Inferring Change Points in High-Dimensional Linear Regression via Approximate Message Passing

Gabriel Arpino, Xiaoqi (Shirley) Liu, Ramji Venkataramanan {ga442, xl394, rv285}@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Change Points Model

Motivation: Vast amounts of time-ordered, non-stationary data, e.g., stock prices

For $i = 1, 2, \ldots, n$,

$\boldsymbol{x}_i, \boldsymbol{\beta}^{(l)} \in \mathbb{R}^p$

$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(1)} + \varepsilon_i$

$i = \eta_1$

$y_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}^{(2)} + \varepsilon_i$

$i = \eta_{L^*-1}$

$\boldsymbol{\eta} := [\eta_1, \ldots, \eta_{L^*-1}]$

L^* unknown, given $L^* < L$

Goal: recover change point locations $\{\eta_l\}_{l=1}^{L^*-1}$ from $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$, and estimate signals $\{\boldsymbol{\beta}^{(l)}\}_{l=1}^{L^*}$

High dimensional regime: $n, p \to \infty$, $n/p \to$ fixed constant δ

Existing work

- penalised maximum-likelihood optimisation + partitioning algorithms e.g., [Rinaldo et al. 2021, Xu et al. 2022]
- complementary sketching [Gao and Wang 2022]

Limitations:

- Restricted to certain signal priors (e.g., sparse $\boldsymbol{\beta}^{(l)}$)
- Provide point estimate, without uncertainty quantifications

Our work

Change point inference via Approximate Message Passing (AMP)

Define $\boldsymbol{B} = [\boldsymbol{\beta}^{(1)} \ \ldots \ \boldsymbol{\beta}^{(L)}]$, rows of $\boldsymbol{B} \overset{\text{iid}}{\sim} p_{\bar{\boldsymbol{B}}}$,

$\boldsymbol{\Theta} := \boldsymbol{X}\boldsymbol{B}$, $\boldsymbol{\Theta} \sim p_{\bar{\boldsymbol{\Theta}}}$

AMP iteratively produces:

$$\boldsymbol{\Theta}^t = \boldsymbol{X}\hat{\boldsymbol{B}}^t - \boldsymbol{R}^{t-1}\left(\boldsymbol{F}^t\right)^\top$$
$$\boldsymbol{R}^t = g^t\left(\boldsymbol{\Theta}^t, \boldsymbol{y}\right)$$

produce residual $\boldsymbol{R}^t$

infer change points

$$\boldsymbol{B}^{t+1} = \boldsymbol{X}^\top \boldsymbol{R}^t - \hat{\boldsymbol{B}}^t\left(\boldsymbol{C}^t\right)^\top$$
$$\hat{\boldsymbol{B}}^t = f^t\left(\boldsymbol{B}^t\right)$$

produce signal estimate $\hat{\boldsymbol{B}}^t$

Theorem (informal):

$$\boldsymbol{B}^t \overset{\mathbb{P}}{\simeq} \boldsymbol{B}\boldsymbol{\nu}_B^t + \boldsymbol{G}_B^t \qquad \text{rows of } \boldsymbol{G}_B^t \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{\kappa}_B^t)$$

$$\boldsymbol{\Theta}^t \overset{\mathbb{P}}{\simeq} \bar{\boldsymbol{\Theta}}\boldsymbol{\nu}_\Theta^t + \boldsymbol{G}_\Theta^t \qquad \text{rows of } \boldsymbol{G}_\Theta^t \overset{\text{iid}}{\sim} N(\boldsymbol{0}, \boldsymbol{\kappa}_\Theta^t)$$

$\boldsymbol{\nu}_B^t, \boldsymbol{\nu}_\Theta^t, \boldsymbol{\kappa}_B^t, \boldsymbol{\kappa}_\Theta^t$ can be computed deterministically; depend on $\boldsymbol{\eta}$

Bayesian approach for choosing denoisers f^t, g^t

Prior info about change points $\boldsymbol{\eta} \sim p_{\bar{\boldsymbol{\eta}}}$

