Tail-summed scores method for inference in the Gaussian sequence model with applications to change-point analysis

Anica Kostic Piotr Fryzlewicz

London School of Economics and Political Science

StatScale ECR Meeting
14-16 December 2022, Brighton, UK

Overview

The Gaussian sequence model

$$X_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is the unknown mean vector

- $\boldsymbol{\mu}$ is a sparse vector
- $\mathcal{S} = \{i : \mu_i \neq 0\}$, $|\mathcal{S}| = k$
- $\sigma^2 = 1$

Inference on $\boldsymbol{\mu}$:

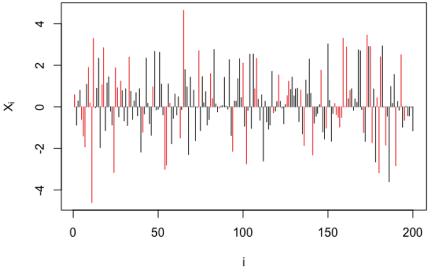
- Signal estimation
- Multiple testing

TSS procedure

Sample: $X_1, \ldots, X_n$
Sequence of thresholds: $\lambda_1, \ldots, \lambda_n$

1. $Y_i = X_i^2,\ i = 1, \ldots, n$
2. $Y_{(n)} \geq \cdots \geq Y_{(1)}$
3. $T_i = \sum_{j=1}^{n-i+1} Y_{(j)}$
4. $\hat{k} = \max_i \{T_i \geq \lambda_i\}$

   $\widehat{\mathcal{S}} = \{j : Y_j \geq Y_{(n-\hat{k}+1)}\}$

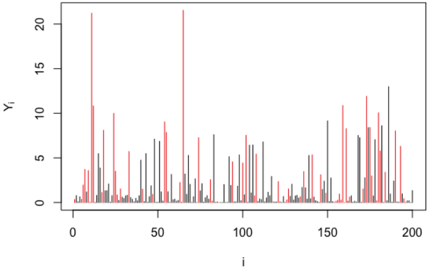


TSS procedure

Sample: $X_1, \dots, X_n$
Sequence of thresholds: $\lambda_1, \dots, \lambda_n$

1. $Y_i = X_i^2,\ i = 1, \dots, n$
2. $Y_{(n)} \geq \dots \geq Y_{(1)}$
3. $T_i = \sum_{j=1}^{n-i+1} Y_{(j)}$
4. $\hat{k} = \max_i \{T_i \geq \lambda_i\}$

   $\widehat{\mathcal{S}} = \{j : Y_j \geq Y_{(n-\hat{k}+1)}\}$

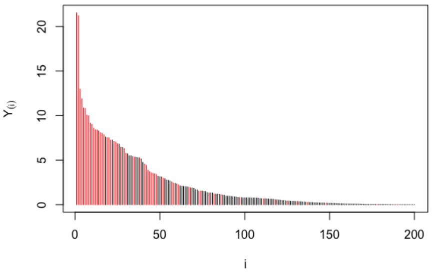


TSS procedure

Sample: $X_1, \ldots, X_n$
Sequence of thresholds: $\lambda_1, \ldots, \lambda_n$

1. $Y_i = X_i^2, \; i = 1, \ldots, n$
2. $Y_{(n)} \geq \cdots \geq Y_{(1)}$
3. $T_i = \sum_{j=1}^{n-i+1} Y_{(j)}$
4. $\hat{k} = \max_i \{ T_i \geq \lambda_i \}$

   $\widehat{\mathcal{S}} = \{ j : Y_j \geq Y_{(n-\hat{k}+1)} \}$

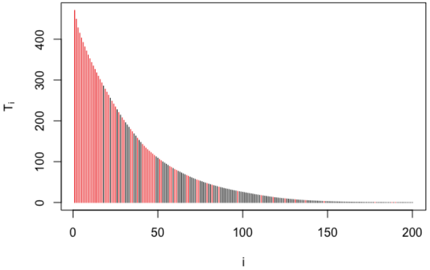


TSS procedure

Sample: $X_1, \dots, X_n$
Sequence of thresholds: $\lambda_1, \dots, \lambda_n$

1. $Y_i = X_i^2,\ i = 1, \dots, n$
2. $Y_{(n)} \geq \dots \geq Y_{(1)}$
3. $T_i = \sum_{j=1}^{n-i+1} Y_{(j)}$
4. $\hat{k} = \max_i \{T_i \geq \lambda_i\}$

