Sparse change detection in high-dimensional linear regression

Tengyao Wang

London School of Economics and Political Science

Departmental Research Day

Jun 2022

Collaborator

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Fengnan Gao
Fudan University

Problem setup

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- Observations $(x_t, y_t) \in \mathbb{R}^p \times \mathbb{R}$ for $t = 1, \ldots, n$ generated from

$$y_t = x_t^\top \beta_t + \epsilon_t,$$

  where $\epsilon_t \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$.
- Coefficients $\beta_1, \ldots, \beta_n$ piecewise constant with changepoints at $z_1, \ldots, z_\nu$

$$\beta_t = \beta^{(r)} \quad \text{for } z_{r-1} < t \leq z_r, 1 \leq r \leq \nu + 1.$$

- **Goal**: estimate the changepoint locations $z_1, \ldots, z_\nu$.
- **Key challenge**: we only assume sparsity of $\beta^{(r)} - \beta^{(r-1)}$ but not $\beta^{(r)}$ themselves.

Landscape of relevant literature

- **High dimensional linear regression**: one of the most fruitful area of statistical research in the past twenty years (Tibshrani, 1996; Fan and Lv, 2010; Bühlmann and van de Geer, 2011; etc)
- Data heterogeneity in high-dimensional linear models (Städler et al., 2010; Krishnamurthy et al., 2019).

Landscape of relevant literature

- **High dimensional linear regression**: one of the most fruitful area of statistical research in the past twenty years (Tibshrani, 1996; Fan and Lv, 2010; Bühlmann and van de Geer, 2011; etc)
- Data heterogeneity in high-dimensional linear models (Städler et al., 2010; Krishnamurthy et al., 2019).

- **Changepoint analysis**: a useful framework for analysing data with temporal heterogeneity (Page, 1955)
- High-dimensional mean change estimation (Cho and Fryzlewicz, 2015; Jirak, 2015; W. and Samworth, 2018; Enikeeva and Harchaoui, 2019; etc)
- Multivariate change in regression (Bai and Perron, 1998; Fryzlewicz, 2021; etc)
- High-dimensional change in regression (Lee et al., 2015; Leonardi and Bühlmann, 2016; Rinaldo et al., 2021; Wang et al., 2021)

Two sample testing of regression coefficients

- Many changepoint procedures are constructed from two-sample testing statistics (e.g. CUSUM statistics for change-in-mean problems)

Two sample testing of regression coefficients

- Many changepoint procedures are constructed from two-sample testing statistics (e.g. CUSUM statistics for change-in-mean problems)

- Two samples $(X^{(1)}, Y^{(1)}) \in \mathbb{R}^{n_1 \times p} \times \mathbb{R}^{n_1}$ and $(X^{(2)}, Y^{(2)}) \in \mathbb{R}^{n_2 \times p} \times \mathbb{R}^{n_2}$, generated from the linear models:

$$\begin{cases} Y^{(1)} = X^{(1)}\beta^{(1)} + \epsilon^{(1)} \\ Y^{(2)} = X^{(2)}\beta^{(2)} + \epsilon^{(2)}, \end{cases}$$

  where $\epsilon^{(1)} \sim N_{n_1}(0, \sigma^2 I_{n_1})$ and $\epsilon^{(2)} \sim N_{n_2}(0, \sigma^2 I_{n_2})$ are independent.
- Given $(X^{(1)}, Y^{(1)})$ and $(X^{(2)}, Y^{(2)})$, we want to test

$$H_0 : \beta^{(1)} = \beta^{(2)} \quad \text{vs} \quad H_1 : \|\beta^{(1)} - \beta^{(2)}\|_2 \geq \rho \text{ and } \|\beta^{(1)} - \beta^{(2)}\|_0 \leq k.$$

Two sample testing of regression coefficients

- Existing works mostly assume sparsity on both $\beta^{(1)}$ and $\beta^{(2)}$ (e.g. Xia, Cai and Cai, 2018)
- But parameter of interest is really

$$\theta := \frac{\beta^{(1)} - \beta^{(2)}}{2}$$

and $\gamma := (\beta^{(1)} + \beta^{(2)})/2$ is a possibly dense nuisance parameter.

