Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Two-Way Truncated Linear Regression Models with Extremely Thresholding Penalization

Hao Yang Teng & Zhengjun Zhang

To cite this article: Hao Yang Teng & Zhengjun Zhang (12 Dec 2022): Two-Way Truncated Linear Regression Models with Extremely Thresholding Penalization, Journal of the American Statistical Association, DOI: 10.1080/01621459.2022.2147074

To link to this article: https://doi.org/10.1080/01621459.2022.2147074

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–17: Theory and Methods
https://doi.org/10.1080/01621459.2022.2147074

Taylor & Francis
Taylor & Francis Group

Two-Way Truncated Linear Regression Models with Extremely Thresholding Penalization

Hao Yang Teng[a] and Zhengjun Zhang[b]

[a]Department of Mathematics and Statistics, Arkansas State University, Jonesboro, AR; [b]Department of Statistics, University of Wisconsin-Madison, Madison, WI

ABSTRACT

This article introduces a new type of linear regression model with regularization. Each predictor is conditionally truncated through the presence of unknown thresholds. The new model, called the two-way truncated linear regression model (TWT-LR), is not only viewed as a nonlinear generalization of a linear model but is also a much more flexible model with greatly enhanced interpretability and applicability. The TWT-LR model performs classifications through thresholds similar to the tree-based methods and conducts inferences that are the same as the classical linear model on different segments. In addition, the innovative penalization, called the extremely thresholding penalty (ETP), is applied to thresholds. The ETP is independent of the values of regression coefficients and does not require any normalizations of regressors. The TWT-LR-ETP model detects thresholds at a wide range, including the two extreme ends where data are sparse. Under suitable conditions, both the estimators for coefficients and thresholds are consistent, with the convergence rate for threshold estimators being faster than $\sqrt{n}$. Furthermore, the estimators for coefficients are asymptotically normal for fixed dimension p. It is demonstrated in simulations and real data analyses that the TWT-LR-ETP model illustrates various threshold features and provides better estimation and prediction results than existing models. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2020
Accepted November 2022

KEYWORDS
Asymptotics; Consistency; Extremely thresholding; High dimension; Interpretability; Predictability; Thresholds

1. Introduction

Linear regression is a widely applied and dominant statistical inference method for studying variable relationships due to its easy computability, interpretability, predictability, and stability (CIPS). In the meantime, many other new developments in the literature extend linear regression models to nonlinear regression models, nonparametric regressions, and semi-parametric regressions, such as generalized additive models. Still, in many applications, linear regressions are preferred.

Due to the desirable properties of the linear regression models, various regularization models have been proposed to deal with high-dimensional datasets and are widely used across different disciplines such as medicine, biology, public health, etc. A vast amount of literature has introduced different penalty functions to improve the estimation of the regression coefficients. Some models which are capable of handling high-dimensional data, but are not limited to, include the least absolute shrinkage and selection operator (Lasso) (Tibshirani 1996) with which Knight and Fu (2000) showed the estimation consistency of the Lasso-type estimators for fixed dimension and asymptotic normality of the estimators, the smoothly clipped absolute deviation (SCAD) (Fan and Li 2001), the elastic net (Zou and Hastie 2005), the adaptive Lasso (Zou 2006), the Dantzig selector (Candès and Tao 2007), the minimax concave penalty (MCP) (Zhang 2010), etc. The regularized linear regression model was later developed to select significant variables from a large number of predictors. These models have gained high popularity, and they have been discussed in many papers and explored in detail; see (Fan et al. 2020) and references therein for more details. It is important to note that these methods assume that the response variable is linearly related to the predictors with different penalty terms.

Change-point and threshold models have broad economic and time-series data applications, primarily close to the linear regression models by sharing the same framework, that is, linear in regression coefficients while introducing additional (nuisance) parameters in the nonlinear framework. There are many different models, such as the regression kink model, the threshold model, the two-phase regression, the segmented regression, and the broken line regression. There is a wide range of literature on the regressions and significant contributions can be found in Hinkley (1969), Feder (1975), Hansen (2000, 2017), Knowles, Siegmund, and Zhang (1991), and Siegmund and Zhang (1993, 1994). These models investigate the nonlinear relationships between the predictors and a response variable and study how the behavior changes at some points. These models have numerous real applications in time series as change-points are more observable for time series data. A key question that arises is about change-points detection and estimation. In some circumstances, the threshold or the change-point can be observed. However, in most cases, thresholds or change-points

CONTACT Zhengjun Zhang zjz@stat.wisc.edu Department of Statistics, University of Wisconsin-Madison, Madison, WI.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

are unknown to us. In the literature, many papers have developed estimation and inference theory with a single unknown threshold (Porter and Yu 2015; Hansen 2017). Porter and Yu (2015) developed the estimation and inference procedures for regression discontinuity with a finite and unknown number of thresholds. The threshold models provide some flexibility in modeling the relationships between the response variable and a set of predictors with thresholds, while another group of predictors has linear relationships with the response variable. Furthermore, most of these models can be reduced to a linear model when all predictors do not contain any threshold.

More attention is given to change-point and threshold detection using penalized regression models, especially in recent literature. Harchaoui and Lévy-Leduc (2012) proposed using a penalized least-square criterion with a ℓ_1 penalty to estimate the location of change-point in one-dimensional piecewise constant signals observed in white noise. Ciuperca (2014) and Zhang, Geng, and Lai (2015) considered a model with multiple change-points under a fixed covariate dimension setting. Lee, Seo, and Shin (2016) presented a model that selects between a linear model and a threshold regression model with a possible change-point in a high-dimensional setting based on a data-dependent ℓ_1 penalty. Leonardi and Bühlmann (2016) proposed a joint estimator for the change-points and coefficients with ℓ_1 regularization on the parameters in different segments for the case of multiple change-points. Kaul, Jandhyala, and Fotopoulos (2019a) considered a threshold model similar to Lee, Seo, and Shin (2016) and developed a two-step estimation procedure for a single change-point and coefficients by including the case of no change-point scenario. Kaul, Jandhyala, and Fotopoulos (2019b) further extended the two-step estimation procedure for multiple change-point detections in a high-dimensional setting. There are other methods in the literature for change-point detection in a high-dimensional setting. Wang and Samworth (2018) and Wang et al. (2021), among others, study the change-point detection in a high-dimensional setting via a projection-based method.

With ultra-high dimensional data becoming more readily available nowadays, there are dimension reduction techniques to handle ultra-high dimensional datasets in the literature. For instance, Fan and Lv (2008) introduced a ranking procedure based on the Pearson correlation to rank and select significant predictors. Subsequently, the Pearson correlation was extended for polynomial transformations of predictors, and the ranking procedure is based on a bootstrap procedure (Hall and Miller 2009). In addition, Li, Zhong, and Zhu (2012) proposed the sure independence screening procedure based on the distance correlation (DC-SIS). Chen et al. (2017) proposed the sure explained variability and independence screening (SEVIS) that incorporates the asymmetric and nonlinear generalized measures of correlation (Zheng, Shi, and Zhang 2012) in the screening process to perform dimension reduction. Other existing methods in the literature can be applied to the ultra-high dimensional datasets but were not mentioned here.

Furthermore, the idea of clustering and segmentation of the regression coefficients to achieve sparsity has been discussed in the literature. Ke, Fan, and Wu (2015) proposed a penalized least squares based method to detect homogeneity by ordering and clustering the regression coefficients through a clustering algorithm in regression via data-driven segmentation (CARDS). Ke, Li, and Zhang (2016) took a different approach to pursue homogeneity from a change-point perspective and considered a latent variable in their work. In addition, Tang and Song (2016) incorporated the idea of homogeneity to identify inter-study homogeneous parameter clusters using the fused lasso. There are several other extensions of the work by Ke, Fan, and Wu (2015) in other model setup and panel data structure. These work include, but are not limited to, panel data using linear model (Wang, Phillips, and Su 2018), nonlinear models (Wang and Su 2021), single-index model (Lian, Qiao, and Zhang 2021).

In this article, we focus on developing a more general threshold model. Our article's contributions to the literature can be concluded in 5-fold. (a) The two-way truncated linear regression (TWT-LR) model contains both linear and nonlinear relationships between different predictors and the response variable. The types of associations (i.e., linear or nonlinear) are modeled through the unknown threshold parameters without the need for prior information on whether the variables contain thresholds or change-points. (b) We introduce a penalty to the two-way truncated linear regression model by penalizing the number of thresholds for each variable with a tuning parameter λ_n to avoid an overfitting problem. The penalization, called extremely thresholding penalty (ETP), introduced in this article is different from the penalty functions in the literature, which penalize the regression coefficients. As a result, the theoretical derivations of the proposed estimators are challenging and nontrivial. Nevertheless, such a new penalty framework can shed new light on a broad area of new theoretical research and applications. (c) The TWT-LR-ETP model is developed to detect thresholds at a wide range of data, including the two extreme ends where data are sparse. (d) The convergence rate of the proposed estimators for the unknown thresholds is faster than the standard parametric rate $\sqrt{n}$ and the estimators for the regression coefficients are shown to be asymptotically normal. (e) Extensive simulation studies show that the TWT-LR-ETP model outperforms the existing models in modeling different types of associations. Furthermore, due to the flexibility in modeling and interpretability, the TWT-LR-ETP model illustrates various threshold features and provides better interpretable results for the four real datasets considered in this article than existing models. The final model, TWT-LR-ETP, can be a practical benchmark for modeling linear and nonlinear associations.

This article is organized as follows. We first introduce the TWT-LR model in Section 2.1 and discuss the flexibilities of the new regression model. Then, in Section 3.1, we will discuss the estimation procedures of the threshold and coefficient parameters of the TWT-LR-ETP model. The asymptotic properties, such as the consistency and asymptotic normality, are discussed in Section 3.2. Next, numerical studies are presented in Section 4. We will first discuss the computational procedures in Section 4.1 and show simulated examples in Section 4.2. Then, in Section 4.3, we will present the analyses and interpretations using a real dataset (with the other three real datasets in a supplementary materials). Finally, the concluding remarks are presented in Section 5. Technical arguments for two main theorems and lemmas are presented in Appendix A, supplementary materials. Additional details of the computational procedure, the results for the numerical experiments, and three additional

datasets are presented in Appendices B and C. Finally, some toy examples are presented in Appendix C, supplementary materials to illustrate the TWT-LR model. Appendices A–C are given in the supplementary materials.

2. Model Specification

In this section, we will begin by introducing some general notations used throughout the article and present the TWT-LR model that is flexible in modeling both linear and nonlinear relationships between a response variable and predictors in the presence of thresholds. We will subsequently present the new regression model that can be expressed in different forms under certain threshold specifications.

2.1. Regression Models with Two-Way Truncated Predictors

Suppose $Y_i, i = 1, 2, \ldots, n$, are the univariate responses; $(X_{i1}, X_{i2}, \ldots, X_{ip})$, $i = 1, 2, \ldots, n$, are p-dimensional predictors. As discussed in the introduction, a predictor X_{ij} may have different linear associations with the response variable depending on its values. As a result, X_{ij} can have different forms in the regression models, or it can be presented as different types of working variables, including itself and its threshold truncated variables. Suppose $c_j = (c_{j,1}, c_{j,2}), j = 1, 2, \ldots, p$, are p bivariate truncation thresholds associated to X_{ij}. Let $\tilde{\boldsymbol{c}} = (c_{1,1}, c_{1,2}, \ldots, c_{p,1}, c_{p,2})'$ be a $2p \times 1$ vector. Denote

$$\tilde{\mathbf{Z}}_i(\tilde{\boldsymbol{c}}) = (1, X_{i1}I(X_{i1} < c_{1,1}), X_{i1}I(X_{i1} > c_{1,2}), X_{i1}, \ldots, X_{ip}I(X_{ip} < c_{p,1}), X_{ip}I(X_{ip} > c_{p,2}), X_{ip})' \tag{2.1}$$

as an $m^* \times 1$ vector where $m^* = 3p + 1$, and $I(\cdot)$ is an indicator function. The parameters $c_{j,1}$ and $c_{j,2}$ in the vector $\tilde{\mathbf{Z}}_i(\tilde{\boldsymbol{c}})$ detect thresholds for a particular jth variable. For instance, when two thresholds are detected for the jth variable, both the parameters $c_{j,1}$ and $c_{j,2}$ take values between $-\infty$ and ∞. To unify all working variables in (2.1) in a two-way truncated forms, we introduce an additional parameter $c_{j,3}$. The parameter $c_{j,3}$ is an additional extreme parameter that only takes values at the two extremes (i.e., $-\infty$ and ∞) for all j. The meanings and functions of each $c_{j,k}$, $k = 1, 2, 3$ will be explained throughout this section using examples. Let $\boldsymbol{c} = (c_{1,1}, c_{1,2}, c_{1,3}, \ldots, c_{p,1}, c_{p,2}, c_{p,3})'$ be a $3p \times 1$ vector. Denote

$$\mathbf{Z}_i(\boldsymbol{c}) = (1, X_{i1}I(X_{i1} < c_{1,1}), X_{i1}I(X_{i1} > c_{1,2}), X_{i1}I(X_{i1} < c_{1,3}), \ldots, X_{ip}I(X_{ip} < c_{p,1}), X_{ip}I(X_{ip} > c_{p,2}), X_{ip}I(X_{ip} < c_{p,3}))' \tag{2.2}$$

as an $m \times 1$ vector where $m = 3p + 1$. Let $\mathbb{Z}(\boldsymbol{c})$ denote the $(n \times m)$ matrix whose ith row is $\mathbf{Z}_i'(\boldsymbol{c})$, and $\boldsymbol{\beta}_0 = (\beta_{0,0}, \beta_{1,1,0}, \beta_{1,2,0}, \beta_{1,3,0}, \ldots, \beta_{j,1,0}, \beta_{j,2,0}, \beta_{j,3,0}, \ldots, \beta_{p,1,0}, \beta_{p,2,0}, \beta_{p,3,0})'$ be an $m \times 1$ vector, where $\beta_{0,0}$ is the intercept, and $(\beta_{j,1,0}, \beta_{j,2,0}, \beta_{j,3,0})$ are unknown regression coefficients associated to X_{ij} and $(c_{j,1,0}, c_{j,2,0}, c_{j,3,0})$ in the new linear model. Let $\boldsymbol{c}_0 = (c_{1,1,0}, c_{1,2,0}, c_{1,3,0}, \ldots, c_{p,1,0}, c_{p,2,0}, c_{p,3,0})'$ be the unknown threshold parameters. For any n-dimensional vector $V = (V_1, \ldots, V_n)'$, define the L_2 norm as $||V||_2 = (\sum_{i=1}^n V_i^2)^{1/2}$. For any $n \times n$ matrix U, $||U||_2$ denotes spectral norm and $||U||_F$ denotes Frobenius norm.

