Journal of Machine Learning Research 20 (2019) 1-40 Submitted 7/18; Revised 4/19; Published 6/19

An Efficient Two Step Algorithm for High Dimensional Change Point Regression Models Without Grid Search

Abhishek Kaul ABHISHEK.KAUL@WSU.EDU
Department of Mathematics and Statistics
Washington State University
Pullman, WA 99164, USA.

Venkata K. Jandhyala JANDHYALA@WSU.EDU
Department of Mathematics and Statistics
Washington State University
Pullman, WA 99164, USA.

Stergios B. Fotopoulos FOTOPO@WSU.EDU
Department of Finance and Management Science
Washington State University
Pullman, WA 99164, USA.

Editor: Ryan Tibshirani

Abstract

We propose a two step algorithm based on ℓ_1/ℓ_0 regularization for the detection and estimation of parameters of a high dimensional change point regression model and provide the corresponding rates of convergence for the change point as well as the regression parameter estimates. Importantly, the computational cost of our estimator is only $2{\cdot}\text{Lasso}(n, p)$, where $\text{Lasso}(n, p)$ represents the computational burden of one Lasso optimization in a model of size (n, p). In comparison, existing grid search based approaches to this problem require a computational cost of at least $n \cdot \text{Lasso}(n, p)$ optimizations. Additionally, the proposed method is shown to be able to consistently detect the case of 'no change', i.e., where no finite change point exists in the model. We allow the true change point parameter τ_0 to possibly move to the boundaries of its parametric space, and the jump size $\|\beta_0 - \gamma_0\|_2$ to possibly diverge as n increases. We then characterize the corresponding effects on the rates of convergence of the change point and regression estimates. In particular, we show that, while an increasing jump size may have a beneficial effect on the change point estimate, however the optimal rate of regression parameter estimates are preserved only upto a certain rate of the increasing jump size. This behavior in the rate of regression parameter estimates is unique to high dimensional change point regression models only. Simulations are performed to empirically evaluate performance of the proposed estimators. The methodology is applied to community level socio-economic data of the U.S., collected from the 1990 U.S. census and other sources.

Keywords: Change point regression, High dimensional models, ℓ_1, ℓ_0 regularization, Rate of convergence, Two phase regression

1. Introduction

Regression models are fundamental to supervised learning and statistical modelling of data collected from scientific phenomena. While applying regression models, one often assumes the regression parameters to be stable over time. However, this assumption may be rigid and may not hold in several environmental, biological and economic models, particularly when data is collected over an extended period of time. There are several approaches to model this dynamic phenomenon in regression parameters. One approach is to let the parameters change at certain unknown time points of the sampling period (Hinkley (1970), Hinkley (1972), Jandhyala and MacNeill (1997), Bai (1997), Jandhyala and Fotopoulos (1999), Fotopoulos et al. (2010) and Jandhyala et al. (2013)). Another closely related approach is to formulate the change point based on one or more covariate thresholds (Hinkley (1969), Koul and Qian (2002) and Koul et al. (2003)). In the literature, it is common to broadly call both as change point regression models. Such dynamic models have been found to have wide ranging applications in all areas of scientific inquiry (Reeves et al. (2007), Lund et al. (2007), and Liu et al. (2013)).

Technological advances in the past two decades have led to the wide availability of large scale/high dimensional data sets in several areas of applications such as genomics, social networking, empirical economics, finance etc. This has led to rapid development of high dimensional statistical methods. A large body of literature has now been developed pertaining to the study of regression models capable of allowing a vastly larger number of parameters p than the sample size n. One of the most successful methods for analysing high dimensional regression models is the Lasso, which is based on the least squares loss and ℓ_1 regularization (Tibshirani (1996)). Innumerable investigations have since been carried out to study the behavior of the Lasso estimator and its various modifications in many different settings (see e.g., Zou (2006), Zhao and Yu (2006), Bickel et al. (2009), Belloni et al. (2011), Belloni et al. (2017a), Kaul (2014), Kaul and Koul (2015), and the references therein). For a general overview on the developments of Lasso and its variants we refer to the monograph of Bühlmann and Van De Geer (2011) and the review article of Tibshirani (2011). All aforementioned articles provide results in a regression setting where the parameters are dynamically stable. In the recent past, work has also been carried out in the context of high dimensional change point models in an ‘only means’ setup, where change occurs in only the mean of time ordered independent random vectors, with the dimension of the observation vector being larger than the number of observations (Cho and Fryzlewicz (2015), Fryzlewicz (2014), and Wang and Samworth (2018) among others). Here the change is characterized in the sense of a dynamic mean vector. Another context in which high dimensional change point models have been investigated is that of a dynamic covariance structure which is related to the study of evolving networks (Gibberd and Roy (2017), and Atchade and Bybee (2017)). In contrast, change point methods for high dimensional linear regression models have received much less attention and only a select few articles have considered this problem in the recent literature (Ciuperca (2014), Zhang et al. (2015), Leonardi and Bühlmann (2016), Lee et al. (2016), and Lee et al. (2018)).

In this paper, we consider a high dimensional linear regression model with a potential change point,

$$y_i = x_i^T\beta_0\mathbf{1}[w_i \le \tau_0] + x_i^T\gamma_0\mathbf{1}[w_i > \tau_0] + \varepsilon_i, \quad i = 1, ..., n. \tag{1.1}$$

The observed variables in model (1.1) are, $y_i \in \mathbb{R}$, the p-dimensional predictors $x_i \in \mathbb{R}^p$, and change inducing variable $w_i \in \mathbb{R}$, $i = 1, .., n$. The unknown parameters of interest are the regression parameters $\beta_0, \gamma_0 \in \mathbb{R}^p$, and the change point $\tau_0 \in \bar{\mathbb{R}}^\star := \mathbb{R} \cup \{-\infty\}$. The change point τ_0 represents a threshold value of the variable w subsequent to which the regression parameter changes from its initial value β_0 to a new value γ_0. Note that, the 'no change' case is allowed by the model (1.1), since we allow $\tau_0 = -\infty$, in its parametric space. In this case, model (1.1) reduces to an ordinary high dimensional linear regression model. The parametric space $\bar{\mathbb{R}}^\star$ of τ is restricted to only contain $-\infty$, and not ∞, since both of these points characterize the 'no change' scenario and are unidentifiable from each other. This differs from the usual characterization of the 'no change' case, which is typically defined by $\beta_0 = \gamma_0$ and $\tau_0 \in \mathbb{R}$, for e.g. in Lee et al. (2016) and Lee et al. (2018). However it should be understood that this difference is only notational and both are characterizing the same null model. It should also be noted that the change point $\tau_0 \in \bar{\mathbb{R}}^*$, may itself depend on n, i.e., as the sample size increases, the change point may shift its location. However, for clarity of exposition, this dependence is notionally suppressed in the rest of this article. Furthermore, we let $p >> n$, so that model (1.1) corresponds to a high dimensional setting. Also, consistent with current literature, we assume that only a total of s components of β_0 and γ_0 are non-zero, i.e., $\|\beta_0\|_0 + \|\gamma_0\|_0 \le s$, where $s < n$.

Recently, models similar to (1.1) have been studied by Ciuperca (2014), Zhang et al. (2015), Leonardi and Bühlmann (2016), Lee et al. (2016) and Lee et al. (2018). Similar to model (1.1), Lee et al. (2016) and Lee et al. (2018) consider a high dimensional model with only a single unknown change point, whereas, Zhang et al. (2015), and Leonardi and Bühlmann (2016) consider a model where multiple change points may be present in the model. The articles Zhang et al. (2015) and Ciuperca (2014) consider a multiple change point setting where the dimension p of the regression parameters is fixed. The common thread in these articles is to provide regularized estimators for the parameters β, γ, τ and study their rates of convergence under different norms. In an earlier work, Wu (2008) provided an information-based criterion for carrying out change point analysis and variable selection in the fixed p setting; this methodology, however is not extendable to the high dimensional case. While these articles make important contributions to this fast emerging area, many aspects of this problem remain to be understood. For example, existing methods may be unable to satisfactorily detect the 'no change' case, these estimation methods may be computationally challenging to implement, and the underlying technical assumptions required for their theoretical validity may be restrictive.

The most commonly applied approach to estimating parameters of models such as (1.1) with a single change point is to consider,

$$(\hat{\beta}, \hat{\gamma}, \hat{\tau}) = \mathop{\arg\min}_{\beta, \gamma \in \mathbb{R}^p, \tau \in \mathcal{T}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \text{loss}(\text{data}, \beta, \gamma, \tau) + \text{pen}(\beta, \gamma, \tau) \right\}, \tag{1.2}$$

where $\text{loss}(\text{data}, \beta, \gamma, \tau)$ is an appropriately chosen loss function, and $\text{pen}(\beta, \gamma, \tau)$ is a suitably defined penalty function on the parameters β, γ, τ $\big($e.g., Lee et al. (2016), Lee et al. (2018) and (5.1) $\big)$. Here $\mathcal{T}$ is a restriction on the parametric space of the change point parameter τ. The loss function in (1.2) is nonconvex and consequently a direct optimization of (1.2) is typically computationally infeasible. To circumvent this difficulty, the space $\mathcal{T}$

is usually broken into a grid of points $\mathcal{T}^*$, and $\hat\beta(\tau),\hat\gamma(\tau)$ are computed for each $\tau\in\mathcal{T}^*$. The estimate $\hat\tau$ of the change point τ_0 is then obtained as that $\tau\in\mathcal{T}^*$ which minimizes the objective function in (1.2) over $\hat\beta(\tau),\hat\gamma(\tau)$. When the loss function is least squares and the penalty is of an ℓ_1-type, the computational cost of the above grid search is $|\mathcal{T}^*|\text{Lasso}(n,p)$, where $|\mathcal{T}^*|$ is typically of order n. Note that this grid search mechanism becomes computationally intensive as n,p increase. In the case of multiple change points, Zhang et al. (2015), and Leonardi and Bühlmann (2016) provide dynamic programming approaches that can estimate the number and locations of change points with the same $n\text{Lasso}(n,p)$ computational cost.

In this article we develop a two step algorithm for detection and estimation of parameters of model (1.1), so that a full grid search is avoided even as the near optimal rates of all parameter estimates are preserved. The idea for developing such an algorithm originates from the following simple and yet surprising numerical observation. Suppose we first choose virtually any initial value $\tau^{(0)}\in\text{Supp}(w)$, separated from its boundaries and then compute regression coefficients $\hat\beta^{(0)}$, $\hat\gamma^{(0)}$ on the binary partition $\big\{i;\, i\in\{1,..,n\},\ w_i\le\tau^{(0)}\big\}$ and $\big\{i;\, i\in\{1,..,n\},\ w_i>\tau^{(0)}\big\}$ respectively. Then a single update of the change point estimate obtained by optimization of the least squares loss over the change point parameter, using the previously obtained regression parameter estimates $\hat\beta^{(0)}$, $\hat\gamma^{(0)}$, yields a very precise estimate of the unknown change point, where the precision of this estimate is indistinguishable from existing full grid search approaches in any uniform sense. This simple numerical observation is very surprising, since it suggests that any initial $\tau^{(0)}$ which carries even a 'fractional amount of information' on the unknown τ_0 (this notion is described precisely later in Section 2), can be utilized to obtain an updated estimate $\hat\tau^{(1)}$ in a single step, which lies in a near optimal neighborhood of τ_0. In other words, the single step update process pulls in the initial guess $\tau^{(0)}$ from a much wider (nearly arbitrary) neighborhood of τ_0, to a near optimal neighborhood of τ_0. The usefulness of this process is immediate, as it removes the necessity of a grid search. The main contribution of this article is to develop a mathematical treatment of this two step algorithm. In doing so we also allow the possibility of 'no change' in the model (1.1).

More precisely, in this article we propose estimators based on ℓ_1/ℓ_0 regularization for the parameters τ_0, β_0, and γ_0 of model (1.1). The proposed methodology completely avoids a grid search approach for locating the unknown change point, consequently has a computational cost of only $2\text{Lasso}(n,p)$, significantly below the $n\text{Lasso}(n,p)$ cost of existing methods. A second important novelty of the proposed method, is its ability to detect the 'no change' case, which is achieved by a ℓ_0 regularization in the change point estimator. From a technical perspective, the rates of convergence associated with the proposed estimators are such that they are optimal for the regression parameter estimates and match the best rate of convergence available in the literature for estimating the change point. Before we describe our proposed methodology in Section 2, we outline below the notations used in this paper.

Notation: Throughout the paper, for any vector $\delta\in\mathbb{R}^p$, $\|\delta\|_0$ represents the number of non-zero components in δ. The norms $\|\delta\|_1$ and $\|\delta\|_2$ represent the standard 1-norm and Euclidean norm, respectively. The norm $\|\delta\|_\infty$ represents the sup norm, i.e., the maximum of absolute values of all elements. For any set of indices $T\subseteq\{1,....,p\}$, let $\delta_T=(\delta_j)_{j\in T}$ represent a sub-vector of δ containing components corresponding to the indices in T. Also, we let $|T|$ represent the cardinality of the set T. The notation $\mathbf{1}[\cdot]$ represents the usual indicator function. We denote by $a \wedge b = \min\{a, b\}$, and $a \vee b = \max\{a, b\}$, for any $a, b \in \mathbb{R}$. In the following, let $\text{Supp}(w)$ represent the support of the distribution of w and $\Phi(\cdot)$ be its cdf. Also denote by $\Phi_{\min}(\tau_0) = \Phi(\tau_0) \wedge (1 - \Phi(\tau_0))$. We shall use the following notation to represent generic constants that may be different from one term to the next. For example, $0 < c_u < \infty$ represent universal constants, whereas $0 < c_m < \infty$ are constants that depend on model parameters such as variance parameters of underlying distributions. The generic constants $0 < c_1, c_2 < \infty$ are used to denote constants that may depend on both c_u, and c_m. Lastly, we shall denote by $\bar{\mathbb{R}}^\star := \mathbb{R} \cup \{-\infty\}$, as the extended Euclidean space with negative infinity included. In the following, without loss of generality we assume that $\text{Supp}(w) = \mathbb{R}$.

2. Methodology and Related Work

We begin this section by describing the proposed methodology for the detection and estimation of parameters of model (1.1). For this purpose we require the following notation. Let for any $\tau \in \bar{\mathbb{R}}^\star$, $\beta, \gamma \in \mathbb{R}^p$,

$$Q(\tau, \beta, \gamma) \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\mathbf{1}[w_i \le \tau] + \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\gamma)^2\mathbf{1}[w_i > \tau].^{a} \qquad (2.1)$$

Additionally define the following Condition I that shall serve as an initializing condition required for the construction of our proposed algorithm.

Condition I: Let $u_n^{(0)}$ be a non-negative sequence defined as,

$$u_n^{(0)} = 1 \wedge c_u\Big(\frac{s\log p}{nl_n^2}\Big)^{\frac{1}{k}}, \quad \text{for any constants, } k \in [1, \infty), \text{ and } c_u > 0,^{b} \qquad (2.2)$$

where $0 < l_n \le 1/2$ is any non-negative sequence. Then, assume that the initializer $\tau^{(0)}$ satisfies,

$$\Phi_{\min}(\tau^{(0)}) \ge c_u l_n, \quad \text{and} \quad |\Phi(\tau^{(0)}) - \Phi(\tau_0)| \le u_n^{(0)}. \qquad (2.3)$$

In the above Condition I, the sequence l_n and the constant k are arbitrary, subject to satisfying Condition A(iii) to follow. The rate of the sequence l_n shall control the ability of Algorithm 1 to detect a finite change point near the boundaries of $\mathbb{R}$. Specifically Algorithm 1 shall be able to detect a finite change point (when it exists), such that $\Phi_{\min}(\tau_0)$ is of order at least that of l_n. In the case where it is assumed that, either $\Phi(\tau_0) = 0$ or $\Phi_{\min}(\tau_0) \ge c_u > 0$, then we can set $l_n \ge c_u$. Here, Algorithm 1 will be able to distinguish between the two cases, whether (a) there is no change point, $\Phi(\tau_0) = 0$ or (b) there is a finite change point $\tau_0 \in \mathbb{R}$ such that $\Phi(\tau_0)$ is bounded below.[c]

Then, the two step algorithm which we propose to obtain change point and regression coefficient estimates is described in Algorithm 1 below.

[a] Here, define $\mathbf{1}[w_i \le \tau] = 0$, for $\tau = -\infty$.

[b] Note here that the constant k is arbitrary, hence it can itself depend on initial $\tau^{(0)}$, i.e., the farther the guess $\tau^{(0)}$ is from τ_0 the larger k can be chosen in order to satisfy Condition I.

[c] The quantities l_n and k are only required for analysis of Algorithm 1. These quantities play no role in its implementation.

Algorithm 1: Detection and estimation of change point and regression parameters

Step 0 (Initialize): Choose any initial value $\tau^{(0)} \in \mathbb{R}$ satisfying Condition I. Compute the initial regression parameter estimates,

$$\big(\hat{\beta}^{(0)}, \hat{\gamma}^{(0)}\big) = \arg\min_{\beta,\gamma\in\mathbb{R}^p} \Big\{ Q(\tau^{(0)}, \beta, \gamma) + \lambda_1 \|(\beta^T, \gamma^T)^T\|_1 \Big\}, \quad \lambda_1 > 0.$$

Step 1: Update $\tau^{(0)}$ to obtain the change point estimate $\hat{\tau}^{(1)}$ where [d],

$$\hat{\tau}^{(1)} = \arg\min_{\tau\in\bar{\mathbb{R}}^{\star}} \Big\{ Q(\tau, \hat{\beta}^{(0)}, \hat{\gamma}^{(0)}) + \mu\|\Phi(\tau)\|_0 \Big\}, \quad \mu > 0. \tag{2.4}$$

Step 2: Update $(\hat{\beta}^{(0)}, \hat{\gamma}^{(0)})$ to obtain regression parameter estimates $(\hat{\beta}^{(1)}, \hat{\gamma}^{(1)})$ where,

$$\big(\hat{\beta}^{(1)}, \hat{\gamma}^{(1)}\big) = \arg\min_{\beta,\gamma\in\mathbb{R}^p} \Big\{ Q(\hat{\tau}^{(1)}, \beta, \gamma) + \lambda_2 \|(\beta^T, \gamma^T)^T\|_1 \Big\}, \quad \lambda_2 > 0.$$

A first concern that may arise to reader regarding **Step 0** of Algorithm 1 pertains to the initializing conditions in (2.3) of Condition I. The first of these conditions is clearly innocuous, all it requires is the initial user chosen $\tau^{(0)}$ to be marginally away from the boundaries of $\mathbb{R}$. The second condition in (2.3) requires that the initial value $\tau^{(0)}$ be in an $u_n^{(0)}$-neighborhood of τ_0. While at first, this might come across as a limitation of the algorithm, however the following discussion shall show how broad this $u_n^{(0)}$-neighborhood truly can be. First note that the constant $k \in [1, \infty)$ is arbitrarily large, subject to Condition A(iii), i.e., this condition is adaptable to the user chosen value of the initializer $\tau^{(0)}$. In other words, the farther the user chosen $\tau^{(0)}$ is from the true change point τ_0, the larger the value of k can be, in order to satisfy this condition. Additionally, note that the largest possible distance (in the cdf scale) between any two $\tau_1, \tau_2 \in \mathbb{R}$, is such that $|\Phi(\tau_1) - \Phi(\tau_2)| \le 1$. Now for $c_u = 1$, consider first the disallowed case of $k = \infty$, then the initial condition is trivially satisfied, since $|\Phi(\tau^{(0)}) - \Phi(\tau_0)| \le 1$. Thus, virtually any initial value in the parametric space of τ_0, separated from its boundaries, will satisfy the required initial condition for a large enough $k \in [1, \infty)$, thereby illustrating that this initial condition is infact very mild. Remarkably, one of our main results shows that, under suitable conditions, the updated change point estimate $\hat{\tau}^{(1)}$ of **Step 1** of Algorithm 1, will satisfy optimal error bounds, irrespective of the value of k. Simply stated, the update in **Step 1** sharpens the initial change point guess from any arbitrary fractional rate to a near optimal rate. The condition (2.3), also provides a precise description of the notion of 'fractional information' mentioned in the introduction section. The sequence $u_n^{(0)}$ forms a metric measuring the amount of information in the guess $\tau^{(0)}$ about τ_0, and the existence of a finite $k < \infty$ provides a way of saying that the guess $\tau^{(0)}$ possesses some fractional amount of information on τ_0.

