# Inferring change points in high-dimensional linear regression via Approximate Message Passing

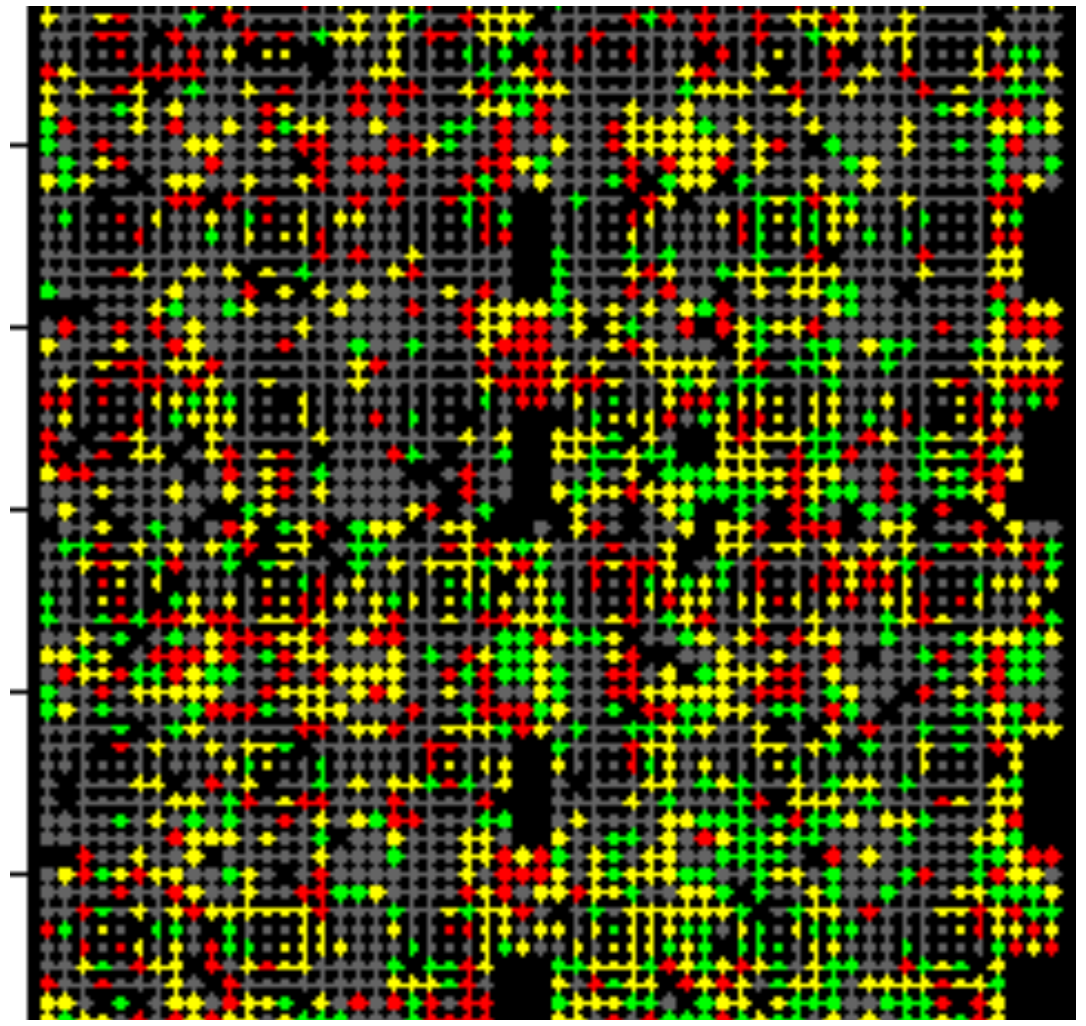
Gabriel Arpino, Xiaoqi (Shirley) Liu,
Ramji Venkataramanan

