A Sharp Oracle Inequality for Graph-Slope

Pierre C. Bellec, Joseph Salmon & **Samuel Vaiter**[1]

2020/01/16 @GIPSA-LAB

[1]CNRS & IMB, Dijon

The One-Minute Talk

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n} \|y - \beta\|^2 + \lambda J(\mathbf{D}^\top \beta)$$

Graph-Lasso ⟵ Lasso [Tibshirani ’95, Donoho '95]

new estimator: **Graph-Slope** ⟵ Slope [Bogdan et al. ’14]

better statistical properties (oracle inequality rate)

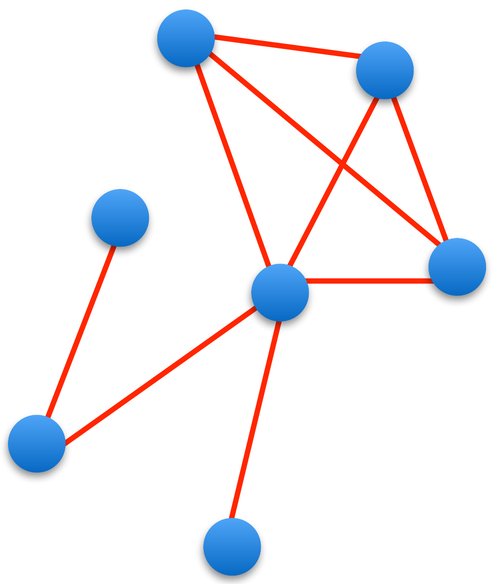
roughly the same computational complexity

"better" (but similar) practical results

Graphs

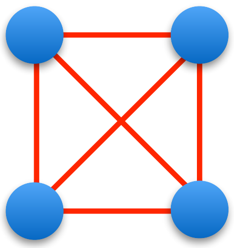
Graph

$$\mathcal{G} = (V, E)$$

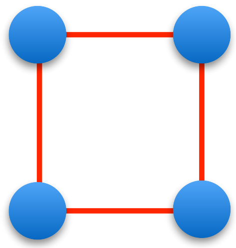


here: **non**-weighted, undirected, connected

Classic graphs (on 4 nodes)



complete



ring

path

Can describe several interactions, e.g.

social networks

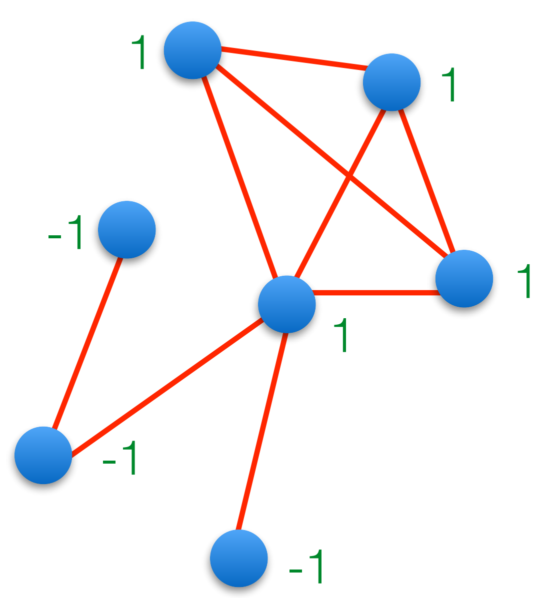
transportation networks

…

Graph (node) signals

Graph

$\mathcal{G} = (V, E)$



Graph signals

$\mathcal{H}(V, \mathbb{R}) \equiv \mathbb{R}^{|V|}$ (euclidean structure)

$\beta^\star : V \to \mathbb{R}$ or $\beta^\star \in \mathbb{R}^{|V|}$

Can describe several quantities, e.g.