   $\widehat{\mathcal{S}} = \{j : Y_j \geq Y_{(n-\hat{k}+1)}\}$

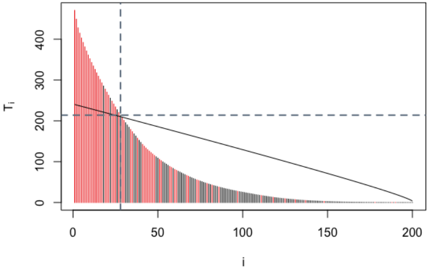


TSS procedure

Sample: $X_1, \dots, X_n$
Sequence of thresholds: $\lambda_1, \dots, \lambda_n$

1. $Y_i = X_i^2,\ i = 1, \dots, n$
2. $Y_{(n)} \geq \cdots \geq Y_{(1)}$
3. $T_i = \sum_{j=1}^{n-i+1} Y_{(j)}$
4. $\hat{k} = \max_i \{T_i \geq \lambda_i\}$

   $\widehat{\mathcal{S}} = \{j : Y_j \geq Y_{(n-\hat{k}+1)}\}$



Choice of thresholds

Chi-square quantiles:

$$\lambda_i = q_{\chi^2_{n-i+1}}(1-\alpha) \tag{2}$$

...or approximations:

$$\lambda_i^{H_i} = n - i + 1 + H_i\sqrt{2(n-i+1)} \tag{3}$$

- $H_i = \sqrt{2\log\frac{1}{\alpha}} + \sqrt{\frac{2}{n-i+1}}\log\frac{1}{\alpha}$
- $H_i = q_{N(0,1)}(1-\alpha)$ or $H_i = 2$
- $H_i = 0$ - asymptotically equal

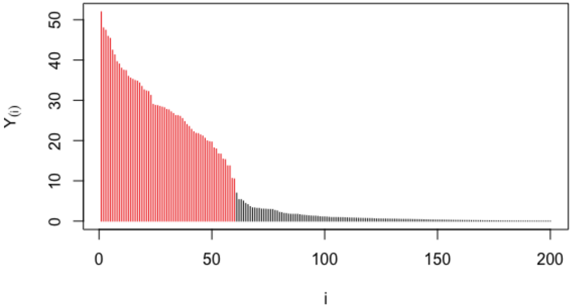
Contiguity property:

$$T_i \leq \lambda_i^{H_i} \implies T_{i+1} \leq \lambda_{i+1}^{H_{i+1}}$$

for any non-decreasing sequence H_i.

Theoretical considerations - perfect separation

- [Duval et al., 2007] considered the TSS procedure with $H_i = 0$
- Perfect separation event: $\Omega_n = \{Y^S_{(1)} \geq Y^{NS}_{(n-k)}\}$
- Perfect separation assumption: $P(\Omega_n) \to 1,\ n \to \infty$
- A sufficient condition: $\mu \geq \sqrt{2\log(n-k)} + \sqrt{2\log k}$



Perfect separation

Let

- $k = \pi_1 n$ for some $\pi_1 \in (0, 1)$
- The perfect separation assumption holds
- $\lambda_i^{H_i}$ such that $H_i(n)/\sqrt{nu_n} \to 0$, where u_n is s.t. $u_n \to 0$, $\sqrt{n}u_n \to \infty$

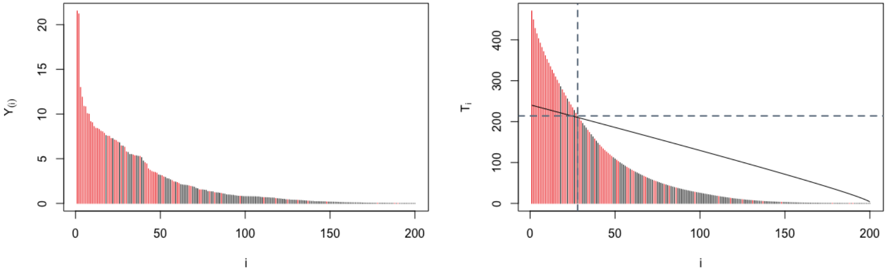
For the TSS stopping time $\hat{k}$ it holds that

