Localizing Changes in High-Dimensional Regression Models

Alessandro Rinaldo
Department of Statistics & Data Science
Carnegie Mellon University

Daren Wang
Department of ACMS
University of Notre Dame

Qin Wen
Department of Statistics
University of Chicago

Rebecca Willett
Department of Statistics
University of Chicago

Yi Yu
Department of Statistics
University of Warwick

Abstract

This paper addresses the problem of localizing change points in high-dimensional linear regression models with piecewise constant regression coefficients. We develop a dynamic programming approach to estimate the locations of the change points whose performance improves upon the current state-of-the-art, even as the dimensionality, the sparsity of the regression coefficients, the temporal spacing between two consecutive change points, and the magnitude of the difference of two consecutive regression coefficient vectors are allowed to vary with the sample size. Furthermore, we devise a computationally-efficient refinement procedure that provably reduces the localization error of preliminary estimates of the change points. We demonstrate minimax lower bounds on the localization error that nearly match the upper bound on the localization error of our methodology and show that the signal-to-noise condition we impose is essentially the weakest possible based on information-theoretic arguments. Extensive numerical results support our theoretical findings, and experiments on real air quality data reveal change points supported by historical information not used by the algorithm.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

1 INTRODUCTION

High-dimensional linear regression modeling has been extensively applied and studied over the last two decades due to the technological advancements in collecting and storing data from a wide range of application areas, including biology, neuroscience, climatology, finance, cybersecurity, to name but a few. There exist now a host of methodologies available to practitioners to fit high-dimensional sparse linear models, and their properties have been thoroughly investigated and are now well understood. See Bühlmann and van de Geer (2011) for recent reviews.

In this paper, we are concerned with a non-stationary variant of the high-dim linear regression model in which the data are observed as a time series and the regression coefficients are piece-wise stationary, with changes occurring at unknown times. We formally introduce our model settings next.

Model 1. *Let the data* $\{(x_t, y_t)\}_{t=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ *satisfy the model*

$$y_t = x_t^\top \beta_t^* + \varepsilon_t, \quad t = 1, \ldots, n$$

where $\{\beta_t^*\}_{t=1}^n \subset \mathbb{R}^p$ *is the unknown coefficient vector,* $\{x_t\}_{t=1}^n$ *are independent and identically distributed mean-zero sub-Gaussian random vectors with* $\mathbb{E}(x_t x_t^\top) = \Sigma$*, and* $\{\varepsilon_t\}_{t=1}^n$ *are independent mean-zero sub-Gaussian random variables with sub-Gaussian parameter bounded by* σ_ε^2 *and independent of* $\{x_t\}_{t=1}^n$*. In addition, there exists a sequence of change points* $1 = \eta_0 < \eta_1 < \ldots < \eta_{K+1} = n$ *such that* $\beta_t^* \neq \beta_{t-1}^*$*, if and only if* $t \in \{\eta_k\}_{k=1}^K$*.*

We consider a high-dimensional framework where the features of the above change-point model are allowed to change with the sample size n; see Assumption 1 below for details. Given data sampled from Model 1, our main task is to develop computationally-efficient

algorithms that can consistently estimate both the unknown number K of change points and the time points $\{\eta_k\}_{k=1}^K$, at which the regression coefficients change. That is, we seek *consistent* estimators $\{\hat{\eta}_k\}_{k=1}^{\widehat{K}}$, such that, as the sample size n grows unbounded, it holds with probability tending to 1 that

$$\widehat{K} = K \quad \text{and} \quad \epsilon = \max_{k=1,\ldots,K} |\widehat{\eta}_k - \eta_k| = o(\Delta),$$

where $\Delta = \min_{k=1,\ldots,K_1} \eta_k - \eta_{k-1}$ is the minimal spacing between consecutive change-points. We refer to the quantity ϵ as the *localization error rate*.

The model detailed above has already been considered in the recent literature. Lee et al. (2016), Kaul et al. (2018), Lee et al. (2018), among others, focused on the cases where there exists at most one true change point. Leonardi and Bühlmann (2016) and Zhang et al. (2015) considered multiple change points and devised consistent change point estimators, albeit with localization error rates worse than the one we establish in Theorem 1. Wang et al. (2019) also allowed for multiple change points in a regression setting and proposed a variant of the wild binary segmentation (WBS) method (Fryzlewicz, 2014), the performances thereof match the one of the procedure we study next. Recently, (Fryzlewicz, 2020) studied inferences for change point regression models in low dimensions. More detailed comparisons are further commentary can be found in Section 3.3.

In this paper, we make several theoretical and methodological contributions, summarized next, that improve the existing literature.

- We provide consistent change point estimators for Model 1. We allow for model parameters to change with the sample size n, including the dimensionality of the data, the entry-wise sparsity of the coefficient vectors, the number of change points, the smallest distance between two consecutive change points, and the smallest difference between two consecutive different regression coefficients. To the best of our knowledge, the theoretical results we provide in this paper are the sharpest in the existing literature. Furthermore, the proposed algorithms, based on the general framework described in (1), can be implemented using dynamic programming approaches and are computationally efficient.
- We devise a additional second step (Algorithm 1), called local refinement, that is guaranteed to deliver an even better localization error rate, even though directly optimizing (1) already provides the sharpest rates among the ones existing in the literature.
- We present information-theoretic lower bounds on both detection and localization, establishing the fundamental limits of localizing change points in Model 1. To the best of our knowledge, this is the first time such results are developed for Model 1. The lower bounds on the localisation and detection nearly match the upper bounds we obtained under mild conditions.
- We present extensive experimental results including simulated data and real data analysis, supporting our theoretical findings, and confirming the practicality of our procedures.

