Statistical Inference for Four-regime Segmented Regression Models

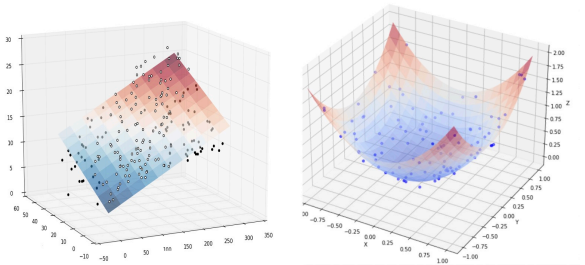
Han Yan

Peking University

Joint work with Song Xi Chen

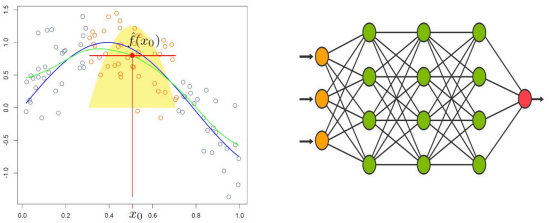
Motivation

Global models



- ✓ Great interpretability and computation simplicity
- ✘ Poor fitting performances

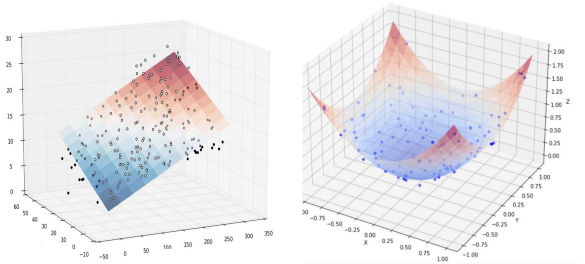
Local models



- ✓ Better fitting performances and model flexibility
- ✘ Lack of interpretability and overfitting

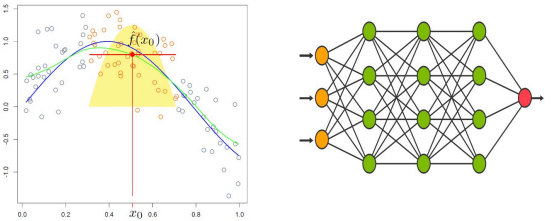
Motivation

Global models



- ✓ Great interpretability and computation simplicity
- ✘ Poor fitting performances

Local models

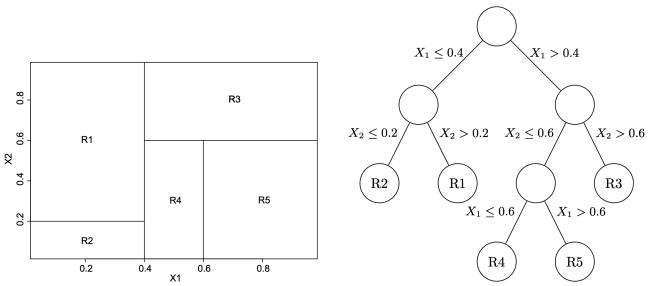


- ✓ Better fitting performances and model flexibility
- ✘ Lack of interpretability and overfitting

As an intermediate, a segmented regression model, where $\mathbb{E}(Y|\boldsymbol{X})$ is homogeneous on several segmented regions, may combine some benefits of the two types of models.

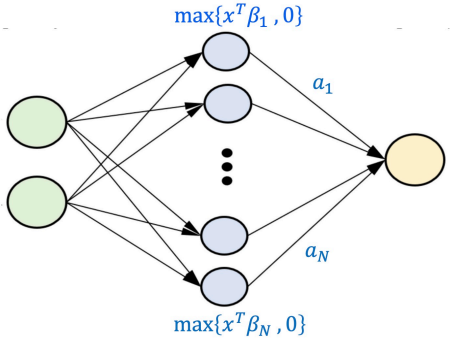
Trees and ReLU Networks as Segmented Models

Regression tree



- Recursively fit a two-region linear model
- Univariate splitting boundary

Shallow ReLU network: $f(\boldsymbol{x}) = \sum_{l=1}^{N} a_l \cdot \sigma(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_l)$

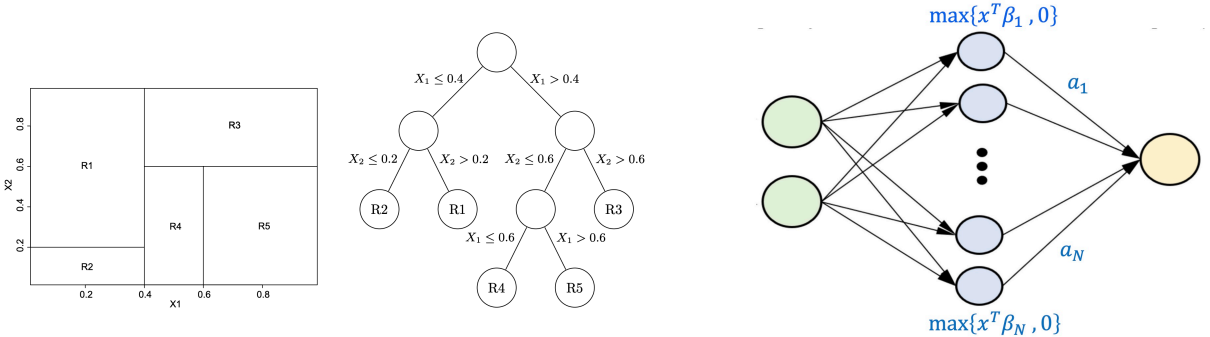


- Linear combination of two-region linear models.
- Can be reparametrized into a segmented model.

