# High-dimensional changepoint estimation with heterogeneous missingness

Tengyao Wang

London School of Economics

LSE Statistics Research Showcase

Jun 2023

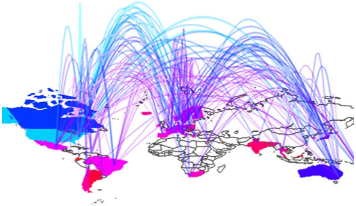
# Collaborators

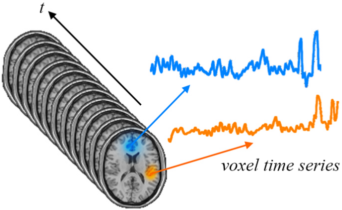
Bertille Follain

Richard Samworth

# High-dimensional changepoint models

- Evolution of technology enables collection of vast amount of time-ordered data
  - Healthcare devices
  - Covid case numbers
  - Network traffic data
  - Trading data of financial instruments





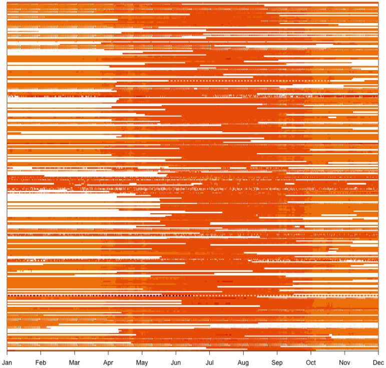

- Changes in the dynamics of the data streams are frequently of interest, leading to a renaissance of research on changepoint analysis.
- Modern data are often high-dimensional in nature — combine high-dimensional statistics with changepoint analysis.

# Missingness in Big Data

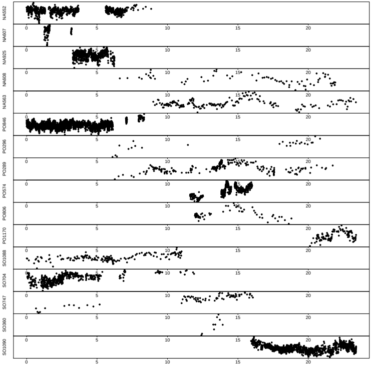
- The irony of Big Data is that missingness plays an even more prominent role.
- Consider running complete-case analysis with an $n \times d$ matrix, where each entry is missing independently with $1\%$ probability.
  - When $d = 5$, around $95\%$ of observations are retained.
  - When $d = 300$, only around $5\%$ of observations are retained.
- In high-dimensional time series models, missingness can also arise due to asynchronous measurements.

# High-dimensional change with missing data

- Our goal is to study the high-dimensional sparse change in mean, but where our data are corrupted by missingness.




French river temperature in 2018




$^{13}C/^{12}C$ in ocean cores 0–23 Ma

- Develop a robust methodology
- Quantify problem difficulty through interaction of signal and missingness

## Problem setup

- Observed data $(X \circ \Omega, \Omega)$
  - Full data matrix $X = (X_{j,t}) \in \mathbb{R}^{p \times n}$
  - Revelation matrix $\Omega = (\omega_{j,t}) \in \{0,1\}^{p \times n}$: $\omega_{j,t} = 1$ if $X_{j,t}$ is observed and 0 otherwise.
- Data distribution:
  - Assume $X_t = (X_{1,t}, \ldots, X_{p,t})^\top \sim N_p(\mu_t, \sigma^2 I_p)$ independently with

$$\mu_1 = \cdots = \mu_z = \mu^{(1)} \quad \text{and} \quad \mu_{z+1} = \cdots = \mu_n = \mu^{(2)}.$$

  - Vector of change $\theta := \mu^{(2)} - \mu^{(1)}$ is sparse in the sense that $\|\theta\|_0 \leq k \ll p$.
- Missingness mechanism:
  - $\omega_{j,t} \sim \mathrm{Bern}(q_j)$ independently, and independent of $X$.

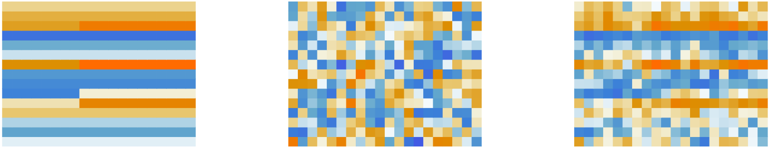
- **Goal**: estimate the changepoint location $z$.