To numerically illustrate this surprising phenomenon, in Section 5 we use the 'no information' initializer $\tau^{(0)} = w^{(0.5)}$, i.e. the 50th percentile of w.[e] Note that this choice is

[d] Note that while the initializing $\tau^{(0)}$ in **Step 0** is chosen in $\mathbb{R}$, however the optimization in **Step 1** is performed over the extended Euclidean space $\bar{\mathbb{R}}^{\star} = \mathbb{R} \cup \{-\infty\}$.

[e] The 50th percentile is the best 'no information' guess, since it is the empirical guess that is equidistant to the ends of the support of w.

the most sensible value of the initializer in the absence of any information about τ_0. These numerical results provide strong evidence to support Algorithm 1 by showing that the precision of the estimates obtained from the proposed method are infact indistinguishable from existing grid search type approaches, and are obtained with a small fraction of the computational burden. Another equivalent way of viewing the above discussion is that, if we pick two distinct initializers $\tau_1^{(0)}$ and $\tau_2^{(0)}$ carrying some fractional information about τ_0, i.e., they satisfy the initializing condition for some $1 \leq k_1 < k_2 < \infty$, respectively ($\tau_1^{(0)}$ is closer to τ_0 than $\tau_2^{(0)}$), then, the corresponding updated change point estimates $\hat{\tau}_1^{(1)}$, $\hat{\tau}_2^{(1)}$ will both be in a near optimal neighborhood of τ_0. This basically implies that the quality of the guess does not influence the updated estimate in its eventual rate of convergence. This observation is also numerically illustrated in Section 5.

Finally, we also mention that a slight relaxation of Condition I is also possible, specifically, the sequence $u_n^{(0)}$ in this condition can instead be relaxed to $u_n^{(0)} = 1 \wedge c_u(\frac{1}{n})^{1/k}$, for any constants $k \in [1, \infty)$ and any $c_u > 0$, while preserving all results to follow. However, Condition I in its current form provides a coherence with other noise bounds that show up in the statement and proof of our results, and thus is a notationally convenient representation. Furthermore, this relaxation of Condition I shall not yield error bounds that are any sharper than those to follow.

A second concern that may arise to the reader regarding implementation of Algorithm 1 is the feasibility of implementing **Step 1**. At first, this optimization seems intractable owing to its nonsmooth, nonconvex (with no apparent convex relaxations) construction. However, upon closer inspection, it is observed that the loss function $Q(\,\cdot\,, \hat{\beta}^{(0)}, \hat{\gamma}^{(0)})$ in **Step 1** is a step function with step changes occurring at any point on the one-dimensional grid $(-\infty, w_1, w_2, ..., w_n)^T$. Secondly, the ℓ_0 term in the objective function only depends on whether $\Phi(\tau)$ is zero or non zero. This implies that the distance between any two τ_1 and τ_2 does not influence the value of the ℓ_0 norm (note that this will not be the case if instead an ℓ_1 norm is used). These two observations together imply that any global optimum achieved in the extended Euclidean space $\bar{\mathbb{R}}^\star$ will also be attained at some point on the finite grid $(-\infty, w_1, w_2, ..., w_n)^T$. An illustration of this step behavior is provided in Figure 1.

A final concern in implementing **Step 1** is that it requires knowledge of the distribution function $\Phi(\cdot)$, which is typically unknown. This concern is also easily overcome upon observing that the objective function in **Step 1** is a step function over the grid $(-\infty, w_1, w_2, ..., w_n)^T$. Specifically, on this grid, the term $\|\Phi(\tau)\|_0 = \|\tau\|_0^*$, where $\|\tau\|_0^* = 1$, if $\tau \in \{w_1, ..., w_n\}$ and $\|\tau\|_0^* = 0$, if $\tau = -\infty$.

In view of the above discussion, **Step 1** of Algorithm 1 can be replaced by the following optimization,

$$\hat{\tau}^{(1)} = \underset{\tau \in \{-\infty\} \cup \{w_1, ..., w_n\}}{\arg\min} \Big\{ Q(\tau, \hat{\beta}^{(0)}, \hat{\gamma}^{(0)}) + \mu \|\tau\|_0^* \Big\}, \quad \mu > 0. \tag{2.5}$$

Thus, the optimization (2.4) in **Step 1** of Algorithm 1 is reduced to the optimization (2.5), which can be easily solved in negligible time by explicitly computing n values of the objective function and then locating the minimum.

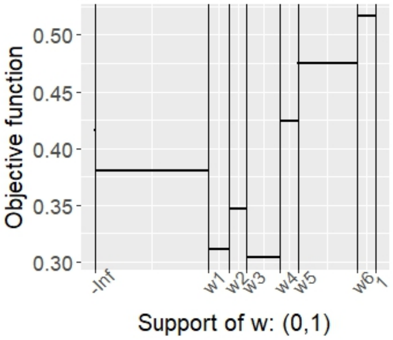

Figure 1: Step behavior of the function $Q(\cdot\,,\hat\beta,\hat\gamma)+\mu\|\Phi(\tau)\|_0$, with $\mu=0.1$, evaluated over grid of points $\{-\infty\}\cup\{0,0.01,...,1\}$. Here $w_i\sim\mathcal{U}(0,1)$, $n=6$, $\tau_0=0.25$, $p=3$, $\beta_0=(1,0,0)^T$, $\gamma_0=(0,1,0)^T$, and we use $\hat\beta=(0.41,0,0)^T$, $\hat\gamma=(0.13,0.92,0)^T$. The realizations $\{w_1,..,w_n\}$ have been sorted (ascending) in the illustration. Observe that step changes occur at $-\{\infty\}\cup\{w_1,w_2,...,w_n\}$.

Another note of interest is the convenience of separability in computing the optimizers $\hat\beta,\hat\gamma$ in **Step 0** and **Step 2**, i.e., for any fixed τ, we can obtain

$$\begin{aligned}\hat\beta(\tau) &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\Big\{\frac{1}{n}\sum_{i;\,w_i\le\tau}(y_i-x_i^T\beta)^2+\lambda_1\|\beta\|_1\Big\}, &(2.6)\\ \hat\gamma(\tau) &= \underset{\beta\in\mathbb{R}^p}{\arg\min}\Big\{\frac{1}{n}\sum_{i;\,w_i>\tau}(y_i-x_i^T\gamma)^2+\lambda_1\|\gamma\|_1\Big\}. &(2.7)\end{aligned}$$

These are ordinary Lasso optimizations and can be carried out by any one of the several methods available in the literature. Some of these methods include, coordinate or gradient descent algorithms (see, e.g. Hastie et al. (2015)), or via interior point methods for linear optimization under second order conic constraints (see, e.g., Koenker and Mizera (2014)).

The main results of this article establish selection consistency of the unknown change point and provide finite sample bounds for the error in estimates obtained from Algorithm 1 under suitable conditions. Let $\xi_n:=\|\beta_0-\gamma_0\|_2$ be the jump size between the pre and post regression parameters. Then the specific results we derive are,

$$\begin{aligned}&(i)\quad \text{If } \Phi(\tau_0)=0,\ \text{ then }\ \Phi(\hat\tau^{(1)})=0, &(2.8)\\ &(ii)\quad \text{If } \Phi_{\min}(\tau_0)\ge c_ul_n,\ \text{ then } |\Phi(\hat\tau^{(1)})-\Phi(\tau_0)|\le t_n:=c_uc_m\max\Big\{\frac{s\log p}{n},\ \frac{1}{(1\vee\xi_n^2)l_n^2}\frac{s\log p}{n}\Big\},\\ &(iii)\quad \big\|\hat\beta^{(1)}-\beta_0\big\|_q\le c_uc_ms^{1/q}\frac{1}{\Phi_{\min}(\tau_0)}\max\Big\{\sqrt{\frac{\log p}{n}},\ \xi_nt_n\Big\},\qquad q=1,2,\\ &(iv)\quad \big\|\hat\gamma^{(1)}-\gamma_0\big\|_q\le c_uc_ms^{1/q}\frac{1}{\Phi_{\min}(\tau_0)}\max\Big\{\sqrt{\frac{\log p}{n}},\ \xi_nt_n\Big\},\qquad q=1,2,\end{aligned}$$

with probability at least $1-c_1\exp(-c_2\log p)$, for n sufficiently large.

These and other related results about estimates from Algorithm 1 are covered in Sections 3 and 4. The sharpness and/or near optimality of the above bounds may be observed from the following special case. Upon letting $\xi_n s\sqrt{\log p/n} \to 0$, and $l_n \ge c_u$, in (2.8), the last three results of (2.8) reduce to,

$$\begin{aligned}
&(i) \quad \text{If } \Phi_{\min}(\tau_0) \ge c_u l_n, \text{ then } |\Phi(\hat{\tau}^{(1)}) - \Phi(\tau_0)| \le c_u c_m \frac{s\log p}{n}, \qquad (2.9)\\
&(ii) \quad \big\|\hat{\beta}^{(1)} - \beta_0\big\|_q \le c_u c_m s^{\frac{1}{q}}\sqrt{\frac{\log p}{n}}, \qquad q = 1, 2,\\
&(iii) \quad \big\|\hat{\gamma}^{(1)} - \gamma_0\big\|_q \le c_u c_m s^{\frac{1}{q}}\sqrt{\frac{\log p}{n}}, \qquad q = 1, 2,
\end{aligned}$$

with probability at least $1 - c_1\exp(-c_2\log p)$, for n sufficiently large.

In an ordinary high dimensional linear regression model without change points, it has been shown that the optimal rate of convergence for regression estimates is $\sqrt{s\log p/n}$ under the ℓ_2 norm (see, e.g.,Ye and Zhang (2010), Raskutti et al. (2011), and Belloni et al. (2017b)). This implies that the rate of convergence of the regression estimates from Algorithm 1 (which stops after one iteration) cannot be uniformly improved upon by carrying out further iterations (over subgaussian distributions). Also, the rate of convergence of the change point estimate in (2.9) is the fastest available rate in the literature. We shall now state the conditions under which the results of this article are derived.

Condition A (assumptions on model parameters):

(i) Let $S = S_1 \cup S_2$, where $S_1 = \{j; \beta_{0j} \ne 0\}$ and $S_2 = \{j; \gamma_{0j} \ne 0\}$. Then for some $s = s_n \ge 1$, we assume that $|S| \le s$.
(ii) The model dimensions s, p, n, satisfy $s\log p/n \to 0$. Additionally, the sequence l_n of Condition I satisfies $s\log p/nl_n^2 \to 0$.
(iii) If a finite change point exists, i.e., $\Phi_{\min}(\tau_0) > 0$, then the sequence l_n and constant $k \in [1, \infty)$ of Condition I satisfy

$$\frac{s}{l_n^2}\Big(\frac{s\log p}{nl_n^2}\Big)^{\frac{1}{k}} \to 0.$$

Additionally, in this case $\Phi_{\min}(\tau_0) \ge c_u l_n$.
(iv) If $\Phi_{\min}(\tau_0) > 0$, then the jump size is bounded below by a constant, i.e, $\xi_n := \|\beta_0 - \gamma_0\|_2 > c > 0$.

Condition B (assumptions on model distributions):

(i) The vectors $x_i = (x_{i1}, ..., x_{ip})^T$, $i = 1, .., n$, are i.i.d subgaussian[f] with mean vector zero, and variance parameter $\sigma_x^2 \le C$. Furthermore, the covariance matrix $\Sigma := Ex_i x_i^T$ has bounded eigenvalues, i.e., $0 < \kappa \le \text{mineigen}(\Sigma) < \text{maxeigen}(\Sigma) \le \phi < \infty$.
(ii) The errors ε_i's are i.i.d. subgaussian with mean zero and variance parameter $\sigma_\varepsilon^2 \le C$.
(iii) The variables w_i, $i = 1, ..., n$ are i.i.d r.v.'s (continuous or discrete), with its cdf $\Phi(a) = P(w_i \le a)$, $a \in \mathbb{R}$.

[f] Recall that for $\alpha > 0$, the random variable η is said to be α-subgaussian if, for all $t \in \mathbb{R}$, $E[\exp(t\eta)] \le \exp(\alpha^2 t^2/2)$. Similarly, a random vector $\xi \in \mathbb{R}^p$ is said to be α-subgaussian if the inner products $\langle \xi, v\rangle$ are α-subgaussian for any $v \in \mathbb{R}^p$ with $\|v\|_2 = 1$.

(iv) The r.v.'s x_i, w_i, ε_i are independent of each other.

Conditions A(i) and A(ii) together form the usual sparsity assumption of high dimensional models. Conditions A(ii) and A(iii) are both restrictions on the model dimensions and in fact A(ii) is implied by A(iii) when it is applicable. However, both conditions are stated here since some of our results in Sections 3 and 4, hold under the weaker Condition A(ii). The Condition A(iii) is on the model parameters and also related to the initial condition of Condition I, via the sequence l_n and the constant $k \in [1, \infty)$. This condition assumes the only additional control on how large a constant k and how small the sequence l_n, can be tolerated by Algorithm 1, given the model dimensions. Heuristically, this condition ensures that the fractional information possessed by the initial guess $\tau^{(0)}$, is not dominated by the noise induced in the linear system due to its large dimensions. Note that Condition A(iii) is only assumed if a change point exists, i.e., $\Phi_{\min}(\tau_0) > 0$. In the case of 'no change' in model (1.1), i.e., $\Phi(\tau_0) = 0$, any initial value $\tau^{(0)}$ satisfying $\Phi(\tau^{(0)}) \geq c_u l_n$, i.e., separated from the boundaries of $\mathbb{R}$, can be used to initialize Algorithm 1. A secondary purpose that Condition A(iii) serves is to ensure that if a finite change point exists, then, to keep $\Phi_{\min}(\tau_0) \geq c_u l_n$ away from the boundaries of $(0, 1)$, whenever a finite change point exists in the model. Note that, this condition does not assume lower boundedness of $\Phi_{\min}(\tau_0)$ as is commonly the case in the literature, since the sequence l_n may converge to zero. Finally, Condition A(iv) requires that if a finite change point exists, then the corresponding jump size ξ_n is bounded below. We also mention here that we do not make any assumptions on the upper bound of ξ_n, and this jump size is allowed to possibly diverge with n.

The subgaussian assumptions in Conditions B(i) and B(ii) are now standard in high dimensional linear regression models and are known to accommodate a large class of random designs. In ordinary high dimensional linear regression, these assumptions are used to establish well behaved restricted eigenvalues of the Gram matrix $\sum x_i x_i^T/n$ (Raskutti et al. (2010), and Rudelson and Zhou (2012)), which are in turn used to derive convergence rates of ℓ_1 regularized estimators (Bickel et al. (2009), and several others). These conditions play a similar role in our change point setup.

One main advantage of the proposed Algorithm 1 over existing methods is its ability to provide near optimal estimates without a grid search. As mentioned earlier in the article, the computational cost of Algorithm 1 is 2Lasso(n, p), significantly below the nLasso(n, p) cost of existing methods and is thus scalable to deal with large data. A novelty of Algorithm 1 in comparison to those proposed in Leonardi and Bühlmann (2016), Lee et al. (2016) and Lee et al. (2018) is its ability to detect the case where $\Phi(\tau_0) = 0$. This is relevant since it removes the necessity to pre-test for the existence of a change point. In contrast, while the methods of Lee et al. (2016), and Lee et al. (2018) are implementable in the case of no change point, they are however unable to detect the absence of the change point. Instead, in this case of $\Phi(\tau_0) = 0$, these methods return a valid $2p$ dimensional estimate $(\hat{\gamma}^T, \hat{\alpha}^T)^T$, where $\alpha_0 = \beta_0 - \gamma_0$, that can be used for predictive purposes using the model (1.1). Note that, the ability to detect the absence of a change point is a stronger statement and may provide additional relevant information, while also preserving the interpretable p dimensional linear regression model in the case where $\Phi(\tau_0) = 0$.

The organization of the remainder of this article is as follows. Section 3 develops preliminary results required for analysis of estimates given by Algorithm 1 and Section 4 provides the main results regarding estimates obtained from Algorithm 1. Proofs of all results are given in Appendix A, while Appendix B consists of some relevant auxiliary results from the literature, stated without proofs. The performance of Algorithm 1 is empirically evaluated in Section 5. In this numerical section, the implementation of Algorithm 1 assumes no prior information of the unknown change point τ_0, additionally we numerically illustrate that the quality of the initial guess has no discernible impact on the final estimates and finally, we also show that the precision of the proposed estimates is indistinguishable from grid search approaches. Section 6 consists of an application of the proposed methodology to socio-economic data of U.S. collected from the 1990 U.S. census, and other sources.

3. Preliminary Results

In this section we present preliminary results that are important for stating and proving our main results in Section 4. First, for any fixed τ, we define

$$\zeta_i(\tau) = \begin{cases} \mathbf{1}[\tau_0 < w_i \le \tau], & \text{if } \tau > \tau_0 \\ \mathbf{1}[\tau \le w_i < \tau_0], & \text{if } \tau \le \tau_0. \end{cases}$$

Then clearly, for any fixed $\tau \in \mathbb{R}$, $E\zeta_i(\tau) := \Phi^*(\tau_0, \tau) = |\Phi(\tau) - \Phi(\tau_0)|$. We shall now state a key result that uniformly controls (over τ) the quantity $n^{-1}\sum_{i=1}^n \zeta_i(\tau)$.

