MULTIPLE CHANGE POINTS DETECTION IN HIGH-DIMENSIONAL DATA

Modhurima Dey Amin*, Texas Tech University and Abhishek Kaul, Washington State University

AEA ASSA Annual Meetings 2021

Motivation

- Consider a linear regression model
  $y_i = x_i^T\beta + \epsilon_i$ where $i = 1, ..., n$ and β is a p dimensional vector
- Assuming "one set of parameters fits all observations" can be misleading because the relationship between the predictor and the response variables may differ over time and space.
- A high dimensional (HD) change point (CP) regression model is,
  $y_i = \sum_{j=1}^{N+1} x_i^T \beta^0_{(j-1)} \mathbf{1}[\tau^0_{j-1} < w_i \leq \tau^0_j] + \epsilon_i$
  where, $j = 1, ..., N \geq 0$ are CPs; τ is the location of CP such that $\tau^0_0 = -\infty; \tau^0_{N+1} = \infty$; w is the change-inducing variable; and $\mathbf{1}[.]$ is an indicator function.
- Studies so far discussed detection of CPs
  - Information based criteria, but not extendable to HD (Wu, 2008)
  - Single change point in HD (Lee et al., 2016, 2018)
  - Multiple change point with fixed regression parameters (Zhang et al., 2015; Leonardi and Buhlmann, 2016)
  - Single change point without grid search in HD (Kaul et al., 2019a)
  - Multiple change points with arbitrary segmentation (Kaul et al., 2019b)

This Study

- Extends the algorithm of Kaul et al. (2019a) to detect multiple change points via sequential binary segmentation based on l_1/l_0 regularization—where a change-inducing variable may switch the regression parameters.
- Which method performs better in HD model: binary versus arbitrary segmentation?

Binary Segmentation (Our Model)

Binary Segmentation Algorithm (Binseg):

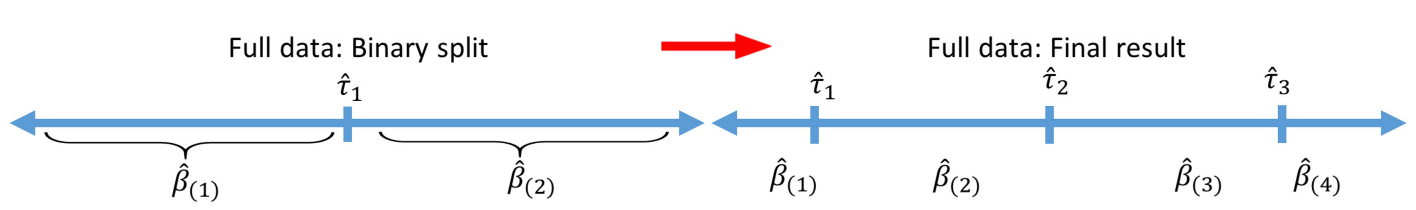
- Step 0: (Initialize) Choose any initial value $\tau^{(0)} \in \mathbb{R}$ and compute the initial regression parameter estimates such that $\lambda_1 > 0$

$$\left(\hat{\beta}_1^0, \hat{\beta}_2^{(0)}\right) = \arg\min_{\beta,\gamma} \left\{ Q(\tau^{(0)}, \beta_1, \beta_2) + \lambda_1 \left\| (\beta_1^T, \beta_2^T)^T \right\|_1 \right\}$$