Throughout this paper, we adopt the following notation. For any set S, $|S|$ denotes its cardinality. For any vector v, let $\|v\|_2$, $\|v\|_1$, $\|v\|_0$ and $\|v\|_\infty$ be its ℓ_2-, ℓ_1-, ℓ_0- and entry-wise maximum norms, respectively; and let $v(j)$ be the jth coordinate of v. For any square matrix $A \in \mathbb{R}^{n\times n}$, let $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ be the smallest and largest eigenvalues of matrix A, respectively. For any pair of integers $s, e \in \{0, 1, \ldots, n\}$ with $s < e$, we let $(s, e] = \{s + 1, \ldots, e\}$ and $[s, e] = \{s, \ldots, e\}$ be the corresponding integer intervals.

2 METHODS

2.1 A Dynamic Programming Approach

To achieve the goal of obtaining consistent change point estimators, we adopt a dynamic programming approach, whiuch we summarize next. Let $\mathcal{P}$ be an integer interval partition of $\{1, \ldots, n\}$ into $K_\mathcal{P}$ intervals, i.e.

$$\mathcal{P} = \big\{\{1, \ldots, i_1 - 1\}, \{i_1, \ldots, i_2 - 1\}, \ldots, \{i_{K_\mathcal{P}-1}, \ldots, i_{K_\mathcal{P}} - 1\}\big\},$$

for some integers $1 < i_1 < \cdots < i_{K_\mathcal{P}} = n + 1$, where $K_\mathcal{P} \geq 1$. For a positive tuning parameter $\gamma > 0$, let

$$\widehat{\mathcal{P}} \in \arg\min_{\mathcal{P}} \left\{ \sum_{I\in\mathcal{P}} \mathcal{L}(I) + \gamma|\mathcal{P}| \right\}, \tag{1}$$

where $\mathcal{L}(\cdot)$ is an appropriate loss function to be specified below, $|\mathcal{P}|$ is the cardinality of $\mathcal{P}$ and the minimization is taken over all possible interval partitions of $\{1, \ldots, n\}$.

The change point estimator resulting from the solution to (1) is simply obtained by taking all the left endpoints of the intervals $I \in \widehat{\mathcal{P}}$, except 1. The optimization problem (1) is known as the *minimal partition problem* and can be solved using dynamic programming with an overall computational cost of order $O(n^2\mathcal{T}(n))$, where $\mathcal{T}(n)$ denotes the computational

cost of solving $\mathcal{L}(I)$ with $|I| = n$ (see e.g. Algorithm 1 in Friedrich et al., 2008).

To specialize the dynamic programming algorithm to the high-dimensional linear regression model of interest by setting the loss function to be (1) with

$$\mathcal{L}(I) = \sum_{t\in I}(y_t - x_t^\top \widehat{\beta}_I^\lambda)^2, \tag{2}$$

where

$$\widehat{\beta}_I^\lambda = \mathop{\arg\min}_{v\in\mathbb{R}^p}\Bigg\{\sum_{t\in I}(y_t - x_t^\top v)^2 + \lambda\sqrt{\max\{|I|, \log(n\vee p)\}}\|v\|_1\Bigg\}, \tag{3}$$

and $\lambda \geq 0$ is a tuning parameter. The penalty is multiplied by the quantity $\max\{|I|, \log(n\vee p)\}$ in order to fulfill certain types of large deviation inequalities that are needed to ensure consistency.

Algorithms based on dynamic programming are widely used in the change point detection literature. Friedrich et al. (2008), Killick et al. (2012), Rigaill (2010), Maidstone et al. (2017), Wang et al. (2018b), among others, studied dynamic programming approaches for change point analysis involving a univariate time series with piecewise-constant means. Leonardi and Bühlmann (2016) examined high-dimensional linear regression change point detection problems by using a version of dynamic programming approach.

2.2 Local Refinement

We will show later in Theorem 1 that the localization error afforded by the dynamic programming approach in (1), (2), and (3) is linear in K, the number of change points. Although the corresponding localization rate is already sharper than any other rates previously established in the literature (see Section 3.3), it is possible to improve it by removing the dependence on K through an additional step, which we refer to as local refinement, detailed in Algorithm 1.

Algorithm 1 takes a sequence of preliminary change point estimators $\{\widetilde{\eta}_k\}_{k=1}^{\widetilde{K}}$, and refines each of the estimator $\widetilde{\eta}_k$ within the interval (s_k, e_k), which is a shrunken version of $(\widetilde{\eta}_{k-1}, \widetilde{\eta}_{k+1})$ (the constants 2/3 and 1/3 specifying the shrinking factor in the definition of $(\widetilde{\eta}_{k-1}, \widetilde{\eta}_{k+1})$ are not special and can be replaced by other values without affecting the rates). The shrinkage is applied to eliminate false positives, which are more likely to occur in the immediate proximity of a preliminary estimate of a change point. Since the refinement is done locally within each disjoint interval, the procedure is parallelizable. A group Lasso penalty is deployed in (4), and this is key to the success of the refinement. Intuitively, the group Lasso penalty integrates the information that the regression coefficients are piecewise-constant within each coordinate. Previously, Wang et al. (2019) also proposed a similar local screening algorithm based on the group Lasso estimators to refine the estimates of the regression change points. While Wang et al. (2019) assumed all the covariates to be uniformly bounded, we show that Algorithm 1 can achieve optimal localization error rates in a more general setting.

