $|C_j(\lambda_{k+1})| = \lambda_{k+1}$ to the active set, then proceed to step $k + 1$.

— (ii) $\hat{\beta}_j(\lambda) = 0$ for some $j \in \hat{S}_k$. In that case, remove j from the active set, then proceed to step $k + 1$.

Two basic properties of the LARS lASSO path, as mentioned in the previous section, are piecewise linearity and continuity with respect to λ. The algorithm and the solutions along its computed path possess a few other nice properties. We begin with a property of the LARS algorithm itself.

Lemma 3.1 *For any* $\boldsymbol{Y}$, $\boldsymbol{X}$, *the LARS algorithm for the lasso path performs at most*

$$\sum_{k=0}^{p} (\begin{pmatrix} p \\ k \end{pmatrix} 2^k = 3^p$$

iterations before termination.

Lemma 3.2 *For any* $\boldsymbol{Y}$, $\boldsymbol{X}$, *the LARS lASSO solution converges to a minimum* ℓ_1*-norm least squares solution as* $\lambda \to 0^+$, *that is,*

$$\lim_{\lambda \to 0^+} \hat{\beta}_{lars}(\lambda) = \hat{\beta}_{ls,\ell_1}$$

, *where* $\hat{\beta}_{ls,\ell_1} \in arg\min_{\beta \in \mathbb{R}^p} \| \boldsymbol{Y} - \boldsymbol{X}\beta \|_2^2$ *and achieves the minimum* ℓ_1 *norm over all such solutions.*

The proofs of this too lemmas can be found in Tibshirani(2012).

Remark LARS has considerable promise, offering speed, interpretability, relatively stable predictions, nearly unbiased inferences, and a nice graphical presentation of coefficient paths. But considerable work is required in order to realize this promise in practice. A number of different approaches have been suggested, both for linear and nonlinear models; further study is needed to determine their advantages and drawbacks. Also various implementations of some of the approaches have been proposed that di?er in speed, numerical stability, and accuracy; these also need further assessment.

Chapter III

Multiple Change-Point Estimation with Total-Variation Penalization

In this chapter, our study will based on the article of Harchaoui and Levy (2010). Change-points detection tasks are pervasive in various fields. The goal is to partition a signal into several homogeneous segments of variable durations, in which some quantity remains approximately constant over time. The authors 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. Their approach consists in reframing this task in a variable selection context. They use a penalized least-squares criterion with a ℓ_1-type penalty for this purpose. They prove some theoretical results on the estimated change-points and on the underlying piecewise constant estimated function. Then, they explain how to implement this method in practice by using the LARS algorithm.

1 Estimation of the Means

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

$$\begin{cases} Y_t = \mu_k^\star + \varepsilon_t, \\ t_{k-1}^\star \leq t \leq t_k^\star - 1, \\ k = 1, ..., K^\star + 1, \\ t = 1, ..., n, \end{cases} \tag{III.1}$$

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 iid zero-mean random variables, having a sub-Gaussian distribution.

We consider here the multiple changes in the mean problem as described in $III.1$. Our purpose

is to estimate the unknown means $\mu_1^\star, ..., \mu_{K+1}^\star$ together with the change points from observations $Y_1, ..., Y_n$. Let us first work with the LASSO formulation to establish the consistency in terms of means estimation. The model *III*.1 can be rewritten as

$$Y^n = X_n\beta^n + \varepsilon^n, \tag{III.2}$$

where $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ is the $n \times 1$ vector of observations, X_n the $n \times n$ lower triangular matrix with nonzero elements equal to one, i.e.

$$X_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \cdots & 1 \end{pmatrix}$$

and $\varepsilon^n = \begin{pmatrix} \varepsilon_1^n \\ \vdots \\ \varepsilon_n^n \end{pmatrix}$ is a zero mean random vector such that the en $\varepsilon_1^n, ..., \varepsilon_n^n$ are *iid* 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 S the set of nonzero components of β_n also the support set of β_n and by its complementary set defined as follows:

$$S = \{k : \beta_k^n \neq 0\} \qquad \text{and} \qquad S^c := 1, ..., n - S. \tag{III.3}$$

With the reformulation *III*.2, the evaluation of the means estimation rate amounts to finding the rate of convergence of $\|X_n(\hat{\beta}^n(\lambda_n) - \beta^n)\|$ to zero, $\hat{\beta}^n(\lambda_n)$ satisfying:

$$\hat{\beta}^n(\lambda_n) = \begin{pmatrix} \hat{\beta}_1{}^n(\lambda_n) \\ \vdots \\ \hat{\beta}_n{}^n(\lambda_n) \end{pmatrix} = \arg\min_{\beta \in \mathbb{R}^n} \{\|Y^n - X_n\beta\|_n^2 + \lambda_n\|\beta\|_1\}. \tag{III.4}$$

Hence, within this framework, we are able to prove the following result regarding the consistency in means estimation of least square-total variation.

Proposition 1.1 *Consider $Y_1, ..., Y_n$ a set of observations following the model described in*

III.2. Assume that the $\varepsilon_1^n, ..., \varepsilon_n^n$ are iid Gaussian random variables with the variance $\sigma^2 > 0$. Assume also that there exists β_{max} such that for all k in A, $|\beta_k^n| \leq \beta_{max}$ the set A being defined in III.3. Then, for all $n \geq 1$ and $C > 2\sqrt{2}$, we obtain that with a probability larger than $1 - n^{1-\frac{C^2}{8}}$, if $\lambda_n = C\sigma\sqrt{\frac{\log n}{n}}$,

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

Proof. By the definition of $\hat{\beta}^n(\lambda_n)$ given by *III.4*, we have

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

Using *III.2*, we get

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

Therefore,

$$\|X_n(\beta^n - \hat{\beta}^n(\lambda_n))\|_n^2 \leq \frac{2}{n}(\hat{\beta}^n(\lambda_n) - \beta^n)^\top X_n^\top\varepsilon^n + \lambda_n\sum_{j\in S}(|\beta_j^n| - |\hat{\beta}_j^n(\lambda_n)|) - \lambda_n\sum_{j\in S^c}\hat{\beta}_j^n(\lambda_n).$$

Observe that

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

Let us define the event

$$\mathcal{E} := \bigcap_{ju1}^n\Big\{\frac{1}{n}\Big|\sum_{i=j}^n\varepsilon_i^n\Big| \leq \frac{\lambda_n}{2}\Big\}.$$

Then, given that the $\varepsilon_1^n, ..., \varepsilon_n^n$ are *iid* zero mean Gaussian variables with finite variance equal to σ^2, we obtain that

$$\begin{aligned}\mathbb{P}(\mathcal{E}^c) &= \sum_{j=1}^n P\Big(\frac{1}{n}\Big|\sum_{i=j}^n\varepsilon_i^n\Big| > \frac{\lambda_n}{2}\Big)\\ &\leq \sum_{i=j}^n\exp\Big(-\frac{n^2\lambda_n^2}{8\sigma^2(n-j+1)}\Big).\end{aligned}$$

Hence, if $\lambda_n = C\sigma\sqrt{\frac{\log n}{n}}$,

$$\mathbb{P}(\mathcal{E}^c) \leq n^{1-\frac{C^2}{8}}.$$

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

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

where S and S^c are defined in *II.4*. Given that

$$\sum_{j=1}^{n} |\hat{\beta}_j^n(\lambda_n) - \beta_j^n| = \sum_{j \in S} |\hat{\beta}_j^n(\lambda_n) - \beta_j^n| - \sum_{j \in S^c} \hat{\beta}_j^n(\lambda_n),$$