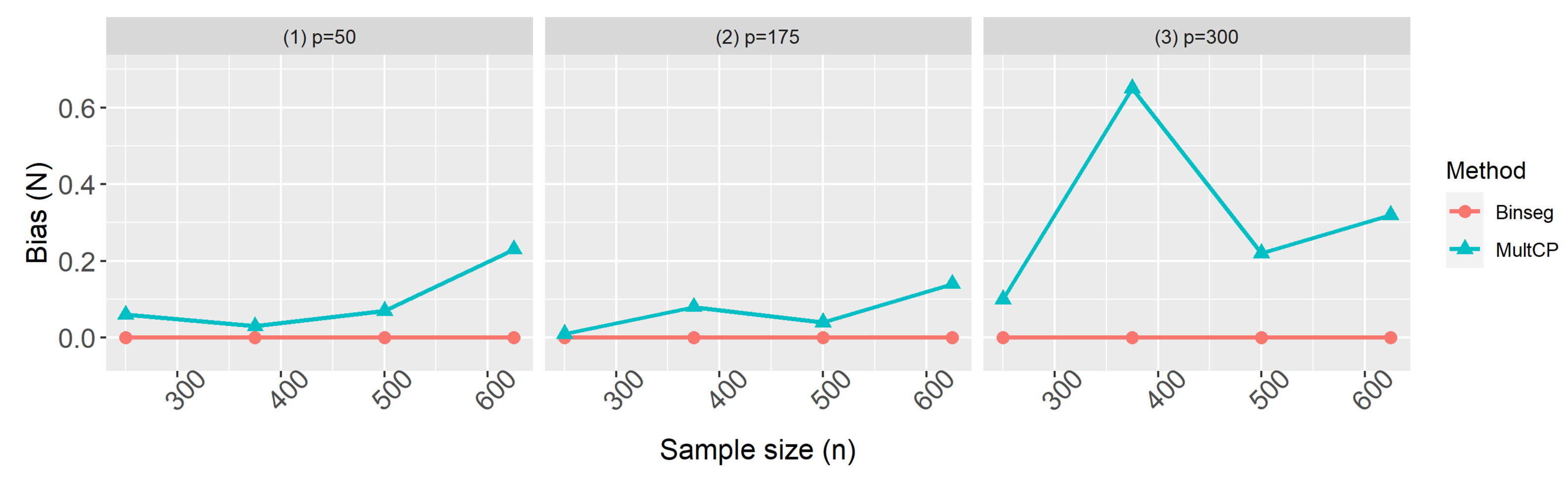
- Step 1: Update $\tau^{(0)}$ to obtain the CP estimate $\hat{\tau}^{(1)}$ such that $\mu > 0$

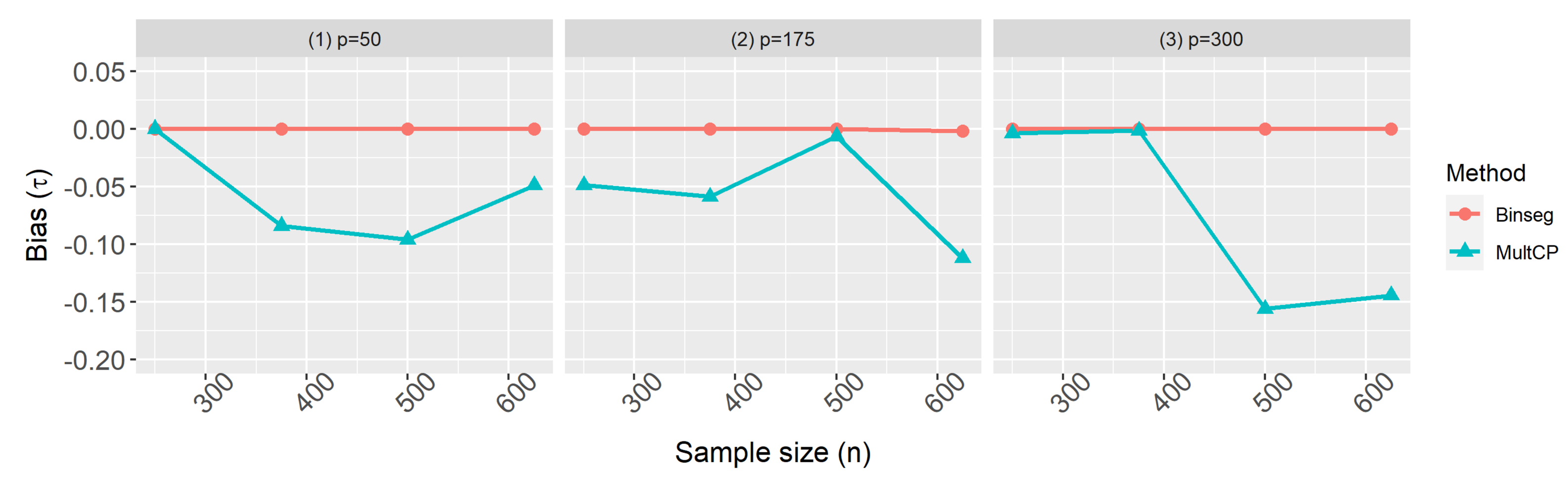
$$\hat{\tau}^{(1)} = \arg\min_{\tau} \left\{ Q\left(\tau, \hat{\beta}_1^{(0)}, \hat{\beta}_2^{(0)}\right) + \mu \|\phi(\tau)\|_0 \right\}$$