Algorithm 1 Local Refinement.

INPUT: Data $\{(x_t, y_t)\}_{t=1}^n$, $\{\widetilde{\eta}_k\}_{k=1}^{\widetilde{K}}$, $\zeta > 0$.
$(\widetilde{\eta}_0, \widetilde{\eta}_{\widetilde{K}+1}) \leftarrow (0, n)$
for $k = 1, \ldots, \widetilde{K}$ **do**
$(s_k, e_k) \leftarrow (2\widetilde{\eta}_{k-1}/3 + \widetilde{\eta}_k/3, \widetilde{\eta}_k/3 + 2\widetilde{\eta}_{k+1}/3)$

$$\begin{aligned}\left(\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\eta}_k\right) \leftarrow \mathop{\arg\min}_{\substack{\eta\in\{s_k+1,\ldots,e_k-1\}\\ \beta_1,\beta_2\in\mathbb{R}^p,\, \beta_1\neq\beta_2}}\Bigg\{&\sum_{t=s_k+1}^{\eta}\left\|y_t - \beta_1^\top x_t\right\|_2^2 \\ &+ \sum_{t=\eta+1}^{e_k}\left\|y_t - \beta_2 x_t\right\|_2^2 \\ &+ \zeta\sum_{i=1}^{p}\sqrt{(\eta - s_k)(\beta_1)_i^2 + (e_k - \eta)(\beta_2)_i^2}\Bigg\}\end{aligned} \tag{4}$$

end for
OUTPUT: $\{\widehat{\eta}_k\}_{k=1}^{\widetilde{K}}$.

3 MAIN RESULTS

In this section, we derive high-probability bounds on the localization errors of our main procedure based on the dynamic programming algorithm as detailed in equations 1, 2, and 3, and of the local refinement procedure of Algorithm 1.

3.1 Assumptions

We begin by stating the assumptions we require in order to derive localization error bounds.

Assumption 1. *Consider the model defined in Model 1. We assume that, for some fixed positive constants C_β, c_x, C_x, ξ, and C_{SNR} the following holds:*

a. *(Sparsity). Let $d_0 = |S|$. There exists a subset $S \subset \{1, \ldots, p\}$ such that*

$$\beta_t^*(j) = 0, \quad t = 1, \ldots, n, \quad j \in S^c = \{1, \ldots, p\} \setminus S.$$

b. *(Boundedness). For some absolute constant $C_\beta > 0$, $\max_{t=1,\ldots,n}\|\beta_t^*\|_\infty \leq C_\beta$.*

c. *(Minimal eigenvalue). We have that* $\Lambda_{\min}(\Sigma) = c_x^2 > 0$ *and* $\max_{j=1,\ldots,p}(\Sigma)_{jj} = C_x^2 > 0$.

d. *(Signal-to-noise ratio). Let* $\kappa = \min_{k=1,\ldots,K+1} \|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2$ *and* $\Delta = \min_{k=1,\ldots,K+1}(\eta_k - \eta_{k-1})$ *be the minimal jump size and minimal spacing defined as follows, respectively. Then,*

$$\Delta\kappa^2 \geq C_{\text{SNR}} d_0^2 K \sigma_\varepsilon^2 \log^{1+\xi}(n \vee p). \tag{5}$$

Assumption 1(**a**) and (**c**) are standard conditions required for consistency of Lasso-based estimators. Assumption 1(**d**) specifies a minimal signal-to-noise ratio condition that allows to detect the presence of a change point. Interestingly, if $K = d_0 = 1$, (5) reduces to $\Delta\kappa^2\sigma_\epsilon^{-2} \gtrsim \log^{1+\xi}(n \vee p)$, matching the information theoretic lower bound (up to constants and logarithmic terms) for the univariate mean change point detection problem (see e.g. Chan and Walther, 2013; Frick et al., 2014; Wang et al., 2018b).

In addition, we have

$$\begin{aligned}\Delta &\geq \frac{C_{\text{SNR}} d_0^2 K \sigma_\epsilon^2 \log^{1+\xi}(n \vee p)}{\kappa^2} \\ &\geq \frac{C_{\text{SNR}} d_0^2 K \sigma_\epsilon^2 \log^{1+\xi}(n \vee p)}{4C_\beta^2 d_0} \\ &\geq \frac{C_{\text{SNR}}}{4C_\beta^2} d_0 K \sigma_\epsilon^2 \log^{1+\xi}(n \vee p),\end{aligned} \tag{6}$$

where the second inequality follows from the bound

$$\kappa^2 = \min_{k=1,\ldots,K+1} \|\beta^*_{\eta_k} - \beta^*_{\eta_{k-1}}\|_2^2 \leq d_0(2C_\beta)^2 = 4C_\beta^2 d_0.$$

If $\Delta = \Theta(n)$ and $K = O(1)$, then (6) becomes $n \gtrsim d_0 \log^{1+\xi}(n \vee p)$, which resembles the effective sample size condition needed in the Lasso estimation literature.