Lemma 3.1 *Let u_n, and v_n be any non-negative sequences such that $v_n \ge c\log p/n$, $c > 0$ and let $\mathcal{T}(\tau_0, u_n) = \big\{\tau :\ \Phi^*(\tau_0,\tau) \le u_n\big\}$ be a u_n-neighborhood of τ_0. Then under Condition B(iii), we have,*

$$(i)\ \sup_{\tau\in\mathcal{T}(\tau_0,u_n)} \frac{1}{n}\sum_{i=1}^n \zeta_i(\tau) \le c_u \max\Big\{\frac{\log p}{n}, u_n\Big\}, \qquad (ii)\ \inf_{\substack{\tau\in\mathbb{R};\\ \Phi^*(\tau_0,\tau)\ge v_n}} \frac{1}{n}\sum_{i=1}^n \zeta_i(\tau) \ge c_u v_n,$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

To proceed further, define for any τ the following set of random indices,

$$n_w := n_w(\tau_0,\tau) = \begin{cases} i \in \{1,...,n\}; & \tau_0 < w_i \le \tau, \quad \text{if } \tau \ge \tau_0, \\ i \in \{1,...,n\}; & \tau \le w_i < \tau_0, \quad \text{if } \tau \le \tau_0. \end{cases} \tag{3.1}$$

Note that the cardinality of the random set n_w is precisely the stochastic term controlled in Lemma 3.1, i.e., $|n_w| = \sum_{i=1}^n \zeta_i(\tau)$. This relation serves to provide bounds on several other stochastic terms considered in subsequent lemmas. The relationship between the cardinality of the random index set n_w and the r.v.'s $\zeta_i(\tau)$, $i = 1,...,n$ has also been used by Kaul et al. (2017) in the context of graphical models with missing data.

Lemma 3.2 *Let u_n be any non-negative sequence and let n_w be the random set of indices as defined in (3.1). Also, let $\mathcal{T}(\tau_0,u_n)$ be a u_n-neighborhood of τ_0 as defined in Lemma 3.1.*

Then under Condition B, we have for any fixed $\delta \in \mathbb{R}^p$ *that,*

$$\begin{aligned}
&(i) \qquad \sup_{\tau\in\mathcal{T}(\tau_0,u_n)} \Big\|\frac{1}{n}\sum_{i\in n_w}\delta^T x_i x_i^T\Big\|_\infty \le c_u c_{m1}\|\delta\|_2 \max\Big\{\frac{\log p}{n},\, u_n\Big\},\\
&(ii) \qquad \sup_{\tau\in\mathcal{T}(\tau_0,u_n)} \frac{1}{n}\sum_{i\in n_w}\delta^T x_i x_i^T\delta \le c_u c_{m1}\|\delta\|_2^2 \max\Big\{\frac{\log p}{n},\, u_n\Big\},\\
&(iii) \qquad \sup_{\tau\in\mathcal{T}(\tau_0,u_n)} \frac{1}{n}\|\sum_{i\in n_w}\varepsilon_i x_i^T\|_\infty \le c_u c_{m2}\sqrt{\frac{\log p}{n}} \max\Big\{\sqrt{\frac{\log p}{n}},\, \sqrt{u_n}\Big\},
\end{aligned}$$

with probability at least $1 - c_1\exp(-c_2\log p)$. *Here* $c_u > 0$ *is a universal constant, and* $c_{m1} = (\phi + \sigma_x + \sigma_x^2)$, $c_{m2} = (\sqrt{\sigma_\varepsilon\sigma_x} + \sigma_\varepsilon\sigma_x)$ *are model constants.*

Finally in order to obtain the desired error bounds (2.8) and (2.9) we require restricted eigenvalue conditions on the gram matrix $\sum_{i=1}^n x_i x_i^T$. For any deterministic set $S \subset \{1,2,...,p\}$, define the collection $\mathbb{A}$ as,

$$\mathbb{A} = \Big\{\delta \in \mathbb{R}^p;\ \|\delta_{S^c}\|_1 \le 3\|\delta_S\|_1,\Big\}. \tag{3.2}$$

Then, Bickel et al. (2009) define the lower restricted eigenvalue condition as,

$$\inf_{\delta\in\mathbb{A}} \frac{1}{n}\sum_{i=1}^n \delta^T x_i x_i^T\delta \ge c_u\kappa\|\delta\|_2^2, \quad \text{for some constant } \kappa > 0. \tag{3.3}$$

Other slightly weaker versions of this condition are also available in the literature such as the compatibility condition of Bühlmann and Van De Geer (2011), and the ℓ_q sensitivity of Gautier and Tsybakov (2011). In the setup of common random designs, it is also well established that condition (3.3) holds with probability converging to 1, see for e.g. Raskutti et al. (2010), and Rudelson and Zhou (2012), for Gaussian designs. In the subgaussian case, the plausibility of this condition is a consequence of a general result stated as Lemma B.2 in Appendix B. Under our high dimensional change point setup, we shall require versions of the restricted eigenvalue condition (3.3). In the following lemma, we shall show that all required conditions hold with probability converging to 1. Among other arguments, the proof of these conditions shall rely on Lemma B.2. In Lemma 3.3 below, the collection $\mathbb{A}$ in (3.2) applies for the set S in Condition A.

Lemma 3.3 (Restricted Eigenvalue Conditions): *Let* u_n, *and* v_n *be any non-negative sequences such that* $v_n \ge c\log p/n$, $c > 0$. *Let* $\mathcal{T}(\tau_0, u_n)$ *be as in Lemma 3.1 and the set* $\mathbb{A}$ *as defined in (3.2) for* S *given in Condition A. Furthermore, define the set* $\mathbb{A}_2 = \Big\{\delta \in \mathbb{R}^p; \|\delta_{S^c}\|_1 \le 3\|\delta_S\|_1 + 3\|\beta_0 - \gamma_0\|_1\Big\}$, *and let any* $\tau \in \mathbb{R}$ *be such that* $\Phi_{\min}^{-1}(\tau)s\log p/n = o(1)$. *Then under Conditions A(i), A(ii), and B, and for* n *sufficiently large, the following re-*

stricted eigenvalue conditions hold with probability at least $1-c_1\exp(-c_2\log p)$,

$$\begin{aligned}
&(i) && \inf_{\delta\in\mathbb{A}}\frac{1}{n}\sum_{i;w_i\le\tau}\delta^T x_i x_i^T\delta \ge c_u\kappa\Phi(\tau)\|\delta\|_2^2,\\
&(ii) && \inf_{\delta\in\mathbb{A}}\frac{1}{n}\sum_{i;w_i>\tau}\delta^T x_i x_i^T\delta \ge c_u\kappa(1-\Phi(\tau))\|\delta\|_2^2,\\
&(iii) && \sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\sup_{\delta\in\mathbb{A}}\frac{1}{n}\sum_{i\in n_w}\delta^T x_i x_i^T\delta \le c_u c_m\|\delta\|_2^2\max\Big\{\frac{s\log p}{n}, u_n\Big\},\\
&(iv) && \inf_{\substack{\tau\in\mathbb{R};\\ \Phi^*(\tau_0,\tau)\ge v_n}}\inf_{\delta\in\mathbb{A}_2}\frac{1}{n}\sum_{i\in n_w}\delta^T x_i x_i^T\delta \ge c_u c_m v_n\|\delta\|_2^2 - c_u c_m\frac{s\log p}{n}\Big(\|\delta\|_2^2+\xi_n^2\Big).
\end{aligned}$$

Before moving on to state our main results in the next section, we make the following remark regarding the role of the set $\mathbb{A}_2$.

Remark 3.1 Note that if $\beta-\beta_0\in\mathbb{A}$, i.e., $\|\beta_{S^c}-\beta_{0S^c}\|_1\le 3\|\beta_S-\beta_{0S}\|_1$, then for $\delta=\beta_0-\gamma_0+\beta-\beta_0$, we have,

$$\|\delta_{S^c}\|_1\le\|\beta_{S^c}-\beta_{0S^c}\|_1\le 3\|\beta_S-\beta_{0S}\|_1\le 3\|\delta_S\|_1+3\|\beta_0-\gamma_0\|_1.$$

Thus $\beta-\beta_0\in\mathbb{A}$, implies the vector $\delta\in\mathbb{A}_2$. This relation is useful in proving Lemma 4.1 and Theorem 4.2 of the next section.

4. Main Results

We are now ready to state our first main result pertaining to the rate of convergence of the regression estimates obtained from (2.6) and (2.7) when τ(possibly random) is in a u_n-neighborhood of τ_0.

Theorem 4.1 *Suppose Conditions A(i), A(ii), and B hold, and consider any* $\tau\in\mathbb{R}$, *satisfying* $\Phi_{\min}^{-1}(\tau)s\log p/n=o(1)$. *Let* $\hat\beta(\tau)$ *and* $\hat\gamma(\tau)$ *be solutions to (2.6) and (2.7). Then,*
(i) When $\Phi(\tau_0)=0$ *and* $\lambda_1\ge c_u c_m\sqrt{\log p/n}$, *for n sufficiently large, we have for* $q=1,2$,

$$\|\hat\beta(\tau)-\gamma_0\|_q\le c_u c_m\frac{1}{\Phi_{\min}(\tau)}s^{1/q}\sqrt{\frac{\log p}{n}},$$

with probability at least $1-c_1\exp(-c_2\log p)$. *The same bound holds for* $\|\hat\gamma(\tau)-\gamma_0\|_q$, $q=1,2$.
(ii) When $\Phi_{\min}(\tau_0)>0$, *let* u_n *be any non-negative sequence satisfying* $u_n=o\big(\Phi_{\min}(\tau_0)\big)$. *Also, let* $\mathcal{T}(\tau_0,u_n)$ *be as defined in Lemma 3.1 and* $\lambda_1=c_u c_m\max\{\sqrt{\log p/n},\ \xi_n u_n\}$. *Then for n sufficiently large, and* $q=1,2$, *the following uniform bound holds,*

$$\sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\|\hat\beta(\tau)-\beta_0\|_q\le c_u c_m\frac{1}{\Phi_{\min}(\tau_0)}s^{1/q}\max\Big\{\sqrt{\frac{\log p}{n}},\|\beta_0-\gamma_0\|_2 u_n\Big\},$$

with probability at least $1-c_1\exp(-c_2\log p)$. *The same uniform upper bound also holds for* $\sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\|\hat\gamma(\tau)-\gamma_0\|_q$, $q=1,2$.

Remark 4.1 As a direct consequence of Theorem 4.1, we obtain the rates of convergence of the regression estimates $\hat{\beta}^{(0)}, \hat{\gamma}^{(0)}$ from **Step 0** of **Algorithm 1**. Specifically, under the conditions of Theorem 4.1,
(i) When $\Phi_{\min}(\tau_0) = 0$,

$$\|\hat{\beta}^{(0)} - \gamma_0\|_q \le c_u c_m \frac{1}{\Phi_{\min}(\tau^{(0)})} s^{1/q} \sqrt{\frac{\log p}{n}}, \qquad q = 1, 2,$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. The same bound holds for $\|\hat{\gamma}^{(0)} - \gamma_0\|_q$, $q = 1, 2$.
(ii) When $\Phi_{\min}(\tau_0) > 0$, and $\Big|\Phi(\tau^{(0)}) - \Phi(\tau_0)\Big| \le u_n^{(0)}$, we have,

$$\|\hat{\beta}^{(0)} - \beta_0\|_q \le c_u c_m \frac{1}{\Phi_{\min}(\tau_0)} s^{1/q} \max\Big\{\sqrt{\frac{\log p}{n}},\ \xi_n u_n^{(0)}\Big\}, \qquad q = 1, 2, \tag{4.1}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. The same bound holds for $\|\hat{\gamma}^{(0)} - \gamma_0\|_q$, $q = 1, 2$. In this case, since $s^{1/q}\xi_n u_n^{(0)} \Big/ \Phi(\tau_0)$ may diverge, these estimates are not guaranteed to be consistent. Nevertheless, (i) and (ii) above play an important role in deriving convergence rates of estimators from subsequent steps of Algorithm 1.

We now turn our attention to establishing selection and estimation results for estimates obtained from **Step 1** and **Step 2** of Algorithm 1. To achieve this goal, we require the following notations. For any $\tau \in \mathbb{R}$, $\beta, \gamma \in \mathbb{R}^p$, let

$$\begin{aligned} R_n(\tau, \beta, \gamma) &= Q(\tau, \beta, \gamma) - Q(\tau_0, \beta, \gamma). \\ S_n(\tau, \beta, \gamma) &= R_n(\tau, \beta, \gamma) + \mu\big(\|\Phi(\tau)\|_0 - \|\Phi(\tau_0)\|_0\big) \end{aligned}$$

Also, for any non-negative u_n, and v_n, define the collection

$$\mathcal{H}(u_n, v_n) = \big\{\tau \in \mathbb{R};\ v_n \le |\Phi(\tau) - \Phi(\tau_0)| \le u_n\big\}$$

Additionally, for any non-negative sequence u_n, we also define the function,

$$F(u_n) = \begin{cases} 0 & \text{if } u_n/\Phi_{\min}(\tau_0) \to 0 \\ 1 & \text{otherwise} \end{cases}. \tag{4.2}$$

Finally, in the following, we denote by $r_n := \max\big\{\sqrt{s\log p/n}, \sqrt{s}\xi_n u_n^{(0)}\big\}\big/\Phi_{\min}(\tau_0)$, in the case where $\Phi_{\min}(\tau_0) > 0$. Notice that r_n is the ℓ_2 rate of estimation error provided in Part(ii) of Remark 4.1. The following lemma provides a uniform lower bound of the expression $S_n(\tau, \beta, \gamma)$, over the collection $\mathcal{H}(u_n, v_n)$, that holds with high probability. This result shall lie at the heart of the argument used to obtain the main results of this article.

Lemma 4.1 *Suppose conditions A and B hold and let u_n be any non negative sequence. Also, let $\hat{\beta}^{(0)}$, $\hat{\gamma}^{(0)}$ be estimates from* **Step 0** *of Algorithm 1. Then,* *(i) When* $\Phi(\tau_0)=0$, *for any* $v_n>0$, *we have*

$$\inf_{\tau\in\mathcal{H}(1,v_n)} S_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})\geq \mu-c_uc_m\frac{s\log p}{n\Phi^2_{\min}(\tau^{(0)})}$$

with probability at least $1-c_1\exp(-c_2\log p)$.

(ii) When $\Phi_{\min}(\tau_0)>0$, *for any* $v_n\geq c\log p/n$, $c>0$, *we have,*

$$\inf_{\tau\in\mathcal{H}(u_n,v_n)} S_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})\geq \xi_n^2\Big(c_uc_mv_n-c_uc_m\frac{s\log p}{n}-\frac{c_uc_m}{1\vee\xi_n}\sqrt{\frac{s\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\sqrt{u_n}\Big\}$$
$$-c_uc_m\frac{r_n^2}{1\vee\xi_n^2}\max\Big\{\frac{s\log p}{n},\,u_n\Big\}-\frac{c_u\mu}{1\vee\xi_n^2}F(u_n)\Big).$$

with probability at least $1-c_1\exp(-c_2\log p)$.

Our main result on rate of convergence of the estimates obtained from **Algorithm 1** is stated in Theorem 4.2 below. While the complete proof of the theorem is given in the appendix, here we provide a sketch of the main idea behind the proof. We show that, for an appropriately chosen regularizer μ, for any $v_n>0$ (in the case where $\Phi(\tau_0)=0$), or for any non-negative sequence v_n slower in rate than those given in (2.8) (in the case where $\Phi_{\min}(\tau_0)>0$), we shall show that,

$$\inf_{\tau;\,v_n\leq\Phi^*(\tau_0,\tau)\leq 1} S_n\big(\tau,\hat\beta^{(0)},\hat\gamma^{(0)}\big)>0,\ \text{for } n \text{ sufficiently large.}$$

with probability at least $1-c_1\exp(-c_2\log p)$. Upon noting that the global optimizer $\hat\tau^{(1)}$ by definition satisfies $S_n(\hat\tau^{(1)},\hat\beta^{(0)},\hat\gamma^{(0)})\leq 0$, we would have shown that the corresponding global optimizer $\hat\tau^{(1)}$ satisfies the relations given in (2.8). Along the way a sequence of recursions are required in order to sequentially sharpen the bound for the change point estimate. Supportive arguments are also required to show that the eventual bound is satisfied with probability at least $1-c_1\exp(-c_2\log p)$. In this process, Remark .1 is quite helpful.

Theorem 4.2 *Suppose Conditions A and B hold and choose* $\lambda_1=c_uc_m\max\{\sqrt{\log p/n},\,\xi_nu_n^{(0)}\}$, *and* $\mu=c_uc_m\big(s\log p/nl_n^2\big)^{1/k^*}$, *where* $k^*=\max\{k,2\}$. *Then for n sufficiently large, the optimizer* $\hat\tau^{(1)}$ *of* **Step 1** *of Algorithm 1 satisfies the following relations.*

(i) When $\Phi(\tau_0)=0$, *then* $\Phi(\hat\tau^{(1)})=0$, *with probability at least* $1-c_1\exp(-c_2\log p)$.

(ii) When $\Phi_{\min}(\tau_0)\geq c_ul_n$, *then,*

$$|\Phi(\hat\tau^{(1)})-\Phi(\tau_0)|\leq t_n:=c_uc_m\max\Big\{\frac{s\log p}{n},\frac{1}{(1\vee\xi_n^2)l_n^2}\frac{s\log p}{n}\Big\},$$

with probability at least $1-c_1\exp(-c_2\log p)$.

The usefulness of Theorem 4.2 is apparent. Despite initializing Algorithm 1 with a $\tau^{(0)}$, which is in an $(s\log p/nl_n^2)^{1/k}$ neighborhood of τ_0, for an nearly arbitrary $k\in[1,\infty)$. (any initial value that posses 'fractional information' of τ_0), the updated change point $\hat\tau^{(1)}$ lies in a near optimal neighborhood of τ_0, irrespective of the value of k (irrespective of the precision of the initial guess). The following theorem provides the rates of convergence of the regression coefficient estimates $\hat\beta^{(1)}$ and $\hat\gamma^{(1)}$ obtained from **Step 2** of Algorithm 1.

