STATISTICAL INFERENCE FOR FOUR-REGIME SEGMENTED REGRESSION MODELS

BY HAN YAN[1,a] AND SONG XI CHEN[2,b]

[1] *Guanghua School of Management, Peking University,* [a] *hanyan@stu.pku.edu.cn*

[2] *Department of Statistics and Data Science, Tsinghua University,* [b] *sxchen@tsinghua.edu.cn*

Segmented regression models offer model flexibility and interpretability as compared to the global parametric and the nonparametric models, and yet are challenging in both estimation and inference. We consider a four-regime segmented model for temporally dependent data with segmenting boundaries depending on multivariate covariates with non-diminishing boundary effects. A mixed integer quadratic programming algorithm is formulated to facilitate the least square estimation of the regression and the boundary parameters. The rates of convergence and the asymptotic distributions of the least square estimators are obtained for the regression and the boundary coefficients, respectively. We propose a smoothed regression bootstrap to facilitate inference on the parameters and a model selection procedure to select the most suitable model within the model class with at most four segments. Numerical simulations and a case study on air pollution in Beijing are conducted to demonstrate the proposed approach, which shows that the segmented models with three or four regimes are suitable for the modeling of the meteorological effects on the $\mathrm{PM}_{2.5}$ concentration.

1. Introduction. Regression analysis is a pivotal tool in modeling the relationship between dependent and independent variables and for prediction purposes. It is often conducted via two types of models: the global parametric and local nonparametric models. The global parametric models, such as the linear and polynomial regression models, have the advantages of interpretability and computation simplicity. However, they often perform poorly due to model misspecification as the underlying model may change over different parts of the domain. To have better adaptability, nonparametric local models facilitated by the kernel smoothing, the wavelets or splines, or the regression trees, have been introduced. The local model's complexities increase with the data's dimension and the sample sizes, elevating the risk of overfitting. The segmented model is a compromise between the global and the local models as they are as interpretable as the global parametric models but have improved model specifications.

Conventional threshold regression model (also called regime switching model) [32] was the first generation of the segmented models. It assumes that the regression function is of form $\mathbb{E}(Y|\boldsymbol{X},Z)=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}+\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\delta}\mathbb{1}(Z>r)$, where Z is an observable scalar variable that can be either a time index or a pre-specified random variable. The threshold regression model has a wide range of applications in empirical research, ranging from modeling effects of shocks to economic systems over the business cycles [27], the dose-response models in biostatistics [29], and in sociological research [5]. Statistical inference of the threshold regression model with a univariate splitting variable has been well developed. [6], [15] and [16] established asymptotic properties of the least squares estimators of the threshold regression models and proposed tests on the threshold effect. As extensions, [12] and [23] introduced the multiple

MSC2020 subject classifications: Primary 62J99, 62H12; secondary 62F12.

Keywords and phrases: Mixed integer programming, Segmented model, Smoothed regression bootstrap, Temporal dependence, Threshold regression.

threshold regression model $\mathbb{E}(Y|\boldsymbol{X}, Z) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \sum_{k=1}^{K} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\delta}_k \mathbb{1}(Z > r_k)$ with K splits and $K+1$ regimes (segments) and investigated the statistical inference problems.

A limitation of the existing threshold regression approach is that the splitting threshold is largely determined by a univariate variable Z. [1] showed difficulties in finding the univariate splitting variable in the analysis of macroeconomic effects of fiscal policies, and [19] indicated that a univariate Z was not suitable to regulate the gene effects on disease risks as the risk of developing a particular disease was due to multiple genes. Recently, [22] and [35] extended the threshold regression to allow regime switching driven by a multivariate random vector Z which is either observable or obtained via a factor model. Although these works overcome the limitation of the univariate split variable, the setting of at most two regimes can be restrictive for some applications. Machine learning methods, such as the convex piece-wise linear fitting, can produce segmented linear regression with unlimited number of regimes. However, as these methods were focused on the fitting performances, the underlying segmented models may not be identifiable with the suggested procedures. The finite mixture models (FMM) proposed by [20] can also produce a subgroup linear model fitting for heterogeneous data. However, the subgroups from the FMMs do not lead to parameterized boundaries, and thus are less interpretable than the segmented linear models.

Our study is motivated from modelling the meteorological effects on PM$_{2.5}$ concentration in Beijing, where a global parametric model is too simple to offer good fitting performances and a nonparametric model may be too local and do not provide sufficient atmospheric interpretation. The air pollution in Beijing is typically governed by different meteorological regimes, namely the removal process by favourable northerly wind which removes PM$_{2.5}$ to a low level, the calm regime between the northerly cleaning and the start of the transported pollution driven under the southerly wind, the pollution growth regime under southerly wind that transports polluted air from the south, and air stagnation regime after the pollution has peaked, followed by the next removal process by the northerly wind. These motivate the four-regime segmented regression model in this work. As the air quality and meteorological data are time series, we consider temporally dependent data in the study.

Motivated by the air pollution problem, we consider four-regime regression models whose splitting hypersplines are determined by linear combinations of two multivariate covariates $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$, where the splitting variables $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ can be any regressors, and the two splitting hyperplanes can intersect. These make the four regime-regression model less restrictive than the multiple threshold regression model of [2] and [23] where the splitting variable Z is univariate, and hence allows not necessarily parallel boundary hyperplanes. The four-regime models include two and three regime models as special cases, where the splitting boundaries are either parallel or two adjacent regimes share the same regression coefficient and hence can be merged.

The main contributions of the study are the following. We first establish the consistency and the asymptotic distributions of the least squares estimators (LSEs) for both the boundary and the regression coefficients under the four-regime regression model with temporally dependent ρ-mixing observations, overcoming challenges posed by (i) the irregular objective function, (ii) the fixed boundary edge effects rather than the diminishing effects commonly treated in the literature and (iii) the unconventional form of the asymptotic distribution for the boundary coefficient vector. It is found that the asymptotic distribution of LSEs for the boundary coefficients is determined by the minimizers of a compound multivariate Poisson process, whose jumps depend on the points near the true hyperplanes, and the boundary coefficient estimators are asymptotically independent of the regression coefficient estimators.

The generalization to the four regimes with two splitting boundaries brings considerable computational challenges. Although the LSE of the conventional threshold regression can be obtained with the grid search method, it is not practical in our setting as the boundaries are defined with multivariate variables. To overcome the challenges, we draw inspiration from [3] and [22] and propose an algorithm based on the mixed integer quadratic programming (MIQP), which is not only computationally efficient but also can be further accelerated by adding an iterative component. It is shown the algorithm can facilitate efficient computation of the LSEs with the rather non-regular form of the least squares objective function.

To permit statistical inference, especially in light of the rather unusual asymptotic distribution for the boundary coefficient estimates, we develop a smoothed regression bootstrap method and establish its consistency for approximating the distribution of the LSEs. Furthermore, the properties of the LSEs under degenerated segmented models with less than four regimes are investigated. In order to find the right segmented models with up to four segments, we propose a model selection method with a backward elimination procedure that is shown to be able to consistently choose the right number of regimes.

The paper is organized as follows. Section 2 introduces the four-regime regression model. Section 3 presents the theoretical properties and the asymptotic distribution of the LSEs for the regression and boundary parameters. In Section 4, we construct a mixed integer quadratic programming (MIQP) algorithm to efficiently compute the LSEs. Section 5 considers inference problems for the four-regime regression model. Section 6 investigates the properties of the proposed estimator under degenerated models with less than four regimes and proposes a model selection method. Sections 7 and 8 report simulation and empirical results, respectively. Section 9 conclude the paper with possible extensions. All technical proofs are relegated to a supplementary material (SM, [33]).

2. Model setup.

We first introduce some notations. We use $\mathbb{1}(\mathcal{A})$ for the indicator function of an event $\mathcal{A}$, $\|\boldsymbol{v}\| = (\sum_{i=1}^d v_i^2)^{1/2}$ for the L_2-norm of vector $\boldsymbol{v} = (v_1, \cdots, v_d)^{\mathrm{T}}$ and $\mathcal{N}(\boldsymbol{v}_0; \delta) = \{\boldsymbol{v} : \|\boldsymbol{v} - \boldsymbol{v}_0\| \le \delta\}$ for the δ-neighborhood of $\boldsymbol{v}$. Define $\boldsymbol{v}_{-1}$ as the sub-vector of $\boldsymbol{v}$ excluding its first element, i.e., $\boldsymbol{v}_{-1} = (v_2, \cdots, v_d)^{\mathrm{T}}$. We use $|E|$ to denote the cardinality of a set E. For any two sets E_1 and E_2, we denote $E_1 \triangle E_2 = (E_1 \setminus E_2) \cup (E_2 \setminus E_1)$ as their symmetric difference.

Let $\{\boldsymbol{W}_t = (Y_t, \boldsymbol{X}_t, \boldsymbol{Z}_{1,t}, \boldsymbol{Z}_{2,t})\}_{t=1}^T$ be a sequence of observations, where Y_t is the response variable to covariates $\boldsymbol{X}_t \in \mathbb{R}^p$ and two partitioning variables $\boldsymbol{Z}_{i,t} \in \mathbb{R}^{d_i}$ for $i = 1$ and 2, which determine the boundaries of the segments or regimes. The variables $\boldsymbol{X}_t$, $\boldsymbol{Z}_{1,t}$ and $\boldsymbol{Z}_{2,t}$ can share common variables. The four-regime regression model is

$$Y_t = \sum_{k=1}^4 \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{\beta}_{k0} \mathbb{1}\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\} + \varepsilon_t, \tag{2.1}$$

where $\boldsymbol{Z}_t$ is the union of variables of $\boldsymbol{Z}_{1,t}$ and $\boldsymbol{Z}_{2,t}$, $\{\boldsymbol{\beta}_{k0}\}_{k=1}^4$ are the regression coefficients, $\{\boldsymbol{\gamma}_{i0}\}_{j=1}^2$ are the boundary coefficients, ε_t is the residual satisfying $\mathbb{E}(\varepsilon_t | \boldsymbol{X}_t, \boldsymbol{Z}_t) = 0$ with a finite second moment, and $R_k(\boldsymbol{\gamma}_0)$ is the k-th region split by the hyperplanes $\{H_{i0} : \boldsymbol{z}_i^{\mathrm{T}} \boldsymbol{\gamma}_{i0} = 0\}_{i=1}^2$ for $\boldsymbol{z}_i \in \mathbb{R}^{d_i}$. The overall parameter of interest is $\boldsymbol{\theta} = (\boldsymbol{\gamma}^{\mathrm{T}}, \boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \cdots, \boldsymbol{\beta}_4^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$. We let $\boldsymbol{\theta}_0$, $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ denote the respective true parameters. For any observation $\boldsymbol{W}_t$, it is the signs of $\boldsymbol{Z}_{1,t}^{\mathrm{T}} \boldsymbol{\gamma}_{10}$ and $\boldsymbol{Z}_{2,t}^{\mathrm{T}} \boldsymbol{\gamma}_{20}$ that determine which regression region it is located at. Denote by $\mathbb{1}_1(U, V) = \mathbb{1}(U > 0, V > 0)$, $\mathbb{1}_2(U, V) = \mathbb{1}(U \le 0, V > 0)$, $\mathbb{1}_3(U, V) = \mathbb{1}(U \le 0, V \le 0)$ and $\mathbb{1}_4(U, V) = \mathbb{1}(U > 0, V \le 0)$. Then we can write Model (2.1) equivalently as

$$Y_t = \sum_{k=1}^4 \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{\beta}_{k0} \mathbb{1}_k(\boldsymbol{Z}_{1,t}^{\mathrm{T}} \boldsymbol{\gamma}_{10}, \boldsymbol{Z}_{2,t}^{\mathrm{T}} \boldsymbol{\gamma}_{20}) + \varepsilon_t, \tag{2.2}$$

which explicitly reflects the role of $\boldsymbol{\gamma}_0$ in Model (2.1).

REMARK 2.1. Although the splitting hyperplanes appears linear, non-linearity may be accommodated by including nonlinear transformed variables in $\boldsymbol{Z}_i (i = 1, 2)$, for instance, $\boldsymbol{Z}_1 = (Z_1, Z_1^2, 1)^{\mathrm{T}}$. The same extension can be conducted to $\boldsymbol{X}$. It is also noted that in the special case of $\boldsymbol{Z}_{1,t}$ having the same distribution with $\boldsymbol{Z}_{2,t}$, the four segments under $\gamma_0 = (\gamma_{10}^{\mathrm{T}}, \gamma_{20}^{\mathrm{T}})^{\mathrm{T}}$ are not distinguishable from that under $\tilde{\gamma}_0 = (\gamma_{20}^{\mathrm{T}}, \gamma_{10}^{\mathrm{T}})^{\mathrm{T}}$. Consequently, $\boldsymbol{\theta}_0$ is only identifiable up to some permutations. To avoid such situation, we assume that the distributions of $\boldsymbol{Z}_{1,t}$ and $\boldsymbol{Z}_{2,t}$ are distinct.

REMARK 2.2. Since the signs of $\boldsymbol{Z}_1^{\mathrm{T}}\gamma_{10}$ and $\boldsymbol{Z}_2^{\mathrm{T}}\gamma_{20}$ determine the regimes in Model (2.1), γ_{10} and γ_{20} have to be normalized in order to be identifiable. For any candidate γ_i of γ_{i0}, we normalize it by its first element $\gamma_{i,1}$, resulting in $\gamma_i =: (1, \widetilde{\gamma}_i)$ where $\widetilde{\gamma}_i$ is assumed to take values in a compact set. As noted in [22], an alternative normalization is $\|\gamma_i\|_2 = 1$. In this study, we employ the former as it has one less parameter.

3. Estimation and asymptotic properties. In this section, we outline the least squares (LS) estimation for $\boldsymbol{\theta}_0$ of the four-regime regression model, and establish the convergence rates of the LS estimators for the regression coefficient $\widehat{\boldsymbol{\beta}}$ and the boundary coefficient $\widehat{\gamma}$ followed by providing their asymptotic distributions.

With the data sample $\{\boldsymbol{W}_t = (Y_t, \boldsymbol{X}_t, \boldsymbol{Z}_{1,t}, \boldsymbol{Z}_{2,t})\}_{t=1}^T$, in view of $\mathbb{E}(\varepsilon_t|\boldsymbol{X}_t, \boldsymbol{Z}_t) = 0$, we define the following least squares criterion function

$$\mathbb{M}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\left\{Y_t - \sum_{k=1}^{4}\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_k \mathbb{1}_k(\boldsymbol{Z}_{1,t}^{\mathrm{T}}\gamma_1, \boldsymbol{Z}_{2,t}^{\mathrm{T}}\gamma_2)\right\}^2 =: \frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \boldsymbol{\theta}), \tag{3.1}$$

and the parameter space is $\Theta = \Gamma_1 \times \Gamma_2 \times \mathcal{B}^4$, where Γ_i is a compact set in $\mathbb{R}^{d_i}$ and the first element of any $\gamma \in \Gamma_i$ is normalized as 1 for each $i = 1, 2$, and $\mathcal{B}$ is a compact set in $\mathbb{R}^p$. Since $\mathbb{M}_T(\boldsymbol{\theta})$ is strictly convex in $\boldsymbol{\beta}$ and piece-wise constant in γ with at most T jumps, it has a unique minimizer $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \cdots, \widehat{\boldsymbol{\beta}}_4^{\mathrm{T}})^{\mathrm{T}}$ for $\boldsymbol{\beta}$, but a set of minimizers for γ, which is denoted as $\widehat{\mathcal{G}}$, such that a LSE $\widehat{\boldsymbol{\theta}} = (\widehat{\gamma}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ satisfies

$$\mathbb{M}_T(\widehat{\boldsymbol{\theta}}) = \inf_{\boldsymbol{\theta}\in\Theta} \mathbb{M}_T(\boldsymbol{\theta}) \ \text{ for any } \ \widehat{\gamma} \in \widehat{\mathcal{G}}. \tag{3.2}$$

It is noted that for any two $\widehat{\gamma}, \widehat{\gamma}' \in \widehat{\mathcal{G}}$, the segmented regimes under the corresponding hyperplanes must be the same, as otherwise the estimated regression coefficients will be distinct. In addition, the set $\widehat{\mathcal{G}}$ is convex since for each $i = 1$ or 2, $\boldsymbol{Z}_{i,t}^{\mathrm{T}}\widehat{\gamma}_i > 0$ and $\boldsymbol{Z}_{i,t}^{\mathrm{T}}\widehat{\gamma}_i' > 0$ imply that $\boldsymbol{Z}_{i,t}^{\mathrm{T}}\widetilde{\gamma}_i > 0$ for all $\widetilde{\gamma}_i = \alpha\widehat{\gamma}_i + (1-\alpha)\widehat{\gamma}_i'$ with $\alpha \in [0, 1]$. In the rest of this section, we investigate the properties of the LS estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\gamma}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ with $\widehat{\gamma} \in \widehat{\mathcal{G}}$.