It can be seen in $\mathbf{Z}_i(\boldsymbol{c})$, a predictor X_{ij} is truncated in two ways: from below and above (i.e., $I(X_{i1} < c_{1,1})$ and $I(X_{i1} > c_{1,2})$) to form two "new" predictors. With the consideration of the extreme parameter in $\mathbf{Z}_i(\boldsymbol{c})$, an additional "new" predictor is included. Since the extreme parameter only takes values at the two extremes, the indicator function for the extreme parameter does not truncate the data. As such, the proposed model can be called a two-way truncated linear regression model (TWT-LR). Let $\boldsymbol{c}_{1,0} = (c_{1,1,0}, c_{2,1,0}, \ldots, c_{p,1,0})'$, $\boldsymbol{c}_{2,0} = (c_{1,2,0}, c_{2,2,0}, \ldots, c_{p,2,0})'$, $\boldsymbol{c}_{3,0} = (c_{1,3,0}, c_{2,3,0}, \ldots, c_{p,3,0})'$ be three $p \times 1$ true threshold and extreme parameter vectors. Denote $\bar{\mathbb{K}}_1 = \mathbb{K}_1^p$, $\bar{\mathbb{K}}_2 = \mathbb{K}_2^p$ and $\bar{\mathbb{K}}_3 = \mathbb{K}_3^p$ as the parameter spaces for the true parameters $\boldsymbol{c}_{1,0}$, $\boldsymbol{c}_{2,0}$, and $\boldsymbol{c}_{3,0}$, respectively, where $\mathbb{C} \subset R$, $\mathbb{K}_1 = \{-\infty\} \cup \mathbb{C}$, $\mathbb{K}_2 = \mathbb{C} \cup \{\infty\}$, $\mathbb{K}_3 = \{-\infty\} \cup \{\infty\}$. Let $\mathbb{B} \subset R^m$ be the parameter space for the true parameters $\boldsymbol{\beta}_0$.

Using the notations defined above, the TWT-LR model is expressed as

$$Y_i = \mathbf{Z}_i'(\boldsymbol{c})\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.3}$$

where $\boldsymbol{c}$ and $\boldsymbol{\beta}$ are defined the same as $\boldsymbol{c}_0$ and $\boldsymbol{\beta}_0$.

We now present the forms and sparsity of the model (2.3) under different threshold specifications in univariate case. If $c_1 = -\infty$, $c_2 = \infty$ and $c_3 = \infty$, X_i is linearly associated to Y_i which is the same as the regular linear regression model and takes the following form

$$Y_i = \beta_0 + \beta_3 X_i + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.4}$$

If $c_1 = -\infty$, $c_2 = \infty$ and $c_3 = -\infty$, the predictor X_i is insignificant. If $c_1 = -\infty$, $-\infty < c_2 < \infty$ and $c_3 = -\infty$, the predictor X_i is a one-way truncated below variable in the model. If $-\infty < c_1 < \infty$, $c_2 = \infty$ and $c_3 = -\infty$, the predictor X_i is a one-way truncated above variable in the model. Accordingly, the TWT-LR model can be expressed as

$$Y_i = \beta_0 + \beta_2 X_i I(X_i > c_2) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.5}$$

and

$$Y_i = \beta_0 + \beta_1 X_i I(X_i < c_1) + \epsilon_i, \quad i = 1, \ldots, n. \tag{2.6}$$

Furthermore, the TWT-LR model with a threshold at c^* takes the following form

$$Y_i = \beta_0 + \beta_1 X_i I(X_i < c^*) + \beta_2 X_i I(X_i > c^*) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.7}$$

where $c_1 = c_2 = c^*$, $-\infty < c^* < \infty$ and $c_3 = -\infty$.

The model (2.7) can be expressed in three other ways: (i) $c_1 \in \mathbb{C}$, $c_2 \in \mathbb{C}$ and $c_3 = \infty$ where $c_1 = c_2 = c^*$, (ii) $c_1 = -\infty$, $c_2 \in \mathbb{C}$ and $c_3 = \infty$ or (iii) $c_1 \in \mathbb{C}$, $c_2 = \infty$ and $c_3 = \infty$. As a result, these cases cause an identifiability issue in the parameter estimation procedure. To resolve the issue, we restrict our case for one threshold parameter estimation to the form specified in model (2.7).

If a predictor has two thresholds at different points c_1 and c_2, the TWT-LR model is expressed as

$$Y_i = \beta_0 + \beta_1 X_i I(X_i < c_1) + \beta_2 X_i I(X_i > c_2) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.8}$$

with $c_3 = -\infty$, $c_1 < c_2$, and

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i I(X_i < c_1) + \beta_2 X_i I(X_i > c_2) + \beta_3 X_i + \epsilon_i, \\ &= \beta_0 + (\beta_1 + \beta_3) X_i I(X_i < c_1) + \beta_3 X_i I(c_1 \le X_i \le c_2) \\ &\quad + (\beta_2 + \beta_3) X_i I(X_i > c_2) + \epsilon_i, \end{aligned} \tag{2.9}$$

where $c_3 = \infty$ for $i = 1, \ldots, n$. When the extreme parameter $c_3 = -\infty$ with two thresholds c_1 and c_2, the data in the middle segment has no linear association between the response variable and the predictor which is shown in model (2.8). On the other hand, when the extreme parameter $c_3 = \infty$ with two thresholds, there is a linear association between the response variable and the predictor in the middle segment of the data shown in model (2.9). Therefore, the extreme parameter c_3 provides more flexibility in modeling the middle segment of the data. For p numbers of predictors, the parameter $c_{j,3}$ controls the significance of a particular segment or variable X_{ij}, that is, to model the middle segment while selecting significant variables. In other words, the parameter $c_{j,3}$ can be viewed as a significance parameter to X_{ij}. In Section 3.1, we will further present and discuss the important roles of the parameter. For this article, we refer the parameter $c_{j,3}$ to as an extreme parameter when only the parameter $c_{j,3}$ is mentioned whereas the parameters $c_{j,1}, c_{j,2}$, and $c_{j,3}$ are referred jointly to as threshold parameters to avoid confusion. We note that $c_{j,1}$ and/or $c_{j,2}$ can take values close to the two extreme ends of the data range of X_{ij}, which is particularly meaningful in some applications. Such thresholds $c_{j,1}$ and $c_{j,2}$ can also be called extreme value thresholds.

The above model equations present how the association between a predictor and a response variable changes and sparsity in our setup when the thresholds are specified differently. It is easy to see the truncation in model (2.3) greatly enhances the applicability of linear regression models, boosts the prediction power, and produces better interpretable and meaningful results. In addition, it is important to emphasize that the extreme parameter in model (2.3) increases the flexibilities in modeling linear, insignificant associations and varying linear associations in three segments. Model (2.3) leads to not only a better applicable new-type regression model but also a new way of variable selection in the literature. Some toy examples are presented in Appendix C, supplementary materials to illustrate the models (2.4)–(2.9). The new theory and methodology of using thresholds in the penalty will be presented in Section 3.

Clearly, differentiating thresholds in different categories in our model is significantly different from the literature. In our model, first, when the thresholds $c_{j,1}$ and $c_{j,2}$ fall in the center of the data distribution, the estimation can be performed to obtain highly accurate parameter estimates with lower uncertainties using readily available models and approaches due to the bulk of data around the thresholds. Second, when thresholds exist in the both extreme ends (i.e., not $-\infty$ and ∞), the behaviors of the predictor can provide important insights into studying the changing associations between the extreme values and the response variable. For example, the extreme events above a certain threshold that exists in the extreme tend to have a higher impact on the response variable. For instance, if the extreme weather conditions surpass an extreme threshold level, the insurance company has to deal with dire impacts brought by the more extreme weather on the damage costs. The extreme weather may have more devastating damage to houses, cars, etc., resulting in a stronger linear association in the extreme end between the extreme weather conditions and damage costs than the weather conditions below the threshold. For such data, the data points are usually sparse in the extreme ends, thereby increasing uncertainties in statistical inferences. We aim to develop the TWT-LR model to detect thresholds that exist in the extreme ends while offering more flexibility in modeling different linear associations and reducing the need to identify the threshold variable in advance. Moreover, the structure of the TWT-LR model can be expressed as a tree structure shown in Figures C10 and C11 in the Appendix, supplementary materials. On the other hand, in our model settings, if there are k covariates with each having two thresholds and an additional extreme threshold, we will have at most 3^k heterogeneous sub-populations in contrast to only one population in a classical linear regression model, which shows great advantages of our new model in modeling the changing trends of the sub-populations. These observations shape the idea of introducing the thresholds in the TWT-LR model and penalty term in our setup, which offers a different approach to modeling from the literature on regularized regression.

3. Estimation and Asymptotic Theory

3.1. Estimation

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)'$. Consider the mean of squared residuals

$$S_n(\boldsymbol{\beta}, \boldsymbol{c}) = n^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{Z}_i'(\boldsymbol{c})\boldsymbol{\beta})^2. \tag{3.1}$$

For $j = 1, 2, \ldots, p$ and $i = 1, 2, \ldots, n$, let

$$W_{ij}(c_j) = [I(X_{ij} < c_{j,1}) + I(X_{ij} > c_{j,2}) + I(X_{ij} < c_{j,3})],$$

where $c_j = (c_{j,1}, c_{j,2}, c_{j,3})$ for all j. The parameter spaces for the true parameters $\boldsymbol{c}_{1,0}$, $\boldsymbol{c}_{2,0}$, $\boldsymbol{c}_{3,0}$, and $\boldsymbol{\beta}_0$ in model (2.3) defined in Section 2.1 are $\mathbb{K}_1$, $\mathbb{K}_2$, $\mathbb{K}_3$, and $\mathbb{B}$. In the previous section, we discussed the identifiability issue for one threshold resulted from the parameter spaces. Subsequently, we respecify the parameter spaces for the threshold and extreme parameters. Denote $\Theta_1 = \mathbb{K}_1 \times \mathbb{K}_2 \times \{-\infty\}$, $\Theta_2 = \mathbb{C}_{12} \times \{\infty\}$, where $\mathbb{C}_{12} = \{(c_1, c_2) : c_1 < c_2, c_1 \in R, c_2 \in R\}$ and $\Theta_3 = \{-\infty\} \times \{\infty\} \times \{\infty\}$ and we further let $\Theta = \Theta_1 \cup \Theta_2 \cup \Theta_3$. The parameter space for the true parameters $\boldsymbol{c}_0$ is $\tilde{\Theta} = \Theta^p$. The parameter space for $\boldsymbol{\beta}_0$ is specified in Section 2. Define the threshold and regression coefficient estimators by

$$(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{c}}) := \underset{\boldsymbol{\beta} \in \mathbb{B}, \boldsymbol{c} \in \tilde{\Theta}}{\operatorname{argmin}} \{S_n(\boldsymbol{\beta}, \boldsymbol{c}) + \frac{\lambda_n}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} W_{ij}(c_j)\}, \tag{3.2}$$

where $\sum_{i=1}^{n} W_{ij}(c_j)$ is known as the extremely thresholding penalty (ETP) function for the jth predictor.

The term $W_{ij}(c_j)$ in ETP takes a value of 0 ($c_{j,1} = -\infty, c_{j,2} = \infty, c_{j,3} = -\infty$, i.e., extremes), 1 (either two of $c_{j,1} = -\infty$, $c_{j,2} = \infty, c_{j,3} = -\infty$, hold) or 2 for every i and j. Since we introduce two thresholds and one extreme parameter to every predictor, to avoid over-fitting the data, we penalize the number of thresholds through our data-dependent penalty function to detect

the thresholds that change the linear associations in different segments caused by X_{ij} with a suitable choice of the tuning parameter λ_n.

The TWT-LR regression model enables us to fit data with linear or nonlinear relationships. A constraint is considered in the form of a penalty when the number of change-points is unknown. The penalty function is introduced to perform clustering on the variables that have linear, nonlinear, or no relationships with the response variable. If a variable is insignificant in predicting the response variable, the penalty function will be zero with suitable regularization. On the other hand, if a variable has a linear association with the response variable, the penalty function is equal to n since the data are not truncated in our setting. Furthermore, if a variable has a nonlinear relationship with the response variable (i.e., the variable is truncated), each of the indicator functions in the penalty function will be less than n, depending on the types of nonlinear relationships. With a small tuning parameter λ_n in the penalty term, two change-points may be detected for all variables, even for the variables that are insignificant. On the other hand, over-penalization in a general sense only detects the most significant changes in the regression coefficients in the variables or none at all. In our model setup, over-penalization forces $c_{j,1}$ and $c_{j,3}$ to the left extreme of the observations, $c_{j,2}$ to the right extreme of the observations, resulting in highly sparse coefficients. With the suitable tuning parameter λ_n that controls the magnitude of the penalty function, we can achieve appropriate variable truncation or clustering and avoid overfitting the data or obtaining highly sparse coefficients which results in the misclassification of the variables.