Theorem 4.3 *Suppose the model (1.1) in the case where a finite change point exists, i.e., $\Phi_{\min}(\tau_0) > 0$. Assume the conditions of Theorem 4.2 and choose $\lambda_2 = c_u c_m \max\big\{\sqrt{\log p/n}, \xi_n t_n\big\}$, where t_n is as defined in Theorem 4.2. Then, the estimates $\hat\beta^{(1)}$ and $\hat\gamma^{(1)}$ of* **Step 2** *of Algorithm 1 satisfy,*

$$(i) \qquad \|\hat\beta^{(1)} - \beta_0\|_q \le c_u c_m s^{1/q} \frac{1}{\Phi_{\min}(\tau_0)} \max\Big\{\sqrt{\frac{\log p}{n}},\ \xi_n t_n\Big\}, \qquad q = 1, 2,$$

$$(ii) \qquad \|\hat\gamma^{(1)} - \gamma_0\|_q \le c_u c_m s^{1/q} \frac{1}{\Phi_{\min}(\tau_0)} \max\Big\{\sqrt{\frac{\log p}{n}},\ \xi_n t_n\Big\}, \qquad q = 1, 2,$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

Remark 4.2 (Interpretation of the rates of Theorem 4.2 and Theorem 4.3) Note that under conditions of Theorem 4.2, and additionally assuming that $\xi_n l_n \ge c_1 > 0$, we have,

$$|\Phi(\hat\tau^{(1)}) - \Phi(\tau_0)| \le c_u c_m \frac{s\log p}{n},$$

with probability at least $1 - c_1\exp(-c_2\log p)$. This observation provides an intuitive statement, saying that an increasing jump size ξ_n can compensate for the location of the unknown change point moving toward the boundaries of $\mathbb{R}$, in effect allowing $\hat\tau^{(1)}$ of Algorithm 1 to approximate the unknown change point (if it exists) at a near optimal rate. In this case, under conditions of Theorem 4.3, the regression estimates of **Step 2** become,

$$\|\hat\beta^{(1)} - \beta_0\|_q \le c_u c_m s^{1/q} \frac{1}{\Phi_{\min}(\tau_0)} \max\Big\{\sqrt{\frac{\log p}{n}},\ \xi_n\frac{s\log p}{n}\Big\}, \qquad q = 1, 2, \tag{4.3}$$

with probability at least $1 - c_1\exp(-c_2\log p)$. The same bound also holds for $\|\hat\gamma^{(1)} - \gamma_0\|_q$, $q = 1, 2$, with the same probability. The bound in the statement (4.3) again provides an interesting observation, where an increasing jump is leading to potentially counteract the precision of the regression estimate $\hat\beta^{(1)}$. First note that, the bound in (4.3) can tolerate an increasing jump size ξ_n such that $\xi_n s\sqrt{\log p/n} \to 0$, while preserving optimality of the rate of convergence, i.e, yielding, $\|\hat\beta^{(1)} - \beta_0\|_2 \le c_u c_m \Phi^{-1}(\tau_0)\sqrt{s\log p/n}$, with high probability. It is only when ξ_n increases faster than $\sqrt{n/s^2\log p}$ that it begins to harm the rate of convergence. This observation is surprising in that it suggests that an increasing jump size always benefits the change point estimate, whereas it benefits the regression coefficient estimates only when the jump size is increasing upto a certain rate. To the best of our knowledge, such a characterization of the effect of the jump size on parameter estimates, which holds only in the high dimensional case, has not been provided in the literature. The illustration in Figure 2 provides an intuitive understanding of this behavior,

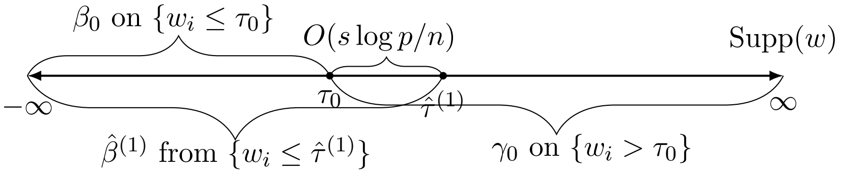
β_0 on $\{w_i \le \tau_0\}$
$O(s\log p/n)$
Supp(w)
$-\infty$
τ_0
$\hat\tau^{(1)}$
∞
$\hat\beta^{(1)}$ from $\{w_i \le \hat\tau^{(1)}\}$
γ_0 on $\{w_i > \tau_0\}$

Figure 2: Illustration of counteracting effect of jump size ξ_n on regression coefficient estimates

From Figure 2, observe that for any finite jump size, the best approximation that our analysis can provide is wherein the error is of order $s\log p/n$.[g] Now the regression estimates of **Step 2** are computed based on the binary partition yielded by the change point estimate $\hat\tau^{(1)}$ of **Step 1**. Consequently, the data based on which the regression estimate $\hat\beta^{(1)}$ of β_0 is obtained, may be corrupted by as much as a fraction $O(s\log p/n)$ of observations where the true regression coefficient is γ_0. Thereby, the higher the jump size ξ_n, the more impact this small corruption will have on the estimate $\hat\beta^{(1)}$. The same argument also holds for the other binary partition. This provides an explanation of the rates observed in Theorem 4.3.

5. Implementation and Numerical Results

The three main objectives of this empirical study are, (i) to evaluate the overall performance of Algorithm 1, i.e., its ability to consistently estimate β_0, γ_0, and a finite τ_0, and compare its performance to a full grid search approach, (ii) to numerically support the theoretically claimed statement, that the estimate $\hat\tau^{(1)}$ is insensitive to the quality of the initial guess $\tau^{(0)}$, and (iii) to evaluate the numerical performance of Algorithm 1 in detecting the 'no change' case, i.e., when $\Phi(\tau_0)=0$.

5.1. Simulation setup

We consider the data generating process (1.1) where ε_i, w_i and x_i are drawn independently satisfying $\varepsilon_i\sim\mathcal{N}(0,\sigma_\varepsilon^2)$, $w_i\sim\mathcal{U}(0,1)$,[h] and $x_i\sim\mathcal{N}(0,\Sigma)$. Here, Σ is a $p\times p$ matrix with elements $\Sigma_{ij}=\rho^{|i-j|}$, $i,j=1,...,p$. We set, $\sigma_\varepsilon=1$ and $\rho=0.5$. The regression parameters of the model are set to be $\beta_0=(1,1,1,1,...,0)^T_{p\times 1}$, and $\gamma_0=(0_{1\times 4},1,1,1,1,0,...,0)^T_{p\times1}$. The metrics of interest are bias and mean squared error of various estimates: For numerical experiments where $\Phi_{\min}(\tau_0)>0$, bias$(\hat\beta)=\|\mathrm{E}(\hat\beta-\beta_0)\|_2$, bias$(\hat\gamma)=\|\mathrm{E}(\hat\gamma-\gamma_0)\|_2$, bias$(\hat\tau)=|\mathrm{E}(\hat\tau-\tau_0)|$, mse$(\hat\beta)=\|\mathrm{E}(\hat\beta-\beta_0)^2\|_2$, mse$(\hat\gamma)=\|\mathrm{E}(\hat\gamma-\gamma_0)^2\|_2$, mse$(\hat\tau)=\mathrm{E}(\hat\tau-\tau_0)^2$, mse$(\Phi(\hat\tau))=\mathrm{E}(\Phi(\hat\tau)-\Phi(\tau_0))^2$. For numerical experiments where $\Phi(\tau_0)=0$, we report $PrM=E(\mathbf{1}[\hat\tau^{(1)}=-\infty])$, i.e., the proportion of times where the 'no change' model is correctly identified. We shall report monte carlo approximations of these metrics based on 100 replications for each combination of model parameters. In the simulations where a finite change point, i.e., $0<\tau_0<1$ is misidentified as 'no change point', i.e., $\hat\tau^{(1)}=-\infty$, $\big($observed to occur sometimes when τ_0 is near the boundaries of $(0,1)$ $\big)$, we do the following operation to maintain fairness of comparisons of the above metrics. In case where $\tau_0<0.5$ and $\hat\tau^{(1)}=-\infty$, then we set $\hat\tau^{(1)}=0$, $\hat\beta^{(1)}=0_{p\times1}$, and when $\tau_0>0.5$ and $\hat\tau^{(1)}=-\infty$, we set $\hat\tau^{(1)}=1$, $\hat\gamma^{(1)}=0_{p\times1}$. Finally, we also report the metric *time*: the average (over replications) computation time [i]. All computations are performed in the software R, R Core Team (2017). All lasso optimizations are performed with the R package 'glmnet', developed by Friedman et al. (2010). We perform two sets of simulations for all combinations of the parameters $n\in\{150,250,350\}$, $p\in\{25,150,250\}$. In the first simulation, we consider finite change points, with $\tau_0\in\{0.169,0.264,...,0.831\}$, (Equally spaced grid of 8 points between 0.169 to 0.831). This is referred to as **Simulation A** in the following. The second simulation considers the case of 'no change' in the model (1.1), i.e., $\tau_0=0$. This simulation is referred

[g] This rate matches the fastest available in the literature

[h] Since $w_i\sim\mathcal{U}(0,1)$, hence $\Phi(\tau)=\tau$, $\tau\in(0,1)$.

[i] CPU: Intel Xeon E5-2609 v3 @ 1.9 GHz, RAM: 128 GB

to as **Simulation B** in the following. Due to the absence of any comparative method that is able to detect the 'no change' case (to the best of our knowledge), we report only the results of our method for this simulation. Note that for each fixed p, the total number of model parameters to be estimated is $2p+1$.

Choice of tuning parameters: The regularizer λ_1 and λ_2 of the Lasso optimizations of **Step 0** and **Step 2** of Algorithm 1 are chosen via a 5-fold cross validation, which is performed internally by the R package 'glmnet'. The regularizer μ of **Step 1** of Algorithm 1 is chosen via the classical BIC criteria. Specifically, $\hat{\tau}(\mu)$ is computed over a grid of values of μ. Then, the value of μ of that minimizes the criteria,

$$\text{BIC}(\mu) = \log\Big(Q\big(\hat{\tau}(\mu), \hat{\beta}(\mu), \hat{\gamma}(\mu)\big)\Big) + \frac{\log n}{n}\|\Phi\big(\hat{\tau}(\mu)\big)\|_0,$$

is chosen. Here $Q(\cdot,\cdot,\cdot)$ is the least squares loss, as defined in (2.1), and $\hat{\beta}(\mu)$, $\hat{\gamma}(\mu)$ represent regression coefficient estimates obtained on the binary partition given by $\hat{\tau}(\mu)$.

In the following, we consider two schemes to choose the initializer $\tau^{(0)}$ of **Algorithm 1**. The first is to set to $w^{(0.5)}$, i.e., the 0.5^{th} empirical quantile of $w=(w_1,..,w_n)^T$. This is done to make the initializer equidistant from the two extremes of the support of w. Note that, in the absence of any information on the unknown τ_0, the choice $\tau^{(0)}=w^{(0.5)}$ is a sensible choice for the initializer. This approach is represented as '**Algorithm 1A**'. As a second scheme, we choose the initializer $\tau^{(0)}$ by setting it to one of values $\{w^{(m)};\, m=0.25, 0.50, 0.75\}$, where $w^{(m)}$ represents the m^{th} empirical quantile of $w=(w_1,...,w_n)^T$. This is done by first computing $\hat{\beta}(\tau)$ and $\hat{\gamma}(\tau)$ in (2.6) and (2.7) for each $\tau = w^{(0.25)}, w^{(0.50)}, w^{(0.75)}$, and finally selecting $\tau^{(0)}$ as the value that minimizes the least squares loss over these three choices. Note that the latter approach has an additional computational burden of two Lasso(n,p) optimizations in comparison to the former. This approach is represented as '**Algorithm 1B**'. Clearly, the initializer in Algorithm 1B will be a closer value to the unknown τ_0 in comparison to the initializer of Algorithm 1A. This shall also help us numerically support our theoretical finding that Algorithm 1 is insensitive to the 'quality' of the initializer. Finally, we also implement the full grid search approach of Lee et al. (2016) in order to serve as a benchmark to compare the performance of the proposed estimates and also to illustrate the dramatic gains in computation time provided by our method. This approach is referred to as **Full grid search** in the following. For completeness, the **Full grid search** estimator of Lee et al. (2016) is described in the notation of this article in the following. The article of Lee et al. (2016) assumes the model $y_i = x_i\delta_0 + x_i^T\eta_0\mathbf{1}[w_i\le\tau_0]$, which is equivalent to the model (1.1) when $\delta_0=\gamma_0$ and $\eta_0=\beta_0-\gamma_0$. Now, let $\tilde{x}_i(\tau) = \big(x_i^T, x_i^T\mathbf{1}(w_i\le\tau)\big)^T_{2p\times 1}$, and the parameter $\alpha=(\gamma, \beta-\gamma)$, where β_0, γ_0 are the true parameter coefficients of the model (1.1), then

$$\begin{aligned}
\hat{\alpha}(\tau) &= \mathop{\arg\min}_{\alpha\in\mathbb{R}^{2p}}\Big\{\frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{x}_i^T(\tau)\alpha)^2 + \lambda\|D(\tau)\alpha\|_1\Big\}, \quad \text{for each } \tau\in\mathcal{T}^*,\\
\hat{\tau} &= \mathop{\arg\min}_{\tau\in\mathcal{T}^*}\Big\{\frac{1}{n}\sum_{i=1}^{n}\big(y_i - \tilde{x}_i^T(\tau)\hat{\alpha}(\tau)\big)^2 + \lambda\|D(\tau)\hat{\alpha}(\tau)\|_1\Big\} \qquad (5.1)
\end{aligned}$$

Table 1: Numerical results of **Algorithm 1A and 1B**, and **Full grid search** for $n \in \{150, 250, 350\}$, $p \in \{25, 150, 250\}$, and $\tau_0 = 0.169$.

Method	**n**	**p**	**bias**$(\hat{\beta})$	**bias**$(\hat{\gamma})$	**bias**$(\hat{\tau})$	**mse**$(\hat{\beta})$	**mse**$(\hat{\gamma})$	**mse**$(\hat{\tau})$	**time**
Algorithm 1A	150	25	0.6595	0.3026	0.0031	0.4564	0.0858	0.0003	0.5792
	150	150	1.5638	0.3530	0.0067	1.4932	0.1044	0.0006	0.8684
	150	250	1.4271	0.3965	0.0193	1.2918	0.1336	0.0021	0.9606
	250	25	0.4354	0.2444	0.0006	0.2135	0.0498	0.0001	0.6068
	250	150	0.5694	0.2717	0.0045	0.2877	0.0599	0.0001	1.3090
	250	250	0.6600	0.2958	0.0015	0.3836	0.0650	0.0001	1.3468
	350	25	0.4193	0.2188	0.0068	0.1758	0.0369	0.0001	0.6515
	350	150	0.5285	0.2362	0.0023	0.2325	0.0415	0.0000	1.5462
	350	250	0.8964	0.2504	0.0028	0.6499	0.0449	0.0001	2.6849
Algorithm 1B	150	25	0.6758	0.2918	0.0030	0.4724	0.0827	0.0002	0.9387
	150	150	1.5803	0.3880	0.0063	1.5061	0.1372	0.0111	1.3351
	150	250	1.4343	0.4167	0.0002	1.3142	0.1648	0.0136	1.4999
	250	25	0.4040	0.2370	0.0020	0.1964	0.0481	0.0001	0.9684
	250	150	0.5666	0.2645	0.0036	0.2779	0.0564	0.0001	2.2686
	250	250	0.6779	0.2961	0.0023	0.4021	0.0648	0.0001	2.1075
	350	25	0.4034	0.2154	0.0081	0.1661	0.0358	0.0001	1.0306
	350	150	0.5209	0.2293	0.0034	0.2228	0.0401	0.0001	2.4129
	350	250	0.8761	0.2459	0.0041	0.6393	0.0442	0.0001	3.5468
Full grid search	150	25	0.7450	0.2240	0.0193	0.5763	0.0612	0.0007	12.7566
	150	150	2.0766	0.4225	0.0531	1.9157	0.1313	0.0068	25.8863
	150	250	2.0300	0.4732	0.0287	1.8209	0.1650	0.0057	31.5806
	250	25	0.5764	0.1750	0.0098	0.3359	0.0341	0.0002	21.0164
	250	150	1.1066	0.2894	0.0023	0.7863	0.0640	0.0002	59.7864
	250	250	1.2875	0.3318	0.0127	0.9927	0.0758	0.0012	88.9717
	350	25	0.5036	0.1588	0.0005	0.2270	0.0238	0.0001	29.4869
	350	150	1.0025	0.2201	0.0057	0.6845	0.0373	0.0006	113.4715
	350	250	1.6075	0.2627	0.0063	1.3631	0.0477	0.0002	144.6366

where $D(\tau) = \text{diag}\big\{\|\tilde{x}^{(j)}(\tau)\|_n,\, j = 1, ..., 2p\big\}$, with $\tilde{x}^{(j)}(\tau)$ representing the j^{th} column of the design matrix $\tilde{x}(\tau) = \big(\tilde{x}_1(\tau), ..., \tilde{x}_n(\tau)\big)^T_{n\times 2p}$. In implementation of this estimator, the search space of the change point is restricted to $\tau \in \mathcal{T}^* = \{w_1, ..., w_n\} \cap (0.1, 0.9)$.

5.2. Results and discussion

The bias and mean squared error (mse) of estimates obtained from **Algorithms 1A, 1B**, and **Full grid search** for **Simulation A**, for all combinations of $n \in \{150, 200, 250\}$, $p \in \{25, 150, 250\}$, and $\Phi(\tau_0) \in \{0.169, 0.67\}$ are presented in Table 1 and Table 2. All results provided are truncated at 10^{-4}. Complete results of the simulation study including all cases for $\tau_0 \in \{0.169, 0.264, ..., 0.831\}$, are available in the supplementary materials of this article. To aid in interpretation, the results on bias are illustrated through Figures 3, 4, 5, 6 and 7. In particular, Figure 3 illustrates bias associated with the change point estimate, Figure 4 and Figure 5 illustrate the bias in $\hat{\beta}^{(1)}$, and $\hat{\gamma}^{(1)}$ respectively. In Figure 6 we illustrate the consistency of the proposed methodology and finally, in Figure 7 we depict the average computation time for the methods implemented in this simulation study. The results of **Simulation B** are reported in Table 3. This table reports the proportion of times the 'no change' model is correctly identified via the metric PrM as described above.

Simulation A: Two important observations from the bias results for the change point estimate depicted in Figures 3 are, first, the proposed **Algorithms 1A and 1B** are indis-

Table 2: Numerical results of **Algorithm 1A and 1B**, and **Full grid search** for $n \in \{150, 250, 350\}$, $p \in \{25, 150, 250\}$, and $\tau_0 = 0.642$.