temperature at a site

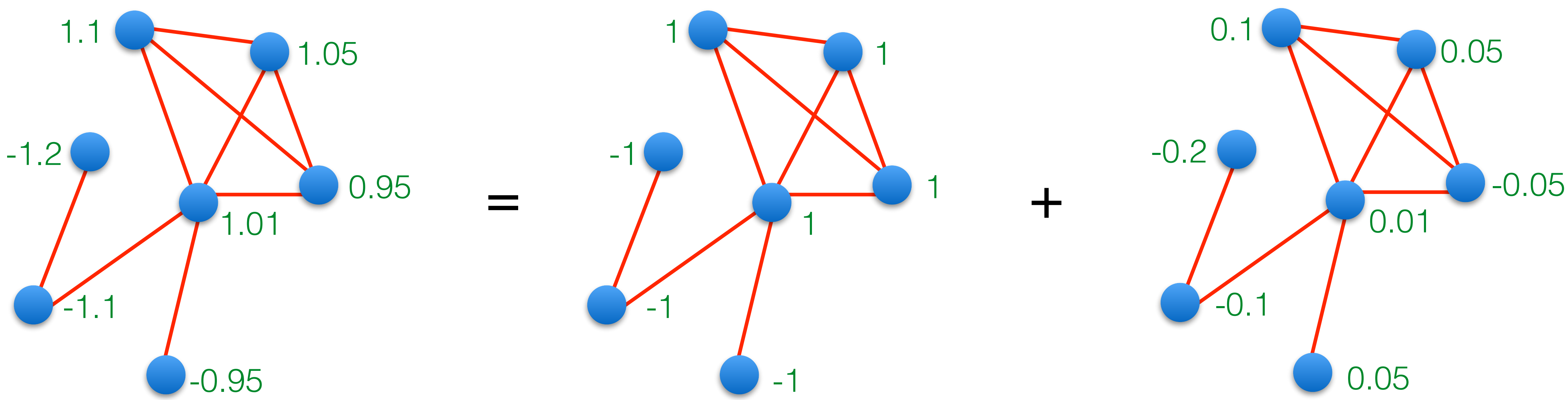
prevalence of illness

intensity of a pixel

…

Noise in a Signal

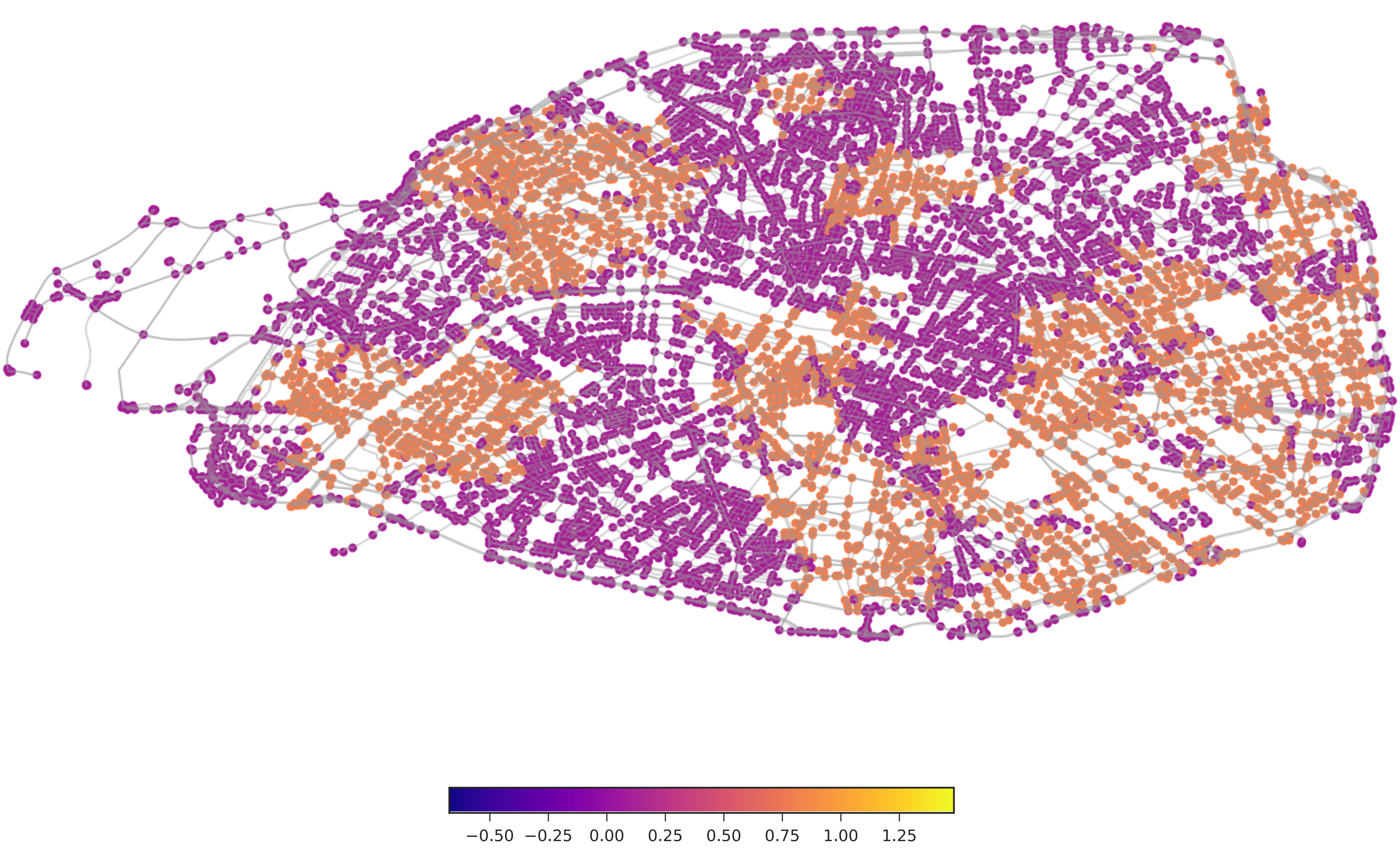
observations = ground truth + noise



y = $\beta^\star$ + ε

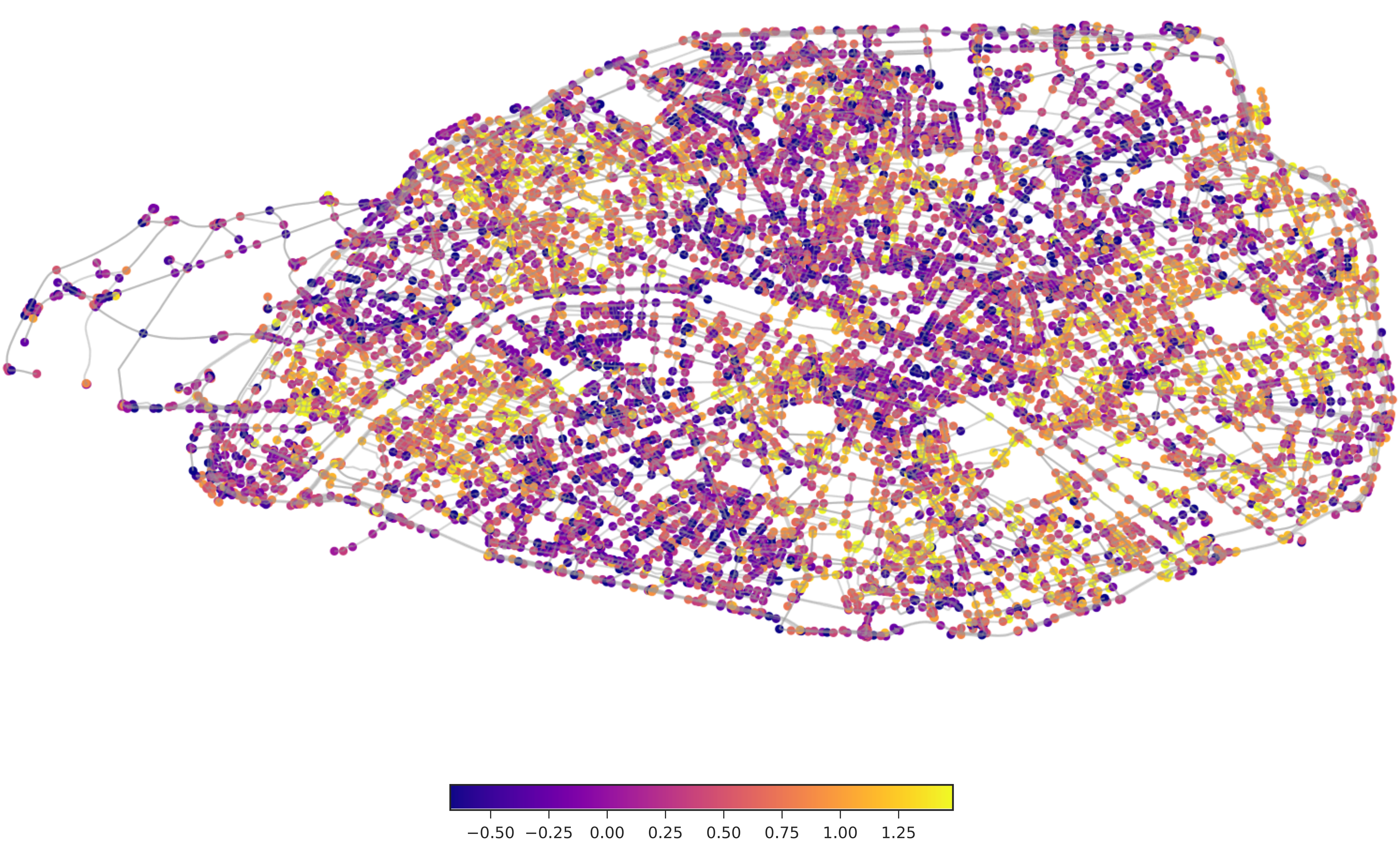
Goal: recover $\beta^\star$ from y

Noise in a Signal



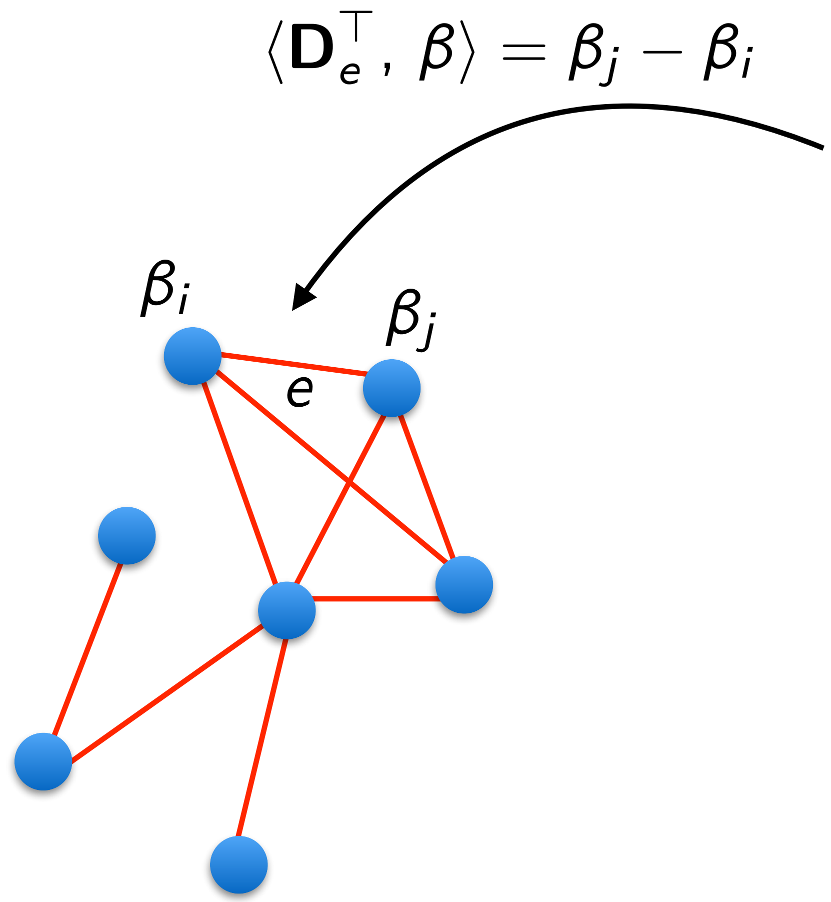
$p = 20108$ (streets), $n = 10205$ (intersections)

Noise in a Signal



$p = 20108$ (streets), $n = 10205$ (intersections)

Incidence Matrix

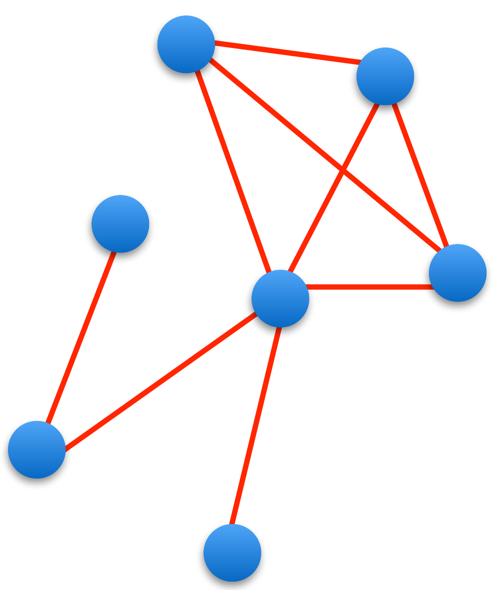


$$(\mathbf{D}^\top)_{e,v} = \begin{cases} +1, & \text{if } v = \min(i,j) \\ -1, & \text{if } v = \max(i,j) \\ 0, & \text{otherwise} \end{cases}$$

$\mathbf{D}^\top \approx \nabla$ in a graph sense

$L = \mathbf{D}\mathbf{D}^\top$ (Laplacian)

Variational Denoising