Trees and ReLU Networks as Segmented Models

Regression tree

ReLU neural network: $f(\boldsymbol{x}) = \sum_{l=1}^{N} a_l \cdot \sigma(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_l)$



- Trees and ReLU networks can be regarded as segmented linear models, if not grow too deep.
- In this perspective, a comprehensive study of the segmented models may improve our understanding of these methods.

Segmented Linear Model

Consider a class of segmented regression models, where there are L splitting hyperplanes $\{\Gamma_0^{(l)} : \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma}_l = 0\}_{l=1}^{L}$, which partition the whole space of $\boldsymbol{Z}$ into K disjoint regions $\{R_k(\boldsymbol{\gamma})\}_{k=1}^{K}$:

$$Y = \sum_{k=1}^{K} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_k \mathbb{1}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma})\} + \varepsilon, \tag{1}$$

- $\boldsymbol{Z} = (Z_1, \cdots, Z_d)^{\mathrm{T}}$ determines the regions and can share common variables with $\boldsymbol{X}$.
- $\mathbb{E}(\varepsilon|\boldsymbol{X}, \boldsymbol{Z}) = 0$, so that $\mathbb{E}\{Y|\boldsymbol{X}, \boldsymbol{Z} \in R_k(\boldsymbol{\gamma})\} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_k$.

Challenges in computation and inference:

- The *non-convexity* introduced by the region indicators brings obstacles to optimization;
- The estimator for $\boldsymbol{\gamma}$ is *not* asymptotic normal. So its statistical inference is not routine.

Existing Studies on Segmented Linear Models

1. **Threshold regression** (Tong, 1990,Hansen, 1996)

   univariate boundary and two regression regions

$$\mathbb{E}(Y|\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\delta}\mathbb{1}(Z > r)$$

2. **Multiple change points regression** (Bai and Perron, 1998, Li and Ling, 2012):

   time index boundaries and multiple regions.

$$\mathbb{E}(Y|\boldsymbol{X}) = \sum_{j=1}^{m+1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_j\mathbb{1}(T_{j-1} < t \leqslant T_j)$$

3. **Two regime regression** (Lee et al., 2021, Yu and Fan, 2021):

   multivariate boundary and two regions.

$$\mathbb{E}(Y|\boldsymbol{X}) = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\delta}\mathbb{1}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{\gamma} > 0)$$

Most studies about statistical inference on segmented regression are still quite restrictive.

Goal of This Study

- Extend the existing segmented regression to a more general class
- Asymptotic results under weakly dependent data
- Valid inference methods for the boundary coefficient
- Choose the optimal number of segmented regions (model selection)
- Efficient computational algorithm
- Application to model the meteorological effects on the $PM_{2.5}$

Model Setup

- In general, suppose there are L splitting hyperplanes $\{\Gamma_0^{(l)} : \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma}_l = 0\}_{l=1}^{L}$ in $\mathbb{R}^d$, then there will be $K = \sum_{i=0}^{\min(L,d)} \binom{L}{i}$ segmented regions since each combination of signs of $\{\boldsymbol{z}^{\mathrm{T}}\boldsymbol{\gamma}_l\}_{l=1}^{L}$ determines a region. When $d > L$, then $K = 2^L$.
- For simplicity of exposition, we henceforth confine $L = 2$, where $K = 4$ if $d \geqslant 2$. Methods and theories for general cases can be extended.
- Suppose the sample $\{Y_t, \boldsymbol{X}_t, \boldsymbol{Z}_t\}_{t=1}^{T}$ is generated by

$$Y_t = \sum_{k=1}^{4} \boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_{k0}\mathbb{1}\left\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\right\} + \varepsilon_t. \tag{2}$$

- ε_t is the residual satisfying $\mathbb{E}(\varepsilon_t|\boldsymbol{X}_t, \boldsymbol{Z}_t) = 0$ with finite second moment.
- The regime indicator $\mathbb{1}\left\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)\right\}$ can be explicitly expressed as $\mathbb{1}_k\left(\boldsymbol{Z}_t^{\mathrm{T}}\boldsymbol{\gamma}_{10}, \boldsymbol{Z}_t^{\mathrm{T}}\boldsymbol{\gamma}_{20}\right)$, where $\{\mathbb{1}_k(u,v)\}_{k=1}^{4}$ are defined as $\mathbb{1}_1(u,v) = \mathbb{1}(u > 0, v > 0)$, $\mathbb{1}_2(u,v) = \mathbb{1}(u \leqslant 0, v > 0)$, $\mathbb{1}_3(u,v) = \mathbb{1}(u \leqslant 0, v \leqslant 0)$ and $\mathbb{1}_4(u,v) = \mathbb{1}(u > 0, v \leqslant 0)$.
- Since the signs of $\boldsymbol{Z}_1^{\mathrm{T}}\boldsymbol{\gamma}_{10}$ and $\boldsymbol{Z}_2^{\mathrm{T}}\boldsymbol{\gamma}_{20}$ determine the regimes in Model (2), the first element of $\boldsymbol{\gamma}_{10}$ and $\boldsymbol{\gamma}_{20}$ are normalized as 1 in order to be identifiable.