Method	**n**	**p**	**bias**($\hat{\beta}$)	**bias**($\hat{\gamma}$)	**bias**($\hat{\tau}$)	**mse**($\hat{\beta}$)	**mse**($\hat{\gamma}$)	**mse**($\hat{\tau}$)	**time**
Algorithm 1A	150	25	0.3170	0.4765	0.0065	0.1041	0.2308	0.0003	0.5855
	150	150	0.3900	0.5719	0.0039	0.1309	0.2819	0.0002	0.7854
	150	250	0.4191	0.5668	0.0046	0.1510	0.2756	0.0004	0.8583
	250	25	0.2470	0.3375	0.0088	0.0548	0.1114	0.0002	0.6177
	250	150	0.2926	0.4248	0.0007	0.0658	0.1486	0.0001	1.1309
	250	250	0.3142	0.4436	0.0039	0.0770	0.1614	0.0001	1.3242
	350	25	0.2277	0.3098	0.0041	0.0429	0.0858	0.0001	0.6748
	350	150	0.2501	0.3573	0.0018	0.0497	0.1012	0.0000	1.6129
	350	250	0.2861	0.3861	0.0018	0.0627	0.1135	0.0001	1.8991
Algorithm 1B	150	25	0.3148	0.4833	0.0064	0.1059	0.2364	0.0002	0.9449
	150	150	0.3826	0.5667	0.0060	0.1275	0.2737	0.0002	1.3024
	150	250	0.4138	0.5617	0.0039	0.1489	0.2724	0.0004	1.4660
	250	25	0.2512	0.3312	0.0121	0.0558	0.1114	0.0003	0.9661
	250	150	0.3029	0.4213	0.0005	0.0683	0.1470	0.0001	2.3502
	250	250	0.3147	0.4274	0.0056	0.0773	0.1490	0.0001	2.2163
	350	25	0.2318	0.3071	0.0043	0.0436	0.0847	0.0001	1.0086
	350	150	0.2525	0.3472	0.0016	0.0502	0.0958	0.0000	2.5538
	350	250	0.2815	0.3766	0.0015	0.0617	0.1087	0.0000	3.4451
Full grid search	150	25	0.2506	0.5229	0.0012	0.0983	0.2482	0.0002	12.9812
	150	150	0.5061	0.9313	0.0012	0.2037	0.5588	0.0009	31.2470
	150	250	0.6217	1.0329	0.0065	0.2572	0.6412	0.0008	35.3190
	250	25	0.2000	0.4208	0.0005	0.0505	0.1477	0.0003	21.5832
	250	150	0.3397	0.7712	0.0116	0.0909	0.3870	0.0003	67.6660
	250	250	0.4072	0.8228	0.0005	0.1229	0.4251	0.0001	105.1549
	350	25	0.1674	0.4083	0.0045	0.0353	0.1284	0.0001	30.2175
	350	150	0.2769	0.5949	0.0032	0.0640	0.2360	0.0001	126.2810
	350	250	0.3299	0.6896	0.0030	0.0841	0.3031	0.0001	200.2775

Table 3: Numerical results of **Simulation B**, where the underlying model is $y_i = x_i^T \gamma_0 + \varepsilon_i$, i.e., the 'no change' case where $\Phi(\tau_0) = 0$. The metric PrM is reported for each combination of n, p.

n	**Algorithm 1A**			**Algorithm 1B**		
	$p = 25$	$p = 150$	$p = 250$	$p = 25$	$p = 150$	$p = 250$
150	0.76	0.87	0.88	0.76	0.87	0.88
250	0.80	0.93	0.91	0.80	0.93	0.91
350	0.85	0.93	0.94	0.85	0.93	0.94

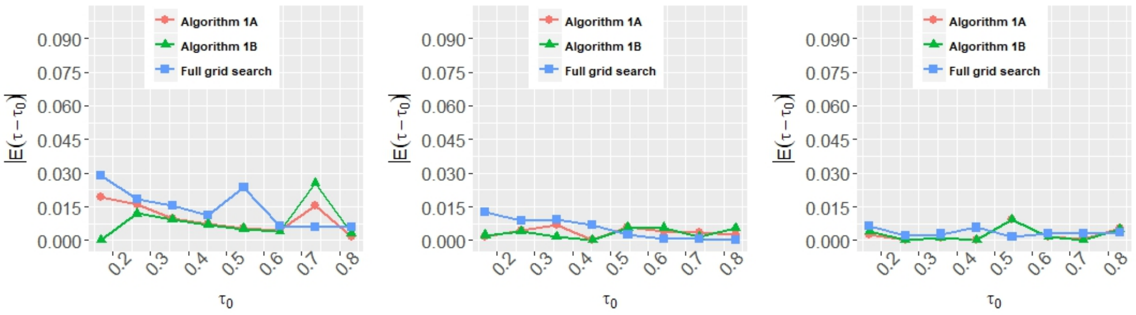

Figure 3: Comparison of bias($\hat{\tau}$) for **Algorithm 1A and 1B** and **Full grid search** across values of τ_0 for $p = 250$. Left panel: $n = 150$, Center panel: $n = 250$, Right panel: $n = 350$.

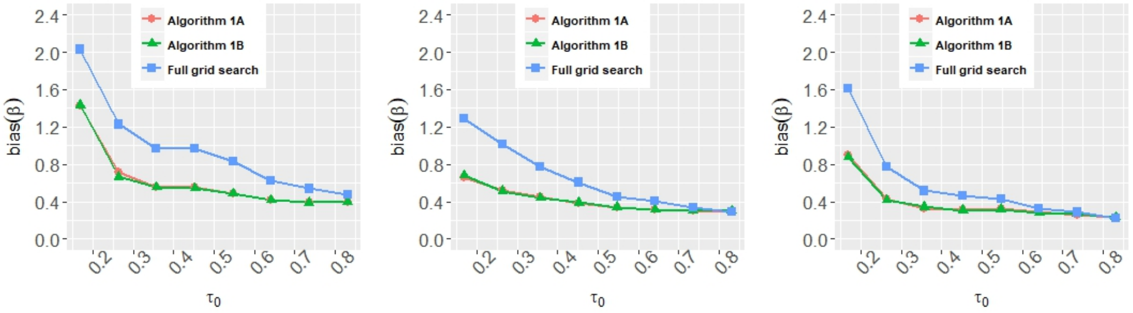

Figure 4: Comparison of bias($\hat{\beta}$) for **Algorithm 1A and 1B**, and **Full grid search** across values of τ_0 for $p = 250$. Left panel: $n = 150$, Center panel: $n = 250$, Right panel: $n = 350$.

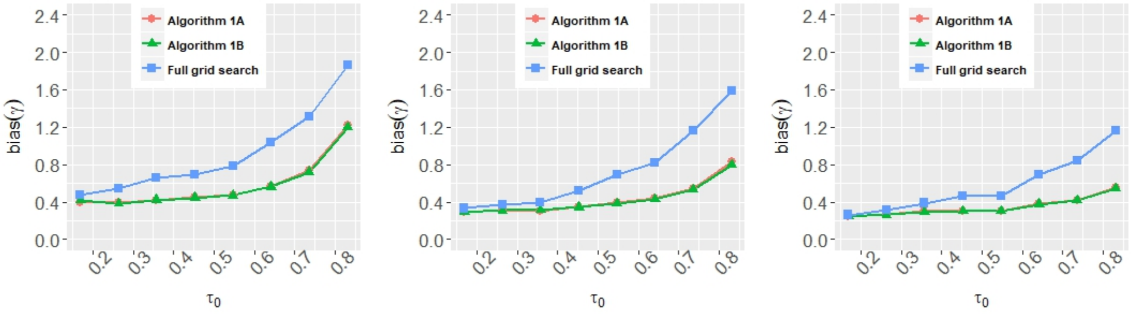

Figure 5: Comparison of bias($\hat{\gamma}$) for **Algorithm 1A and 1B**, and **Full grid search** across values of τ_0 for $p = 250$. Left panel: $n = 150$, Center panel: $n = 250$, Right panel: $n = 350$.

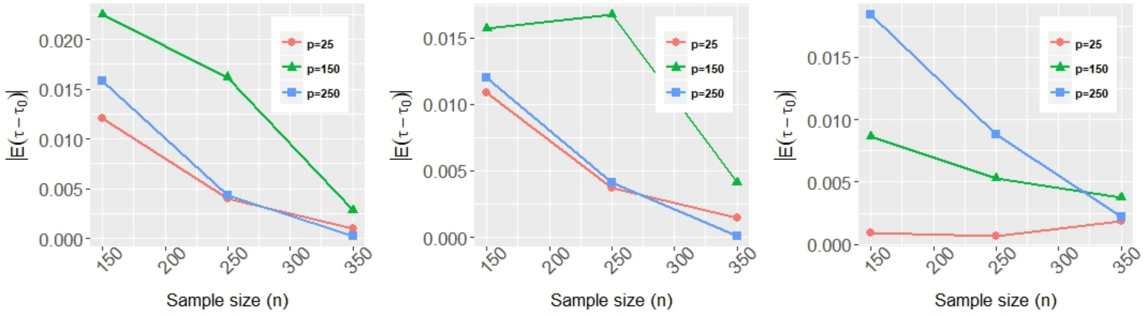

Figure 6: Illustration of consistency of implemented methods with $\tau_0 = 0.264$. Left panel: **Algorithm 1A**, Center panel: **Algorithm 1B**, and Right panel: **Full grid search**

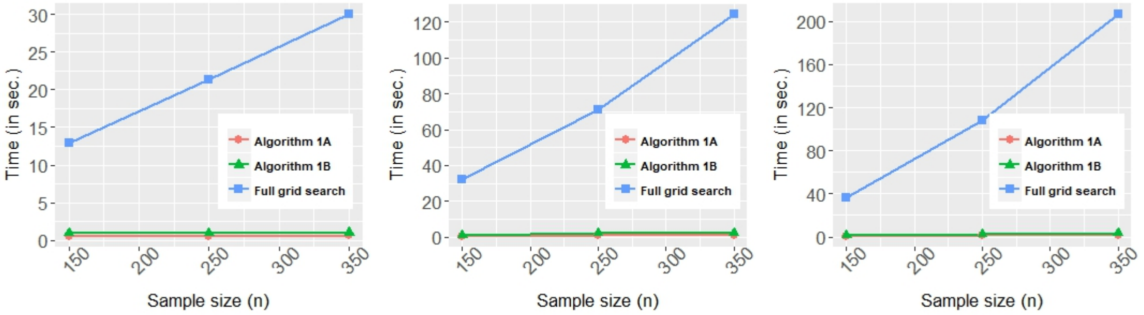

Figure 7: Comparison of computation time (in seconds) for **Algorithm 1A and 1B**, and **Full grid search** across values of n for $\tau_0 = 0.547$ Left panel: $p = 25$, Center panel: $p = 150$, Right panel: $p = 250$. Note that, these times are computed as averages over 100 replications of each method running in parallel over 12 cores. Running a single instance of any method is two to three times faster. Reported computation times include time taken to choose tuning parameters.

tinguishable from the **Full grid search** approach of Lee et al. (2016). Second, it is also observed that the proposed **Algorithms 1A and 1B** are indistinguishable from each other in terms of the bias in the change point estimate. Recall that, Algorithm 1B was designed in a way so that the starting value is always closer to τ_0 in comparison to **Algorithm 1A**. Despite a better initial value, no uniform improvement is observed in Algorithm 1B. This supports our theoretical result, that the quality of the initial value does not impact Algorithm 1, and it yields near optimal estimates with any initializing value containing any fractional information on the unknown change point. The bias results for the regression coefficient estimates depicted in Figures 4 and Figure 5 suggest that the proposed methodology yields a uniformly lower bias at all considered cases of τ_0. One possible reason for this behavior is that the design variable in our methodology are constructed as $z_i = (z_i^{(1)T}, z_i^{(2)T})^T$ where $z_i^{(1)} = x_i\mathbf{1}[w_i \le \tau_0]$, and $z_i^{(2)} = x_i\mathbf{1}[w_i \le \tau_0]$, which are orthogonal to each other, in contrast, the design variables in the methodology of Lee et al. (2016), the design variables are constructed as $z_i^{(1)} = x_i$ and $z_i^{(2)} = \mathbf{1}[w_i \le \tau_0]$, which may be highly correlated. It is also clear from Figure 6 that the in bias in change point estimates from **Algorithm 1A and 1B** and **Full grid search** progressively shrinks with increasing values of n, thereby illustrating the consistency of the proposed methodology. Finally, in Figure 7 we illustrate the dramatic differences in the overall computation times in the implementation of the compared approaches. In the largest considered data set, the average time for computation of **Algorithm 1A and 1B** was $\approx$ 3seconds, as opposed to the full grid search which required $\approx$ 200seconds to implement. Note that the reported computation times include the time taken for choosing all required tuning parameters for each method.

Simulation B: The results of **Simulation B** reported in Table 3 are in accordance with expectations. The proposed methods are able to detect the 'no change' scenario with $\approx$ 85% accuracy in all considered cases. Selection consistency is also observed, i.e. the proportion of correct identifications is seen to increase with n. Finally, both **Algorithm**

1A and 1B are seen to provide the exact same results, which is again not surprising since the only difference in these two methods is the choice of the initial value.

6. Application

In this section, we apply our proposed methodology to the 'Communities and Crime' data set of Redmond and Baveja (2002), available publicly at `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`. This data contains: (i) socio-economic data at a community level from across the entire United states, and is collected from the 1990 US Census, (ii) law enforcement data from the 1990 US Law Enforcement Management and Administrative Statistics survey, and (iii) crime data from the 1995 FBI Uniform Crime. The full data set contains 1994 observations and 128 variables. The dependent variable of interest is the total number of violent crimes per one hundred thousand population, which is calculated using the population and the sum of the crime variables that are considered violent crimes: murder, rape, robbery, and assault. The remaining variables are quantitative measurements on socio-economic variables such as the median (community level) income per household, percentage of people aged 16 and over who are employed, percentage of households with public assistance, percent of population who have immigrated within the last 10 years, amongst many others. This data was recently analyzed by Leonardi and Bühlmann (2016) for detecting and identifying change points in covariates when the change point(s) are modeled over locations. In this study we are interested in identifying changes in covariates when the change occurs through a change inducing variable. Specifically, we consider tow cases, (1) when the change inducing variable is assumed to be the population for the community, and (2) when the change inducing variable is assumed to be the median household income for the community, in an effort to investigate whether violent crime at a community level is influenced by distinct socio-economic factors below and above a certain threshold of population or the median household income, and also to estimate the threshold level at which such a transition occurs.

The full data set consists of $n = 1994$ observations and $p = 128$ variables, which have been normalized to $[0, 1]$ scale. The normalization process is described in the webpage whose link has been provided at the beginning of this section. This normalized data is pre-processed by deleting observations with any missing values, and by eliminating predictors that are highly correlated with other predictor variables. After the pre-processing, we obtain a filtered data set with $n = 319$ communities. The remaining data is then mean centered and scaled columnwise in order to remove the need for an intercept term in the regression, mainly to be consistent with model (1.1). Finally, predictor variables having a significant correlation with the change inducing variable have also been dropped from the analysis. This process yields a refined data set with $p = 75$ predictor variables (excluding the change inducing variable) in the case where the change inducing variable is 'population' and $p = 77$ in the case where the change inducing variable is 'median income'

We apply the proposed Algorithm 1 to the data under consideration with the initializer chosen as the 50^{th} percentile of the change inducing variable, i.e, Algorithm 1A described in Section 5. The regularizer's λ_1, λ_2 and μ are chosen via cross validation and the classical BIC criteria respectively, as described in Section 5. Table 4 summarizes estimation and variable selection results for the regression coefficients of the assumed model (1.1), in the

Table 4: Summary of analysis of 'Communities and Crime' data set. Change inducing variable (w): population, model size: $n = 319$, $p = 75$. Estimated change point is $\hat{\tau}^{(1)} = 0.24$, which is the 73^{rd} percentile of the population variable. The table lists all estimated non-zero regression coefficients truncated at 10^{-4}.

Variable	**Description**	**Coefficient ($\hat{\beta}_{\hat{S}}$)** **(pre change)**	**Coefficient ($\hat{\gamma}_{\hat{S}}$)** **(post change)**
racepctblack	% of population that is african american	0.0322	0.0000
racePctWhite	% of population that is caucasian	-0.1844	0.0000
pctWWage	% of households with wage or salary income in 1989	-0.0505	0.0000
pctWInvInc	% of households with investment / rent income in 1989	-0.0909	0.0000
PctPopUnderPov	% of people under the poverty level	0.0686	0.0000
PctEmploy	% of people 16 and over who are employed	-0.0069	0.0000
PctIlleg	% of kids born to never married	0.3105	0.1734
PctHousLess3BR	% of housing units with less than 3 bedrooms	0.0199	0.0000
PctHousOccup	% of housing occupied	-0.0359	0.0000
PctVacantBoarded	% of vacant housing that is boarded up	0.0000	0.0743
NumStreet	# of homeless people counted in the street	0.0000	0.1597
LemasSwFTFieldOps	# of sworn full time police officers in field operations	0.0000	-0.0229
PolicReqPerOffic	total requests for police per police officer (0/1)	0.0229	0.0000
LemasGangUnitDeploy	gang unit deployed	0.0196	0.0000

Table 5: Summary of analysis of 'Communities and Crime' data set. Change inducing variable (w): median income, model size: $n = 319$, $p = 77$. Estimated change point is $\hat{\tau}^{(1)} = -\infty$, i.e., no change detected in the model w.r.t. w. The table lists all estimated non-zero regression coefficients truncated at 10^{-4}.

Variable	Description	Coefficient ($\hat{\gamma}_{\hat{S}}$)
racepctblack	% of population that is african american	0.0207
racePctWhite	% of population that is caucasian	-0.1863
pctWWage	% of households with wage or salary income in 1989	-0.0245
pctWInvInc	% of households with investment / rent income in 1989	-0.1115
PctIlleg	% of kids born to never married	0.3010
PctHousLess3BR	% of housing units with less than 3 bedrooms	0.0230
HousVacant	# of vacant households	0.0045
PctHousOccup	% of housing occupied	-0.0352
PctVacantBoarded	% of vacant housing that is boarded up	0.0132
NumStreet	# of homeless people counted in the street	0.0919
PolicCars	# of police cars	0.0178

case where the change inducing variable is 'population' and Table 5 summarizes the results of the case where the change inducing variable is the 'median income'. In the first case with the change inducing variable as 'population', we find a change point estimate $\hat{\tau}^{(1)} = 0.23$ which is the 73^{rd} percentile of the population variable. A noteworthy observation in this case about the estimated pre and post coefficients $\hat{\beta}^{(1)}$, $\hat{\gamma}^{(1)}$ from Table 4 is the near disjoint nature of the features influencing violent crime across the threshold $\hat{\tau}^{(1)}$ of the population variable. In the second case, where the change inducing variable is 'median income' the proposed method detects 'no change' in the model, i.e., yields a ordinary linear regression model for this case.

Acknowledgement

We thank three anonymous referee's for insighfut comments and suggestions that led to a significant improvement to the results and presentation of the manuscript.

