Robust multiscale estimation of time-average variance for time series segmentation

StatScale ECR Event 2022

Euan T. McGonigle[1]
Joint work with Haeran Cho[1]

15th December 2022

[1] School of Mathematics, University of Bristol
euan.mcgonigle@bristol.ac.uk

https://euanmcgonigle.github.io

Motivation

- Mean change point detection (CPD) is a well-studied area of statistics with a vast volume of recent literature.
- A Gaussian i.i.d. error assumption is common, which is unlikely to hold in practice.
- Less attention given to the case of autocorrelated errors, although growing literature.
- I'll discuss a robust approach to noise estimation that can be used with multiscale CPD algorithms.

The change in mean problem

We consider the univariate change in mean model:

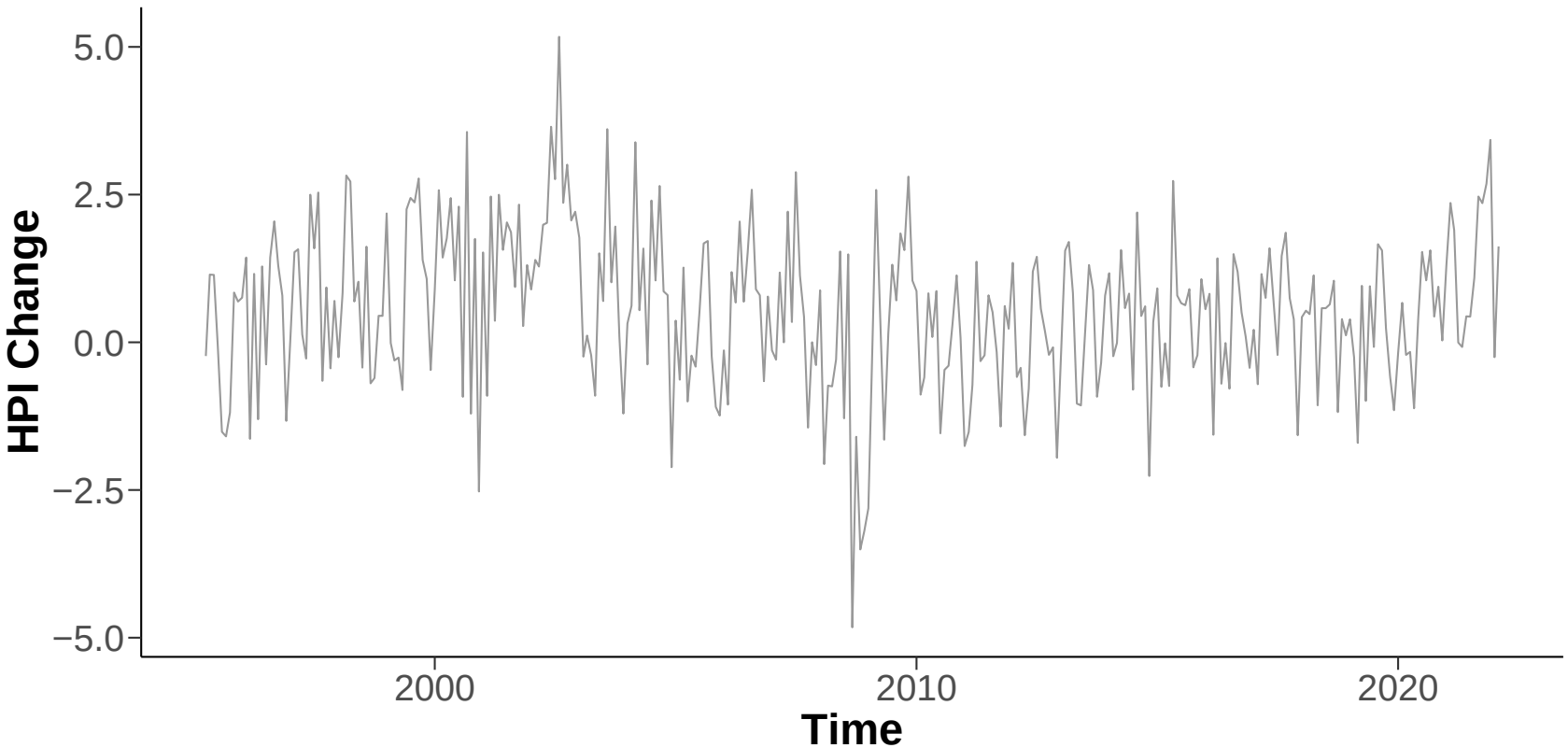
$$X_t = f_t + \varepsilon_t = f_0 + \sum_{i=1}^{q} \mu_i \mathbb{I}(t \geq \tau_i + 1) + \varepsilon_t, \qquad 1 \leq t \leq n.$$

- f_t is piecewise constant with q change points $\tau_1, \ldots, \tau_q$.
- The errors $\{\varepsilon_t\}_{t=1}^n$ are weakly stationary with $\mathbb{E}(\varepsilon_t) = 0$.
- $\{\epsilon_t\}_{t=1}^n$ can be serially correlated and heavy-tailed.

