0. 論文

Electronic Journal of Statistics
Vol. 10 (2016) 2000–2038
ISSN: 1935-7524
DOI: 10.1214/16-EJS1155

Change-point detection in panel data via double CUSUM statistic

Haeran Cho

School of Mathematics, University of Bristol, UK
e-mail: haeran.cho@bristol.ac.uk

タイトル：Change-point detection in panel data via double CUSUM statistic 著者： Haeran Cho

arXiv投稿日： 2016/11/25 学会/ジャーナル：Electronic Journal of Statistics (2016)

1. どんなもの？

- CUSUM統計量を用いたパネルデータの変化点検出手法
- バイナリーセグメンテーションを用いた手法
    - 必ず収束することが保証されている
    - ブートストラップ法を用いる
        - 検定手法の一つ，近似分布を用いて検定を行う
- 扱うモデルとしては何らかの平均値$f_{j,t}$に対してノイズが載る形を仮定している
- j番目の系列のt時点での観測は以下の形

$$x_{j,t} = f_{j,t} + \epsilon_{j,t}$$

- CUSUM統計量のn系列の集合から変化点を検索することによって、同時にn次元データをセグメントする

2. 先行研究

- CUSUM統計量を用いた変化点検出手法
- 基本的なCUSUM統計量は以下の形で書ける $$\mathcal{X}_{s,b,e}^j = \frac{1}{\sigma_j}\sqrt{\frac{(b-s+1)(e-b)}{e-s+1}}\left(\frac{1}{b-s+1}\sum_{t=s}^b x_{j,t} - \frac{1}{e-b}\sum_{t=b+1}^e x_{j,t}\right)$$
- s:始点，b:変化時点，e:終点
- 変化点が一つあるものとして導出

$$\mathcal{T}_{1,T}^{\mathrm{HH}} = \max_{b\in[1,T)} \frac{1}{\sqrt{n}} \frac{b(T-b)}{T^2} \sum_{j=1}^{n} \{(\mathcal{X}_{1,b,T}^j)^2 - 1\},$$

    -
- 断面的に疎な変化点を検出する

$$\mathcal{T}_{1,T}^{\mathrm{scan}} = \max_{b\in[1,T)} \max_{1\le m\le n} \frac{1}{T_m\sqrt{2m}} \sum_{j=1}^{m} \{(\mathcal{X}_{1,b,T}^{(j)})^2 - 1\},$$

- 時間的依存と系列間の依存を考慮した手法

$$\mathcal{T}_{1,T}^{\mathrm{Jirak}} = \max_{b\in[1,T)} \max_{1\le j\le n} \sqrt{\frac{b(T-b)}{T}} |\mathcal{X}_{1,b,T}^{j}|,$$

  - これはガンベル分布，もしくはブートストラップの極値によって計算される閾値と検定統計量が比較される
- 上記の手法のほとんどはCUSUMの最大値，和のいずれかを取るのでCUSUM統計値の構造に適応していない
  - 高次元設定に置ける変化点検出では劣った性能をもたらす可能性あり
- 閾値化したCUSUM統計量の使用

$$\mathcal{T}_{1,T}^{\mathrm{SBS}}(\pi_T) = \max_{b\in[1,T)} \sum_{j=1}^{n} |\mathcal{X}_{1,b,T}^{j}| \cdot \mathbb{I}(|\mathcal{X}_{1,b,T}^{j}| > \pi_T)$$

  - 設定するべきハイパーパラメータの量が大きすぎることが問題

3. コアアイディア

- 以下の問題を最適化するように変化点検出を行う

$$\widehat{\eta} = \arg \max_{b\in[s,e)} \max_{1\le m\le n} \mathcal{D}_m^{\varphi}(\{|\mathcal{X}_{s,b,e}^{(j)}|\}_{j=1}^{n}).$$

- $$\mathcal{D}_m^{\varphi} = \left\{\frac{m(2n-m)}{2n}\right\}^\varphi \frac{1}{m}\sum_{j=1}^{m}\left(|\mathcal{X}_{s,b,e}^{(j)}| - \frac{1}{2n-m}\sum_{j=m+1}^{n}|\mathcal{X}_{s,b,e}^{(j)}|\right)$$
- ただし，$|\mathcal{X}^{(j)}|$は大きい順にソートされているものとする
- 一例