3.1. *Identification and consistency.* Here we discuss the identification of $\boldsymbol{\theta}_0$ and establish the consistency of the LSEs $\widehat{\boldsymbol{\theta}}$. Let $\boldsymbol{W} = (Y, \boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$ follow the stationary distribution $\mathbb{P}_0$ of $\boldsymbol{W}_t$, and $q_i = \boldsymbol{Z}_i^{\mathrm{T}}\gamma_{i0}$ for $i = 1$ and 2 to indicate whether $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2)$ is located on the true hyperplane $H_{i0} : \boldsymbol{Z}_i^{\mathrm{T}}\gamma_{i0} = 0$ or not. Let $\mathcal{S}(i)$ be the set consisting of index pairs (k, h) if $R_k(\gamma_0)$ and $R_h(\gamma_0)$ are two adjacent regions split by H_{i0}. Specifically, $\mathcal{S}(1) = \{(1, 2), (2, 1)(3, 4), (4, 3)\}$ and $\mathcal{S}(2) = \{(1, 4), (4, 1), (2, 3), (3, 2)\}$ according to the provision in the lines above (2.2). Furthermore, let $\boldsymbol{Z}$ be the union vector of variables in $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$

ASSUMPTION 1 (temporal dependence). (i) The time series $\{\boldsymbol{W}_t\}_{t\geq 1}$ is strictly stationary and ρ-mixing with the mixing coefficient $\rho(t) \leq c\alpha^t$ for finite positive constants c and

$\alpha \in (0,1)$, where $\rho(t) = \sup_{s,t\geq 1}\left\{\sup \mathrm{Corr}(f,g) : f \in \Omega_1^s, g \in \Omega_{s+t}^{\infty}\right\}$, where Ω_i^j denotes the σ-filed generated by $\{\boldsymbol{W}_t : i \leq t \leq j\}$. (ii) $\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$, where $\mathcal{F}_{t-1}$ is a filtration generated by $\{(\boldsymbol{X}_i, \boldsymbol{Z}_{1i}, \boldsymbol{Z}_{2i}, \varepsilon_{i-1}) : i \leq t\}$.

ASSUMPTION 2 (identification). For $i \in \{1,2\}$ and $k,h \in \{1,\cdots,4\}$, (i) $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are not identically distributed. (ii) There exists a $j \in \{1,\cdots,d_i\}$ such that $\mathbb{P}(|q_i| \leq \epsilon|\boldsymbol{Z}_{-j,i}) > 0$ almost surely for $\boldsymbol{Z}_{-j,i}$ and for any $\epsilon > 0$, where $\boldsymbol{Z}_{-j,i}$ is the vector after excluding $\boldsymbol{Z}_i$'s jth element; without loss of generality, assume $j = 1$. (iii) For any $\boldsymbol{\gamma} \in \Gamma_1 \times \Gamma_2$ and $\mathbb{P}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0) \cap R_h(\boldsymbol{\gamma})\} > 0$, the smallest eigenvalue of $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}|\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0) \cap R_h(\boldsymbol{\gamma})\} \geq \lambda_0$ for some constant $\lambda_0 > 0$. (iv) For $(k,h) \in \mathcal{S}(i)$, $\|\boldsymbol{\beta}_{k0} - \boldsymbol{\beta}_{h0}\| > c_0$ for some constant $c_0 > 0$.

ASSUMPTION 3. (i) $\mathbb{E}(Y^4) < \infty$, $\mathbb{E}(\|\boldsymbol{X}\|^4) < \infty$ and $\max_{i=1,2}\mathbb{E}(\|\boldsymbol{Z}_i\|) < \infty$. (ii) For each $i = 1$ and 2, $\mathbb{P}(\boldsymbol{Z}_i^{\mathrm{T}}\gamma_1 < 0 < \boldsymbol{Z}_i^{\mathrm{T}}\gamma_2) \leq c_1\|\gamma_1 - \gamma_2\|$ if $\gamma_1, \gamma_{2,} \in \mathcal{N}(\gamma_{i0};\delta_0)$, for some constants $\delta_0, c_1 > 0$.

Assumption 1 (i) prescribes the strict stationarity and ρ-mixing condition on the time series, as used in the existing time-series threshold regression literature ([16] and [22]). It is noted that such a decaying rate is only required in deriving the limiting distribution of $\widehat{\boldsymbol{\gamma}}$, which can be relaxed to the polynomial decay for Theorem 3.1 and Theorem 3.2. Assumption 1 (ii) imposes a martingale difference condition for the noises, which is standard for time series regressions.

Assumption 2 is for the identification of $\boldsymbol{\theta}_0$. Specifically, without Assumption 2 (i), $(\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$ are not distinguishable from $(\boldsymbol{\gamma}_2^{\mathrm{T}}, \boldsymbol{\gamma}_1^{\mathrm{T}})^{\mathrm{T}}$ as discussed in Remark 2.1. It is noted that the methods and theories in the rest of the papers are applicable without such a condition, while a permutation for $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$ is possibly required. Section F of the SM ([33]) provides sufficient conditions for Assumption 2 (ii), which ensures there are positive probability of observations located around the true splitting hyperplanes. Discrete variables can be accommodated in $\boldsymbol{Z}_i$, as long as it includes at least one continuous variable, say $Z_{1,i}$. Otherwise, if all the splitting variables are discretely distributed, then $\mathbb{E}\{m(\boldsymbol{W},\boldsymbol{\theta})\}$ will be piece-wise constant and $\boldsymbol{\gamma}_0$ will not be identifiable. Assumption 2 (iii) guarantees that the splittings by candidate hyperplanes do not lead to degenerated covariance matrices, which is needed for the identification of $\boldsymbol{\beta}_0$. Assumption 2 (iv) means that adjacent regimes have distinguishable regression coefficients so that the splitting effect of each hyperplane is strictly bounded away 0, which is similar to the fixed threshold effect models treated in [6] and [35]. Assumption 3 (i) is a moment condition, and (ii) means $\mathbb{P}(\boldsymbol{Z}_i^{\mathrm{T}}\boldsymbol{\gamma} < 0)$ is continuous at $\boldsymbol{\gamma}_{i0}$, implying that $\mathbb{E}\{m(\boldsymbol{W};\boldsymbol{\theta})\}$ is continuous at the true parameter $\boldsymbol{\theta}_0$.

The identification of $\boldsymbol{\theta}_0$ is formally ensured in the following proposition.

PROPOSITION 3.1. *Under Assumptions 1 and 2,* $\mathbb{E}\{m(\boldsymbol{W},\boldsymbol{\theta})\} > \mathbb{E}\{m(\boldsymbol{W},\boldsymbol{\theta}_0)\}$ *for any* $\boldsymbol{\theta} \in \Theta$ *and* $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

The proposition ensures that despite the multiple LS estimates $\widehat{\boldsymbol{\gamma}}$, the underlying $\boldsymbol{\gamma}_0$ is unique. The following theorem shows that any LSE estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ defined in (3.2) are consistent to $\boldsymbol{\theta}$. It is worth noting that though there exist infinitely many solutions $\widehat{\boldsymbol{\gamma}}$ which are collected in the convex set $\widehat{\mathcal{G}}$, the consistency of each $\widehat{\boldsymbol{\gamma}}$ can be guaranteed, which implies that the solution set $\widehat{\mathcal{G}}$ is a local neighborhood of $\boldsymbol{\gamma}_0$ with a shrinking radius.

THEOREM 3.1. *Under Assumptions 1–3, let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ for any $\widehat{\boldsymbol{\gamma}} \in \widehat{\mathcal{G}}$, , then $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$ as $T \to \infty$.*

With the estimated splitting hyperplanes, each datum can be classified into one of the four estimated regimes $\{R_k(\widehat{\boldsymbol{\gamma}})\}_{k=1}^4$. Besides the estimation accuracy of $\boldsymbol{\theta}_0$, the classification accuracy is also an important criterion. It is shown next that the estimated regime $R_k(\widehat{\boldsymbol{\gamma}})$ is consistent to the true regime $R_k(\boldsymbol{\gamma}_0)$ for each $k = 1, \cdots, 4$.

COROLLARY 3.1. *Under the conditions of Theorem 3.1, $\mathbb{P}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0) \bigtriangleup R_k(\widehat{\boldsymbol{\gamma}})\} \to 0$ as $T \to \infty$ for all $k \in \{1, \cdots, 4\}$.*

3.2. *Convergence rates and asymptotic distributions.* We first study the convergence rates of the LSEs $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, which require the following conditions.

ASSUMPTION 4. (i) For $i = 1$ and 2, there exist constants $\delta_1, c_2 > 0$ such that if $\epsilon \in (0, \delta_1)$ then $\mathbb{P}(|q_i| < \epsilon | \boldsymbol{Z}_{-1,i}) \geq c_2 \epsilon$ almost surely. (ii) For $i = 1$ and 2, there exists a neighborhood $\mathcal{N}_i = \mathcal{N}(\boldsymbol{\gamma}_{i0}; \delta_2)$ of $\boldsymbol{\gamma}_{i0}$ for some $\delta_2 > 0$, such that $\inf_{\boldsymbol{\gamma} \in \mathcal{N}_i} \mathbb{E}(\|\boldsymbol{X}^{\mathrm{T}} \boldsymbol{\delta}_{kh,0}\| | \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma} = 0) > 0$ almost surely for each $(k, h) \in \mathcal{S}(i)$, where $\boldsymbol{\delta}_{kh,0} = \boldsymbol{\beta}_{k0} - \boldsymbol{\beta}_{h0}$. (iii) $\mathbb{P}(\boldsymbol{Z}_1^{\mathrm{T}} \boldsymbol{\gamma}_1 < 0 < \boldsymbol{Z}_1^{\mathrm{T}} \boldsymbol{\gamma}_2, \boldsymbol{Z}_2^{\mathrm{T}} \boldsymbol{\gamma}_3 < 0 < \boldsymbol{Z}_2^{\mathrm{T}} \boldsymbol{\gamma}_4) \leq c_3 \|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\| \|\boldsymbol{\gamma}_3 - \boldsymbol{\gamma}_4\|$ for some constant $c_3 > 0$ if $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2 \in \mathcal{N}_1$ and $\boldsymbol{\gamma}_3, \boldsymbol{\gamma}_4 \in \mathcal{N}_2$. (iv) $\sup_{\boldsymbol{\gamma} \in \mathcal{N}_i} \mathbb{E}(\|\boldsymbol{X}\|^8 | \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma} = 0) < \infty$ and $\sup_{\boldsymbol{\gamma} \in \mathcal{N}_i} \mathbb{E}(\varepsilon^8 | \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma} = 0) < \infty$ almost surely.

Assumption 4 (i) strengthens Assumption 2 (i) and is satisfied when the conditional density $f_{q_i | \boldsymbol{Z}_{-1,i}}(q)$ is continuous and bounded away from 0 at $q = 0$ almost surely. Assumption 4 (ii) ensures there is a jump of the regression surface at the splitting hyperplane, which is similar to Assumption D3 of [35] and Assumption 4.(iii) of [22]. Assumption 4 (iii) controls the probability of data near the cross regions of the two hyperplanes, whose sufficient condition is presented in Section F of the SM ([33]). Assumption 4 (iv) requires that $\|\boldsymbol{X}\|$ and ε has a finite moment of the order 8 around the hyperplanes.

The next theorem establishes the rates of convergence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, followed by the convergence rate of the proportions of misclassifications.

THEOREM 3.2. *Under Assumptions 1–4, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(1/\sqrt{T})$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(1/T)$ for any $\widehat{\boldsymbol{\gamma}} \in \widehat{\mathcal{G}}$.*

COROLLARY 3.2. *Under the conditions of Theorem 3.2, $\mathbb{P}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0) \bigtriangleup R_k(\widehat{\boldsymbol{\gamma}})\} = O(1/T)$ for all $k \in \{1, \cdots, 4\}$.*

The theorem, whose proof is in Section B of the SM ([33]), shows that the regression coefficient estimator $\widehat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}_0$ at the standard $\sqrt{T}$-rate, while the boundary parameter estimator $\widehat{\boldsymbol{\gamma}}$, despite having multiple solutions, converges to $\boldsymbol{\gamma}_0$ at the faster T-rate. The super convergence rate attained by $\widehat{\boldsymbol{\gamma}}$ is quite typical for the boundary parameter estimators, for instance, the maximum likelihood estimator for the boundary parameter of uniform distributions, the LS estimator of models with a jump in the conditional density [8], the threshold regression model [6] and the two-regime regression model with a fixed threshold effect [35]. An intuition for the fast convergence of $\widehat{\boldsymbol{\gamma}}$ is that the discontinuity of the regression planes is highly informative for the inference of $\boldsymbol{\gamma}$. It is noted that in the shrinking threshold effect setting $\boldsymbol{\beta}_{10} - \boldsymbol{\beta}_{20} = \boldsymbol{c} T^{-\alpha}$ with $\boldsymbol{c} \neq 0$ and $0 < \alpha < \frac{1}{2}$ adopted by [16] and [22], the convergence rate of $\widehat{\boldsymbol{\gamma}}$ is slower at $T^{1-2\alpha}$.

To present the asymptotic distributions of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, we define for each $k \in \{1, \cdots, 4\}$,

$$B_k = \mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\mathbb{1}(\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0)\} \text{ and } \Sigma_k = B_k^{-1}\mathbb{E}\left\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\varepsilon^2\mathbb{1}(\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0)\right\} B_k^{-1}.$$

Let $q_{i,t} = \boldsymbol{Z}_{i,t}^{\mathrm{T}}\gamma_{i0}$ and $q_i = \boldsymbol{Z}_i^{\mathrm{T}}\gamma_{i0}$ for $i = 1$ and 2. Denote by $s_i^{(k)} = (-1)^{\mathbb{1}(q_i \le 0,\ \forall \boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0))}$ be the sign of q_i for $\boldsymbol{Z} = (\boldsymbol{Z}_1^{\mathrm{T}}, \boldsymbol{Z}_2^{\mathrm{T}}) \in R_k(\boldsymbol{\gamma}_0)$. For instance, $s_1^{(1)} = s_2^{(1)} = 1$ and $s_1^{(2)} = -1, s_2^{(2)} = 1$. If $R_k(\boldsymbol{\gamma}_0)$ and $R_h(\boldsymbol{\gamma}_0)$ are adjacent such that $(k,h) \in \mathcal{S}(i)$ for $i = 1$ or 2, let

$$\xi_t^{(k,h)} = \left(\boldsymbol{\delta}_{kh,0}^{\mathrm{T}}\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\delta}_{kh,0} + 2\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\delta}_{kh,0}\varepsilon_t\right)\mathbb{1}\left\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0) \cup R_h(\boldsymbol{\gamma}_0)\right\} \tag{3.3}$$

where $\boldsymbol{\delta}_{kh,0} = \boldsymbol{\beta}_{k0} - \boldsymbol{\beta}_{h0}$. Let $\boldsymbol{Z}_{-1,i,t}$ be the random vector of $\boldsymbol{Z}_{i,t}$ excluding its first element. Suppose $(q_i, \boldsymbol{Z}_{-1,i}, \xi^{(k,h)})$ follows the stationary distribution of $(q_{i,t}, \boldsymbol{Z}_{-1,i,t}, \xi_t^{(k,h)})$. We denote $F_{q_i|\boldsymbol{Z}_{-1,i}}(q|\boldsymbol{Z}_{-1,i})$ and $F_{\xi^{(k,h)}|q_i,\boldsymbol{Z}_{-1,i}}(\xi|q_i, \boldsymbol{Z}_{-1,i})$ as the conditional distributions of q_i on $\boldsymbol{Z}_{-1,i}$ and $\xi^{(k,h)}$ on $(q_i, \boldsymbol{Z}_{-1,i})$, respectively, and the corresponding conditional densities are $f_{q_i|\boldsymbol{Z}_{-1,i}}(q|\boldsymbol{Z}_{-1,i})$ and $f_{\xi^{(k,h)}|q_i,\boldsymbol{Z}_{-1,i}}(\xi|q_i, \boldsymbol{Z}_{-1,i})$, respectively. Let $\mathcal{Z}_{-1,i}$ be the support of the distribution of $\boldsymbol{Z}_{-1,i}$. The following is needed for the weak convergence of $\widehat{\boldsymbol{\gamma}}$.