Aim: estimate locations and number of changepoints.

Motivation – monthly house price changes

- Monthly house price percentages in Taunton and Somerset.
- Mean changes and/or autocorrelation?



Detecting changes via localised testing

Local approaches for CPD:

- Scan data for candidate change point estimators.
- Typically involve computing a test statistic of the form $\widehat{\sigma}_{s,e}^{-1}|\mathcal{T}_{s,k,e}|$, where

$$\mathcal{T}_{s,k,e} = \sqrt{\frac{(k-s)(e-k)}{e-s}}\left(\frac{1}{k-s}\sum_{t=s+1}^{k} X_t - \frac{1}{e-k}\sum_{t=k+1}^{e} X_t\right),$$

  and $\widehat{\sigma}_{s,e}$ is a estimator of $\sigma_{s,e}$, a measure of variability of $\{X_t\}_{t=s}^{e}$.
- Are often multiscale in nature.

Detecting changes via localised testing

Local approaches:

- Compute $\mathcal{T}_{s,k,e}$ over a range of intervals in a method-specific way.
- Often compare $\widehat{\sigma}_{s,e}^{-1}|\mathcal{T}_{s,k,e}|$ to a threshold D to test for changes.
- Threshold is often theoretically motivated.
- Examples include multiscale MOSUM, WBS, SBS, ...

Key challenge: separating genuine change points from fluctuations due to noise by careful selection of the estimator $\widehat{\sigma}_{s,e}^2$.

Main idea: use a multiscale estimator of variability that is more suitable for multiscale change point algorithms.

Scale-dependent variability

In multiscale methods, a (single) estimator of the **long-run variance**

$$\sigma^2 = \lim_{n\to\infty} \text{Var}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\varepsilon_t\right)$$

is often used to estimate the noise level $\sigma^2_{s,e}$.

We instead estimate $\sigma^2_{s,e}$ using a scale-dependent measure of variability; the **time-average variance constant (TAVC)**:

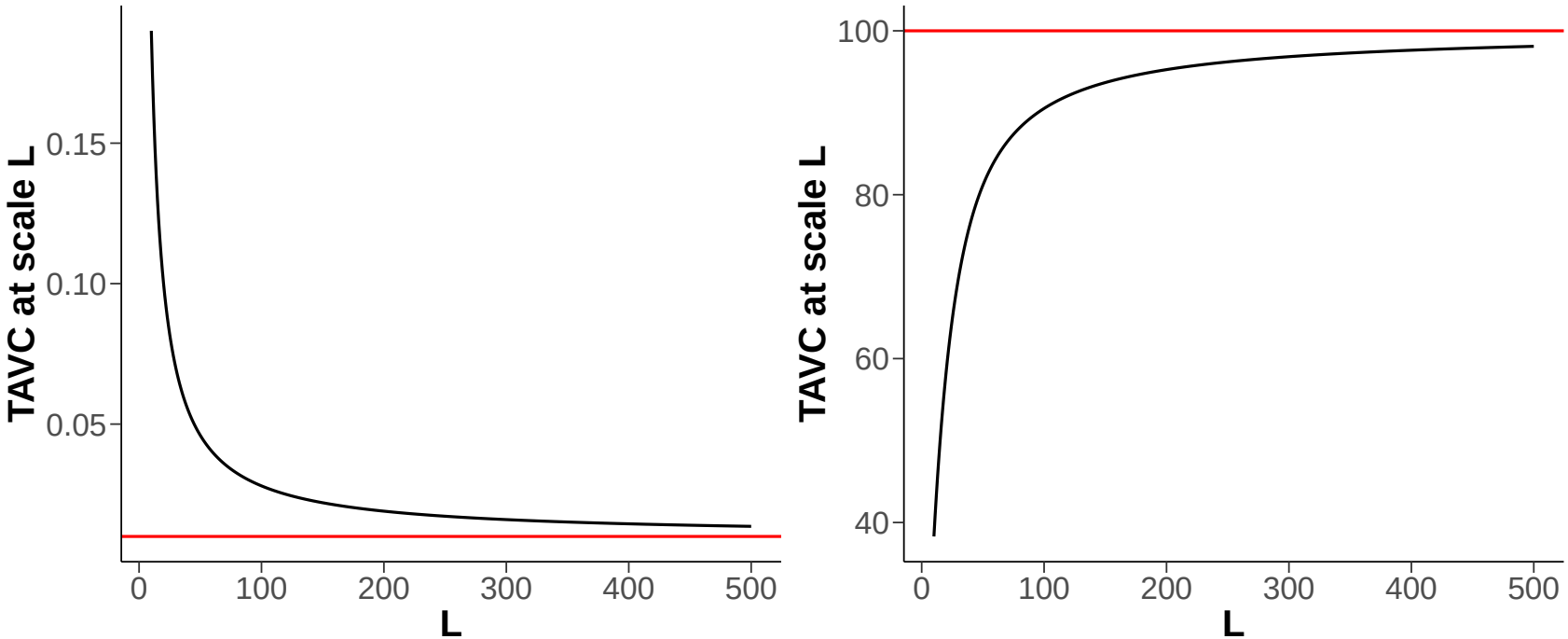
$$\sigma^2_L = \text{Var}\left(\frac{1}{\sqrt{L}}\sum_{t=1}^{L}\varepsilon_t\right),$$