$$P\left(\left|\frac{\hat{k}}{n} - \pi_1\right| \geq u_n\right) \to 0, \quad n \to \infty$$

$$FDR \to 0$$

$$FNR \to 0$$

Weak signal - no assumptions on the strength



Weak signal - no assumptions on the strength

- **Stop after the mixing starts**
  Let j be the position of the largest non-signal variable in the decreasingly sorted sample. It holds that

$$P\left(\hat{k} \leq j\right) \to 0, \quad n \to \infty.$$

- **Stop before the mixing ends**
  The probability of the TSS method overestimating the number of signals k is small:

$$P\left(\hat{k} > k\right) \leq 1 - F_{\chi^2_{n-k}}(\lambda_{k+1}).$$

Simulation results

The TSS procedure does not control the FDR when the signal is weak.

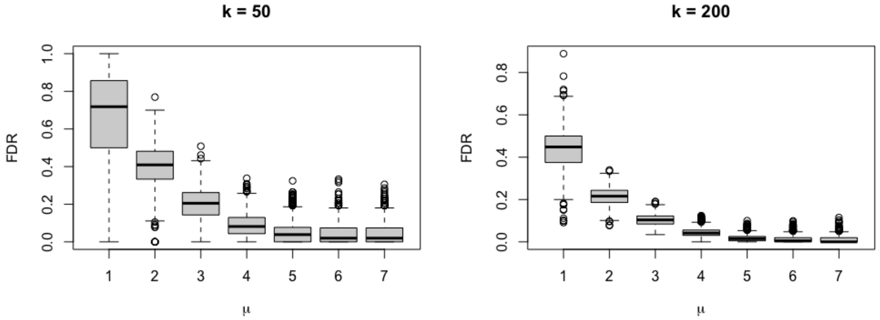


Figure: Boxplots of the estimated FDR for the TSS procedure with $H = 0$, $n = 1000$.

Simulation results

The TSS procedure works well as a signal estimation method if the signal is weak.

	$\mu = 2$					$\mu = 3$				
Number nonzero	50	100	200	300	500	50	100	200	300	500
TSS $H = 2$	**206**	**377**	**662**	904	1300	275	434	689	888	1190
TSS $H = 0$	209	**367**	**634**	**859**	**1226**	**227**	374	609	798	1078
UNI	**203**	404	805	1206	2000	366	725	1433	2122	3391
FDR $q = 0.05$	314	383	785	1150	1795	313	519	836	1085	1460
FDR $q = 0.2$	227	421	673	900	1230	**226**	**350**	**527**	**659**	**850**
EBT	**203**	386	681	**874**	**944**	242	**361**	**494**	**585**	**749**
SURE	**203**	400	803	1206	2012	312	700	1443	2180	3640

Table: The estimated l_2 risk of different thresholding estimators based on $N = 1000$ repetitions for the sample of size $n = 1000$ from the Gaussian sequence model.

Summary and extensions

- Proposed is a new thresholding signal estimation/Multiple testing procedure for the Gaussian sequence model
  - Consider values in groups
  - Stop when the remaining signal is too weak
- Choose the sequence of thresholds depending on the objective:
  - FWER control
  - Signal estimation
  - FDR control
  - Estimating the l_2 norm of the quadratic functional [Collier et al., 2017]

$$\hat{\mu}_{\text{minimax}} = \sum_{i=1}^{n} X_i^2 - n$$

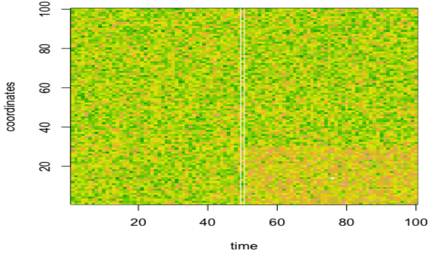
- Different transformation of the data?