Estimation

- 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

$$\mathbb{M}_T(\boldsymbol{\theta}) = \frac{1}{T}\sum_{t=1}^{T}\left\{Y_t - \sum_{k=1}^{4}\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\beta}_k \mathbb{1}_k(\boldsymbol{Z}_{1,t}^{\mathrm{T}}\boldsymbol{\gamma}_1, \boldsymbol{Z}_{2,t}^{\mathrm{T}}\boldsymbol{\gamma}_2)\right\}^2 =: \frac{1}{T}\sum_{t=1}^{T} m(\boldsymbol{W}_t, \boldsymbol{\theta}), \tag{3}$$

  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 $\boldsymbol{\gamma} \in \Gamma_i$ 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 $\boldsymbol{\gamma}$ 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 $\boldsymbol{\gamma}$, which is denoted as $\widehat{\mathcal{G}}$, such that a LSE $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ satisfies

$$\mathbb{M}_T(\widehat{\boldsymbol{\theta}}) = \inf_{\boldsymbol{\theta}\in\Theta} \mathbb{M}_T(\boldsymbol{\theta}) \quad \text{for any} \quad \widehat{\boldsymbol{\gamma}} \in \widehat{\mathcal{G}}. \tag{4}$$

Identification and consistency

Let $\boldsymbol{W} = (Y, \boldsymbol{X}, \boldsymbol{Z}_1, \boldsymbol{Z}_2)$ follow the stationary distribution $\mathbb{P}_0$ of $\boldsymbol{W}_t$, and $q_i = \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma}_{i0}$ for $i = 1$ and 2 to indicate whether $\boldsymbol{Z}$ is located on the true hyperplane H_{i0} or not.

Assumption 1 (temporal dependence) *(i) The time series $\{\boldsymbol{W}_t\}_{t \geqslant 1}$ is stationary and α-mixing with the mixing coefficient $\alpha(t) \leqslant c\rho^t$ for finite positive constants c and $\rho \in (0, 1)$. (ii) $\mathbb{E}(\varepsilon_t | \mathcal{F}_{t-1}) = 0$, where $\mathcal{F}_{t-1}$ is a filtration generated by $\{(\boldsymbol{X}_i, \boldsymbol{Z}_{1i}, \boldsymbol{Z}_{2i}, \varepsilon_{i-1}) : i \leqslant t\}$.*

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 $\mathbb{P}(|q_i| \leqslant \epsilon | \boldsymbol{Z}_{-j,i}) > 0$ 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 $\mathbb{E}\{\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} | \boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0) \cap R_h(\boldsymbol{\gamma})\} \geqslant \lambda_0$ 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, $\mathbb{P}(\boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma}_1 < 0 < \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma}_2) \leqslant c_1 \|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\|$ if $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \in \mathcal{N}(\boldsymbol{\gamma}_{i0}; \delta_0)$, for some constants $\delta_0, c_1 > 0$.*

Identification and consistency

The identification of $\boldsymbol{\theta}_0$ under the least squared criterion is ensured in the following proposition.

Proposition 1 (Identification)

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 following theorem shows that any LSE estimators $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ defined in (4) are consistent to $\boldsymbol{\theta}$. It is worth noting that though there exist infinitely many solutions $\widehat{\gamma}$ which are collected in the convex set $\widehat{\mathcal{G}}$, the consistency of each $\widehat{\gamma}$ can be guaranteed, which implies that the solution set $\widehat{\mathcal{G}}$ is a local neighborhood of γ_0 with a shrinking radius.

Theorem 1 (Consistency)

Under Assumptions 1–3, let $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\gamma}}^{\mathrm{T}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ *for any* $\widehat{\gamma} \in \widehat{\mathcal{G}}$, , *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 $\{R_k(\widehat{\boldsymbol{\gamma}})\}_{k=1}^4$.