for a given scale $L = e - s$.

Large gap between σ^2 and σ_L^2

Left: MA(1) process $\varepsilon_t^{(1)} = W_t - 0.9W_{t-1}$, $W_t \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$.

Right: AR(1) process $\varepsilon_t^{(2)} = 0.9\varepsilon_{t-1}^{(2)} + W_t$, $W_t \sim_{\text{i.i.d.}} \mathcal{N}(0,1)$.



Consequence: false positives (left) or false negatives (right).

Robust estimation of multiscale TAVC

Let $G = L/2$ denote block size. For starting point $b \in \{0, \ldots, G-1\}$ with $N_1(b) = \lfloor (n-b-G)/G \rfloor$ blocks, define

$$\bar{X}_{j,b} = \frac{1}{G} \sum_{t=jG+b+1}^{(j+1)G+b} X_t, \quad \text{and} \quad \xi_{j,b} = \frac{G(\bar{X}_{j,b} - \bar{X}_{j-1,b})^2}{2},$$

for $j = 1, \ldots, N_1(b)$.

If there were no mean changes, the block-based average

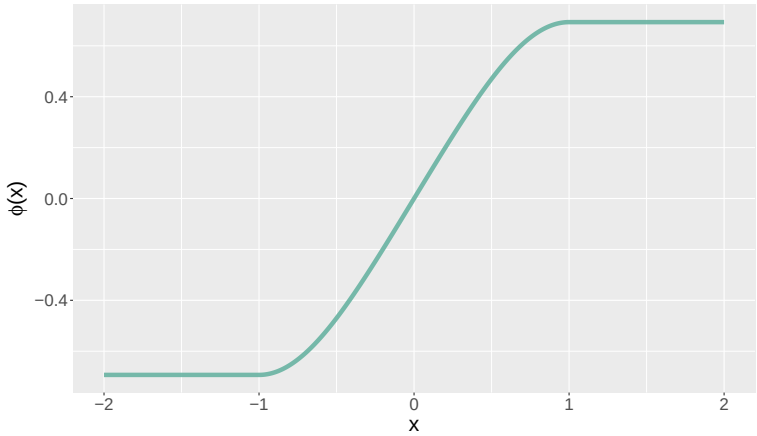
$$\widetilde{\sigma}_L^2 = \frac{1}{N_1(b)} \sum_{j=1}^{N_1(b)} \xi_{j,b}$$

could be used to estimate σ_L^2.

TAVC estimation – influence function approach

To deal with mean changes, we use an **influence function** approach[1]:

$$\phi(x) = \begin{cases} -\log(2) & x \leq -1, \\ \log(1 + x + x^2/2) & -1 \leq x \leq 0, \\ -\log(1 - x + x^2/2) & 0 \leq x \leq 1, \\ \log(2) & x \geq 1. \end{cases}$$



The function $|\phi|$ is bounded above by $\log(2)$, and ϕ satisfies

$$x - x^2/2 \leq \phi(x) \leq x + x^2/2,$$

which allows us to control the influence of the change points.

[1] Catoni (2012). Challenging the empirical mean and empirical variance: A deviation study.
Chen, Wang, Wu. (2021). Inference of breakpoints in high-dimensional time series.