Appendix A: Proofs

Proof of Lemma 3.1: Let $\tau_1 > \tau_0$ be a boundary point on the right of τ_0, such that $\Phi^*(\tau_0,\tau_1) = u_n$. Then recall that

$$\zeta_i(\tau_1) = \mathbf{1}[\tau_0 < w_i \leq \tau_1], \qquad \Phi^*(\tau_0,\tau) = \Phi(\tau_1) - \Phi(\tau_0).$$

Also, note that $p_n := E\zeta_i(\tau_1) = \Phi^*(\tau_0,\tau_1)$. Since ζ_i, $i = 1,...,n$ are Bernoulli r.v.'s, for any $s > 0$, the moment generating function is given by $E(\exp(s\zeta_i)) = q_n + p_n\exp(s)$, where $q_n = 1 - p_n$. Applying the Chernoff Inequality, we obtain,

$$P\big(\sum_{i=1}^n \zeta_i(\tau_1) > t + np_n\big) = P\big(e^{\sum_{i=1}^n s\zeta_i(\tau_1)} > e^{(st+snp_n)}\big) \leq e^{-s(t+np_n)}[q_n + p_n e^s]^n.$$

Now in order to show,

$$P\Big(\frac{1}{n}\sum_{i=1}^n \zeta_i(\tau_1) \leq c_u \max\Big\{\frac{\log p}{n}, u_n\Big\}\Big) \quad \geq \quad 1 - c_1\exp(-c_2\log p). \tag{A.1}$$

We divide the argument into two cases. First, for any arbitrary constant $c_u > 0$, we let $\Phi^*(\tau_0,\tau_1) \geq c_u\log p/n$, upon choosing $t = n\Phi^*(\tau_0,\tau_1)$ we obtain,

$$P\big(\sum_{i=1}^n \zeta_i(\tau_1) > 2n\Phi^*(\tau_0,\tau_1)\big) \leq e^{[-2sn\Phi^*(\tau_0,\tau_1)]}[1 + (\Phi^*(\tau_0,\tau_1))(e^s - 1)]^n.$$

Using the deterministic inequality $(1+x)^k \leq \exp(kx)$, for any $k, x > 0$, we obtain that

$$P\big(\sum_{i=1}^n \zeta_i(\tau_1) > 2n\Phi^*(\tau_0,\tau_1)\big) \leq e^{-2sn\Phi^*(\tau_0,\tau_1)}e^{(e^s-1)n\Phi^*(\tau_0,\tau_1)} \leq e^{-c_2\log p}.$$

The inequality to the right follows by choosing $s = \log 2$, which maximizes the function $f(s) = 2s - e^s + 1$ and provides a positive value at the maximum, and by using the restriction $\Phi^*(\tau_0,\tau) \geq c_u\log p/n$. Next we let $\Phi^*(\tau_0,\tau_1) < c_u\log p/n$. Here choose $t = c_u\log p$ to obtain,

$$P\big(\sum_{i=1}^n \zeta_i(\tau_1) > c_u\log p + n\Phi^*(\tau_0,\tau_1)\big) \leq e^{[-sc_u\log p - sn\Phi^*(\tau_0,\tau_1)]}[1 + (\Phi^*(\tau_0,\tau_1))(e^s - 1)]^n \tag{A.2}$$

Calling upon the inequality $(1+x)^k \leq \exp(kx)$, for any $k, x > 0$, we can bound the RHS of (A.2) from above by $\exp\big[-sc_u\log p + (e^s - s - 1)\log p\big]$. Now $s = \log(1 + c_u)$ provides a positive value at the maximum, since it maximizes $f(s) = (1 + c_u)s - e^s + 1$. Then for any $c_u > 0$, we obtain,

$$P\big(\sum_{i=1}^n \zeta_i(\tau_1) > c_u\log p + n\Phi^*(\tau_0,\tau_1)\big) \quad \leq \quad e^{-c_2\log p}.$$

Upon combining both cases, (A.1) follows by noting $\Phi^\star(\tau_0,\tau_1)=u_n$.

Now repeating the same argument for a fixed boundary point τ_2 on the left of τ_0, such that $\Phi(\tau_0)-\Phi(\tau_2)=u_n$, and applying a union bound we obtain,

$$P\Big(\max_{\tau\in\{\tau_1,\tau_2\}}\frac{1}{n}\sum_{i=1}^{n}\zeta_i(\tau)\le c_u\max\Big\{\frac{\log p}{n},u_n\Big\}\Big)\ge 1-c_1\exp(-c_2\log p). \tag{A.3}$$

It remains to show that (A.1) holds uniformly over $\mathcal{T}(\tau_0,u_n)$. For this, we begin by noting that for any $\tau\in\mathcal{T}(\tau_0,u_n)$, where $\tau>\tau_0$ we have $\zeta_i(\tau)=\mathbf{1}\big[w_i\in(\tau_0,\tau]\big]\le\mathbf{1}\big[w_i\in(\tau_0,\tau_1]\big]$. Similarly for any $\tau\in\mathcal{T}(\tau_0,u_n)$ where $\tau<\tau_0$ we have $\zeta_i(\tau)\le\mathbf{1}\big[w_i\in[\tau_2,\tau_0)\big]$. Thus

$$\sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\frac{1}{n}\sum_{i=1}^{n}\zeta_i(\tau)\le\max_{\tau\in\{\tau_1,\tau_2\}}\frac{1}{n}\sum_{i=1}^{n}\zeta_i(\tau). \tag{A.4}$$

Part (i) of this lemma follows by combining (A.4) with the bound in (A.3).

To prove Part (ii) we use a lower bound for sums of non-negative r.v.s' stated in Lemma B.3. This result was originally proved by Maurer (2003). For a fixed right boundary point $\tau_1>\tau_0$ such that $\Phi(\tau_1)-\Phi(\tau_0)=v_n$, set $t=v_n/2$ in Lemma B.3. Then we have

$$P\Big(\frac{1}{n}\sum_{i=1}^{n}\zeta_i(\tau_1)\le\frac{v_n}{2}\Big)\le\exp\Big(-nv_n\Big)\le c_1\exp(-c_2\log p),$$

where the last inequality follows from $v_n\ge c_u\log p/n$. We obtain the same bound applying a similar argument for the left boundary point $\tau_2<\tau_0$ such that $\Phi(\tau_0)-\Phi(\tau_2)=v_n$. Now applying an elementary union bound we obtain

$$P\Big(\min_{\tau\in\{\tau_1,\tau_2\}}\frac{1}{n}\sum_{i=1}^{n}\zeta_i(\tau)\ge c_uv_n\Big)\ge 1-c_1\exp(-c_2\log p). \tag{A.5}$$

Finally to obtain uniformity over $\tau\in\big\{\tau;\,\Phi^*(\tau_0,\tau)\ge v_n\big\}$ note that for $\tau>\tau_0$, we have $\zeta_i(\tau)=\mathbf{1}\big[w_i\in(\tau_0,\tau]\big]\ge\mathbf{1}\big[w_i\in(\tau_0,\tau_1]\big]$ and for any $\tau<\tau_0$, we have $\zeta_i(\tau)=\mathbf{1}\big[w_i\in[\tau,\tau_0)\big]\ge\mathbf{1}\big[w_i\in[\tau_2,\tau_0)\big]$. This implies that

$$\inf_{\{\tau;\,\Phi^*(\tau_0,\tau)\ge v_n\}}\frac{1}{n}\sum_{i=1}^{n}\zeta_i(\tau)\ge\min_{\tau\in\{\tau_1,\tau_2\}}\frac{1}{n}\sum_{i=1}^{n}\zeta_i(\tau). \tag{A.6}$$

Part(ii) follows by combining (A.5) and (A.6). This complete the proof of Lemma 3.1. □

Proof of Lemma 3.2: We begin with the proof of Part (i). Note that the RHS of the inequality in Part (i) is normalized by the ℓ_2 norm of δ. Hence, without loss of generality we can assume $\|\delta\|_2=1$. Now, the proof of this lemma relies on $|n_w|=\sum_{i=1}^n\zeta_i(\tau)$, where $\zeta_i(\tau)$ are as defined for Lemma 3.1. Note that if $|n_w|=0$ then Lemma 3.2 holds trivially with probability 1, thus without loss of generality we shall assume that $|n_w|>0$. Now, for any fixed $\tau\in\mathcal{T}(\tau_0,u_n)$, we have

$$\Big\|\frac{1}{n}\sum_{i\in n_w}\delta^Tx_ix_i^T\Big\|_\infty\le\frac{|n_w|}{n}\Big\|\frac{1}{|n_w|}\sum_{i\in n_w}\delta^Tx_ix_i^T\Big\|_\infty \tag{A.7}$$

The second key observation is that under Condition A(iv) and by properties of conditional expectations (see e.g. Lemma B.4), the conditional probability $P_w(\cdot) = P(\cdot\,|w)$ can be bounded by treating w as a constant. Thus,

$$P_w\Big(\Big\|\frac{\sum_{i\in n_w}\delta^T x_i x_i^T}{|n_w|} - \delta^T\Sigma\Big\|_\infty > t\Big) \le 6p\exp(-c_u|n_w|\min\big\{\frac{t^2}{\sigma_x^4},\frac{t}{\sigma_x^2}\big\})$$

where the above probability bound is obtained by an application of Part (ii) of Lemma 14 of Loh and Wainwright (2012): supplementary materials. This lemma is reproduced as Lemma B.1 in the Appendix. Now choosing $t = c_u\max\Big\{\sigma_x^2\sqrt{\frac{\log p}{|n_w|}},\sigma_x\frac{\log p}{|n_w|}\Big\}$ we obtain,

$$P_w\left(\Big\|\frac{\sum_{i\in n_w}\delta^T x_i x_i^T}{|n_w|}\Big\|_\infty \le \|\delta^T\Sigma\|_\infty + c_u\max\Big\{\sigma_x^2\sqrt{\frac{\log p}{|n_w|}},\sigma_x\frac{\log p}{|n_w|}\Big\}\right)$$
$$\ge 1 - c_1\exp(-c_2\log p). \qquad \text{(A.8)}$$

The result in (A.8) together with (A.7) yields,

$$P_w\left(\Big\|\frac{1}{n}\sum_{i\in n_w}\delta^T x_i x_i^T\Big\|_\infty \le \frac{|n_w|}{n}\|\delta^T\Sigma\|_\infty + \frac{|n_w|}{n}c_u\max\Big\{\sigma_x^2\sqrt{\frac{\log p}{|n_w|}},\sigma_x\frac{\log p}{|n_w|}\Big\}\right)$$
$$\ge 1 - c_1\exp(-c_2\log p). \qquad \text{(A.9)}$$

Taking expectations on both sides of the inequality (A.9) and observing that the RHS of the conditional probability (A.9) is free of w, we obtain,

$$P\left(\Big\|\frac{1}{n}\sum_{i\in n_w}\delta^T x_i x_i^T\Big\|_\infty \le \frac{|n_w|}{n}\|\delta^T\Sigma\|_\infty + \frac{|n_w|}{n}c_u\max\Big\{\sigma_x^2\sqrt{\frac{\log p}{|n_w|}},\sigma_x\frac{\log p}{|n_w|}\Big\}\right)$$
$$\ge 1 - c_1\exp(-c_2\log p) \qquad \text{(A.10)}$$

On the other hand, we have by the result of Lemma 3.1 that with probability at least $1 - c_1\exp(-c_2\log p)$ that $\sup_{\tau\in\mathcal{T}}|n_w|/n \le c_u\max\{\log p/n, u_n\}$. Also, it is straightforward to see that $\|\delta^T\Sigma\|_\infty \le c_u\phi$, for some constant $c_u > 0$. Thus with the same probability we have the bound,

$$\sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\frac{|n_w|}{n}\|\delta^T\Sigma\|_\infty \le c_u\phi\max\Big\{\frac{\log p}{n},u_n\Big\}. \qquad \text{(A.11)}$$

By applying Part (i) of Lemma 3.1 we also have the following bound with probability at least $1 - c_1\exp(-c_2\log p)$,

$$\sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\frac{|n_w|}{n}\sqrt{\frac{\log p}{|n_w|}} \le c_u\sqrt{\frac{\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\sqrt{u_n}\Big\} \le c_u\max\Big\{\frac{\log p}{n},u_n\Big\}. \qquad \text{(A.12)}$$

The final inequality follows upon noting that if $\sqrt{\log p/n}\sqrt{u_n} \ge u_n$ then $u_n \le \log p/n$. Finally also note that $\sup_{\tau\in\mathcal{T}}(|n_w|/n)(\log p/|n_w|) \le \log p/n$. Part (i) of the lemma follows by combining these results together with the bounds (A.11) and (A.12) in (A.10). The proofs of Part (ii) and Part (iii) are similar and are thus omitted. □

Proof of Lemma 3.3 To prove Part (i), first define $z_i = x_i \mathbf{1}[w_i \leq \tau]$. Clearly z_i is also subgaussian with the same variance parameter as x_i's, i.e., σ_x^2. Furthermore, since by assumption $\Phi(\tau) > 0$, thus $\Sigma_z = E z_i z_i^T = \Phi(\tau)\Sigma_x$, which implies that $\lambda_{\min}(\Sigma_z) = \Phi(\tau)\lambda_{\min}(\Sigma_x) \geq \Phi(\tau)\kappa$. Similarly $\lambda_{\max}(\Sigma_z) \leq \Phi(\tau)\phi$. Now applying Lemma B.2 we obtain

$$\frac{1}{n}\sum_{\{i;w_i\leq\tau\}} \delta^T x_i x_i^T \delta = \frac{1}{n}\sum_{i=1}^{n} \delta^T z_i z_i^T \delta \geq c_u \kappa \Phi(\tau)\|\delta\|_2^2 - c_u \frac{1}{\Phi(\tau)}\frac{\log p}{n}\|\delta\|_1^2, \qquad \text{(A.13)}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. Since $\delta \in \mathbb{A}$, it is straightforward to see that $\|\delta\|_1^2 \leq c_u s\|\delta\|_2^2$. This together with Condition A(ii) yields Part (i). The proof of Part (ii) is quite similar. To prove Part (iii), we shall invoke the arguments seen in the proof of Lemma 3.2. Consider,

$$\sup_{\tau\in\mathcal{T}(\tau_0,u_n)} \sup_{\delta\in\mathbb{A}} \frac{1}{n}\sum_{i\in n_w} \delta^T x_i x_i^T \delta = \sup_{\tau\in\mathcal{T}(\tau_0,u_n)} \frac{|n_w|}{n} \sup_{\delta\in\mathbb{A}} \frac{1}{|n_w|}\sum_{i\in n_w} \delta^T x_i x_i^T \delta \qquad \text{(A.14)}$$

Let $P_w(\cdot\,)$ denote the conditional probability $P(\cdot\,|w)$, where $w = (w_1, ..., w_n)^T$. Then using Lemma B.2 we have

$$P_w\left(\sup_{\delta\in\mathbb{A}} \frac{1}{|n_w|}\sum_{i\in n_w} \delta^T x_i x_i^T \delta \leq \frac{3\phi}{2}\|\delta\|_2^2 + c_u c_m \frac{\log p}{|n_w|}\|\delta\|_1^2\right) \geq 1 - c_1 \exp(-c_2 \log p). \quad \text{(A.15)}$$

Noting that the above probability on the RHS of (A.15) is free of w, taking expectations on both sides we obtain,

$$P\left(\sup_{\delta\in\mathbb{A}} \frac{1}{|n_w|}\sum_{i\in n_w} \delta^T x_i x_i^T \delta \leq \frac{3\phi}{2}\|\delta\|_2^2 + c_u c_m \frac{\log p}{|n_w|}\|\delta\|_1^2\right) \geq 1 - c_1 \exp(-c_2 \log p). \quad \text{(A.16)}$$

Recall from Lemma 3.1 that $\sup_{\tau\in\mathcal{T}(\tau_0,u_n)} |n_w|/n \leq c_u \max\{\log p/n, u_n\}$, with probability at least $1 - c_1 \exp(-c_2 \log p)$. Combining this result with (A.16) and substituting into (A.14) we obtain

$$\begin{aligned}\sup_{\tau\in\mathcal{T}(\tau_0,u_n)} \sup_{\delta\in\mathbb{A}} \frac{1}{n}\sum_{i\in n_w} \delta^T x_i x_i^T \delta &\leq c_u \phi \|\delta\|_2^2 \max\left\{\frac{\log p}{n}, u_n\right\} + c_u c_m \frac{s\log p}{n}\|\delta\|_2^2 \\ &\leq c_u c_m \|\delta\|_2^2 \max\left\{\frac{s\log p}{n}, u_n\right\}\end{aligned}$$

with probability at least $1 - c_1 \exp(-c_2 \log p)$. This completes the proof of Part (iii). Proof of Part (iv) is based on similar arguments. First applying the same conditional argument as above, Part (i) of Lemma B.2 yields,

$$P\left(\inf_{\delta\in\mathbb{A}_2} \frac{1}{|n_w|}\sum_{i\in n_w} \delta^T x_i x_i^T \delta \geq \frac{\kappa}{2}\|\delta\|_2^2 - c_u c_m \frac{\log p}{|n_w|}\|\delta\|_1^2\right) \geq 1 - c_1 \exp(-c_2 \log p). \quad \text{(A.17)}$$

Since $v_n \geq c\log p/n$, Part (ii) of Lemma 3.1 gives,

$$\inf_{\substack{\tau\in\mathbb{R};\\ \Phi(\tau_0,\tau)\geq v_n}} \inf_{\delta\in\mathbb{A}_2} \frac{1}{n}\sum_{i\in n_w} \delta^T x_i x_i^T\delta \geq c_u\kappa\|\delta\|_2^2 v_n - c_u c_m\frac{\log p}{n}\|\delta\|_1^2, \tag{A.18}$$

with probability at least $1-c_1\exp(-c_2\log p)$. By the definition of the set $\mathbb{A}_2$ together with the fact that $\|\beta_0-\gamma_0\|_0\leq s$, we also have

$$\|\delta\|_1^2 \leq c_u s(\|\delta\|_2^2 + \|\beta_0-\gamma_0\|_2^2). \tag{A.19}$$

Finally, substituting (A.19) in (A.18) we obtain

$$\inf_{\substack{\tau\in\mathbb{R};\\ \Phi(\tau_0,\tau)\geq v_n}} \inf_{\delta\in\mathbb{A}_2} \frac{1}{n}\sum_{i\in n_w} \delta^T x_i x_i^T\delta \geq c_u\kappa\|\delta\|_2^2 v_n - c_u c_m\frac{s\log p}{n}\|\delta\|_2^2 - c_u c_m\frac{s\log p}{n}\|\beta_0-\gamma_0\|_2^2$$

with probability at least $1-c_1\exp(-c_2\log p)$. This completes the proof of the lemma. □

Proof of Theorem 4.1: To prove part (i), first note that when $\Phi_{\min}(\tau_0)=0$, the model (1.1) reduces to an ordinary linear regression model with regression coefficient γ_0. Thus for any $\tau\in\mathbb{R}$, the estimates $\hat\beta(\tau)$ and $\hat\gamma(\tau)$ are ordinary Lasso estimates on the binary partitioned data (y_i,z_i), where $z_i = x_i\mathbf{1}[w_i\leq\tau]$, and $z_i = x_i\mathbf{1}[w_i>\tau]$, respectively. Also note that by assumption $\Phi_{\min}^{-1}(\tau)s\log p/n = o(1)$, thus the restricted eigenvalue condition of Part (i) and Part (ii) of Lemma 3.3 are applicable. The remaining arguments to prove the desired bounds are the same as typically used to derive bounds for Lasso estimates, such as those given in Chapter 6 of Bühlmann and Van De Geer (2011), these arguments are also similar to those to follow for the proof of Part (ii) and are thus omitted.