Another way to interpret the signal-to-noise ratio Assumption 1(**d**) is to introduce a normalized jump size $\kappa_0 = \kappa/\sqrt{d_0}$, which leads to the equivalent condition

$$\Delta\kappa_0^2 \geq C_{\text{SNR}} d_0 K \sigma_\epsilon^2 \log^{1+\xi}(n \vee p).$$

Analogous constraints on the model parameters are required in other change point detection problems, including high-dimensional mean change point detection (Wang and Samworth, 2018), high-dimensional covariance change point detection (Wang et al., 2017), sparse dynamic network change point detection (Wang et al., 2018a), high-dimensional regression change point detection (Wang et al., 2019), to name but a few. Note that in these aforementioned papers, when variants of wild binary segmentation (Fryzlewicz, 2014) were deployed, additional knowledge is needed to get rid of K in the lower bound of the signal-to-noise ratio. We refer the reader to Wang et al. (2018a) for more discussions regarding this point.

The constant ξ is needed to guarantee consistency when Δ is of the same irder as n but can be set to zero if $\Delta = o(n)$. We may instead replace it with a weaker condition of the form

$$\Delta\kappa^2 \gtrsim C_{\text{SNR}} d_0^2 K\{\log(n \vee p) + a_n\},$$

where $a_n \to \infty$ arbitrarily slow as $n \to \infty$. We stick with the signal-to-noise ratio condition (5) for simplicity.

3.2 Localization Rates

We are now ready to state one of the main results of the paper.

Theorem 1. *Assume Model 1 and the conditions in Assumption 1. Then, the change point estimators* $\{\widetilde{\eta}_k\}_{k=1}^{\widehat{K}}$ *obtained as a solution to the dynamic programming optimization problem given in (1), (2), and (3) with tuning parameters*

$$\lambda = C_\lambda \sigma_\varepsilon \sqrt{d_0 \log(n \vee p)}$$

and

$$\gamma = C_\gamma \sigma_\epsilon^2 (K+1) d_0^2 \log(n \vee p),$$

are such that

$$\begin{aligned}\mathbb{P}\left\{\widehat{K} = K, \max_{k=1,\ldots,K} |\widetilde{\eta}_k - \eta_k| \leq \frac{KC_\epsilon d_0^2 \sigma_\epsilon^2 \log(n \vee p)}{\kappa^2}\right\} \\ \geq 1 - C(n \vee p)^{-c},\end{aligned} \tag{7}$$

where $C_\lambda, C_\gamma, C_\epsilon, C, c > 0$ *are absolute constants depending only on* C_β, C_x, *and* c_x.

The above result implies that, with probability tending to 1 as n grows,

$$\begin{aligned}\max_{k=1,\ldots,K} \frac{|\hat{\eta}_k - \eta_k|}{\Delta} &\leq \frac{KC_\epsilon d_0^2 \sigma_\epsilon^2 \log(n \vee p)}{\kappa^2 \Delta} \\ &\leq \frac{C_\epsilon}{C_{\text{SNR}} \log^{\xi}(n \vee p)} \to 0,\end{aligned}$$

where in the second inequality we have used Assumption 1(**c**). Thus, the localization error converges to zero in probability.

It is worth emphasizing that the bound in (7) along with Model 1 provide a *family* of rates, depending on how the model parameters (p, d_0, κ, Δ, K ands σ_ϵ) scale with n.

The tuning parameter λ affects the performance of the Lasso estimator. The second tuning parameter γ prevents overfitting while searching the optimal partition

as a solution to the problem (1). In particular, γ is determined by the squared ℓ_2-loss of the Lasso estimator and is of order $\lambda^2 d_0$. We will elaborate more on this point in the supplementary materials.

We now turn to the analysis of the local refinement algorithm, which takes as input a preliminary collection of change point estimators $\{\widetilde{\eta}_k\}_{k=1}^K$ such that $\max_{k=1,\ldots,K} |\widetilde{\eta}_k - \eta_k|$, such as the ones returned by our estimator based on the dynamic programming approach. The only assumption required for local refinement is that the localization error of the preliminary estimators be a small enough fraction of the minimal spacing Δ (see (8) below). Then local refinement returns an improved collection of change point estimators $\{\widehat{\eta}_k\}_{k=1}^K$ with a vanishing localization error rate of order $O\left(\frac{d_0 \log(n\vee p)}{n\kappa^2}\right)$. Interestingly, the initial estimators need not be consistent in order for local refinement to work: all that is required is essentially that the each of the working intervals in Algorithm 1 contains one and only one true change point. This fact allows us to refine the search within each working intervals separately, yielding better rates.

In particular, if we use the outputs of (1), (2), and (3) as the inputs of Algorithm 1, then it follows from (7) and (5) that, for any $k \in \{1, \ldots, K\}$,

$$\begin{aligned} s_k - \eta_{k-1} &> \frac{2}{3}\widetilde{\eta}_{k-1} + \frac{1}{3}\widetilde{\eta}_k - \widetilde{\eta}_{k-1} - \epsilon \\ &= \frac{1}{3}(\widetilde{\eta}_k - \widetilde{\eta}_{k-1}) - \epsilon > \Delta/3 - 5\epsilon/3 > 0 \end{aligned}$$

and

$$\begin{aligned} s_k - \eta_k &< \frac{2}{3}\widetilde{\eta}_{k-1} + \frac{1}{3}\widetilde{\eta}_k - \widetilde{\eta}_k + \epsilon \\ &= -\frac{2}{3}(\widetilde{\eta}_{k-1} - \widetilde{\eta}_k) + \epsilon < -2\Delta/3 + 5\epsilon/3 < 0. \end{aligned}$$