(Informal): choose f^t as the Bayes-optimal estimator of signal $\boldsymbol{B}\boldsymbol{\nu}_B^t$ embedded in Gaussian noise $\boldsymbol{G}_B^t$:

$$f_j^t(\boldsymbol{B}_j^t) = \mathbb{E}\left[\bar{\boldsymbol{B}} \,|\, \bar{\boldsymbol{B}}\bar{\boldsymbol{\nu}}_B^t + \bar{\boldsymbol{G}}_{B,j}^t = \boldsymbol{B}_j^t\right]$$

where $\bar{\boldsymbol{\nu}}_B^t := \mathbb{E}_{\bar{\boldsymbol{\eta}}}[\boldsymbol{\nu}_B^t(\bar{\boldsymbol{\eta}})]$

rows of $\bar{\boldsymbol{G}}_B^t \sim N(\boldsymbol{0}, \bar{\boldsymbol{\kappa}}_B^t)$, $\bar{\boldsymbol{\kappa}}_B^t := \mathbb{E}_{\bar{\boldsymbol{\eta}}}[\boldsymbol{\kappa}_B^t(\bar{\boldsymbol{\eta}})]$

(Informal): choose g^t using prior knowledge $\boldsymbol{\eta} \sim p_{\bar{\boldsymbol{\eta}}}$ of change point locations:

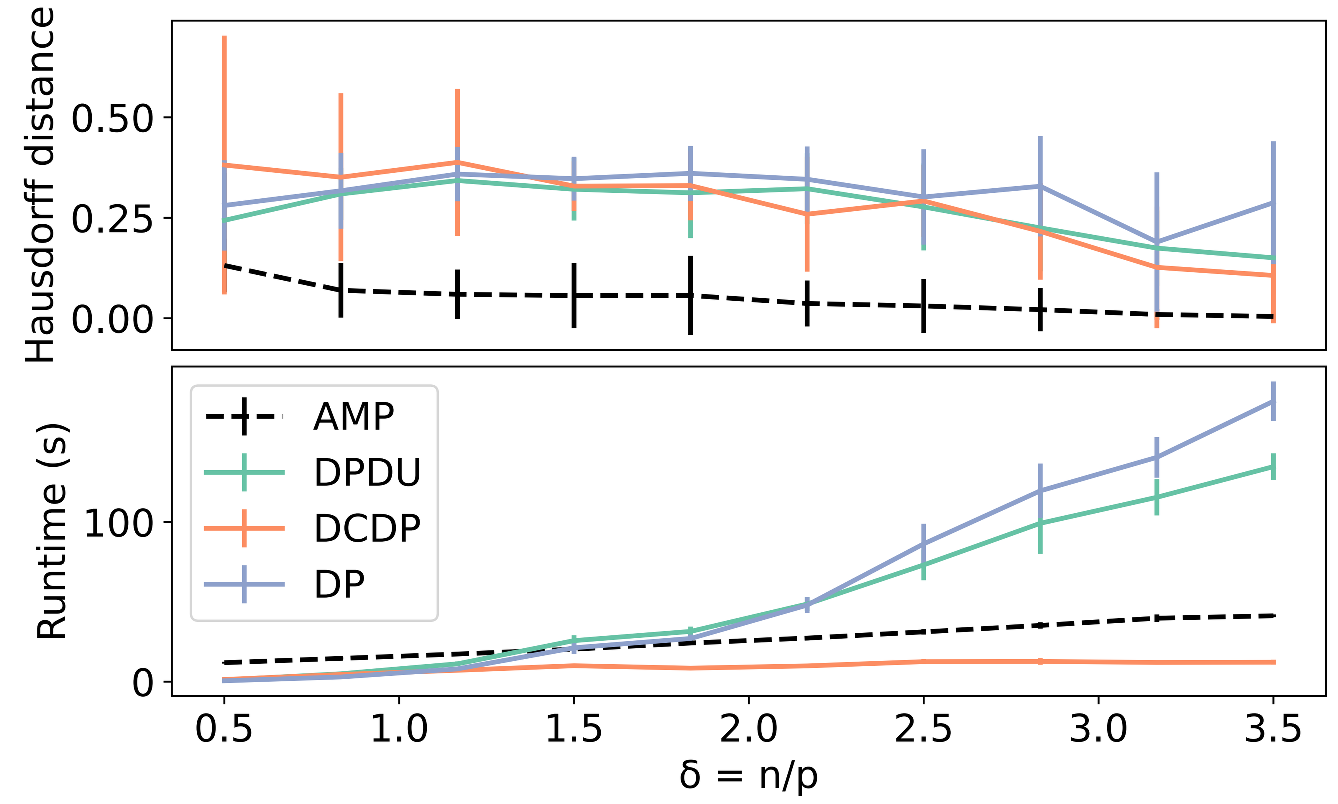
$$g_i^t(\boldsymbol{\Theta}^t, \boldsymbol{y}) = \mathbb{E}\left[\bar{\boldsymbol{\Theta}}_i \,|\, \bar{y}_i(\bar{\boldsymbol{\Theta}}_i, \bar{\boldsymbol{\eta}}) = y_i,\ \bar{\boldsymbol{\Theta}}_i\bar{\boldsymbol{\nu}}_\Theta^t + \bar{\boldsymbol{G}}_{\Theta,i}^t = \boldsymbol{\Theta}_i^t\right] \cdot \text{other terms}$$

