High-dimensional data segmentation in regression settings permitting heavy tails and temporal dependence

Dom Owens[1]

Compass CDT & Institute for Statistical Science, University of Bristol

December 15, 2022

[1] dom.owens@bristol.ac.uk

Joint work with my supervisor, Dr. Haeran Cho

Introduction

- High dimensional linear regression models are widely used and studied
- Often applied to time series data
- Assumes stationarity of conditional relationship $\mathrm{E}[Y_t|\mathbf{x}_t]$
- This is unrealistic!

Motivating Example

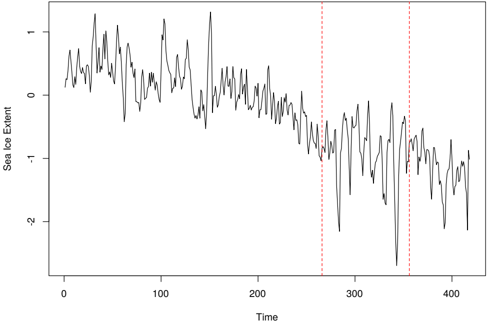


Figure: Monthly-adjusted Arctic sea ice extent, 1984-2018. Estimated change points marked in red.

Contributions

For the piecewise stationary regression model, we propose MOSEG, a novel 2-step algorithm for estimating change point numbers and locations.
This is

- Minimax-optimal under Gaussian design
- Consistent under heavy tails and dependence (functional dependence)
- Consistent under multiscale changes (Large & frequent / Small & rare in same series) with a bottom-up extension
- Lowest cost and runtime of all competing methods

Piecewise-Stationary Sparse Model

We observe $(Y_t, \mathbf{x}_t)$, $t = 1, \ldots, n$, with $\mathbf{x}_t = (X_{1t}, \ldots, X_{pt})^\top \in \mathbb{R}^p$ where

$$Y_t = \begin{cases} \mathbf{x}_t^\top \boldsymbol{\beta}_0 + \varepsilon_t & \text{for } \theta_0 = 0 < t \le \theta_1, \\ \mathbf{x}_t^\top \boldsymbol{\beta}_1 + \varepsilon_t & \text{for } \theta_1 < t \le \theta_2, \\ \vdots & \\ \mathbf{x}_t^\top \boldsymbol{\beta}_q + \varepsilon_t & \text{for } \theta_q < t \le n = \theta_{q+1}, \end{cases}$$

- For all j, $\boldsymbol{\beta}_{j-1} \neq \boldsymbol{\beta}_j$
- Possibly $p \gg n$
- $\boldsymbol{\beta}_j$ is sparse - at most $\mathfrak{s}$ non-zero entries
- Noise ε_t satisfies $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma_\varepsilon^2 \in (0, \infty)$
- $(\mathbf{x}_t, \varepsilon_t)$ possibly heavy tailed and dependent

Method: Step 1: Detector

Scan the data with detector

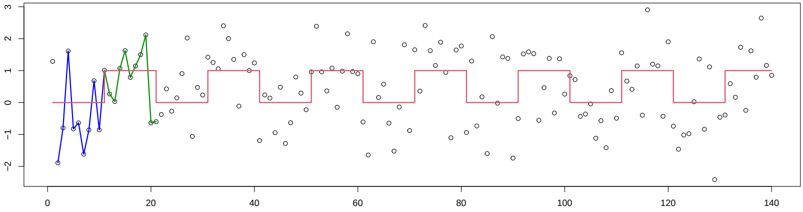
$$T_k(G) = \sqrt{\frac{G}{2}} \left| \widehat{\boldsymbol{\beta}}_{k,k+G} - \widehat{\boldsymbol{\beta}}_{k-G,k} \right|_2 ,$$

where G is chosen bandwidth, using Lasso solutions

$$\widehat{\boldsymbol{\beta}}_{s,e}(\lambda) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{t=s+1}^{e} (Y_t - \mathbf{x}_t^\top \boldsymbol{\beta})^2 + \lambda \sqrt{e-s} |\boldsymbol{\beta}|_1$$

for k in grid $\mathcal{T} \subset \{k : G \leq k \leq n - G\}$

Test signal `teeth10` (top) and MOSUM detector with $G = 10$:



Method: Step 1: Detector

Select local maximisers $\{\widetilde{\theta}_j\}_{j=1}^{\widehat{q}}$ with $T_{\widetilde{\theta}_j}$ exceeding threshold D as Step 1 estimators

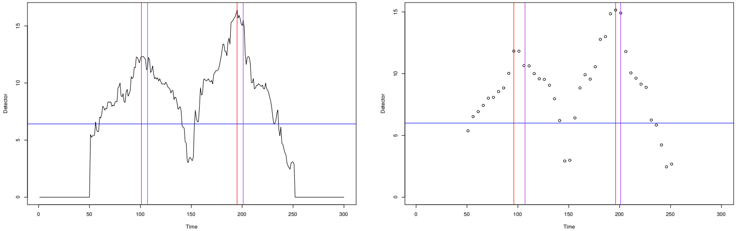


Figure: Left: $\mathcal{T} = \{G, 2G, \ldots, n-G\}$.
Right: $\mathcal{T} = \{G, (11/10)G, \ldots, n-G\}$.
Step 1 estimators in red; Step 2 in purple.

Method: Step 2: Location Refinement

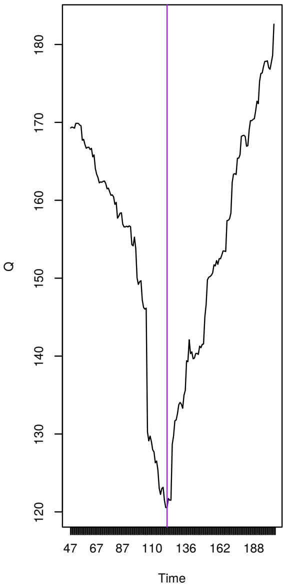
For $\widetilde{\theta}_j$, pick $\widehat{\boldsymbol{\beta}}^{\mathsf{L}}_j = \widehat{\boldsymbol{\beta}}_{0 \vee (\widetilde{\theta}^{\mathsf{L}}_j - G), \widetilde{\theta}^{\mathsf{L}}_j}$ and $\widehat{\boldsymbol{\beta}}^{\mathsf{R}}_j = \widehat{\boldsymbol{\beta}}_{\widetilde{\theta}^{\mathsf{R}}_j, (\widetilde{\theta}^{\mathsf{R}}_j + G) \wedge n}$ from either side.
Plug each $\{\widetilde{\theta}_j\}_{j=1}^{\widehat{q}}$ into left/right loss

$$Q\left(k; \widehat{\boldsymbol{\beta}}^{\mathsf{L}}, \widehat{\boldsymbol{\beta}}^{\mathsf{R}}\right) = \sum_{t=\widetilde{\theta}_j - G + 1}^{k} (Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}^{\mathsf{L}})^2 + \sum_{t=k+1}^{\widetilde{\theta}_j + G} (Y_t - \mathbf{x}_t^\top \widehat{\boldsymbol{\beta}}^{\mathsf{R}})^2$$

selecting $\widehat{\theta}_j = \arg\min_k Q$ as Step 2 estimator[2].

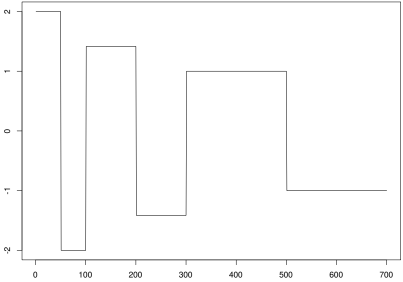
[2] Kaul et al. (2019) An efficient two step algorithm for high dimensional change point regression models without grid search. JMLR

Method: Step 2: Location Refinement



Multiscale Method

How should we pick the bandwidth? Changes could be *multiscale*, with size $\Delta^{(2)} = \min_j \delta_j^2 \cdot \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1})$ (where $\delta_j = |\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}|_2$)