TAVC estimation – influence function approach

The robust estimator $\widehat{\sigma}^2_{L,b}$ of the TAVC at scale L and starting point b is given by the solution of the M-estimation equation

$$h_{L,b}(u) = \frac{1}{vN_1(b)} \sum_{j=1}^{N_1(b)} \phi\left(v(\xi_{j,b} - u)\right) = 0,$$

where $v > 0$ is loosely speaking a scaling factor.

- If there are multiple solutions, any of them may be chosen.
- Can be extended heuristically to nonstationary noise setting.

TAVC estimation – theoretical analysis

Consistency of the estimator requires conditions on the noise $\{\varepsilon_t\}_{t=1}^n$.

Key assumptions:

(i) $\{\varepsilon_t\}_{t=1}^n$ has a linear representation:

$$\varepsilon_t = \sum_{k=0}^{\infty} a_k \eta_{t-k},$$

where $\{\eta_t\}_{t\in\mathbb{Z}}$ are zero-mean i.i.d. r.v.s and $|a_k| \leq \gamma(k+1)^{-\beta}$ for $\beta > 2.5$ and $\gamma > 0$ for all $k \geq 0$.

(ii) The increments $\{\eta_t\}_{t\in\mathbb{Z}}$ satisfy *either*:

(A) $\|\eta_1\|_r = (\mathbb{E}(|\eta_1|^r))^{1/r} < \infty$ for some $r > 4$ (bounded moment).

(B) For $C_\eta > 0$, $\kappa \geq 0$, $\|\eta_1\|_r \leq C_\eta r^\kappa$ for all $r \geq 1$ (sub-Weibull).

TAVC estimation – theoretical analysis

Provided that $v \asymp \sqrt{qG/n}$, for any starting point $b \in \{0, \ldots, G-1\}$, under assumption

(A) (bounded moment)

$$\left|\widehat{\sigma}^2_{L,b} - \sigma^2_L\right| = \mathcal{O}_P\left(\sqrt{\frac{Lq}{n}} + \max\left\{\left(\frac{L}{n}\right)^{\frac{r-2}{r+2}}, \sqrt{\frac{L\log(n)}{n}}\right\}\right) + \mathcal{O}(L^{-1}).$$

(B) (sub-Weibull)

$$\left|\widehat{\sigma}^2_{L,b} - \sigma^2_L\right| = \mathcal{O}_P\left(\sqrt{\frac{Lq}{n}} + \sqrt{\frac{L\log^{4\kappa+3}(n)}{n}}\right) + \mathcal{O}(L^{-1}).$$

Additionally,

$$\left|\sigma^2_L - \sigma^2\right| = \mathcal{O}(L^{-1}).$$

TAVC estimation – picking a maximum L

- Error due to bias ($\mathcal{O}(L^{-1})$ term) decreases with increasing L.
- Estimation error increases with L due to mean changes and decrease in number of available blocks.

To balance the two, set a **maximum time-scale** M when estimating σ_L^2:

- If the change point detector $\mathcal{T}_{s,k,e}$ involves $e - s \leq M$, then scale $\mathcal{T}_{s,k,e}$ with the estimator of σ_{e-s}^2, the TAVC at scale $L = e - s$.
- If $e - s > M$, use the estimator of σ_M^2 instead, which satisfies

$$|\sigma_{e-s}^2 - \sigma_M^2| = \mathcal{O}(M^{-1}).$$

TAVC estimation – setting b and v

Setting the starting point b:

- For greatest robustness, we compute $\widehat{\sigma}^2_{L,b}$ for all $b \in \{0, \ldots, G-1\}$.
- Then, we set $\widehat{\sigma}^2_L = \text{Median}(\sigma^2_{L,1}, \ldots, \sigma^2_{L,G-1})$.

Setting the scaling v:

- We set $v = v_b = \hat{v}_b^{-1}\sqrt{G/n}$,
- $\hat{v}_b$ is an estimator of variability of the $\{\xi_{j,b}\}_{j=1}^{N_1(b)}$.
- We use trimmed mean or scaled median of the $\{\xi_{j,b}\}_{j=1}^{N_1(b)}$.

TAVC estimation – extension to locally stationary noise

If the noise is nonstationary, we can instead try to estimate a **time-varying** TAVC:

$$\sigma_L^2(k) = \text{Var}\left(\frac{1}{\sqrt{L}} \sum_{t=k-G+1}^{k+G} \varepsilon_t\right).$$

- Can proceed as before, but using a moving window.
- Much more challenging to work well in practice – "doubly reduced" sample size.
- No theoretical guarantees in this setting ☹.