Corollary 1

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

Convergence rate

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 $\mathbb{P}(|q_i| < \epsilon | \boldsymbol{Z}_{-1,i}) \geqslant c_2\epsilon$ 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 $\inf_{\boldsymbol{\gamma} \in \mathcal{N}_i} \mathbb{E}(\|\boldsymbol{X}\| | \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma} = 0) > 0$ almost surely. (iii) $\mathbb{P}(\boldsymbol{Z}_1^{\mathrm{T}} \boldsymbol{\gamma}_1 < 0 < \boldsymbol{Z}_1^{\mathrm{T}} \boldsymbol{\gamma}_2, \boldsymbol{Z}_2^{\mathrm{T}} \boldsymbol{\gamma}_3 < 0 < \boldsymbol{Z}_2^{\mathrm{T}} \boldsymbol{\gamma}_4) \leqslant c_3 \|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_2\| \|\boldsymbol{\gamma}_3 - \boldsymbol{\gamma}_4\|$ 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) For some constant $r > 8$, $\sup_{\boldsymbol{\gamma} \in \mathcal{N}_i} \mathbb{E}(\|\boldsymbol{X}\|^r | \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma} = 0) < \infty$ and $\sup_{\boldsymbol{\gamma} \in \mathcal{N}_i} \mathbb{E}(\varepsilon^r | \boldsymbol{Z}_i^{\mathrm{T}} \boldsymbol{\gamma} = 0) < \infty$ almost surely.*

Under the above assumptions we can obtain

$$\mathbb{M}(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{\theta}_0) \gtrsim \|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|^2 + \|\boldsymbol{\gamma}_0 - \boldsymbol{\gamma}\|$$

and two maximal inequalities regarding the empirical processes $\sup_{\boldsymbol{\theta} \in \Theta} |\mathbb{M}_T(\boldsymbol{\theta}) - \mathbb{M}(\boldsymbol{\theta})|$, which lead to the following convergence rate.

Theorem 2 (Convergence rate)

Under Assumptions 1–4, $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(T^{-\frac{1}{2}})$ and $\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\| = O_p(T^{-1})$ for any $\widehat{\boldsymbol{\gamma}} \in \widehat{\mathcal{G}}$.

Asymptotic Distributions

We define the following quantities to present the asymptotic distributions of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$.

- For each $k \in \{1, \cdots, 4\}$, let

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

- Let $q_{i,t} = \boldsymbol{Z}_{i,t}^{\mathrm{T}}\boldsymbol{\gamma}_{i0}$ for $i = 1$ and 2, and $s_i^{(k)} \in \{-1, 1\}$ be the sign of $q_{i,t}$ for $\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0)$.
- If two adjacent regions $R_k(\boldsymbol{\gamma}_0)$ and $R_h(\boldsymbol{\gamma}_0)$ are divided by H_i, we denote $(k, h) \in \mathcal{S}(i)$ and let

$$\xi_t^{(k,h)} = \left(\boldsymbol{\delta}_{kh,0}^{\mathrm{T}}\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\delta}_{kh,0} + 2\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{\delta}_{kh,0}\varepsilon_t\right)\mathbb{1}\left\{\boldsymbol{Z}_t \in R_k(\boldsymbol{\gamma}_0) \cup R_h(\boldsymbol{\gamma}_0)\right\} \tag{5}$$

  where $\boldsymbol{\delta}_{kh,0} = \boldsymbol{\beta}_{k0} - \boldsymbol{\beta}_{h0}$.
- 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.

Asymptotic Distributions

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

$$D(\boldsymbol{v}) = \sum_{i=1,2} \sum_{k,h \in \mathcal{S}(i)} \sum_{\ell=1}^{\infty} \xi_{\ell}^{(k,h)} \mathbb{1}\left\{ s_i^{(k)} \left(J_{i,\ell}^{(k,h)} + (\boldsymbol{Z}_{i,\ell}^{(k,h)})^{\mathrm{T}} \boldsymbol{v}_{-1,i} \right) \leqslant 0 < s_i^{(k)} J_{i,\ell}^{(k,h)} \right\}, \tag{6}$$

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

Remark: $D(\boldsymbol{v})$ is a sum of multivariate compound Poisson processes and only depends on pairs of adjacent regions.

Intuitively, this is because $D(\boldsymbol{v})$ largely relies on those points lying in a local neighbourhood 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.

Asymptotic Distributions

- Let $\mathcal{G}_D = \{\boldsymbol{v}_m : D(\boldsymbol{v}_m) \leqslant D(\boldsymbol{v}) \text{ if } \boldsymbol{v} \neq \boldsymbol{v}_m\}$ 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 $\mathcal{G}_D$.
- 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 $C(\mathcal{A}) = \int_{\boldsymbol{v}\in\mathcal{A}} \boldsymbol{v} d\boldsymbol{v} / \int_{\boldsymbol{v}\in\mathcal{A}} d\boldsymbol{v}$.
- 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 $\boldsymbol{\gamma}$.

Assumption 5 *(i) For $i = 1$ and 2, there exist constants $\delta_3, c_4 > 0$ such that $\mathbb{P}(|q_{i,t}| \leqslant \delta_3, |q_{i,t+j}| \leqslant \delta_3) \leqslant c_4 \{\mathbb{P}(|q_{i,t}| \leqslant \delta_3)\}^2$ uniformly for $t \geqslant 1$ and $j \geqslant 1$; (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 $c_4 \leqslant f_{q_i|\boldsymbol{Z}_{-1,i}}(0|\boldsymbol{z}_{-1,i}) \leqslant c_5$ 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 $f_{\xi^{(k,h)}|q_i,\boldsymbol{Z}_{-1,i}}(\xi|0, \boldsymbol{z}_{-1,i}) \leqslant c_6$ for a constant $c_6 > 0$; (iv) $\mathcal{Z}_{-1,i}$ is a compact subset of $\mathbb{R}^{d_i-1}$.*