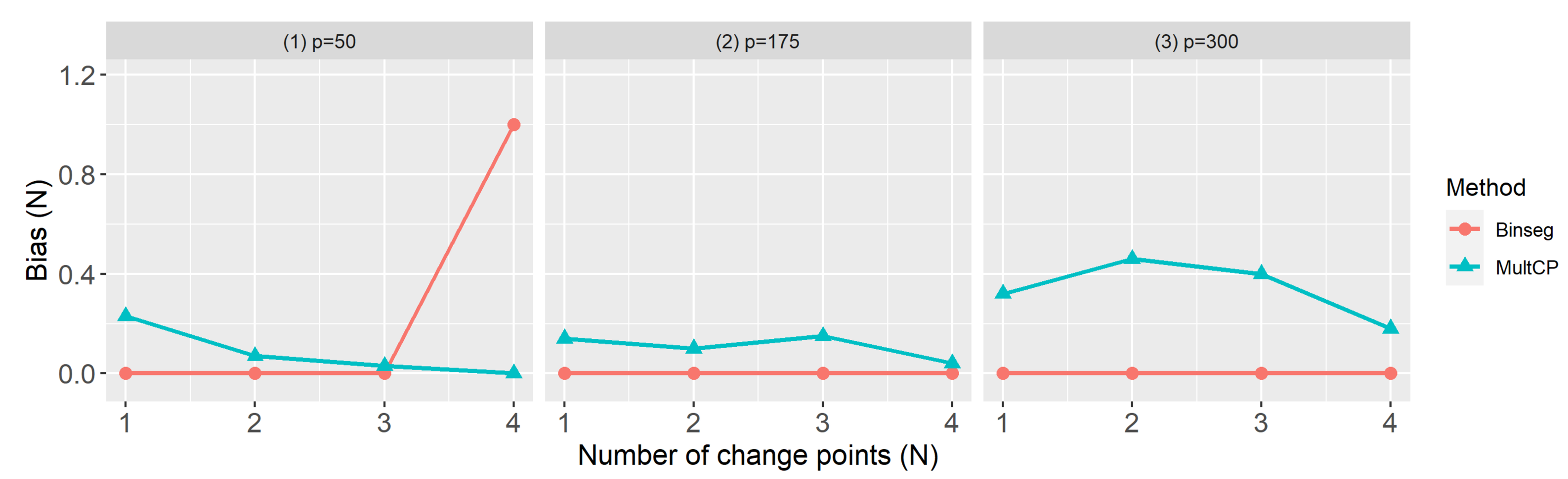
- Step 2: Update $(\hat{\beta}_1^{(0)}, \hat{\beta}_2^{(0)})$ to obtain $(\hat{\beta}_1^{(1)}, \hat{\beta}_2^{(1)})$
- Step 3: Repeat in each segment until no further change point is found.