Solution: run algorithm with multiple bandwidths $\mathcal{G} = \{G_1, \ldots, G_H\}$, merge results bottom-up

Assumptions

(1) $\mathrm{Cov}(\mathbf{x}_t) = \boldsymbol{\Sigma}_x$ has bounded eigenvalues

(2) Deviation bounds hold for $\left|\frac{1}{\sqrt{e-s}}\sum_{t=s+1}^{e}\varepsilon_t\mathbf{x}_t\right|_\infty$ and $\left|\frac{1}{\sqrt{e-s}}\sum_{t=s+1}^{e}(Y_t - \mathbf{x}_t^\top\beta^*_{s,e})\mathbf{x}_t\right|_\infty$

(3) Restricted strong convexity holds on all large enough pairs $e - s \geq C_0\rho_{n,p}^2$

- $\rho_{n,p} = \log^{2\gamma+3/2}(p \vee n), \gamma > 0$ under heavy tails[3]
- $\rho_{n,p} = \log^{1/2}(p \vee n)$ under (sub)Gaussian design

(4) Bandwidth

- $2G \leq \min_j(\theta_j - \theta_{j-1})$
- Multiscale: For each θ_j, $4G_{(j)} \leq \min(\theta_{j+1} - \theta_j, \theta_j - \theta_{j-1})$

and $\min_j \delta_j^2 G$ grows fast enough

[3] $\gamma < 1$ sub-exponential, $\gamma < 3/2$ sub-Weibull

Assumptions: moments and dependence

- Deviation bounds and Restricted strong convexity hold under bounded functional dependence (Zhang and Wu 2017), which holds under fairly general conditions on moments/dependence
- Example: vector moving average $\begin{bmatrix} \mathbf{x}_t \\ \varepsilon_t \end{bmatrix} = \sum_{\ell=0}^{\infty} \mathbf{D}_\ell \boldsymbol{\xi}_{t-\ell}$,
  - $|D_{\ell,ik}|$ decay algebraically as $\ell \to \infty$
  - innovations $\boldsymbol{\xi}_t$ (i) have finite moments, or (ii) are Gaussian

Results

- Under Gaussianity of $\boldsymbol{\xi}_t$, we have optimal (up to log factors)
  - Detection rate (Step 1): If $\min_j \delta_j^2 G \geq c\mathfrak{s}\log(p \vee n)$ then $\widehat{q} = q$
  - Localisation rate (Step 2): $\max_j \delta_j^2 |\widehat{\theta}_j - \theta_j| \leq C\mathfrak{s}\log(p \vee n)$
- Under heavy tails, we have localisation rate $C(\mathfrak{s}\log(p \vee n))^{4\gamma+3}$
- Step 1 estimators localise at sub-optimal rate
- In the multiscale setting, the rates are the same

Competitors

Typically, $\text{Lasso}(a, b) = O(b^3 + ab^2)$

	Multiscale	Computational complexity
MOSEG	No	$O(\frac{n}{rG} \cdot \text{Lasso}(G, p))$
MOSEG.MS	Yes	$O(\frac{n}{rG_1} \cdot \text{Lasso}(n, p))$
Wang et. al. 2021	No	$O(n \log^2(n) \cdot \text{GroupLasso}(n, p))$
Kaul et. al. 2019	No	$O(\tilde{q} \cdot \text{Lasso}(n, p) + \text{SA}(\tilde{q}))$
Xu et. al. 2022	No	$O(n^2 \text{Lasso}(n, p))$

Computation: selecting threshold and λ

Select $(\lambda, \widehat{q})$ using sample splitting

(1) Split into odd/even folds for testing/training

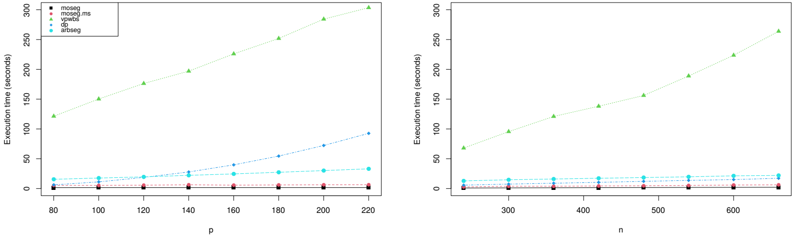
(2) Order candidate changes $\theta_{(1)}, \ldots, \theta_{(\widetilde{q}_0)}$ by descending detector value