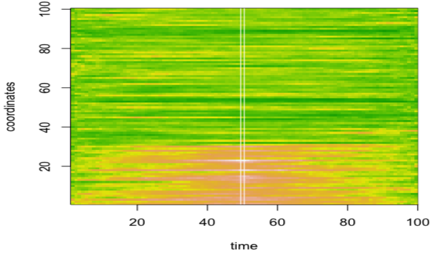
Change-point applications

$$X_{ij} = \mu_{ij} + \varepsilon_{ij}, \quad i = 1, \ldots, p \quad j = 1, \ldots, n,$$
$$\mathcal{S} = \{j : \mu_{i,\tau+1} \neq \mu_{i,\tau}\} \subset \{1, ..., p\}.$$

CUSUM transform the data matrix $X \in \mathbb{R}^{p \times n} \mapsto Z \in \mathbb{R}^{p \times (n-1)}$ is

$$Z_{ij} = \sqrt{\frac{j(n-j)}{n}} \left(\frac{1}{n-j} \sum_{l=j+1}^{n} X_{il} - \frac{1}{j} \sum_{l=1}^{j} X_{il} \right)$$





Change-point applications

$$Z_{ij} \sim N(\theta_{ij}, 1),$$

$$Z_{\cdot,\tau} \sim N(\theta_{\cdot,\tau}, I_d), \text{ where } \theta_{\cdot,\tau} = \sqrt{\frac{\tau(n-\tau)}{n}}(\mu_{\cdot,\tau+1} - \mu_{\cdot,\tau})$$

This separates the coordinates into two groups based on the distribution of $Z_{i,\tau}$:

$$Z_{i,\tau} \sim \begin{cases} N(0,1), & i \notin \mathcal{S} \\ N(\theta_{i,\tau}, 1), & i \in \mathcal{S} \end{cases} \tag{4}$$

Change-point applications - inference on $\mathcal{S}$

Consistent estimation of $\mathcal{S}$

- [Jirak, 2015]

As a part of the change-point estimation algorithm

- [Cho and Fryzlewicz, 2015], [Cho, 2016]
- [Wang and Samworth, 2018]

Higher Criticism-like statistic for change-point detection

- [Jeng et al., 2012]

Copy number variation (CNV) data

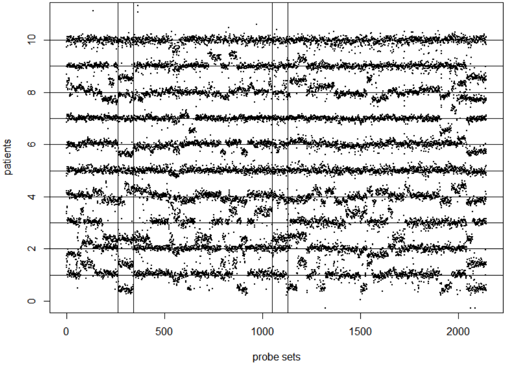


Figure: Log-intensity-ratio values for 10 individuals at 2215 different loci from the study of 43 patients with a bladder tumour. Positive values indicate duplication, while negative indicate deletion at a given locus.

- Variants can be shared among a large number of patients or individual-specific

Change-point applications - inference on $\mathcal{S}$

- Estimating $\mathcal{S}$
  - Improving the accuracy of a change-point estimator
  - For further research on the affected profiles

  Multiple testing or signal estimation methods

- Estimating $|\mathcal{S}|$
  - Estimating the sparsity of the problem
  - Estimating the frequency of copy number variants across a population to get a measure of the extent of the changes studied

  Methods for estimating the proportion of the false null hypotheses in multiple testing

References I

Cho, H. (2016).
Change-point detection in panel data via double cusum statistic.
Electron. J. Statist., 10:2000–2038.

Cho, H. and Fryzlewicz, P. (2015).
Multiple-change-point detection for high dimensional time series via sparsified binary segmentation.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77(2):475–507.

Collier, O., Comminges, L., and Tsybakov, A. B. (2017).
Minimax estimation of linear and quadratic functionals on sparsity classes.
The Annals of Statistics, 45(3):923–958.

References II

Duval, M., Delmas, C., Laurent, B., and Robert-Granié, C. C. (2007).
A procedure based on partial sums of order statistics to detect differentially expressed genes.
Working paper or preprint, available at https://hal.archives-ouvertes.fr/hal-00302355/file/Duval_etal_arxiv.pdf.

Jeng, X. J., Cai, T. T., and Li, H. (2012).
Simultaneous discovery of rare and common segment variants.
Biometrika, 100(1):157–172.

Jirak, M. (2015).
Uniform change point tests in high dimension.
The Annals of Statistics, 43(6):2451–2483.

References III

Wang, T. and Samworth, R. J. (2018).
High dimensional change point estimation via sparse projection.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(1):57–83.