For the proof of Part (ii) where $\Phi_{\min}(\tau_0)>0$, we only prove the uniform bound for the error in estimate $\|\hat\beta(\tau)-\beta_0\|_q$. The proof for $\|\hat\gamma(\tau)-\gamma_0\|_q$ is nearly identical. First, for any $\tau\in\mathcal{T}(\tau_0,u_n)$, note that by Lemma 3.2, we have,

$$\begin{aligned}\sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\Big\|\frac{1}{n}\sum_{i;w_i\leq\tau}\varepsilon_i x_i^T\Big\|_\infty &\leq \|\frac{1}{n}\sum_{i;w_i\leq\tau_0}\varepsilon_i x_i^T\| + \sup_{\tau\in\mathcal{T}(\tau_0,u_n)}\Big\|\frac{1}{n}\sum_{i\in n_w}\varepsilon_i x_i^T\Big\|_\infty\\ &\leq c_u c_m\sqrt{\frac{\log p}{n}} + c_u c_m\sqrt{\frac{\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\sqrt{u_n}\Big\}\\ &\leq c_u c_m\sqrt{\frac{\log p}{n}} \end{aligned}\tag{A.20}$$

with probability at least $1-c_1\exp(-c_2\log p)$. Also we have for any $\beta\in\mathbb{R}^p$ and $\tau\in\mathbb{R}$,

$$\begin{aligned}\frac{1}{n}\sum_{i;w_i\leq\tau}(y_i-x_i^T\beta)^2 &= \frac{1}{n}\sum_{i;w_i\leq\tau}(y_i-x_i^T\beta_0-x_i^T(\beta-\beta_0))^2\\ &= \frac{1}{n}\sum_{i;w_i\leq\tau}\tilde\varepsilon_i^2 - \frac{2}{n}\sum_{i;w_i\leq\tau}\tilde\varepsilon_i x_i^T(\beta-\beta_0) + \frac{1}{n}\sum_{i;w_i\leq\tau}\|x_i^T(\beta-\beta_0)\|_2^2.\end{aligned}\tag{A.21}$$

Here $\tilde\varepsilon_i=\varepsilon_i$, for $i\in\{i;w_i\leq\tau_0\}$ and $\tilde\varepsilon_i = \varepsilon_i - x_i^T(\beta_0-\gamma_0)$ for $i\in\{i;w_i>\tau_0\}$. Now by the definition of $\hat\beta(\tau)$, it follows that

$$\frac{1}{n}\sum_{i;w_i\leq\tau}(y_i-x_i^T\hat\beta(\tau))^2+\lambda_1\|\hat\beta(\tau)\|_1 \leq \frac{1}{n}\sum_{i;w_i\leq\tau}(y_i-x_i^T\beta_0)^2+\lambda_1\|\beta_0\|_1. \tag{A.22}$$

Applying (A.21) in (A.22) and carrying out some algebraic operations we get

$$
\begin{aligned}
&\frac{1}{n}\sum_{i;w_i\le\tau}\|x_i^T(\hat\beta(\tau)-\beta_0)\|_2^2+\lambda_1\|\hat\beta(\tau)\|_1\\
&\le\lambda_1\|\beta_0\|_1+\Big|\frac{2}{n}\sum_{i;w_i\le\tau}\varepsilon_i x_i(\hat\beta(\tau)-\beta_0)\Big|+\Big|\frac{2}{n}\sum_{\tau_0<w_i\le\tau}(\beta_0-\gamma_0)x_ix_i^T(\hat\beta(\tau)-\beta_0)\Big|\\
&\le\Big\|\frac{2}{n}\sum_{i;w_i\le\tau}\varepsilon_i x_i\Big\|_\infty\|\hat\beta(\tau)-\beta_0\|_1+\Big\|\frac{2}{n}\sum_{i\in n_w}(\beta_0-\gamma_0)^Tx_ix_i^T\Big\|_\infty\|\hat\beta(\tau)-\beta_0\|_1+\lambda_1\|\beta_0\|_1\\
&\le c_uc_m\|\beta_0-\gamma_0\|_2\max\Big\{\frac{\log p}{n},u_n\Big\}\|\hat\beta(\tau)-\beta_0\|_1+c_uc_m\max\sqrt{\frac{\log p}{n}}\|\hat\beta(\tau)-\beta_0\|_1+\lambda_1\|\beta_0\|_1\\
&\le\lambda\|\hat\beta(\tau)-\beta_0\|_1+\lambda_1\|\beta_0\|_1. \qquad\qquad (A.23)
\end{aligned}
$$

Here $\lambda=c_uc_m\max\Big\{\sqrt{\log p/n},\ \|\beta_0-\gamma_0\|_2u_n\Big\}$. The first term of the second to last inequality follows from (A.20) and the second term from Part (i) of Lemma 3.2. The bound (A.23) holds uniformly over $\tau\in\mathcal{T}(\tau_0,u_n)$ with probability at least $1-c_1\exp(-c_2\log p)$. Observe that the first term on the LHS of inequalities (A.23) is nonnegative, therefore $\lambda_1\|\hat\beta\|_1\le\lambda\|\hat\beta-\beta_0\|_1+\lambda_1\|\beta_0\|$. Choosing $\lambda_1\ge2\lambda$ leads to the inequality $\|\hat\beta_{S^c}\|_1\le3\|\hat\beta_S-\hat\beta_{0S}\|_1$, by elementary triangle inequalities, see for e.g. Lemma 6.3 of Bühlmann and Van De Geer (2011). Thus $\delta=\hat\beta-\beta_0\in\mathbb{A}$ and thus the first three inequalities of Lemma 3.3 are now applicable. From (A.23) we obtain,

$$
\begin{aligned}
\frac{2}{n}\sum_{i;w_i\le\tau_0}\|x_i^T(\hat\beta(\tau)-\beta_0)\|_2^2-\frac{2}{n}\sum_{i\in n_w}\|x_i^T(\hat\beta(\tau)-\beta_0)\|_2^2 &\le 3\lambda_1\|\hat\beta(\tau)-\beta_0\|_1\\
&\le 3\sqrt{s}\lambda_1\|\hat\beta(\tau)-\beta_0\|_2 \quad (A.24)
\end{aligned}
$$

Bounding the terms on the LHS of (A.24) by applying Part (i) and (iii) of Lemma 3.3 together with the assumption $u_n=o(\Phi(\tau_0))$, yields

$$
c_uc_m\Phi(\tau_0)\|\hat\beta-\beta_0\|_2^2\le3\sqrt{s}\lambda_1\|\hat\beta-\beta_0\|_2
$$

This directly implies $\|\hat\beta(\tau)-\beta_0\|_2\le c_uc_m\sqrt{s}\lambda_1$. The ℓ_1 bound $\|\hat\beta(\tau)-\beta_0\|_1\le\sqrt{s}\|\hat\beta-\beta_0\|_2$, follows from the previously shown result that $\hat\beta(\tau)-\beta_0\in\mathbb{A}$. To complete the proof of Part (ii), note that all bounds in the above arguments hold uniformly over $\mathcal{T}(\tau_0,u_n)$, consequently the final bound holds uniformly over $\mathcal{T}(\tau_0,u_n)$. □

Proof of Lemma 4.1: We begin with proving Part (i), where $\Phi(\tau_0)=0$, in this case, for any $\tau\in\mathbb{R}$, we have

$$\begin{aligned}
nR_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)}) &= \sum_{i;w_i\le\tau}(y_i-x_i^T\hat\beta^{(0)})^2+\sum_{i;w_i>\tau}(y_i-x_i^T\hat\gamma^{(0)})^2-\sum_{i;w_i>\tau_0}(y_i-x_i^T\hat\gamma^{(0)})^2\\
&= \sum_{i;w_i\le\tau}(y_i-x_i^T\hat\beta^{(0)})^2-\sum_{i;\tau_0<w_i\le\tau}(y_i-x_i^T\hat\gamma^{(0)})^2\\
&= \sum_{i;w_i\le\tau}(\hat\beta^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\beta^{(0)}-\gamma_0)-\sum_{i;\tau_0<w_i\le\tau}(\hat\gamma^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\gamma^{(0)}-\gamma_0)\\
&\quad -2\sum_{i;w_i\le\tau}\varepsilon_ix_i^T(\hat\beta^{(0)}-\gamma_0)+2\sum_{i;w_i\le\tau}\varepsilon_ix_i^T(\hat\gamma^{(0)}-\gamma_0)\\
&\ge -\sum_{i\in n_w}(\hat\gamma^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\gamma^{(0)}-\gamma_0)-2\sum_{i;w_i\le\tau}\varepsilon_ix_i^T(\hat\beta^{(0)}-\gamma_0)\\
&\quad +2\sum_{i;w_i\le\tau}\varepsilon_ix_i^T(\hat\gamma^{(0)}-\gamma_0) \qquad\qquad (A.25)
\end{aligned}$$

Now, by the result in Remark 4.1, we have that $\|\hat\beta^{(0)}-\gamma_0\|_2,\|\hat\gamma^{(0)}-\gamma_0\|\le c_uc_m\sqrt{s\log p/n\Phi^2_{\min}(\tau^{(0)})}$, with probability at least $1-c_1\exp(-c_2\log p)$. Additionally from the proof of Theorem 4.1, it has also been shown that $\hat\beta^{(0)}-\gamma_0$ and $\hat\gamma^{(0)}-\gamma_0$ lie in the set $\mathcal{A}$ of (3.2), with the same probability. Thus the bounds of Lemma 3.2 are applicable. Substituting these bounds in (A.25), for any $v_n>0$, we obtain uniformly over $\mathcal{H}(1,v_n)$ that,

$$\inf_{\tau\in\mathcal{H}(u_n,v_n)}R_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})\ge -c_uc_m\frac{s\log p}{n\Phi^2_{\min}(\tau^{(0)})}-c_uc_m\frac{s\log p}{n\Phi_{\min}(\tau^{(0)})}$$

Finally, recall that $S_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})=R_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})+\mu\big(\|\Phi(\tau)\|_0-\|\Phi(\tau_0)\|_0\big)$, and since in this case $\|\Phi(\tau)\|_0=0$, and for any $\tau\in\mathcal{H}(1,v_n)$, we have $\|\Phi(\tau)\|_0=1$ (since $v_n>0$), hence the statement of Part (i) follows directly.

To prove Part (ii), where $\Phi(\tau_0)>0$, we divide the argument into two cases. First consider the case where $\tau\in\mathcal{H}(u_n,v_n)$, with $\tau\ge\tau_0$, here,

$$\begin{aligned}
nR_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)}) &= nQ(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})-nQ(\tau_0,\hat\beta^{(0)},\hat\gamma^{(0)})\\
&= \sum_{i\in\tau_0<w_i\le\tau}(y_i-x_i^T\beta)^2-\sum_{i\in\tau_0<w_i\le\tau}(y_i-x_i^T\gamma)^2 \qquad (A.26)
\end{aligned}$$

Recall by construction of model (1.1), $\varepsilon_i=y_i-x_i^T\gamma_0$, for $i;w_i>\tau_0$. Using this relation in (A.26) and performing some algebraic manipulation we have that

$$\begin{aligned}
R_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)}) &= \frac{1}{n}\sum_{i\in n_w}(\hat\beta^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\beta^{(0)}-\gamma_0)-\frac{1}{n}\sum_{i\in n_w}(\hat\gamma^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\gamma^{(0)}-\gamma_0)\\
&\quad -\frac{2}{n}\sum_{i\in n_w}\varepsilon_ix_i^T(\hat\beta^{(0)}-\gamma_0)+\frac{2}{n}\sum_{i\in n_w}\varepsilon_ix_i^T(\hat\gamma^{(0)}-\gamma_0)\\
&:= T1+T2+T3+T4 \qquad\qquad (A.27)
\end{aligned}$$

Substituting bounds for term $(T1)$-$(T4)$ given in Lemma .1 (stated after this proof), we obtain,

$$\inf_{\tau\in\mathcal{H}(u_n,v_n)} R_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)}) \geq c_uc_m\xi_n^2v_n - c_uc_m\xi_n^2\frac{s\log p}{n}$$
$$-c_uc_m\xi_n\sqrt{\frac{s\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\sqrt{u_n}\Big\} - c_uc_mr_n^2\max\Big\{\frac{s\log p}{n},\,u_n\Big\}.$$

Now, note that for any $\tau\in\mathbb{R}$, we have that $\|\Phi(\tau)\|_0-\|\Phi(\tau_0)\|_0\leq 1$. Also, when $u_n/\Phi(\tau_0)\to 0$, for any $\tau\in\mathcal{H}(u_n,v_n)$, the quantities $\Phi(\tau_0)$ and $\Phi(\tau)$ will have the same sign, consequently $\|\Phi(\tau)\|_0-\|\Phi(\tau_0)\|_0=0$. In effect, we have for any $\tau\in\mathcal{H}(u_n,v_n)$, that, $\|\Phi(\tau)\|_0-\|\Phi(\tau_0)\|_0\leq F(u_n)$. Using this relation in the definition of $S(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})$, together with the assumption that $\xi_n>c_u$ we obtain,

$$\inf_{\tau\in\mathcal{H}(u_n,v_n)} S_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)}) \geq \xi_n^2\Big(c_uc_mv_n - c_uc_m\frac{s\log p}{n} - \frac{c_uc_m}{1\vee\xi_n}\sqrt{\frac{\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\sqrt{u_n}\Big\}$$
$$-c_uc_m\frac{r_n^2}{1\vee\xi_n^2}\max\Big\{\frac{s\log p}{n},\,u_n\Big\} - \frac{c_u\mu}{1\vee\xi_n^2}F(u_n)\Big).$$

with probability at least $1-c_1\exp(-c_2\log p)$. This completes the proof of this lemma. □

Lemma .1 *Suppose the conditions of Lemma 4.1 and let the terms $T1$, $T2$, $T3$ and $T4$ be as defined in (A.27). Then for n sufficiently large, we have the following bounds,*

$$(i)\qquad \inf_{\tau\in\mathcal{H}(u_n,v_n)}|T1| \geq c_uc_m\xi_n^2v_n - \xi_n^2\frac{s\log p}{n}$$

$$(ii)\qquad \sup_{\tau\in\mathcal{H}(u_n,v_n)}|T2| \leq c_uc_mr_n^2\max\Big\{\frac{s\log p}{n},\,u_n\Big\}$$

$$(iii)\qquad \sup_{\tau\in\mathcal{H}(u_n,v_n)}|T3| \leq c_uc_m\xi_n\sqrt{\frac{s\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\,\sqrt{u_n}\Big\}$$

$$(iv)\qquad \sup_{\tau\in\mathcal{H}(u_n,v_n)}|T4| \leq c_uc_mr_n\sqrt{\frac{s\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\,\sqrt{u_n}\Big\}$$

with probability at least $1-c_1\exp(-c_2\log p)$.

Proof of Lemma .1: Consider the term $T1=n^{-1}\sum_{i\in n_w}(\hat\beta^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\beta^{(0)}-\gamma_0)$. First, recall from the proof of Theorem 4.1 that $\hat\beta^{(0)}-\beta_0\in\mathbb{A}$, with probability at least $1-c_1\exp(-c_2\log p)$. Thus, as described in Remark 3.1, we have that $\delta=\hat\beta^{(0)}-\gamma_0=\hat\beta^{(0)}-\beta_0+\beta_0-\gamma_0\in\mathbb{A}_2$ with the same probability. Now applying Part (iv) of Lemma 3.3 we obtain,

$$\inf_{\tau\in\mathcal{H}(u_n,v_n)}\frac{1}{n}\sum_{i\in n_w}\delta^Tx_ix_i^T\delta \geq c_uc_mv_n\|\delta\|_2^2 - c_uc_m\frac{s\log p}{n}\big(\|\delta\|_2^2+\xi_n^2\big)$$

with probability at least $1-c_1\exp(-c_2\log p)$. Applying the algebraic inequality, $\|\delta_1+\delta_2\|_2^2\geq\|\delta_1\|_2^2+\|\delta_2\|_2^2-2\|\delta_1\|_2\|\delta_2\|_2$ which is applicable for any $\delta_1,\delta_2\in\mathbb{R}^p$, we obtain,

$\|\hat\beta^{(0)}-\beta_0+\beta_0-\gamma_0\|_2^2\geq r_n^2+\xi_n^2-2r_n\xi_n$. Now, by definition of r_n, we also have that $r_n=o(1)\xi_n$, thus $\|\hat\beta^{(0)}-\beta_0+\beta_0-\gamma_0\|_2^2\geq c_u\xi_n^2$, for n large. Similarly, using the inequality $\|\delta_1+\delta_2\|_2^2\leq\|\delta_1\|_2^2+\|\delta_2\|_2^2+2\|\delta_1\|_2\|\delta_2\|_2$, we can show that for n large, $\|\hat\beta^{(0)}-\beta_0+\beta_0-\gamma_0\|_2^2\leq c_u\xi_n^2$. Substituting these bounds back in (A.28) we obtain the result of Part (i). Next consider Part (ii), where we have $T2=n^{-1}\sum_{i\in n_w}(\hat\gamma^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\gamma^{(0)}-\gamma_0)$. Recall from the proof of Theorem 4.1, $\hat\gamma^{(0)}-\gamma_0\in\mathbb{A}$. Now applying Part (iii) of Lemma 3.3 we obtain,

$$\sup_{\tau\in\mathcal{H}(u_n,v_n)}\frac{1}{n}\sum_{i\in n_w}(\hat\gamma^{(0)}-\gamma_0)^Tx_ix_i^T(\hat\gamma^{(0)}-\gamma_0)\leq c_uc_mr_n^2\max\Big\{\frac{s\log p}{n},\,u_n\Big\}$$

with probability at least $1-c_1\exp(-c_2\log p)$. This proves Part (ii). The proof of Part (iii) follows by an application of Part (iii) of Lemma 3.2, i.e.,

$$\begin{aligned}\sup_{\tau\in\mathcal{H}(u_n,v_n)}\Big|\frac{1}{n}\sum_{i\in n_w}\varepsilon_ix_i^T(\hat\beta-\gamma_0)\Big|&\leq\sup_{\tau\in\mathcal{H}(u_n,v_n)}\Big\|\frac{1}{n}\sum_{i\in n_w}\varepsilon_ix_i^T\Big\|_\infty\|(\hat\beta^{(0)}-\gamma_0)\|_1\\&\leq\sup_{\tau\in\mathcal{H}(u_n,v_n)}\Big\|\frac{1}{n}\sum_{i\in n_w}\varepsilon_ix_i^T\Big\|_\infty\sqrt{s}\big(\|(\hat\beta^{(0)}-\beta_0)\|_2+\|\beta_0-\gamma_0\|_1\big)\\&\leq c_uc_mr_n\sqrt{\frac{s\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\,\sqrt{u_n}\Big\}+c_uc_m\xi_n\sqrt{\frac{s\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\,\sqrt{u_n}\Big\}\end{aligned}$$

with probability at least $1-c_1\exp(-c_2\log p)$. This proves Part (iii). The proof of Part (iv) is very similar and is thus omitted. □

Proof of Theorem 4.2: First consider Part (i), where $\Phi(\tau_0)=0$. Applying Part (i) of Lemma 4.1 for any $v_n>0$, we have,

$$\inf_{\tau\in\mathcal{H}(1,v_n)}S_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})\geq\mu-c_uc_m\frac{s\log p}{n\Phi_{\min}^2(\tau^{(0)})}$$

with probability at least $1-c_1\exp(-c_2\log p)$. Recall the choice of $\mu=c_uc_m\big(s\log p/nl_n^2\big)^{1/k^*}$, and the initializing condition $\Phi_{\min}(\tau^{(0)})\geq c_ul_n$. Consequently, $\inf_{\tau\in\mathcal{H}(1,v_n)}S_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})>0$, for n sufficiently large, with the same probability. This implies that $\hat\tau^{(1)}\notin\mathcal{H}(1,v_n)$ for any $v_n>0$. Thereby proving that $\Phi(\hat\tau^{(1)})=0$ is the only remaining possibility with the same probability. This completes the proof of Part (i).