University of Cambridge

International Conference on Machine Learning (ICML) 2024, July 2024

Vast amounts of **time-ordered, non-stationary** data

Stock prices


Gene expression

Want to understand how underlying **data generating mechanisms** change over time

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

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

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

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

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

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

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

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

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

$$i = \eta_1$$

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

$$\vdots$$

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

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

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

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

**Scaling regime**: $n, p \to \infty$, $n/p \to$ fixed constant $\delta$

$$\hat{\eta}_1, \ldots, \hat{\eta}_{L^*-1} = \underset{\tilde{\eta}_1 \leq \tilde{\eta}_2 \leq \ldots \leq \tilde{\eta}_{L^*-1}}{\arg\min} \sum_{l=1}^{L^*} \min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p} \sum_{i=\tilde{\eta}_{l-1}+1}^{\tilde{\eta}_l} \left( y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} \right)^2$$

search through different partitioning

Residual Sum of Squares

$p \ll n$

$$\hat{\eta}_1, \ldots, \hat{\eta}_{L^*-1} = \underset{\tilde{\eta}_1 \le \tilde{\eta}_2 \le \ldots \le \tilde{\eta}_{L^*-1}}{\arg\min} \sum_{l=1}^{L^*} \min_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p} \sum_{i=\tilde{\eta}_{l-1}+1}^{\tilde{\eta}_l} \left( y_i - \boldsymbol{x}_i^\top \tilde{\boldsymbol{\beta}} \right)^2$$

search through different partitioning

~~Residual Sum of Squares~~ ~~$p \ll n$~~

penalised maximum likelihood e.g., LASSO

$n, p \to \infty$

[Zhang et al. 2015, Leonardi and Bühlmann 2016, Lee et al. 2016, Rinaldo et al. 2021]

Sparse prior: penalised maximum likelihood + partitioning

[Zhang et al. 2015, Leonardi and Bühlmann 2016, Lee et al. 2016, Rinaldo et al. 2021]

Sparse prior: penalised maximum likelihood + partitioning

[Zhang et al. 2015, Leonardi and Bühlmann 2016, Lee et al. 2016, Rinaldo et al. 2021]

Sparse difference prior: complementary sketching

[Wang et al. 2021, Gao and Wang 2022]

Limitations:

- Restricted to certain signal priors
- Provide point estimates, without uncertainty quantification

We address these limitations via:

Approximate Message Passing (AMP)

[Donoho et al. 2009, Bayati and Montanari 2011, Rangan 2011, Berthier et al. 2019, Celentano and Montanari 2022, Feng et al. 2022, Gerbelot and Berthier 2023]

# Aside: AMP for sparse recovery

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

entries of $\boldsymbol{\beta} \overset{\text{iid}}{\sim} p_{\bar{\beta}}$

entries of $\boldsymbol{X} \overset{\text{iid}}{\sim} N(0,1/n)$

modified residual: $\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t + b_t \boldsymbol{r}^{t-1}$

signal estimate: $\boldsymbol{\beta}^{t+1} = \eta_{\substack{\text{soft}\\ \text{thres}}}\left(\boldsymbol{\beta}^t + \boldsymbol{X}^\top \boldsymbol{r}^t; \theta_t\right)$