ASSUMPTION 5. (i) For $i = 1$ and 2, there exist constants $\delta_3, c_4 > 0$ such that $\mathbb{P}(|q_{i,t}| \le \delta_3, |q_{i,t+j}| \le \delta_3) \le c_4\{\mathbb{P}(|q_{i,t}| \le \delta_3)\}^2$ uniformly for $t \ge 1$ and $j \ge 1$; (ii) For each $\boldsymbol{z}_{-1,i} \in \mathcal{Z}_{-1,i}$, the conditional density $f_{q_i|\boldsymbol{Z}_{-1,i}}(q|\boldsymbol{z}_{-1,i})$ is continuous at $q = 0$ and $c_4 \le f_{q_i|\boldsymbol{Z}_{-1,i}}(0|\boldsymbol{z}_{-1,i}) \le c_5$ for some constants $c_4, c_5 > 0$; (iii) For each $\xi \in \mathbb{R}$ and $\boldsymbol{z}_{-1,i} \in \mathcal{Z}_{-1,i}$, the conditional density $f_{\xi^{(k,h)}|q_i,\boldsymbol{Z}_{-1,i}}(\xi|q_i, \boldsymbol{z}_{-1,i})$ is continuous at $q_i = 0$ and $f_{\xi^{(k,h)}|q_i,\boldsymbol{Z}_{-1,i}}(\xi|0, \boldsymbol{z}_{-1,i}) \le c_6$ for a constant $c_6 > 0$; (iv) $\mathcal{Z}_{-1,i}$ is a compact subset of $\mathbb{R}^{d_i-1}$.

Assumption 5 (i) is a non-clustering condition that states the probability of two points are both located near the splitting hyperplane H_{i0} is of a smaller order compared to that of just one point is located near H_{i0}, which curbs the clustering of extreme events and is similar to Condition C.4 of [7]. Assumption 5 (ii) and (iii) are on the conditional densities $f_{q_i|\boldsymbol{Z}_{-1,i}}$ and $f_{\xi^{(k,h)}|q_i,\boldsymbol{Z}_{-j,i}}$, respectively, which are used to characterize behaviors of the points near H_{i0}. The compactness of $\mathcal{Z}_{-1,i}$ is required by the limiting theory of point processes ([28] and [8]), which may be attained by trimming $\boldsymbol{Z}_{-1,i,t}$ or empirical quantile transformation.

The asymptotic distribution of $\widehat{\boldsymbol{\gamma}}$ needs the following stochastic process

$$D(\boldsymbol{v}) = \sum_{i=1,2}\sum_{k,h\in\mathcal{S}(i)}\sum_{\ell=1}^{\infty}\xi_{i,\ell}^{(k,h)}\mathbb{1}\left\{J_{i,\ell}^{(k,h)} + (\boldsymbol{Z}_{i,\ell}^{(k,h)})^{\mathrm{T}}\boldsymbol{v}_{-1,i} \le 0 < J_{i,\ell}^{(k,h)}\right\}, \tag{3.4}$$

for $\boldsymbol{v} = (\boldsymbol{v}_1^{\mathrm{T}}, \boldsymbol{v}_2^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^{d_1+d_2}$, where $\{(\xi_{i,\ell}^{(k,h)}, \boldsymbol{Z}_{i,\ell}^{(k,h)})\}_{\ell=1}^{\infty}$ are independent copies of $(\bar{\xi}_i^{(k,h)}, \boldsymbol{Z}_{-1,i})$ with $\bar{\xi}_i^{(k,h)} \sim F_{\xi^{(k,h)}|q_i,\boldsymbol{Z}_{-1,i}}(\xi|0, \boldsymbol{Z}_{-1,i})$, and $J_{i,\ell}^{(k,h)} = \mathcal{J}_{i,\ell}^{(k,h)}/f_{q_i|\boldsymbol{Z}_{-1,i}}(0|\boldsymbol{Z}_{i,\ell}^{(k,h)})$ with $\mathcal{J}_{i,\ell}^{(k,h)} = s_i^{(k)}\sum_{n=1}^{\ell}\mathcal{E}_{i,n}^{(k,h)}$ and $\{\mathcal{E}_{i,n}^{(k,h)}\}_{n=1}^{\infty}$ are independent unit exponential variables which are independent of $\{(\xi_{i,\ell}^{(k,h)}, \boldsymbol{Z}_{i,\ell}^{(k,h)})\}_{\ell=1}^{\infty}$. Moreover, $\{(\xi_{i,\ell}^{(k,h)}, \boldsymbol{Z}_{i,\ell}^{(k,h)}, J_{i,\ell}^{(k,h)})\}_{\ell=1}^{\infty}$ are mutually independent with respect to $i = 1, 2$ and $(k,h) \in \mathcal{S}(i)$.

Let $\mathcal{G}_D = \{\boldsymbol{v}_m : D(\boldsymbol{v}_m) \le D(\boldsymbol{v})$ if $\boldsymbol{v} \ne \boldsymbol{v}_m\}$ be the set of minimizers for $D(\boldsymbol{v})$. Since $D(\boldsymbol{v})$ is a piece-wise constant random function, there are infinitely many elements in $\mathcal{G}_D$. Such a phenomenon also appears in the threshold regression, where the minimizers of the process, that is a special case of (3.4), are attained in an interval, whose left endpoint is commonly used as a representative, which is not applicable to our case since $\mathcal{G}_D$ is a polyhedron. As treated in [35], we use the centroid of $\mathcal{G}_D$ as the representative. For any set $\mathcal{A}$ of

d-dimensional vectors, the centroid of $\mathcal{A}$ is $C(\mathcal{A}) = \int_{\boldsymbol{v}\in\mathcal{A}} \boldsymbol{v} d\boldsymbol{v} / \int_{\boldsymbol{v}\in\mathcal{A}} d\boldsymbol{v}$, which can be geometrically interpreted as the center of mass of the set $\mathcal{A}$. Let $\boldsymbol{\gamma}_D^c = C(\mathcal{G}_D)$ and $\widehat{\boldsymbol{\gamma}}^c = C(\widehat{\mathcal{G}})$, where $\widehat{\mathcal{G}}$ is the set for LS estimators for $\boldsymbol{\gamma}$. The former will define the limit of $\widehat{\boldsymbol{\gamma}}^c$ as shown in Theorem 3.3. Numerically, $\widehat{\boldsymbol{\gamma}}^c$ can be approximated by the average of N elements of $\widehat{\mathcal{G}}$ for a sufficiently large N. The following theorem establishes the asymptotic distributions of $\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k0})$ and $T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0)$.

THEOREM 3.3 (Asymptotic distribution). *Under Assumptions 1-5, we have (i)* $\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k0}) \xrightarrow{d} \boldsymbol{N}(0, \Sigma_k)$ *for* $k = 1, \cdots, 4$ *and* $T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0) \xrightarrow{d} \boldsymbol{\gamma}_D^c$*; (ii)* $\{\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k0})\}_{k=1}^4$ *and* $\{T(\widehat{\boldsymbol{\gamma}}_i^c - \boldsymbol{\gamma}_{i0})\}_{i=1}^2$ *are asymptotically independent.*

REMARK 3.1. The limiting process $D(\boldsymbol{v})$ is derived by the asymptotics of the point process induced by $\{(\xi_t^{(k,h)}, \boldsymbol{Z}_{-1,i,t}, Tq_{i,t})\}_{t=1}^T$. The process $D(\boldsymbol{v})$ can be regarded as a multivariate compound Poisson process, whose jump sizes are $\{\xi_{i,\ell}^{(k,h)}\}_{\ell=1}^\infty$ and jump locations are determined by the counting measure induced by $\{(J_{i,\ell}^{(k,h)}, \boldsymbol{Z}_{i,\ell}^{(k,h)})\}_{\ell=1}^\infty$. Intuitively, this is because $D(\boldsymbol{v})$ largely relies on those points lying in a local neighborhood of the true splitting hyperplanes, whose $|q_{i,t}|$ are on the order of $O(T^{-1})$, which are rare events with their occurrences asymptotically governed by a Poisson process. In the case of univariate threshold model where $\boldsymbol{Z}_i = (Z, 1)^\mathrm{T}$ and $\boldsymbol{\gamma}_{i0} = (1, \gamma_{i0})^\mathrm{T}$ so that $\boldsymbol{Z}_{-1,i} = 1$ and $q_i = Z - \gamma_{i0}$, it can be seen that $D(\boldsymbol{v})$ coincides with the compound Poisson process established in [6]. Theorem 3.3 also extends the result of [35] to accommodate the temporal-dependent data and multiple splitting hyperplanes. The analysis is technically more involved than the existing literature of the fixed effect threshold regression due to the challenge of the multivariate boundaries and the dependence of the observations. To tackle these challenges, we exploit large sample theory for the extreme values and point processes ([25] and [28]), as well as the epi-convergence in distribution ([21]), which is more general than the classic uniform convergence in distribution and allows for more general discontinuity, as outlined in the SM ([33]). The techniques used in the proof may be used to analyze the asymptotic of other extreme type statistics that can be expressed as some functional of a multivariate point process with temporal-dependent sequences.

REMARK 3.2. The asymptotic independence of $T(\widehat{\boldsymbol{\gamma}}_1^c - \boldsymbol{\gamma}_{10})$ and $T(\widehat{\boldsymbol{\gamma}}_2^c - \boldsymbol{\gamma}_{20})$ was shown for the univaraite multiple-regime threshold model ([23]). Theorem 3.3 reveals that this can be extended to multiple splitting hyperplanes, provided that the probability of data locating at the crossing region of the two hyperplanes is negligible as reflected in Assumption 4 (iii). As shown in the proof, the empirical point process induced by $\{(\xi_t^{(k,h)}, \boldsymbol{Z}_{-1,i,t}, Tq_{i,t}), i = 1, 2, (k,h) \in \mathcal{S}(i)\}_{t=1}^T$ is asymptotic Poisson, whose arrivals can be divided into different segments, depending on whether they belong to the same pair $(k,h) \in \mathcal{S}(i)$ or not, where $\mathcal{S}(i)$ is the set of index pairs of adjacent regions split by the i-th hyperplane. Hence, the limiting Poisson process can be thinned into several asymptotic independent child processes, which further implies the asymptotic independence of $T(\widehat{\boldsymbol{\gamma}}_1^c - \boldsymbol{\gamma}_{10})$ and $T(\widehat{\boldsymbol{\gamma}}_2^c - \boldsymbol{\gamma}_{20})$. As a building block, we established a thinning theorem for Poisson processes for the α-mixing sequences, which might be useful in its own right. The asymptotic independence of $\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k0})$ and $T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0)$ can be explained by the fact that the former is asymptotically a sum of terms with each term being asymptotically negligible. Hence $\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k0})$ should not depend on the stochastically bounded number of points near the hyperplanes that determine the distribution of $T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0)$ ([18]).

It is also noted that the temporal dependence structure of the observed time series does not show up in the asymptotic distributions of $T(\widehat{\gamma}^c - \gamma_0)$ and $\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k0})$. That regarding $\sqrt{T}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k0})$ is due to the martingale difference condition $\mathbb{E}(\varepsilon_t|\mathcal{F}_{t-1}) = 0$ as far as the asymptotic variance of $\widehat{\boldsymbol{\beta}}_k$ is concerned, which is commonly the case in other related studies [6, 23]. That on the $T(\widehat{\gamma}^c - \gamma_0)$ is because the asymptotic distribution of $\widehat{\gamma}^c$ is determined by the empirical point process induced by the points near the underlying splitting hyperplanes, which satisfies Meyer's condition ([25]) for rare events of mixing sequences and ensures the limiting process being Poisson as in the case of independent observations.

4. Computation. The computation of the LSE for $\widehat{\boldsymbol{\theta}}$ by minimizing (3.2) is quite challenging due to the non-regularity of $m(\boldsymbol{W}_t, \boldsymbol{\theta})$ that makes the most commonly used optimization algorithms unworkable. We overcome the difficulty via the mixed integer quadratic programming (MIQP), which optimizes a quadratic objective function with linear constraints over points in polyhedral sets whose components can be both integer and continuous variables; see [4] and [3] for details. For the two-regime regression, [22] expressed the LS problem as an MIQP problem to improve the computation efficiency. The inclusion of the second boundary in the current study brings challenges. If formulated directly using the approach of [22], it would make the objective function quartic rather than quadratic. We will formulate a MIQP for the two-boundary problem to facilitate the computation.

To make the notations compact, we define $I_{k,t} = \mathbb{1}\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma})\}$ for any candidate $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^{\mathrm{T}}, \boldsymbol{\gamma}_2^{\mathrm{T}})^{\mathrm{T}}$ and $k = 1, \cdots, 4$. Let $X_{t,i}$ be the i-th element of $\boldsymbol{X}_t$ and $\beta_{k,i}$ be the i-th element of $\boldsymbol{\beta}_k$. It can be noted that the irregularity of $\mathbb{M}_T(\boldsymbol{\theta})$ in (3.2) is brought by the indicators $\{I_{k,t}\}$. If we define $\ell_{k,i,t} = I_{k,t}\beta_{k,i}$ for $i = 1, \cdots, p$, then $\mathbb{M}_T(\boldsymbol{\theta})$ can be expressed as

$$\mathbb{V}_T(\boldsymbol{\ell}) = \frac{1}{T}\sum_{t=1}^{T}\left(Y_t - \sum_{k=1}^{4}\sum_{i=1}^{p} X_{t,i}\ell_{k,i,t}\right)^2 \tag{4.1}$$

which is quadratic with respect to $\boldsymbol{\ell} = \{\ell_{k,i,t} : k = 1, \cdots, 4;\ i = 1, \cdots, p;\ t = 1, \cdots, T\}$.

Since the constraints of an MIQP have to be linear, while $\ell_{k,i,t} = I_{k,t}\beta_{k,i}$ is non-linear, it is necessary to introduce linear constraints to ensure that $\{\ell_{k,i,t}\}$ have a one-to-one correspondence to the unknown parameters $\{\boldsymbol{\beta}_k\}_{k=1}^4$ and$\{\gamma_j\}_{j=1}^2$. As $\boldsymbol{\beta}_k$ belongs to a compact set, there exist constants L_i and U_i such that $L_i \le \beta_{k,i} \le U_i$. By imposing constraints

$$I_{k,t}L_i \le \ell_{k,i,t} \le I_{k,t}U_i \quad \text{and} \quad L_i(1 - I_{k,t}) \le \beta_{k,i} - \ell_{k,i,t} \le U_i(1 - I_{k,t}), \tag{4.2}$$

it can be verified that (4.2) holds if and only if $\ell_{k,i,t} = I_{k,t}\beta_{k,i}$ under the condition that $I_{k,t} \in \{0, 1\}$. That $\ell_{k,i,t} = I_{k,t}\beta_{k,i}$ implies (4.2) is obvious. To appreciate the other way, note that if $I_{k,t} = 1$, $\ell_{k,i,t} = \beta_{k,i}$; otherwise if $I_{k,t} = 0$, $\ell_{k,i,t} = 0$. In either cases, $\ell_{k,i,t} = I_{k,t}\beta_{k,i}$.

The next goal is to relate $I_{k,t} = \mathbb{1}\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma})\}$ to the boundary coefficients $\{\gamma_j\}_{j=1}^2$. Let $g_{j,t} = \mathbb{1}(\boldsymbol{Z}_{j,t}^{\mathrm{T}}\gamma_j > 0)$. We first express $g_{j,t}$ by linear constraints in γ_j, so as to link $I_{k,t}$ with $g_{j,t}$ via linear inequalities. Let $M_{j,t} = \max_{\gamma \in \Gamma_j} |\boldsymbol{Z}_{j,t}^{\mathrm{T}}\gamma|$ which can be readily computed via linear programming. Then,

$$(g_{j,t} - 1)(M_{j,t} + \epsilon) < \boldsymbol{Z}_{j,t}^{\mathrm{T}}\gamma_j \le g_{j,t}M_{j,t} \tag{4.3}$$

hold by the definition of $g_{j,t}$, where $\epsilon > 0$ is a small predetermined constant. On the other hand, let $g_{j,t}$ be a binary variable that satisfies (4.3). Then, $g_{j,t} = 1$ and the first inequality implies that $Z_{j,t}^{\mathrm{T}}\gamma_k > 0$; and $g_{j,t} = 0$ and the second inequality implies that $Z_{j,t}^{\mathrm{T}}\gamma \le 0$. Thus, (4.3) are equivalent to $g_{j,t} = \mathbb{1}(\boldsymbol{Z}_{j,t}^{\mathrm{T}}\gamma_j > 0)$.