# The `MissInspect` methodology

# Motivation of methodology

- The `inspect` method (W. and Samworth, 2018) works in the fully observed case:
  - Aggregate component series by finding a projection direction well-aligned with the vector of change.
  - Project data along this direction into a univariate series.
  - Estimate changepoint by looking at the CUSUM transform of the projected series.

# Recap of the `inspect` methodology



$$\boldsymbol{\mu} \quad + \quad W \quad = \quad X$$

For $a \in \mathbb{S}^{p-1}$,

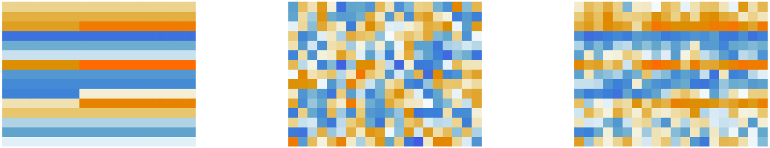
$$a^\top X_t \sim N(a^\top \boldsymbol{\mu}, \sigma^2).$$

Optimal projection direction is $\theta/\|\theta\|_2$.

# Recap of the `inspect` methodology

Use CUSUM transformation $\mathcal{T} : \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times(n-1)}$ for temporal aggregation:

$$[\mathcal{T}(M)]_{j,t} := \sqrt{\frac{t(n-t)}{n}}\biggl(\frac{1}{n-t}\sum_{r=t+1}^{n} M_{j,r} - \frac{1}{t}\sum_{r=1}^{t} M_{j,r}\biggr).$$

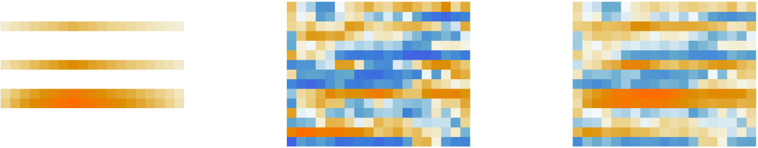


$$\boldsymbol{\mu} \qquad + \qquad W \qquad = \qquad X$$

# Recap of the `inspect` methodology

Use CUSUM transformation $\mathcal{T} : \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times(n-1)}$ for temporal aggregation:

$$[\mathcal{T}(M)]_{j,t} := \sqrt{\frac{t(n-t)}{n}}\left(\frac{1}{n-t}\sum_{r=t+1}^{n} M_{j,r} - \frac{1}{t}\sum_{r=1}^{t} M_{j,r}\right).$$

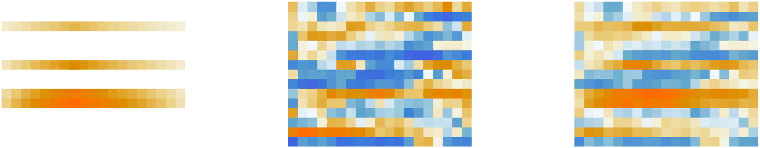


$$\mathcal{T}(\boldsymbol{\mu}) \qquad + \qquad \mathcal{T}(W) \qquad = \qquad \mathcal{T}(X)$$

# Recap of the `inspect` methodology

Use CUSUM transformation $\mathcal{T} : \mathbb{R}^{p\times n} \to \mathbb{R}^{p\times(n-1)}$ for temporal aggregation:

$$[\mathcal{T}(M)]_{j,t} := \sqrt{\frac{t(n-t)}{n}}\biggl(\frac{1}{n-t}\sum_{r=t+1}^{n} M_{j,r} - \frac{1}{t}\sum_{r=1}^{t} M_{j,r}\biggr).$$



$$A \qquad + \qquad E \qquad = \qquad T$$

Define $A := \mathcal{T}(\boldsymbol{\mu})$, $E := \mathcal{T}(W)$ and $T := \mathcal{T}(X)$.

# Recap of the `inspect` methodology

- For a single changepoint, $A = \theta\gamma^\top$ for some $\gamma \in \mathbb{R}^{n-1}$.
- Oracle projection direction $\theta/\|\theta\|_2$ is the leading left singular vector of $A$.
- We could therefore estimate $v$ by

$$\hat{v}_{\max,k} \in \operatorname*{argmax}_{u \in \mathbb{S}^{p-1}(k)} \|u^\top T\|_2.$$

However, computing $\hat{v}_{\max,k}$ is **NP-hard**.