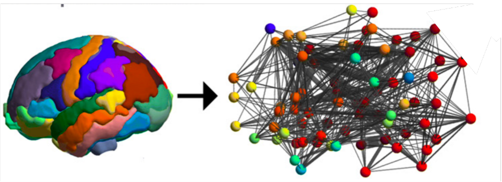
Two sample testing of regression coefficients

- Existing works mostly assume sparsity on both $\beta^{(1)}$ and $\beta^{(2)}$ (e.g. Xia, Cai and Cai, 2018)
- But parameter of interest is really

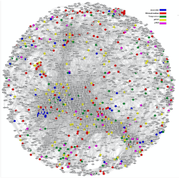
$$\theta := \frac{\beta^{(1)} - \beta^{(2)}}{2}$$

and $\gamma := (\beta^{(1)} + \beta^{(2)})/2$ is a possibly dense nuisance parameter.

- **Application**: testing whether two networks formulated by Gaussian graphical models are the same.
  - Gene-gene interaction network
  - Foreign exchange network model



Bansal et al. (*Sci. Adv.* 2019)



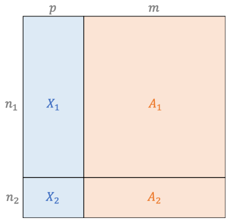
Chen et al. (*PLOS ONE*, 2015)

Complementary sketching

- **Procedure:** Given data $X^{(1)}, X^{(2)}, Y^{(1)}, Y^{(2)}$, set $m := n_1 + n_2 - p$
  1. Construct $A_1 \in \mathbb{R}^{n_1 \times m}$ and $A_2 \in \mathbb{R}^{n_2 \times m}$ such that $\binom{A_1}{A_2}$ has orthonormal columns orthogonal to the column space of $\binom{X^{(1)}}{X^{(2)}}$.
  2. Compute

$$W := \begin{pmatrix} A_1^\top & -A_2^\top \end{pmatrix} \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \in \mathbb{R}^{m \times p},$$

$$Z := \begin{pmatrix} A_1^\top & A_2^\top \end{pmatrix} \begin{pmatrix} Y^{(1)} \\ Y^{(2)} \end{pmatrix} \in \mathbb{R}^{m}.$$



- Similar to orthogonal sketching, but sketches the covariate matrix and the response vector in opposite ways in the second block.

Complementary sketching

- Observe that

$$\begin{aligned} Z &= A_1^\top Y^{(1)} + A_2^\top Y^{(2)} = A_1^\top X^{(1)}\beta^{(1)} + A_2^\top X^{(2)}\beta^{(2)} + A_1\epsilon^{(1)} + A_2\epsilon^{(2)} \\ &= A_1^\top X^{(1)}\theta + \cancel{A_1^\top X^{(1)}}\gamma - A_2^\top X^{(2)}\theta + \cancel{A_2^\top X^{(2)}}\gamma + A_1\epsilon^{(1)} + A_2\epsilon^{(2)} \\ &= W\theta + \xi, \end{aligned}$$

where $\xi \sim N_m(0, \sigma^2 I_m)$.

- We have reduced the two-sample testing problem to a one-sample problem of sample size m without the nuisance parameter.

- Observe that

$$\begin{aligned} Z &= A_1^\top Y^{(1)} + A_2^\top Y^{(2)} = A_1^\top X^{(1)}\beta^{(1)} + A_2^\top X^{(2)}\beta^{(2)} + A_1\epsilon^{(1)} + A_2\epsilon^{(2)} \\ &= A_1^\top X^{(1)}\theta + \cancel{A_1^\top X^{(1)}\gamma} - A_2^\top X^{(2)}\theta + \cancel{A_2^\top X^{(2)}\gamma} + A_1\epsilon^{(1)} + A_2\epsilon^{(2)} \\ &= W\theta + \xi, \end{aligned}$$

  where $\xi \sim N_m(0, \sigma^2 I_m)$.
- We have reduced the two-sample testing problem to a one-sample problem of sample size m without the nuisance parameter.