# Aside: AMP for sparse recovery

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

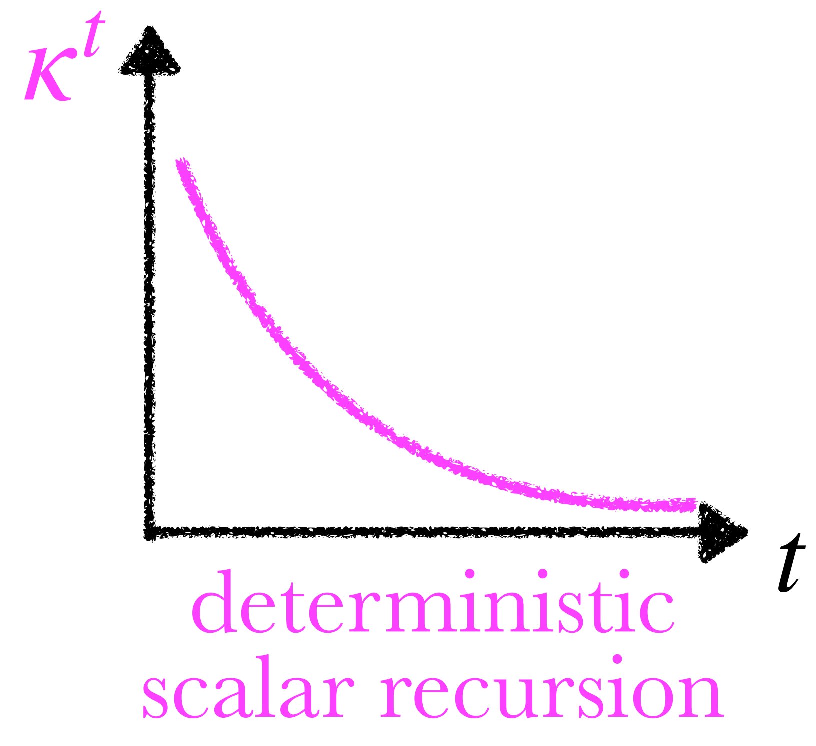
entries of $\boldsymbol{\beta} \overset{\text{iid}}{\sim} p_{\bar{\beta}}$

entries of $\boldsymbol{X} \overset{\text{iid}}{\sim} N(0, 1/n)$

modified residual: $\boldsymbol{r}^t = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^t + b_t \boldsymbol{r}^{t-1}$ — debias term ($+b_t \boldsymbol{r}^{t-1}$)

signal estimate: $\boldsymbol{\beta}^{t+1} = \eta_{\text{soft thres}}\left(\boldsymbol{\beta}^t + \boldsymbol{X}^\top \boldsymbol{r}^t; \theta_t\right)$

distribution of entries of $\boldsymbol{\beta}^t + \boldsymbol{X}^\top \boldsymbol{r}^t - \boldsymbol{\beta}$ $\rightarrow N(0, \kappa^t)$



# AMP for change point inference

Signal is $\boldsymbol{B} = \begin{bmatrix} \uparrow & & \uparrow \\ \beta^{(1)} & \dots & \beta^{(L)} \\ \downarrow & & \downarrow \end{bmatrix} \sim p_{\bar{B}}$ and define $\boldsymbol{\Theta} := \boldsymbol{X}\boldsymbol{B}$.

$g^t$ explicitly uses prior info about change points $\boldsymbol{\eta} \sim p_{\bar{\eta}}$

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

$$\boldsymbol{R}^t = g^t\left(\boldsymbol{\Theta}^t, \boldsymbol{y}\right)$$

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 $\hat{\boldsymbol{B}}^t$

Bayes-optimal $f^t\left(\boldsymbol{B}^t\right) = \mathbb{E}\left[\bar{\boldsymbol{B}} \mid \bar{\boldsymbol{B}} + \boldsymbol{G}_{\boldsymbol{B}}^t = \boldsymbol{B}^t\right]$

Theorem (informal)

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

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

# AMP for change point inference

Signal is $\boldsymbol{B} = \begin{bmatrix} \uparrow & & \uparrow \\ \beta^{(1)} & \ldots & \beta^{(L)} \\ \downarrow & & \downarrow \end{bmatrix} \sim p_{\bar{B}}$ and define $\boldsymbol{\Theta} := X\boldsymbol{B}$.