To prove Part (ii), for any $v_n\geq s\log p/n$, we apply Part (ii) of Lemma 4.1 on the set $\mathcal{H}(1,v_n)$, to obtain,

$$\begin{aligned}\inf_{\tau\in\mathcal{H}(1,v_n)}S_n(\tau,\hat\beta^{(0)},\hat\gamma^{(0)})\geq\xi_n^2\Big(&c_uc_mv_n-c_uc_m\frac{s\log p}{n}-\frac{c_uc_m}{1\vee\xi_n}\sqrt{\frac{s\log p}{n}}\\&-c_uc_m\frac{r_n^2}{1\vee\xi_n^2}-\frac{c_u\mu}{1\vee\xi_n^2}\Big),\end{aligned}$$

with probability at least $1-c_1\exp(-c_2\log p)$. Note that, by Condition A(iii) we have that $(s/l_n^2)u_n^{(0)}=o(1)$. Then, upon choosing,

$$v_n\geq v_n^*\quad:=\quad c_uc_m\max\Big\{\frac{s\log p}{n},\frac{1}{1\vee\xi_n}\Big(\frac{s\log p}{nl_n^2}\Big)^{1/k^*}\Big\}$$

for some $c_u > 0$, we have that $\inf_{\tau\in\mathcal{H}(1,v_n)} S_n(\tau,\hat{\beta}^{(0)},\hat{\gamma}^{(0)}) > 0$, for n sufficiently large. This follows by the choice of $\mu = (s\log p/nl_n^2)^{1/k^*}$, and by $r_n^2/\xi_n^2 < v_n^*$, which in turn follows from Condition A(iii). This implies that $\hat{\tau}^{(1)} \notin \mathcal{H}(1,v_n^*)$, i.e., $|\Phi(\hat{\tau}^{(1)}) - \Phi(\tau_0)| \le v_n^*$ with probability at least $1 - c_1\exp(-c_2\log p)$. Note that if $\frac{1}{1\vee\xi_n}\Big(\frac{s\log p}{nl_n^2}\Big)^{1/k^*} \le \frac{s\log p}{n}$, then the result is already proved. Else, reset $u_n = v_n^*$ and reapply the above argument for any $v_n \ge s\log p/n$, to obtain,

$$\inf_{\tau\in\mathcal{H}(u_n,v_n)} S_n(\tau,\hat{\beta}^{(0)},\hat{\gamma}^{(0)}) \ge \xi_n^2\Big(c_uc_m v_n - c_uc_m\frac{s\log p}{n} - \frac{c_uc_m}{1\vee\xi_n}\sqrt{\frac{s\log p}{n}}\max\Big\{\sqrt{\frac{\log p}{n}},\sqrt{u_n}\Big\} - c_uc_m\frac{r_n^2}{\xi_n^2}\max\Big\{\frac{s\log p}{n},\, u_n\Big\}\Big) \qquad .$$

Here, the term $F(u_n) = 0$, since for $u_n = v_n^*$, the sign of $\Phi(\hat{\tau}^{(1)})$ is the same as that of $\Phi(\tau_0)$, for n large. Now, upon choosing,

$$v_n \ge v_n^* := c_uc_m\max\Big\{\frac{s\log p}{n},\ \frac{1}{1\vee\xi_n^{1+\frac{1}{2k^*}}}\Big(\frac{s\log p}{nl_n^2}\Big)^{a_2}\Big\},\ \text{with, } a_2 = \min\Big\{\frac{1}{2}+\frac{1}{2k^*},\frac{1}{k^*}+\frac{1}{k^*}\Big\},$$

we obtain that for n large, $\inf_{\tau\in\mathcal{H}(u_n,v_n^*)} S_n(\tau,\hat{\beta}^{(0)},\hat{\gamma}^{(0)}) > 0$, with probability at least $1 - c_1\exp(-c_2\log p)$. Consequently $\hat{\tau}^{(1)} \notin \mathcal{H}(u_n,v_n^*)$, i.e., $|\Phi(\hat{\tau}^{(1)}) - \Phi(\tau_0)| \le v_n^*$. Note that, by using the above recursive argument, we have tightened the desired rate at each step. As seen earlier, if the second term of the maximum expression is smaller than the first, then the proof is done. Else, continuing these recursions, by resetting u_n to the bound of the previous recursion, and applying Part (ii) of Lemma 4.1, we can obtain for the m^{th} recursion that

$$\begin{aligned} |\Phi(\hat{\tau}^{(1)}) - \Phi(\tau_0)| &\le c_uc_m\max\Big\{\frac{s\log p}{n},\ \frac{1}{1\vee\xi_n^{b_m}}\Big(\frac{s\log p}{n}\Big)^{a_m}\Big\}, \quad \text{where,}\\ a_m &= \min\Big\{\frac{1}{2}+\frac{a_{m-1}}{2},\frac{1}{k^*}+a_{m-1}\Big\}, \quad \text{and} \quad b_m = 1 + b_{m-1}/2, \end{aligned}$$

with $a_1 = b_1 = 1/k^*$. Note that, despite the recursions in the above argument, the probability of the bound obtained after every recursion is maintained to be at least $1 - c_1\exp(-c_2\log p)$, this follows from Remark .1. To finish the proof, note that $k^* \in [2,3]$, $a_m = 1/2 + a_{m-1}/2$, $\forall m$ and when $k^* > 3$, $a_m = 1/2 + a_{m-1}/2$, for m large enough. Finally, if we continue the above recursions an infinite number of times we obtain $a_\infty = \sum_{m=1}^{\infty} 1/2^m = 1$, and $b_\infty = 1 + \sum_{m=1}^{\infty} 1/2^m = 2$. This finishes the proof of this theorem. □.

Remark .1 *(Observation utilized in the proof of Theorem 4.2):* The proof of Theorem 4.2 relies on a recursive application of Lemma 4.1, this in turn requires a recursive application of the bounds of Lemma .1, where the probability of all bounds holding simultaneously at each recursion being at least $1 - c_1\exp(-c_2\log p)$. Despite these recursions (potentially infinite) the result from the final recursion continues to hold with probability at least $1 - c_1\exp(-c_2\log p)$. To see this, let $u_n \to 0$ be any positive sequence and let $\{a_j\} \to a_\infty$, $j \to \infty$, $0 < a_j \le 1$, be any strictly increasing sequence over $j = 1,2,....$. Then define sequences $u_n^j = u_n^{a_j}$, $j = 1,2...$. Here note that $u_n^{j+1} = o(u_n^j)$, $j = 1,...,$ i.e., each sequence converges to zero faster than the preceding one. Let $\mathcal{E}_{u^1}, \mathcal{E}_{u^2}...$ be events, each with probability $1 - c_1 \exp(-c_2 \log p)$, on which the upper bounds of Lemma .1 hold for each $u_n^1, u_n^2, ...$ respectively. Clearly, on the intersection of events $\mathcal{E}_{u^1} \cap \mathcal{E}_{u^2} \cap ....$, all upper bounds of Lemma .1 hold simultaneously over any sequence u_n^j, $j = 1, ..., \infty$ Now, note that by the construction of these sequences, and that these are all upper bounds, the following containment holds $\mathcal{E}_{u^1} \supseteq \mathcal{E}_{u^2} \supseteq ... \supseteq \mathcal{E}_{u^\infty}$. This implies that on the event $\mathcal{E}_{u^\infty}$ all bounds of Lemma .1 hold simultaneously for any sequence $\{u_n^j\}$, $j = 1, ..., \infty$. Here $\mathcal{E}_{u^\infty}$ represents the set corresponding to the sequence $u_n^\infty = u_n^{a_\infty}$. Also, by a single application of Lemma .1, $P(\mathcal{E}_{u^\infty}) \geq 1 - c_1 \exp(-c_2 \log p)$. The same argument can be made for the lower bound of Lemma .1, with the direction of the containment switched.

Proof of Theorem 4.3: Recall that the result of Part (ii) of Theorem 4.1 is a uniform result over the set $\mathcal{T}(\tau_0, u_n)$. The proof of this theorem is now a direct application of Part (ii) of Theorem 4.1, since by the result of Theorem 4.2, we have that $\hat{\tau}^{(1)} \in \mathcal{T}(\tau_0, t_n)$, with probability at least $1 - c_1 \exp(-c_2 \log p)$. □

Appendix B: Auxiliary lemma's

Here we restate without proof the technical lemma's from the literature which have been used in the analysis presented in this manuscript.

Lemma B.1 *If $X \in \mathbb{R}^{n \times p_1}$ is a zero mean subgaussian matrix with parameters (Σ_x, σ_x^2), then for any fixed (unit) vector in $v \in \mathbb{R}^{p_1}$, we have*

$$(i)\ P\Big(\Big|\|Xv\|_2^2 - E\|Xv\|_2^2\Big| \geq nt\Big) \leq \exp\Big(-cn \min\Big\{\frac{t^2}{\sigma_x^4}, \frac{t}{\sigma_x^2}\Big\}\Big)$$

Moreover, if $Y \in \mathbb{R}^{n \times p_2}$ is a zero mean subgaussian matrix with parameters (Σ_y, σ_y^2), then

$$(ii)\ P\Big(\|\frac{Y^T X}{n} - \mathrm{cov}(y_i, x_i)\|_\infty \geq t\Big) \leq 6p_1 p_2 \exp\Big(-cn \min\Big\{\frac{t^2}{\sigma_x^2 \sigma_y^2}, \frac{t}{\sigma_x \sigma_y}\Big\}\Big)$$

where x_i, y_i are the i^{th} rows of X and Y respectively. In particular, if $n \geq c \log p$, then

$$(iii)\ P\Big(\|\frac{Y^T X}{n} - \mathrm{cov}(y_i, x_i)\|_\infty \geq c\sigma_x \sigma_y \sqrt{\frac{\log p}{n}}\Big) \leq c_1 \exp(-c_2 \log p).$$

This lemma provides tail bounds on subexponential r.v.'s and is as stated in Lemma 14 of Loh and Wainwright (2012): supplementary materials. The first part of this lemma is a restatement of Proposition 5.16 of Vershynin (2010) and the other two part are derived via algebraic manipulations of the product under consideration. The following is another useful result from Loh and Wainwright (2012) which provides control on restricted eigenvalues of the gram matrix.

Lemma B.2 *Let $z_i \in \mathbb{R}^p$, $i = 1, ..., n$ be i.i.d subgaussian random vectors with variance parameter σ_z^2 and covariance $\Sigma_z = E z_i z_i^T$. Also, let $\lambda_{\min}(\Sigma_z)$ and $\lambda_{\max}(\Sigma_z)$ be the minimum* *and maximum eigenvalues of the covariance matrix respectively. Then,*

$$(i)\quad \frac{1}{n}\sum_{i=1}^{n}\delta^T z_i z_i^T\delta \geq \frac{\lambda_{\min}(\Sigma_z)}{2}\|\delta\|_2^2 - c_u\lambda_{\min}(\Sigma_z)\max\Big\{\frac{\sigma_z^4}{\lambda_{\min}^2(\Sigma_z)},1\Big\}\frac{\log p}{n}\|\delta\|_1^2,\quad \forall\delta\in\mathbb{R}^p,$$

$$(ii)\quad \frac{1}{n}\sum_{i=1}^{n}\delta^T z_i z_i^T\delta \leq \frac{3\lambda_{\max}(\Sigma_z)}{2}\|\delta\|_2^2 + c_u\lambda_{\min}(\Sigma_z)\max\Big\{\frac{\sigma_z^4}{\lambda_{\min}^2(\Sigma_z)},1\Big\}\frac{\log p}{n}\|\delta\|_1^2,\quad \forall\delta\in\mathbb{R}^p,$$

with probability at least $1 - c_1\exp(-c_2\log p)$.

A proof of this result in an errors-in-variables setting result can be found in the supplementary material of Loh and Wainwright (2012), However Lemma B.2 can be seen to follow as a special case (substitute $\sigma_w = 0$ in Lemma 1 of Loh and Wainwright (2012): supplementary materials).

Lemma B.3 *Let the* $\{X_i\}_{i=1}^m$ *be independent random variables,* $EX_i^2 < \infty$, $X_i \geq 0$. *Set* $S = \sum_{i=1}^n X_i$ *and let* $t > 0$. *Then*

$$P\Big(ES - S \geq t\Big) \leq \exp\Big(\frac{-t^2}{2\sum_{i=1}^n EX_i^2}\Big)$$

This result is as stated in Theorem 1 of Maurer (2003), it provides a lower bound on a sum of positive independent r.v.'s.

Lemma B.4 *Suppose* X *and* Y *are independent random variables. Let* ϕ *be a function with* $E|\phi(X,Y)| < \infty$ *and let* $g(x) = E\phi(x,Y)$, *then*

$$E\big(\phi(X,Y)|X\big) = g(X)$$

This is an elementary result on conditional expectations. A straightforward proof can be found in Example 1.5. page 222, Durrett (2010).

References

Yves Atchade and Leland Bybee. A scalable algorithm for gaussian graphical models with change-points. *arXiv preprint arXiv:1707.04306*, 2017.

Jushan Bai. Estimation of a change point in multiple regression models. *Review of Economics and Statistics*, 79(4):551–563, 1997.

Alexandre Belloni, Victor Chernozhukov, and Lie Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

Alexandre Belloni, Victor Chernozhukov, Abhishek Kaul, Mathieu Rosenbaum, and Alexandre B Tsybakov. Pivotal estimation via self-normalization for high-dimensional linear models with error in variables. *arXiv preprint arXiv:1708.08353*, 2017a.

Alexandre Belloni, Mathieu Rosenbaum, and Alexandre B Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):939–956, 2017b.

Peter J Bickel, Yaacov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications.* Springer Science & Business Media, 2011.

Haeran Cho and Piotr Fryzlewicz. Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(2):475–507, 2015.

Gabriela Ciuperca. Model selection by lasso methods in a change-point model. *Statistical Papers*, 55(2):349–374, 2014.

Rick Durrett. *Probability: theory and examples.* Cambridge university press, 2010.

Stergios B Fotopoulos, Venkata K Jandhyala, Elena Khapalova, et al. Exact asymptotic distribution of change-point mle for change in the mean of gaussian sequences. *The Annals of Applied Statistics*, 4(2):1081–1104, 2010.

Jerome H Friedman, TJ Hastie, and RJ Tibshirani. glmnet: lasso and elastic-net regularized generalized linear models, 2010b. *URL http://CRAN. R-project. org/package= glmnet. R package version*, pages 1–1, 2010.

Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

Eric Gautier and Alexandre Tsybakov. High-dimensional instrumental variables regression and confidence sets. *arXiv preprint arXiv:1105.2454*, 2011.

Alex J Gibberd and Sandipan Roy. Multiple changepoint estimation in high-dimensional gaussian graphical models. *arXiv preprint arXiv:1712.05786*, 2017.

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations.* CRC press, 2015.

David V Hinkley. Inference about the intersection in two-phase regression. *Biometrika*, 56 (3):495–504, 1969.

David V Hinkley. Inference about the change-point in a sequence of random variables. *Biometrika*, 1970.

David V Hinkley. Time-ordered classification. *Biometrika*, 59(3):509–523, 1972.

Venkata Jandhyala, Stergios Fotopoulos, Ian MacNeill, and Pengyu Liu. Inference for single and multiple change-points in time series. *Journal of Time Series Analysis*, 34(4):423–446, 2013.

Venkata K Jandhyala and Stergios B Fotopoulos. Capturing the distributional behaviour of the maximum likelihood estimator of a changepoint. *Biometrika*, 86(1):129–140, 1999.

Venkata K Jandhyala and Ian B MacNeill. Iterated partial sum sequences of regression residuals and tests for changepoints with continuity constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(1):147–156, 1997.

Abhishek Kaul. Lasso with long memory regression errors. *Journal of Statistical Planning and Inference*, 153:11–26, 2014.

Abhishek Kaul and Hira L Koul. Weighted 1-penalized corrected quantile regression for high dimensional measurement error models. *Journal of Multivariate Analysis*, 140:72–91, 2015.

Abhishek Kaul, Ori Davidov, and Shyamal D Peddada. Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics*, 18(3):422–433, 2017.

Roger Koenker and Ivan Mizera. Convex optimization in r. *Journal of Statistical Software*, 60(5):1–23, 2014.

Hira L Koul and Lianfen Qian. Asymptotics of maximum likelihood estimator in a two-phase linear regression model. *Journal of Statistical Planning and Inference*, 108(1-2): 99–119, 2002.

Hira L Koul, Lianfen Qian, and Donatas Surgailis. Asymptotics of m-estimators in two-phase linear regression models. *Stochastic Processes and their Applications*, 103(1):123–154, 2003.

Sokbae Lee, Myung Hwan Seo, and Youngki Shin. The lasso for high dimensional regression with a possible change point. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):193–210, 2016.

Sokbae Lee, Yuan Liao, Myung Hwan Seo, and Youngki Shin. Oracle estimation of a change point in high-dimensional quantile regression. *Journal of the American Statistical Association*, 0(0):1–11, 2018. doi: 10.1080/01621459.2017.1319840. URL `https://doi.org/10.1080/01621459.2017.1319840`.

Florencia Leonardi and Peter Bühlmann. Computationally efficient change point detection for high-dimensional regression. *arXiv preprint arXiv:1601.03704*, 2016.

Biao Liu, Carl D Morrison, Candace S Johnson, Donald L Trump, Maochun Qin, Jeffrey C Conroy, Jianmin Wang, and Song Liu. Computational methods for detecting copy number variations in cancer genome using next generation sequencing: principles and challenges. *Oncotarget*, 4(11):1868, 2013.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Ann. Statist.*, 40(3):1637–1664, 06 2012. doi: 10.1214/12-AOS1018. URL `https://doi.org/10.1214/12-AOS1018`.

Robert Lund, Xiaolan L Wang, Qi Qi Lu, Jaxk Reeves, Colin Gallagher, and Yang Feng. Changepoint detection in periodic and autocorrelated time series. *Journal of Climate*, 20 (20):5178–5190, 2007.

Andreas Maurer. A bound on the deviation probability for sums of non-negative random variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1):15, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE transactions on information theory*, 57 (10):6976–6994, 2011.

Michael Redmond and Alok Baveja. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, 141(3):660–678, 2002.

Jaxk Reeves, Jien Chen, Xiaolan L Wang, Robert Lund, and Qi Qi Lu. A review and comparison of changepoint detection techniques for climate data. *Journal of Applied Meteorology and Climatology*, 46(6):900–915, 2007.

Mark Rudelson and Shuheng Zhou. Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1, 2012.

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

Robert Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Tengyao Wang and Richard J Samworth. High dimensional change point estimation via sparse projection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):57–83, 2018.

Yuehua Wu. Simultaneous change point analysis and variable selection in a regression problem. *Journal of Multivariate Analysis*, 99(9):2154–2171, 2008.

Fei Ye and Cun-Hui Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540, 2010.

Bingwen Zhang, Jun Geng, and Lifeng Lai. Multiple change-points estimation in linear regression models via sparse group lasso. *IEEE Trans. Signal Processing*, 63(9):2209–2224, 2015.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563, 2006.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.