(3) For $q = 0, 1, \ldots, \widetilde{q}_0$, fit model with q "biggest" changes on training set, predict for testing set

(4) Pick $(\lambda, \widehat{q})$ in $\Lambda \times \{0, \ldots, \widetilde{q}_0\}$ minimising error

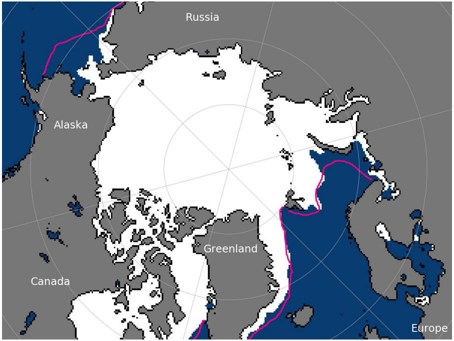
Coordinate descent (glmnet) gives $\widehat{\beta}_j(\lambda)$ for $\lambda \in \Lambda$ for free

Runtime



moseg
moseg.ms
vpwbs
dp
arbseg
Execution time (seconds)
0
50
100
150
200
250
300
80
100
120
140
160
180
200
220
p
Execution time (seconds)
300
400
500
600
n

Sea ice extent



4

- Arctic sea ice extent is retreating
- Influences the Arctic ecosystem
- Can model this as a dynamical system with e.g. weather covariates
- Piecewise stationarity is useful and interpretable

[4] Coulombe, P. G. and Göbel, M. (2021). Arctic amplification of anthropogenic forcing: a vector autoregressive analysis. Journal of Climate

Sea ice extent

$n = 418$ monthly observations, $p = 55$ features

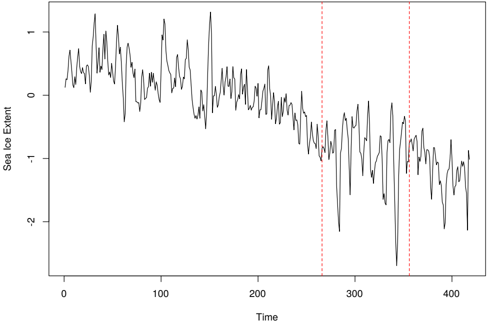


Figure: Monthly-adjusted Arctic sea ice extent, 1984-2018. Estimated change points marked in red.

Sea ice extent

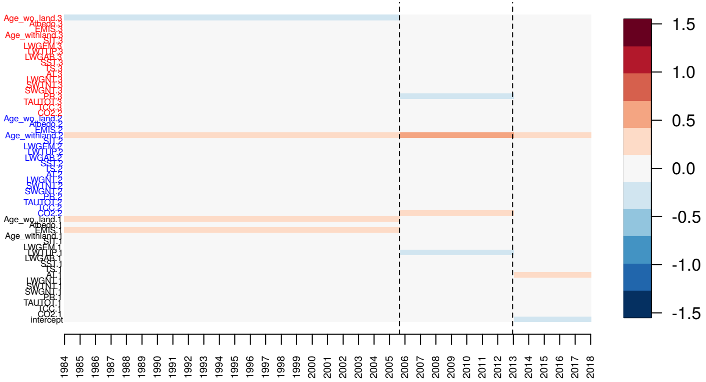


Figure: Parameter estimates from each estimated segment obtained by MOSEG.MS. Variables at different lags are coloured differently in the y-axis.

Conclusion

- We propose a two-step method for data segmentation under the sparse regression model
- Achieves minimax optimal detection and localisation, and is consistent under dependence and heavy tails
- Extends to multiscale changes
- Cheaper and faster than competitors
- Preprint available on ArXiv, R package MOSEG on github