## Recap of the `inspect` methodology

- We obtain a computationally efficient projection direction via convex relaxation.

$$\max_{u\in\mathbb{S}^{p-1}(k)} \|u^\top T\|_2 = \max_{u\in\mathbb{S}^{p-1}(k),w\in\mathbb{S}^{n-2}} u^\top T w$$
$$= \max_{u\in\mathbb{S}^{p-1},w\in\mathbb{S}^{n-2},\|u\|\le k} \langle uw^\top, T\rangle = \max_{M\in\mathcal{M}} \langle M, T\rangle,$$

where $\mathcal{M} := \{M : \|M\|_* = 1, \mathrm{rk}(M) = 1, \mathrm{nnzr}(M) \le k\}$.

- Therefore, a convex relaxation of the above optimisation problem is to compute

$$\hat{M} \in \operatorname*{argmax}_{M\in\mathcal{S}_1} \bigl\{\langle M, T\rangle - \lambda\|M\|_1\bigr\},$$

where $\mathcal{S}_1 = \{M \in \mathbb{R}^{p\times(n-1)} : \|M\|_* \le 1\}$.

- Estimate $\theta/\|\theta\|_2$ by the leading left singular vector of $\hat{M}$.

# Motivation of methodology

- The `inspect` method (W. and Samworth, 2018) works in the fully observed case:
  - Aggregate component series by finding a projection direction well-aligned with the vector of change.
  - Project data along this direction into a univariate series.
  - Estimate changepoint by looking at the CUSUM transform of the projected series.
- In the presence of missingness
  - Projection of data with missingness does not make sense.
  - But the notion of CUSUM transformation can be extended to missing data setting.
  - Project the CUSUM transformation instead.

# MissCUSUM transform

- Writing

$$L_{j,t} := \sum_{r=1}^{t} \omega_{j,t}, \quad R_{j,t} := \sum_{j=n-t+1}^{n} \omega_{j,t}, \quad N_j := L_{j,n} = R_{j,n}.$$

- The MissCUSUM transformation $\mathcal{T}^{\mathrm{Miss}} : \mathbb{R}^{p\times n} \times \{0,1\}^{p\times n} \to \mathbb{R}^{p\times(n-1)}$ is defined such that $T_\Omega = \mathcal{T}^{\mathrm{Miss}}(X,\Omega)$ satisfies

$$(T_\Omega)_{j,t} := \sqrt{\frac{L_{j,t}R_{j,n-t}}{N_j}}\biggl(\frac{1}{R_{j,n-t}}\sum_{r=t+1}^{n}(X\circ\Omega)_{j,r} - \frac{1}{L_{j,t}}\sum_{r=1}^{t}(X\circ\Omega)_{j,r}\biggr),$$

  when $\min\{L_{j,t}, R_{j,t}\} > 0$ and 0 otherwise.
- When the data are fully-observed, i.e. $\Omega$ is an all-one matrix, $\mathcal{T}^{\mathrm{Miss}}$ reduces to the standard CUSUM transformation.

## How to aggregate signal

- Given the MissCUSUM transformed matrix $T_\Omega = \mathcal{T}^{\mathrm{Miss}}(X, \Omega)$, we want to find a good projection direction to aggregate signal across coordinates.
- $T_\Omega$ can be viewed as a perturbation of $A_\Omega$, the MissCUSUM transformation of $(\mathbb{E}(X) \circ \Omega, \Omega)$.
- $A_\Omega$ can in turn be viewed as a perturbation of the rank one matrix with a leading left singular vector $\theta \circ \sqrt{\boldsymbol{q}}$.
- This suggests an 'oracle projection direction' of $\theta \circ \sqrt{\boldsymbol{q}} / \|\theta \circ \sqrt{\boldsymbol{q}}\|$.

# Estimating the oracle projection direction

- We can estimate $\theta \circ \sqrt{\boldsymbol{q}} / \|\theta \circ \sqrt{\boldsymbol{q}}\|$ by looking at 'sparse leading left singular vector' of $T_\Omega$

$$\max_{(v,w) \in \mathbb{R}^p \times \mathbb{R}^{n-1}} \quad v^\top T_\Omega w \qquad \text{subject to} \quad \|v\|_0 \leq k.$$

- Problem non-convex and requires knowledge of $k$.
- W. and Samworth (2018) adopts a semidefinite relaxation approach to convexify the problem. But this the fact that $A_\Omega$ is not rank one means the semi-definite relaxation is too coarse in this case.
- We instead relax it into a bi-convex problem

$$(\hat{v}, \hat{w}) \in \operatorname*{argmax}_{(v,w) \in \mathbb{R}^p \times \mathbb{R}^{n-1}} \left\{ v^\top T_\Omega w - \lambda \|v\|_1 \right\}$$

- Additional benefit: directly exploits the row sparsity pattern.