The objective function given in (3.2) is convex in $\boldsymbol{\beta}$ but nonconvex in $\boldsymbol{c}$. It is more computationally convenient to estimate $\boldsymbol{c}$ first through a combination of concentration and grid search similar to Lee, Seo, and Shin (2016) and Hansen (2017) which is typically used in the threshold literature. The estimation of $\boldsymbol{c}$ is given by

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmin}}\{S_n(\hat{\boldsymbol{\beta}}(\boldsymbol{c}), \boldsymbol{c}) + \frac{\lambda_n}{n}\sum_{j=1}^{p}\sum_{i=1}^{n} W_{ij}(c_j)\}, \tag{3.3}$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{c})$ are the least-squares coefficients for fixed $\boldsymbol{c}$. The computational procedure will be discussed in details in Section 4.1. Different from the conventional threshold literature, we introduce a penalty as seen in (3.3) to help in the estimation procedure. Minimizing equation (3.3) requires $\hat{\boldsymbol{\beta}}(\boldsymbol{c})$ for any fixed $\boldsymbol{c}$ which is given as

$$\hat{\boldsymbol{\beta}}(\boldsymbol{c}) = [\mathbb{Z}'(\boldsymbol{c})\mathbb{Z}(\boldsymbol{c})]^{-1}\mathbb{Z}'(\boldsymbol{c})\mathbf{Y}, \tag{3.4}$$

where $\mathbb{Z}(\boldsymbol{c})$ is a $(n \times m)$ matrix whose ith row is $\mathbf{Z}_i'(\boldsymbol{c})$. However, (3.4) will not be well-defined when $I(X_{ij} < c_{j,1}) = 0$, $I(X_{ij} > c_{j,2}) = 0$ or $I(X_{ij} < c_{j,3}) = 0$ for fixed $c_{j,1}$, $c_{j,2}$, and $c_{j,3}$ for some j and all i. As a result, minimizing (3.3) will run into some theoretical and computational issues. In the literature, there have been extensive work to overcome invertibility problems for least-squares coefficients. Similar to the ridge regression, we propose

$$\hat{\boldsymbol{\beta}}(\boldsymbol{c}) = [\mathbb{Z}'(\boldsymbol{c})\mathbb{Z}(\boldsymbol{c}) + \delta\mathbb{M}]^{-1}\mathbb{Z}'(\boldsymbol{c})\mathbf{Y}, \tag{3.5}$$

where $\mathbb{M}$ is an m-by-m identity matrix and δ is a tuning parameter. Alternatively, we propose using pseudoinverse by letting $\delta \to 0$ for faster computational time in our estimation procedure which we will further discuss the advantages in Section 4.1. Let $\mathbb{Z}^N(\boldsymbol{c})$ be a $n \times (r+1)$ matrix with nonzero columns. Without loss of generality, assume the first r numbers of columns in matrix $\mathbb{Z}^N(\boldsymbol{c})$ are nonzero where $r_1 = \#\{j : c_{j,1} \neq -\infty\}$, $r_2 = \#\{j : c_{j,2} \neq \infty\}$, $s = \#\{j : c_{j,3} \neq -\infty\}$ and $r = r_1 + r_2 + s$. We obtain

$$\lim_{\delta\to 0}[\mathbb{Z}'(\boldsymbol{c})\mathbb{Z}(\boldsymbol{c}) + \delta\mathbb{M}]^{-1} = \lim_{\delta\to 0}\begin{bmatrix}\mathbb{Z}^{N'}(\boldsymbol{c})\mathbb{Z}^N(\boldsymbol{c}) + \delta\mathbb{I}_{(r+1)\times(r+1)} & \mathbf{0} \\ \mathbf{0} & \delta\mathbb{I}_{(3p-r)\times(3p-r)}\end{bmatrix}^{-1}.$$

Denote $\boldsymbol{Z}_i^{*'}(\boldsymbol{c})$ as the ith row of matrix $\mathbb{Z}^*(\boldsymbol{c})$ with $\boldsymbol{Z}_i^*(\boldsymbol{c}) = [\boldsymbol{Z}_i^{N'}(\boldsymbol{c}), \underset{\sim}{\mathbf{0}}']'$ where $\boldsymbol{Z}_i^{N'}(\boldsymbol{c})$ is the ith row of matrix $\mathbb{Z}^N(\boldsymbol{c})$ and $\underset{\sim}{\mathbf{0}}$ is a $(3p - r) \times 1$ zero vector. For a better representation, we can always rearrange significant variables in the following way. Since we assume there are r numbers of nonzero columns in matrix $\mathbb{Z}^N(\boldsymbol{c})$, the following form is the same as the equation above.

$$\begin{aligned}&\lim_{\delta\to 0}\begin{bmatrix}\mathbb{Z}^{N'}(\boldsymbol{c})\mathbb{Z}^N(\boldsymbol{c}) + \delta I_{(r+1)\times(r+1)} & \mathbf{0} \\ \mathbf{0} & \delta\mathbb{I}_{(3p-r)\times(3p-r)}\end{bmatrix}^{-1} \\ &= \begin{bmatrix}\lim_{\delta\to 0}[\mathbb{Z}^{N'}(\boldsymbol{c})\mathbb{Z}^N(\boldsymbol{c}) + \delta I_{(r+1)\times(r+1)}]^{-1} & \mathbf{0} \\ \mathbf{0} & \lim_{\delta\to 0}[\delta\mathbb{I}_{(3p-r)\times(3p-r)}]^{-1}\end{bmatrix} \\ &= \begin{bmatrix}[\mathbb{Z}^{N'}(\boldsymbol{c})\mathbb{Z}^N(\boldsymbol{c})]^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{O}^+\end{bmatrix} \\ &= [\mathbb{Z}^{*'}(\boldsymbol{c})\mathbb{Z}^*(\boldsymbol{c})]^+\end{aligned}$$

where $[\mathbb{Z}^{*'}(\boldsymbol{c})\mathbb{Z}^*(\boldsymbol{c})]^+$ and $\boldsymbol{O}^+$ are Moore-Penrose pseudoinverse of $\mathbb{Z}^{*'}(\boldsymbol{c})\mathbb{Z}^*(\boldsymbol{c})$ and zero matrix, respectively.

Since we rearranged the rows in $[\mathbb{Z}^{*'}(\boldsymbol{c})\mathbb{Z}^*(\boldsymbol{c})]^+$, we also need to rearrange the coefficients. Let $\hat{\boldsymbol{\beta}}^*$ and $\boldsymbol{\beta}_0^*$ be the coefficient estimators and true parameters. The model equation (3.4) is now rewritten as

$$\hat{\boldsymbol{\beta}}^*(\boldsymbol{c}) = [\mathbb{Z}^{*'}(\boldsymbol{c})\mathbb{Z}^*(\boldsymbol{c})]^+\mathbb{Z}^{*'}(\boldsymbol{c})\mathbf{Y} \tag{3.6}$$

and the model equation (3.3) has the following form

$$\hat{\boldsymbol{c}} = \underset{\boldsymbol{c}}{\operatorname{argmin}}\{S_n(\tilde{\boldsymbol{\beta}}(\boldsymbol{c}), \boldsymbol{c}) + \frac{\lambda_n}{n}\sum_{j=1}^{p}\sum_{i=1}^{n} W_{ij}(c_j)\}, \tag{3.7}$$

where $\tilde{\boldsymbol{\beta}}(c)$ follows (3.5) with $\mathbb{Z}^{*'}(\boldsymbol{c})\mathbb{Z}^*(\boldsymbol{c})$ matrix or its limit case when $\delta \to 0$ in (3.6). The use of (3.5) provides better precision in estimation but with higher computational time. On the other hand, Equation (3.6) achieves faster computational time with slightly lower precision. The two equations can be viewed as a tradeoff between computational time and precision.

Given the different features of the threshold parameters and the coefficients, we implement a two-step estimation procedure. The first step is to estimate the threshold parameters $\boldsymbol{c}$ using

(3.7) and the second step is to estimate the regression coefficients $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{c}})$ using (3.8) where $\hat{\boldsymbol{c}}$ is obtained from the first step.

$$\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{c}}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\frac{1}{n}(\mathbf{Y}-\mathbb{Z}(\hat{\boldsymbol{c}})\boldsymbol{\beta})'(\mathbf{Y}-\mathbb{Z}(\hat{\boldsymbol{c}})\boldsymbol{\beta})\right\} = [\mathbb{Z}'(\hat{\boldsymbol{c}})\mathbb{Z}(\hat{\boldsymbol{c}})]^{-1}\mathbb{Z}'(\hat{\boldsymbol{c}})\mathbf{Y}. \tag{3.8}$$

As emphasized earlier, the computation of the inverse in Equation (3.8) has an invertibility issue when $I(X_{ij} < \hat{c}_{j,1}) = 0$ a.s., $I(X_{ij} > \hat{c}_{j,2}) = 0$ a.s. or $I(X_{ij} < \hat{c}_{j,3}) = 0$ a.s. (i.e., $\hat{c}_{j,1} = -\infty, \hat{c}_{j,2} = \infty, \hat{c}_{j,3} = -\infty$) for $i = 1, 2, \ldots, n$ and some j which are possible in our setup. When $I(X_{ij} < \hat{c}_{j,1}) = 0$ a.s., $I(X_{ij} > \hat{c}_{j,2}) = 0$ a.s. or $I(X_{ij} < \hat{c}_{j,3}) = 0$ a.s. for all i and some j, the corresponding columns of matrix $\mathbb{Z}(\hat{\boldsymbol{c}})$ consist of zero vectors. Therefore, the term $\mathbb{Z}(\hat{\boldsymbol{c}})\boldsymbol{\beta}$ in the objective function of (3.8) can be reduced to $\mathbb{Z}_R(\hat{\boldsymbol{\tau}})\boldsymbol{\beta}_R$ in the objective function of (3.9). We propose the coefficient estimators corresponding to predictors in (2.2) as follows. If $I(X_{ij} < \hat{c}_{j,1}) = 0$ a.s., $I(X_{ij} > \hat{c}_{j,2}) = 0$ a.s. or $I(X_{ij} < \hat{c}_{j,3}) = 0$ a.s. for all i and some j, the corresponding coefficient estimators are set to $\hat{\beta}_{j,1}(\hat{c}_{j,1}) = 0$ a.s., $\hat{\beta}_{j,2}(\hat{c}_{j,2}) = 0$ a.s. or $\hat{\beta}_{j,3}(\hat{c}_{j,3}) = 0$ a.s.. Optimizing (3.8) is equivalent to

$$\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}}) = \underset{\boldsymbol{\beta}_R}{\operatorname{argmin}}\left\{\frac{1}{n}(\mathbf{Y}-\mathbb{Z}_R(\hat{\boldsymbol{\tau}})\boldsymbol{\beta}_R)'(\mathbf{Y}-\mathbb{Z}_R(\hat{\boldsymbol{\tau}})\boldsymbol{\beta}_R)\right\} = [\mathbb{Z}'_R(\hat{\boldsymbol{\tau}})\mathbb{Z}_R(\hat{\boldsymbol{\tau}})]^{-1}\mathbb{Z}'_R(\hat{\boldsymbol{\tau}})\mathbf{Y} \tag{3.9}$$

where $\hat{\boldsymbol{\tau}}$'s are the threshold estimators ($-\infty < \hat{\tau}_{j,1}, \hat{\tau}_{j,2} < \infty$) for all j and $\boldsymbol{\beta}_R$ is a vector of regression coefficients of $\mathbb{Z}_R(\hat{\boldsymbol{\tau}})$. The matrix $\mathbb{Z}_R(\hat{\boldsymbol{\tau}})$ can be partitioned into $[\mathbb{X}_1(\hat{\boldsymbol{\tau}}), \mathbb{X}_2]$ where the matrix $\mathbb{X}_1(\hat{\boldsymbol{\tau}})$ contains variables with thresholds and the matrix $\mathbb{X}_2$ contains variables without thresholds (i.e., $\hat{c}_{j,1} = -\infty$ a.s., $\hat{c}_{j,2} = \infty$ a.s., $\hat{c}_{j,3} = \infty$ a.s. for some j). The dimension of $\hat{\boldsymbol{\tau}}$ is less than or equal to the dimension of $\hat{\boldsymbol{c}}$. For this article, let r_1 and r_2 be the numbers of predictors truncated below and above and s be the numbers of predictors without thresholds.

So far, it has been clear that the proposed model and its estimation procedure are significantly different from the models in the literature. There are some main advantages of the TWT-LR-ETP model with some interesting aspects to consider. We need not predetermine the number of variables with thresholds and the number of thresholds of any variable. Moreover, since our penalty function does not involve the regression coefficients, standardizing the predictors, which is a standard step in the regularized linear regression literature, is not necessary for our setup. The intercept term is included in our model. In addition, the regularized regression literature has often applied penalization to regression coefficients. As a result, there is a tradeoff between predictors with smaller regression coefficients and predictors with larger regression coefficients in real data applications with small or moderate sample sizes. We note that the TWT-LR model does not involve such a tradeoff scenario. Furthermore, it is easier to solve for the coefficient estimators with the closed-form solution. Besides, if $\hat{\beta}_{j,1}(\hat{c}_{j,1}) = 0$ a.s., $\hat{\beta}_{j,2}(\hat{c}_{j,2}) = 0$ a.s. or $\hat{\beta}_{j,3}(\hat{c}_{j,3}) = 0$ a.s., for some j, we only need to focus on developing the asymptotic properties of $\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}})$ which will be presented in the next section.

3.2. Asymptotic Theory

We first present the assumptions before stating and discussing the asymptotic theory for our framework. Additional notations are introduced. Let $\lambda_{\min}(A)$ denote the smallest eigenvalue of matrix A. Denote $\boldsymbol{Z}_i^N(\boldsymbol{c})$ as a vector with nonzero entries for variables with thresholds and $\boldsymbol{Z}'_{R_i}(\boldsymbol{\tau})$ as the ith row vector in the matrix $\mathbb{Z}_R(\boldsymbol{\tau})$ defined in the previous section. Let $\boldsymbol{\tau}$ be a $(r_1+r_2)\times 1$ vector containing thresholds (i.e., the points at which the linear associations change) and denote $\boldsymbol{\tau}_0$ as the unknown threshold parameters. We recall that the threshold parameters in $\boldsymbol{\tau}$ and $\boldsymbol{\tau}_0$ are also elements of $\boldsymbol{c}$ and $\boldsymbol{c}_0$, respectively. Let $\boldsymbol{c}_1 = (c_{1,1}, c_{2,1}, \ldots, c_{p,1})'$, $\boldsymbol{c}_2 = (c_{1,2}, c_{2,2}, \ldots, c_{p,2})'$ and $\boldsymbol{c}_3 = (c_{1,3}, c_{2,3}, \ldots, c_{p,3})'$ be three $p \times 1$ vectors.