$$(\mathbf{D}^\top)_{e,v} = \begin{cases} +1, & \text{if } v = \min(i,j) \\ -1, & \text{if } v = \max(i,j) \\ 0, & \text{otherwise} \end{cases}$$

$\mathbf{D}^\top \approx \nabla$ in a graph sense

$L = \mathbf{D}\mathbf{D}^\top$ (Laplacian)

Variational methods

compromise

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{2n}\|y - \beta\|^2 + \lambda J(\mathbf{D}^\top \beta)$$

data fidelity

convex "regularization"

Examples

$J(\cdot) = \langle \cdot, \cdot \rangle$ Laplacian

$J(\cdot) = \| \cdot \|_1$ Graph-Lasso

Outline

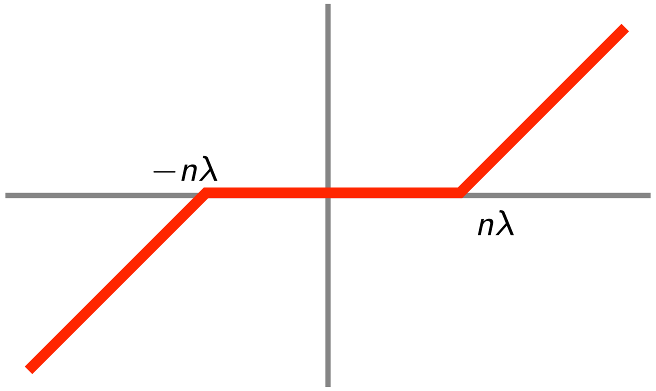
1) Graph-Lasso and Graph-Slope
2) Theoretical result: an oracle inequality
3) How to solve the problem ?
4) Some experiments

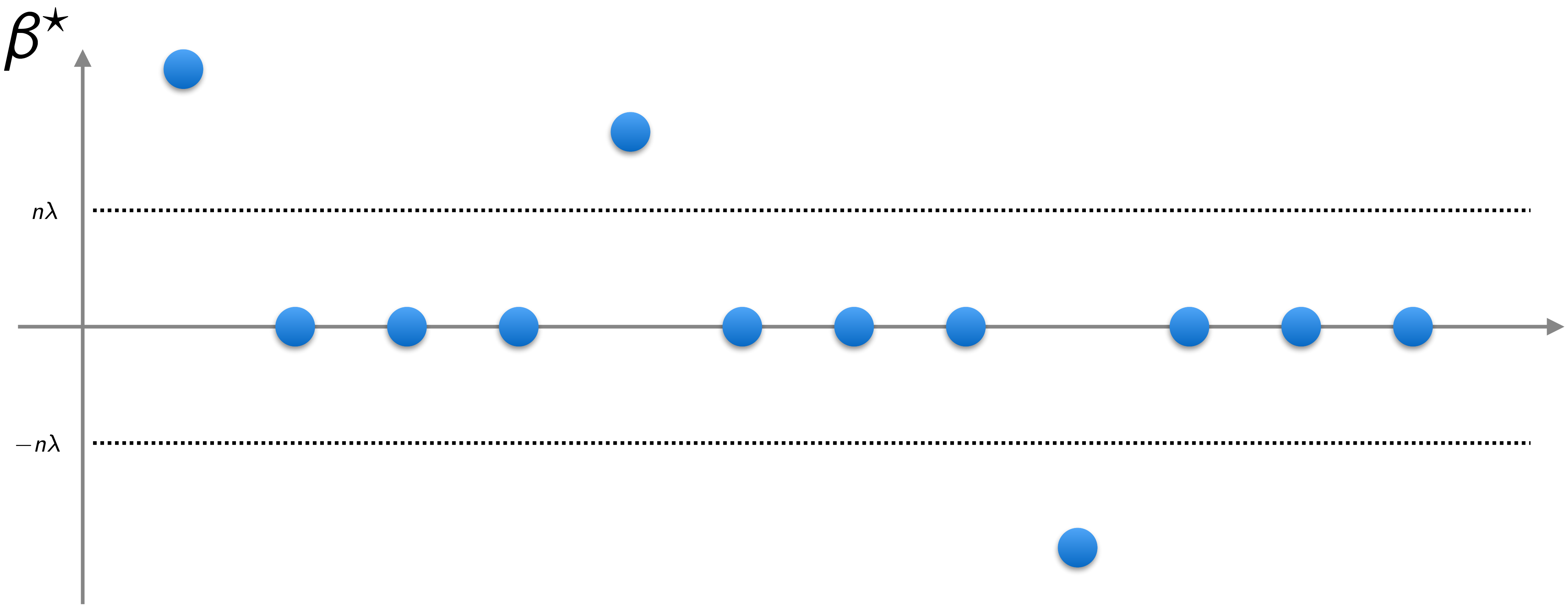
Graph-Slope

Soft-Thresholding

Standard Lasso (denoising case = soft-thresholding)