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

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

Which gives the desired result. ∎

Note that in Proposition 1.1, 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 $\frac{1}{\sqrt{n}}$ 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 least squares total variation (LS-TV) instead of its LASSO counterpart, and write model *III.1* as

$$\begin{cases} Y_t = u_t^\star + \varepsilon_t, \\ u_t^\star = \mu_k^\star,\ t_{k-1}^\star \leq t \leq t_k^\star - 1, \\ k = 1, ..., K^\star + 1, \\ t = 1, ..., n, \end{cases} \tag{III.5}$$

The vector $u^\star(\lambda_n) = \begin{pmatrix} u_1^\star(\lambda_n) \\ \vdots \\ u_n^\star(\lambda_n) \end{pmatrix}$ can be estimated by using a criteria based on total variation penalty as following:

$$\hat{u}(\lambda_n) = \begin{pmatrix} \hat{u}_1(\lambda_n) \\ \vdots \\ \hat{u}_n(\lambda_n) \end{pmatrix} = \arg\min_{u\in\mathbb{R}^n} \left\{ ||Y^n - u||_n^2 + \lambda_n \sum_{i=1}^{n-1} |u_{i+1} - u_i| \right\} \tag{III.6}$$

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 1.2 *Consider $Y_1, ..., Y_n$ a set of observations following the model described in III.5 where the $\varepsilon_1^n, ..., \varepsilon_n^n$ are iid zero mean Gaussian variables with finite variance equal to $\sigma^2 > 0$. Assume also that $\hat{u}(\lambda_n)$ defined in III.6 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 $\lambda_n = \sigma(A\frac{\sqrt{B}}{2}(K_{max}\log n)^{\frac{1}{2}} n^{-\frac{3}{2}} - \sigma(2K_{max}+1)^{\frac{1}{2}} n^{-\frac{3}{2}}$,*

$$\mathbb{P}\left(\|\hat{u} - u^\star\|_n \geq \sigma(BK_{max}\frac{\log n}{n})^{\frac{1}{2}} \right) \leq K_{max} n^{\{1 - \frac{B(1-A)^2}{8}\}K_{max}}. \tag{III.7}$$

Proof. 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 *III.6*, we get:

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

Using Model *III.5*, the previous inequality can be rewritten as follows:

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

Using the Cauchy Schwarz inequality, we obtain

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

Thus, defining $\dot{G()}$ fro $v \in \mathbb{R}^n$ by

$$G(v) := \frac{\left(\sum_{i=1} \varepsilon_i(v_i - u_i^\star)\right)}{\sigma\sqrt{n}\|v - u^\star\|_n}.$$

We have

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

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

$$\begin{aligned}\mathbb{P}\Big(\|\hat{u} - u^\star\|_n \geq \alpha_n\Big) &\leq P\Big(n\lambda_n + \sigma n^{-\frac{1}{2}} G(\hat{u}) \geq \frac{\alpha_n}{2}\Big) \\ &\leq \sum_{K=1}^{K_{max}} n^K \mathbb{P}\Big(\sup_{v\in S_K} G(v) \geq n^{\frac{1}{2}}\sigma^{-1}\frac{\alpha_n}{2} - n^{\frac{3}{2}}\sigma^{-1}\lambda_n\Big). \qquad \text{(III.8)}\end{aligned}$$

Using that, $Var(G(v)) = 1$, for all v in $\mathbb{R}^n$, we obtain by using an inequality due to Cirelson,Ibragimov,and Sudakov in the same way as in the proof of theorem 1 in BirgÃ© and Massart (2001), that for all $\gamma > 0$,

$$\mathbb{P}\Big(\sup_{v\in S_K} G(v) \geq E[\sup_{v\in S_K} G(v)] \geq +\gamma\Big) \leq \exp(\frac{-\gamma}{2}). \qquad \text{(III.9)}$$

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

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

Using the Cauchy Schwarz inequality, we derive

$$\sup_{v\in S_K} G(v) \leq \sup_{\alpha\in\mathbb{R}^D} \frac{\sum_{i=1}^{n}\varepsilon_i\Big(\sum_{j=1}^{D}\alpha_j\psi_{j,i}\Big)}{\sigma\sqrt{n}\Big(\sum_{j=1}^{D}\alpha_j^2\Big)^{\frac{1}{2}}} \tag{III.10}$$

$$\leq \Big(\sigma^2 n\Big)^{-\frac{1}{2}}\left\{\sum_{j=1}^{D}\left(\sum_{i=1}^{n}\varepsilon_i\psi_{j,i}\right)^2\right\}^{\frac{1}{2}}. \tag{III.11}$$

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

$$\mathbf{E}\Big[\sup_{v\in S_K} G(v)\Big] \leq (2K_{max}+1)^{\frac{1}{2}}. \tag{III.12}$$

Using *III.8*, *III.9*, and *III.12* with $\gamma = n^{\frac{1}{2}}\sigma^{-1}\frac{\alpha_n}{2} - n^{\frac{3}{2}}\sigma^{-1}\times\lambda_n - (2K_{max}+1)^{\frac{1}{2}}$, we have

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

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

$$n^{\frac{3}{2}}\sigma^{-1}\lambda_n + (2K_{max}+1)^{\frac{1}{2}} = A\frac{n^{\frac{1}{2}}\alpha_n}{2\sigma}.$$

It yields,

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

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

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 $O(n^{-\frac{1}{2}})$.

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

2 Estimation of the Change-Point Locations

In this section, we aim at estimating the change-point locations from the observations $(Y_1, ..., Y_n)$ satisfying Model *III.2*. The change-point estimates that we propose to study are obtained from the $\hat{\beta}_j(\lambda_n)$ is satisfying the criterion *III.4* as follows. Let us define the set of active variables by

$$\hat{S}(\lambda_n) := \{j \in \{1, ..., n\} : \hat{\beta}_j(\lambda_n) \neq 0\}.$$

Moreover, we define the change-point estimators by $\hat{t}_j(\lambda_n)$ satisfying

$$\hat{S}(\lambda_n) = \left\{\hat{t}_1(\lambda_n), ..., \hat{t}_{|\hat{S}(\lambda_n)|}(\lambda_n)\right\},$$

where

$$\hat{t}_1(\lambda_n) < ... < \hat{t}_{|\hat{S}(\lambda_n)|}(\lambda_n),$$

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

Let us now detail the assumptions under which the becoming theoretical results in the sequel are established. Define

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

We impose the following assumptions.

Assumption 1: The $\varepsilon_1, ..., \varepsilon_n$ are iid zero-mean random variables with $\mathrm{Var}[\varepsilon_1] = \sigma^2$ satisfying: there exisits a positive constant β such for all $v \in \mathbb{R}$,

$$\mathbb{E}[\exp(v\varepsilon_1)] \leq \exp(\beta v^2).$$

Assumption 2: The sequence $\{\delta_n\}_{n \geq 1}$ is a nonincreasing and positive sequence tending to zero as n tends to infinity and satisfying

$$\frac{n\delta_n (J^\star_{min})^2}{\log n} \to \infty, \; as \; n \to \infty.$$

Assumption 3: The change points $t_1^\star, ..., t_{K^\star}^\star$ satisfy

$$n\delta_n \leq I_{min}^\star, \text{ for all } n \geq 1.$$

Assumption 4: The sequence of regularization parameters $\{\lambda_n\}_{n\geq 1}$ is such that

$$\frac{n\lambda_n}{n\delta_n J_{min}^\star} \to 0, \, as \, n \to \infty.$$

We first state a lemma arising from the Karush Kuhn Tucker (KKT) conditions of the optimization problem stated in *III.4* which will be useful in the proof of the consistency of the procedure in the sequel.