g_i^t is explicitly dependent on $\bar{\boldsymbol{\eta}}$, row-wise **non-separable** due to sequential structure in $\bar{\boldsymbol{\eta}}$

Comparison with existing work

sparse $p_{\bar{\boldsymbol{B}}}$, $L^* = 3$, $p = 200$

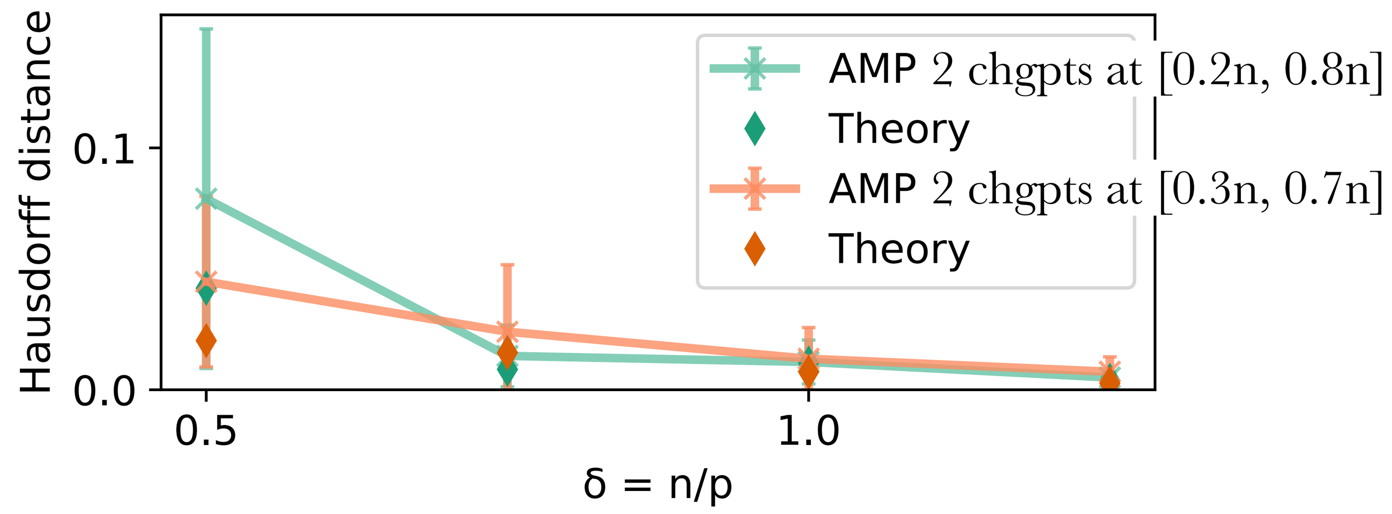
$d_H(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}})/n$