- Let $\tilde{W}$ be W with columns normalised to have unit ℓ_2 norms. If θ is sparse, then the test

$$\psi_{\lambda,\tau} := \mathbb{1}\{\|\mathbf{hard}(\tilde{W}^\top Z, \lambda)\|_2^2 \geq \tau\},$$

  with suitably chosen tuning parameters can be shown to be minimax rate optimal (Gao and W., 2021).

Single changepoint estimation

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- Inspired by the two-sample testing problem, we construct $A \in \mathbb{O}^{n\times(n-p)}$ whose columns span the orthogonal complement of the column space of X.

- For any $t \in [n-1]$, form sketched design matrx

$$W_t := A_{(0,t]}^\top X_{(0,t]} - A_{(t,n]}^\top X_{(t,n]} = 2A_{(0,t]}^\top X_{(0,t]} \in \mathbb{R}^{m\times p}$$

- The sketched response is

$$\begin{aligned} Z := A^\top Y &= A_{(0,z]}^\top (X_{(0,z]}\beta^{(1)} + \epsilon_{(0,z]}) + A_{(z,n]}^\top (X_{(z,n]}\beta^{(2)} + \epsilon_{(z,n]}) \\ &= A_{(0,z]}^\top X_{(0,z]}(\theta+\gamma) - A_{(z,n]}^\top X_{(z,n]}(\theta-\gamma) + \xi = W_z\theta + \xi, \end{aligned}$$

- Reduced to finding t such that W_t forms a 'best' sparse linear approximation of Z.

Changepoint algorithms

- Let $Q = (Q_1, \ldots, Q_{n-1})^\top$ be defined such that

$$Q_t := \tilde{W}_t^\top Z \propto \mathsf{Corr}(W_t, Z),$$

where $\tilde{W}_t := W_t\{\mathrm{diag}(W_t^\top W_t)\}^{-1/2}$. Estimate changepoint by

$$\hat{z}^{\mathrm{corr}} := \operatorname*{argmax}_t \|\mathbf{soft}(Q_t, \lambda)\|_2^2.$$

Algorithm 1: Pseudocode for change-point estimation

Input: $X \in \mathbb{R}^{n\times p}, Y \in \mathbb{R}^n$ satisfying $n > p$, $\lambda \geq 0$, $\alpha > 0$

Set $m \leftarrow n - p$;

Form $A \in \mathbb{O}^{n\times m}$ with columns orthogonal to the column space of X;

Compute $Z \leftarrow A^\top Y$;

Set $W_0 = \mathbf{0}_{m\times p}$;

for $1 \leq t \leq n-1$ **do**

  Compute $W_t \leftarrow W_{t-1} + 2a_t x_t^\top$;

  Compute $Q_t = \{\mathrm{diag}(W_t^\top W_t)\}^{-1/2} W_t^\top Z$;

  Compute $H_t \leftarrow \|\mathrm{soft}(Q_t, \lambda)\|_2$;

Output: $\hat{z} := \arg\max_{\alpha n < t < (1-\alpha)n} H_t$.

- Let $Q = (Q_1, \ldots, Q_{n-1})^\top$ be defined such that

$$Q_t := \tilde{W}_t^\top Z \propto \mathrm{Corr}(W_t, Z),$$

where $\tilde{W}_t := W_t\{\mathrm{diag}(W_t^\top W_t)\}^{-1/2}$. Estimate changepoint by