Lemma 2.1 *Consider $Y_1, ..., Y_n$ a set of observations following the Model(2). Then, the change-points estimators $(\hat{t}_1(\lambda_n), ..., \hat{t}_n(\lambda_n))$ and $(\hat{u}_1(\lambda_n), ..., \hat{u}_n(\lambda_n))^\top$ defined by $\hat{u}_i(\lambda_n) = (X_n\hat{\beta}^n)_i$, where X_n is a $n \times n$ matrix nonzero elements equal to one and the$(\hat{\beta}^n)_{1\leq i\leq n}$ are obtained by..., satisfy*

$$\sum_{i=\hat{t}_\ell(\lambda_n)}^{n} Y_i - \sum_{i=\hat{t}_l(\lambda_n)}^{n} \hat{u}_i = \frac{n\lambda_n}{2}\hat{\alpha}_\ell, \forall \ell = 1, ..., |\hat{S}(\lambda_n)|. \tag{III.13}$$

and

$$\left| \sum_{i=j}^{n} Y_i - \sum_{i=j}^{n} \hat{u}_i \right| \leq \frac{n\lambda_n}{2}, \forall j = 1, ..., n. \tag{III.14}$$

Using the convention,

$$\begin{cases} \hat{\alpha}_\ell = +1, \hat{u}_{\hat{t}_\ell(\lambda_n)} > \hat{u}_{\hat{t}_\ell(\lambda_n)-1}; \\ \hat{\alpha}_\ell = -1, otherwise. \end{cases}$$

The vector $\hat{u}(\lambda_n) = (\hat{u}_1(\lambda_n), ..., \hat{u}_n(\lambda_n))^\top$ has the following additional property:

$$\begin{cases} \hat{u}_t(\lambda_n) = \hat{\mu}_k, \qquad \hat{t}_{k-1}(\lambda_n) \leq t \leq \hat{t}_k(\lambda_n) - 1, \\ k = 1, ..., |\hat{S}(\lambda_n)| + 1. \end{cases} \tag{III.15}$$

Proof. A necessary and sufficient condition for a vector $\hat{\beta}$ in $\mathbb{R}^n$ to minimize the function 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 $\Phi(\beta)$ at the point $\hat{\beta}$, that is, the

following KKT Optimality conditions

$$\begin{cases} \left(X_n^\top\left(Y_n - X_n\hat{\beta}\right)\right)_j = \frac{n\lambda_n}{2} sign(\hat{\beta}_j), \text{ if } \hat{\beta}_j \neq 0, \\ \left|\left(X_n^\top\left(Y_n - X_n\hat{\beta}\right)\right)_j\right| \leq \frac{n\lambda_n}{2} sign(\hat{\beta}_j), \text{ if } \hat{\beta}_j = 0. \end{cases}$$

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

Now, 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 2.2 *Let $\{\varepsilon_i\}_{1\leq i\leq n}$ be a sequence of random variables satisfying **Assumption 1**. If $\{v_n\}_{n\geq 1}$ and $\{x_n\}_{n\geq 1}$ are two positive sequence such that $\frac{v_n x_n^2}{\log n} \to \infty$ as $n \to \infty$, then*

$$\mathbb{P}\left(\max_{1\leq r_n<s_n\leq n;\,|r_n-s_n|\geq v_n}\left|(s_n-r_n)^{-1}\sum_{i=r_n}^{s_n-1}\varepsilon_i\right| \geq x_n\right) \to 0, \; as\, n\to\infty. \tag{III.16}$$

Proof. In the remainder, for any sequence of random variables, say, $Z_1, ..., Z_n$, we shall use the following notation:

$$Z(r;s) := \sum_{i=r}^{s} Z_i \quad \text{for any } 1 \leq r < s \leq n. \tag{III.17}$$

Using the notation introduced in *III.17*, we have

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

Using **Assumption 1**, it yields that for all$\delta > 0$,

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

Since the sharpest bound holds for $\delta = \frac{x_n}{2\beta}$, we get

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

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

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

Therefore, it yields that

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

which completes the proof. ∎

Proposition 2.1 *Let $Y_1, ..., Y_n$ be a set of observations satisfying Model III.1 then under the **Assumptions 1** to **4**, the change-points estimators $\{\hat{t}_1(\lambda_n), ..., \hat{t}_{|\hat{S}(\lambda_n)|}(\lambda_n)\}_{n\geq 1}$ satisfy, if $|\hat{S}(\lambda_n)| = K^\star$ with probability tending to one:*

$$\mathbb{P}\Big(\max_{1\leq k\leq K^\star} |\hat{t}_k - t_k^\star| \leq n\delta_n\Big) \to 1, \quad n \to \infty. \tag{III.18}$$

Proof. Since

$$\mathbb{P}(\max_{1\leq k\leq K^\star} |\hat{t}_k - t_k^\star| > n\delta_n) \leq \sum_{k=1}^{K^\star} P(|\hat{t}_k - t_k^\star| > n\delta_n),$$

it suffices to prove that for all $k = 1, ..., K^\star$,

$$\begin{cases} \mathbb{P}(A_{n,k}) \to 0, \\ \text{where } A_{n,k} := \{|\hat{t}_k - t_k^\star| \geq n\delta_n)\}. \end{cases}$$

Defining the set C_n by

$$C_n := \left\{\max_{1\leq k\leq K^\star} |\hat{t}_k - t_k^\star| < \frac{I^\star_{min}}{2}\right\}. \tag{III.19}$$

Hence, it is enough to prove, simultaneously, that

$$\begin{cases} \mathbb{P}(A_{n,k} \cap C_n) \to 0, \quad n \to \infty, \\ \mathbb{P}(A_{n,k} \cap C_n^c) \to 0, \quad n \to \infty. \end{cases}$$

Note that (III.19) implies that

$$t^\star_{k-1} < \hat{t}_k < t^\star_{k+1}, \quad \text{for all} \quad k = 1, ..., K^\star. \tag{III.20}$$

Let us consider the first case where $\hat{t}_k \leq t^\star_k$.

We begin by proving the first statement, i.e., $\mathbb{P}(A_{n,k} \cap C_n) \to 0, \quad as\, n \to \infty$.

Applying $refeq : III12$ in Lemma 2.1 (KKT) with $j = t^\star_k$ and $III.14$ in Lemma 2.1 with $\ell = k$ gives, respectively,

$$\left| \sum_{i=t^\star_k}^{n} Y_i - \sum_{i=t^\star_k}^{n} \hat{u}_i \right| \leq \frac{n\lambda_n}{2},$$

and

$$\sum_{i=\hat{t}_k(\lambda_n)}^{n} Y_i - \sum_{i=\hat{t}_k(\lambda_n)}^{n} \hat{u}_i = \frac{n\lambda_n}{2}\hat{\alpha}_l.$$

Using the additional property of the $\hat{u}_i$ we get that for all $k = 1, ..., K^\star$,

$$\begin{cases} -\frac{n\lambda_n}{2} \leq \sum_{i=t^\star_k}^{n} Y_i - \sum_{i=t^\star_k}^{n} \hat{u}_i \leq \frac{n\lambda_n}{2} \\ -\frac{n\lambda_n}{2} \leq \sum_{i=\hat{t}_k(\lambda_n)}^{n} Y_i - \sum_{i=\hat{t}_k(\lambda_n)}^{n} \hat{u}_i \leq \frac{n\lambda_n}{2} \end{cases}$$

It implies that

$$-n\lambda_n \leq \sum_{i=\hat{t}_k(\lambda_n)}^{t^\star_k - 1} Y_i - \sum_{i=\hat{t}_k(\lambda_n)}^{t^\star_k - 1} \hat{u}_i \leq n\lambda_n.$$

Hence, by using the Model(2),

$$\left| \sum_{i=\hat{t}_k(\lambda_n)}^{t^\star_k - 1} (\varepsilon_i + \mu^\star_k) - \sum_{i=\hat{t}_k(\lambda_n)}^{t^\star_k - 1} \hat{u}_i \right| \leq n\lambda_n,$$

also we have

$$\left| (t^\star_k - \hat{t}_k)\mu^\star_k + \varepsilon(\hat{t}_k; t^\star_k - 1) - \sum_{i=\hat{t}_k(\lambda_n)}^{\hat{t}_{k+1}-1} \hat{u}_i + \sum_{i=t^\star_k}^{\hat{t}_{k+1}-1} \hat{u}_i \right| \leq n\lambda_n.$$