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n} \|y - \beta\|^2 + \lambda \|\beta\|_1$$

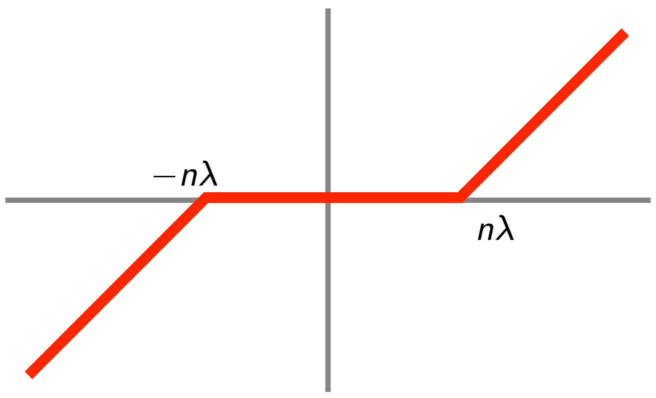


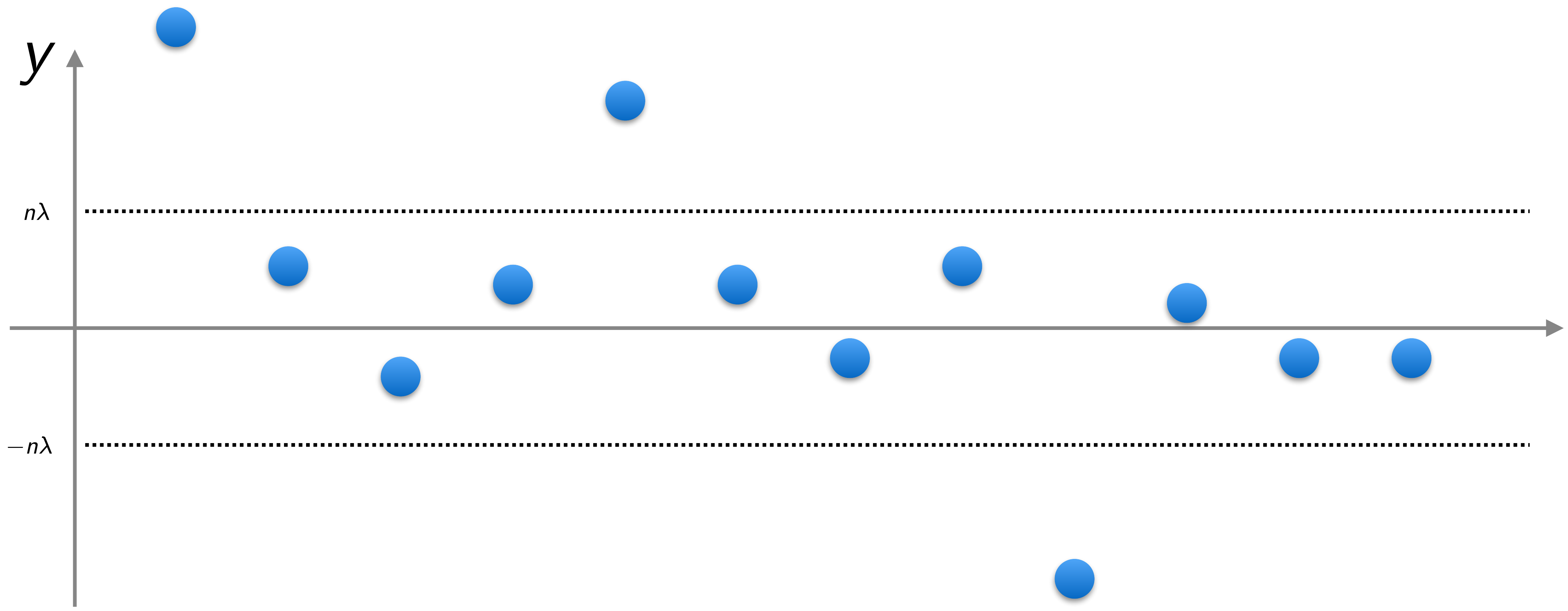


Soft-Thresholding

Standard Lasso (denoising case = soft-thresholding)

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n} \|y - \beta\|^2 + \lambda \|\beta\|_1$$