Corollary 2. *Assume the same conditions of Theorem 1. Let $\{\widetilde{\eta}_k\}_{k=1}^K$ be a set of time points satisfying*

$$\max_{k=1,\ldots,K} |\widetilde{\eta}_k - \eta_k| \le \Delta/7. \tag{8}$$

Let $\{\widehat{\eta}_k\}_{k=1}^{\widehat{K}}$ be the change point estimators generated from Algorithm 1 with $\{\widetilde{\eta}_k\}_{k=1}^K$ and

$$\zeta = C_\zeta \sqrt{\log(n \vee p)}$$

as inputs. Then,

$$\begin{aligned} \mathbb{P}\left\{ \widehat{K} = K, \max_{k=1,\ldots,K} |\hat{\eta}_k - \eta_k| \le \frac{C_\epsilon d_0 \log(n\vee p)}{\kappa^2} \right\} \\ \ge 1 - Cn^{-c}, \end{aligned}$$

where $C_\zeta, C_\epsilon, C, c > 0$ are absolute constants depending only on $C_\beta, \mathcal{M}$ and c_x.

Compared to the localization error given in Theorem 1, the improved localization error guaranteed by the local refinement algorithm does not have a direct dependence on K, the number of change points. The intuition for this is as follows. First, due to the nature of the change point detection problem, there is a natural group structure. This justifies the use of the group Lasso-type penalty, which reduces the localization error by bringing down d_0^2 to d_0. Second, using condition (8), there is one and only one true change point in every working interval used by the local refinement algorithm. The true change points can then be estimated separately using K independent searches, in such a way that the final localization rate that does not depend on the number of searches, namely K.

3.3 Comparisons

We now discuss how our contributions compared with the results of Wang et al. (2019) and of Leonardi and Bühlmann (2016), which investigate the same high-dimensional change-point linear regression model.

Wang et al. (2019) proposed different algorithms, all of which are variants of wild binary segmentation, with or without additional Lasso estimation procedures. Those methods inherit both the advantages and the disadvantages of WBS. Compared with dynamic programming, WBS-based methods require additional tuning parameters such as randomly selected intervals as inputs. With these additional tuning parameters, Theorem 1 in Wang et al. (2019) achieved the same statistical accuracy in terms of the localization error rate as Theorem 1 above. In terms of computational cost, the methods in Wang et al. (2019) are of order $O(K^2 n \cdot \text{Lasso}(n))$, where K, n and $\text{Lasso}(n)$ denote the number of change points, the sample size and the computational cost of Lasso algorithm with sample size n, respectively, while the dynamic programming approach of this paper is of order $O(n^2 \cdot \text{Lasso}(n))$. Thus, when $K \lesssim \sqrt{n}$, the algorithm in Wang et al. (2019) is computationally more efficient, but when $K \gtrsim \sqrt{n}$, the method in this paper has smaller complexity.

Leonardi and Bühlmann (2016) analysed two algorithms, one based on a dynamic programming approach, and the other on binary segmentation, and claimed that they both yield the same localization, which is, in our notation,

$$\sum_{k=1}^K |\hat{\eta}_k - \eta_k| \lesssim \frac{d_0^2 \sqrt{n \log(np)}}{\kappa^2}. \tag{9}$$

Note that, the error bound in Leonardi and Bühlmann (2016) is originally of the form $\sum_{k=1}^K |\hat{\eta}_k - \eta_k| \lesssim$

$\frac{d_0\sqrt{n\log(np)}}{\kappa^2}$ under a slightly stronger assumption than ours. In the more general settings of Assumption 1, the localization error bound of Leonardi and Bühlmann (2016) is of the form (9), based on personal communication with the authors.

It is not immediate to directly compare the sum of all localization errors, used by Leonardi and Bühlmann (2016), with the maximum localization error, which is the target in this paper. Using a worst-case upper bound, Theorem 1 yields that

$$\sum_{k=1}^{K}|\widehat{\eta}_k-\eta_k|\lesssim\frac{K^2d_0^2\sigma_\varepsilon^2\log(n\vee p)}{\kappa^2}.$$

In light of Corollary 2, this error bound can be sharpened, using the local refinement Algorithm 1 to

$$\sum_{k=1}^{K}|\widehat{\eta}_k-\eta_k|\lesssim\frac{Kd_0\sigma_\varepsilon^2\log(n\vee p)}{\kappa^2}.$$

As long as $K^2\lesssim\sqrt{\frac{n}{\log(np)}}$, or, using the local refinement algorithm, $K\lesssim\sqrt{\frac{n}{\log(np)}}$, our localization rates are better than the one implied by (9). It is not immediate to compare directly the assumptions used in Leonardi and Bühlmann (2016) with the ones we formulate here due to the different ways we use to present them. For instance, the conditions in Theorem 3.1 of Leonardi and Bühlmann (2016) imply, in our notation, that condition

$$\Delta\gtrsim\sqrt{n\log(p)}$$

is needed for consistency, even if the sparsity parameter $d_0=\Theta(1)$. However in our case, in view of (5), if we assume $d_0=\kappa=\Theta(1)$, then we only require $\Delta\gtrsim\log^{1+\xi}(n\vee p)$ for consistency.

3.4 Lower Bounds

In Section 3.2, we show that as long as

$$\kappa^2\Delta\gtrsim d_0^2K\sigma_\varepsilon^2\log^{1+\xi}(n\vee p),$$

we demonstrate provide change point estimators with localization errors upper bounded by

$$d_0\sigma_\varepsilon^2\kappa^{-2}\log(n\vee p).$$

In this section, we show that no algorithm is guaranteed to be consistent in the regime

$$\kappa^2\Delta\lesssim d_0\sigma_\varepsilon^2,$$

and otherwise, a minimax lower bound on the localization errors is

$$d_0\sigma_\varepsilon^2\kappa^{-2}.$$

These findings are formally stated next, in Lemmas 3 and 4, respectively.