Assumption 1. (a) $\boldsymbol{c} \in \tilde{\Theta}$ where $\mathbb{C}$ and $\mathbb{C}_{12}$ are compact. (b) $(\boldsymbol{X}'_i, \epsilon_i)$, $i = 1, 2, \ldots, n$ are iid with (i) $E(\epsilon_i|X_{ij}) = 0$ a.s., $\text{var}(\epsilon_i|X_{ij}) = \sigma^2 < \infty$ a.s., $E(\epsilon_i^4|X_{ij}) < \infty$ a.s. (ii) $E|X_{ij}|^r < \infty$ for $r = 1, 2, 3, 4$. (c) For $c_{j,k} \neq c_{j,k,0}$, (i) there exists a $\eta > 0$ such that $||\tilde{\boldsymbol{F}}(\boldsymbol{c}) - \tilde{\boldsymbol{F}}(\boldsymbol{c}_0)||_2 > \eta$ where $\tilde{\boldsymbol{F}}$ is a vector of distribution functions of random variables $\boldsymbol{X}$. (ii) $E(\boldsymbol{Z}'_i(\boldsymbol{c})\boldsymbol{\beta} - \boldsymbol{Z}'_i(\boldsymbol{c}_0)\boldsymbol{\beta}_0)^2 > 0$ for $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. (d) For $\tau_{j,k} \neq \tau_{j,k,0}$, (i) $\lambda_{\min}(E[(\boldsymbol{Z}_{R_i}(\boldsymbol{\tau}) - \boldsymbol{Z}_{R_i}(\boldsymbol{\tau}_0))(\boldsymbol{Z}_{R_i}(\boldsymbol{\tau}) - \boldsymbol{Z}_{R_i}(\boldsymbol{\tau}_0))']) > 0$. (ii) $0 < E|Z_{ij}(\tau_{j,k}) - Z_{ij}(\tau_{j,k,0})|^r < \infty$ for $r = 1, 2, 3, 4$ where $Z_{ij}(\tau_{j,k})$ and $Z_{ij}(\tau_{j,k,0})$ are elements of $\boldsymbol{Z}_{R_i}(\boldsymbol{\tau})$ and $\boldsymbol{Z}_{R_i}(\boldsymbol{\tau}_0)$. (iii) $\lambda_{\min}\big(E[\boldsymbol{Z}_{R_i}(\boldsymbol{\tau}_0)\boldsymbol{Z}'_{R_i}(\boldsymbol{\tau}_0)]\big) > 0$.

In the literature on threshold models, the assumption that the parameter space for thresholds is compact is a common assumption since the models are built to model predictors with thresholds. Similarly, we assume compactness for the threshold parameter space when thresholds exist in our setting. Additionally, our model is proposed to model a group of predictors that has linear or no relationships with the response variable. We also consider cases where the parameters can be $-\infty$ and ∞, as shown in Section 2. Since the extreme parameter $c_{j,3}$ only takes two values (i.e., $-\infty$ and ∞) for all j, we need not assume compactness for the parameter space of $\boldsymbol{c}_3$.

Assumption 1(b) is a common assumption in a regression setup in the literature, with additional assumptions on the higher moments being finite. In the regularized regression literature with a random design setting, $E(X_{ij}) = 0$ and $E(X_{ij}^2) = 1$ are common assumptions and $E(Y_i) = 0$ forces the model to exclude the intercept, leading to a wrong or restrictive model when a nonzero intercept exist. Due to the constraint set by some regularized regressions such as Lasso, penalizing the coefficients will depend on the magnitude of the coefficients, so standardization is necessary. Additionally, standardizing the predictors can also be seen in other variable selection methods. Some variable selection methods, such as SIS mentioned in the introduction, depend on the relative magnitude of the coefficients to rank the covariates by their importance. As such, standardizing the predictors is necessary. In our setting, we need not standardize the predictors as we do not penalize the regression coefficients. As a result, the intercept is included in our setup, as mentioned as one of the advantages in the previous section. Moreover, we do not compare the coefficients based on their magnitudes. Instead, the coefficients in our setup are affected by the locations of the thresholds. For these reasons,

standardizing the predictors is unnecessary in our setup, which may be viewed as a big advantage. Assumptions (c) and (d) are conditions for showing theoretical properties. Assumption (d)(iii) is a full-rank condition needed for nondegenerate asymptotic distribution.

Under no threshold setting, the TWT-LR model can be rewritten as a linear model

$$Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, 2, \ldots, n.$$

In this setting, the associations between the response variable and the predictors do not change at any points. The parameter estimators given below can be derived by OLS

$$\hat{\boldsymbol{\beta}} = [\mathbb{X}'\mathbb{X}]^{-1}\mathbb{X}'\mathbf{Y}. \tag{3.10}$$

The asymptotic properties, such as consistency and asymptotic normality of the above parameter estimators, have been well established in the literature. When $p_n \gg n$, many dimension reduction methods such as the SIS, DC-SIS, SEVIS, etc., can be applied to obtain $d\,(< n)$ significant predictors. Consistency and asymptotic normality results will still hold.

Under the multiple thresholds setting when $\boldsymbol{\tau}_0$ is known in advance, since the TWT-LR-ETP model does not penalize on the regression coefficients, the parameter estimators in the following form can also be derived by OLS

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\tau}_0) = [\mathbb{Z}'(\boldsymbol{\tau}_0)\mathbb{Z}(\boldsymbol{\tau}_0)]^{-1}\mathbb{Z}'(\boldsymbol{\tau}_0)\mathbf{Y}. \tag{3.11}$$

In this setting, it can be viewed as a generalization of a linear model where there are two linear curves when there exists one threshold and three linear curves when there exist two thresholds for a given predictor. Therefore, consistency and asymptotic normality will still hold with some modifications to the previous case. Similarly, when $p_n \gg n$, dimension reduction methods can be applied to obtain $d\,(< n)$ significant predictors, and the asymptotic results will hold. If the data have multiple thresholds with nonlinear structure, linear model-based dimension reduction methods are not recommended for variable screening since they only work well under a linear model setting.

For data with multiple thresholds structure, if thresholds are easily observed or known for those variables, it is suitable to use the known threshold setting. However, in many cases, thresholds are not observed for data like gene expression. Despite performing the dimension reduction procedure, they remain difficult to detect. Consequently, our article focuses on multiple unknown thresholds and consider a fixed dimension of p. If $p_n = o(n)$ or $p_n \gg n$, we apply the SEVIS for dimension reduction. We assume that the dimension does not vary with n after variable screening. Next, we establish the consistency and asymptotic normality for the estimators. The proofs are presented in Appendix A, supplementary materials. For the following theorems and lemmas, the convergence in probability and distribution are represented by $\xrightarrow{p}$ and $\xrightarrow{d}$, respectively.

Theorem 1 (Consistency). For $p < n$ and if $\lambda_n \to 0$ and $n\lambda_n \to \infty$, under Assumption 1, we have $\hat{\boldsymbol{c}} \xrightarrow{p} \boldsymbol{c}_0$ and $\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}}) \xrightarrow{p} \boldsymbol{\beta}_R$.

Lemma 1. For $p < n$, if $\sqrt{n}\lambda_n \to \lambda_0$ where $\lambda_0 \geq 0$, then $n||\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0||_2 = O_p(1)$.

Lemma 2. For $p < n$, $\sqrt{n}||\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}}) - \hat{\boldsymbol{\beta}}_R(\boldsymbol{\tau}_0)||_2 \xrightarrow{p} 0$.

Lemma 1 shows that the convergence rate of $\hat{\boldsymbol{\tau}}$ is n under unknown threshold parameters $\boldsymbol{\tau}_0$. If $\boldsymbol{\tau}_0$ is known, Lemma 1 is not needed. Lemma 2 shows that the parameter estimators under known and unknown thresholds converge at a rate of $\sqrt{n}$.

Theorem 2 (Asymptotic Normality). By Lemmas 1 and 2, for $p < n$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}}) - \boldsymbol{\beta}_R) \xrightarrow{d} N(0, \Sigma),$$

where $\Sigma = \sigma^2 E[\boldsymbol{Z}_{R_1}(\boldsymbol{\tau}_0)\boldsymbol{Z}'_{R_1}(\boldsymbol{\tau}_0)]^{-1}$.

Remark. By Lemma 2, the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}}) - \boldsymbol{\beta}_R)$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_R(\boldsymbol{\tau}_0) - \boldsymbol{\beta}_R)$ will be the same as the bias between $\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}})$ and $\hat{\boldsymbol{\beta}}_R(\boldsymbol{\tau}_0)$ goes to zero with the rate of $\sqrt{n}$. Lemmas 1 and 2 guarantee that the asymptotic distribution of parameter estimators $\sqrt{n}(\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}}) - \boldsymbol{\beta}_R)$ will not be an issue.

4. Numerical Studies

4.1. Computational Procedure

In this section, we will present a computational and estimation procedure for simulations and real data analyses. A flowchart is used in the supplementary file Appendix B, supplementary materials to illustrate the computational procedure. We will discuss the initial groupings of the variables and present the estimation procedure subsequently.

The computational procedure for the parameter estimation of the TWT-LR-ETP model consists of two main parts. Part 1 involves group classifications of the predictors (group 1—variables that are linearly related to the response variable, group 2—variables that have changing associations in different segments, group 3—insignificant variables), which determine the initial values for the variables in different groups. Some conditions are imposed on the regression coefficient and threshold parameters to group the variables. After the groups are determined for the predictors, initial values are set based on the groups. In this section, we will discuss the initial values for groups 1–3 later. In Part 2, the threshold and regression coefficient parameters are estimated using the two-step estimation procedure described in Section 3.1. More details of the two-step estimation procedure are also presented in this section. We summarize our procedure in Algorithm 1 and then discuss each part and step in detail.

In the first part of the estimation procedure, we perform an initial grouping for the variables. The variables are grouped into one of the three groups (linear, nonlinear, and no relationship). Initial estimates of the threshold parameters $c_{j,1}, c_{j,2}$ are set to the middle point of the ordered observations for each variable and $c_{j,3}$ is set to $-\infty$. The preliminary coefficient estimates are then computed based on the TWT-LR model. Subsequently, the variables are classified into one of the three groups by using the preliminary coefficient estimates with different conditions on the regression coefficients. The variables are placed in the first group as having a linear association with the response variable if all three conditions are satisfied $|\hat{\beta}_{j,1}| > t_1$, $|\hat{\beta}_{j,2}| > t_1$, and

$|\hat{\beta}_{j,1} - \hat{\beta}_{j,2}| < t_2$, where t_1 and t_2 are predetermined parameters. Furthermore, the variables are grouped into the second group where the linear associations change for the variables with the response variable if (i) $|\hat{\beta}_{j,1}| > t_1$ or (ii) $|\hat{\beta}_{j,2}| > t_1$ with $|\hat{\beta}_{j,1} - \hat{\beta}_{j,2}| > t_2$ are satisfied. If both conditions where $|\hat{\beta}_{j,1}| < t_1$ and $|\hat{\beta}_{j,2}| < t_1$ are satisfied, the variables are placed in the third group. To ensure the same conditions can be applied to all variables, we standardize the predictors for initial variable groupings. We note that the standardization of the predictors is only applied to perform the initial groupings, and the standardization step is not applied to the main estimation procedure. We provide some discussions on the use of standardization following the discussions of all predetermined parameters.

Algorithm 1 Procedure of the parameter estimation using the TWT-LR-ETP model

Input:

1: Training data $\{(x_i, y_i)\}_{i=1}^n$, the number of variables p, the number of grids for groups 1-3, the number of subgrids for group 2 and the number of iterations I.

Part 1 (Group classifications of the variables):

2: Estimate the regression coefficients using the TWT-LR model (see model (2.3)). Classify the variables based on the preliminary conditions on the magnitudes of the regression coefficients and pairwise differences of the regression coefficients in different segments for each variable. Perform the variable group classifications by imposing additional conditions on the regression coefficient and threshold parameters and then update the groups using the TWT-LR-ETP model (see model (3.2)).

3: Set initial values for the threshold parameters denoted as $\hat{\boldsymbol{c}}^{(0)}$ based on the groups.

Part 2 (Step 1—Estimation of the threshold parameters):

4: **for** $i = 1 : I$ **do**

5: **for** $j = 1 : p$ **do**

6: Obtain the threshold parameter estimates $\hat{\boldsymbol{c}}_j^{(i)}$ for variable j by optimizing the objective function (3.7) using the threshold parameter estimates $\hat{\boldsymbol{c}}_k^{(i-1)}$, $k \in \{2, \ldots, p\}$ for $j = 1$, using $\hat{\boldsymbol{c}}_k^{(i)}$, $k \in \{1, \ldots, j-1\}$ and $\hat{\boldsymbol{c}}_{k'}^{(i-1)}$, $k' \in \{j+1, \ldots, p\}$ for $1 < j < p$, and using $\hat{\boldsymbol{c}}_k^{(i)}$, $k \in \{1, \ldots, p-1\}$ for $j = p$.

7: **end for**

8: **end for**

Step 2—Estimation of the regression coefficient parameters:

9: Obtain the regression coefficient estimates $\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}})$ using model (3.9).