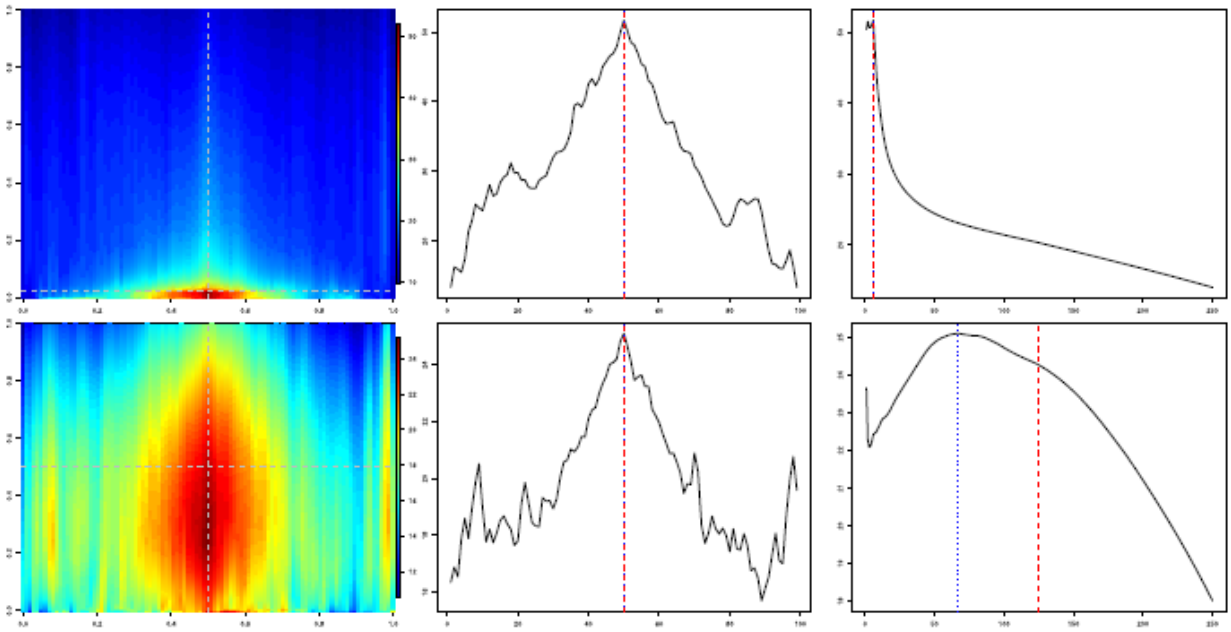

FIG 1. $(m_1, \delta_1) = ([\log n], 0.24)$ *(top) and* $(m_1, \delta_1) = ([0.5n], 0.05)$ *(bottom); the heat map of* $\mathcal{D}_m^{\varphi}(\{|\mathcal{X}_{1,b,T}^{(j)}|\}_{j=1}^n)$ *over* b *(x-axis) and* m *(y-axis) (left), the pointwise maximum of* $\mathcal{D}_m^{\varphi}(\{|\mathcal{X}_{1,b,T}^{(j)}|\}_{j=1}^n)$ *over* m *for* $b \in [1, T)$*, with broken lines indicating* η_1 *and the dotted ones* $\widehat{\eta}_1$ *(middle), and* $\mathcal{D}_m^{\varphi}(\{|\mathcal{X}_{1,\widehat{\boldsymbol{\eta}}_1,T}^{(j)}|\}_{j=1}^n)$*,* $1 \le m \le n$ *with broken lines indicating* m_1 *and dotted ones* $\widehat{m}_1$*.*

-

4. どうやって有効だと検証した？

- ARMAモデルで生成したもの，複雑なノイズを載せたものの二つで評価を行う

TABLE 2
Type I error when $\alpha = 0.05$*;* $n = 100$ *(top) and* $n = 250$ *(bottom).*