$g^t$ explicitly uses prior info about change points $\boldsymbol{\eta} \sim p_{\bar{\eta}}$

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

$$\boldsymbol{R}^t = g^t\left(\boldsymbol{\Theta}^t, \boldsymbol{y}\right)$$

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 $\hat{\boldsymbol{B}}^t$

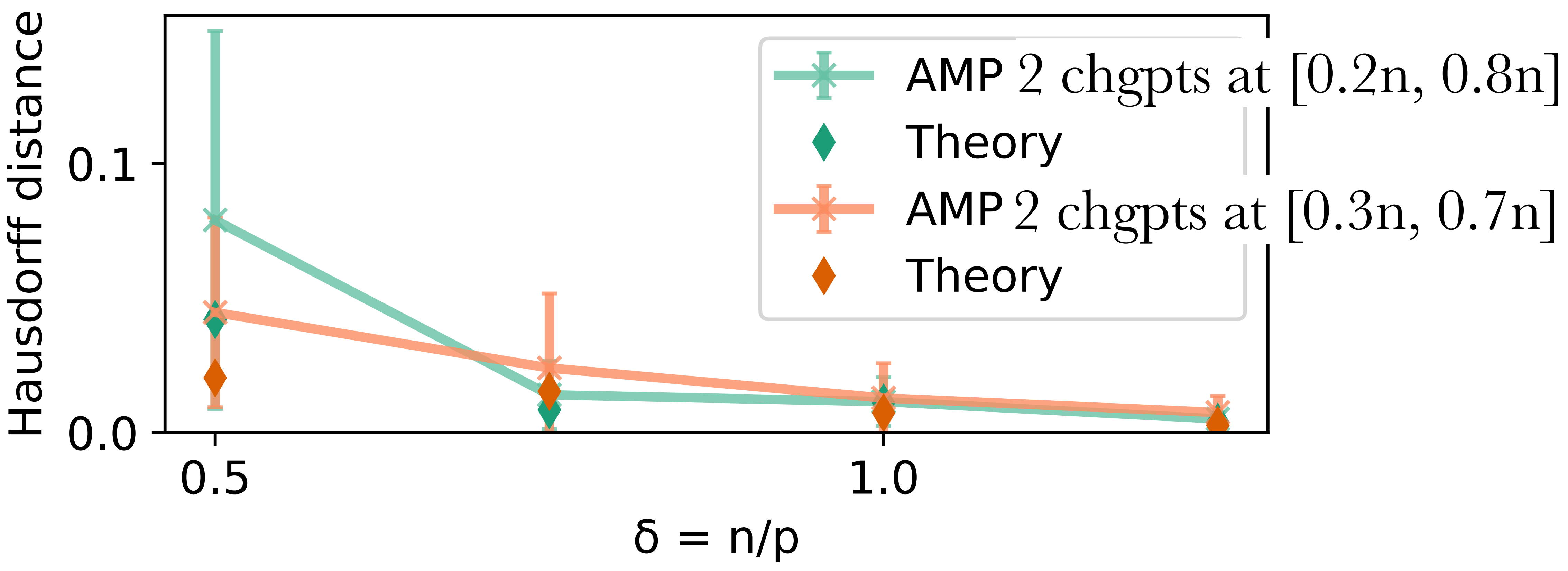
Proposition (informal) (Hausdorff distance $d_H$)