Lemma 3. *Let $\{(x_t,y_t)\}_{t=1}^T\subset\mathbb{R}^p\times\mathbb{R}$ satisfy Model 1 and Assumption 1, with $K=1$. In addition, assume that $\{x_t\}_{t=1}^n\overset{iid}{\sim}\mathcal{N}(0,I_p)$ and $\{\varepsilon_t\}_{t=1}^n\overset{iid}{\sim}\mathcal{N}(0,\sigma_\varepsilon^2)$. Let $P^T_{\kappa,\Delta,\sigma_\varepsilon,d}$ be the corresponding joint distribution. For any $0<c<\frac{2}{8e+1}$, consider the class of distributions*

$$\mathcal{P}=\Big\{P^T_{\kappa,\Delta,\sigma_\varepsilon,d}:\Delta=\min\left\{\lfloor cd_0\sigma_\varepsilon^2\kappa^{-2}\rfloor,\lfloor T/4\rfloor\right\},\\ 2cd_0\max\{d_0,2\}\le\Delta\Big\}.$$

There exists a $T(c)$, which depends on c, such that for all $T\ge T(c)$,

$$\inf_{\widehat{\eta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P(|\widehat{\eta}-\eta(P)|)\ge\Delta,$$

where $\eta(P)$ is the location of the change point of distribution P and the infimum is over all estimators of the change point.

Lemma 3 shows that if $\kappa^2\Delta\lesssim d_0\sigma_\varepsilon^2$, then

$$\frac{\inf_{\widehat{\eta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P(|\widehat{\eta}-\eta(P)|)}{\Delta}\ge1,$$

which implies that the localization error is not a vanishing fraction of Δ as the sample size grows unbounded.

Lemma 4. *Let $\{(x_t,y_t)\}_{t=1}^T\subset\mathbb{R}^p\times\mathbb{R}$ satisfy Model 1 and Assumption 1, with $K=1$. In addition, assume $\{x_t\}_{t=1}^n\overset{iid}{\sim}\mathcal{N}(0,I_p)$ and $\{\varepsilon_t\}_{t=1}^n\overset{iid}{\sim}\mathcal{N}(0,\sigma_\varepsilon^2)$. Let $P^T_{\kappa,\Delta,\sigma_\varepsilon,d}$ be the corresponding joint distribution. For any diverging sequence ζ_T, consider the class of distributions*

$$\mathcal{P}=\left\{P^T_{\kappa,\Delta,\sigma_\varepsilon,d}:\Delta=\min\left\{\lfloor\zeta_Td_0\sigma_\varepsilon^2\kappa^{-2}\rfloor,\lfloor T/4\rfloor\right\}\right\}.$$

Then

$$\inf_{\widehat{\eta}}\sup_{P\in\mathcal{P}}\mathbb{E}_P(|\widehat{\eta}-\eta(P)|)\ge\frac{cd_0\sigma_\varepsilon^2}{\kappa^2},$$

where $\eta(P)$ is the location of the change point of distribution P, the infimum is over all estimators of the change point and $c>0$ is an absolute constant.

Recalling all the results we have obtained, the change point localization task is either impossible when $\kappa^2\Delta\lesssim d_0\sigma_\varepsilon^2$ (in the sense that no algorithm is guaranteed to be consistent) or, when $\kappa^2\Delta\gtrsim d_0^2K\sigma_\varepsilon^2\log^{1+\xi}(n\vee p)$, it can be solved by our algorithms at nearly a minimax optimal rate.

In the intermediate case

$$d_0\sigma_\varepsilon^2\lesssim\kappa^2\Delta\lesssim d_0^2K\sigma_\varepsilon^2\log^{1+\xi}(n\vee p)$$

we are unable to provide a result one way or another. However, we remark that, if in addition, in Theorem 1,

we assume $\kappa \leq C$, for an absolute constant $C > 0$, then we are able to weaken the condition from $\kappa^2\Delta \gtrsim d_0^2 K\sigma_\varepsilon^2 \log^{1+\xi}(n\vee p)$ to $\kappa^2\Delta \gtrsim d_0 K\sigma_\varepsilon^2 \log^{1+\xi}(n\vee p)$, by almost identical arguments. This shows that, under the additional conditions

$$\max\{\kappa,\, K\} \leq C,$$

for an absolute constant $C > 0$, the condition is nearly optimal, off by a logarithmic factor.

4 NUMERICAL EXPERIMENTS

In this section, we investigate the numerical performances of our proposed methods, with efficient binary segmentation algorithm (EBSA) of Leonardi and Bühlmann (2016) as the competitor. We compare four methods: dynamic programming (DP, see 1, 2, and 3), EBSA, local refinement (Algorithm 1) initialized by DP (DP.LR), and local refinement (Algorithm 1) initialized by EBSA (EBSA.LR).