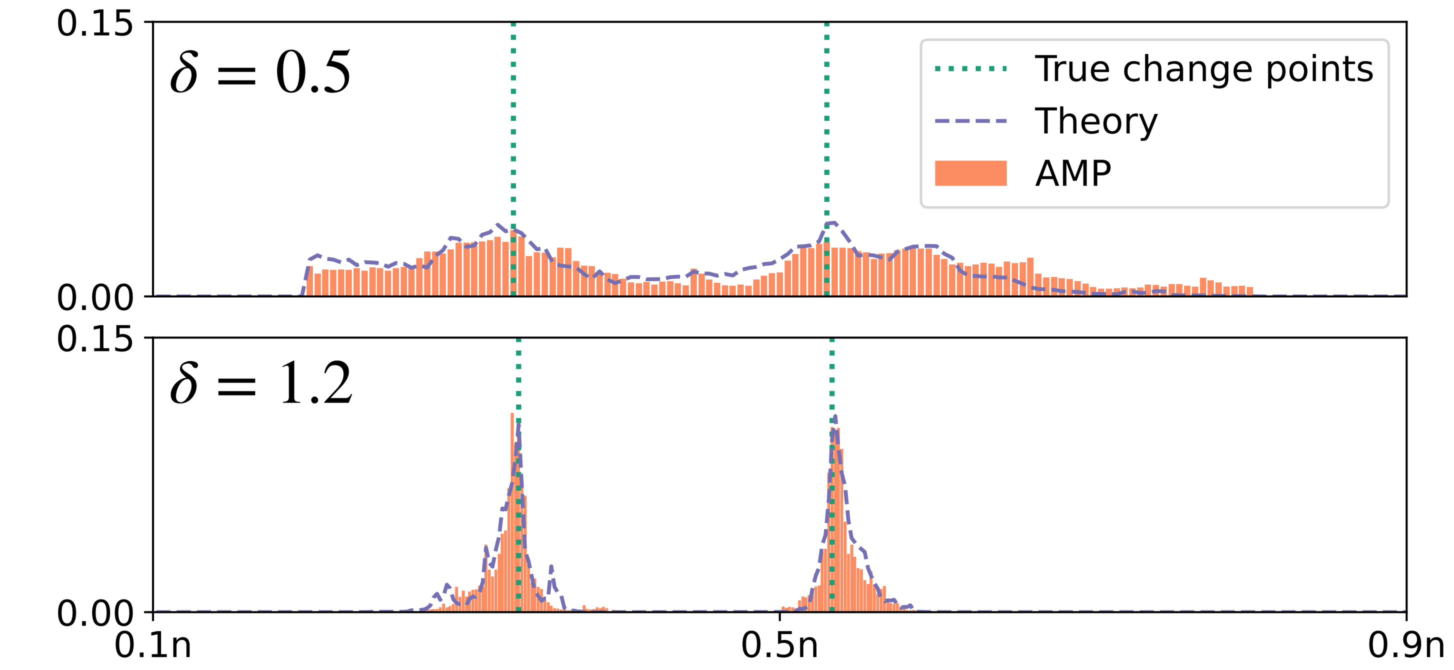
Estimation performance (Hausdorff distance d_H)

Proposition (informal):

$$d_H\left(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\boldsymbol{\Theta}^t, \boldsymbol{y})\right)/n \overset{\mathbb{P}}{\to} \mathbb{E}_{\boldsymbol{V}_\Theta^t, \bar{\boldsymbol{\Theta}}}\left[d_H\left(\boldsymbol{\eta}, \hat{\boldsymbol{\eta}}(\bar{\boldsymbol{\Theta}}\boldsymbol{\nu}_\Theta^t + \boldsymbol{G}_\Theta^t, \bar{\boldsymbol{y}}(\bar{\boldsymbol{\Theta}}, \boldsymbol{\eta}))\right)/n\right]$$



Inference performance (posterior)



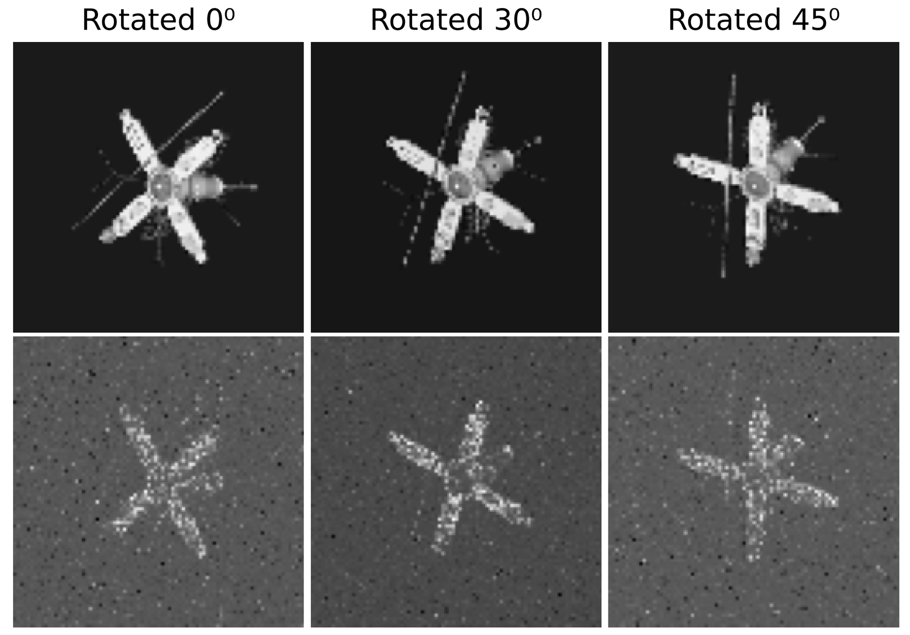
$p_{\bar{\boldsymbol{B}}} = N(\boldsymbol{0}, \boldsymbol{I})$, 2 chgpts at $n/3, 8n/15$, $p = 400$