Finally, we construct constraints which are linear in $\{g_{j,t}\}_{j=1}^2$ and equivalent to $I_{k,t} = \mathbb{1}\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma})\}$. Since each regime $R_k(\boldsymbol{\gamma})$ can be written as $R_k(\boldsymbol{\gamma}) = \{(\boldsymbol{z}_1, \boldsymbol{z}_2) : s_j^{(k)}\boldsymbol{z}_j^{\mathrm{T}}\gamma_j >$

$0,\ j=1,2\}$, where $s_j^{(k)}\in\{-1,1\}$ is the sign of $\boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{\gamma}_j$ for the points belonging in $R_k(\boldsymbol{\gamma})$, we can write $I_{k,t}=\prod_{j=1}^{2}\mathbb{1}(s_j^{(k)}\boldsymbol{Z}_{j,t}^{\mathrm{T}}\boldsymbol{\gamma}_j>0)$, which can be linked to $\{g_{j,t}\}_{j=1}^{2}$ via

$$I_{k,t}=\prod_{j=1}^{2}\mathbb{1}\left(s_j^{(k)}\boldsymbol{Z}_{j,t}^{\mathrm{T}}\boldsymbol{\gamma}_j>0\right)=\prod_{j=1}^{2}\left\{s_j^{(k)}g_{j,t}+(1-s_j^{(k)})/2\right\}, \tag{4.4}$$

where the first equality is by the definition of $I_{k,t}$, and the second equality can be directly verified. Since the right-hand side of (4.4) is a product of two factors taking values in $\{0,1\}$, it can be shown that (4.4) is equivalent to the following linear constraints

$$I_{k,t}\geq\sum_{j=1}^{2}\left\{s_j^{(k)}g_{j,t}+(1-s_j^{(k)})/2\right\}-1 \text{ and } I_{k,t}\leq s_j^{(k)}g_{j,t}+(1-s_j^{(k)})/2. \tag{4.5}$$

for $j=1$ and 2 and $k\in\{1,\cdots,4\}$.

In summary, via the linear constraints (4.2), (4.3) and (4.5), we transform the original LS problem (2.2) to a MIQP problem formulated as following.

Let $\boldsymbol{g}=\{g_{j,t}:j=1,2,t=1,\cdots,T\}$, $\boldsymbol{\mathcal{I}}=\{I_{k,t}:k=1,\cdots,4,t=1,\cdots,T\}$ and $\boldsymbol{\ell}=\{\ell_{k,i,t}:k=1,\cdots,4,i=1,\cdots,p,t=1,\cdots,T\}$. Solve the following problem:

$$\min_{\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{g},\boldsymbol{\mathcal{I}},\boldsymbol{\ell}}\frac{1}{T}\sum_{t=1}^{T}\left(Y_t-\sum_{k=1}^{4}\sum_{i=1}^{p}X_{t,i}\ell_{k,i,t}\right)^2 \tag{4.6}$$

(4.7)

$$\text{subject to}\begin{cases}\boldsymbol{\beta}_k\in\mathcal{B},\ \ \boldsymbol{\gamma}_j\in\Gamma_j,\ \ g_{j,t}\in\{0,1\},\ \ I_{k,t}\in\{0,1\},\ \ L_i\leq\beta_{k,i}\leq U_i,\\ (g_{j,t}-1)(M_{j,t}+\epsilon)<\boldsymbol{Z}_{j,t}^{\mathrm{T}}\boldsymbol{\gamma}_j\leq g_{j,t}M_{j,t},\ I_{k,t}L_i\leq\ell_{k,i,t}\leq I_{k,t}U_i,\\ L_i(1-I_{k,t})\leq\beta_{k,i}-\ell_{k,i,t}\leq U_i(1-I_{k,t}),\\ I_{k,t}\leq s_j^{(k)}g_{j,t}+(1-s_j^{(k)})/2,\ \ I_{k,t}\geq\displaystyle\sum_{j=1}^{2}\left\{s_j^{(k)}g_{j,t}+(1-s_j^{(k)})/2\right\}-1,\end{cases}$$

for $k=1,\cdots,4,j=1,2,i=1,\cdots,p$ and $t=1,\cdots,T$.

The above optimization problem can be solved quite efficiently with modern mixed integer optimization softwares such as GUROBI and CPLEX. The next theorem, whose proof is in Section C of the SM ([33]), shows that the formulated MIQP is equivalent to the original LS problem.

THEOREM 4.1. *For any small $\epsilon>0$ in (4.7), let $\tilde{\boldsymbol{\theta}}=(\tilde{\gamma}^{\mathrm{T}},\tilde{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ be a solution of the MIQP defined with (4.6) and (4.7), then $\mathbb{M}_T(\widehat{\boldsymbol{\theta}})=\mathbb{M}_T(\tilde{\boldsymbol{\theta}})$ where $\widehat{\boldsymbol{\theta}}$ is a solution in (3.2).*

Theorem 4.1 indicates that any $\widetilde{\boldsymbol{\gamma}}$ satisfying (4.6) and (4.7) is an element of $\widehat{\mathcal{G}}$, the solution set for the LS estimators for $\boldsymbol{\gamma}_0$. Since for any $\{g_{j,t}\}\in\{0,1\}^{2T}$, there are infinitely many $\boldsymbol{\gamma}_j$ $(j=1,2)$ that satisfy the constraint in the second line of (4.7), we can output multiple solutions $\{\widetilde{\boldsymbol{\gamma}}_n=(\widetilde{\boldsymbol{\gamma}}_{n1}^{\mathrm{T}},\widetilde{\boldsymbol{\gamma}}_{n2}^{\mathrm{T}})^{\mathrm{T}}\}_{n=1}^{N}$ of the above MIQP for a sufficiently large N, and use their average as an approximation for the centroid $\widehat{\boldsymbol{\gamma}}^c$ of the set $\widehat{\mathcal{G}}$ as advocated in [35]. We display a scatter plot of the multiple solutions from a simulation experiment reported in Section H.2 of the SM ([33]), which appeared to be uniformly distributed. However, it requires further investigation to understand the detailed mechanism regarding how the multiple elements of $\widehat{\mathcal{G}}$ are produced by the MIQP solver.

REMARK 4.1. It is noted that the above algorithm requires prior specifications of (L_i, U_i), the upper and lower bound for $\beta_{k,i}$. In practice, we can first standardize $\{\boldsymbol{X}_t\}_{t=1}^T$ and specify a sufficient large parameter interval (L_i, U_i) to ensure it contains the true value. Alternatively, we can employ the data-driven method proposed in [3] that estimates $\max\{|L_i|, |U_i|\}$ via the convex quadratic optimization. Besides the proposed MIQP algorithm, the MCMC-based method as used in [35] for the two-regime regression can also be adapted to minimize the LS criterion $\mathbb{M}_T(\boldsymbol{\theta})$, which avoids the specification of the parameter bounds but requires more intensive computations since it is a simulation-based method. A comprehensive comparison between the MIQP and MCMC algorithms for segmented regressions would require more work and we leave it to further study.

REMARK 4.2. As indicated in [22], the MIQP may be slow when the dimension of $\boldsymbol{X}_t$ and the sample size T are large. As an alternative, we present a block coordinate descent (BCD) algorithm for the four-regime model in Section C of the SM ([33]), which minimizes the LS criterion with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ iteratively. At each step, the update for $\boldsymbol{\gamma}$ given $\boldsymbol{\beta}$ is via a mixed integer linear programming (MILP), which is easier to solve than the MIQP. The update for $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}$ is by linear regression in each candidate regime. Hence, the BCD is computationally more efficient than the MIQP that jointly optimizes $(\boldsymbol{\gamma}, \boldsymbol{\beta})$. However, there is no guarantee that the BCD converges to the global optimal solution without a consistent initialization. Simulations to compare the two algorithms are presented in the SM ([33]), which show that the BCD with proper initial values can produce close solutions to that of the MIQP with significantly reduced running time.

5. Smoothed regression bootstrap.

We now consider the statistical inference problems for $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$. The inference for $\boldsymbol{\beta}_0$ is quite standard due to the asymptotic normality of $\widehat{\boldsymbol{\beta}}$, while that for the boundary coefficient $\boldsymbol{\gamma}_0$ is much more challenging since the asymptotic distribution of $T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0)$ has a much-involved form and is hard to simulate.

A natural idea for the inference of $\boldsymbol{\gamma}_0$ is to employ the bootstrap. However, as shown in [31] and [34], neither the nonparametric, the residual, nor the wild bootstrap is consistent in approximating the distribution of estimator for the change points in change point models or the threshold in threshold regression models. The failure of these bootstrap methods can be explained as follows. As pointed out in Remark 3.2, only the data around the boundary hyperplanes is informative for the inference on $\boldsymbol{\gamma}_0$. Thus the bootstrap sampling distribution $\widehat{\mathbb{P}}_T$, when conditional on the original data, must approximate the true distribution $\mathbb{P}_0$ in the neighborhood of the true hyperplanes. For the identification of $\boldsymbol{\gamma}_0$, $\mathbb{P}_0$ must have a positive probability on any local region around the underlying boundaries, as reflected in Assumption 2 (ii). However, conditional on the original data, the bootstrap distribution $\widehat{\mathbb{P}}_T$ is discrete under either the nonparametric, the residual, or the wild bootstrap, which fails to mirror $\mathbb{P}_0$. As a remedy, we present a smoothed regression bootstrap method and prove its theoretical validity.

Suppose that Y is generated according to the following segmented linear regression model with heteroscedastic error

$$Y = \sum_{k=1}^{4} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0 \mathbb{1}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0)\} + \sigma_0(\boldsymbol{X}, \boldsymbol{Z})\, e, \tag{5.1}$$

where e has a continuous distribution and is independent of $(\boldsymbol{X}, \boldsymbol{Z})$ with $\mathbb{E}(e) = 0$ and $\mathbb{E}\left(e^2\right) = 1$, and $\sigma_0^2(\boldsymbol{X}, \boldsymbol{Z})$ is a conditional variance function representing possible heteroskedasticity. Model (5.1) is a refinement of Model (2.1) with more detailed structure on the residuals. If it is believed that the error is homogeneous within each region $R_k(\boldsymbol{\gamma}_0)$ so that $\varepsilon = \sigma_k \mathbb{1}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0)\}e$ for some $\sigma_k > 0$, as assumed in [34], then the nonparametric estimation for $\sigma_0(\boldsymbol{x}, \boldsymbol{z})$ is not required and σ_k can be estimated with the sample standard deviation of the fitted residuals in the k-th region.

Let $F_0(\boldsymbol{x}, \boldsymbol{z})$ be the distribution function of $(\boldsymbol{X}, \boldsymbol{Z})$, whose density function is $f_0(\boldsymbol{x}, \boldsymbol{z})$. We estimate $F_0(\boldsymbol{x}, \boldsymbol{z})$ and $\sigma_0(\boldsymbol{x}, \boldsymbol{z})$ nonparametrically with the kernel smoothing. Specifically, let $K_1(\cdot)$ and $K_2(\cdot)$ be a p-dimensional and a $(d_1 + d_2)$-dimensional kernel functions, respectively. Let $G_i(\boldsymbol{u}) = \int_{-\infty}^{\boldsymbol{u}} K_i(\boldsymbol{u}) d\boldsymbol{u}$ for $i = 1, 2$. The kernel smoothing estimator for $F_0(\boldsymbol{x}, \boldsymbol{z})$ is given by

$$\widetilde{F}_0(\boldsymbol{x}, \boldsymbol{z}) = \frac{1}{T} \sum_{t=1}^{T} G_1\left(\frac{\boldsymbol{X}_t - \boldsymbol{x}}{h_1}\right) G_2\left(\frac{\boldsymbol{Z}_t - \boldsymbol{z}}{h_2}\right),$$

where h_1 and h_2 are smoothing bandwidths.

With the LS estimator $(\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\beta}})$, the estimated residuals are $\widehat{\varepsilon}_t = Y_t - \sum_{k=1}^{4} \boldsymbol{X}^{\mathrm{T}} \widehat{\boldsymbol{\beta}} \mathbb{1}\{\boldsymbol{Z}_t \in R_k(\widehat{\boldsymbol{\gamma}})\}$. The conditional variance function $\sigma_0^2(\boldsymbol{x}, \boldsymbol{z})$ can be estimated via the local linear approach proposed by [10]. For any given $(\boldsymbol{x}, \boldsymbol{z})$, the local linear estimator $\widetilde{\sigma}^2(\boldsymbol{x}, \boldsymbol{z}) = \widehat{\alpha}$, which is defined by

$$(\widehat{\alpha}, \widehat{\boldsymbol{\eta}}) = \arg\min_{(\alpha, \boldsymbol{\eta})} \sum_{t=1}^{T} \left\{\widehat{\varepsilon}_t^2 - \alpha - \left((\boldsymbol{X}_t - \boldsymbol{x})^{\mathrm{T}}, (\boldsymbol{Z}_t - \boldsymbol{z})^{\mathrm{T}}\right) \boldsymbol{\eta}\right\}^2 K_1\left(\frac{\boldsymbol{X}_t - \boldsymbol{x}}{b_1}\right) K_2\left(\frac{\boldsymbol{Z}_t - \boldsymbol{z}}{b_2}\right),$$

where $\boldsymbol{\eta} \in \mathbb{R}^{p+d_1+d_2}$, and b_1 and b_2 are smoothing bandwidths. Let $\widehat{e}_t = \widehat{\varepsilon}_t / \widetilde{\sigma}(\boldsymbol{X}_t, \boldsymbol{Z}_t)$ and $\tilde{e}_t = \widehat{e}_t - \bar{e}_T$, where $\bar{e}_T = \sum_{t=1}^{T} \widehat{e}_t / T$. Denote $\widehat{G}(e)$ as the empirical distribution of $\{\tilde{e}_t\}_{t=1}^{T}$.

We need the following conditions on the underlying stationary distribution and its density functions, the kernel functions, and the smoothing bandwidths to facilitate the Bootstrap procedure.

ASSUMPTION 6. (i) The stationary distribution F_0 of $(\boldsymbol{X}_t, \boldsymbol{Z}_t)$ has a compact support and is absolute continuous with density $f_0(\boldsymbol{x}, \boldsymbol{z})$ which is bounded and $\inf_{\boldsymbol{x}, \boldsymbol{z}} f_0(\boldsymbol{x}, \boldsymbol{z}) > 0$.

(ii) The conditional variance function $\sigma_0^2(\boldsymbol{x}, \boldsymbol{z})$ is bounded and $\inf_{\boldsymbol{x}, \boldsymbol{z}} \sigma_0^2(\boldsymbol{x}, \boldsymbol{z}) > 0$.

(iii) The kernels $K_1(\cdot)$ and $K_2(\cdot)$ are symmetric density functions which are Lipshitz continuous and have bounded supports. The smoothing bandwidths satisfy $h_i, b_i \to 0$ for $i = 1$ and 2, and $T(\log T)^{-1} h_1^p h_2^{d_1+d_2} \to \infty$ and $T(\log T)^{-1} b_1^p b_2^{d_1+d_2} \to \infty$ as $T \to \infty$.

Under Assumptions 1 and 6, it can be shown that $\sup_{\boldsymbol{x}, \boldsymbol{z}} \|\widetilde{F}_0(\boldsymbol{x}, \boldsymbol{z}) - F_0(\boldsymbol{x}, \boldsymbol{z})\| \xrightarrow{p} 0$, and $\sup_{\boldsymbol{x}, \boldsymbol{z}} \|\widetilde{\sigma}^2(\boldsymbol{x}, \boldsymbol{z}) - \sigma_0^2(\boldsymbol{x}, \boldsymbol{z})\| \xrightarrow{p} 0$, following the uniform convergence results of kernel density and regression estimators for mixing sequences, say [14]. In addition, the above assumptions also ensure the uniform convergence of the density $\widetilde{f}_0$ of the kernel estimator $\widetilde{F}_0$ to the true density function f_0, which is required in establishing the consistency of the smoothed regression bootstrap. If $(\boldsymbol{X}, \boldsymbol{Z})$ is of high dimensions we can also employ machine learning methods that are adaptive to high dimensional features, such as the deep neural networks, to estimate $f_0(\boldsymbol{x}, \boldsymbol{z})$ and $\sigma_0(\boldsymbol{x}, \boldsymbol{z})$, as long as their uniform convergence can be guaranteed.