The evaluation metric considered is the scaled Hausdorff distance between the estimators $\{\hat{\eta}_k\}_{k=1}^{\hat{K}}$ and the truth $\{\eta_k\}_{k=1}^{K}$. To be specific, we report $d(\hat{\mathcal{C}}, \mathcal{C}) = n^{-1}\mathcal{D}(\hat{\mathcal{C}}, \mathcal{C})$, where

$$\mathcal{D}(\hat{\mathcal{C}}, \mathcal{C}) = \max\{\max_{\hat{\eta}\in\hat{\mathcal{C}}} \min_{\eta\in\mathcal{C}} |\hat{\eta} - \eta|, \max_{\eta\in\mathcal{C}} \min_{\hat{\eta}\in\hat{\mathcal{C}}} |\hat{\eta} - \eta|\},$$

$\mathcal{C} = \{\eta_k\}_{k=1}^{K}$ and $\hat{\mathcal{C}} = \{\hat{\eta}_k\}_{k=1}^{\hat{K}}$.

We consider both simulated data and a real-life public dataset on air quality indicators in Taiwan. The implementations for our approaches can be found at https://github.com/Willett-Group/changepoint_regression.

4.1 Tuning Parameter Selection

We adopt a cross-validation approach to choosing tuning parameters. Let samples with odd indices be the training set and even ones be the validation set. Recall that for the DP, we have two tuning parameters λ and γ, which we tune using a brute-force grid search. For each pair of tuning parameters, we conduct DP on the training set and obtain estimated change points. Within each estimated segment of the training set, we obtain $\hat{\beta}_t$ by (3). On the validation set, let $\hat{y}_t = x_t^\top \hat{\beta}_t$ and calculate the validation loss $(n/2)^{-1}\sum_{t \mod 2\equiv 0}(\hat{y}_t - y_t)^2$. The pair (λ, γ) is chosen to be the one corresponding to the lowest validation loss.

As for the simulated data, we use some prior knowledge of the truth to save some computational cost. To be specific, we let the odd index set be partitioned by the true change points and estimate β_t on these intervals. We then plot the mean squared errors of $\hat{\beta}_t$ across a range of values of λ and obtain an "optimal" λ. We choose the grid range of λ around the "optimal" λ. This step is to approximately locate the range of λ's value but this step will not be used in the real data experiment. The same procedure is conducted for the tuning parameter selection in EBSA.

For the local refinement algorithm, we let the estimated change points of DP or EBSA be the initializers of the local refinement algorithm. We then regard the initialization algorithm and local refinement as a self-contained method and tune all three parameters λ, γ, and ζ jointly. The tuning procedure is almost the same as we described above, except that we use

$$\widehat{\beta}_I^\lambda = \arg\min_{v\in\mathbb{R}^p} \left\{ \sum_{t\in I} (y_t - x_t^\top v)^2 + \zeta\sqrt{|I|}\|v\|_1 \right\}$$

to estimate β_t.

4.2 Simulations

Throughout this section, we let $n = 600$, $p = 200$, $K = 4$, $\Sigma = I$ and $\sigma_\epsilon = 1$. The true change points are at 121, 221, 351 and 451. Let $\beta_0 = (\beta_{0i}, i = 1, \ldots, p)^\top$, with $\beta_{0i} = 2^{-1}d_0^{-1/2}\kappa$, $i \in \{1, \ldots, d_0\}$, and zero otherwise. Let

$$\beta_t = \begin{cases} \beta_0, & t \in \{1, \ldots, 120\}, \\ -\beta_0, & t \in \{121, \ldots, 220\}, \\ \beta_0, & t \in \{221, \ldots, 350\}, \\ -\beta_0, & t \in \{351, \ldots, 450\}, \\ \beta_0, & t \in \{451, \ldots, 600\}. \end{cases}$$

We let $\kappa \in \{4, 5, 6\}$ and $d_0 \in \{10, 15, 20\}$. For each pair of κ and d_0, the experiment is repeated 100 times. The results are reported in Figure 4.2 and in Table 1 of Appendix 3.1.

Generally speaking, DP outperforms EBSA, and LR significantly improves upon EBSA and DP when DP doesn't give accurate results. LR is comparable with DP when the initial points estimated by DP are already good enough. Note that since for EBSA.LR we tune EBSA to optimize EBSA.LR's performance, it may well happen that the estimated number of change points K from EBSA.LR is much different than from EBSA.

4.3 Air Quality Data

In this subsection, we consider the air quality data from https://www.kaggle.com/nelsonchu/

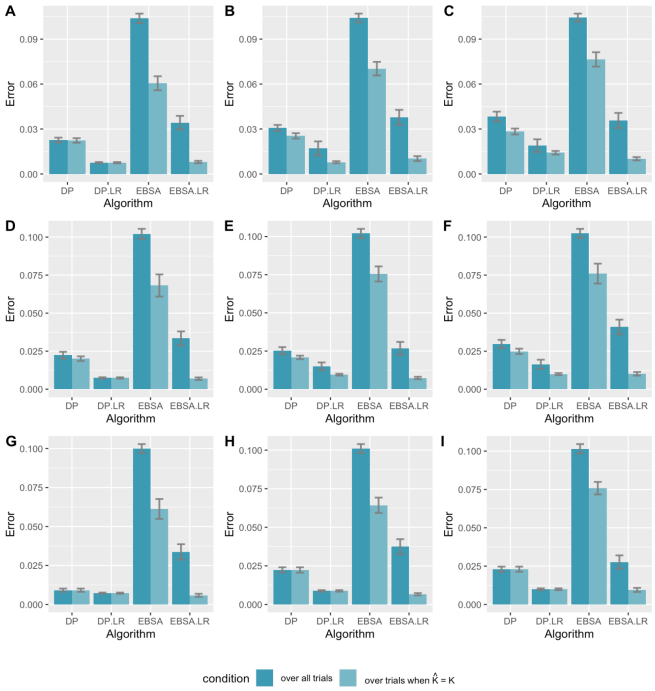

Figure 1: Bar plots for simulation results in Section 4.2. Plots A-C are settings with $\kappa = 4$ and $d_0 \in \{10, 15, 20\}$; Plots D-F are settings with $\kappa = 5$ and $d_0 \in \{10, 15, 20\}$; and Plots G-I are settings with $\kappa = 6$ and $d_0 \in \{10, 15, 20\}$.