# The `MissInspect` algorithm

**Algorithm 1:** Pseudocode of the `MissInspect` algorithm

**Input:** $X_\Omega = X \circ \Omega \in \mathbb{R}^{p\times n}$, $\Omega \in \{0,1\}^{p\times n}$, $\lambda > 0$

$T_\Omega \leftarrow \mathcal{T}^{\mathrm{Miss}}(X_\Omega, \Omega)$;

Find $(\hat{v}, \hat{w}) \in \mathrm{argmax}_{\tilde{v}\in\mathbb{B}^{p-1}, \tilde{w}\in\mathbb{B}^{n-2}} \bigl\{\langle T_\Omega, \tilde{v}\tilde{w}^\top\rangle - \lambda\|\tilde{v}\|_1\bigr\}$;

$\hat{z} \leftarrow \mathrm{median}\Bigl(\mathrm{argmax}_{t\in[n-1]}\bigl|(\hat{v}^\top T_\Omega)_t\bigr|\Bigr)$;

**Output:** $\hat{z}$

**Algorithm 2:** Pseudocode for an iterative procedure optimising (2)

**Input:** $T_\Omega \in \mathbb{R}^{p\times(n-1)}$, $\lambda \in \bigl(0, \|T_\Omega\|_{2\to\infty}\bigr)$

$\tilde{v} \leftarrow$ leading left singular vector of $T_\Omega$;

**repeat**

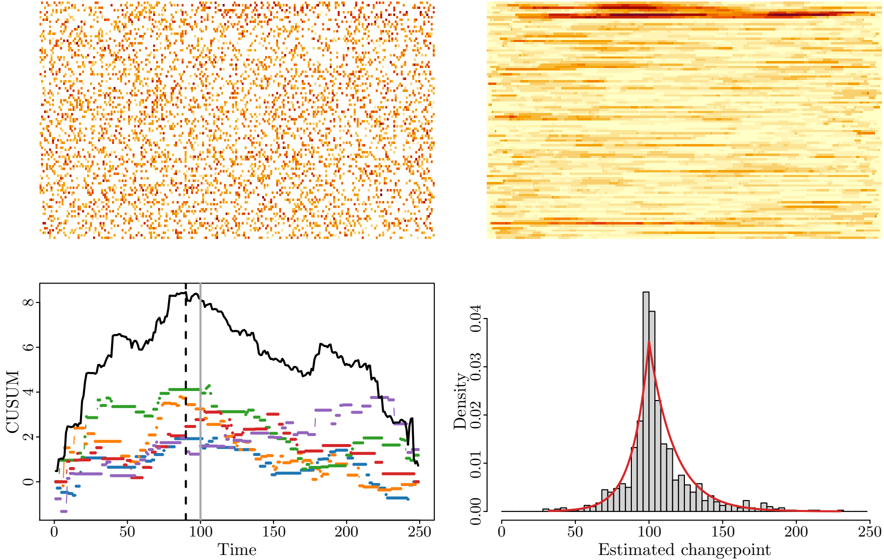
  $\tilde{w} \leftarrow \frac{T_\Omega^\top \tilde{v}}{\|T_\Omega^\top \tilde{v}\|_2}$;

  $\tilde{v} \leftarrow \frac{\mathrm{soft}(T_\Omega \tilde{w}, \lambda)}{\|\mathrm{soft}(T_\Omega \tilde{w}, \lambda)\|_2}$;

**until** *convergence*;

**Output:** $(\hat{v}, \hat{w}) = (\tilde{v}, \tilde{w})$

# Illustration of the algorithm in action




Parameters: $p = 100$, $n = 250$, $z = 100$, $k = 10$, $\|\theta\|_2 = 2$, $q_j = 0.2 \ \forall\, j$