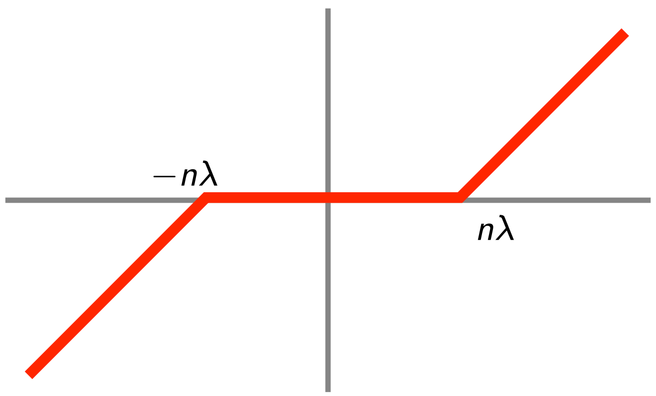


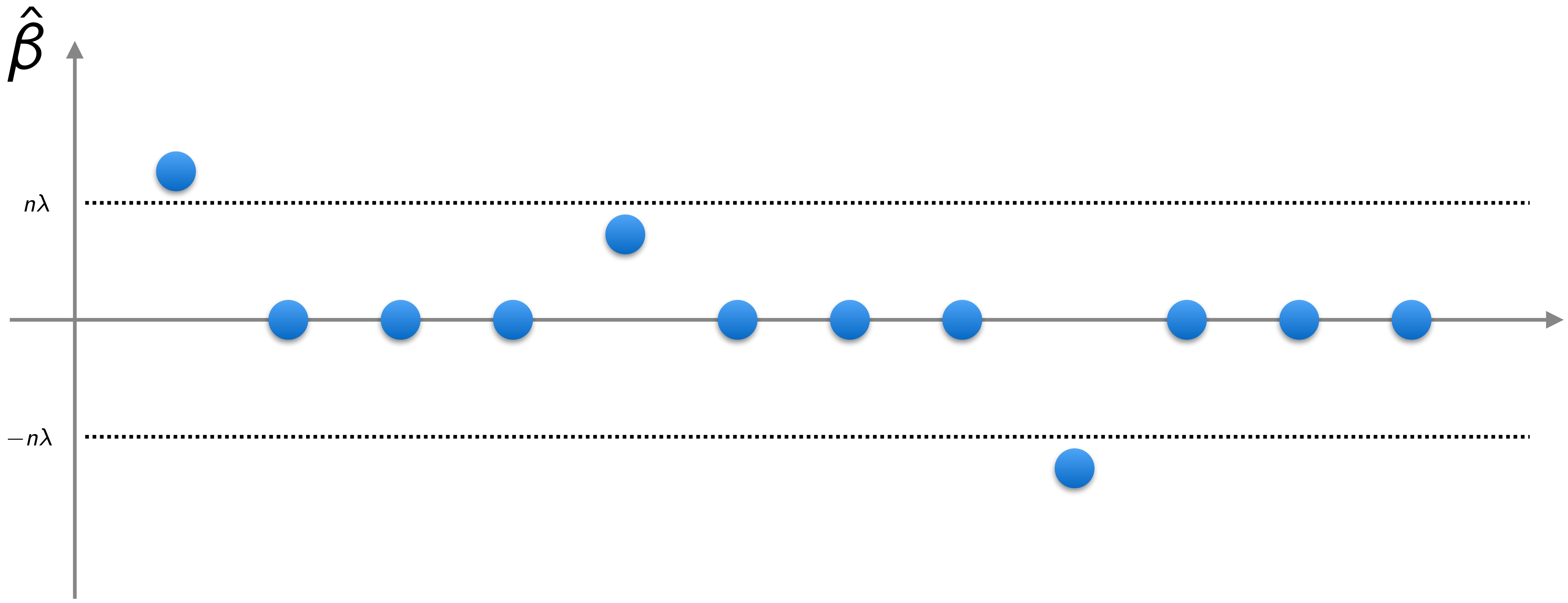


Soft-Thresholding

Standard Lasso (denoising case = soft-thresholding)

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \ \frac{1}{2n} \|y - \beta\|^2 + \lambda \|\beta\|_1$$

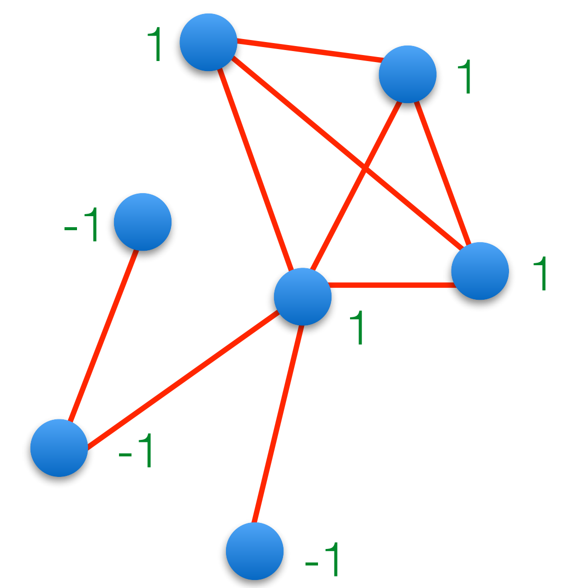




Graph-Lasso

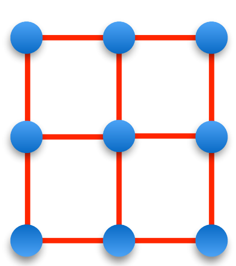
Graph-Lasso (denoising case)

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n} \|y - \beta\|^2 + \lambda \|\mathbf{D}^\top \beta\|_1$$



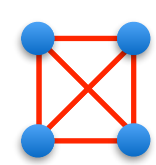
sparsity ➤ sparsity across edges

Grid



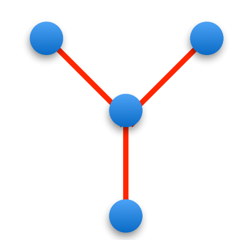
TV 2D

Complete



Clustered Lasso

Star



Stratified data

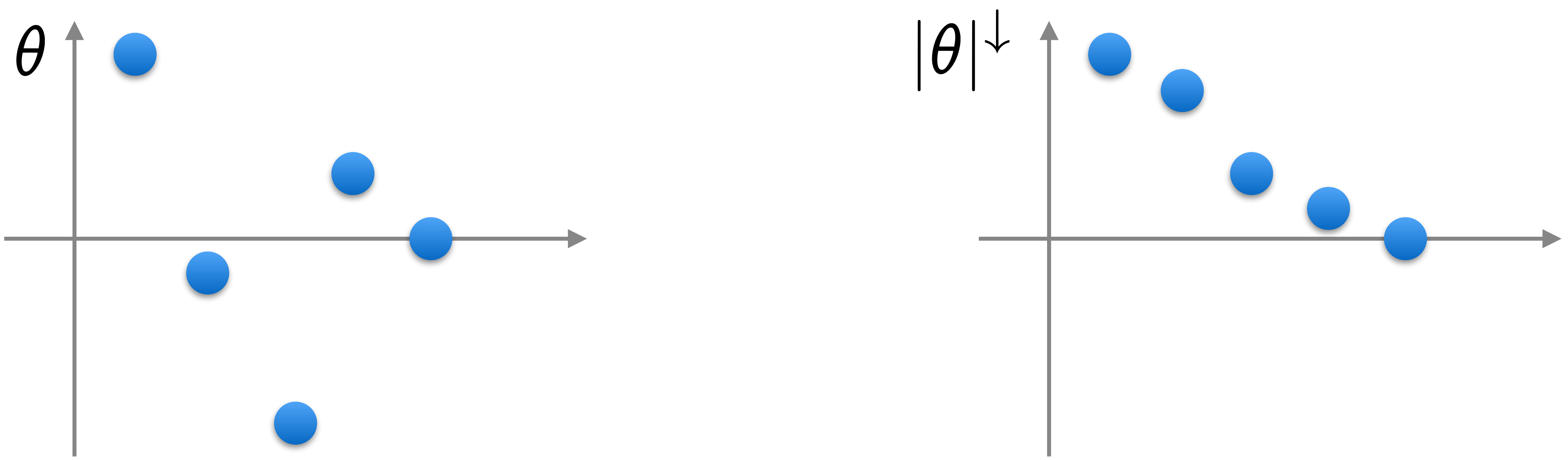
Slope

Idea: it is harsh to threshold all values the same way

$$\lambda \in \mathbb{R}_+ \longrightarrow \lambda \in \mathbb{R}_+^p \text{ s.t } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

$$\lambda \| \cdot \|_1 \longrightarrow \| \cdot \|_{[\lambda]} \text{ defined as}$$

$$\|\theta\|_{[\lambda]} = \sum_{j=1}^{p} \lambda_j |\theta|_j^{\downarrow}$$



[Bogdan et al. '14, Zeng-Figueiredo '14]

Slope

Idea: it is harsh to threshold all values the same way

$$\lambda \in \mathbb{R}_+ \longrightarrow \lambda \in \mathbb{R}_+^p \text{ s.t } \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$$

$\lambda \| \cdot \|_1 \longrightarrow \| \cdot \|_{[\lambda]}$ defined as

Ordered ℓ^1-norm

$$\|\theta\|_{[\lambda]} = \sum_{j=1}^{p} \lambda_j |\theta|_j^{\downarrow}$$

Proposition

$\theta \mapsto \|\theta\|_{[\lambda]}$ is a norm

Graph-Slope

how to compute?

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^n} \frac{1}{2n} \|y - \beta\|^2 + \|\mathbf{D}^\top \beta\|_{[\lambda]}$$

how to choose?

Theory

How to choose the weights?

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n} \|y - \beta\|^2 + \|\mathbf{D}^\top \beta\|_{[\lambda]}$$

Parameter selection is hard, even when only 1!