$L = L^* = 3$, $p_{\bar{\boldsymbol{\eta}}}$ = uniform prior over all configs with chgpts at least $n/5$ apart

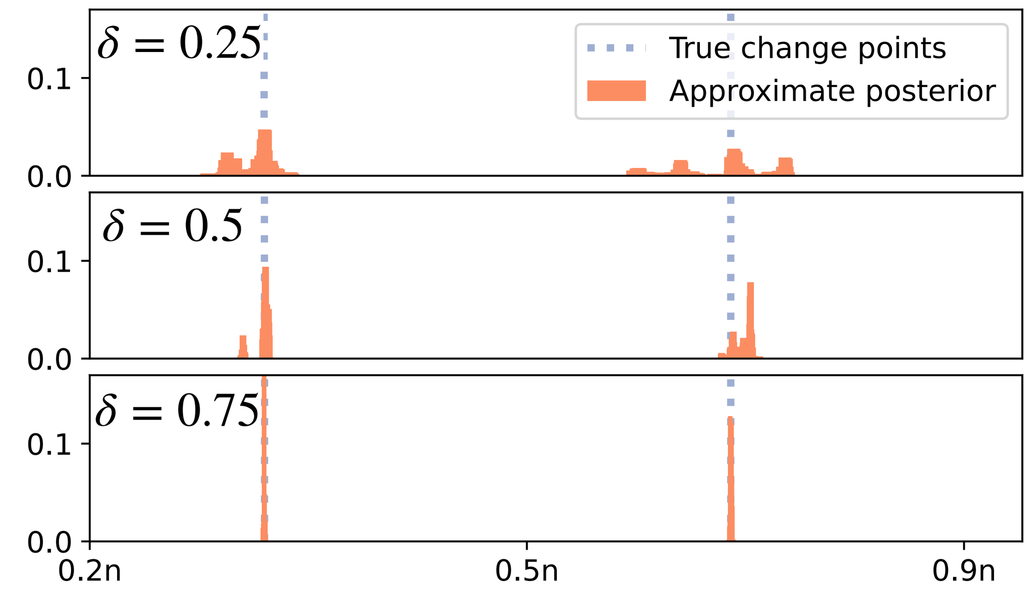
Experiments on real images

Ground truth signals:

Signals' estimate $\hat{\boldsymbol{B}}^t$ after 10 AMP iterations:



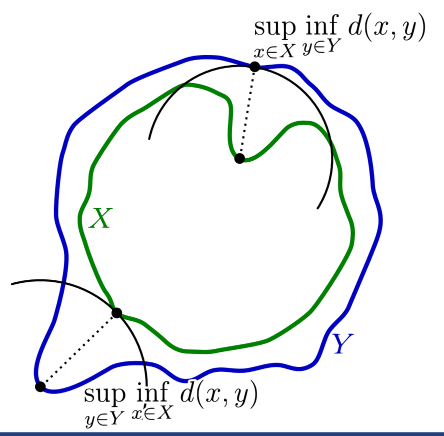
Posterior:



Hausdorff distance d_H

Measure of distance between two subsets X, Y of a metric space:

$$d_H(X, Y) := \max\left\{\sup_{x \in X} d(x, Y), \sup_{y \in Y} d(X, y)\right\}$$



References

- Rinaldo, A., Wang, D., Wen, Q., Willett, R., and Yu, Y. **Localizing Changes in High-Dimensional Regression Models.** In Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, 2021.
- Xu, H., Wang, D., Zhao, Z., and Yu, Y. **Change point inference in high-dimensional regression models under temporal dependence**, 2022. arXiv:2207.12453.
- Gao, F. and Wang, T. **Sparse change detection in high-dimensional linear regression**, 2022. arXiv:2208.06326.