Asymptotic Distributions

Theorem 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* $\{T(\widehat{\boldsymbol{\gamma}}_j^c - \boldsymbol{\gamma}_{j0})\}_{j=1}^2$ *are asymptotically independent.*

- The limiting process $D(\boldsymbol{v})$ is derived by the asymptotics of the point process induced by $\{(\xi_t^{(k,h)}, \boldsymbol{Z}_{-1,i,t}, Tq_{i,t})\}_{t=1}^T$, using large sample theories for extreme values.
- To accommodate discontinuities of the processes, we employed the epi-convergence in distribution (Knight, 1999), which is more general than the classic uniform convergence in distribution.
- 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 by establishing a thinning theorem of the Poisson process for the α-mixing sequences.

Computation

- Recall that the objective function is non-convex:

$$\mathbb{M}_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \left\{ Y_t - \sum_{k=1}^{4} \boldsymbol{X}_t^{\mathrm{T}} \boldsymbol{\beta}_k \mathbb{1}_k(\boldsymbol{Z}_t^{\mathrm{T}} \boldsymbol{\gamma}_1, \boldsymbol{Z}_t^{\mathrm{T}} \boldsymbol{\gamma}_2) \right\}^2 .$$

- For the univariate threshold model $Y = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} + \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\delta} I(Z > r) + \varepsilon$, the LSE of the $\hat{r}$ is often obtained via grid search, which is not suitable for multivariate boundaries.
- Since $\mathbb{M}_T(\boldsymbol{\theta})$ involves an indicator function, it is natural to consider transforming the original problem into a mixed integer quadratic programming (MIQP) problem.

- The general form of MIQP:

$$\begin{aligned} \min \quad & \boldsymbol{x}^{\mathrm{T}}\mathbf{H}\boldsymbol{x} + \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x} \\ \text{s.t.} \quad & \mathbf{A}\boldsymbol{x} \leqslant \mathbf{b} \\ & \boldsymbol{x} \in \mathbb{Z}^p \times \mathbb{R}^{n-p} \end{aligned}$$

- Key idea of forming the MIQP: use reparametrization and introduce some linear inequalities to guarantee the MIQP is equivalent to the original problem.

Mixed integer quadratic programming

The original LS problem (4) is reformulated to an MIQP problem as follows.
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:

$$\min_{\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{g}, \boldsymbol{\mathcal{I}}, \boldsymbol{\ell}} \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - \sum_{k=1}^{4} \sum_{i=1}^{p} X_{t,i} \ell_{k,i,t} \right)^2 \tag{7}$$

$$\text{subject to } \begin{cases} \boldsymbol{\beta}_k \in \mathcal{B}, \quad \boldsymbol{\gamma}_j \in \Gamma_j, \quad g_{j,t} \in \{0, 1\}, \quad I_{k,t} \in \{0, 1\}, \quad L_i \leqslant \beta_{k,i} \leqslant U_i, \\ (g_{j,t} - 1)(M_{j,t} + \epsilon) < \boldsymbol{Z}_{j,t}^{\mathrm{T}} \boldsymbol{\gamma}_j \leqslant g_{j,t} M_{j,t}, \ I_{k,t} L_i \leqslant \ell_{k,i,t} \leqslant I_{k,t} U_i, \\ L_i (1 - I_{k,t}) \leqslant \beta_{k,i} - \ell_{k,i,t} \leqslant U_i (1 - I_{k,t}), \\ I_{k,t} \leqslant s_j^{(k)} g_{j,t} + (1 - s_j^{(k)})/2, \quad I_{k,t} \geqslant \sum_{j=1}^{2} \left\{ s_j^{(k)} g_{j,t} + (1 - s_j^{(k)})/2 \right\} - 1, \end{cases} \tag{8}$$

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

Mixed integer quadratic programming

The following theorem shows that the formulated MIQP is equivalent to the original LS problem.

Theorem 4

For any small $\epsilon > 0$ in (8), let $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\gamma}}^{\mathrm{T}}, \tilde{\boldsymbol{\beta}}^{\mathrm{T}})^{\mathrm{T}}$ be a solution of the MIQP defined with (7) and (8), then $\mathbb{M}_T(\widehat{\boldsymbol{\theta}}) = \mathbb{M}_T(\tilde{\boldsymbol{\theta}})$ where $\widehat{\boldsymbol{\theta}}$ is a solution in (4).

- It can be efficiently solved via modern optimization software, such as `GUROBI` and `CPLEX`.
- We also proposed a blockwise coordinate descent version for large-dimension scenarios.

Inference for the Boundary Coefficient

- Since the inference for $\boldsymbol{\beta}_0$ is standard, we focus on the inference for the boundary coefficient $\boldsymbol{\gamma}_0$.
- The asymptotic distribution of $T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0)$ is nonpivotal and hard to simulate.
- A natural idea for is to employ the **bootstrap**. However, the nonparametric, the residual, and the wild bootstrap are all **failed** to consistently approximate the distribution of $T(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0)$.