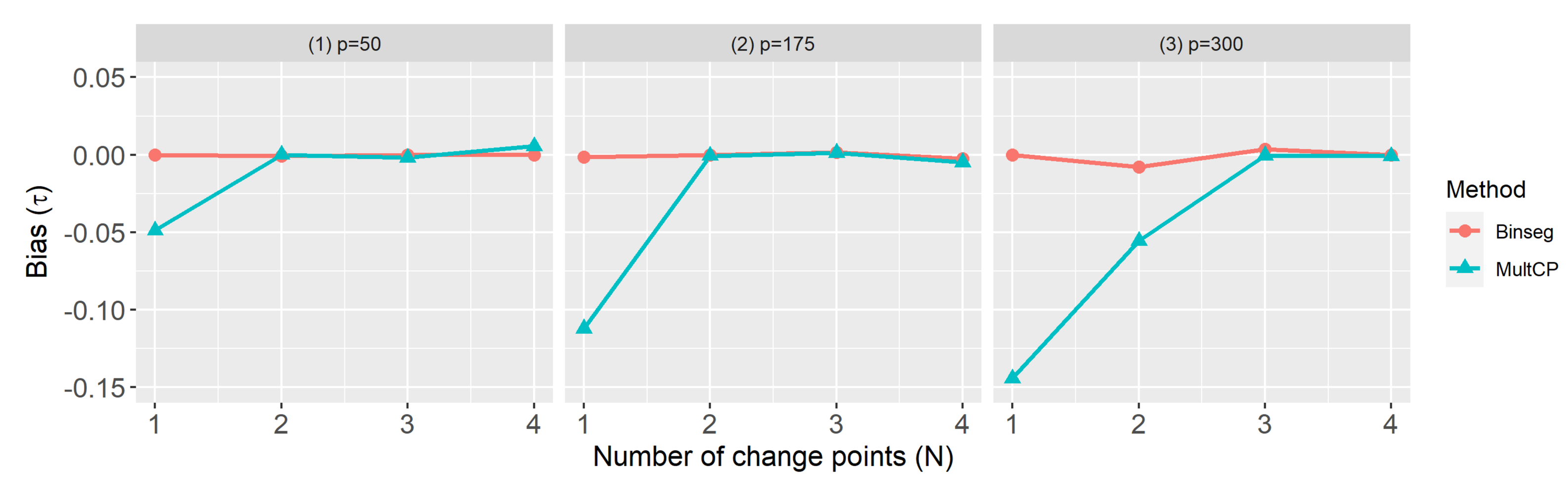


Performance: Binseg vs. MultCP









Simulation Findings

- Binseg improves over MultCP in detecting the true number of change points N and their locations τ (=the threshold values of w).
- Binseg tends to produce lower bias in τ and N regardless of the size of n, N, and p.
- The bias in MultCP estimates generally increases with p and n but decreases with N.

Contact Information: Modhurima Dey Amin, Ph.D.
Assistant Professor, Texas Tech University
Email: Modhurima.Amin@ttu.edu

Arbitrary Segmentation (Benchmark Model)

Arbitrary segmentation algorithm (MultCP):

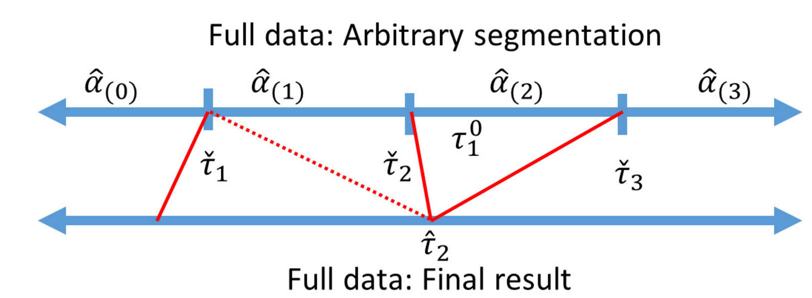
- Step 0: Choose any number of change points $\check{N} \geq 1 \vee N$ and any vector $\check{\tau} = \left(\check{\tau}_1, \ldots, \check{\tau}_{\check{N}}\right)^T \in \mathbb{R}^{\check{N}}$. Compute initial regression estimates $\hat{\alpha}_{(j)}$, for each $j = 0, \ldots, \check{N}$, where $\lambda_0 > 0$

$$\hat{\alpha}_j = \arg\min_{\alpha} \left\{ \frac{1}{n} Q^*\left(\alpha, \check{\tau}_{j-1}, \check{\tau}_j\right) + \lambda_0 \|\alpha\|_1 \right\}$$

- Step 1: Update $\check{\tau}$ to obtain estimate $\hat{\tau}$ such that $\mu > 0$.

$$\hat{\tau} = \arg\min_{\tau} \left\{ Q\left(\check{N}, \hat{\alpha}, \tau\right) + \mu \sum_{j=1}^{\check{N}} \left\| d\left(\tau_{j-1}, \tau_j\right) \right\|_0 \right\}$$

- Step 2: Let $\hat{\mathcal{T}} := \hat{\mathcal{T}}(\hat{\tau})$ and update the estimated number of change points to $\tilde{N} = |\hat{\mathcal{T}}|$, and recover the corresponding locations of change points as the subset $\hat{\mathcal{T}}$.



Empirical Application

- Does the relationship between violent crimes and socio-economic factors is influenced by a change-inducing variable?
- Cross-sectional data on communities and crime in the United States (Census, LEMAS survey, FBI: 1995). After pre-processing, $n = 319$ communities and $p = 77$ predictors.
- $y =$ per capita violent crimes in a community, which was calculated as the sum of violent crimes including murder, rape, robbery, and assault. $x =$ socio-demographic and law and order variables, e.g., median income, law enforcement, per capita number of police officers, % of officers assigned to drug units, working population, homeless population.

Table: Summary of empirical estimation

	Mean	SD	$\hat{N}$	$\hat{\tau}_0$	$\hat{\tau}_1$	$\hat{\tau}_2$
Dependent variable (y)						
Total violent crimes per 100k population	0.238	0.233				
Change-inducing variable (w)						
Median household income	0.361	0.209	0	NA		
Population of the community	0.058	0.127	1	0.23		
Households with social security income (%)	0.471	0.174	2	0.47	0.63	
Mean persons per rental household	0.404	0.189	3	0.32	0.36	0.41

- MultCP does not detect any change point.
- Binseg detects one, two and three change points in the parameters induced by population, social security, and rental density respectively.
- Policies aimed at crime reduction should undertake separate measures across communities based on the estimated thresholds.