- by hand!
- by cross-validation
- **using theoretical results (e.g. MSE rate)**

Main Result: Oracle Inequality

Assume $\lambda_1 \geqslant \ldots \geqslant \lambda_p \geqslant 0$ are such that the event

$$\frac{1}{\sqrt{n}}\|\mathbf{D}^\dagger \varepsilon\|^*_{[\lambda]} \leqslant 1/2$$

has pr. $\geqslant 1/2$. Then, $\forall \delta \in (0,1)$ we have with pr. $\geqslant 1 - 2\delta$

$$\|\hat{\beta} - \beta^\star\|_n^2 \leqslant \inf_{s \in [p]} \left[\inf_{\substack{\beta \in \mathbb{R}^n \\ \|\mathbf{D}^\top \beta\|_0 \leqslant s}} \|\beta - \beta^\star\|_n^2 + \left(\frac{3\Lambda(\boldsymbol{\lambda}, s)}{\kappa(s)} + \frac{\sigma + 2\sigma\sqrt{2\log(1/\delta)}}{\sqrt{n}} \right)^2 \right]$$

compatibility factor ~ [Hutter-Rigollet '16]

$$\kappa(s) \triangleq \inf_{v \in \mathbb{R}^n : 3\Lambda(\boldsymbol{\lambda}, s)\|\mathbf{D}^\top v\|_2 > \sum_{j=s+1}^p \lambda_j |\mathbf{D}^\top v|_j^\downarrow} \left(\frac{\|v\|_n}{\|\mathbf{D}^\top v\|_2} \right)$$

$$\Lambda(\boldsymbol{\lambda}, s) = \left(\sum_{j=1}^s \lambda_j^2 \right)^{1/2}$$

Main Result: Oracle Inequality

$$\|\hat{\beta} - \beta^\star\|_n^2 \leqslant \inf_{s \in [p]} \Bigg[\inf_{\substack{\beta \in \mathbb{R}^n \\ \|\mathbf{D}^\top \beta\|_0 \leqslant s}} \|\beta - \beta^\star\|_n^2$$

Choice of Weights

How to guarantee

$$\frac{1}{\sqrt{n}}\|\mathbf{D}^{\dagger}\varepsilon\|_{[\lambda]}^{*} \leqslant 1/2?$$

Two possible ways:

1) be smart enough from the theory!

2) use Monte Carlo estimation

Choice of Weights: the Smart Way©

inverse scaling factor [Hutter-Rigollet '16]

$$\rho(\mathcal{G}) = \max_{j \in [p]} \|(\mathbf{D}^\top)^\dagger e_j\|_n$$

Assume that $\lambda_1 \geqslant \ldots \geqslant \lambda_p \geqslant 0$ satisfy for any $j \in [p]$

$$\lambda_j \geqslant 8\sigma\rho(\mathcal{G})\sqrt{\frac{\log(2p/j)}{n}}.$$

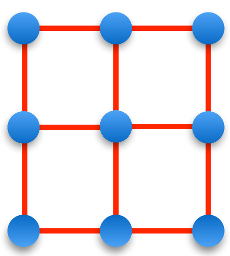
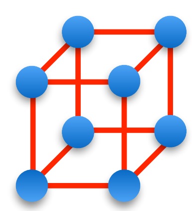
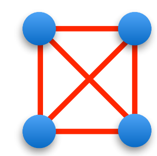
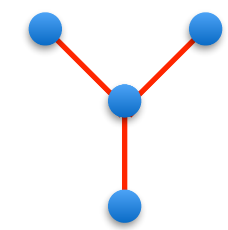
Then, for any $\delta \in (0,1)$, the oracle inequality holds with probability at least $1 - 2\delta$.

What about computing the inverse scaling factor ?

Inverse Scaling Factor

inverse scaling factor [Hutter-Rigollet '16]

$$\rho(\mathcal{G}) = \max_{j \in [p]} \|(\mathbf{D}^\top)^\dagger e_j\|_n$$

Grid	Hypercube	Complete	Star
			
$\rho(\mathcal{G}) \lesssim \log n$	$\rho(\mathcal{G}) \leqslant 1$	$\rho(\mathcal{G}) \lesssim 1/n$	$\rho(\mathcal{G}) \leqslant 1$

Generic graph

If $\lambda_2 > 0$, then $\rho(\mathcal{G}) \leqslant \sqrt{2}/\lambda_2$

Fiedler eigenvalue of the Laplacian

Oracle Inequality, Simplified

Assume that $\|\mathbf{D}^\top \beta^\star\| = s^\star$ and let

$$\lambda_j = 8\sigma\rho(\mathcal{G})\sqrt{\frac{\log(2p/j)}{n}}.$$

Then, for any $\delta \in (0,1)$, we have with pr. at least $1-2\delta$.

$$\|\hat{\beta} - \beta^\star\|_n^2 \leqslant \frac{\sigma^2}{n}\left(\frac{48\rho(\mathcal{G})^2 s^\star}{\kappa(s^\star)^2}\log\left(\frac{2ep}{s^\star}\right) + 2 + 16\log\left(\frac{1}{\delta}\right)\right)$$

Graph-Slope rate

$$\log\left(\frac{2ep}{s^\star}\right)$$

Graph-Lasso rate

$$\log\left(\frac{ep}{\delta}\right)$$

[Hutter-Rigollet '16]

Choice of Weights: MC Estimation

$$g_j = e_j^\top \mathbf{D}^\dagger \varepsilon / \sqrt{n} \quad \longrightarrow \quad |g|_1^\downarrow \geq \cdots \geq |g|_p^\downarrow$$

$$\max_{j=1,\dots,p} \left(|g|_j^\downarrow \,/\, \lambda_j \right) \leq 1/2 \quad \Longrightarrow \quad \frac{1}{\sqrt{n}} \|\mathbf{D}^\dagger \varepsilon\|_{[\lambda]}^* \leq 1/2$$

1) Estimate the law $\mathbb{P}$ of ε (say $\mathcal{N}(0, \sigma^2 \mathrm{Id})$)

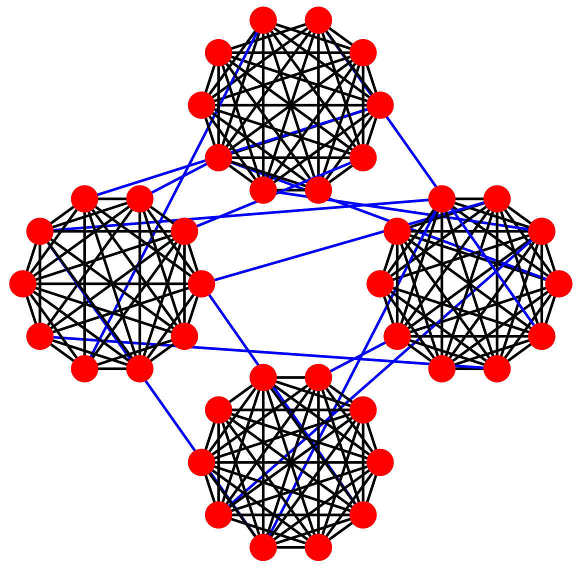
2) λ_j choose as (quantile evaluation of $\mathbb{P}$)