Combining TAVC estimator with change point algorithms

- Estimator can be combined with many multiscale CPD algorithms.
- Could be used on any threshold-based approach using interval generation and CUSUM-based test statistics.
- Consistency of $\widehat{\sigma}_L^2$ implies consistency of CPD algorithm.
- Focus on multiscale MOSUM and WBS2 here.
- Estimator can be used for other problems, e.g. variance estimation ($L = 2$), nonparametric regression.

MOSUM with TAVC estimator

- For bandwidth G, the MOSUM detector $T_G(k)$ is given by

$$\mathcal{T}_G(k) = \widehat{\sigma}_{2G}^{-1}\mathcal{T}_{k-G,k,k+G} = \frac{\widehat{\sigma}_{2G}^{-1}}{\sqrt{2G}}\left(\sum_{t=k+1}^{k+G} X_t - \sum_{t=k-G+1}^{k} X_t\right)$$

  for $k = G, \ldots, n-G$.

- $\widehat{\sigma}_{2G}^{-1}$ is computed by TAVC estimator (using $\widehat{\sigma}_M^2$ if $2G > M$).

- Changes declared as all local maximisers with $|\mathcal{T}_G(k)| > D_n(G, \alpha)$.

- Multiple bandwidths incorporated using bottom-up merging.

WBS2 with TAVC estimator

- Generate set of R intervals $\mathcal{R} = \{(s_i, e_i)\}_{i=1}^{R}$ of $\{X_t\}_{t=1}^{n}$ randomly or deterministically. Identify

$$(s_\circ, k_\circ, e_\circ) = \arg\max_{\substack{(s,k,e):\, s<k<e \\ (s,e)\in\mathcal{R}}} \frac{|\mathcal{T}_{s,k,e}|}{\widehat{\sigma}_{e-s}} \quad \text{with} \quad \frac{|\mathcal{T}_{s,k,e}|}{\widehat{\sigma}_{e_\circ - s_\circ}} > D$$

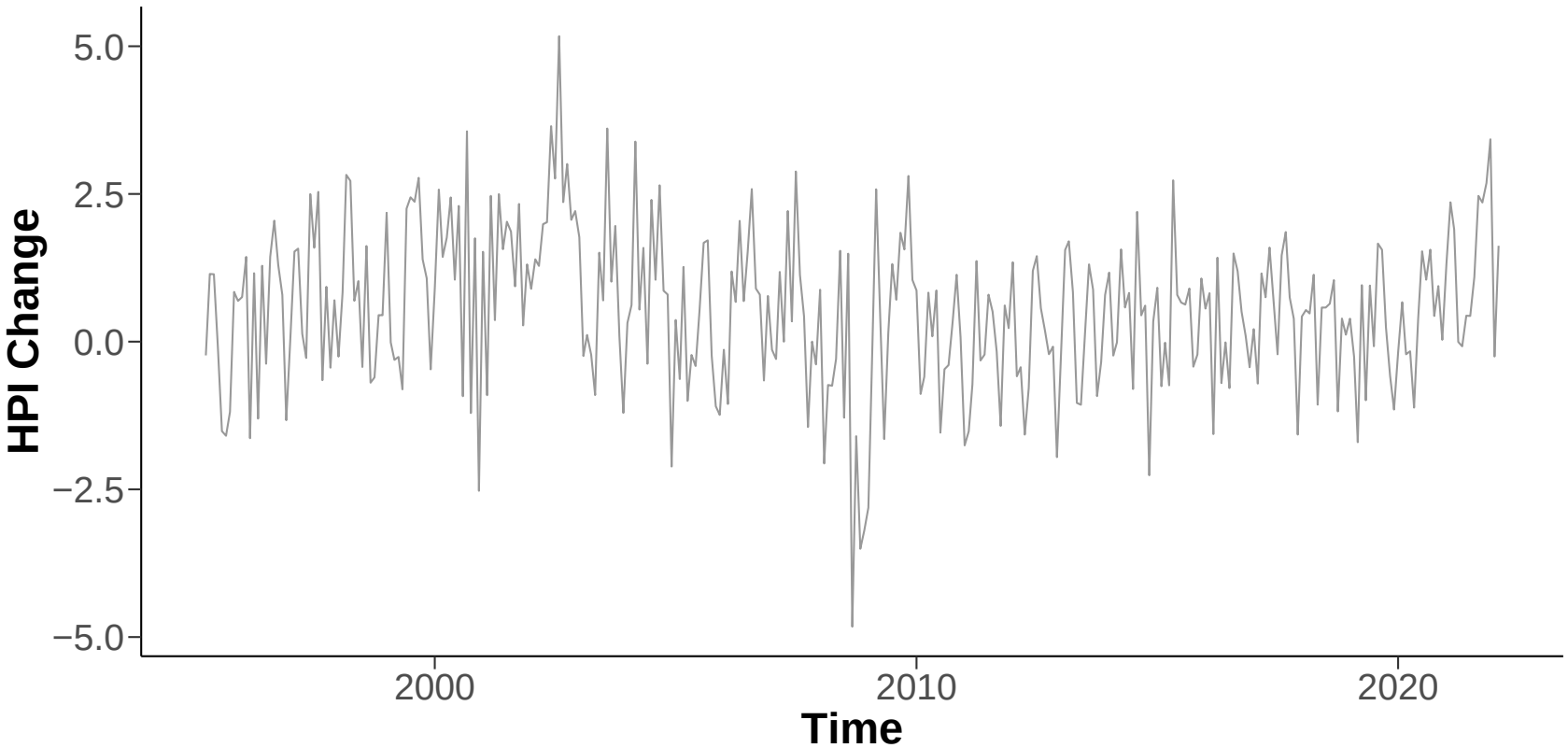
- $\widehat{\sigma}_{e-s}$ is computed by TAVC estimator (using $\widehat{\sigma}_M^2$ if $e - s > M$).