After the preliminary variable groupings, we update the groups with additional conditions. The initial values for the third group are set to the left and right extremes of the observations for the parameters $c_{j,1}$ and $c_{j,2}$, respectively, and $c_{j,3}$ is set to $-\infty$. The threshold parameters are estimated for groups 1 and 2 via a grid-search procedure and the regression coefficients are estimated by OLS. The initial groupings of the variables are updated using some conditions on the threshold parameters. The variables are assigned to group 1 if (1) $\hat{c}_{j,3} = -\infty$, $|\hat{c}_{j,1} - \hat{c}_{j,2}| < t_3$, $|\hat{\beta}_{j,1}| > t_1$, $|\hat{\beta}_{j,2}| > t_1$ and $|\hat{\beta}_{j,1} - \hat{\beta}_{j,2}| < t_2$ or (2) $\hat{c}_{j,3} = \infty$, $|\hat{c}_{j,1} - \hat{c}_{j,2}| < t_3$, $|\hat{\beta}_{j,3}| > t_1$, $|\hat{\beta}_{j,1} + \hat{\beta}_{j,3}| > t_1$, $|\hat{\beta}_{j,2} + \hat{\beta}_{j,3}| > t_1$, $|\hat{\beta}_{j,1}| < t_2$, $|\hat{\beta}_{j,2}| < t_2$ and $|\hat{\beta}_{j,1} - \hat{\beta}_{j,2}| < t_2$ or (3) $\hat{c}_{j,3} = \infty$, $\hat{c}_{j,1} = -\infty$ and $\hat{c}_{j,2} = \infty$ are satisfied where t_3 is a predetermined parameter. Furthermore, the variables are placed into group 2 if one of the three conditions is satisfied. (1) $\hat{c}_{j,3} = -\infty$, $|\hat{c}_{j,1} - \hat{c}_{j,2}| > t_3$ with (i) $|\hat{\beta}_{j,1}| > t_1$ or (ii) $|\hat{\beta}_{j,2}| > t_1$ or (iii) $|\hat{\beta}_{j,1} - \hat{\beta}_{j,2}| > t_2$. (2) $\hat{c}_{j,3} = \infty$, $|\hat{c}_{j,1} - \hat{c}_{j,2}| > t_3$ with (i) $|\hat{\beta}_{j,1} + \hat{\beta}_{j,3}| > t_1$ or (ii) $|\hat{\beta}_{j,2} + \hat{\beta}_{j,3}| > t_1$ or (iii) $|\hat{\beta}_{j,3}| > t_1$ or (iv) $|\hat{\beta}_{j,1}| > t_2$ or (v) $|\hat{\beta}_{j,2}| > t_2$ or (vi) $|\hat{\beta}_{j,1} - \hat{\beta}_{j,2}| > t_2$. (3) $\hat{c}_{j,3} = -\infty$ with (i) $|\hat{\beta}_{j,1}| > t_1$ or (ii) $|\hat{\beta}_{j,2}| > t_1$. The condition (3) imposed for group 2 is to include cases where there is only one threshold. The variables are grouped into the third group if (1) $\hat{c}_{j,3} = -\infty$, $|\hat{\beta}_{j,1}| < t_1$ and $|\hat{\beta}_{j,2}| < t_1$ or (2) $\hat{c}_{j,3} = \infty$, $|\hat{\beta}_{j,1} + \hat{\beta}_{j,3}| < t_1$, $|\hat{\beta}_{j,2} + \hat{\beta}_{j,3}| < t_1$ and $|\hat{\beta}_{j,3}| < t_1$ or (3) $\hat{c}_{j,1} = -\infty$, $\hat{c}_{j,2} = \infty$ and $\hat{c}_{j,3} = -\infty$ are satisfied. In addition, for variables that are not assigned a group, we assign the variables to group 3. The idea of using the initial groupings with the conditions is inspired by Ke, Fan, and Wu (2015) and adapted to our model setup with additional conditions specified above. Incorporating the initial groupings in our case reduces the computational time substantially. We have different grid arrangements for different groups by extracting some useful prior information on the variables using the preliminary estimates. Moreover, if the initial values set based on the groups are close to the true unknown threshold parameters for some variables, fewer iterations are needed for convergence and we need not search thoroughly for the tuning parameter λ_n. In contrast, if we do not incorporate the idea of initial groupings, a more thorough search for the tuning parameter λ_n is needed to obtain the estimates of the regression coefficients with low variance and bias since λ_n has multiple functions in our setup which has been discussed in Section 3.1. Similar to Ke, Fan, and Wu (2015) and Ke, Li, and Zhang (2016), we select the parameters t_1 and t_2 via Bayesian information criteria (BIC). Ke, Li, and Zhang (2016) reports that the estimation procedure depends heavily on the choice of the predetermined parameter in their model setup. However, it is not the case in our setup. If the parameters t_1 and t_2 are too small, the computational time increases since denser grids are assigned to more variables as compared to having larger values of t_1 and t_2. Since the conditions on the magnitudes of the regression coefficients in the three segments are to separate significant variables from insignificant variables and the condition on the pairwise differences of the regression coefficients of the three segments is to distinguish variables with thresholds from variables without thresholds, large values of t_1 and t_2 are not required. In other words, the search of t_1 and t_2 can be restricted to smaller values (e.g., from 0.2 to 0.7 with equally spaced grid 0.1 or 0.2). In addition, unlike the estimation procedures in the literature, we need not perform BIC for the entire estimation procedure to select t_1 and t_2 as our estimation procedure does not depend heavily on the choices of t_1 and t_2. With the updated initial groupings, the initial threshold parameter values are set differently based on the respective groups.

Subsequently, we specify the initial values for different groups of the variables after the initial groupings are updated. The initial values for the threshold parameters $c_{j,1}$ and $c_{j,2}$ are set to the

left and right extremes of the observations, respectively and the initial value for $c_{j,3}$ is set to ∞ for the variables in group 1. For variables in group 2, the initial value for the threshold parameter $c_{j,1}$ is set to the middle point of the data and the initial values for the parameters $c_{j,2}$ and $c_{j,3}$ are set to ∞ and $-\infty$, respectively. In group 3, the initial values for the threshold parameters $c_{j,1}$ and $c_{j,2}$ are set similar to the initial values specified for group 1 and the initial value of $c_{j,3}$ is set to $-\infty$.

The change-points are estimated via a grid-search procedure. The regression coefficients are estimated via OLS. Subsequently, the parameters t_1 and t_2 in the intermediate step are selected via BIC. Through our simulation experiments, the parameters t_1 and t_2 selected provide reasonable groupings. The parameter t_3 is set to 1 in our case. Since the predictors are standardized and thresholds are, in most cases, close to the middle of the data, the predictors may be misclassified as group 1 with large values of t_3 or group 2 with small values of t_3. For a standardized predictor, the interval for one standard deviation away from the mean contains most observations (e.g., more than 50% data and about 68% for a Gaussian distribution). If the variable has two change-points close to the center of the data distribution, it is also unlikely that the change-points are close to each other. As a result, the choice of t_3 is reasonable where the proportions of data points between the two thresholds are about 25% and 34% in the case of a Gaussian distribution, assuming the variables have change-points that are close to the middle of the data. If two thresholds exist at two extreme ends, the choice of t_3 is also reasonable to classify the variables according to the conditions discussed above. We note that the parameters t_1, t_2, and t_3 are only used in the initial step to obtain the initial groupings and the parameters are not used in the final estimation procedure. It is important to note that we need not standardize the predictors for the estimation procedure discussed subsequently. Furthermore, the use of the predetermined parameters does not affect our theoretical results.

In this step, the thresholds for each variable are then estimated using the penalized regression with the same number of grids. This step is viewed as an update to the initial groupings of the variables to boost efficiency in the computational time. The conditions of the variable groupings and initial values for the respective groups are similar to the conditions specified above. After updating the variables' groups, the variables' parameters in group 2 are estimated first, followed by the variables in groups 1 and 3.

Subsequently, we present the estimation procedure for the TWT-LR-ETP model. A grid-search approach is commonly used in the threshold literature to estimate the threshold parameters. However, the approach can be computationally heavy if both p and n are large. Therefore, we propose some modifications to the conventional grid-search method that is efficient to obtain the tuning parameter and estimate the threshold parameters in the two-step estimation procedure. The usual grid-search procedure divides the parameter space into equally spaced grids and finds the parameter(s) value that optimizes the objective function. In our setting, the grid-search procedure is based on data-driven information. Since the data points are treated as grids, the grids in our setting are not equally spaced. Once the value that optimizes the objective function is chosen, the grid-search procedure in our setting will be performed in the neighborhood of the value to search for the optimal value. Due to increasing computational costs as the number of grids increases, certain data points are used as grid points where a grid contains 10%–20% of the data points. Besides, the initial groupings are used to determine the number of grids for each variable. For example, denser grids are specified for variables in group 2 and sparser grids are adopted for variables in groups 1 and 3. For instance, the number of grids for group 2 is set to 10 to 15 while the numbers of grids for groups 1 and 3 are set to 5 to 10. Furthermore, for group 2, denser grids are specified in the middle part of the data. For instance, if 10 grids are used, we introduce denser subgrids from the third grid to the eighth grid.

A more detailed discussion on the grid-search procedure is presented at a given level of the tuning parameter λ_n. First, all predictors are imposed with three thresholds. Then, we apply a two-step estimation procedure for the parameter estimation. We first estimate the threshold parameters $\boldsymbol{c}$ using (3.7) via a grid-search approach. Since the threshold parameter $c_{j,3}$ only takes two values, we only apply the grid-search approach to estimate $c_{j,1}$ and $c_{j,2}$. Let $x_{(i),j}$ be the ith ordered sample for the truncated above and below of predictor j and the data are divided into g numbers of grids. Subsequently, the middle point of each grid that minimizes the objective function (3.7) is used to determine the grid or region of the global optimum. The estimate of the threshold $\hat{c}_{j,1}$ is the observed sample point that minimizes the objective function in the selected grid. The procedure is repeated for the jth covariate truncated below and for the other covariates. After nth iteration, the insignificant terms will be dropped from the model correspondingly, that is, $\hat{c}_{j,1} = -\infty$, $\hat{c}_{j,2} = \infty$ or $\hat{c}_{j,3} = -\infty$. The coefficient estimates $\hat{\boldsymbol{\beta}}_R(\hat{\boldsymbol{\tau}})$ are computed via (3.9).

The computational costs of estimating the threshold parameters in Step 1 of the two-step estimation procedure are at most $2p(G+1)I\text{Ridge}(n,p)$ where G is the maximum number of grids, I is the number of iterations and $\text{Ridge}(n,p)$ is the computational cost to obtain the closed form solution of the regression coefficient parameters using ridge regression for each grid to search for the threshold parameters with sample size n and number of variables p. The computational costs of estimating the regression coefficient parameters in Step 2 are q^3+nq^2 where $q = r_1 + r_2 + s$, r_1, r_2, and s are defined in Section 3.1. We denote the computational costs of estimating the threshold and regression coefficient parameters for the two-step estimation procedure using the TWT-LR-ETP model as $\text{TWT}(n,p,G)$. We apply the K-fold cross-validation to obtain the tuning parameter λ_n. The computational costs are $Kl_1\text{TWT}(n^*,p,G)$ where K is the number of folds, l_1 is the number of grid points for λ_n and n^* is the sample size for each fold in the K-fold cross-validation for λ_n. The computational cost for the tuning parameter δ is $l_2\text{TWT}(n,p,G)$ where l_2 is the number of grid points for δ. Furthermore, the computational costs for choosing t_1 and t_2 are at most $Ul_3\text{TWT}(n,p,G)$ and $Ul_4\text{TWT}(n,p,G)$, respectively, where U is the number of initial grouping updates, l_3 and l_4 are the numbers of grid points for t_1 and t_2.

In addition, the computational costs for the procedure depend on the dimension p, number of grids g, number of grids used for the tuning parameter λ_n and tuning an additional parameter δ. For larger p or denser grids, we use the Moore-

Penrose pseudoinverse to obtain the coefficient estimates using (3.6) in the intermediate steps by leaving out the terms where $\hat{c}_{j,1} = -\infty$, $\hat{c}_{j,2} = \infty$ or $\hat{c}_{j,3} = -\infty$. Since the thresholds are at the extremes (i.e., $\hat{c}_{j,1} = -\infty$, $\hat{c}_{j,2} = \infty$ or $\hat{c}_{j,3} = -\infty$), the computational costs can be reduced by estimating the significant terms in the intermediate steps to obtain the estimates of the threshold parameters. After the threshold parameters are estimated, the regression estimates are computed using (3.9).

For simulation studies in our setup, the SEVIS mentioned in the introduction is adopted to perform dimension reduction for high-dimensional cases. The SEVIS is applied to all p covariates to select covariates which are highly correlated with the response variable and reduce the dimension to d. The estimation procedure for the final d-dimensional predictors will be similar to the procedure mentioned above.

4.2. Simulated Examples

In this section, some simulation studies are conducted to assess the performance of the TWT-LR-ETP model. For $j = 1, \ldots, p$, we let $\tilde{x}_{i,j} = x_{i,j}\{I(x_{i,j} < c_{j,1}), I(x_{i,j} > c_{j,2}), I(x_{i,j} < c_{j,3})\}$ since these related variables form a group naturally (Li, Zhong, and Zhu, 2012). Also, the variability in the response variable explained by predictors is through $\tilde{x}_{i,j}$.

Suppose first that $(x_{i,1}, \ldots, x_{i,p})'$ is a p-dimensional vector generated from $N(0, I_{p\times p})$ and the error term ϵ_i is generated from $N(0, 1)$. The number of active variables with thresholds is set to 5 while the number of active variables without thresholds is set to 4 and the total number of regressors is 12 with 3 inactive variables. Two thresholds for the first variable are set as $c_{1,1,0} = c_{1,2,0} = 0$ and $c_{1,3,0}$ is set to $-\infty$. The thresholds $c_{2,1,0}$ and $c_{2,2,0}$ are set to -0.8 and 0.8, respectively with $c_{2,3,0} = \infty$ for the second variable. The thresholds $c_{j,1,0}$ and $c_{j,2,0}$ are set to -0.8 and 0.8 with $c_{j,3,0} = -\infty$ for $j = 3, 4, 5$. The coefficients for covariates with thresholds are set to $\beta_{1,1,0} = 5$, $\beta_{1,2,0} = -5$, $\beta_{1,3,0} = 0$, $\beta_{2,1,0} = 6$, $\beta_{2,2,0} = 6$, $\beta_{2,3,0} = -4$, $\beta_{3,1,0} = 5$, $\beta_{3,2,0} = 0$, $\beta_{3,3,0} = 0$, $\beta_{4,1,0} = 0$, $\beta_{4,2,0} = -5$, $\beta_{4,3,0} = 0$, $\beta_{5,1,0} = 5$, $\beta_{5,2,0} = -5$, $\beta_{5,3,0} = 0$. The coefficients for active variables without thresholds are set to (-5, 5, -5, 5). The number of observations is set to $n = 500, 600, 800, 1200$. The parameter estimation is performed based on the two-step estimation procedure and the root mean squared error (RMSE) are computed based on the sample size $n = 300, 400, 600, 1000$ which are used to estimate the parameters while the remaining 200 observations are used as testing datasets to compute the root mean squared prediction error (RMSPE).

In addition, correlations between covariates are considered. The numbers of active variables with thresholds and without thresholds, the parameter values of the thresholds and regression coefficients are set the same as above. The p-dimensional vector $(x_{i,1}, \ldots, x_{i,p})'$ is jointly generated from a multivariate normal distribution $N(0, \Sigma)$ with (i) $\Sigma_{i,j} = 0.3^{|i-j|}$ (ii) $\Sigma_{i,j} = 0.5^{|i-j|}$ where $\Sigma_{i,j}$ denotes the correlation between ith and jth covariates with the diagonal entries being equal to 1. Other settings are the same as above. The simulation study with the parameter settings specified above and different correlation structures is referred to as numerical experiment 1. The results are presented in Figures 1 and 2 where different colored lines are used to represent different candidate models. We also produce RMSE and RMSPE results for $p = 110$ and $p = 1100$ with the parameter settings specified above.