So

$$\left| (\hat{t}_k - t^\star_k)(\mu^\star_{k+1} - \mu^\star_k) + \varepsilon(\hat{t}_k; t^\star_k - 1) + (\hat{t}_k - t^\star_k)(\hat{\mu}_{k+1} - \mu^\star_{k+1}) \right| \leq n\lambda_n.$$

Therefore on $C_n \cap \{\hat{t}_k \leq t_k^\star\}$ we have

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

Defining the event

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

It follows that

$$C_n \cap \{\hat{t}_k \leq t_k^\star\} \subset C_{n,k},$$

also

$$A_{n,k} \cap C_n \cap \{\hat{t}_k \leq t_k^\star\} = A_{n,k} \cap C_n \cap \{\hat{t}_k \leq t_k^\star\} \cap C_{n,k}.$$

Therefore,

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

By **Assumption 4**, $\frac{n\lambda_n}{n\delta_n J_{min}^\star} < \frac{1}{3}$, for n large enough, leading to $\mathbb{P}(A_{n,k,1}) \to 0$ as $n \to \infty$. By lemma 1.2.2 with $x_n = \frac{J_{min}^\star}{3}$, $v_n = n\delta_n$ and **Assumption 2**, $\mathbb{P}(A_{n,k,3}) \to 0$ as $n \to \infty$.

Let us now address $\mathbb{P}(A_{n,k,2})$. Using Lemma 2.1 (KKT) with $j = \frac{t_k^\star + t_{k+1}^\star}{2}$ and with $j = t_k^\star$, and using the triangle inequality, it follows that

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

Since we are in the event $C_n \cap \{\hat{t}_k \leq t_k^\star\}$, so we get

$$\hat{t}_k \leq t_k^\star \leq i \leq \frac{t_k^\star + t_{k+1}^\star}{2} - 1 \leq \hat{t}_{k+1} - 1,$$

and $\hat{u}_i \equiv \hat{\mu}_{k+1}$ within the interval $[t_k^\star, \frac{(t_k^\star + t_{k+1}^\star)}{2} - 1]$, which gives,

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

This implies that

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

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

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

which is arbitrary small if $n\lambda_n < \frac{I_{min}^\star \cdot J_{min}^\star}{12}$ for n large enough, and, by lemma 1.2.2 if $\frac{I_{min}^\star \cdot (J_{min}^\star)^2}{\log(n)} \to \infty$, as n tends to infinity. The last two conditions hold thanks to **Assumptions 2**, **3** and **4**. Since the proof in the case $\hat{t}_k \geq t_k^\star$ follows from the 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 $\mathbb{P}(A_{n,k} \cap C_n^c) \to 0$. Recall that by definition of c_n given in (1.7),

$$C_n^c := \left\{\max_{1 \leq k \leq K^\star} |\hat{t}_k - t_k^\star| \geq \frac{I_{min}^\star}{2}\right\}.$$

We split $\mathbb{P}(A_{n,k} \cap C_n^c)$ in three terms as following

$$\mathbb{P}(A_{n,k} \cap C_n^c) = \mathbb{P}(A_{n,k} \cap D_n^{(l)}) + \mathbb{P}(A_{n,k} \cap D_n^{(m)}) + \mathbb{P}(A_{n,k} \cap D_n^{(r)}),$$

where

$$\begin{aligned} D_n^{(l)} &:= \{\text{there exists } p \in \{1,...,K^\star\} : \hat{t}_p \leq t^\star_{p-1}\} \cap C_n^c \\ D_n^{(m)} &:= \{\text{for all } k \in \{1,...,K^\star\} : t^\star_{k-1} \leq \hat{t}_k \leq t^\star_{k+1}\} \cap C_n^c, \\ D_n^{(r)} &:= \{\text{there exists } p \in \{1,...,K^\star\} : \hat{t}_p \geq t^\star_{p+1}\} \cap C_n^c. \end{aligned}$$

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

$$\mathbb{P}(A_{n,k} \cap D_n^{(m)} \cap \{\hat{t}_k \leq t^\star_k\}) \leq \mathbb{P}(A_{n,k} \cap B_{k+1,k} \cap D_n^{(m)}) + \sum_{l=k+1}^{K^\star} \mathbb{P}(C_{l,l} \cap B_{l+1,l} \cap D_n^{(m)}), \quad \text{(III.21)}$$

where

$$\begin{cases} B_{p,q} := \{\hat{t}_p - t^\star_q \geq \frac{I^\star_{min}}{2}\}, \\ \text{with the convention } B_{K^\star+1,K^\star} := \{n - t^\star_{K^\star} \geq \frac{I^\star_{min}}{2}\}, \\ C_{p,q} := \{t^\star_p - \hat{t}_q \geq \frac{I^\star_{min}}{2}\}. \end{cases}$$

Let us now prove that the first term in the right hand side of (1.9) tends to zero as n tends to infinity , the argumetns for adressing the other terms being similar.

Using *III*.13 and *III*.14 in Lemma 2.1 with $j = t^\star_k$ and $\ell = k$, on the hand and *III*.14 in Lemma 2.1 with $j = t^\star_k$ and *III*.13 in Lemma 2.1 with $\ell = k+1$ on the other hand, we obtain, respectively:

$$|\hat{t}_k - t^\star_k||\hat{\mu}_{k+1} - \mu^\star_k| \leq n\lambda_n + |\varepsilon(\hat{t}_k; t^\star_k - 1)|, \quad \text{(III.22)}$$

and

$$|\hat{t}_{k+1} - t^\star_k||\hat{\mu}_{k+1} - \mu^\star_{k+1}| \leq n\lambda_n + |\varepsilon(t^\star_k; \hat{t}_{k+1} - 1)|. \quad \text{(III.23)}$$

In the one hand, we have

$$\begin{aligned} |\mu^\star_{k+1} - \mu^\star_k| &= |(\hat{\mu}_{k+1} - \mu^\star_k) - (\hat{\mu}_{k+1} - \mu^\star_{k+1})| \\ &\leq |\hat{\mu}_{k+1} - \mu^\star_k| + |\hat{\mu}_{k+1} - \mu^\star_{k+1}| \\ &\leq \frac{n\lambda_n}{|\hat{t}_k - t^\star_k|} + \frac{|\varepsilon(\hat{t}_k; t^\star_k - 1)|}{|\hat{t}_k - t^\star_k|} + \frac{n\lambda_n}{|\hat{t}_{k+1} - t^\star_k|} + \frac{|\varepsilon(t^\star_k; \hat{t}_{k+1} - 1)|}{|\hat{t}_{k+1} - t^\star_k|} \\ &\leq \frac{n\lambda_n}{n\delta_n} + \frac{|\varepsilon(\hat{t}_k; t^\star_k - 1)|}{|\hat{t}_k - t^\star_k|} + \frac{n\lambda_n}{|\hat{t}_{k+1} - t^\star_k|} + \frac{|\varepsilon(t^\star_k; \hat{t}_{k+1} - 1)|}{|\hat{t}_{k+1} - t^\star_k|} \end{aligned}$$

and in the other hand

$$\begin{aligned} |\hat{t}_{k+1} - t_k^\star| &= |(t_{k+1}^\star - t_k^\star) - (t_{k+1}^\star - \hat{t}_{k+1})| \\ &\geq |t_{k+1}^\star - t_k^\star| - |t_{k+1}^\star - \hat{t}_{k+1}| \\ &\geq I_{min}^\star - |\hat{t}_{k+1} - t_{k+1}^\star| \\ &\geq \frac{I_{min}^\star}{2} \end{aligned}$$

Defining the even E_n by

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

E_n occurs with probability equal to one. Therefore we obtain

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

By **Assumptions 2**, **3** and **4**, $\mathbb{P}(A_{n,k} \cap B_{k+1,k} \cap D_n^{(m)}) \to 0$, as n tends to infinity, which concludes that $\mathbb{P}(A_{n,k} \cap D_n^{(m)}) \to 0$.