The bootstrap procedure to approximate the distributions of $\{T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0), \sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}$ is as follows.

Step 1: First, generate $\{(\boldsymbol{X}_t^*, \boldsymbol{Z}_t^*)\}_{t=1}^{T}$ independently from $\widetilde{F}(\boldsymbol{x}, \boldsymbol{z})$ and $\{e_t^*\}_{t=1}^{T}$ independently from $\widehat{G}(e)$, respectively. Then, generate $Y_t^* = \sum_{k=1}^{4} (\boldsymbol{X}_t^*)^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_k \mathbb{1}\{\boldsymbol{Z}_t^* \in R_k(\widehat{\boldsymbol{\gamma}}^c)\} + \widetilde{\sigma}(\boldsymbol{X}_t^*, \boldsymbol{Z}_t^*) e_t^*$ to obtain bootstrap resample $\{(Y_t^*, \boldsymbol{X}_t^*, \boldsymbol{Z}_t^*)\}_{t=1}^{T}$.

Step 2: Compute the LSEs based on $\{(Y_t^*, \boldsymbol{X}_t^*, \boldsymbol{Z}_t^*)\}_{t=1}^T$, where $\widehat{\boldsymbol{\beta}}^*$ is the LSE for $\boldsymbol{\beta}_0$ and $\{\widehat{\gamma}_i^*\}_{i=1}^N$ are the LSEs for γ_0 for a sufficiently large N. Let $\widehat{\gamma}^{*c} = \sum_{i=1}^N \widehat{\gamma}_i^*/N$.

Step 3: Repeat the above two steps B times for a large positive integer B to obtain $\{\widehat{\gamma}_b^{*c}\}_{b=1}^B$ and $\{\widehat{\boldsymbol{\beta}}_b^*\}_{b=1}^B$, and use the empirical distribution of $\left\{T(\widehat{\gamma}_b^{*c} - \widehat{\gamma}^c), \sqrt{T}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})\right\}_{b=1}^B$ as an estimate of the distribution of $\{T\left(\widehat{\gamma}^c - \gamma_0\right), \sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}$.

As in the original LS problem, the LSEs for γ_0 based on each bootstrap resample are attained on a convex set $\widehat{\mathcal{G}}^*$. Therefore, in Step 2 we approximate the centroid of $\widehat{\mathcal{G}}^*$ by the average of N elements in $\widehat{\mathcal{G}}^*$. Denote the distribution of $\{T\left(\widehat{\gamma}^c - \gamma_0\right), \sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}$ as $\mathcal{L}_T$ and the empirical distribution of $\left\{T(\widehat{\gamma}_b^{*c} - \widehat{\gamma}^c), \sqrt{T}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})\right\}_{b=1}^B$ as $\mathcal{L}_{T,B}$. The validity of the smoothed regression bootstrap is established in the following theorem.

THEOREM 5.1. *Suppose that Assumptions 1-6 hold. Then* $\rho\left(\mathcal{L}_{T,B}, \mathcal{L}_T\right) \xrightarrow{p} 0$ *as* $B, T \to \infty$*, for any metric* ρ *that metrizes weak convergence of distributions.*

The proof of the theorem is in Section D of the SM ([33]) by first establishing sufficient conditions for a consistent bootstrap scheme for approximating $\mathcal{L}_T$, followed by showing that the smoothed regression bootstrap satisfies these conditions. With the above result, confidence regions and hypothesis testings about γ_0 and $\boldsymbol{\beta}_0$ can be readily conducted via the empirical distribution of the smoothed bootstrap estimates $\mathcal{L}_{T,B}$.

REMARK 5.1. We exploit the parametric regression model in the bootstrap resampling, under which the mixing-dependent structure of the observed data does not show up in the asymptotic distributions as shown in Theorem 3.3. As discussed in [17], if one has a parametric model that reduces the data generating process to independence sampling, then the parametric bootstrap has properties that are essentially the same as they are when the observations are independently distributed. Therefore, in the resampling procedure, the temporal dependence of the original data is not necessary to be explicitly taken into account.

REMARK 5.2. In addition to the smoothed regression bootstrap, there are two alternative methods which may be applicable for inference of γ_0. One is the block subsampling method proposed by [26], which was adopted by [13] in the threshold autoregressive models. Another is the nonparametric posterior confident interval approach based on the Markov Chain Monte Carlo (MCMC) adopted by [35] for inference on the two-regime regression model. Whether these methods work for the current four-regime segmented regression with fixed boundary effects and dependent data are interesting future research topics.

6. Degenerated models and model selection.

Model (2.1) assumes that there are four segments divided by two boundary hyperplanes where the adjacent regimes have distinct regression coefficients. However, it is possible that the underlying regimes are degenerated with less than four regimes. In this section, we show that the LS estimator (3.2) attains desirable convergence properties even in the degenerated cases, and propose a model selection method for choosing the underlying model.

Given the data sample $\{(Y_t, \boldsymbol{X}_t, \boldsymbol{Z}_{1,t}, \boldsymbol{Z}_{2,t})\}_{t=1}^T$ for $\boldsymbol{Z}_{1,t} \in \mathcal{Z}_1$ and $\boldsymbol{Z}_{2,t} \in \mathcal{Z}_2$, there are five possible degenerated models as follows in addition to the four regime model (2.1).

(a.1). Three-regime model with non-intersected splitting hyperplanes:

$$Y_t = \sum_{k=1}^3 \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{\beta}_{k0} \mathbb{1}\left\{\boldsymbol{Z}_t \in R_k(\gamma_0)\right\} + \varepsilon_t, \tag{6.1}$$

where the two hyperplanes H_1 and H_2 have no intersection on $\mathcal{Z}_1 \times \mathcal{Z}_2$. Without loss of generality, we suppose that $\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\gamma}_{10} \le \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\gamma}_{20}$ for all $(\boldsymbol{z}_1, \boldsymbol{z}_2) \in (\mathcal{Z}_1 \times \mathcal{Z}_2)$. Then, $R_1(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\gamma}_{10} > 0\}, R_2(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\gamma}_{10} \le 0, \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\gamma}_{20} > 0\}$ and $R_3(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\gamma}_{20} \le 0\}$. The conventional multi-threshold models (e.g., [12] and [23]) correspond to this case.

(a.2). Three-regime regression model with intersected splitting hyperplanes:

$$Y_t = \sum_{k=1}^{3} \boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_{k0}\mathbb{1}(\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0) + \varepsilon_t, \tag{6.2}$$

where $R_1(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{j,0} > 0, \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{\gamma}_{j,0} > 0\}, R_2(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{j,0} > 0, \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{\gamma}_{j,0} \le 0\}$ and $R_3(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{\gamma}_{j,0} \le 0\}$ for $i \ne j \in \{1, 2\}$. Geometrically, one side of the hyperplane $H_j : \boldsymbol{z}_j^{\mathrm{T}}\boldsymbol{\gamma}_{j,0} = 0$ is split by $H_i : \boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\gamma}_{i,0} = 0$ that does not extend to the other side of H_j.

(b.1). Two-regime regression model with one splitting hyperplane:

$$Y_t = \sum_{k=1}^{2} \boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_{k0}\mathbb{1}\left\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\right\} + \varepsilon_t, \tag{6.3}$$

where $(\boldsymbol{z}, \boldsymbol{\gamma}_0)$ is either $(\boldsymbol{z}_1, \boldsymbol{\gamma}_{10})$ or $(\boldsymbol{z}_2, \boldsymbol{\gamma}_{20})$ and $R_1(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma}_0 > 0\}$ and $R_2(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma}_0 \le 0\}$, which are the same as the two-regime models of [22] and [35].

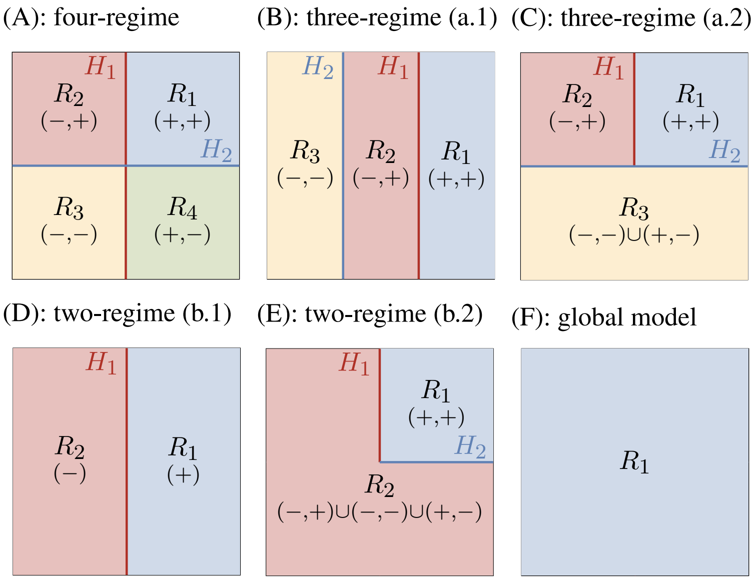
(b.2). Two-regime regression model with two splitting hyperplanes:

$$Y_t = \sum_{k=1}^{2} \boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_{k0}\mathbb{1}\left\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\right\} + \varepsilon_t, \tag{6.4}$$

where $R_1(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\gamma}_{10} > 0, \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\gamma}_{20} > 0\}$ and $R_2\{\boldsymbol{\gamma}_0\} = \mathcal{Z}_1 \times \mathcal{Z}_2 \setminus R_1(\boldsymbol{\gamma}_0)$.

(c). Global linear model:

$$Y_t = \boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_0 + \varepsilon_t, \tag{6.5}$$


(A): four-regime

(B): three-regime (a.1)

(C): three-regime (a.2)

(D): two-regime (b.1)

(E): two-regime (b.2)

(F): global model

Fig 1: Illustrations of segmented models with no more than four regimes. The signs of $(\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\gamma}_1, \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\gamma}_2)$ for each region are indicated below the region names.

Figure 1 illustrates the segmented models with no more than four regimes, which can be expressed in a unified form

$$Y_t = \sum_{k=1}^{K_0} \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{\beta}_{k0} \mathbb{1}\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\} + \varepsilon_t, \tag{6.6}$$

where the number of regimes $1 \le K_0 < 4$ and the number of splitting hyperplanes $L_0 \le 2$. In particular, $R_k(\boldsymbol{\gamma}_0) = \mathcal{Z}_1 \times \mathcal{Z}_2$ for the global linear model ($K_0 = 1$), the splitting coefficient $\boldsymbol{\gamma}_0 = \boldsymbol{\gamma}_{10}$ or $\boldsymbol{\gamma}_{20}$ when $L_0 = 1$, and $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{10}^{\mathrm{T}}, \boldsymbol{\gamma}_{20}^{\mathrm{T}})^{\mathrm{T}}$ when $L_0 = 2$.

Let $\widehat{\mathcal{B}} = \{\widehat{\boldsymbol{\beta}}_k\}_{k=1}^4$ and $\widehat{\mathcal{G}} = \{\widehat{\boldsymbol{\gamma}}_j\}_{j=1}^2$ be the LS estimators for the regression and the boundary coefficients, respectively, obtained under the four-regime regression model (3.2). To measure the estimation accuracy of the four-regime algorithms for less than four regime models, we need a distance of the true parameters of possibly degenerated models to the set of the LS estimates under the four-regime model. To this end, we define a distance between a vector $\boldsymbol{v}$ and a set of vectors $\widehat{\mathcal{V}} = \{\hat{\boldsymbol{v}}_j\}_{j=1}^J$ as $d(\boldsymbol{v}, \widehat{\mathcal{V}}) = \min_j \|\boldsymbol{v} - \hat{\boldsymbol{v}}_j\|_2$. The following theorem establishes the convergence of the LS estimators to the underlying parameters by showing that the distance of the true parameters of the degenerated models to the set of the LSEs under the four-regime model convergences to zero.

THEOREM 6.1. *For Model (6.6) with K_0 regimes and L_0 splitting hyperplanes, where $1 \le K_0 < 4$ and $0 \le L_0 \le 2$, under Assumption 1 and Assumptions S2-S4 in the SM ([33]), which adapt Assumptions 3–4 to the degenerate model settings, then for each $\boldsymbol{\beta}_{k0}$ with $1 \le k \le K_0$, $d(\boldsymbol{\beta}_{k0}, \widehat{\mathcal{B}}) = O_p(1/\sqrt{T})$. If $L_0 = 1$, then $d(\boldsymbol{\gamma}_0, \widehat{\mathcal{G}}) = O_p(1/T)$. If $L_0 = 2$, then $d(\boldsymbol{\gamma}_{i0}, \widehat{\mathcal{G}}) = O_p(1/T)$ for each $i = 1$ and 2. Moreover, for any of the degenerated models with $K_0 < 4$ regimes, there exists an index set $\mathcal{Q}_k \subset \{1, \cdots, 4\}$ such that $\mathbb{P}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0) \triangle \cup_{i \in \mathcal{Q}_k} R_i(\widehat{\boldsymbol{\gamma}})\} = O(1/T)$ for each $1 \le k \le K_0$.*

The theorem shows that under each of the degenerated models, the estimated boundaries and the regression coefficients obtained under (3.2) of the four-regime model are consistent to the true parameters in the sense of the diminishing distance between the true parameters and the sets of the estimates. A remaining issue is to identify the true number of regimes so that more precise segmented regression can be conducted. In the following, we introduce a model selection procedure to attain the purpose.

The last part of Theorem 6.1 suggests that each true regime $R_k(\boldsymbol{\gamma}_0)$ can either be consistently estimated by some $R_i(\widehat{\boldsymbol{\gamma}})$ if $|\mathcal{Q}_k| = 1$, which occurs when $R_k(\boldsymbol{\gamma}_0)$ has two boundaries, such as the first two regimes in Figure 1 (C), or there are some redundant estimated segments in $R_k(\boldsymbol{\gamma}_0)$, which happens if $R_k(\boldsymbol{\gamma}_0)$ has a single boundary while an unnecessary estimated hyperplane splits through $R_k(\boldsymbol{\gamma}_0)$. If the latter case is true, then $|\mathcal{Q}_k| > 1$ and there exist two adjacent estimated regimes $R_i(\widehat{\boldsymbol{\gamma}})$ and $R_h(\widehat{\boldsymbol{\gamma}})$ with $i, h \in \mathcal{Q}_k$, whose corresponding $\widehat{\boldsymbol{\beta}}_i$ and $\widehat{\boldsymbol{\beta}}_h$ both consistently estimate $\boldsymbol{\beta}_{k0}$. Under such a case, merging $R_i(\widehat{\boldsymbol{\gamma}})$ with $R_h(\widehat{\boldsymbol{\gamma}})$ as one regression regime will asymptotically not lead to an increased sum of squared residuals (SSR). Otherwise, if the regression models on $R_i(\widehat{\boldsymbol{\gamma}})$ and $R_h(\widehat{\boldsymbol{\gamma}})$ are distinct, then merging these two regimes will deteriorate the fitting performance. Such a property hints that the true model with $K_0 < 4$ can be selected via a backward elimination procedure.

Starting from the estimated four-regime model, we try recursively finding the best pairs of adjacent regimes to be merged, under a criterion that the merging leads to the minimal increase in the fitting errors, as defined in (6.7) below. Via conducting the optimal regime merging recursively, we obtain four candidate regression models with the number of regimes from $K = 4$ to $K = 1$. In the second step, the optimal number of regimes K is selected based on a criterion function (6.8) that combines a goodness-of-fit measure and a penalty for over-segmentation.