For the tuning parameters, we set 10 grids for group 2 and 6 grids for groups 1 and 3. For group 2, we consider three additional grids from the first to the eighth grid to estimate the threshold parameter $c_{j,1}$ and from the third to the tenth grid for $c_{j,2}$. Denser grids are considered at both ends to detect the thresholds near the extreme ends. The tuning parameter λ_n is searched from 70 equally spaced values from 0.1 to 1.5 by a 5-fold cross-validation procedure. For $p < 50$, the tuning parameter δ is selected using Akaike information criteria (AIC). Another algorithm's computational challenge is selecting the optimal tuning parameter λ_n. The computational time can be further improved using the parallel computing toolbox available in many software. We use MATLAB® software to produce the numerical results. Multiple parallel computing methods are readily available. We adopt a parallel computing tool that is accessible and feasible to all users. Each task is run in parallel using the user's computer cores. The simulations in this article are conducted using a computer with ten-core processors. Besides the RMSE and RMSPE, we also report the computational time of each simulation. The TWT-LR-ETP model is compared to Lasso, SCAD, and MCP to show the model performances when thresholds are unknown, or no prior information on the thresholds is available in the data. With more predictors considered, the detection of thresholds through visual plots can be challenging (see Appendix C, supplementary materials). Furthermore, different from the threshold models with a threshold variable in the literature, the TWT-LR-ETP model includes all variables without the need to predetermine the threshold variable. In this article, the performance of the TWT-LR-ETP model is compared to the threshold models in Lee, Seo, and Shin (2016) and Kaul, Jandhyala, and Fotopoulos (2019b). The results for the classical linear regression models with the Lasso, SCAD, and MCP penalties are obtained using the existing packages in R. The estimation for the candidate models, such as the threshold models, are run using R programming software.

As we see from the results, the TWT-LR-ETP model outperforms the Lasso, SCAD, MCP, and OLS shown in the plots. RMSE and RMSPE of our model range from 0 to 3, while most of the cases range from 0 to 1.5. For the Lasso, SCAD, MCP, and OLS, RMSE and RMSPE are mostly above 5, while most of them remain above 6. Based on our observations, the Lasso, SCAD, and MCP cannot select most of the variables with thresholds. Hence, the RMSE and RMSPE are larger due to the large coefficients specified in the setup. However, additional results with smaller coefficients set (in Appendix B, supplementary materials) show smaller differences in RMSE and RMSPE between the Lasso, SCAD, or MCP and our model. We further note that when we apply our model under known thresholds setting with the Lasso, SCAD, or MCP penalty functions, these models are able to select the active variables. These observations suggest that for a given dataset where some predictors contain thresholds and are significantly similar to the settings we specified, existing methods will not be able to select most of the significant covariates or provide poor estimates for the

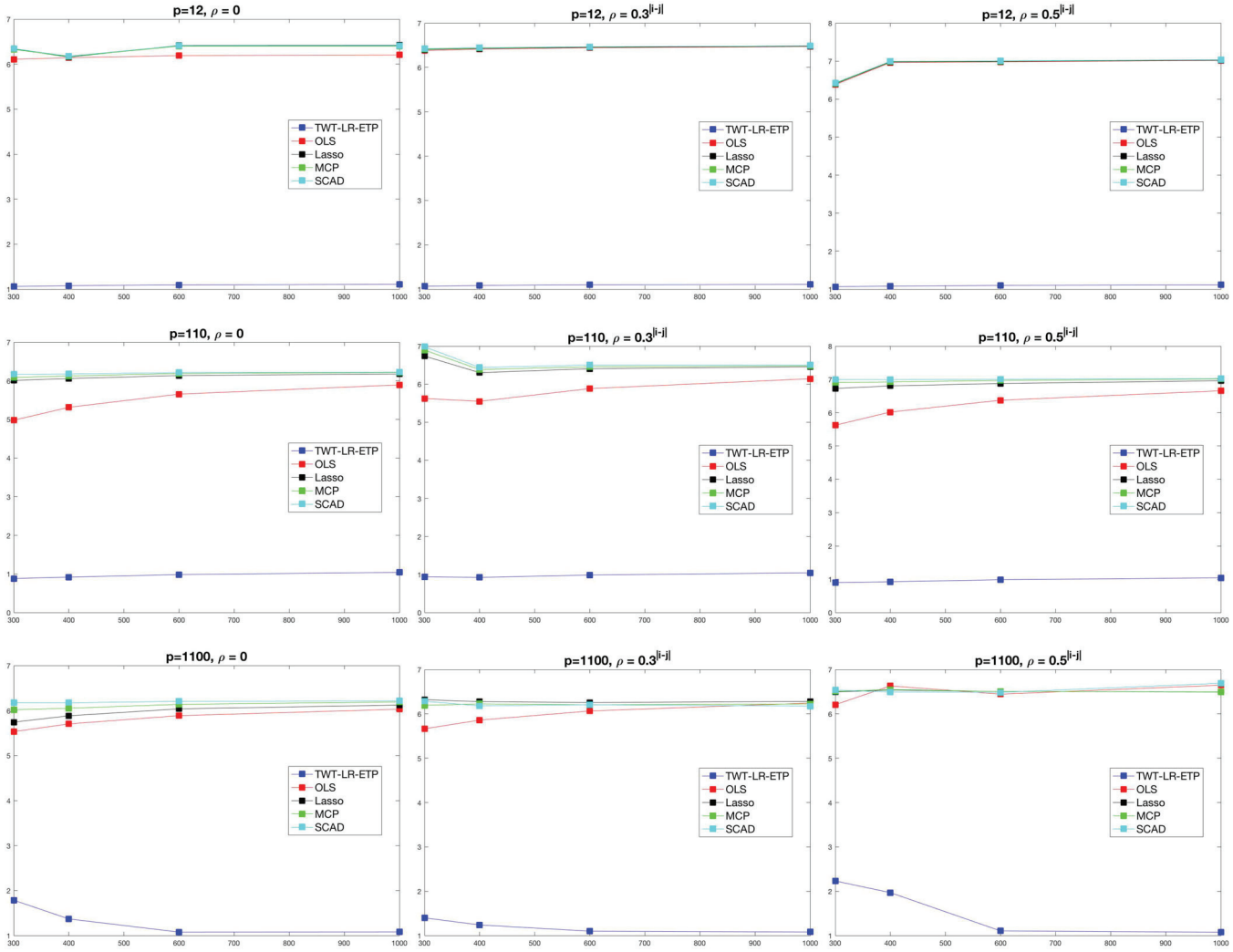

Figure 1. The RMSE is produced for the TWT-LR-ETP, Lasso, SCAD, MCP, and OLS under the simulation setting in numerical experiment 1.

regression coefficients. As a result, the TWT-LR-ETP model should be considered due to its applicability, interpretability, and predictability.

Based on Figures 1 and 2, RMSE and RMSPE are relatively stable after 600 for the training set. We consider the sample size to be 600 for the training set and 200 for the testing set for additional simulations with a different setting on the regression coefficients. We extend the number of active variables without thresholds to 10 and the coefficients are set to

$$\beta_{j,0} = \begin{cases} -5 & \text{for } j = 2k-1, \\ 5 & \text{for } j = 2k \end{cases}$$

for k=1,2,…,5 where $\beta_{j,0}$ is the coefficient of the jth variable without thresholds. Other parameter settings are specified similarly to the setting in numerical experiment 1. The process is repeated for $p = 110$ and $p = 1100$. The simulation setting is referred to as Case A.

In addition, we consider smaller regression coefficients in Case B. The coefficients of 5 active variables with thresholds are set to $\beta_{1,1,0} = 1$, $\beta_{1,2,0} = -1$, $\beta_{1,3,0} = 0$, $\beta_{2,1,0} = -3$, $\beta_{2,2,0} = -3$, $\beta_{2,3,0} = 1$, $\beta_{3,1,0} = 1$, $\beta_{3,2,0} = 0$, $\beta_{3,3,0} = 0$, $\beta_{4,1,0} = 0$, $\beta_{4,2,0} = -1$, $\beta_{4,3,0} = 0$, $\beta_{5,1,0} = 1$, $\beta_{5,2,0} = -1$, $\beta_{5,3,0} = 0$. The coefficients of four active variables without thresholds are set to $(-1, 1, -1, 1)$. The process is repeated for $p = 110$ and $p = 1100$ with other parameter settings specified in numerical experiment 1.

In Case C, we set the number of active variables without thresholds to 10 with the regression coefficients

$$\beta_{j,0} = \begin{cases} -1 & \text{for } j = 2k-1, \\ 1 & \text{for } j = 2k \end{cases}$$

for k=1,2,…,5 while other settings are specified similarly to numerical experiment 1. We tabulate these additional results in Tables B1–B3 in the Appendix, supplementary materials. Based on the results, the TWT-LR-ETP model outperforms the Lasso, SCAD and MCP from Case A through Case C.

We further consider data generated from a linear model (i.e., no thresholds) with 10 active variables using different sets of parameter coefficients $(-5, 5, \ldots, -5, 5)$ and $(-1, 1, \ldots, -1, 1)$ in Case D and Case E, respectively. The dimensions of the predictors and correlation structures are set similarly to numerical experiment 1. The results in Tables B4 and B5 (in Appendix B, supplementary materials) show that the TWT-LR-ETP model produces results that are comparable to the Lasso, SCAD and MCP.

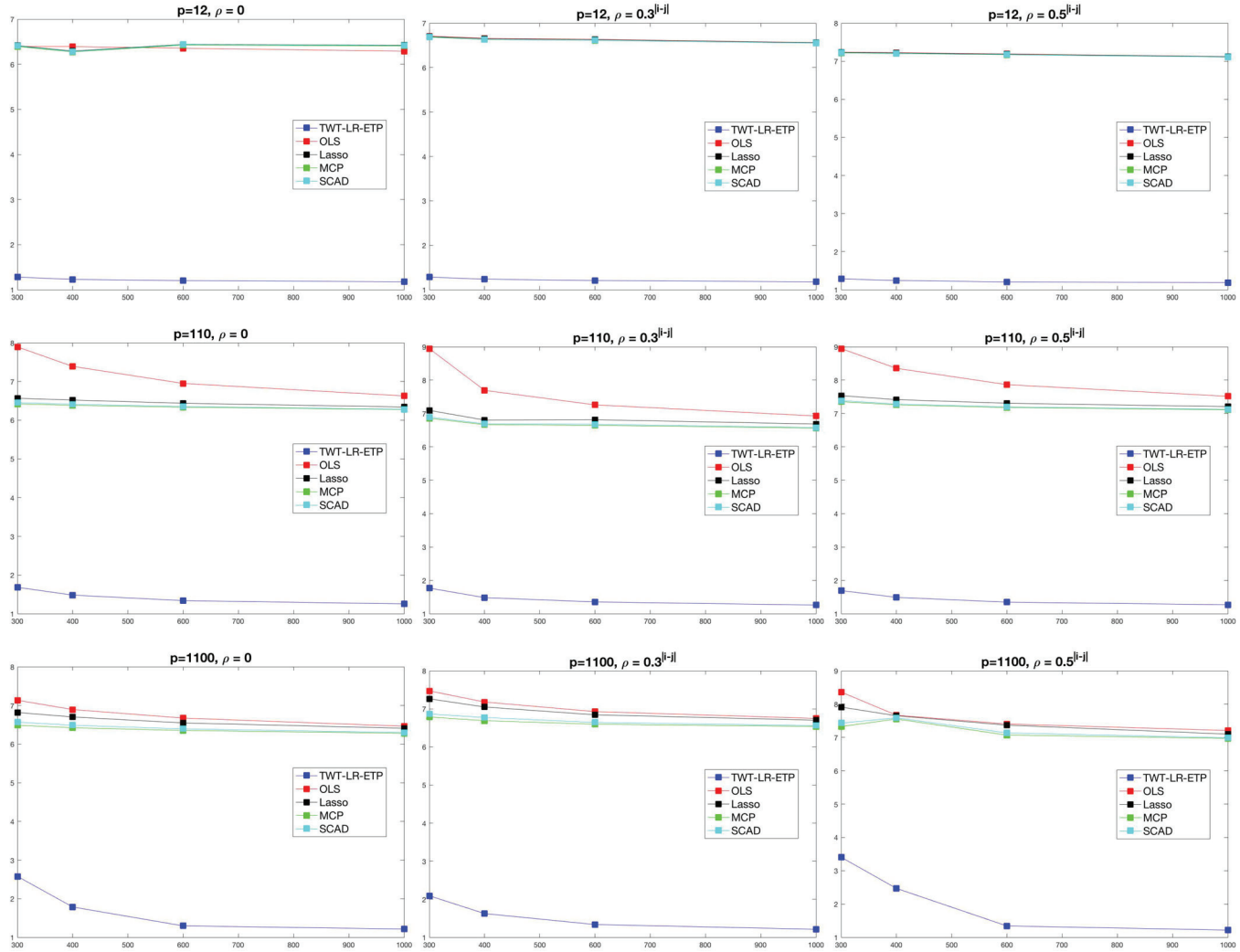

Figure 2. The RMSPE is produced for the TWT-LR-ETP, Lasso, SCAD, MCP, and OLS under the simulation setting in numerical experiment 1.

In Case F, the threshold parameters are set to $c_{j,1,0} = -1$, $c_{j,2,0} = 1$ and $c_{j,3,0} = -\infty$ for $j = 1, 2, 3, 4, 5$. The coefficients for covariates with thresholds are set to $\beta_{1,1,0} = \beta_{1,2,0} = 5$, $\beta_{1,3,0} = 0$, $\beta_{2,1,0} = \beta_{2,2,0} = -5$, $\beta_{2,3,0} = 0$, $\beta_{3,1,0} = \beta_{3,2,0} = 5$, $\beta_{3,3,0} = 0$, $\beta_{4,1,0} = 0$, $\beta_{4,2,0} = -5$, $\beta_{4,3,0} = 0$, $\beta_{5,1,0} = 5$, $\beta_{5,2,0} = 0$, $\beta_{5,3,0} = 0$. The parameter coefficients for 10 active variables without thresholds are set to $(-5, 5, \ldots, -5, 5)$. Other settings are considered similarly as in numerical experiment 1. In Case G, we consider similar settings in Case F with regression coefficients $\beta_{1,1,0} = \beta_{1,2,0} = 1$, $\beta_{1,3,0} = 0$, $\beta_{2,1,0} = \beta_{2,2,0} = -1$, $\beta_{2,3,0} = 0$, $\beta_{3,1,0} = \beta_{3,2,0} = 1$, $\beta_{3,3,0} = 0$, $\beta_{4,1,0} = 0$, $\beta_{4,2,0} = -1$, $\beta_{4,3,0} = 0$, $\beta_{5,1,0} = 1$, $\beta_{5,2,0} = 0$, $\beta_{5,3,0} = 0$. Based on the RMSE and RMSPE in Tables B6 and B7 (in Appendix B, supplementary materials), the TWT-LR-ETP model outperforms the Lasso, SCAD and MCP.