- Threshold $D = C\sqrt{2\log(n)}$ where C is a universal constant.

- If change exists, partition data into $\{X_t\}_{t=1}^{k_\circ}$ and $\{X_t\}_{t=k_\circ+1}^{n}$ and repeat.

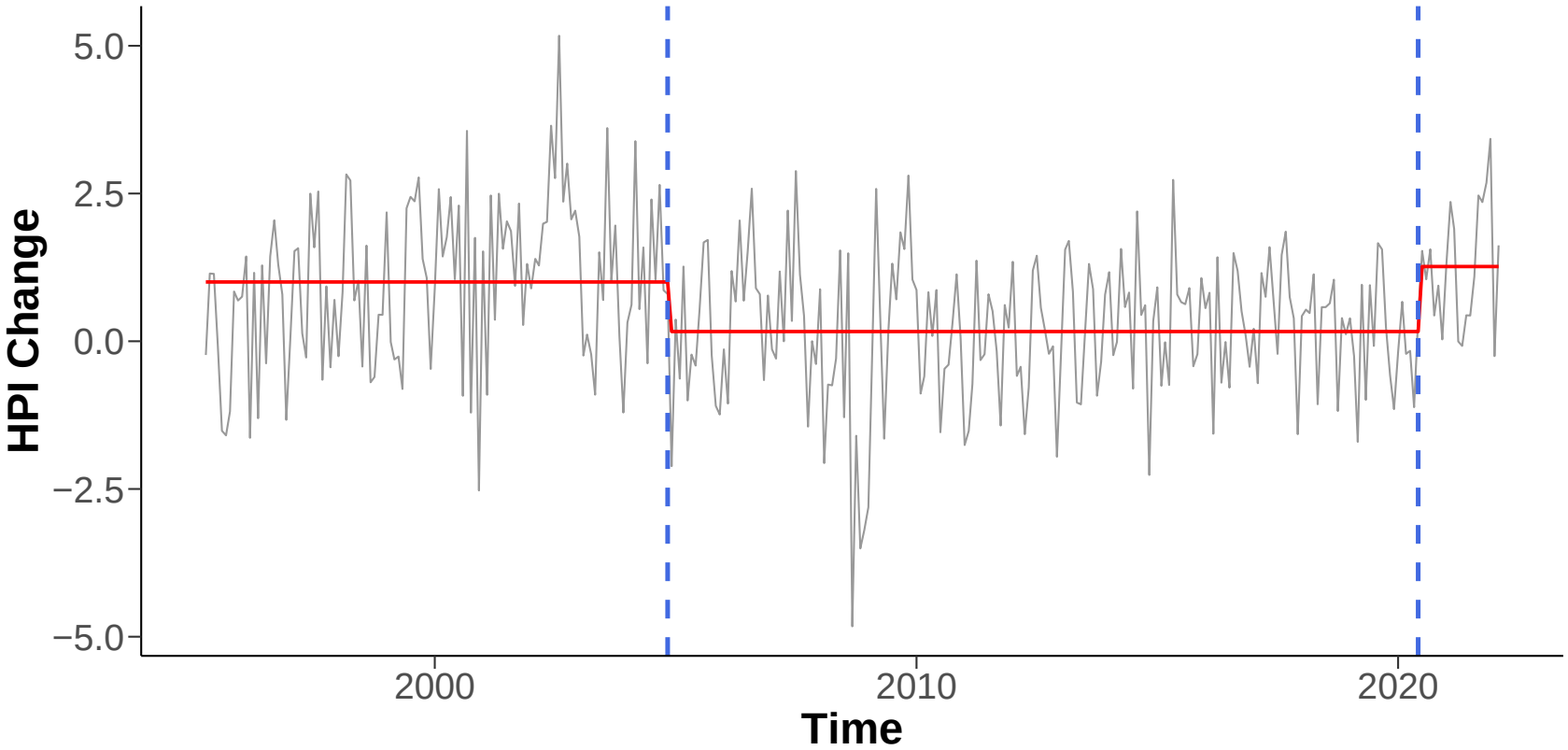
Data example - monthly house price changes