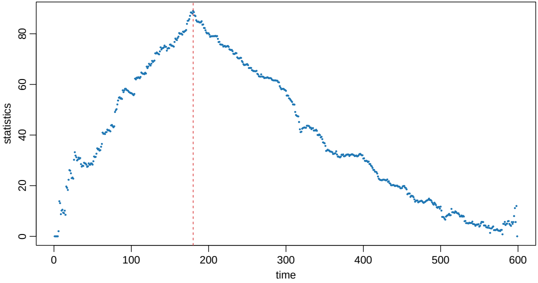
$$\hat{z}^{\mathrm{corr}} := \operatorname*{argmax}_t \|\mathbf{soft}(Q_t, \lambda)\|_2^2.$$

- Let $Q = (Q_1, \ldots, Q_{n-1})^\top$ be defined such that

$$Q_t := \tilde{W}_t^\top Z \propto \mathsf{Corr}(W_t, Z),$$

where $\tilde{W}_t := W_t\{\mathrm{diag}(W_t^\top W_t)\}^{-1/2}$. Estimate changepoint by

$$\hat{z}^{\mathrm{corr}} := \operatorname*{argmax}_t \|\mathbf{soft}(Q_t, \lambda)\|_2^2.$$



Changepoint algorithms

- Let $Q = (Q_1, \ldots, Q_{n-1})^\top$ be defined such that

$$Q_t := \tilde{W}_t^\top Z \propto \mathrm{Corr}(W_t, Z),$$

where $\tilde{W}_t := W_t\{\mathrm{diag}(W_t^\top W_t)\}^{-1/2}$. Estimate changepoint by

$$\hat{z}^{\mathrm{corr}} := \underset{t}{\mathrm{argmax}}\, \|\mathbf{soft}(Q_t, \lambda)\|_2^2.$$

- Alternatively, let $\hat{v}$ be the leading left singular vector of $\mathbf{soft}(Q, \lambda)$, estimate

$$\hat{z}^{\mathrm{proj}} := \underset{t}{\mathrm{argmax}}(\hat{v}^\top Q_t).$$

- We can also simply run Lasso on (W_t, Z) to find the best fit

$$\hat{z}^{\mathrm{lasso}} := \underset{t}{\mathrm{argmin}}\, \|Z - W_t\hat{\theta}_t(\lambda)\|_2^2,$$

where $\hat{\theta}_t(\lambda) := \mathrm{argmin}_\theta\{\frac{1}{2m}\|Z - W_t\theta\|_2^2 + \lambda\|\theta\|_1\}$.

- Gaussian Orthogonal Ensemble design matrices
- $\theta^{(1)}$ sampled as a Gaussian vector
- $\theta^{(2)} - \theta^{(1)}$ randomly generated k-sparse vector with ℓ_2 norm ρ.