# Theoretical guarantees

# Projection direction estimation

- Let $\tau := n^{-1}\min\{z, n-z\}$. Define the 'observation rate-weighted signal $\ell_2$ norm':

$$\|\theta\|_{2,\boldsymbol{q}} := \left(\sum_{j=1}^{p} \theta_j^2 q_j\right)^{1/2}$$

**Proposition.** Let $(\hat{v}, \hat{w})$ be the optimiser in Step 1 of Algorithm 1, applied with $\lambda = 2\sigma\sqrt{n\log(pn)}$. Then

$$\mathbb{P}\left\{\sin\angle(\hat{v}, \theta\circ\sqrt{\boldsymbol{q}}) \leq \frac{64\sigma}{\tau\|\theta\|_{2,\boldsymbol{q}}}\sqrt{\frac{k\log(pn)}{n}} + \frac{112\|\theta\|_2}{\tau\|\theta\|_{2,\boldsymbol{q}}}\sqrt{\frac{6\log(kn)}{n}}\right\} \geq 1 - \frac{6}{kn}.$$

- First term represents estimation error caused by noise in data: $\|\theta\|_{2,\boldsymbol{q}}/\sigma$ is the signal-to-noise ratio
- Second term reflects error due to incomplete observation: $\|\theta\|_{2,\boldsymbol{q}}^2/\|\theta\|_2^2$ may be regarded as 'signal-weighted observation probability'.

# Rate of location estimation

- With a good projection direction estimator, `MissInspect` algorithm produces good changepoint location estimator.
- We analyse a sample-splitting variant of Algorithm 1
  - Odd time points for projection direction estimation
  - Even time points for changepoint estimation after projection
- Two different rates of convergence of the location estimator depending on how much we are willing to assume on $\boldsymbol{q}$:
  - slow rate: algorithm works well even if some coordinates are almost completely missing.
  - fast rate: when at least a logarithmic number of observations are seen in each coordinate.

# Slow and fast rates

**Theorem.** Set tuning parameter $\lambda = 2\sigma\sqrt{n\log(pn)}$. There exists universal constants $c, C, C_1, C_2$ such that if

$$\frac{1}{\tau}\sqrt{\frac{\log(pn)}{n}}\biggl(\frac{\sigma\sqrt{k}}{\|\theta\|_{2,\boldsymbol{q}}} + \frac{\|\theta\|_2}{\|\theta\|_{2,\boldsymbol{q}}}\biggr) \leq c,$$

then

$$\mathbb{P}\biggl\{\frac{|\hat{z} - z|}{n\tau} \leq C\sqrt{\frac{\log(kn)}{n\tau}}\biggl(\frac{\sigma}{\|\theta\|_{2,\boldsymbol{q}}} + \frac{\|\theta\|_2}{\|\theta\|_{2,\boldsymbol{q}}}\biggr)\biggr\} \geq 1 - \frac{22}{n}.$$

If in addition, $n\tau^2 \min_j q_j \geq C_1 k\log(pn)$, then

$$\mathbb{P}\biggl\{\frac{|\hat{z} - z|}{n\tau} \leq \frac{C_2\log(pn)}{n\tau}\biggl(\frac{\sigma^2}{\|\theta\|_{2,\boldsymbol{q}}^2} + \frac{\|\theta\|_\infty^2}{\|\theta\|_{2,\boldsymbol{q}}^2}\biggr)\biggr\} \geq 1 - \frac{23}{n}.$$

# Lower bound

- Let $P_{n,p,z,\theta,\sigma,\boldsymbol{q}}$ denote all distributions satisfying our modelling assumption.
- Let $\hat{\mathcal{Z}}$ be the set of all estimators of $z$.

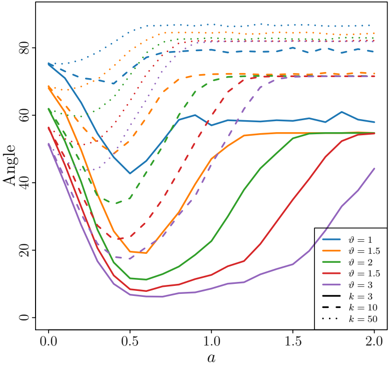
**Theorem.** Let $M \geq 1$ satisfy $\|\theta\|_\infty \leq M \min_{j\in[p]:\theta_j\neq 0} |\theta_j|$. If $\max\{\sigma^2, \|\theta\|_\infty^2/(2M^2)\} \geq \|\theta\|_{2,\boldsymbol{q}}^2$, then there exists $c > 0$, depending only on $M$, such that for $n \geq 4$,