> **Reason:** The bootstrap sampling distribution $\widehat{\mathbb{P}}_T$ must approximate the true distribution $\mathbb{P}_0$ in the neighbourhood of the boundary hyperplanes. However, $\mathbb{P}_0$ is continuous around the true boundaries so that $\boldsymbol{\gamma}_0$ is identifiable, while conditional on the original data, the bootstrap distribution $\widehat{\mathbb{P}}_T$ is discrete under the aforementioned bootstraps.

- Our remedy is to first smooth the data, then resampling from the smoothed distribution.

Smoothed regression bootstrap

- Suppose that $\boldsymbol{W} \sim \mathbb{P}_0$ follows the segmented model with heteroscedastic errors:

$$Y = \sum_{k=1}^{4} \boldsymbol{X}^{\mathrm{T}} \boldsymbol{\beta}_0 \mathbb{1}\{\boldsymbol{Z} \in R_k(\boldsymbol{\gamma}_0)\} + \sigma(\boldsymbol{X}, \boldsymbol{Z}) e, \tag{9}$$

  where $e \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{Z})$ with $\mathbb{E}(e) = 0, \mathbb{E}(e^2) = 1$, and $\sigma^2(\boldsymbol{X}, \boldsymbol{Z})$ is the conditional variance.

- 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 $i = 1, 2$. Our kernel estimator for $F_0(\boldsymbol{x}, \boldsymbol{z})$ is

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

- For any given $(\boldsymbol{x}, \boldsymbol{z})$, the local linear estimator $\widetilde{\sigma}^2(\boldsymbol{x}, \boldsymbol{z}) = \widehat{\alpha}$, which is defined by

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

- Let $\widehat{e}_t = \widehat{\varepsilon}_t / \widetilde{\sigma}(\boldsymbol{X}_t, \boldsymbol{Z}_t)$ and $\widehat{G}(e)$ be the empirical distribution of the centered $\{\widehat{e}_t\}_{t=1}^{T}$.

Smoothed regression bootstrap

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 the α-mixing sequences.

Smoothed Bootstrap: Procedure

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 $\{\widehat{\boldsymbol{\gamma}}_i^*\}_{i=1}^N$ are the LSEs for $\boldsymbol{\gamma}_0$ for a sufficiently large N. Let $\widehat{\boldsymbol{\gamma}}^{*c} = \sum_{i=1}^N \widehat{\boldsymbol{\gamma}}_i^*/N$.

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

Denote the distribution of $\{T(\widehat{\boldsymbol{\gamma}}^c - \boldsymbol{\gamma}_0), \sqrt{T}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\}$ as $\mathcal{L}_T$ and the empirical distribution of $\left\{T(\widehat{\boldsymbol{\gamma}}_b^{*c} - \widehat{\boldsymbol{\gamma}}^c), \sqrt{T}(\widehat{\boldsymbol{\beta}}_b^* - \widehat{\boldsymbol{\beta}})\right\}_{b=1}^B$ as $\mathcal{L}_{T,B}$.

Theorem 5

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

Degenerated Models

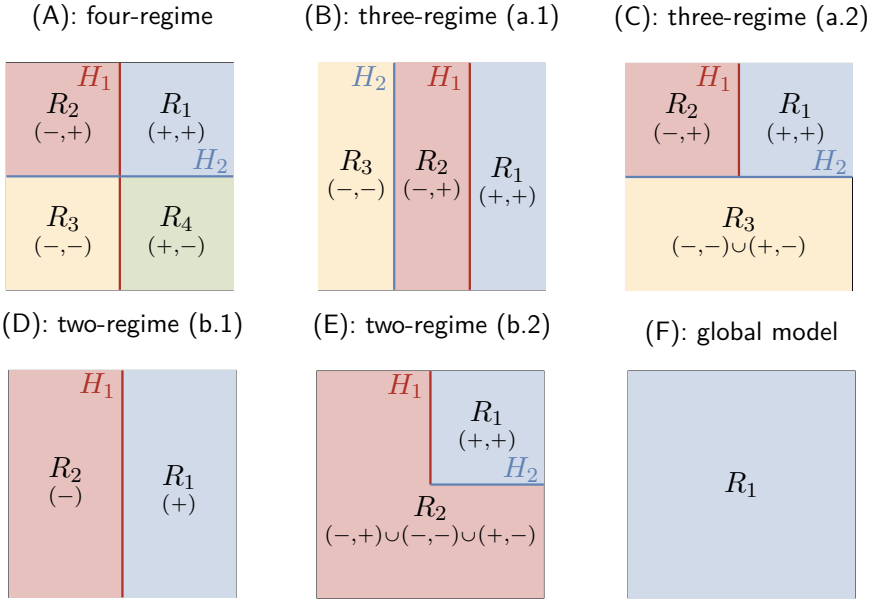


Figure: Segmented models with no more than four regimes. The signs of $(\boldsymbol{z}_1^{\mathrm{T}}\boldsymbol{\gamma}_1, \boldsymbol{z}_2^{\mathrm{T}}\boldsymbol{\gamma}_2)$ for each region are indicated below the region names.