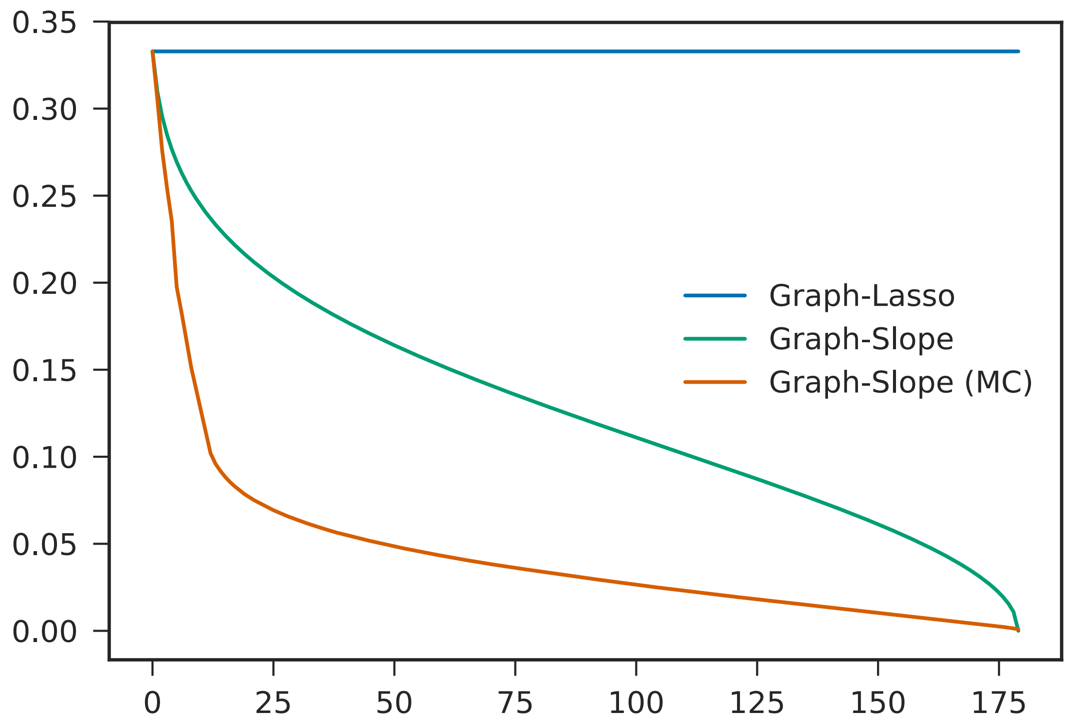
$$\mathbb{P}(2|g|_j^\downarrow \leqslant \lambda_j) \geqslant 1 - 1/3p$$

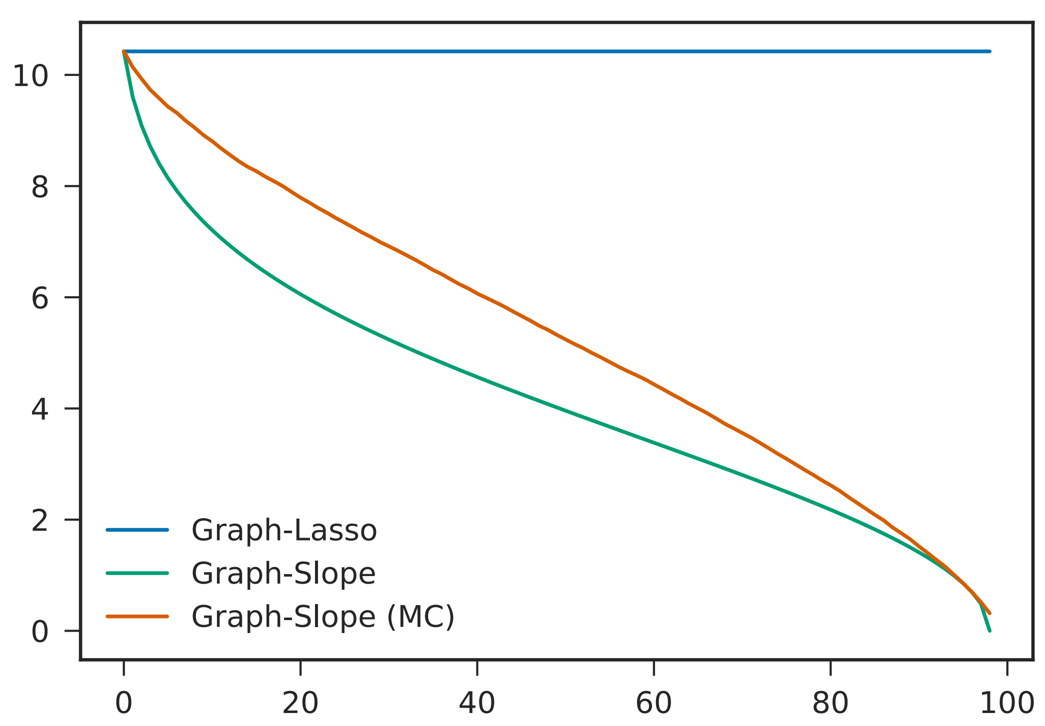
3) And voila !

(typically just choose the .95 quantile)

Choice of Weights: Examples







Optimization

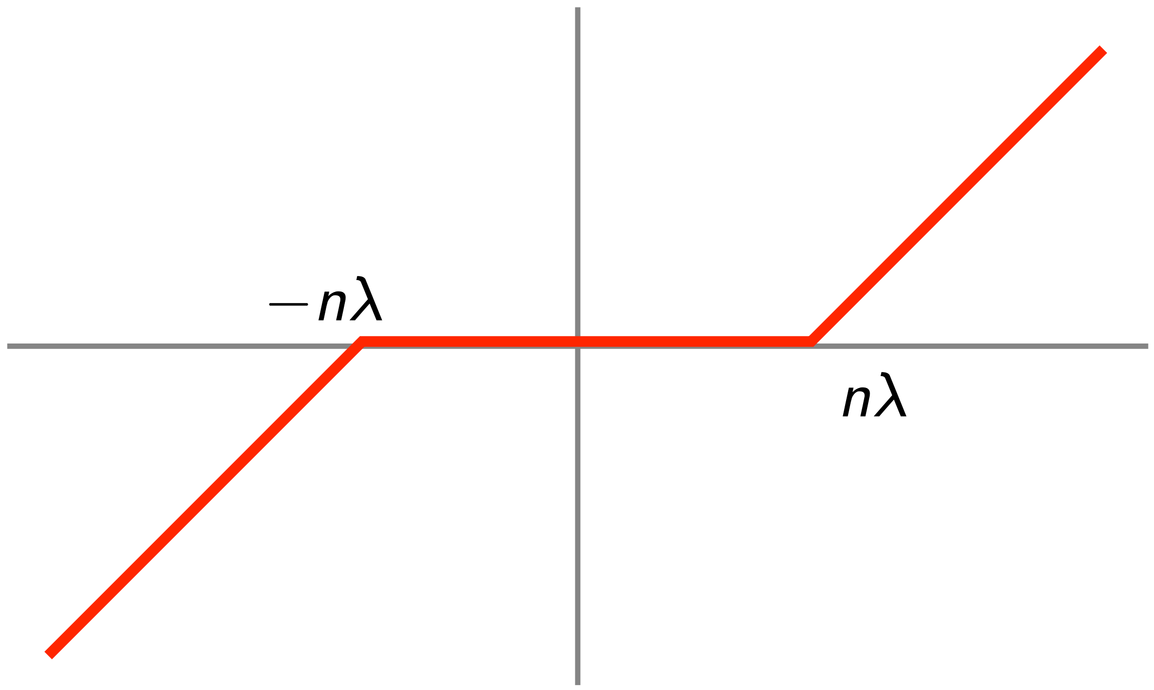
Soft-Thresholding

Standard Lasso

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2n} \|y - \beta\|^2 + \lambda \|\beta\|_1$$

Proximity operator = soft-thresholding

$$\hat{\beta} = \mathrm{Prox}_{n\lambda \|\cdot\|_1}(y)$$



Two Difficulties

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \; \frac{1}{2n}\|y - \beta\|^2 + \|\mathbf{D}^\top \beta\|_{[\lambda]}$$

Linear operator in the norm

→ "dualization"

This is not the ℓ^1 norm!