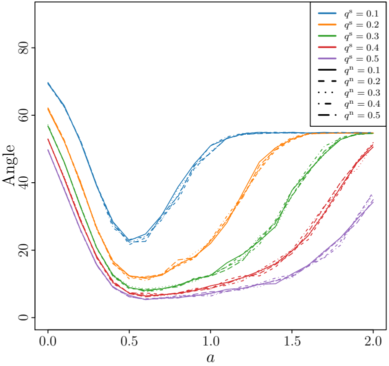
$$\inf_{\tilde{z}\in\hat{\mathcal{Z}}} \max_{z\in[n-1]} \mathbb{E}_{P_{n,p,z,\theta,\sigma,\boldsymbol{q}}} \frac{|\tilde{z}(X\circ\Omega,\Omega) - z|}{n\tau} \geq \frac{c}{n\tau} \min\biggl\{\frac{\sigma^2}{\|\theta\|_{2,\boldsymbol{q}}^2} + \frac{\|\theta\|_\infty^2}{\|\theta\|_{2,\boldsymbol{q}}^2}, n\biggr\}.$$

# Numerical studies

# Choice of the tuning parameter

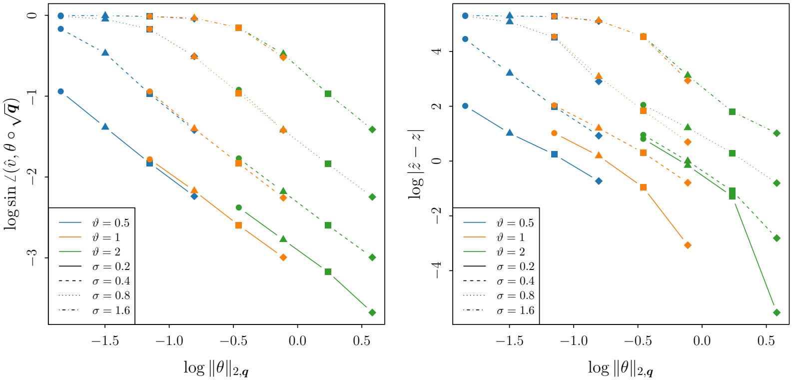
- The tuning parameter $\lambda = 2\sigma\sqrt{n\log(pn)}$ is convenient for theoretical analysis but often too conservative in practice.
- Examine the performance of the projection direction estimator for $\lambda = a\sigma\sqrt{n\log(pn)}$ by varying $a$.
- Best choice around $a = 1/2$.



# Validation of theory

- We show via simulation that the quantity $\|\theta\|_{2,\boldsymbol{q}}$ indeed captures the appropriate interaction between signal and missingness in this problem.




Parameters: $n = 1200$, $p = 1000$, $z = 400$, $k = 3$, $\boldsymbol{q} = q\mathbf{1}_p$ with $q \in \{0.1, 0.2, 0.4, 0.8\}$.

## Comparison with a competitor

- `ImputeInspect` algorithm
  - First impute missing data using the `softImpute` matrix completion algorithm (since the mean matrix of $X \circ \Omega$ is low-rank)
  - Then run the `inspect` procedure on the imputed data.
- Compare both projection direction estimator quality and changepoint location estimation accuracy.