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

$$\begin{aligned}
\mathbb{P}(D_n{}^{(l)}) &= \mathbb{P}\Bigg(\Big\{ \exists p \in \{1, ..., K^\star\} : \hat{t_p} \leq t^\star_{p-1} \Big\} \bigcap C_n^c \Bigg) \\
&= \mathbb{P}\Bigg(\Big\{ \bigcup_{k=1}^{K^\star} \max_{1 \leq l \leq K^\star} \{\hat{t_l} \leq t^\star_{l-1}\} = k \Big\} \bigcap C_n^c \Bigg) \\
&= \sum_{k=1}^{K^\star} \mathbb{P}\Bigg(\Big\{ \{ \max_{1 \leq l \leq K^\star} \{\hat{t_l} \leq t^\star_{l-1}\} = k \Big\} \bigcap C_n^c \Bigg) \\
&\leq \sum_{k=1}^{K^\star - 1} 2^{k-1} \mathbb{P}\Big(\{\hat{t_k} \leq t^*_{k-1}\} \bigcap \bigcap_{m \geq k}^{K^\star - 1} \{\hat{t}_{m+1} > t^\star_m\} \bigcap C_n^c \Big) + 2^{K^\star - 1} \mathbb{P}(\{\hat{t}_{K*} \leq t^\star_{K^\star - 1}\}) \\
&\leq \sum_{k=1}^{K^\star - 1} 2^{k-1} \mathbb{P}\Big(\{\hat{t_k} \leq t^*_{k-1}\} \bigcap \bigcap_{m \geq k}^{K^\star - 1} \{\hat{t}_{m+1} > t^\star_m\} \bigcap C_n^c \Big) + 2^{K^\star - 1} \mathbb{P}\Big(\{t^\star_{K^\star} - \hat{t}_{K*} > \frac{I^\star_{min}}{2}\} \Big) \\
&\leq \Sigma(n, K^\star) + 2^{K^\star - 1} \mathbb{P}\Big(\{t^\star_{K^\star} - \hat{t}_{K*} > \frac{I^\star_{min}}{2}\} \Big).
\end{aligned}$$

where,
$\Sigma(n, K^\star) := \sum_{k=1}^{K^\star - 1} 2^{k-1} \mathbb{P}\Big(\{\hat{t_k} \leq t^*_{k-1}\} \bigcap \bigcap_{m \geq k}^{K^\star - 1} \{\hat{t}_{m+1} > t^\star_m\} \bigcap C_n^c \Big)$.

Straightforward,

$$
\begin{aligned}
\Sigma(n,K^\star) &\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\mathbb{P}\Big(\{\hat{t_k}\leq t^\star_{k-1}\}\bigcap\bigcap_{m\geq k}^{K^\star-1}\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{\max_{1\leq m\leq K^\star}\{t^\star_m-\hat{t}_m\geq\frac{I^\star_{min}}{2}\}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\mathbb{P}\Big(\{\hat{t_k}\leq t^\star_{k-1}\}\bigcap\bigcap_{m\geq k}^{K^\star-1}\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{\max_{k\leq m\leq K^\star}\{t^\star_m-\hat{t}_m\geq\frac{I^\star_{min}}{2}\}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\mathbb{P}\Big(\{\hat{t_k}\leq t^\star_{k-1}\}\bigcap\bigcap_{m\geq k}^{K^\star-1}\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{\bigcup_{m\geq k}^{K^\star-1}\{t^\star_m-\hat{t}_m\geq\frac{I^\star_{min}}{2}\}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\mathbb{P}\Big(\{\bigcap_{m\geq k}^{K^\star-1}\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{\bigcup_{m\geq k}^{K^\star-1}\{t^\star_m-\hat{t}_m>\frac{I^\star_{min}}{2}\}\}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\mathbb{P}\Big(\bigcap_{m\geq k}^{K^\star-1}\bigcup_{m\geq k}^{K^\star-1}\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{t^\star_m-\hat{t}_m>\frac{I^\star_{min}}{2}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\mathbb{P}\Big(\{\bigcup_{m\geq k}^{K^\star-1}\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{(t^\star_m-\hat{t}_m)>\frac{I^\star_{min}}{2}\}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\sum_{m\geq k}^{K^\star-1}\mathbb{P}\Big(\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{(t^\star_m-\hat{t}_m)>\frac{I^\star_{min}}{2}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\sum_{m\geq k}^{K^\star-1}\mathbb{P}\Big(\{\hat{t}_{m+1}>t^\star_m\}\bigcap\{(t^\star_m-\hat{t}_m)>\frac{I^\star_{min}}{2}\}\Big)\\
&\leq \sum_{k=1}^{K^\star-1} 2^{k-1}\sum_{m\geq k}^{K^\star-1}\mathbb{P}\Big(\{\{\hat{t}_{m+1}-t^\star_m>\frac{I^\star_{min}}{2}\}\bigcup\{\hat{t}_{m+1}-t^\star_m<I^\star_{min}/2\}\}\bigcap\{(t^\star_m-\hat{t}_m)>\frac{I^\star_{min}}{2}\}\Big)\\
&\leq 2^{K^\star-1}\sum_{k=1}^{K^\star-1}\sum_{m\geq k}^{K^\star-1}\mathbb{P}\Big(\{(t^\star_m-\hat{t}_m)>\frac{I^\star_{min}}{2}\}\bigcap\{\hat{t}_{m+1}-t^\star_m>\frac{I^\star_{min}}{2}\}\Big).
\end{aligned}
$$

It follows that,

$$
\begin{aligned}
\mathbb{P}({D_n}^{(l)}) \leq\ & 2^{K^\star-1}\sum_{k=1}^{K^\star-1}\sum_{m\geq k}^{K^\star-1}\mathbb{P}\Big(\{(t^\star_m-\hat{t}_m)>\frac{I^\star_{min}}{2}\}\bigcap\{\hat{t}_{m+1}-t^\star_m>\frac{I^\star_{min}}{2}\}\\
&+2^{K^\star-1}\mathbb{P}\Big(\{t^\star_{K^\star}-\hat{t}_{K*}>\frac{I^\star_{min}}{2}\}\Big)\Big). \qquad \text{(III.24)}
\end{aligned}
$$

Consider one term of the sum in the right-hand side of $refeq : III16$. Using *III.22* and *III.23*

with $k = m$, we obtain

$$\begin{aligned}\mathbb{P}\Big(\Big\{t_m^\star - \hat{t}_m > \frac{I_{min}^\star}{2}\Big\} \cap \Big\{\hat{t}_{m+1} - t_m^\star > \frac{I_{min}^\star}{2}\Big\}\Big) \;\leq\; & \mathbb{P}\Big(\frac{4n\lambda_n}{I_{min}^\star} \geq \frac{|\mu_{m+1}^\star - \mu_m^\star|}{3}\Big) \\ & + \mathbb{P}\Big(\Big\{(t_m^\star - \hat{t}_m)^{-1}\Big|\varepsilon(\hat{t}_m; t_m^\star - 1)\Big| \geq \frac{|\mu_{m+1}^\star - \mu_m^\star|}{3}\Big\} \\ & \quad \bigcap \Big\{t_m^\star - \hat{t}_m \geq \frac{I_{min}^\star}{2}\Big\}\Big) \\ & + \mathbb{P}\Big(\Big\{(\hat{t}_{m+1} - t_m^\star)^{-1}\Big|\varepsilon(t_m^\star; \hat{t}_{m+1} - 1)\Big| \geq \frac{|\mu_{m+1}^\star - \mu_m^\star|}{3}\Big\} \\ & \quad \bigcap \Big\{\hat{t}_{m+1} - t_m^\star \geq \frac{I_{min}^\star}{2}\Big\}\Big)\end{aligned}$$