air-quality-in-northern-taiwan. It collects environment information and air quality data from Northern Taiwan in 2015. We choose the PM10 in Banqiao as the response variables, with covariates being the temperature, the CO level, the NO level, the NO_2 level, the NO_x level, the rainfall quantity, the humidity quantity, the SO_2 level, the ultraviolet index, the wind speed, the wind direction and the PM10 levels in Guanyin, Longtan, Taoyuan, Xindian, Tamsui, Wanli and Keelung District. We transfer the hourly data into daily by averaging across 24 hours. After removing all dates containing missing values, we obtain a data set with $n = 343$ days and $p = 18$ covariates. Our goal is to detect potential change points of this data set and determine if they are consistent with the historical information.

We standardize the data so that the variance of $y_t = 1$, for all t. We then conduct DP, DP.LR, EBSA and EBSA.LR. DP estimates 2 change points which are March 16th and November 1st, 2015. No change points are detected by EBSA. DP.LR and EBSA.LR both detect May 15th and October 25th, 2015 as the change points.

The first change point detected by DP.LR and EBSA.LR seems to correspond with the first strong-enough typhoon near Northern Taiwan in 2015, which happened during May 6th-20th (e.g. Wikipedia, 2020a). The second change points from EBSA.LR, DP.LR and DP are relatively close and they all could be explained by the severe air pollution at the beginning of November in Taiwan, which reached the hazardous purple alert on November 8th (e.g. Wikipedia, 2020b). The visualization is shown in Figure 2.

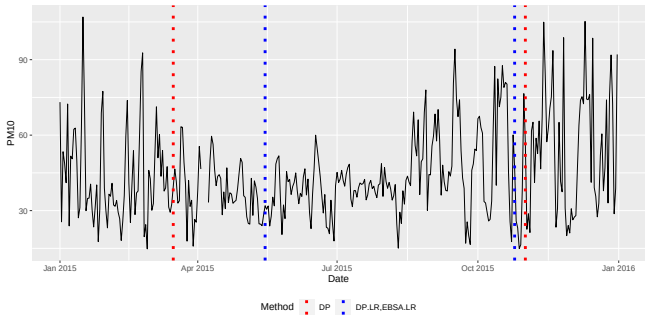

Figure 2: Real air quality example. When we tune EBSA for its standalone performance, we obtain zero change points, but when we tune EBSA to optimize the EBSA.LR's performance, we obtain two change points.

5 DISCUSSION

In this paper, we in fact provide a general framework for analyzing general regression-type change point localization problems that include the linear regression model above as a special case. The analysis in this paper may be utilized as a blueprint for more complex change point localization problems. In our analysis, we develop a new and refined toolbox for the change point detection community to study more complex data generating mechanisms above and beyond linear regression models.

References

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

Hock Peng Chan and Guenther Walther. Detection with the scan and the average likelihood ratio. *Statistica Sinica*, 1(23):409–428, 2013.

Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:495–580, 2014.

Felix Friedrich, Angela Kempe, Volkmar Liebscher,

and Gerhard Winkler. Complexity penalized m-estimation: Fast computation. *Journal of Computational and Graphical Statistics*, 17:201–204, 2008.

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

Piotr Fryzlewicz. Narrowest significance pursuit: inference for multiple change-points in linear models. *arXiv preprint arXiv:2009.05431*, 2020.

Abhishek Kaul, Venkata K Jandhyala, and Stergios B Fotopoulos. Parameter estimation for high dimensional change point regression models without grid search. *arXiv preprint arXiv:1805.03719*, 2018.

Rebecca Killick, Paul Fearnhead, and Idris A. Eckley. Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

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*, 113(523): 1184–1194, 2018.

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

R. Maidstone, T. Hocking, G. Rigaill, and P. Fearnhead. On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27:519–533, 2017.

G. Rigaill. Pruned dynamic programming for optimal multiple change-point detection. *arXiv preprint arXiv:1004.0887*, 2010.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*, 2017.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal change point detection and localization in sparse dynamic networks. *arXiv preprint arXiv:1809.09602*, 2018a.

Daren Wang, Yi Yu, and Alessandro Rinaldo. Univariate mean change point detection: Penalization, cusum and optimality. *arXiv preprint arXiv:1810.09498*, 2018b.

Daren Wang, Kevin Lin, and Rebecca Willett. Statistically and computationally efficient change point localization in regression settings. *arXiv preprint arXiv:1906.11364*, 2019.

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.

Wikipedia. 2015 pacific typhoon season. https://en.wikipedia.org/wiki/2015_Pacific_typhoon_season, 2020a.

Wikipedia. Air pollution in taiwan. https://en.wikipedia.org/wiki/Air_pollution_in_Taiwan, 2020b.

Bingwen Zhang, Jun Geng, and Lifeng Lai. Change-point estimation in high dimensional linear regression models via sparse group lasso. In *Communication, Control, and Computing (Allerton), 2015 53rd Annual Allerton Conference on*, pages 815–821. IEEE, 2015.