# Comparison with a competitor

| $\nu$ | $k$ | $\vartheta$ | $\angle(\hat{v}^{\mathrm{MI}}, \theta \circ \sqrt{\boldsymbol{q}})$ | $\angle(\hat{v}^{\mathrm{II}}, \theta)$ | $\lvert\hat{z}^{\mathrm{MI}} - z\rvert$ | $\lvert\hat{z}^{\mathrm{II}} - z\rvert$ | $\lvert\hat{z}^{\mathrm{IMI}} - z\rvert$ | $\lvert\hat{z}^{\mathrm{GLR}} - z\rvert$ |
|---|---|---|---|---|---|---|---|---|
| 0.1 | 3 | 1 | **71.4** | 86.8 | **141.7** | 468.0 | 184.1 | 212.4 |
| 0.1 | 3 | 2 | **40.6** | 56.7 | **36.5** | 304.8 | 147.7 | 139.8 |
| 0.1 | 3 | 3 | **26.1** | 40.1 | **14.5** | 257.5 | 101.0 | 66.6 |
| 0.1 | 44 | 1 | **82.6** | 88.9 | **185.9** | 468.9 | 187.6 | 209.4 |
| 0.1 | 44 | 2 | **63.5** | 83.2 | **66.9** | 404.5 | 133.7 | 118.3 |
| 0.1 | 44 | 3 | **49.0** | 72.8 | **18.7** | 308.6 | 90.8 | 52.0 |
| 0.1 | 2000 | 1 | **86.5** | 88.2 | **180.0** | 485.0 | 184.1 | 219.6 |
| 0.1 | 2000 | 2 | **76.9** | 87.6 | **121.2** | 457.3 | 138.9 | 137.5 |
| 0.1 | 2000 | 3 | **67.7** | 82.9 | 50.4 | 376.9 | 79.2 | **41.0** |
| 0.5 | 3 | 1 | **32.3** | 81.0 | **11.9** | 358.4 | 150.8 | 176.0 |
| 0.5 | 3 | 2 | **13.6** | 42.1 | **1.6** | 7.2 | 44.8 | 10.5 |
| 0.5 | 3 | 3 | **9.6** | 24.8 | **0.7** | 6.9 | 7.6 | 2.1 |
| 0.5 | 44 | 1 | **62.7** | 88.4 | **50.1** | 438.5 | 159.4 | 207.1 |
| 0.5 | 44 | 2 | **37.3** | 73.6 | **2.3** | 174.2 | 41.8 | 7.3 |
| 0.5 | 44 | 3 | **26.9** | 58.1 | **0.7** | 1.8 | 3.3 | 1.6 |
| 0.5 | 2000 | 1 | **77.5** | 88.6 | **114.3** | 448.1 | 162.5 | 202.9 |
| 0.5 | 2000 | 2 | **59.2** | 85.5 | **6.7** | 338.6 | 40.6 | 6.8 |
| 0.5 | 2000 | 3 | **52.0** | 72.4 | **1.7** | 48.2 | 3.9 | **1.7** |

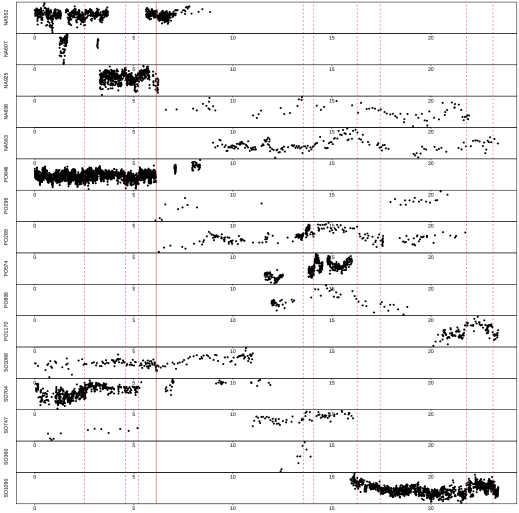
Parameters: $n = 1200$, $p = 2000$, $z = 400$, $q_1, \ldots, q_p \overset{\text{iid}}{\sim} \mathrm{Beta}\bigl(10\nu, 10(1-\nu)\bigr)$

# Real data analysis

- Oceanographic dataset covering the Neogene geological period (Samworth and Poore, 2005; Poore et al., 2006).
- Cores were extracted from North Atlantic, Pacific and Southern Oceans measuring ratio of abundance of $^{13}\mathrm{C}$ to $^{12}\mathrm{C}$ isotope ratio in microfossils at different depths (proxy for geological age).
- 7369 observations at 6295 distinct time points.
- Due to physical constraints and heterogeneity in the analysis carried out in different cores, appropriate to treat the series as data with missingness.

# Real data analysis

- The most prominent change at 6.13Ma was previously identified as a time of rapid change in oceanographic current flows (Poore et al., 2006).



# Summary

- We propose a new method for high-dimensional changepoint estimation in the presence of missing data.
- A good projection direction for aggregation is estimated after applying a MissCUSUM transformation to the data.
- Theory reveals interesting interaction between signal and missingness in this problem.
- **R** package available on https://github.com/wangtengyao/MissInspect.

Main reference


- Follain, B., Wang, T. and Samworth, R. J. (2021) High-dimensional changepoint estimation with heterogeneous missingness. *arXiv preprint*, arxiv:2108.01525.

Thank you!