- Possibly nonstationary noise, due to e.g. financial crash.
- Use the time-varying TAVC estimator along with WBS2.



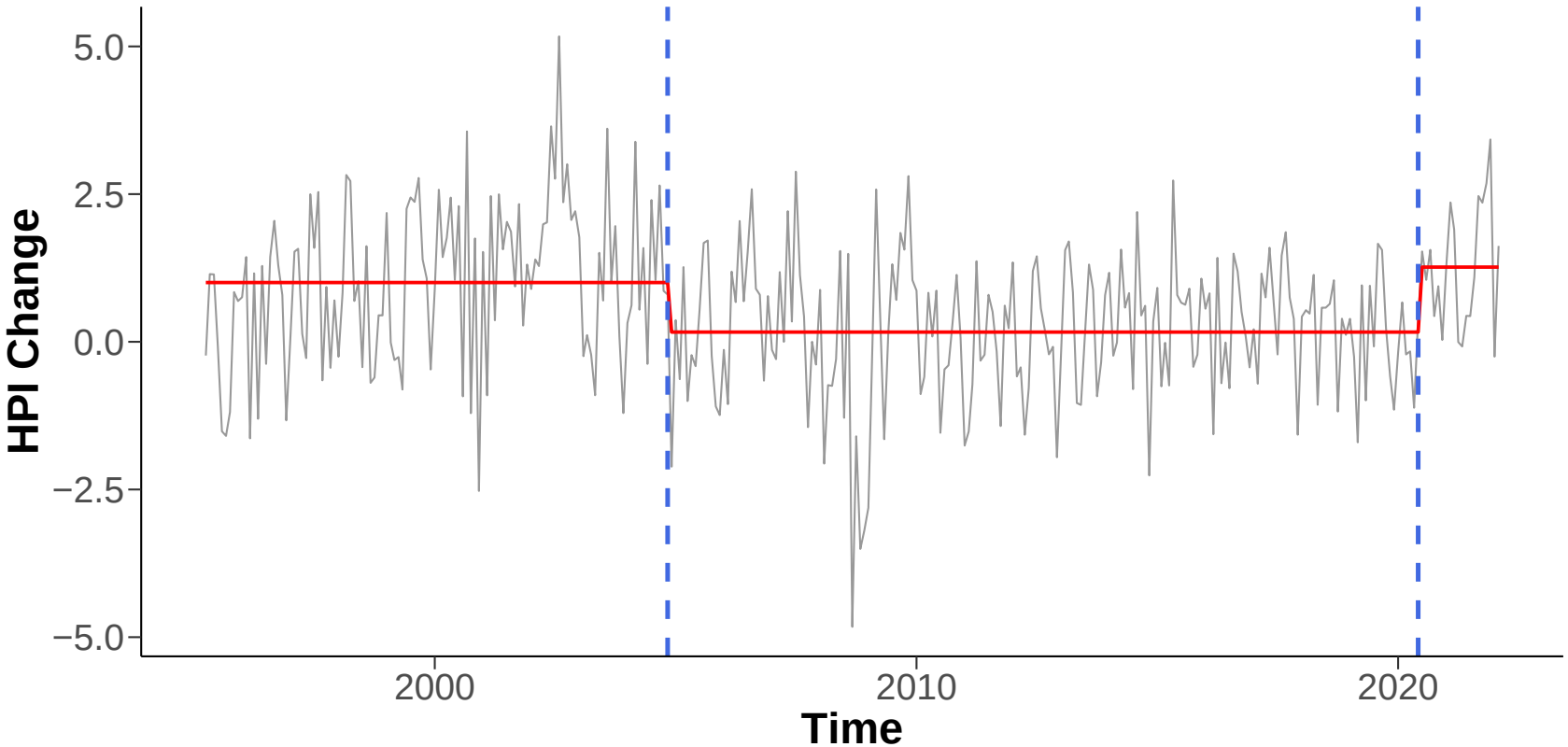
Data example - detected changes

- We find 2 changes.
- No changes in the favourite change point time period of the 2007-2008 financial crash.