Question: How will LSE with $L = 2, K = 4$ behave in the degenerated cases with $L \leqslant 2, K < 4$?

Properties under Degenerated Models

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. For a set $\mathcal{H} = \{\boldsymbol{h}_j\}_{j=1}^J$ and a vector $\boldsymbol{v}$, let $d(\boldsymbol{v}, \mathcal{H}) = \min_j \|\boldsymbol{v} - \boldsymbol{h}_j\|_2$.

Theorem 6

For degenerated models with K_0 regimes and L_0 splitting hyperplanes, where $1 \leqslant K_0 < 4$ and $0 \leqslant L_0 \leqslant 2$ under Assumption 1 and Assumptions S2-S4 in the SM, which adapt Assumptions 3–4 to the degenerate model settings, then for each $\boldsymbol{\beta}_{k0}$ with $1 \leqslant k \leqslant K_0$, $d(\boldsymbol{\beta}_{k0}, \widehat{\mathcal{B}}) = O_p(1/\sqrt{T})$. 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 $1 \leqslant k \leqslant K_0$.

- The estimated boundaries and the regression coefficients obtained under (4) of the four-regime model are consistent to the true parameters of degenerated models.
- Each true region asymptotically equals to a estimated region or a union of some estimated regions.

Backward elimination fitting

Starting from the estimated four-regime model, we recursively find the best pairs of adjacent regimes to be merged, under a criterion that the merging leads to the minimal increase in the fitting errors.

For $K = 4, 3, 2$, recursively define

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

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 $\widehat{R}_{\widehat{i}}^{(K)}$ and $\widehat{R}_{\widehat{h}}^{(K)}$ if

$$(\widehat{i}, \widehat{h}) = \underset{(i,h)\in\mathcal{A}_K}{\arg\min}\, D_T^{(K)}(i,h), \tag{10}$$

followed by labelling 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 $\{\widehat{\boldsymbol{\beta}}_k^{(K-1)}\}_{k=1}^{K-1}$.

Model selection

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

$$\widehat{K} = \underset{1 \leqslant K \leqslant 4}{\arg\min} \{\log(\frac{S_T(K)}{T}) + \frac{\lambda_T}{T} K\} \tag{11}$$

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

Theorem 7

Under the assumptions of Theorem 6, and $\lambda_T \to \infty, \lambda_T/T \to 0$ *as* $T \to \infty$, *then* $\widehat{K}$ *selected in* (11) *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* $k \in \{1, \cdots, K_0\}$.

Simulation: Four-Regime Model

TABLE 1

Empirical average estimation errors $\|\gamma_0 - \widehat{\gamma}\|_2$ and $\|\beta_0 - \widehat{\beta}\|_2$ (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					
	$\psi = 0$		$\psi = 0.2$		$\psi = 0.4$		$\psi = 0.8$		$\psi = 0.2$		$\psi = 0.4$		$\psi = 0.8$	
	$\widehat{\gamma}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\beta}$	$\widehat{\gamma}$	$\widehat{\beta}$
200	0.94	6.68	0.92	6.66	0.88	6.43	0.88	5.9	0.93	6.63	0.9	6.49	0.85	6.14
	(0.59)	(1.7)	(0.58)	(1.68)	(0.6)	(1.56)	(0.61)	(2.24)	(0.56)	(1.63)	(0.54)	(1.66)	(0.52)	(1.8)
400	0.45	4.55	0.45	4.55	0.45	4.4	0.43	3.98	0.44	4.46	0.43	4.38	0.43	4.06
	(0.3)	(1.1)	(0.3)	(1.11)	(0.27)	(1.17)	(0.29)	(1.53)	(0.28)	(1)	(0.33)	(1.07)	(0.28)	(1.21)
800	0.25	3.11	0.24	3.09	0.22	2.97	0.22	2.64	0.23	3.11	0.25	3.03	0.22	2.81
	(0.16)	(0.66)	(0.15)	(0.66)	(0.14)	(0.66)	(0.14)	(0.96)	(0.14)	(0.66)	(0.16)	(0.65)	(0.15)	(0.72)
1600	0.11	2.2	0.11	2.18	0.12	2.11	0.11	1.88	0.11	2.17	0.11	2.11	0.11	1.97
	(0.07)	(0.46)	(0.07)	(0.47)	(0.08)	(0.5)	(0.07)	(0.77)	(0.07)	(0.45)	(0.07)	(0.47)	(0.07)	(0.54)

Simulation: model selection

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)

Simulation: smoothed regression bootstrap

TABLE 3

Empirical coverage probabilities and widths (×100 in parentheses) of the 95% confidence intervals for five projected parameters $\{\gamma_0^{\mathrm{T}} \boldsymbol{d}_i\}_{i=1}^5$ 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)

Case Study

Goal: To explore the relationship between meterological variables and $PM_{2.5}$.