For the initial model with four regimes, define

$$S_T(4) = \sum_{t=1}^{T} [Y_t - \sum_{k=1}^{K} \boldsymbol{X}_t^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_k^{(4)} \mathbb{1}\{\boldsymbol{Z}_t \in \widehat{R}_k^{(4)}\}]^2$$

to be the sum of square residual (SSR) of the estimated four-regime model. For $K = 4, 3, 2$, recursively define

$$\begin{aligned} & D_T^{(K)}(i,h) \\ &= \min_{\boldsymbol{\beta} \in \mathcal{B}} \sum_{t=1}^{T} [Y_t - \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{\beta} \mathbb{1}\{\boldsymbol{Z}_t \in \widehat{R}_i^{(K)} \cup \widehat{R}_h^{(K)}\}]^2 - \sum_{t=1}^{T} [Y_t - \sum_{k=i,h} \boldsymbol{X}_t^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_k^{(K)} \mathbb{1}\{\boldsymbol{Z}_t \in \widehat{R}_k^{(K)}\}]^2 \end{aligned}$$

to be the increment in the SSR after merging $\widehat{R}_i^{(K)}$ and $\widehat{R}_h^{(K)}$. Let $\mathcal{A}_K$ be the pair of indices for the adjacent segments of $\{\widehat{R}_k^{(K)}\}$. We merge the segments $\widehat{R}_{\hat{i}}^{(K)}$ and $\widehat{R}_{\hat{h}}^{(K)}$ if

$$(\widehat{i}, \widehat{h}) = \operatorname*{arg\,min}_{(i,h) \in \mathcal{A}_K} D_T^{(K)}(i,h), \tag{6.7}$$

followed by labeling the merged region and the remaining regions as $\{\widehat{R}_k^{(K-1)}\}_{k=1}^{K-1}$, and we denote the estimated regression coefficients to these $K-1$ regimes by $\{\widehat{\boldsymbol{\beta}}_k^{(K-1)}\}_{k=1}^{K-1}$. Then, define the SSR of the $(K-1)$-segment submodel as

$$S_T(K-1) = S_T(K) + D_T^{(K)}(\widehat{i}, \widehat{h}).$$

After obtaining the $S_T(K)$ for $K = 2, 3, 4,$ we select the number of segments $\widehat{K}$ as

$$\widehat{K} = \operatorname*{arg\,min}_{1 \le K \le 4} \{\log(\frac{S_T(K)}{T}) + \frac{\lambda_T}{T} K\} \tag{6.8}$$

and output the estimated regimes and regression coefficients accordingly. The following theorem shows that the above selection algorithm has the model selection consistency.

THEOREM 6.2. *Under the assumptions of Theorem 6.1, and* $\lambda_T \to \infty, \lambda_T/T \to 0$ *as* $T \to \infty$, *then* $\widehat{K}$ *selected in* (6.8) *satisfies* $\mathbb{P}(\widehat{K} = K_0) \to 1$ *as* $T \to \infty$. *In addition,* $\mathbb{P}\{\widehat{R}_k^{(\widehat{K})} \triangle R_k(\boldsymbol{\gamma}_0)\} = O(1/T)$ *and* $\|\widehat{\boldsymbol{\beta}}_k^{(\widehat{K})} - \boldsymbol{\beta}_{k0}\| = O_p(1/\sqrt{T})$ *for any* $k \in \{1, \cdots, K_0\}$.

Theorem 6.2 indicates that with the probability approaching 1, the selected number of regimes $\widehat{K}$ coincides with the true number K_0, and as a by-product, the corresponding estimated regimes and the regression coefficients converge to their underlying counterparts. If the regularization parameter is chosen as $\lambda_T = \log T$, the (6.8) corresponds to the Bayesian information criterion (BIC) [30].

REMARK 6.1. There are two existing approaches for carrying out the model selection for the segmented models. One is by conducting pairwise linearity tests. Specifically, for each adjacent regimes $R_i(\widehat{\boldsymbol{\gamma}})$ and $R_h(\widehat{\boldsymbol{\gamma}})$ under the four-regime model, one can test for the hypothesis $H_0 : \boldsymbol{\beta}_{i0} = \boldsymbol{\beta}_{h0}$ via two-regime linearity tests, such as the score-type test of [35]. However, implementing such tests are computationally demanding, as the test statistics have to be formulated via supremum or averaging over $\boldsymbol{\gamma} \in \Gamma$, as $\boldsymbol{\gamma}$ is not identifiable under the null hypothesis of no splitting within $R_i(\widehat{\boldsymbol{\gamma}}) \cup R_h(\widehat{\boldsymbol{\gamma}})$, which is known as the Davis problem [9].

The other is the forward sequential fitting procedure for model selection of multi-threshold regression models [12], which requires optimization for the splitting (boundary) coefficients in each step. Compared with these two methods, the proposed model selection method has two advantages. One is that it has quite readily computation without having to do the bootstrap for the model selection; and the other is that we only need to estimate the splitting coefficients for the initial four-segment model once and for all, as the submodels with fewer regimes are selected via (6.7) without the need to conduct non-convex optimization as in the forward sequential fitting procedure.

7. Simulation Study.

In this section, we present results from simulation experiments designed to investigate the performance of the proposed estimation and inference procedures for the four-regime and the degenerated less than four regime models.

7.1. *Estimation under the four-regime model.*

We first conducted simulations under the four-regime model (2.1) such that the sample was generated according to

$$Y_t = \sum_{k=1}^{4} \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{\beta}_{k0} \mathbb{1}_k(\boldsymbol{Z}_{1,t}^{\mathrm{T}}\boldsymbol{\gamma}_{10}, \boldsymbol{Z}_{2,t}^{\mathrm{T}}\boldsymbol{\gamma}_{20}) + \varepsilon_t, \quad t = 1, \cdots, T, \tag{7.1}$$

where $\boldsymbol{X}_t = (\tilde{\boldsymbol{X}}_t^{\mathrm{T}}, 1)^{\mathrm{T}}$ with $\tilde{\boldsymbol{X}}_t = (X_{1,t}, X_{2,t}, X_{3,t})^{\mathrm{T}}$ and $\boldsymbol{Z}_{j,t} = (\tilde{\boldsymbol{Z}}_{j,t}^{\mathrm{T}}, 1)^{\mathrm{T}}$ with $\tilde{\boldsymbol{Z}}_{j,t} = (Z_{j,1,t}, Z_{j,2,t})^{\mathrm{T}}$ for $j = 1, 2$. The noises were generated as $\varepsilon_t = \sigma(\boldsymbol{X}_t, \boldsymbol{Z}_t)e_t$ with $\sigma(\boldsymbol{X}_t, \boldsymbol{Z}_t) = 1 + 0.1X_{1,t}^2 + 0.1Z_{1,1,t}^2$ and $\{e_t\}_{t=1}^T$ being generated independently from the standard normal distribution and independent of $\{\boldsymbol{X}_t, \boldsymbol{Z}_t\}_{t=1}^T$. The regression coefficients of the four regimes were $\boldsymbol{\beta}_{10} = (1, 1, 1, 1)^{\mathrm{T}}, \boldsymbol{\beta}_{20} = (-3, -2, -1, 0), \boldsymbol{\beta}_{30} = (0, 1, 3, -1)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{40} = (2, -1, 0, 2)^{\mathrm{T}}$, and the two boundary coefficients $\boldsymbol{\gamma}_{10} = (1, -1, 0)^{\mathrm{T}}$ and $\boldsymbol{\gamma}_{20} = (1, 1, 0)^{\mathrm{T}}$, respectively.

We considered three settings for $\boldsymbol{X}_t$ and $\boldsymbol{Z}_{j,t}$: independence, dependence with autoregressive (AR) and moving average (MA) models, respectively. Let $\boldsymbol{V}_t = (\tilde{\boldsymbol{X}}_t^{\mathrm{T}}, \tilde{\boldsymbol{Z}}_{1,t}^{\mathrm{T}}, \tilde{\boldsymbol{Z}}_{2,t}^{\mathrm{T}})^{\mathrm{T}}$. For the independence setting, we generated $\{\boldsymbol{V}_t\}_{t=1}^T \overset{\text{i.i.d.}}{\sim} \boldsymbol{N}(\boldsymbol{0}, \Sigma_V)$, where $\Sigma_V = (\sigma_{ij})_{i,j=1,\cdots,7}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.1$ if $i \neq j$. For the AR dependence, $\boldsymbol{V}_t = \psi \boldsymbol{V}_{t-1} + \boldsymbol{u}_t$, where $\{\boldsymbol{u}_t\}_{t=1}^T \overset{\text{i.i.d.}}{\sim} \boldsymbol{N}(\boldsymbol{0}, \Sigma_V)$ and the dependence level $\psi \in \{0.2, 0.4, 0.8\}$. For the MA scenario, we generated $\boldsymbol{V}_t = \psi \boldsymbol{u}_{t-1} + \boldsymbol{u}_t$, where $\{\boldsymbol{u}_t\}_{t=1}^T \overset{\text{i.i.d.}}{\sim} \boldsymbol{N}(\boldsymbol{0}, \Sigma_V)$ and ψ took values in $\{0.2, 0.4, 0.8\}$, respectively. The simulation experimented with four sample sizes: $\{200, 400, 800, 1600\}$, and the experiments were repeated 500 times for each sample size and dependence setting.

Table 1 reports the average L_2 estimation errors under the three temporal settings (independence, AR(1) and MA(1)) and different dependence levels ($\psi = 0.2, 0.4, 0.8$) for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. It suggests that under the three dependence settings the estimation errors of $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\beta}}$ both decreased as the sample size T was increased, indicating the convergence of the estimation in both the regression and the splitting boundary coefficients. The table also suggests that the magnitudes of the estimation errors were comparable across the three temporal settings with different dependence levels, which support the result of Theorem 3.3 that the temporal dependence in $\{\boldsymbol{X}_t, \boldsymbol{Z}_{1,t}, \boldsymbol{Z}_{2,t}\}_t^T$ does not have leading order effects on the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. Moreover, Table 1 shows that the simulated averages of $\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2$ were approximately halved once the sample size was doubled, while the reduction in $\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\|_2$ was much slower, confirming the faster convergence rates of $\widehat{\gamma}$.

TABLE 1

Empirical average estimation errors $\|\boldsymbol{\gamma}_0 - \widehat{\boldsymbol{\gamma}}\|_2$ and $\|\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}\|_2$ (multiplied by 10), under the independence (IND), auto-regressive (AR) and moving average (MA) settings with different dependence level ψ for $\{\boldsymbol{X}_t, \boldsymbol{Z}_{1,t}, \boldsymbol{Z}_{2,t}\}_{t=1}^{T}$. The numbers inside the parentheses are the standard errors of the simulated averages.

T	IND		AR						MA					
	$\psi = 0$		$\psi = 0.2$		$\psi = 0.4$		$\psi = 0.8$		$\psi = 0.2$		$\psi = 0.4$		$\psi = 0.8$	
	$\widehat{\boldsymbol{\gamma}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{\boldsymbol{\beta}}$	$\widehat{\boldsymbol{\gamma}}$	$\widehat{\boldsymbol{\beta}}$
200	0.94 (0.59)	6.68 (1.7)	0.92 (0.58)	6.66 (1.68)	0.88 (0.6)	6.43 (1.56)	0.88 (0.61)	5.9 (2.24)	0.93 (0.56)	6.63 (1.63)	0.9 (0.54)	6.49 (1.66)	0.85 (0.52)	6.14 (1.8)
400	0.45 (0.3)	4.55 (1.1)	0.45 (0.3)	4.55 (1.11)	0.45 (0.27)	4.4 (1.17)	0.43 (0.29)	3.98 (1.53)	0.44 (0.28)	4.46 (1)	0.43 (0.33)	4.38 (1.07)	0.43 (0.28)	4.06 (1.21)
800	0.25 (0.16)	3.11 (0.66)	0.24 (0.15)	3.09 (0.66)	0.22 (0.14)	2.97 (0.66)	0.22 (0.14)	2.64 (0.96)	0.23 (0.14)	3.11 (0.66)	0.25 (0.16)	3.03 (0.65)	0.22 (0.15)	2.81 (0.72)
1600	0.11 (0.07)	2.2 (0.46)	0.11 (0.07)	2.18 (0.47)	0.12 (0.08)	2.11 (0.5)	0.11 (0.07)	1.88 (0.77)	0.11 (0.07)	2.17 (0.45)	0.11 (0.07)	2.11 (0.47)	0.11 (0.07)	1.97 (0.54)

7.2. *Estimation under models with less than four regimes.* We next investigated the performances of the proposed estimation based on the four-regime model when the underlying model was degenerated with less than four regimes. The data generating process for $\{\boldsymbol{X}_t, \boldsymbol{Z}_{1,t}, \boldsymbol{Z}_{2,t}, \varepsilon_t\}_{t=1}^{T}$ was largely the independence setting used in Section 7.1. For the three-regime model (6.1) with non-intersected splitting hyperplanes, we let $\boldsymbol{\gamma}_{10} = (1, 0, -1)^{\mathrm{T}}, \boldsymbol{\gamma}_{20} = (1, 0, 1)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{10} = (1, 1, 1, 1)^{\mathrm{T}}, \boldsymbol{\beta}_{20} = (-3, -2, -1, 0)^{\mathrm{T}}, \boldsymbol{\beta}_{30} = (0, 1, 3, -1)^{\mathrm{T}}$. For the three-regime model (6.2) with intersected splitting hyperplanes, we let $\boldsymbol{\gamma}_{10} = (1, 1, 0)^{\mathrm{T}}, \boldsymbol{\gamma}_{20} = (1, -1, 0)^{\mathrm{T}}$ while H_{10} does not extend to the positive side of H_{20}, and $\{\boldsymbol{\beta}_{k0}\}_{k=1}^{3}$ were the same as above. The parameters for the two-regime model (6.3) with one splitting hyperplane were set as $\boldsymbol{\gamma}_0 = (1, 1, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_{10} = (1, 1, 1, 1)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{20} = (-3, -2, -1, 0)^{\mathrm{T}}$. For the two-regime model (6.4) with two splitting hyperplanes, we set the splitting coefficients as the same as the four-regime model (7.1), and $R_1(\boldsymbol{\gamma}_0) = \{\boldsymbol{z} : \boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\gamma}_{10} > 0, \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\gamma}_{20} > 0\}$ and $R_2(\boldsymbol{\gamma}_0) = \mathcal{Z}_1 \times \mathcal{Z}_2 \setminus R_1(\boldsymbol{\gamma}_0)$, where the regression coefficients are $\boldsymbol{\beta}_{10} = (1, 1, 1, 1)^{\mathrm{T}}$ and $\boldsymbol{\beta}_{20} = (-3, -2, -1, 0)^{\mathrm{T}}$, respectively. Finally, the regression coefficients for the global linear model (6.5) were $\boldsymbol{\beta}_0 = (1, 1, 1, 1)^{\mathrm{T}}$.

The simulation results are reported in Tables S2 of Section H.2 in the SM ([33]). They show that for all the models with less than four regimes, the empirical averages of $\sum_i d(\boldsymbol{\gamma}_{i0}, \widehat{\mathcal{G}})$ and $\sum_k d(\boldsymbol{\beta}_{k0}, \widehat{\mathcal{B}})$ all diminished to 0 at similar rates as those in Table 1, where $\widehat{\mathcal{G}}$ and $\widehat{\mathcal{B}}$ are the sets of estimators obtained under the four-regime model for the splitting and regression coefficients, respectively. These confirmed the results in Theorem 6.1. In addition, to evaluate the cost of not knowing the number of the underlying regimes, we also estimated $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_0$ in the oracle setting, in which the true model forms were known. It was found that estimation errors of $\boldsymbol{\gamma}_0$ under the four-regime model fitting were about the same as that obtained under the oracle models, which was because the four-regime estimator can efficiently use the data points located near the underlying boundaries as the oracle estimators did. Moreover, as shown in Figures S2 and S3 of the SM ([33]), if the estimated four-regime model produced redundant segments within a true regime, then the discrepancy between the estimated regression coefficients on these redundant segments converged to 0, which verified the idea used in the optimal merger strategy for the backward elimination procedure in the model selection.

7.3. *Model selection.* We then conducted simulation experiments to examine the performance of the proposed model selection method in Section 6. We considered the true number of regimes ranging from $K_0 = 4$ to $K_0 = 1$, where the parameters for the model with $K_0 = 4$ were the same as Model (7.1) and those for $K_0 = 3$ and $K_0 = 2$ were Model 6.1 and Model 6.3, respectively, in Section 7.2. More simulation results for Model (6.2) and Model (6.5) ($K_0 = 1$) were reported in Table S3 of the SM.