Besides, we compare the performance of the TWT-LR-ETP model with the threshold models for a single threshold in Lee, Seo, and Shin (2016) and multiple thresholds in Kaul, Jandhyala, and Fotopoulos (2019b). We consider a variable with a threshold at $c_{1,1,0} = c_{1,2,0} = 1$ ($c_{1,3,0} = -\infty$) with the parameter coefficients $\beta_{1,1,0} = 5$, $\beta_{1,2,0} = -5$ and $\beta_{1,3,0} = 0$. The parameter coefficients for 10 active variables without thresholds are set to $(-5, 5, \ldots, -5, 5)$ in Case H. For Case I, the parameter coefficients for the variable with a threshold are set to $\beta_{1,1,0} = 1$, $\beta_{1,2,0} = -1$ and $\beta_{1,3,0} = 0$ while the parameter coefficients for 10 active variables without thresholds are set to $(-1, 1, \ldots, -1, 1)$. In addition, in Case J and Case K, we consider a variable with two thresholds at $c_{1,1,0} = -1$ and $c_{1,2,0} = 1$ ($c_{1,3,0} = -\infty$) with other parameter settings similar to Case H and Case I, respectively. The results in Tables B8–B11 (in Appendix B, supplementary materials) show that the TWT-LR-ETP model outperforms the candidate models from Case H to Case K.

Furthermore, we include a simulation where the thresholds are at the two extreme ends for comparison purposes in Cases L and M. We consider two variables with two thresholds at $c_{j,1,0} = -1.8$, $c_{j,2,0} = 1.8$, $c_{j,3,0} = \infty$ for $j = 1, 2$ with the parameter coefficients $\beta_{1,1,0} = 3$, $\beta_{1,2,0} = 3$, $\beta_{1,3,0} = -2$, $\beta_{2,1,0} = -3$, $\beta_{2,2,0} = -3$ and $\beta_{2,3,0} = 2$ in Case L. The parameter coefficients for 10 active variables without thresholds are set to $(-1, 1, \ldots, -1, 1)$. Other parameter settings follow from Case H. In Case M, we consider one variable with two thresholds at the two extreme ends, one variable with one threshold at the left extreme end and another variable with one threshold at the right extreme end. The threshold parameters are set to $c_{1,1,0} = -1.8$, $c_{1,2,0} = 1.8$, $c_{1,3,0} = \infty$, $c_{2,1,0} = -1.8$, $c_{2,2,0} = -1.8$, $c_{2,3,0} =$

$-\infty$, $c_{3,1,0} = 1.8$, $c_{3,2,0} = 1.8$, $c_{3,3,0} = -\infty$ with the parameter coefficients $\beta_{1,1,0} = 3$, $\beta_{1,2,0} = 3$, $\beta_{1,3,0} = -2$, $\beta_{2,1,0} = 1$, $\beta_{2,2,0} = -2$, $\beta_{2,3,0} = 0$, $\beta_{3,1,0} = 2$, $\beta_{3,2,0} = -1$, $\beta_{3,3,0} = 0$. Other parameter settings are set similar to Case L. The results in Tables B12–B13 (in Appendix B, supplementary materials) show that the TWT-LR-ETP model outperforms the candidate models for Cases L and M.

In addition, we produce the computational time for each simulation in the Appendix, supplementary materials. However, we do not report the computational time for the Lasso, SCAD, and MCP in Figures 1 and 2 as they are similar to the computational time in other simulation settings. The TWT-LR-ETP model has a longer computational time for all simulation settings. However, a higher computational cost is expected as we search for possible thresholds or change-points for every variable without any prior information on whether the variables contain thresholds or change-points, which is the case in handling new real datasets. The computational cost can be reduced if we acquire prior knowledge of the variables with possible thresholds by updating the initial groups of the variables in the computational procedure or estimate the thresholds for the variables directly, which is a common practice for most thresholds or change-point modeling.

Furthermore, since the estimation results depend on the choice of the tuning parameter λ_n as discussed in Section 3.1, the choice of the tuning parameter λ_n, which is selected via a K-fold cross-validation procedure, may not yield optimal estimation results in the case of smaller regression coefficients. For the variables with change-points, if most data points in the same segments are placed in the same fold to train or validate the data to select λ_n, the choice of λ_n may result in some change-points being undetected in the case of smaller regression coefficients. This is due to the fact that the samples in each segment are not split evenly across the folds. As a result, the change-points can go undetected for certain folds as there are insufficient data points in some segments to detect the change-points especially when the regression coefficients are small, leading to higher cross-validation error and poor estimation. Dividing the data points evenly into different folds can be challenging since the values of the predictors may fall into different segments for multiple predictors. Therefore, our best approach is to split the data points randomly into different folds before performing the K-fold cross-validation procedure. Similarly, it is also suggested to permute and select the data points randomly as the same issue might arise when splitting the data into training and testing datasets.

Moreover, under the classical linear regression models (i.e., without thresholds), we also illustrate simulation examples to demonstrate the Lasso, SCAD, and MCP fittings outperform the TWT-LR-ETP fitting. The simulation settings are taken from examples in Breheny and Huang (2011) since the MCP and SCAD used for comparison are based on the algorithms in the article. Breheny and Huang (2011) showed through simulation examples that the SCAD and MCP outperform the Lasso for large regression coefficients but not necessarily for small coefficients. The covariate values are generated independently from the standard normal distribution. Two of the nonzero coefficients are set to $+z$, and two other nonzero coefficients are set to $-z$. Here, we set $z = 0.3, 0.5, 0.8$ for comparison purposes. The total numbers of predictors are set to 30 and 90 as in the simulation of the article. In setting 1 of the article, comparisons based on the prediction error were not reported. We set the sample size to 300, where we use a sample size of 200 for estimation and 100 for prediction. The RMSE and RMSPE are computed for each of the settings and the ratios of the medians of the RMSEs are computed, for example, $\text{Ratio} = \frac{\text{RMSE for the TWT-LR-ETP model}}{\text{RMSE for the fitted model by competing approach}}$. The results are shown in the Figures B2 and B3 (in Appendix B, supplementary materials). The plots show that the ratios for the MCP are the highest, followed by the SCAD and the Lasso based on the ratios of the RMSE and RMSPE for $p = 30, 90$. For smaller regression coefficients without threshold, the Lasso, SCAD, and MCP outperform the TWT-LR-ETP model as the ratios are above 1. In addition, it is observed that the ratios of RMSE and RMSPE decrease for larger regression coefficients, suggesting the estimation and prediction outperform the TWT-LR-ETP model as compared to the Lasso, SCAD, and MCP.

4.3. Real Data

In this section, we will present an analysis of a real dataset, show the results where thresholds are detected at extreme ends, and the interpretability aspects of our model as mentioned in the introduction.

4.3.1. Cancer Mortality

Numerous research studies have been conducted on studying factors and genes related to cancers in past decades as cancer is a leading cause of death worldwide, according to the World Health Organization. Some recent work on identifying key genes related to different types of cancers, but are not limited to, Zhou et al. (2020), Gao et al. (2020), Song et al. (2021), Zhang (2021), and Zhang (2022).

This section presents the results and analyses of a cancer mortality dataset. It is of interest to investigate factors related to cancer mortality at the county level. The cancer mortality measures death rate from cancer per capita (100,000) at county level. The dataset can be obtained from the United States Census Bureau and National Cancer Institute websites. After performing data cleaning, there are 742 data points with 13 variables in the dataset. The training dataset consists of 602 data points, and the remaining data points are used as the testing dataset. First, the regularization parameter λ_n is chosen using a 5-fold cross-validation procedure. Next, an initial tuning parameter λ_n is obtained after searching through 70 values. Then, a more thorough search in the neighborhood of the initial tuning parameter is performed to obtain the optimal tuning parameter.

The relationships between two predictors and cancer mortality are depicted in Figure 3. Based on Figure 3(a), there is an overall increasing trend, and there are changes in the linear association between cancer mortality and incidence rate at 350 and 520. Figure 3(b) shows a general decreasing trend with potential change-points at around \$38,000 and \$60,000.

Figure 4 shows the RMSE and RMSPE from different candidate models. From the plots, we can see that TWT-LR-ETP performs the second (third) best in terms of RMSE and RMSPE, respectively, among seven approaches, including

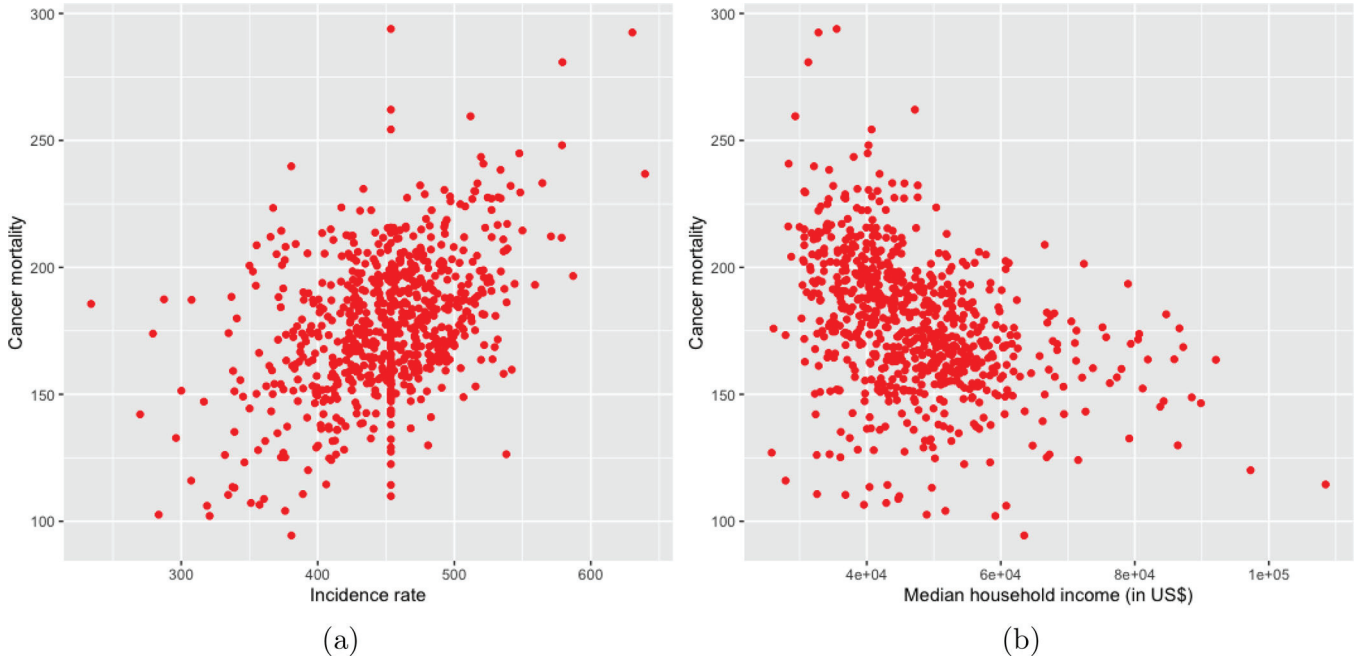

Figure 3. Panel (a) shows the scatterplot between cancer mortality and incidence rate. Panel (b) shows the scatterplot between cancer mortality and median household income (in US$).

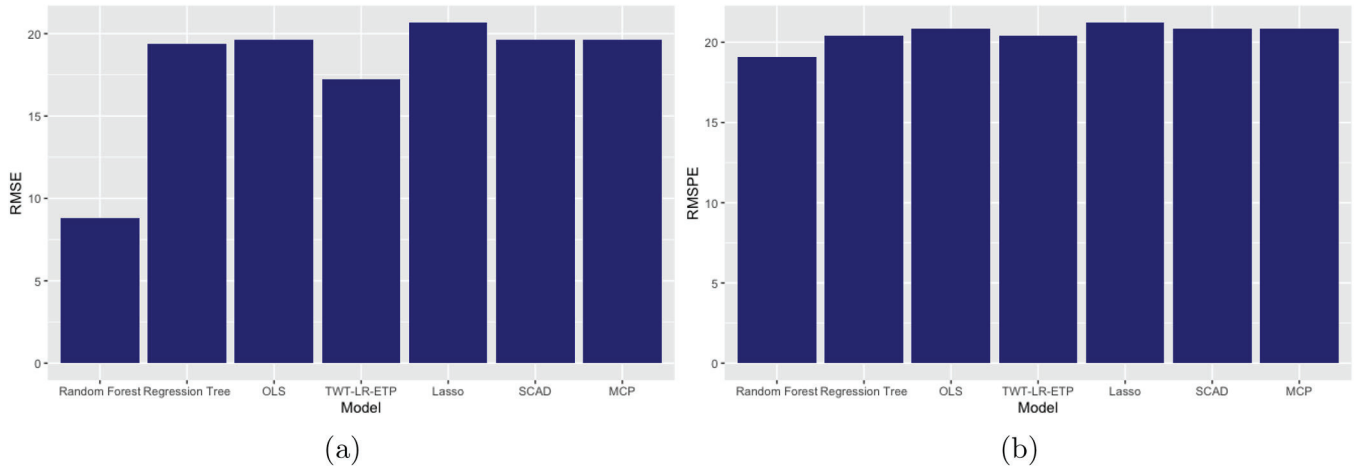

Figure 4. Panels (a) and (b) show the RMSE and RMSPE based on different candidate models for the cancer mortality dataset.

six other widely applied approaches. In the second dataset of COVID-19 in the supplementary materials, TWT-LR-ETP performs the best and the result is 21.6% better than the second-best performance by Random Forrest in terms of RMSE (Table C16). Table 1 displays partial estimation results for the cancer mortality data. A complete table of the results and a table containing the descriptions of the variables can be found in Appendix B, supplementary materials. In the table, we report the coefficient estimates, standard errors, threshold estimates, and percentages of the thresholds in the distribution of the empirical distribution for each predictor. Based on the results, the tuning parameter $\hat{\lambda}_n$ is 2.18. The computational time for this real dataset is 71.317 sec.

Table 1. Partial estimation results for cancer mortality dataset.