- **Response:** $PM_{2.5}$;
  Covariates: temperature (TEMP), dew point temperature (DEWP), pressure (PRES), cumulative wind speed (IWS), wind direction (NE, NW, SE, SW, CV), boundary layer height (BLH) and the one-hour lagged PM2.5 term (Lag).
- **Time range:** January 1, 2019 to December 31, 2019.
- **Testing data:** 11-th day to the 20-th day of each month;
  Training data: the rest samples of each month.
- **Comparison models:**
  1. the global linear regression (GLR);
  2. the two-regime model (2-REG) of Lee et al. (2021), Yu and Fan (2021);
  3. the four-regime model (4-REG);
  4. the linear regression tree (LRT) of Breiman et al. (1984);
  5. the multivariate adaptive regression splines (MARS) of Friedman (1991)

Case Study: Model Comparision

Fig 2: Mean squared errors (MSE) for $PM_{2.5}$ 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.

Average rank of RMSE. The 4-REG achieved the best out-of-sample performances.

	GLR	2-REG	4-REG	LRT	MARS
Train	5	3.75	2.5	1.75	2
Test	3	2	1	4.75	4.25

Case Study: Splitting Boundaries

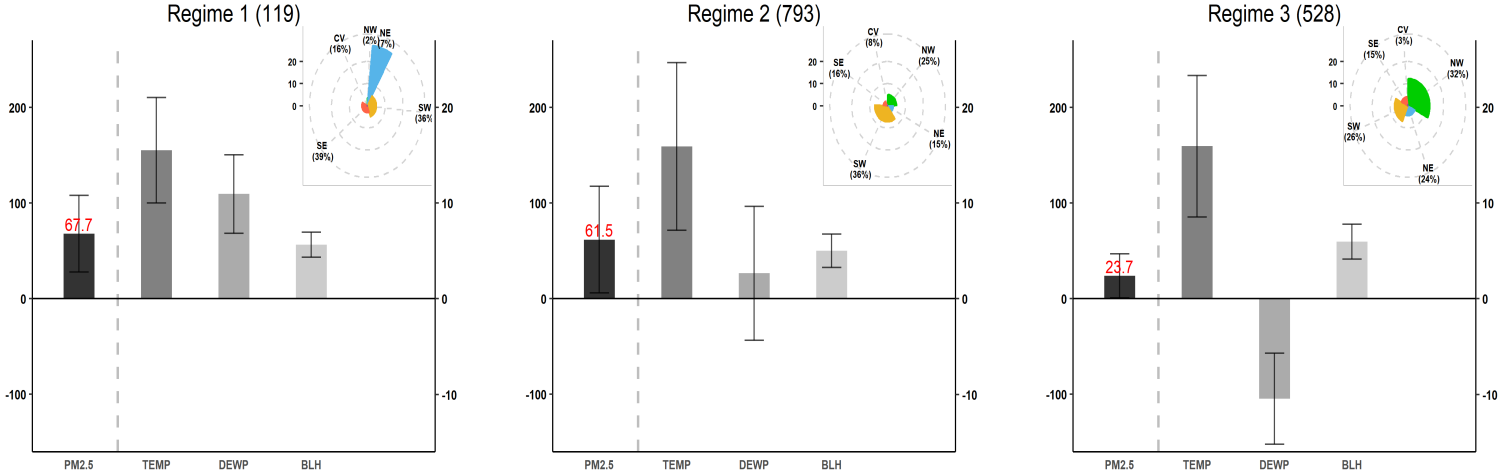
(a) Spring

Table 4: Estimated coefficients of the splitting boundaries and cos of the angle ϕ between the two boundaries. The coefficients were normalized such that 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	

- The DEWP and the wind-related variables were the most important variables in determining the estimated boundaries.
- The splitting boundaries are determined empirically by multivariate covariates, while the boundary variable has to be user-specified in classic threshold regression.

Case Study: Regime Splitting in Spring



The data-driven regime-splitting results had clear **atmosphere physics interpretation**:

Regime 1: high DEWP & CV $\Longrightarrow$ pollution state; Regime 2: lower DEWP & reduced CV $\Longrightarrow$ transitional state; Regime 3: lowest DEWP & mainly northerly wind $\Longrightarrow$ cleaning state

Summary

The segmented regression model is an intermediate between the global and local model, but has been underexplored. In this work, we contribute to this area in several aspects:

- propose a more flexible class of segmented models that extends threshold/change point regression;
- the asymptotics under weakly dependent data;
- smoothed regression bootstrap for inference;
- model selection with consistency guarantee;
- MIQP-based computation method.

Future interests:

- nonlinear splitting boundaries;
- high-dimensional segmented regressions;
- extension to tree-based regressions;
- . . .

Thank you!

References

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Routledge.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.*, 19(1):1–141. With discussion and a rejoinder by the author.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica*, 64(2):413–430.

Knight, K. (1999). Epi-convergence and stochastic equisemicontinuity. Preprint.

Lee, S., Liao, Y., Seo, M. H., and Shin, Y. (2021). Factor-driven two-regime regression. *Ann. Statist.*, 49(3):1656–1678.

Li, D. and Ling, S. (2012). On the least squares estimation of multiple-regime threshold autoregressive models. *J. Econometrics*, 167(1):240–253.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press.

Yu, P. and Fan, X. (2021). Threshold regression with a threshold boundary. *Journal of Business & Economic Statistics*, 39(4):953–971.