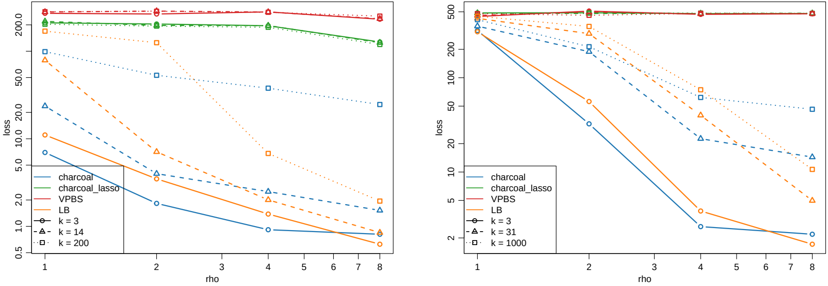


Figure: Left: $n = 600, p = 200, z = 180$; Right: $n = 1200, p = 1000, z = 120$

Empirical performances

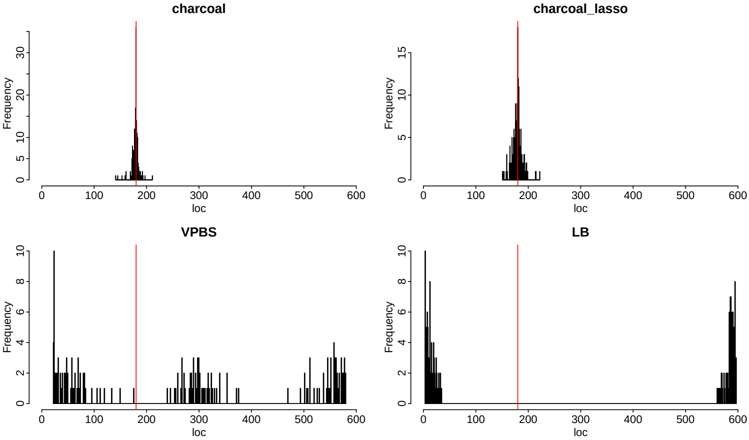


Figure: $n = 600$, $p = 200$, $z = 180$, $k = 14$, $\rho = 2$

Empirical performances

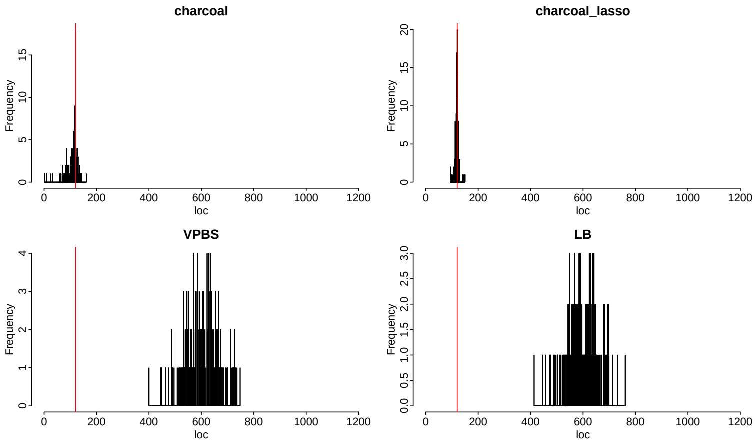
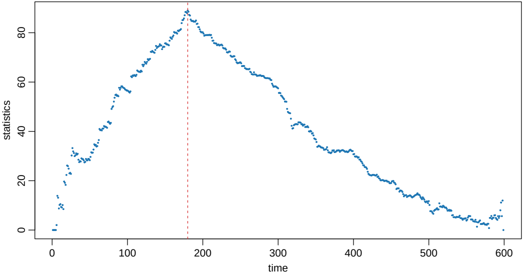


Figure: $n = 1200$, $p = 1000$, $z = 120$, $k = 31$, $\rho = 8$

Theoretical analysis

- $\tilde{W}_t^\top Z = \{\mathrm{diag}(W_t^\top W_t)\}^{-1/2}\big(W_t^\top W_z\theta + W_t^\top \xi\big)$
- **Key step**: show that $W_t^\top W_z$ is close to $4t(n-z)(n-p)n^{-2}I_p$ in k-operator norm uniformly over t.
- Hence $H_t := \|\textbf{soft}(\tilde{W}_t^\top Z, \lambda)\|_2$ is close to $\tilde{H}_t := \|(\tilde{W}_t^\top W_z)_{S,S}\theta_S\|_2$, which can be in turn shown to be approximately

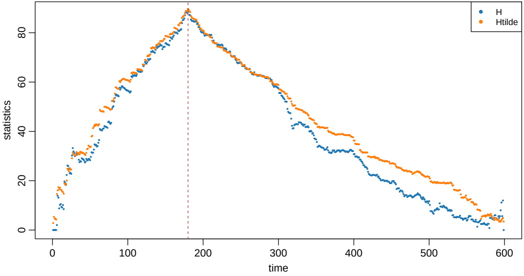
$$h_t := \frac{4(n-p)}{n}\left\{\sqrt{\frac{t}{n(n-t)}}(n-z)\mathbb{1}_{\{t\le z\}} + \sqrt{\frac{n-t}{nt}}z\mathbb{1}_{\{t>z\}}\right\}.$$