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

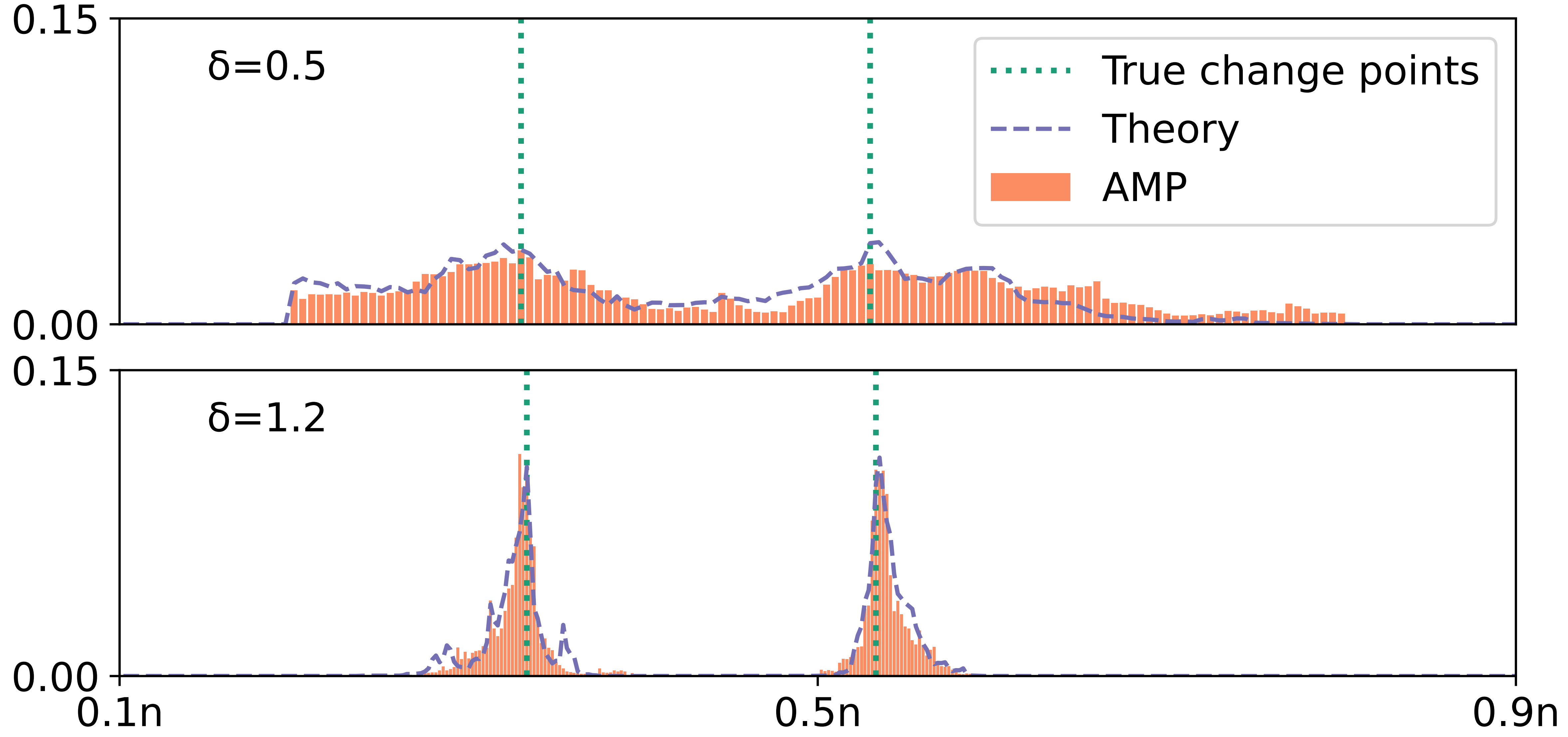
Similar result in terms of posterior $p(\boldsymbol{\eta} \,|\, \boldsymbol{\Theta}^t, \boldsymbol{y})$