		$T = 100$						$T = 250$					
	ϱ/ϱ_h	$\mathcal{T}^0$	$\mathcal{T}^{1/2}$	$\tilde{\mathcal{T}}$	$\mathcal{T}^{\text{Jirak}}$	$\mathcal{T}^{\text{EH}}$	$\mathcal{T}^{\text{SBS}}$	$\mathcal{T}^0$	$\mathcal{T}^{1/2}$	$\tilde{\mathcal{T}}$	$\mathcal{T}^{\text{Jirak}}$	$\mathcal{T}^{\text{EH}}$	$\mathcal{T}^{\text{SBS}}$
(N1)	0.2	0.06	0.05	0.06	0.15	0.08	0	0.01	0.07	0.01	0.11	0.1	0
	0.5	0.04	0.03	0.04	0.19	0.08	0	0.02	0.06	0.02	0.14	0.11	0
(N2)	0.5	0.08	0.04	0.07	0.13	0.16	0	0.01	0.08	0.01	0.18	0.23	0
	0.9	0.04	0.04	0.04	0.1	0.57	0	0.04	0.05	0.05	0.05	0.64	0
(N1)	0.2	0.06	0	0.06	0.15	0.07	0	0.07	0.03	0.07	0.18	0.09	0
	0.5	0.06	0.01	0.04	0.21	0.09	0	0.12	0.01	0.1	0.22	0.05	0
(N2)	0.5	0.07	0.04	0.05	0.04	0.35	0	0.02	0.05	0.05	0.08	0.34	0
	0.9	0.04	0.07	0.07	0.1	0.61	0	0.05	0.05	0.05	0.06	0.75	0

  -
  - type I errorの比較
  - 提案法(左3つ)が有意水準に押さえることに成功しているのに対し，その他の手法は制御できていない
  - 一般的にサンプル数(おそらく系列長)が大きくなると検出力は増す
  - 多重変化点検出時でも同様の結果が得られた
- S&P100データセット
  - 系列数88, 系列長252のデータセット
  - このデータの対数スケールは平均ゼロの分散構造が変化するデータセット仮定されることが多い
    - 条件付き異種混合モデルを用いて条件付き分散を持つもの
  - 単一の変化点の検出ができた
    - 過去最悪のスタートをマークしたものの検出を行っている(大きく下落したことが示唆されている)

    - 何らかの正当性はあるだろう

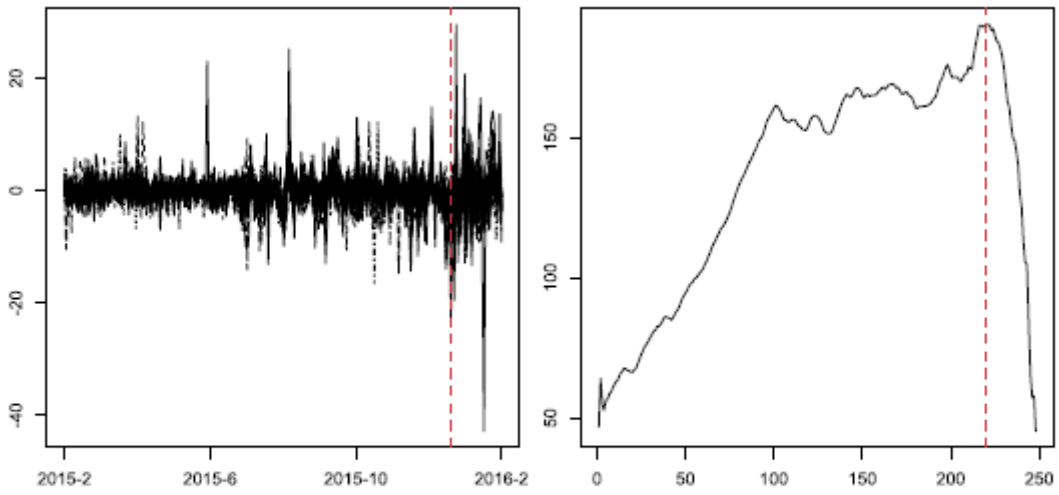

FIG 4. *Log returns of S&P 100 component stock prices ($y_{i,t}$) between February 24, 2015 and February 23, 2016 (left); pointwise maximum of $\widetilde{\mathcal{D}}_m(\{|\mathcal{X}_{1,b,T}^{(j)}|\}_{j=1}^n)$ over $b = 1,\ldots,T-1$ (right); the vertical broken line denotes the location of the estimated change-point.*

5. データセット

- S&P100データセット

6. 疑問点

- 変化点検出+統計的有意性の評価を行っている
    - 近似分布を使っていることに注意
    - 何らかの問題が生じる(系列数を増やすと問題があるとか？)

7. 次に読むべき論文は？

- CUSUM統計量を用いた多重変化点検出
    - Multiscale and multilevel technique for consistent segmentation of nonstationary time series

キーワード

- CUSUM統計量
- 多次元系列
- 変化点検出
- 統計的仮説検定
- ブートストラップ
- バイナリセグメンテーション