By **Assumptions 2**, **3**, and **4**,

$$\mathbb{P}\Big(\Big\{t_m^\star - \hat{t}_m > \frac{I_{min}^\star}{2}\Big\} \bigcap \Big\{\hat{t}_{m+1} - t_m^\star > \frac{I_{min}^\star}{2}\Big\}\Big) \to 0, \quad as\, n \to \infty.$$

Let us now consider the last term in the right hand of *III.24*. By using the observations *III.22* and *III.23* with $k = K^\star$ leads to

$$\begin{aligned}\mathbb{P}\Big(\Big\{t_{K^\star}^\star - \hat{t}_{K^\star} > \frac{I_{min}^\star}{2}\Big\}\Big) \;\leq\; & P\Big(\frac{3n\lambda_n}{I_{min}^\star} \geq \frac{|\mu_{K^\star+1}^\star - \mu_{K^\star}^\star|}{3}\Big) \\ & + \mathbb{P}\Big(\Big\{(t_{K^\star}^\star - \hat{t}_{K^\star})^{-1}\Big|\varepsilon(\hat{t}_{K^\star}; t_{K^\star}^\star - 1)\Big| \geq \frac{|\mu_{K^\star+1}^\star - \mu_{K^\star}^\star|}{3}\Big\} \\ & \quad \bigcap \Big\{t_{K^\star}^\star - \hat{t}_{K^\star} \geq \frac{I_{min}^\star}{2}\Big\}\Big) \\ & + \mathbb{P}\Big(\Big\{(n - t_{K^\star+1}^\star)^{-1}\Big|\varepsilon(t_{K^\star}^\star; n)\Big| \geq \frac{|\mu_{K^\star+1}^\star - \mu_{K^\star}^\star|}{3}\Big\}\Big)\end{aligned}$$

By **Assumptions 2**, **3**, and **4**,

$$\mathbb{P}\Big(\Big\{t_{K^\star}^\star - \hat{t}_{K^\star} > \frac{I_{min}^\star}{2}\Big\}\Big) \to 0, \quad as\, n \to \infty,$$

which gives

$$\mathbb{P}({D_n}^{(l)}) \to 0, \quad as\, n \to \infty.$$

In a similar way, we can prove that $\mathbb{P}({D_n}^{(r)}) \to 0$, as $n \to \infty$ which yields that $\mathbb{P}(A_{n,k} \cap C_n^c) \to 0$ and concludes the proof. ∎

Under the assumptions of Proposition 2.1, the $\hat{\tau}_k$ defined for all $k \in \{1, ..., K^\star\}$ by $\hat{t}_k = [n\hat{\tau}_k]$

are consistent estimators of the $\tau_k^\star$ defined by $t_k^\star = [n\tau_k^\star]$, for all $k \in \{1, ..., K^\star\}$ with the rate δ_n.
Note that with $\delta_n = \frac{(\log n)^2}{n}$, $J_{min}^\star \geq (\log n)^{\frac{1}{4}}$, $\lambda_n = \frac{\log n}{n}$ or $\lambda_n = \frac{\log n}{n^{\frac{3}{2}}}$, the **Assumptions 2** to **4** are satisfied leading thus to a rate of order $\frac{(\log n)^2}{n}$ for the estimation of the $\hat{\tau_k}$. With this choice of parameters, the authors 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$\frac{1}{n}$ according to Yao and Au(1989).

3 Estimation of the change-Point's Number

In Proposition 2.1, the number of estimated change points is assumed to be equal to the number of change points. since this information is not in general available, the authors propose to evaluate the distance between the set $\hat{T}_{n,K} := \left\{\hat{t}_1, ..., \hat{t}_K\right\}$ of K estimated change points and the set of true change points $T_n^\star := \left\{t_1^\star, ..., t_{K^\star}^\star\right\}$.
Let us define the distance $\mathcal{E}(.\|.)$, for any two sets A and B, by

$$\mathcal{E}(A\|B) := \sup_{b\in B} \inf_{a\in A} |a - b| \tag{III.25}$$

Obviously, when $K = K^\star$, Proposition 2.1 implies that, under the same assumptions $\mathcal{E}(\hat{T}_{n,K}\|T_n^\star) \leq n\delta_n$ and $\mathfrak{E}(T_n^\star\|\hat{T}_{n,K}) \leq n\delta_n$ with probability tending to one as n tends to infinity. In the case where $K > K^\star$, the authors proves in the following proposition that $\mathcal{E}(\hat{T}_{n,K}\|T_n^\star) \leq n\delta_n$ with probability tending to one as n tends to infinity.

Proposition 3.1 *Let $Y_1, ..., Y_n$ be a set of observations satisfying Model III.1 then under **Assumptions 1, 2, 4** and if $\frac{n\delta_n J_{min}^{\star 2}}{\log(\frac{n^3}{\lambda_n{}^2})} \to \infty$, the change-points estimators $\left\{\hat{t}_1(\lambda_n), ..., \hat{t}_{|\hat{S}(\lambda_n)|}(\lambda_n)\right\}_{n\geq 1}$ defined below, satisfy, if $|\hat{S}(\lambda_n)| \leq K^\star$ with probability tending to one:*

$$\mathbb{P}\left(\mathcal{E}(\hat{T}_{n,|\hat{S}(\lambda_n)|}\|T_n^\star) \leq n\delta_n\right) \to 1, \qquad as \quad n \to \infty. \tag{III.26}$$

Proof. By Lemma 2 of Meinshausen and Yu (2009), it yields with probability tending to one

$$|\hat{S}(\lambda)| \leq C\frac{n}{\lambda_n^2}, \tag{III.27}$$

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

$$P\left(\left\{\mathcal{E}(\hat{T}_{n,|\hat{S}(\lambda_n)|}\|T_n^\star) \geq n\delta_n\right\} \bigcap \left\{|\hat{S}(\lambda_n)| \geq K_{max}^\star Big\}\right) \to 0, \, as \, n \to \infty$$

, It is enough to prove that

$$P\Big(\Big\{\mathcal{E}(\hat{T}_{n,|\hat{S}(\lambda_n)|}\|T_n^\star) \geq n\delta_n\Big\}\bigcap\Big\{K^\star \leq |\hat{S}(\lambda)| \leq C\frac{n}{\lambda_n^2}\Big\}\Big) \to 0, \, as\, n \to \infty.$$

We observe that

$$\begin{aligned}\mathbb{P}\Big(\Big\{\mathcal{E}(\hat{T}_{n,|\hat{S}(\lambda_n)|}\|T_n^\star) \geq n\delta_n\Big\}\bigcap\Big\{K^\star \;\leq\; C\frac{n}{\lambda_n^2}\Big\}\Big) &\leq \mathbb{P}\Big(\mathcal{E}(\hat{T}_{n,K^\star|}\|T_n^\star) \geq n\delta_n\Big) \\ &\quad + \sum_{K>K^\star}^{C\frac{n}{\lambda_n^2}} \mathbb{P}\Big(\mathcal{E}(\hat{T}_{n,K^\star|}\|T_n^\star) \geq n\delta_n\Big). \quad \text{(III.28)}\end{aligned}$$