Variables	TWT-LR-ETP	
	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{c}}$
Incidence rate	−0.044 (0.015)	352.5 (3.91%)
	0.016 (0.006)	511.1 (90.8%)
	0.151 (0.022)	∞ (100%)
Income	0.000197 (0.000083)	38221 (20.62%)
	0.000182 (0.000078)	69430 (95.55%)
	−0.0008 (0.0002)	∞ (100%)
Percent public health coverage	−0.580 (0.216)	23.4 (5.53%)
	0.365 (0.081)	34.8 (44.47%)
	−0.924 (0.276)	∞ (100%)

Subsequently, we will discuss and interpret the estimation results of the TWT-LR-ETP model. For counties with a cancer incidence rate lower than 352.5, the estimated regression coefficient is 0.107 (i.e., refer to the term $\hat{\beta}_{j,1} + \hat{\beta}_{j,3}$ in model (2.9)). The estimated regression coefficient is 0.151 for counties with a cancer incidence rate between 352.5 and 511.1. When the cancer incidence rate is higher than 511.1, the estimated regression coefficient is 0.167 (i.e., refer to the term $\hat{\beta}_{j,2} + \hat{\beta}_{j,3}$

in model (2.9)). The estimated change-points at 352.5 and 511.1 are detected at the two extreme ends where the first change-point is located at 3.91% from the left of the data distribution and the second change-point is detected at 9.2% from the right of the data distribution (with a cumulative percentage of 90.8% from the left), respectively. In the extreme value theory context, the thresholds detected within 10% from the two extreme ends can be considered as thresholds at the two extreme endpoints. Furthermore, cancer mortality at county level decreases by 0.000603 for every unit increase in the median household income below $38,221. The estimated regression coefficient is −0.0008 for the median household earning between $38,221 and $69,430. The cancer mortality decreases by 0.000618 for every unit increase in the median household income above $69,430. The estimated threshold at $69,430 is detected at the right extreme end which is located at 4.45% from the right of the data distribution (with a cumulative percentage of 95.55% from the left). In addition, cancer mortality decreases by 1.504 when the percentage of public health coverage is below 23.4. When the percentage of public health coverage is between 23.4 and 34.8, cancer mortality decreases by 0.924. As the percentage of public health coverage increases beyond 34.8, cancer mortality decreases by 0.559. The estimated threshold at 23.4 is detected at the left extreme end which is 5.53% from the left of the data distribution. The variables discussed show the cases of thresholds detected at the two extreme ends, right extreme ends and left extreme ends. In addition, there are thresholds detected in the middle of the data distribution shown in Table 1 and Table B15 (in Appendix B, supplementary materials). These findings illustrate the flexibility and capability of the TWT-LR-ETP model for detecting thresholds at the extreme ends. The results for other variables in Table B15 can be interpreted similarly. The results suggest the cancer mortality is related to the overall economic and social welfare of a county. The cancer mortality can be improved by either providing better public healthcare coverage to families that cannot afford healthcare or reducing poverty. Furthermore, the economic welfare such as education, income, etc., that are related to the cancer mortality should also be enhanced.

5. Conclusion

This article proposes a more general class of threshold model with regularization, which allows us to capture significant linear and nonlinear relationships between variables due to the flexibilities of the two-way truncated linear regression with an extremely thresholding penalty (TWT-LR-ETP) model. The TWT-LR-ETP model is capable of detecting thresholds at the two extreme ends where data are sparse through simulation studies and a real cancer mortality dataset. Our model maintains the CIPS properties mentioned in the introduction, and it is highly flexible in modeling data with no threshold, one threshold, and two thresholds while controlling the number of thresholds through the penalty, which does not involve the regression coefficients. Therefore, standardizing the predictors is not necessary. It is substantiated in simulation studies that our model is useful, especially in the presence of variables with thresholds, as the relationships between the response variable and predictors are not always linear. In addition, it is also shown using a socio-economic dataset, medical research example, and real-world business problem in Appendix C, supplementary materials that our model provides highly interpretable results, which are important in studying the underlying experience and making better business decisions and medical treatments.

Moreover, it is also established that the model has desired theoretical properties such as consistency and asymptotic normality under appropriate conditions. Therefore, throughout the article, we focused on the estimation consistency of the TWT-LR-ETP model, both theoretically and computationally, in both simulations and real data analyses.

Furthermore, we note that the way the TWT-LR-ETP model performs variable selection is not the same as the traditional methods. Since our primary goal is not to propose a new variable selection method as there are various existing dimension reduction techniques, the results depend heavily on the performance of the SEVIS dimension reduction method in a high-dimensional setting. After this work, it provides us with some directions for future research work. In this article, we focus on proposing, developing the estimation properties of the TWT-LR-ETP model, and exploring the estimation performance of the model in the simulation section. The theoretical properties of variable selection using the TWT-LR-ETP are yet to be developed, which can be a potential future work. Besides, the computational procedure in this article can be further customized and enhanced to explore variable selection performance. As we have shown in this article that popular penalized regression-based methods such as the Lasso, the SCAD, and the MCP, which depend on a linear model assumption, are not able to capture nonlinear relationships between the response variable and the predictors, further developing this new model to perform a high-dimensional variable selection of active variables while performing variable selection for variables with thresholds can be a future research work. In addition, the computational procedure has a high computational cost as the dimension of covariates increases. Therefore, a more efficient algorithm can be developed for a higher p dimension.

In addition, since we only impose two thresholds on each predictor, the model can be extended to handle multiple thresholds (i.e., multi-way truncation), making the extended model more complex. For instance, m numbers of thresholds are considered for each variable. Using the notation defined in Section 2.1, each variable can be truncated using the thresholds, for example, the variable X_{ij} can be truncated to $T_{ij} = (X_{ij}I(c_{j,1} < X_{ij} \leq c_{j,2}), X_{ij}I(c_{j,2} < X_{ij} \leq c_{j,3}), \ldots, X_{ij}I(c_{j,m-1} < X_{ij} \leq c_{j,m}))$ where T_{ij} contains the truncations of the variable X_{ij}. The random covariate vector $\mathbf{Z}_i(\boldsymbol{c}) = (1, T_{i1}, \ldots, T_{ip})'$ is a $(mp+1) \times 1$ vector. Here, an additional assumption is imposed on the order of the change-points that is $c_{j,1} \leq c_{j,2} \leq \cdots \leq c_{j,m}$ for all j. The types of associations between the response variable and predictors are similar to the discussions in Section 2.1 with slight modifications on the threshold parameters.

The TWT-LR model is reduced to a linear regression model when there is no change-point. On the other hand, the TWT-LR model can be expressed as a tree structure, as illustrated in Figures C10 and C11 in Appendix, supplementary materials, but the TWT-LR model cannot be generalized to every tree structure. Furthermore, different from the popular tree-based methods, we do not fit the intermediate steps. To further illustrate the idea of performing classification similar to the tree-

based method, we consider two predictors with truncation at c_1^* for the first variable and c_2^* for the second variable, respectively, as in Figure C10. The data can be truncated into four different quadrants that can be represented using a tree structure shown in Figure C11. The truncation performed using the TWT-LR model can be explained using a tree structure.

The new regression model is continuous when the response variable and predictor are linearly related or for one threshold without intercept. The regression is discontinuous otherwise. As it is not always easy to determine if a regression continuity or discontinuity should be used, and it is not the main focus of this article, the regression continuity can be considered as future work. In addition, the estimation procedure will be different in the regression continuity setting.

Supplementary Materials

The supplementary materials contain codes, 4 datasets used in the article, a readme file and an online supplement containing theoretical justifications, simulation results and real data results.

Acknowledgments

The authors thank the editor, the associate editor, and two anonymous referees for their tremendous efforts and insightful comments that substantially improve the article's quality and presentation.

Funding

The partial support from NSF-DMS-2012298 is also acknowledged.

References

Breheny, P., and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, with Applications to Biological Feature Selection," *The Annals of Applied Statistics*, 5, 232–253. [13]

Candès, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation when p is much Larger than n," *The Annals of Statistics*, 35, 2313–2351. [1]

Chen, M., Lian, Y., Chen, Z., and Zhang, Z. (2017), "Sure Explained Variability and Independence Screening," *Journal of Nonparametric Statistics*, 29, 849–883. [2]

Ciuperca, G. (2014), "Model Selection by Lasso Methods in a Change-Point Model," *Statistical Papers*, 55, 349–374. [2]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1]

Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020), *Statistical Foundations of Data Science*, Boca Raton, FL: Chapman and Hall/CRC. [1]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [2]

Feder, P. I. (1975), "On Asymptotic Distribution Theory in Segmented Regression Problems," *The Annals of Statistics*, 3, 49–83. [1]

Gao, M., Kong, W., Huang, Z., and Xie, Z. (2020), "Identification of Key Genes Related to Lung Squamous Cell Carcinoma using Bioinformatics Analysis," *International Journal of Molecular Sciences*, 21. DOI: 10.3390/ijms21082994. [13]

Hall, P., and Miller, H. (2009), "Using Generalized Correlation to Effect Variable Selection in Very High Dimensional Problems," *Journal of Computational and Graphical Statistics*, 18, 533–550. [2]

Hansen, B. E. (2000), "Sample Splitting and Threshold Estimation," *Econometrica*, 68, 575–603. [1]

——— (2017), "Regression Kink with an Unknown Threshold," *Journal of Business and Economic Statistics*, 35, 228–240. [1,2,5]

Harchaoui, Z., and Lévy-Leduc, C. (2012), "Multiple Change-Point Estimation with a Total Variation Penalty," *Journal of the American Statistical Association*, 106, 1480–1493. [2]

Hinkley, D. V. (1969), "Inference about the Intersection in Two-Phase Regression," *Biometrika*, 56, 495–504. [1]

Kaul, A., Jandhyala, V. K., and Fotopoulos, S. B. (2019a), "An Efficient Two Step Algorithm for High Dimensional Change Point Regression Models without Grid Search," *Journal of Machine Learning Research*, 20, 1–40. [2]

Kaul, A., Jandhyala, V. K., and Fotopoulos, S. B. (2019b), "Detection and Estimation of Parameters in High Dimensional Multiple Change Point Regression Model via ℓ_1/ℓ_0 Regularization and Discrete Optimization," arXiv:1906.04396. [2,10,12]

Ke, Y., Li, J., and Zhang, W. (2016), "Structure Identification in Panel Data Analysis," *The Annals of Statistics*, 44, 1193–1233. [2,8]

Ke, Z. T., Fan, J., and Wu, Y. (2015), "Homogeneity Pursuit," *Journal of the American Statistical Association*, 110, 175–194. [2,8]

Knight, K., and Fu, W. (2000), "Asymptotics for Lasso-Type Estimators," *The Annals of Statistics*, 28, 1356–1378. [1]

Knowles, M., Siegmund, D., and Zhang, H. (1991), "Confidence Regions in Semilinear Regression," *Biometrika*, 78, 13–31. [1]

Lee, S., Seo, M. H., and Shin, Y. (2016), "The Lasso for High-Dimensional Regression with a Possible Change-Point," *Journal of the Royal Statistical Society*, Series B, 78, 193–210. [2,5,10,12]

Leonardi, F., and Bühlmann, P. (2016), "Computationally Efficient Change Point Detection for High-Dimensional Regression," arXiv preprint arXiv:1601.03704. [2]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [2,10]

Lian, H., Qiao, X., and Zhang, W. (2021), "Homogeneity Pursuit in Single Index Models based Panel Data Analysis," *Journal of the Royal Statistical Society*, Series B, 39, 386–401. [2]

Porter, J., and Yu, P. (2015), "Regression Discontinuity Designs with Unknown Discontinuity Points: Testing and Estimation," *Journal of Econometrics*, 189, 132–147. [2]

Siegmund, D. O., and Zhang, H. (1993), "The Expected Number of Local Maxima of a Random Field and the Volume of Tubes," *The Annals of Statistics*, 21, 1948–1966. [1]

Siegmund, D. O., and Zhang, H. (1994), "Confidence Regions in Broken Line Regression," *IMS Lecture Notes, Monograph Series*, 23, 292–316. [1]

Song, Z., Zhang, Y., Chen, Z., and Zhang, B. (2021), "Identification of Key Genes in Lung Adenocarcinoma based on a Competing Endogenous RNA Network," *Oncology Letters*, 60. DOI: 10.3892/ol.2020.12322. [13]

Tang, L., and Song, P. X. (2016), "Fused Lasso Approach in Regression Coefficients Clustering — Learning Parameter Heterogeneity in Data Integration," *Journal of Machine learning Research*, 17, 1–23. [2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

Wang, D., Zhao, Z., Lin, K., and Willet, R. (2021), "Statistically and Computationally Efficient Change Point Localization in Regression Settings," *Journal of Machine Learning Research*, 22, 1–46. [2]

Wang, T., and Samworth, R. J. (2018), "High Dimensional Change Point Estimation via Sparse Projection," *Journal of the Royal Statistical Society*, Series B, 80, 57–83. [2]

Wang, W., Phillips, P. C., and Su, L. (2018), "Homogeneity Pursuit in Panel Data Models: Theory and Application," *Journal of Applied Econometrics*, 33, 797–825. [2]

Wang, W., and Su, L. (2021), "Identifying Latent Group Structures in Nonlinear Panels," *Journal of Econometrics*, 220, 272–295. [2]

Zhang, B., Geng, J., and Lai, L. (2015), "Multiple Change-Points Estimation in Linear Regression Models via Sparse Group Lasso," *IEEE Transactions on Signal Processing*, 63, 2209–2224. [2]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [1]

Zhang, Z. (2021), "Functional Effects of Four or Fewer Critical Genes Linked to Lung Cancers and New Subtypes Detected by a New Machine Learning Classifier," *Journal of Clinical Trials*, 11, 001. [13]

Zhang, Z. (2022), "Lift the Veil of Breast Cancers using 4 or Fewer Critical Genes," *Cancer Informatics*, 21, 1–11. [13]

Zheng, S., Shi, N.-Z., and Zhang, Z. (2012), "Generalized Measures of Correlation for Asymmetry, Nonlinearity, and Beyond," *Journal of the American Statistical Association*, 107, 1239–1252. [2]

Zhou, J., Mu, M., Xing, Y., Xin, Z., Danting, L., Yafeng, L., Jun, X., Wangfa, H., Lijun, Z., Jing, W., and Dong, H. (2020), "Identification of Key Genes in Lung Adenocarcinoma and Establishment of Prognostic Mode," *Frontiers in Molecular Biosciences*, 7. DOI: 10.3389/fmolb.2020.561456. [13]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [1]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [1]