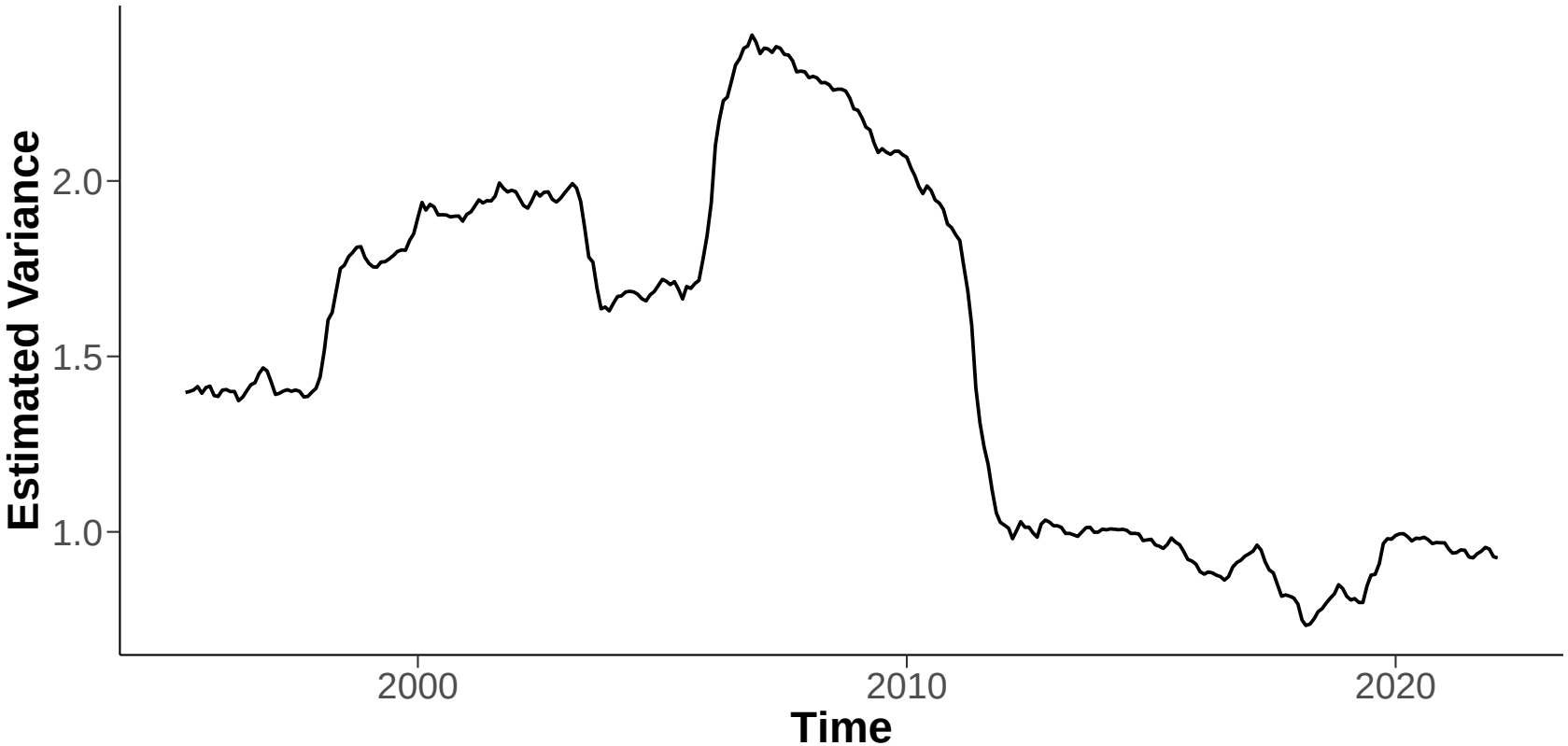
Data example - detected changes

- Methods like DepSMUCE, WCM.gSa have similar results, but either don't detect last change or detect changes in crash period.
- Last change linked to "race for space" during pandemic.



Data example - estimated variance

- Instead of mean changes during the crash, we find variance changes.
- No ground truth, but a plausible interpretation.



Concluding remarks

Summary:

- Mean CPD methods frequently struggle with serial dependence.
- Multiscale approach to robust noise estimation appropriate for many popular multiscale CPD methods.
- Full details in paper: on arXiv and in Computational Statistics & Data Analysis. Code available on Github.

Further work:

- Incorporate within other methods, e.g. dynamic programming?
- Theoretical analysis for locally stationary noise.