The first term of the right-hand side of *III.28* tends to zero as n tends to infinity since it is upper bounded by $\mathbb{P}\Big(\max_{1\leq k\leq K^\star}|\hat{t}_k - t_k^\star| \leq n\delta_n\Big)$ which tends to zero by Proposition 2.1. Let us now focus on the second term on the right-hand side of *III.28*. Note that

$$\begin{aligned}\sum_{K>K^\star}^{C\frac{n}{\lambda_n^2}} \mathbb{P}\Big(\mathcal{E}(\hat{T}_{n,K^\star|}\|T_n^\star) \geq n\delta_n\Big) &\leq \sum_{K>K^\star}^{C\frac{n}{\lambda_n^2}}\sum_{k=1}^{K_{max}^\star} \mathbb{P}\Big(\forall\, 1\leq l\leq K,\, |\hat{t}_k - t^\star|_k \leq n\delta_n\Big) \\ &:= \sum_{K>K^\star}^{C\frac{n}{\lambda_n^2}}\sum_{k=1}^{K_{max}^\star} \mathbb{P}\Big(E_{n,k,1}\Big) + \mathbb{P}\Big(E_{n,k,2}\Big) + \mathbb{P}\Big(E_{n,k,3}\Big),\end{aligned}$$

where

$$\begin{aligned}E_{n,k,1} &:= \Big\{\forall\, 1\leq l\leq K,\, |\hat{t}_k - t_k^\star| \leq n\delta_n \quad \text{and } \hat{t}_l < t_k^\star\Big\} \\ E_{n,k,2} &:= \Big\{\forall\, 1\leq l\leq K,\, |\hat{t}_k - t_k^\star| \leq n\delta_n \quad \text{and } \hat{t}_l > t_k^\star\Big\} \\ E_{n,k,3} &:= \Big\{\exists\, 1\leq l\leq K,\, |\hat{t}_k - t_k^\star| \leq n\delta_n, \Big\{|\hat{t}_{l+1} - t_l^\star| \geq n\lambda_n\Big\} \quad \text{and } \Big\{\hat{t}_l < t_k^\star < \hat{t}_{l+1}\Big\}\Big\}.\end{aligned}$$

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

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

Applying *III.14* in Lemma 2.1 with $j = t^\star$ and

Let us now address to $\mathbb{P}\Big(E_{n,k,2}^{(2)}\Big)$. Using *III.14* in Lemma 2.1 with $j = t_k^\star$ and with $j = t_{k+1}^\star$, we get

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

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

$$\begin{aligned}\mathbb{P}(\{|\hat{\mu}_{K+1}-\mu^\star_{k+1}|\geq\frac{|\mu^\star_{k+1}-\mu^\star_k|}{3}\}) & \\ &\leq \mathbb{P}(n\lambda_n\geq(t^\star_{k+1}-t^\star_k)\frac{|\mu^\star_{k+1}-\mu^\star_k|}{6}\}) \\ &\quad+\mathbb{P}(\{|\frac{\varepsilon(t^\star_k;t^\star_{k+1}-1)}{t^\star_{k+1}-t^\star_k}|\geq\frac{|\mu^\star_{k+1}-\mu^\star_k|}{6}\}),\end{aligned}$$

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

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

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

where

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

As for addressing $\mathbb{P}\Big(E_{n,k,3}\bigcap\Big\{\hat{t}_K>t^\star_{k-1}\Big\}\Big)$, we have to use twice Lemma 2.1. For $\mathbb{P}\Big(E^{(1)}_{n,k,3}\Big)$, we first use **??** and **??** in Lemma 2.1 with $j=t^\star_k$ and $\ell=l$, respectively. Second, we use **??** and **??** in Lemma 2.1 with $j=t^\star_k$ and $\ell=l+1$, respectively. For $P\Big(E^(2)_{n,k,3}\Big)$, we first use Lemma 2.1 with $j=t^\star_k$ and $\ell=l$. Second, we use Lemma 2.1 with $j=t^\star_k$ and $j=t^\star_{k+1}$. For $\mathbb{P}(E^(3)_{n,k,3})$, we first use Lemma 2.1 with $j=t^\star_{k-1}$ and $j=t^\star_k$. Second, we use Lemma 2.1 with $j=t^\star_k$ and $\ell=l+1$. Finally, for $\mathbb{P}(E^(4)_{n,k,3})$, we first use Lemma 2.1 with $j=t^\star_{k-1}$ and $j=t^\star k$. Second, we use Lemma 2.1 with $j=t^\star_k$ and $j=t^\star_{k+1}$.

■

Note that with $\delta_n=\frac{(\log n)^2}{n}$, $J^\star_{min}\geq(\log n)^{\frac{1}{4}}$, $\lambda_n=\frac{\log n}{n}$ or $\lambda_n=\frac{\log n}{n^{\frac{3}{2}}}$, the **Assumptions 3**,

Note that with $\delta_n = \frac{(\log n)^2}{n}$, $J^\star_{min} \geq (\log n)^{\frac{1}{4}}$, $\lambda_n = \frac{\log n}{n}$ or $\lambda_n = \frac{\log n}{n^{\frac{3}{2}}}$, the **Assumptions 3**, **4** and $\frac{n\delta_n J^{\star 2}_{min}}{\log(\frac{n^3}{\lambda_n^2})} \to \infty$, are fullfilled of Proposition 2.2.

4 Least Squares-Total Variation with LARS

In this section, we detail the process of the LARS implemented to the method of LS-TV. For the sake of simplicity, one can shall describe here the algorithm where he looks for K_{max} change points, K_{max} being a known upper bound on the true number of change points

Suppose that we have performed $k-1$ iterations in the algorithm, then the current set of estimated change points, that is, the active set in the variable selection framework, is $\hat{T}_{n,k-1} = \left\{\hat{t}_1, ..., \hat{t}_{k-1}\right\}$ and the current set of estimated segment levels is $\left\{\hat{u}_1(k-1), ..., \hat{u}_n(k-1)\right\}$.

We are now describing the computational requirements of the kth iteration of the algorithm.

First, we look to the next change point $\hat{t}_k$ to add to $\hat{T}_{n,k-1}$ yielding the largest discrepancy with the true signal. This requires, given $\left\{\hat{u}_1(k-1), ..., \hat{u}_n(k-1)\right\}$, the computation of the n cumulative sums $\left\{\sum_{i=j}^{n} \hat{u}_i(k-1)\right\}_{j=1,...,n}$. These cumulative sums may actually be computed in $O(n)$ operations in time, using the simple recursion $\sum_{i=j}^{n} \hat{u}_i(k-1) = \sum_{i=j+1}^{n} \hat{u}_i(k) + \hat{u}_j(k)$. 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 has $O(n + \log(n))$ time complexity.

Second, we have to compute the descent direction, which involves the multiplication of the inverse of $k \times k$-matrix by a k-long vector. Indeed, $\mathbf{X}_k$ is a matrix which consists of the columns of $\mathbf{X}$ indexed by the element of $\hat{T}_{n,k}$ and $\mathbf{1}_k$ denotes the vector of dimension k with each component equal to one. Given the current set of change-points $\hat{T}_{n,k}$, the inverse maybe computed in $O(k^2)$ operations, since the entries of the inverse matrix of size $k \times k$ are available in close form beforehand. Then, the multiplication of $k \times k$-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 $O(K^2_{max})$.

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, $\alpha_j = sign(\hat{u}_{j+1}(k) - \hat{u}_j(k))$. 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 $\hat{\gamma} = \widetilde{\gamma}$ is unknown in general, and no theoretically grounded upper bound on N was provided in the literature so far.
Finally, the implementation of LS-TV based upon the LAR/LASSO algorithm runs in at most $O(K_{max}^3 + K_{max} n \log(n))$ in time.