TABLE 2

Empirical model selection results under 500 *replications. The performances were evaluated by the average estimated number of regimes* $\widehat{K}$*, the discrepancy between the true regimes and the estimated regimes* $D(\mathcal{R}, \widehat{\mathcal{R}})$ *and the* L_2 *estimation error of regression coefficients* $D(\mathcal{B}, \widehat{\mathcal{B}})$*. The penalty parameter* λ_T *was chosen in* $\{5, 5\log(T), 5\log^2(T)\}$*. The numbers inside the parentheses are the standard errors of the simulated averages.*

Model	T	$\lambda_T = 5$			$\lambda_T = 5\log(T)$			$\lambda_T = 5\log^2(T)$		
		$\widehat{K}$	$D(\mathcal{R}, \widehat{\mathcal{R}})$	$D(\mathcal{B}, \widehat{\mathcal{B}})$	$\widehat{K}$	$D(\mathcal{R}, \widehat{\mathcal{R}})$	$D(\mathcal{B}, \widehat{\mathcal{B}})$	$\widehat{K}$	$D(\mathcal{R}, \widehat{\mathcal{R}})$	$D(\mathcal{B}, \widehat{\mathcal{B}})$
Model (2.1) ($K_0 = 4$)	200	4.00 (0.00)	0.03 (0.02)	0.61 (0.12)	3.99 (0.08)	0.03 (0.04)	0.62 (0.16)	2.78 (0.87)	0.87 (0.91)	2.24 (1.05)
	400	4.00 (0.00)	0.01 (0.01)	0.41 (0.08)	4.00 (0.00)	0.01 (0.01)	0.41 (0.08)	3.92 (0.27)	0.05 (0.13)	0.53 (0.43)
	800	4.00 (0.00)	0.01 (0.00)	0.29 (0.05)	4.00 (0.00)	0.01 (0.00)	0.29 (0.05)	4.00 (0.00)	0.01 (0.00)	0.29 (0.05)
	1600	4.00 (0.00)	0.00 (0.00)	0.20 (0.04)	4.00 (0.00)	0.00 (0.00)	0.20 (0.04)	4.00 (0.00)	0.00 (0.00)	0.20 (0.04)
Model (6.1) ($K_0 = 3$)	200	3.44 (0.50)	0.12 (0.11)	0.50 (0.11)	3.00 (0.00)	0.02 (0.02)	0.48 (0.11)	2.85 (0.38)	0.13 (0.30)	0.75 (0.69)
	400	3.39 (0.49)	0.10 (0.11)	0.34 (0.07)	3.00 (0.00)	0.01 (0.01)	0.33 (0.07)	3.00 (0.00)	0.01 (0.01)	0.33 (0.07)
	800	3.33 (0.47)	0.08 (0.11)	0.23 (0.05)	3.00 (0.00)	0.01 (0.00)	0.22 (0.05)	3.00 (0.00)	0.01 (0.00)	0.22 (0.05)
	1600	3.33 (0.47)	0.08 (0.11)	0.16 (0.03)	3.00 (0.00)	0.00 (0.00)	0.16 (0.03)	3.00 (0.00)	0.00 (0.00)	0.16 (0.03)
Model (6.3) ($K_0 = 2$)	200	3.38 (0.59)	0.14 (0.11)	0.35 (0.10)	2.03 (0.17)	0.01 (0.01)	0.30 (0.08)	2.00 (0.00)	0.01 (0.01)	0.30 (0.08)
	400	3.54 (0.51)	0.13 (0.11)	0.24 (0.07)	2.01 (0.08)	0.01 (0.01)	0.20 (0.05)	2.00 (0.00)	0.01 (0.00)	0.20 (0.05)
	800	3.53 (0.53)	0.12 (0.11)	0.16 (0.04)	2.00 (0.06)	0.00 (0.00)	0.14 (0.04)	2.00 (0.00)	0.00 (0.00)	0.14 (0.04)
	1600	3.50 (0.55)	0.13 (0.12)	0.12 (0.03)	2.00 (0.00)	0.00 (0.00)	0.10 (0.03)	2.00 (0.00)	0.00 (0.00)	0.10 (0.03)

Table 2 reports three model selection performance measures for the simulation, namely (i) the estimated number of regimes $\widehat{K}$, (ii) the discrepancy between the true regimes and the estimated regimes measured by

$$D(\mathcal{R}, \widehat{\mathcal{R}}) = \sum_{k=1}^{K_0} \min_{1 \le h \le \widehat{K}} \left\{ T^{-1} \sum_{t=1}^{T} \left| \mathbb{1}\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\} - \mathbb{1}\{\boldsymbol{Z}_t \in R_h(\widehat{\boldsymbol{\gamma}})\} \right| \right\},$$

where $\mathcal{R} = \{R_k(\boldsymbol{\gamma})\}_{k=1}^{K_0}$ and $\widehat{\mathcal{R}} = \{R_k(\widehat{\boldsymbol{\gamma}})\}_{k=1}^{\widehat{K}}$, and (iii) the L_2 estimation error of regression coefficients, quantified by $D(\mathcal{B}, \widehat{\mathcal{B}}) = \sum_{k=1}^{K_0} \min_{1 \le h \le \widehat{K}} \|\boldsymbol{\beta}_{k0} - \widehat{\boldsymbol{\beta}}_h\|$. To evaluate the impact of the penalty parameter λ_T in (6.8), we presented the results under three different choices: $\lambda_T = 5, 5\log(T)$ and $5\log^2(T)$.

Table 2 shows that, for the constant penalty $\lambda_T = 5$, although the estimated number of regimes $\widehat{K}$ was consistent under $K_0 = 4$, it tended to select overly segmented models when $K_0 < 4$. Both $\lambda_T = 5\log(T)$ and $5\log^2(T)$ led to consistent estimated $\widehat{K}$ for all models, which confirmed the assertion in Theorem 6.2 that λ_T satisfying $\lambda_T \to \infty$ and $\lambda_T/T \to 0$ leads to model selection consistency. It was also noted that while the last two penalties were consistent, for smaller sample sizes, the selection performance with $\lambda_T = 5\log(T)$ was superior to that with $\lambda_T = 5\log^2(T)$ when $K_0 \geq 3$, while the latter penalty had better selection accuracy when $K \leq 2$. Such a phenomenon may be understood since a larger penalty tends to encourage under-segmentations. In addition, both $D(\mathcal{R}, \widehat{\mathcal{R}})$ and $D(\mathcal{B}, \widehat{\mathcal{B}})$ diminished to 0 when $\widehat{K}$ was correctly selected, indicating that the model specification procedure was able to not only consistently identify K_0, but also led to consistent estimates of regimes and the corresponding regression coefficients, as shown in Theorem 6.2.

7.4. *Smoothed regression bootstrap.* We now report simulation results designed to evaluate the empirical performance of the smoothed regression bootstrap.

The data generating model for $\{Y_t, \boldsymbol{X}_t, \boldsymbol{Z}_{1,t}, \boldsymbol{Z}_{2,t}\}_{t=1}^{T}$ was the same as the independent setting in Section 7.1, but $(\tilde{\boldsymbol{X}}_t^{\mathrm{T}}, \tilde{\boldsymbol{Z}}_{1,t}^{\mathrm{T}}, \tilde{\boldsymbol{Z}}_{2,t}^{\mathrm{T}})^{\mathrm{T}}$ was truncated over a 7-dimensional region $[-2,2]^7$ to ensure the distribution of the covariates was compactly supported as required in Assumption 6. The product Gaussian kernel was used as the kernel function with the smoothing bandwidths h_i and $b_i (i = 1, 2)$ for $\tilde{F}_0(\boldsymbol{x}, \boldsymbol{z})$ and $\tilde{\sigma}^2(\boldsymbol{x}, \boldsymbol{z})$ were chosen by the cross-validation method ([10]). As a comparison, we also conducted the wild bootstrap procedure ([24]), which is a commonly used bootstrap method in regression. Different from the smoothed regression bootstrap, the wild bootstrap does not resample the covariates and the resampled residuals $\varepsilon_t^* = d_t^* \widehat{\varepsilon}_t$, where $\widehat{\varepsilon}_t$ was the estimated residual and d_t^* followed a two-point distribution. Both the smoothed regression bootstrap and the wild bootstrap were based on $B = 500$ resamples for each simulation run. As there are infinitely many solutions for $\hat{\boldsymbol{\gamma}}$ from the MIQP algorithm, for each bootstrap resample, we outputted $N = 100$ solutions for the LSE of $\boldsymbol{\gamma}_0$ and used their average as $\widehat{\boldsymbol{\gamma}}_b^{*c}$.

TABLE 3
Empirical coverage probabilities and widths ($\times 100$ in parentheses) of the 95% confidence intervals for five projected parameters $\{\tilde{\boldsymbol{\gamma}}^{\mathrm{T}} \boldsymbol{d}_i\}_{i=1}^{5}$ obtained with the smoothed regression bootstrap (Smooth) and the wild bootstrap (Wild) based on 500 resamples.

T	$\boldsymbol{d}_1$		$\boldsymbol{d}_2$		$\boldsymbol{d}_3$		$\boldsymbol{d}_4$		$\boldsymbol{d}_5$	
	Smooth	Wild	Smooth	Wild	Smooth	Wild	Smooth	Wild	Smooth	Wild
200	0.92 (6.76)	0.87 (3.57)	0.97 (6.91)	0.87 (3.91)	0.93 (5.78)	0.90 (4.02)	0.93 (6.20)	0.83 (3.44)	0.96 (6.86)	0.86 (3.56)
400	0.95 (3.31)	0.86 (1.69)	0.94 (3.57)	0.83 (1.89)	0.97 (2.56)	0.86 (1.94)	0.94 (3.37)	0.88 (1.73)	0.97 (3.69)	0.85 (1.75)
800	0.93 (1.70)	0.85 (0.83)	0.96 (1.76)	0.87 (0.99)	0.94 (1.68)	0.88 (1.00)	0.96 (1.72)	0.88 (0.86)	0.96 (1.80)	0.87 (0.76)
1600	0.95 (0.81)	0.83 (0.40)	0.94 (0.86)	0.88 (0.51)	0.95 (0.89)	0.90 (0.53)	0.96 (0.85)	0.84 (0.41)	0.94 (0.79)	0.85 (0.42)

To evaluate the quality of the two bootstrap schemes, we constructed 95% confidence intervals (CIs) for $\tilde{\gamma}_0 = (\boldsymbol{\gamma}'_{-1,10}, \boldsymbol{\gamma}'_{-1,20})^{\mathrm{T}} = (-1, 0, 1, 0)^{\mathrm{T}}$ projected on five directions $\{\boldsymbol{d_i}\}_{i=1}^{5}$ where $\boldsymbol{d_i} = \boldsymbol{e}_i$ for $i = 1, \ldots, 4$ and $\boldsymbol{d}_5 = \sum_{i=1}^{4} \boldsymbol{d}_i/2$, and $\boldsymbol{e}_i = (e_{i1}, \cdots, e_{i4})^{\mathrm{T}}$ with $e_{ii} = 1$ and $e_{ij} = 0$ if $j \neq i$. Table 3 reports the coverage probabilities and widths of the nominal 95% CIs for $\tilde{\gamma}_0^{\mathrm{T}} \boldsymbol{d}_i$ based on the smoothed regression bootstrap and the wild bootstrap, respectively. It is shown that the smoothed regression bootstrap had satisfactory coverage as its empirical coverage levels were quite close to the nominal 95% level under large sample sizes for all the five projection directions. This verified the consistency of the proposed bootstrap procedure in Theorem 5.1. On the other hand, the wild bootstrap had substantial under-coverage, and its coverage was not improved with the increases of the sample sizes. The comparison between the two bootstrap schemes reveals that for the inference of γ_0, it is crucial to conduct resampling from a smoothed distribution, as advocated in Section 5.

8. Case Study.

Air quality is naturally affected by meteorological regimes as the latter defines the atmospheric dispersion conditions. We demonstrate here that the four-regime regression model is well suited for $\mathrm{PM}_{2.5}$ modeling in Beijing.

We considered hourly $\mathrm{PM}_{2.5}$ data from Wanshouxigong site in central Beijing with the meteorological data from the nearest weather observation site being used. The study period was from December 1, 2018 to November 30, 2019, which encompassed four seasons. The meteorological data included the air temperature (TEMP), dew point temperature (DEWP), surface air pressure (PRES), the cumulative wind speed (IWS) at a direction and wind direction (WD). Cumulative rainfall (RAIN) was included in summer, however not in the other three seasons due to a lack of it. The categorical wind direction (WD) took five values: Northwesterly (NW), Northeasterly (NE), Southwesterly (SW), Southeasterly (SE) and calm and variable (CV). We also used the boundary layer height (BLH), which defines the vertical dispersion property, from European Centre for Medium-Range Weather Forecasts (ECMWF).

To investigate the in-sample and out-of-sample performances, the data were divided to the training and testing sets, where the testing sets consisted of the data from the 11-th to the 20-th days of a month and the training sets included the rest of the data in the month. $\mathrm{PM}_{2.5}$ was regressed on covariates TEMP, DEWP, PRES, $\log(\mathrm{BLH})$, IWS, WD as well as the $\mathrm{PM}_{2.5}$ at the previous hour (Lag $\mathrm{PM}_{2.5}$). For the wind direction, NW, NE, SW and SE were set as dummy covariates with the CV as the baseline.

Along with the proposed four-regime model (4-REG), the global linear regression (GLR), the two-regime model (2-REG) [22] and [35], the linear regression tree (LRT) ([37]) and the multivariate adaptive regression splines (MARS) [11] were also considered. For 2-REG and 4-REG, the splitting boundaries were determined by TEMP, DEWP, $\log(\mathrm{BLH})$, IWS, and the four wind directions NE, NW, SE and SW with the coefficients standardized so that the intercept term being 1.

Fig 2: Mean squared errors (MSE) for $\mathrm{PM}_{2.5}$ on the training (red) and testing (green) sets for each season of five models, including global linear regression (GLR), two-regime model (2-REG), four-regime model (4-REG), linear regression tree (LRT) and multivariate adaptive regression splines (MARS), with model ranks (in increasing order of the MSEs) marked on top of the bars.

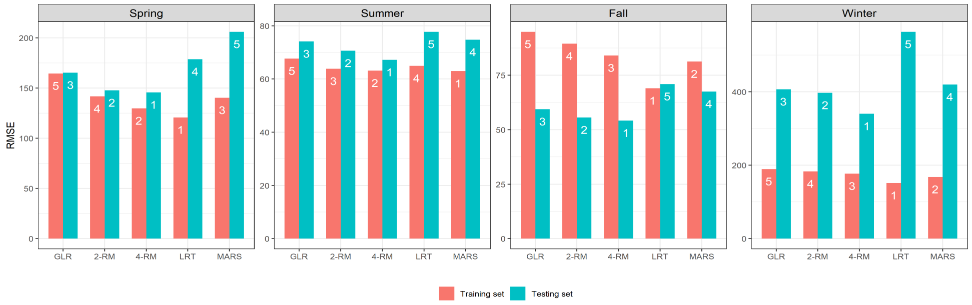


Figure 2 summarizes the in-sample and out-of-sample MSEs of these models in each season. Within the training sets, LRT or MARS achieved the lowest MSE among the five models with the average rank being 1.75 and 2, respectively. Here, rank 1 indicates the best performance. The average rank of the 4-REG in the training groups was 2.5, while those of the 2-REG and GLR ranked the lowest in all seasons. However, LRT and MARS had the highest prediction MSEs on the testing sets, even worse than the benchmark GLR for all seasons, indicating they were severely over-fitted. The segmented linear models, 4-REG and 2-REG, were the best two in terms of out-of-sample performances, with the 4-REG achieving the lowest predictive errors consistently in all seasons.

The estimated 4-regimes models in the spring, summer and fall seasons all had three regimes, as the fourth estimated regime had zero sample size in the three seasons. A further examination suggested that the two estimated boundaries had no intersections over the sample regions, which corresponded to Model (6.1) and reflected the fact that the proposed LS criterion based on the four-regime model may be able to produce a three-regime model if the latter offers better fit. The winter had four estimated regimes. The estimated regression coefficients and their 95% confidence intervals are given in Figure S4 of the SM ([33]).