Theoretical analysis

- $\tilde{W}_t^\top Z = \{\mathrm{diag}(W_t^\top W_t)\}^{-1/2}\big(W_t^\top W_z\theta + W_t^\top \xi\big)$
- **Key step**: show that $W_t^\top W_z$ is close to $4t(n-z)(n-p)n^{-2}I_p$ in k-operator norm uniformly over t.
- Hence $H_t := \|\mathbf{soft}(\tilde{W}_t^\top Z, \lambda)\|_2$ is close to $\tilde{H}_t := \|(\tilde{W}_t^\top W_z)_{S,S}\theta_S\|_2$, which can be in turn shown to be approximately

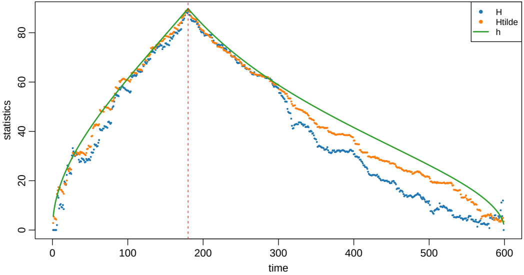
$$h_t := \frac{4(n-p)}{n}\left\{\sqrt{\frac{t}{n(n-t)}}(n-z)\mathbb{1}_{\{t\le z\}} + \sqrt{\frac{n-t}{nt}}z\mathbb{1}_{\{t>z\}}\right\}.$$



Theoretical analysis

- $\tilde{W}_t^\top Z = \{\mathrm{diag}(W_t^\top W_t)\}^{-1/2}\big(W_t^\top W_z\theta + W_t^\top \xi\big)$
- **Key step**: show that $W_t^\top W_z$ is close to $4t(n-z)(n-p)n^{-2}I_p$ in k-operator norm uniformly over t.
- Hence $H_t := \|\mathbf{soft}(\tilde{W}_t^\top Z, \lambda)\|_2$ is close to $\tilde{H}_t := \|(\tilde{W}_t^\top W_z)_{S,S}\theta_S\|_2$, which can be in turn shown to be approximately

$$h_t := \frac{4(n-p)}{n}\left\{\sqrt{\frac{t}{n(n-t)}}(n-z)\mathbb{1}_{\{t\le z\}} + \sqrt{\frac{n-t}{nt}}z\mathbb{1}_{\{t>z\}}\right\}.$$



Assumptions

(A1) Random design: $x_t \sim N_p(0, I_p)$ independently for $t = 1, \ldots, n$

(A2) Asymptotic regime: n, z, p satisfies $p < n$ and $z/n \to \tau \in (0,1)$ and $(n-p)/n \to \eta \in (0,1)$ as $n \to \infty$.

Theorem. Assume Conditions (A1) and (A2). Suppose that $\|\theta\|_2 \leq 1$, $k \leq p/2$. There exists $c, C > 0$, depending only on τ, η, such that if $\lambda > c\sigma \log p$, then asymptotically with probability 1, for all but finitely many n's, we have

$$\frac{|\hat{z}^{\mathrm{corr}} - z|}{n} \leq \frac{C\lambda\sqrt{k}}{\sqrt{n}\|\theta\|_2}.$$

Theoretical analysis

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Theorem. Under the same condition as above, There exists $c, C > 0$, depending only on τ, η, such that if $\lambda > c\sigma \log p$, then asymptotically with probability 1, for all but finitely many n's, we have

$$\sin \angle(\hat{v}^{\mathrm{proj}}, \theta) \leq \frac{C\lambda\sqrt{k}}{\sqrt{n}\|\theta\|_2}.$$

Hence, $\hat{z}^{\mathrm{proj}}$ satisfies

$$\frac{|\hat{z} - z|}{n} \leq \frac{C\lambda^2\sqrt{k}\log p}{\sqrt{n}\|\theta\|_2^2}.$$