# Simulation results

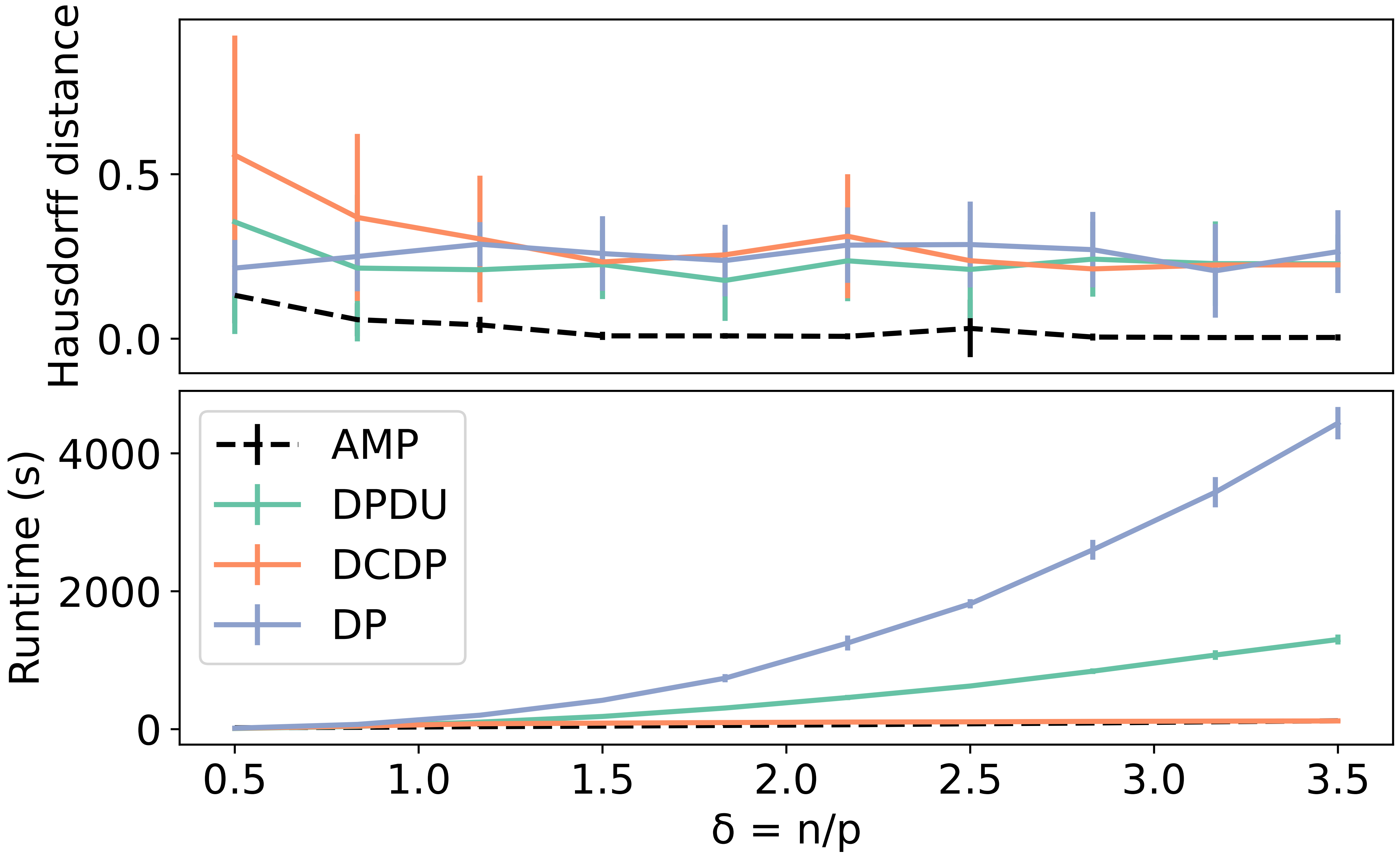
## Hausdorff distance




## Posterior



# Comparison with other algorithms



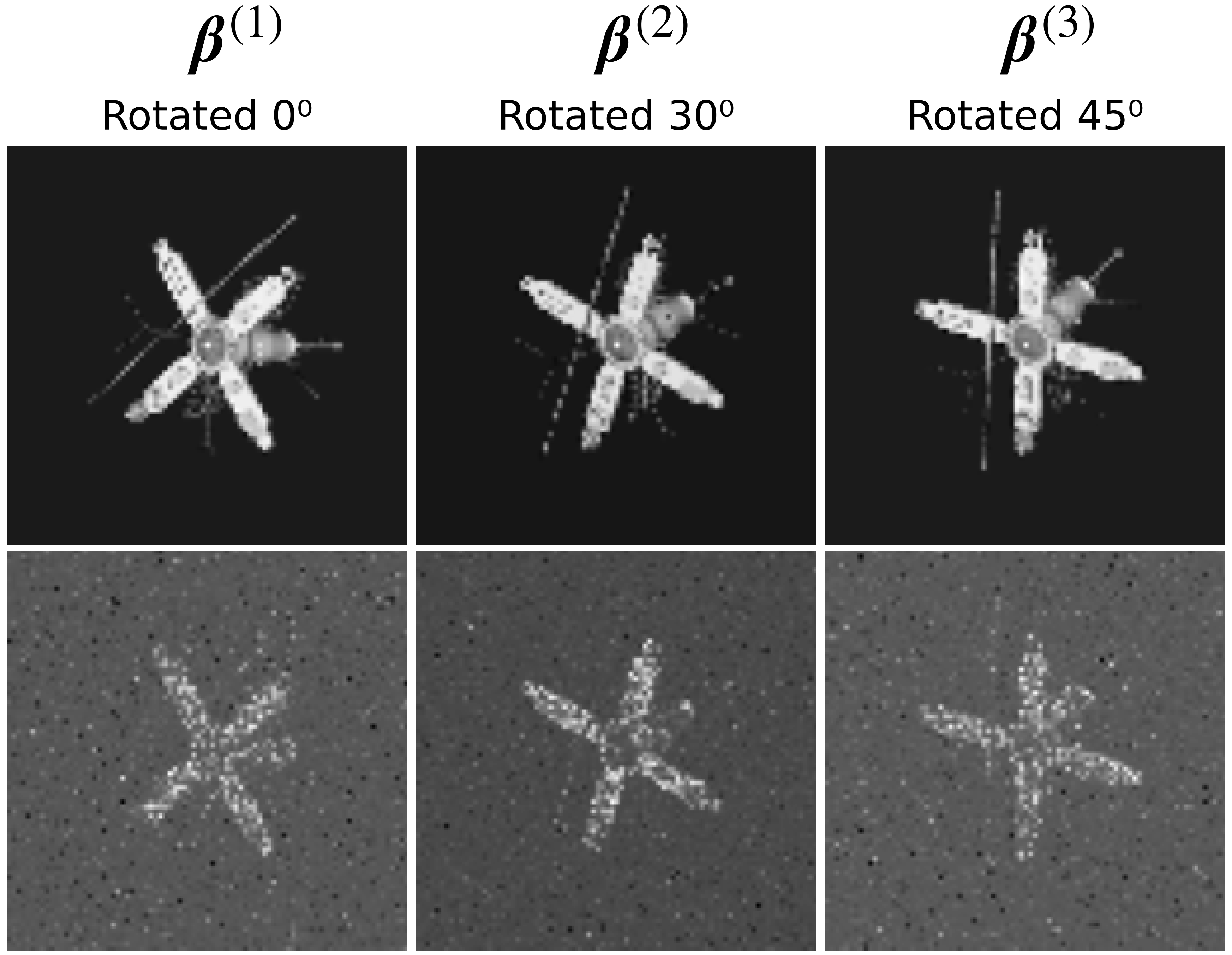

Computational cost: LASSO+partitioning: $O(n^2p^2)$, $O(n^3)$

AMP: $O(np)$

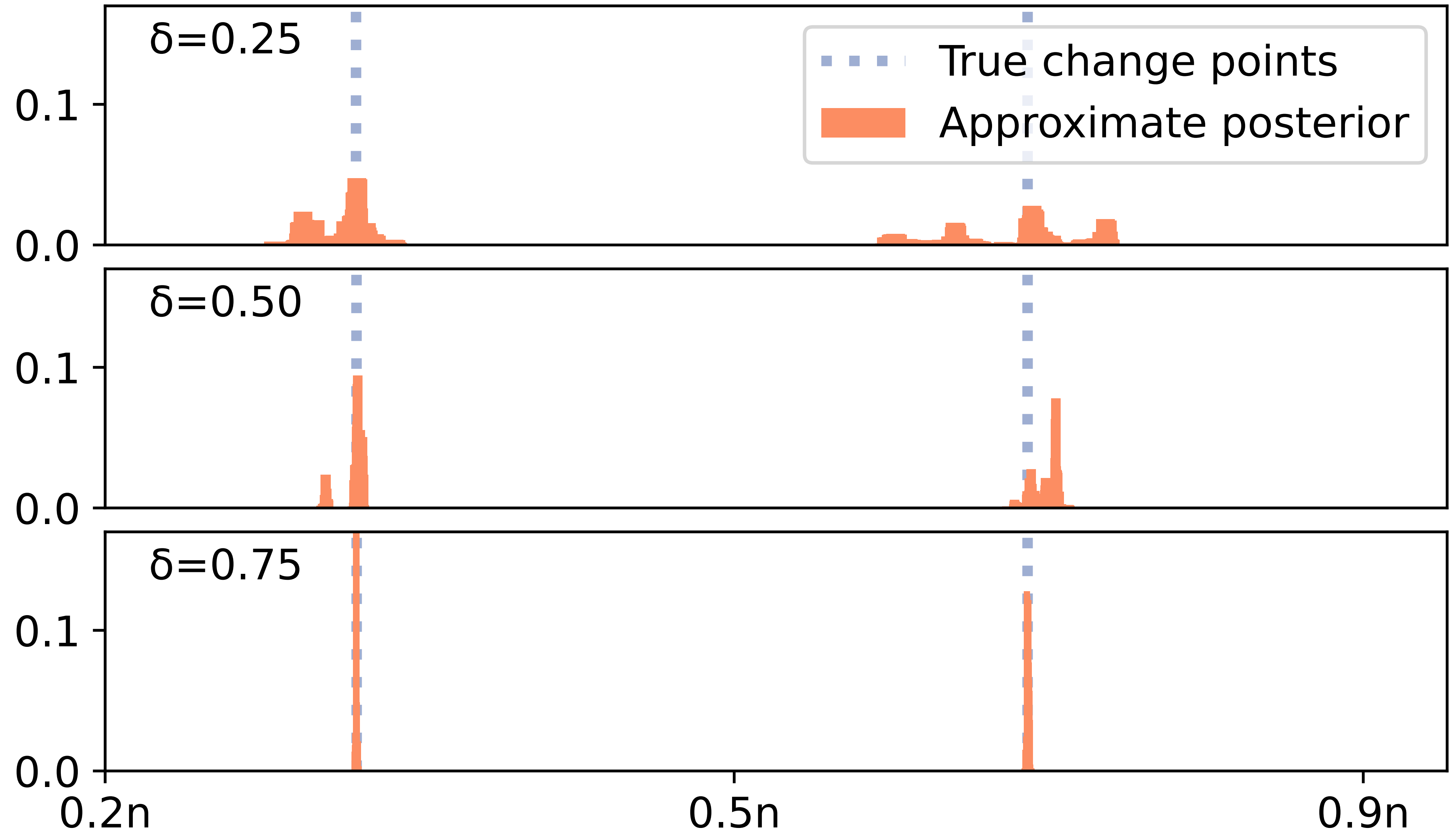
# Experiments on real images

Ground truth signals:

Estimated signals after 10 AMP iterations:




Posterior:



# Summary


First work to apply AMP algorithms to change point inference:

$(n, p \to \infty, n/p \to \delta)$

- Bayesian approach to choosing denoisers $g^t, f^t$
- Near-optimal computational complexity: $O(np)$
- Exact asymptotic performance guarantees (e.g., Hausdorff distance, posterior)

**Future work:**

generalised linear models,

beyond iid Gaussian $\boldsymbol{X}$ matrices,

online change point detection…

# Hausdorff distance

For $X, Y \subset$ metric space $(M, d)$,

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

where $d(a, B) := \inf_{b \in B} d(a, b)$