TABLE 4

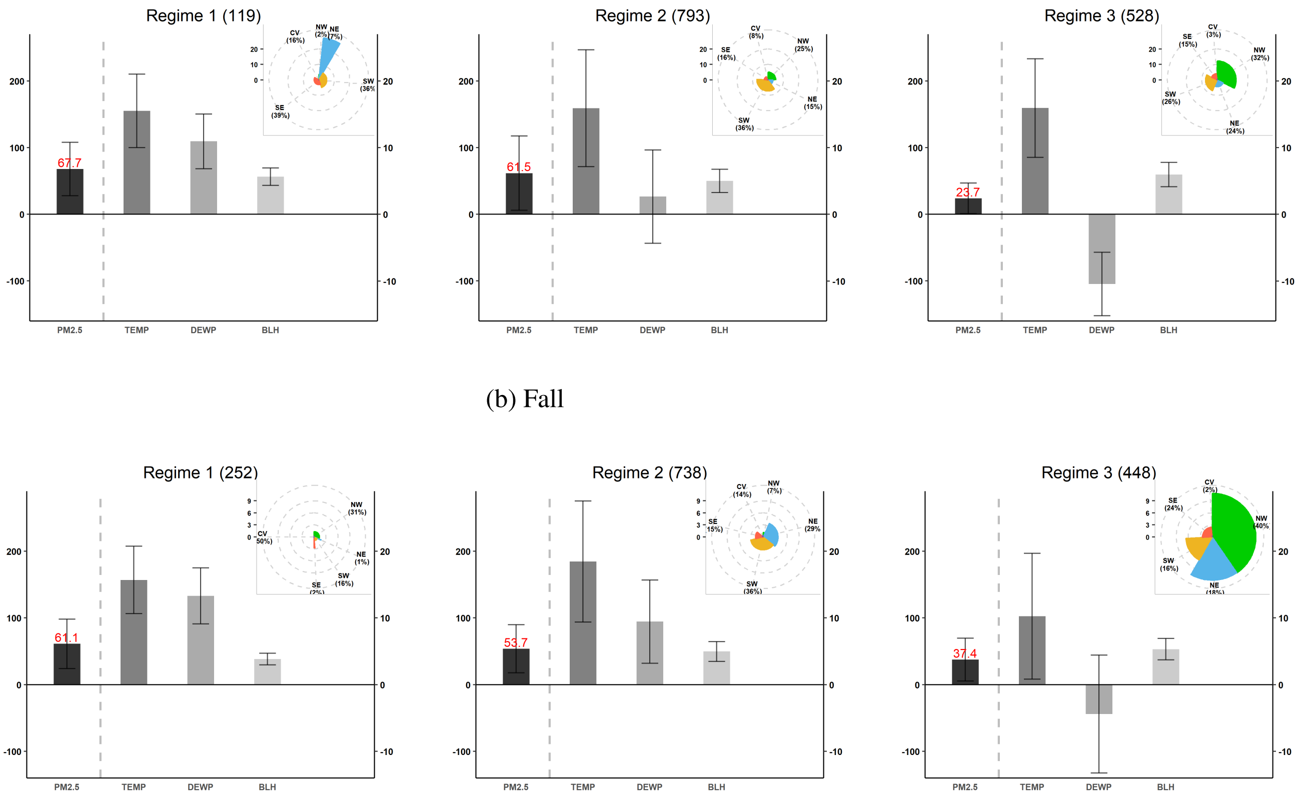
Estimated coefficients of the splitting boundaries and cos *of the angle* ϕ *between the two boundaries. The coefficients were normalized such that the coefficients of the intercept terms were* 1. *All the covariates were standardized such that their sample means were* 0 *and standard deviations were* 1 *in each season.*

Season	γ	TEMP	DEWP	IWS	log(BLH)	NE	NW	SE	SW	$\cos\phi$
Spring	1	1.3	-2.5	-0.0	-0.4	0.9	0.3	0.1	0.0	0.78
	2	0.4	-0.5	-0.1	-0.1	0.6	0.6	0.1	0.3	
Summer	1	1.0	5.5	-12.9	-0.0	-12.7	-15.0	-8.9	-9.0	0.75
	2	0.4	0.2	-0.2	0.0	-0.7	-0.7	-0.7	-0.7	
Fall	1	0.7	-1.0	0.3	-0.1	0.5	-0.0	0.3	0.0	0.65
	2	-0.5	1.6	-1.0	0.0	0.1	-1.6	-1.3	-0.1	
Winter	1	0.2	-0.5	0.6	-0.2	0.2	0.4	0.4	-0.4	0.45
	2	0.0	-0.6	0.2	-0.4	1.2	1.4	0.3	1.0	

Table 4 reports the estimated coefficients of the two splitting boundaries for each season as well as the cosine of the dihedral angle (denoted as ϕ) between the two boundary hyperplanes. It can be seen that $\cos\phi$ for the first three seasons were relatively larger than that in winter, which explains why the boundary hyperplanes of these three seasons were non-intersected. Table 4 indicates that the DEWP and the wind-related variables were the most influential in determining the slopes of the estimated boundaries due to their absolute coefficient values as the γ was normalized. This reveals an attraction of the proposed regime-splitting mechanism in that the splitting boundaries are determined empirically by multivariate covariates, which contrasts to the threshold regression where the boundary variable has to be user-specified.

Figure 3 displays summary statistics of $\mathrm{PM}_{2.5}$ and the meteorological variables under the three regimes in the spring and fall seasons, as well as the rose plots for the wind directions and the average integrated wind speed (IWS). It shows that the segmented regression picked up three meteorological regimes on $\mathrm{PM}_{2.5}$ where Regime 1 corresponded to the pollution state with high DEWP and high proportion of Calm and Variable wind (CV) which are known to encourage the secondary generation of PM2.5 and unfavorable static atmospheric diffusion, Regime 2 was a transitional state between the clean and high pollution states with reduced DEWP and CV, and Regime 3 was a cleaning state dominated by the northerly wind which brought cleaner and cooler air from the north. Results of the other two seasons and analysis are provided in Figure S5 of the SM.

Fig 3: Bar and rose plots for key variables under each estimated regimes in spring and fall 2019. The height of the bars indicate the sample means with imposed line segments indicating twice of the sample deviations above and below the means. The rose plots display the distribution of wind directions (width of angles) and average speed (length of radius). Sample sizes of each regime is reported in the subtitle.



(a) Spring

(b) Fall

9. Discussion. This paper develops a statistical inference approach for four-regimes segmented linear models, which broadens the scope of the two-regime models of [22] and [35], and can attains valid inference for degenerated models with less than four regimes. The proposed segmented model is shown to produce better in-sample and out-sample results for the air quality data in Beijing and produced regime-splitting results which had clear atmospheric physics interpretation.

There are two possible extensions which may be considered in future research. One is to allow endogeneity which may be encountered in economic and social behavior applications. If $\boldsymbol{X}_t$ is endogenous and $\boldsymbol{Z}_t$ is exogenous, $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ can be consistently estimated with instrument variables $\boldsymbol{V}_t$ and the two-stage least squares estimation (2SLS) by first regressing $\boldsymbol{X}_t$ on $\boldsymbol{V}_t$, and then using the fitted $\hat{\boldsymbol{X}}_t$ to substitute $\boldsymbol{X}_t$ in the four-regime model. The LS estimation via the MIQP and the inference methods for the four-regime model presented in this paper is still applicable. However, the 2SLS is no longer working if $\boldsymbol{Z}_t$ is endogenous as discussed in [36], who proposed a conditioning and re-centering approach which might be extended to the four-regime model. Specifically, let $g(\boldsymbol{X}_t, \boldsymbol{Z}_t) = \boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_{10} + \mathbb{E}(\varepsilon_t|\boldsymbol{X}_t, \boldsymbol{Z}_t)$, $\boldsymbol{\delta}_{k0} = \boldsymbol{\beta}_{k0} - \boldsymbol{\beta}_{10}$ for $k \neq 1$, and $e_t = \varepsilon_t - \mathbb{E}(\varepsilon_t|\boldsymbol{X}_t, \boldsymbol{Z}_t)$, then Model (2.1) can be written as $Y_t = g(\boldsymbol{X}_t, \boldsymbol{Z}_t) + \sum_{k=1}^{3} \boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\delta}_{k0}\mathbb{1}\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\} + e_t$, which is a partially linear segmented model, where γ_0 is identifiable without instrument variables. However, the integrated difference kernel estimator used in [36] was designed for univariate threshold, and it is interesting to see how it can be extended to multivariate γ_0. Alternatively, one may consider estimating γ_0 via the mixed integer programming with the nonlinear $\mathbb{E}(\varepsilon_t|\boldsymbol{X}_t, \boldsymbol{Z}_t)$ part approximated via sieve functions. How to solve these issues in the context of the four-regime model requires further investigation.

Another extension is for segmented models with $L > 2$ splitting hyperplanes. In general, the L splitting hyperplanes in $\mathbb{R}^d$ can lead to as many as $K_L = \sum_{i=0}^{\min(L,d)} \binom{L}{i}$ segments, as shown in Section G of the SM ([33]). It is clear that the investigations in this study for the two boundary case provide vital understanding to the general cases. For example, if we consider an extension to the case of having three hyperplanes in $\mathbb{R}^d$, we can fit a segmented model with $K = \sum_{i=0}^{\min(3,d)} \binom{3}{i}$ regimes by the least squares estimation, whose criterion function would have the same form as (3.1). The backward selection procedure in Section 6 can be employed to specify the optimal number of regimes, and the smoothed regression bootstrap is still able to facilitate the inference for γ_0 and $\boldsymbol{\beta}_0$. Furthermore, the proof for the asymptotic distributions of the least squares estimators can be modified to suit the more general segmented models. The main challenge for the general cases is the complicated model form and demanding computation costs caused by the increase of L, requiring efforts in further studies. On the other hand, as K_L grows exponentially with respect to L if $d > L$ and polynomially if $d \leq L$, there would be little need to consider segmented models with large L and d as the nonparametric local models (regression trees, etc) may be better suited.

Acknowledgements. The research was partially supported by National Natural Science Foundation of China grants 12292980, 12292983 and 92358303.

SUPPLEMENTARY MATERIAL

Supplement to "Statistical Inference on Four-Regime Segmented Regression Models"
In the supplementary material, we present technical details, proofs and additional results of the simulations and the case study.

REFERENCES

[1] Auerbach, A. J. and Gorodnichenko, Y. (2012). Measuring the Output Responses to Fiscal Policy. *American Economic Journal: Economic Policy* **4** 127. https://doi.org/10.1257/pol.4.2.1

[2] Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. https://doi.org/10.2307/2998540 MR1616121

[3] Bertsimas, D., King, A. and Mazumder, R. (2016). Best subset selection via a modern optimization lens. *Ann. Statist.* **44** 813–852. https://doi.org/10.1214/15-AOS1388 MR3476618

[4] Bertsimas, D. and Weismantel, R. (2005). *Optimization over Integers* **13**. Dynamic Ideas Belmont.

[5] Card, D., Mas, A. and Rothstein, J. (2008). Tipping and the Dynamics of Segregation. *The Quarterly Journal of Economics* **123** 177–218. https://doi.org/10.1162/qjec.2008.123.1.177

[6] Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Ann. Statist.* **21** 520–533. https://doi.org/10.1214/aos/1176349040 MR1212191

[7] Chernozhukov, V. and Fernández-Val, I. (2011). Inference for extremal conditional quantile models, with an application to market and birthweight risks. *The Review of Economic Studies* **78** 559–589. https://doi.org/10.1093/restud/rdq020

[8] Chernozhukov, V. and Hong, H. (2004). Likelihood estimation and inference in a class of nonregular econometric models. *Econometrica* **72** 1445–1480. https://doi.org/10.1111/j.1468-0262.2004.00540.x MR2077489

[9] Davies, R. B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **74** 33–43. https://doi.org/10.1093/biomet/74.1.33 MR885917

[10] Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika* **85** 645–660. https://doi.org/10.1093/biomet/85.3.645 MR1665822

[11] FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141. With discussion and a rejoinder by the author. https://doi.org/10.1214/aos/1176347963 MR1091842

[12] GONZALO, J. and PITARAKIS, J.-Y. (2002). Estimation and model selection based inference in single and multiple threshold models. *J. Econometrics* **110** 319–352. Long memory and nonlinear time series (Cardiff, 2000). https://doi.org/10.1016/S0304-4076(02)00098-2 MR1928308

[13] GONZALO, J. and WOLF, M. (2005). Subsampling inference in threshold autoregressive models. *J. Econometrics* **127** 201–224. https://doi.org/10.1016/j.jeconom.2004.08.004 MR2156333

[14] GYÖRFI, L., HÄRDLE, W., SARDA, P. and VIEU, P. (1989). *Nonparametric curve estimation from time series*. *Lecture Notes in Statistics* **60**. Springer-Verlag, Berlin.

[15] HANSEN, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64** 413–430. https://doi.org/10.2307/2171789 MR1375740

[16] HANSEN, B. E. (2000). Sample splitting and threshold estimation. *Econometrica* **68** 575–603. https://doi.org/10.1111/1468-0262.00124 MR1769379

[17] HÄRDLE, W., HOROWITZ, J. and KREISS, J.-P. (2003). Bootstrap methods for time series. *International Statistical Review* **71** 435–459.

[18] HSING, T. (1995). On the asymptotic independence of the sum and rare values of weakly dependent stationary random variables. *Stochastic Process. Appl.* **60** 49–63. https://doi.org/10.1016/0304-4149(95)00054-2 MR1362318

[19] JIANG, Z., DU, C., JABLENSKY, A., LIANG, H., LU, Z., MA, Y. and TEO, K. L. (2014). Analysis of schizophrenia data using a nonlinear threshold index logistic model. *PloS ONE* **9** e109454. https://doi.org/10.1371/journal.pone.0109454

[20] KHALILI, A. and CHEN, J. (2007). Variable selection in finite mixture of regression models. *J. Amer. Statist. Assoc.* **102** 1025–1038. https://doi.org/10.1198/016214507000000590 MR2411662

[21] KNIGHT, K. (1999). Epi-convergence and stochastic equisemicontinuity. Preprint.

[22] LEE, S., LIAO, Y., SEO, M. H. and SHIN, Y. (2021). Factor-driven two-regime regression. *Ann. Statist.* **49** 1656–1678. https://doi.org/10.1214/20-aos2017 MR4298876

[23] LI, D. and LING, S. (2012). On the least squares estimation of multiple-regime threshold autoregressive models. *J. Econometrics* **167** 240–253. https://doi.org/10.1016/j.jeconom.2011.11.006 MR2885449

[24] LIU, R. Y. (1988). Bootstrap procedures under some non-i.i.d. models. *Ann. Statist.* **16** 1696–1708. https://doi.org/10.1214/aos/1176351062 MR964947

[25] MEYER, R. M. (1973). A Poisson-type limit theorem for mixing sequences of dependent "rare" events. *Ann. Probability* **1** 480–483. https://doi.org/10.1214/aop/1176996941 MR350816

[26] POLITIS, D. N. and ROMANO, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Statist.* **22** 2031–2050. https://doi.org/10.1214/aos/1176325770 MR1329181

[27] POTTER, S. M. (1995). A nonlinear approach to US GNP. *Journal of Applied Econometrics* **10** 109-125. https://doi.org/10.1002/jae.3950100203

[28] RESNICK, S. I. (2008). *Extreme values, regular variation and point processes*. *Springer Series in Operations Research and Financial Engineering*. Springer, New York. MR2364939

[29] SCHWARTZ, P. F., GENNINGS, C. and CHINCHILLI, V. M. (1995). Threshold models for combination data from reproductive and developmental experiments. *Journal of the American Statistical Association* **90** 862–870. https://doi.org/10.1080/01621459.1995.10476585

[30] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR468014

[31] SEIJO, E. and SEN, B. (2011). Change-point in stochastic design regression and the bootstrap. *Ann. Statist.* **39** 1580–1607. https://doi.org/10.1214/11-AOS874 MR2850213

[32] TONG, H. (1983). *Threshold Models in Non-linear Time Series Analysis. Lecture Notes in Statistics, No. 21*. Springer-Verlag.

[33] YAN, H. and CHEN, S. X. (2024). Supplement to "Statistical Inference for Four-Regime Segmented Regression Models".

[34] YU, P. (2014). The bootstrap in threshold regression. *Econometric Theory* **30** 676–714. https://doi.org/10.1017/S0266466614000012 MR3205610

[35] YU, P. and FAN, X. (2021). Threshold regression with a threshold boundary. *Journal of Business & Economic Statistics* **39** 953–971. https://doi.org/10.1080/07350015.2020.1740712

[36] YU, P. and PHILLIPS, P. C. B. (2018). Threshold regression with endogeneity. *J. Econometrics* **203** 50–68. https://doi.org/10.1016/j.jeconom.2017.09.007 MR3758327

[37] ZEILEIS, A., HOTHORN, T. and HORNIK, K. (2008). Model-based recursive partitioning. *J. Comput. Graph. Statist.* **17** 492–514. https://doi.org/10.1198/106186008X319331 MR2439970