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Theorem. Under the same condition as above, There exists $c, C > 0$, depending only on τ, η, such that if $\lambda > c\sigma \log p$, then asymptotically with probability 1, for all but finitely many n's, we have

$$\sin \angle(\hat{v}^{\mathrm{proj}}, \theta) \leq \frac{C\lambda\sqrt{k}}{\sqrt{n}\|\theta\|_2}.$$

Hence, a sample-splitting variant of $\hat{z}^{\mathrm{proj}}$ satisfies

$$\frac{|\hat{z} - z|}{n} \leq \frac{C\lambda\sqrt{k}\log p}{\sqrt{n}\|\theta\|_2}.$$

Summary

- It is possible to estimate sparse changes in high-dimensional regression coefficients, even if the coefficients themselves are dense.
- Use complementary sketching to eliminate nuisance parameter.
- Future work
  - Multiple changepoints / non-GOE design
  - Can the rate of convergence be improved?
  - Theory for $\hat{z}^{\mathrm{lasso}}$?

Summary

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

- It is possible to estimate sparse changes in high-dimensional regression coefficients, even if the coefficients themselves are dense.
- Use complementary sketching to eliminate nuisance parameter.
- Future work
  - Multiple changepoints / non-GOE design
  - Can the rate of convergence be improved?
  - Theory for $\hat{z}^{\mathrm{lasso}}$?

- Main references:

Gao, F. and Wang, T. (2021) Two-sample testing of high-dimensional linear regression coefficients via complementary sketching. *arXiv preprint*, arxiv:2011.13624.

Gao, F. and Wang, T. (2022) Sparse change detection in high-dimensional linear regression. *In preparation.*

Thank you!

References

- Bai, J. and Perron, P. (1998) Estimating and testing linear models with multiple structural changes. *Econometrica*, **66**, 47–78.
- Bühlmann, P. and van de Geer, S. (2011) *Statistics for High-dimensional Data: Methods, Theory and Applications.* Springer.
- Cho, H. and Fryzlewicz, P. (2015) Multiple changepoint detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B*, **77**, 475–507.
- Enikeeva, F. and Harchaoui, Z. (2019) High-dimensional change-point detection under sparse alternatives. *Ann. Statist.*, **47**, 2051–2079.
- Fan, J. and Lv, J. (2010) A selective overview of variable selection in high dimensional feature space. *Statist. Sinica*, **20**, 101–148.
- Fryzlewicz, P. (2021) Narrowest significant pursuit: inference for multiple changepoints in linear models. *arXiv preprint*, arxiv: 2009.05431.
- Krishnamurthy, A., Mazumdar, A., McGregor, A. and Pal, S. (2019) Sample complexity of learning mixture of sparse linear regressions. *Adv. Neur. Inform. Proc. Sys.*, **32**.
- Lee, S., Seo, M. H. and Shin, Y. (2016) The lasso for high dimensional regression with a possible change point. *J. Roy. Statist. Soc., Ser. B*, **78**, 193–210.

References

- Leonardi, F. and Bühlmann, P. (2016) Computationally efficient change point detection for high-dimensional regression. *arXiv preprint*, arXiv:1601.03704.
- Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527
- Rinaldo, A., Wang, D., Wen, Q., Willett, R. and Yu, Y. (2021) Localizing changes in high-dimensional regression models. In *International Conference on Artificial Intelligence and Statistics*, 2089-2097).
- Städler, N., Bühlmann, P. and van de Geer, S. (2010) ℓ_1-penalization for mixture regression models. *Test*, **19**, 209–256
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc., Ser. B*, **58**, 267–288.
- Wang, D., Zhao, Z., Lin, K. Z. and Willett, R. (2021) Statistically and computationally efficient change point localization in regression settings. *J. Mach. Learn. Res.*, **22**, 1–46.
- Wang, T. and Samworth, R. J. (2018) High-dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., Ser. B*, **80**, 57–83.
- Xia, Y., Cai, T. and Cai, T. T. (2018) Two-sample tests for high-dimensional linear regression with an application to detecting interactions. *Statist. Sinica*, **28**, 63–92.