LS-TV with LAR/LASSO

Initialization, $k = 0$.
(a) Set $\hat{T}_{n,0} = \emptyset$.
(b) Set $\hat{u}_i(0) = 0$, for all $i = 1, ..., n$.
While $k < K_{max}$.
(a) *Change-Points Addition*:
Find $\hat{t}_k$ such that

$$\hat{t}_k = \arg \max_{t \in \{1,...,n\} - \hat{T}_{n,k-1}} \Big| \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \hat{u}_i(k-1) \Big|.$$

(b) *Descent Direction Computation*:

$$\mathbf{w}_k = \Big(X_k^\top X_k \Big)^{-1} \mathbf{1}_k.$$

(c) *Descent Step Search*:
Search for $\hat{\gamma}$ such that

$$\hat{\gamma} = \min_{t \in \{1,...,n\} - \hat{T}_{n,k}} \left(\frac{\sum_{i=t}^{n} (Y_i - \sum_{i=t}^{n} \hat{u}_i(k))}{1 - \sum_{i=t}^{n} w_{k,i}}, \frac{\sum_{i=t}^{n} (Y_i + \sum_{i=t}^{n} \hat{u}_i(k))}{1 + \sum_{i=t}^{n} w_{k,i}} \right).$$

(d) *Zero-Crossing Check*:
If

$$\hat{\gamma} > \widetilde{\gamma} := \min_j \Big(\alpha_j w_{k,i} \Big)^{-1} \Big(\sum_{i=t}^{n} \hat{u}_i(k) \Big),$$

then decrease $\hat{\gamma}$ down to $\hat{\gamma} = \widetilde{\gamma}$, and remove $\widetilde{t}$ from $\hat{T}_{n,k}$, where

$$\widetilde{t} := \arg\min_j \Big(\alpha_j w_{k,i} \Big)^{-1} \Big(\sum_{i=t}^{n} \hat{u}_i(k) \Big).$$

References

[1] Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2012), " Optimization with Sparsity-Inducing Penalties ", Foundations and Trends in Machine Learning, Vol. 4 (1), pp. $1 - 106$.

[2] Bickel, P. J., Ritov, Y. and Tsybakov, A. (2009), " Simultaneous Analysis of LASSO and Dantzig selector ", The Annals of Statistics, Vol. 37(4), pp. $1705 - 1732$. (2001)

[3] Birgé, L. and Massart, P.(2001), " Gaussian Model Selection ", J. Eur. Math. Soc. Vol. 3, pp. $203 - 268$.

[4] Boyd, S. and Vandenberghe, L. (2004), " Convex Optimization ", Cambridge University Press, Cambridge,.

[5] Buehlmann, P. and Van de Geer, S. (2011), " Statistics for High-Dimensional Data: Methods, Theory and Applications ", Springer Verlag Berlin.

[6] Cai, T., Xu, G. and Zhang, J. (2009), " On Recovery of Sparse Signals via ℓ_1-minimization ", IEEE Transactions on Information Theory, Vol. 57(7), pp. $3388 - 3397$.

[7] Candès, E. and Tao, T. (2007), " The Dantzig Selector: Statistical Estimation when p is much arger than n", The Annals of Statistics, Vol. 35, pp. $2313 - 2351$.

[8] Chen, X. (2012), " LASSO-Type Sparse Regression and High-dimensional Gaussian Graphical Models ", PhD Thesis.

[9] Cirelson, B.S., Ibragimov, I.A. and Sudakov, V.N. (1976), " Norms of Gaussian Sample Functions ", Proc. 3rd Japan-USSR Symp. Probab. Theory, Taschkent 1975, Lect. NotesMath. 550, pp. $20 - 41$.

[10] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004), " Least Angle Regression ", The Annals of Statistics, Vol. 32, pp. $407 - 499$.

[11] Fan, J. and Li, R. (2006), " Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery ", in J. L. M. Sanz-Sole, J. Soria and J. Verdera, eds, International Congress of Mathematicians, Vol. 3, European Mathematical Society, Freiburg, pp. 595 – 622.

[12] Fu, W. J.(1998), " Penalized Regressions: The Bridge Versus The LASSO ", J. Comput. Graph. Stat., Vol. 7(3), pp. 397 – 416.

[13] Harchaoui, Z. and Lévy-Leduc, C. (2008), " Catching Change-Points with LASSO ", Advances in Neural Information Processing Systems (NIPS).

[14] Harchaoui, Z. and Lévy-Leduc, C. (2010), " Multiple Change-Point Estimation with a Total-variation Penalty ", Journal of the American Statistical Association, pp. 105 – 492.

[15] Hastie, T., Tibshirani, R. and Friedman, J. (2008), " The Elements of Statistical Learning: Data Mining, Inference, and Prediction ", second edn, Springer.

[16] Hesterberg, T., Choi, N., Meier, L. and Fraley, C (2008), " Least Angle and ℓ_1-Penalized Regression: A Review ", Statistics Surveys Vol. 2, pp. 61 – 93.

[17] Hoerl, A. and Kennard, R. (1988), " Ridge regression", Encyclopedia of Statistical Sciences, Vol. 8, Wiley, New York, pp. 129 – 136.

[18] Knight, K. and Fu, W. (2000), " Asymptotics for LASSO-Type Estimators ", The Annals of Statistics, Vol. 28(5), pp. 1356 – 1378.

[19] Leng, C., Lin, Y. and Wahba, G. (2006), " A Note on the LASSO and Related Procedures in Model Selction", Statistica Sinica, Vol. 16, pp. 1273 – 1284.

[20] Lounici, K. (2008), " Sup-norm Convergence Rate and Sign Concentration Property of LASSO and Dantzig Estimators", Electronic Journal of Statistics, Vol. 2, pp. 90 – 102.

[21] Mairal, J. and Yu. B (2012), " Complexity Analysis of the LASSO Regularization Path ", International Conference on Machine Learning.

[22] March, D. (2011), " Statistical Models for High-dimensional Data Selection and Assessment ", PhD Thesis.

[23] Meinshausen, N. (2007), " Relaxed LASSO ", Computational Statistics and Data Analysis, Vol. 52(1), pp. 374 – 393.

[24] Meinshausen, N. and Yu, B. (2009), " LASSO-Type Recovery of Sparse Representations for high-dimensional data ", The Annals of Statistics, Vol. 37(1), pp. 246 – 270.

[25] Osborne, M.R., Presnell, B. and Turlach, B.A.(2000), " A New Approach to Variable Selection in Least Squares Problems ", IMA Journal of Numerical Analysis, Vol. 20,pp. 389 − 403.

[26] Saharon, R and Zhu, J. (2007), " Piecewise Linear Regularized Solution Paths ", The Annals of Statistics 35:1012-1030.

[27] Tibshirani,R.(1996), " Regression Shrinkage and Selection via the LASSO ", Journal of the Royal Statistical Society Series B-Methodological, Vol. 58(1), pp. 267 − 288.

[28] Tibshirani, J. R.(2012), " The LASSO Problem and Uniqueness ".

[29] Van de Geer, S. (2010), " ℓ_1-Regularization in High-dimensional Statistics ", Proceedings of the International Congress of Mathematiciansm Hyderabad, India.

[30] Wainwright, M. J. (2006), " Sharp Thresholds for High-dimensional and Noisy Recovery of sparsity ", Technical Report 708, Department of Statistics, UC Berkeley.

[31] Yao, Y. and Au, S. T. (1989), " Least-Squares Estimation of a Step Function ", Sankhya: The Indian Journal of Statistics, Series A (1961-2002), Vol. 51, No. 3 , pp. 370 − 381

[32] Zhao, P. and Yu, B. (2006), " On Model Selection Consistency of LASSO ", Journal of Machine Learning Research, Vol. 7, pp. 2451 − 2563.

[33] Zou, H. (2006), " The Adaptive LASSO and its Oracle Properties", Journal of the American Statistical Association, Vol. 101, pp. 1418 − 1429.