→ computational trick

Duality

$$\min_{\beta \in \mathbb{R}^n} f(\beta) + g(\mathbf{D}^\top \beta)$$

$$\Big\uparrow \qquad \Big\downarrow$$

$$\min_{\theta \in \mathbb{R}^p} f^\star(\mathbf{D}\theta) + g^\star(-\theta)$$

Fenchel transform

$$f^\star(x) = \sup_z \langle x, z \rangle - f(z)$$

Duality

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2n}\|y - \beta\|^2 + \|\mathbf{D}^\top \beta\|_{[\lambda]}$$

$$\uparrow \quad \downarrow$$

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|\mathbf{D}\theta - y\|_2^2 - \frac{1}{2n}\|y\|_2^2 \quad \text{subject to} \quad \frac{1}{n}\|\theta\|_{[\lambda]}^* \leqslant 1$$

$$\uparrow \quad \downarrow$$

$$\min_{\theta \in \mathbb{R}^p} \frac{1}{2n}\|\mathbf{D}\theta - y\|_2^2 + \iota_{\left\{\theta \,:\, \frac{1}{n}\|\theta\|_{[\lambda]}^* \leqslant 1\right\}}(\theta)$$

Forward-Backward on the Dual

$$\min_{\theta \in \mathbb{R}^p} \bar{f}(\theta) + \bar{g}(\theta)$$

FB iterations (in practice we use FISTA with dual gap stopping criterion, but nevermind)

$$\theta^k = \text{Prox}_{\tau \bar{g}}(\theta^k - \tau \nabla \bar{f}(\theta^k))$$

implicit step fixed point — *explicit step* gradient descent

$\rightarrow$ converges to $\hat{\theta}$ if $\tau < 2/\|D\|$

So what about $\text{Prox}_{\tau \bar{g}}$?

$$\min_{\theta \in \mathbb{R}^p} \underbrace{\frac{1}{2n}\|\mathbf{D}\theta - y\|_2^2}_{\bar{f}(\theta)} + \underbrace{\iota_{\left\{\theta \,:\, \frac{1}{n}\|\theta\|_{[\lambda]}^* \leqslant 1\right\}}(\theta)}_{\bar{g}(\theta)}$$

The Prox I: Moreau Decompositon

$$\mathrm{Prox}_{\tau\iota_{\left\{\theta\,:\,\frac{1}{n}\|\theta\|^*_{[\lambda]}\leqslant 1\right\}}}$$

The Prox I: Moreau Decompositon

$$\mathrm{Prox}_{\tau \iota_{\left\{\theta \,:\, \frac{1}{n}\|\theta\|^*_{[\lambda]} \leqslant 1\right\}}} = \Pi_{\left\{\theta \,:\, \frac{1}{n}\|\theta\|^*_{[\lambda]} \leqslant \frac{1}{\tau}\right\}}$$

$$\| \quad \longleftarrow \quad \text{Moreau decomposition}$$

$$\mathrm{Id} - \tau\, \mathrm{Prox}_{\frac{1}{\tau}\|\cdot\|_{[\lambda]}}\left(\frac{\cdot}{\tau}\right)$$

$$\downarrow$$

Thanks to [Bogdan et al. '14] or [Zeng-Figueiredo '14], we know how to compute it!

The Prox II: Isotonic Regression

Assume that $(y_j - \lambda_j)$ is a positive and decreasing sequence, then

$$\mathrm{Prox}_{\|\cdot\|_{[\lambda]}}(u) = \underset{\theta \in \mathbb{R}^p}{\mathrm{argmin}} \ \frac{1}{2}\|u - \lambda - \theta\|^2$$

$$\text{subject to} \quad \theta_1 \geqslant \theta_2 \geqslant \ldots \geqslant \theta_p$$

⟶ well-known problem, isotonic regression

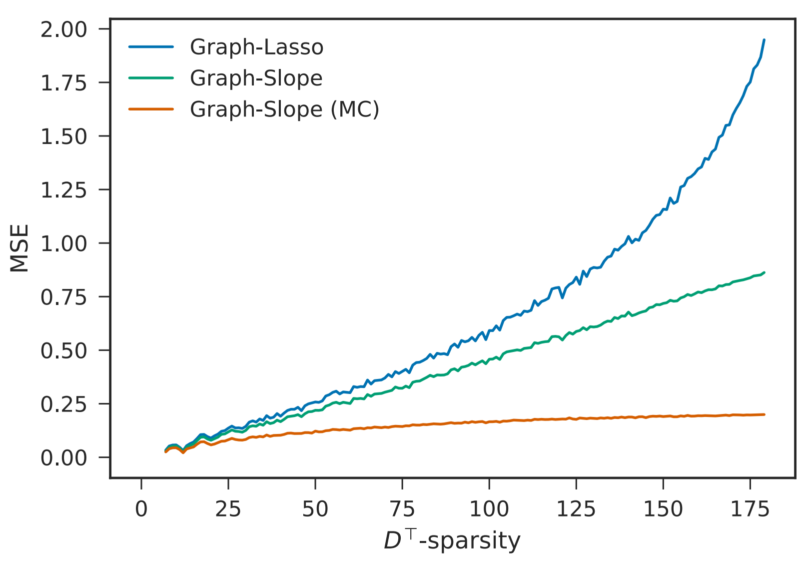
⟶ several solutions,
including PAVA (Pool Adjacent Violators Algorithm)

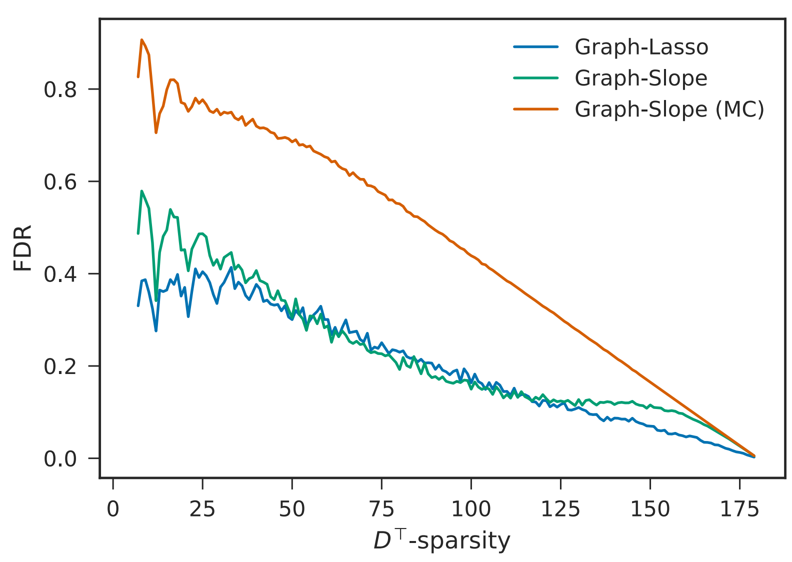
Soft-thresholding	Isotonic regression
$O(p)$	$O(p)$
trivially parallelized	more difficult

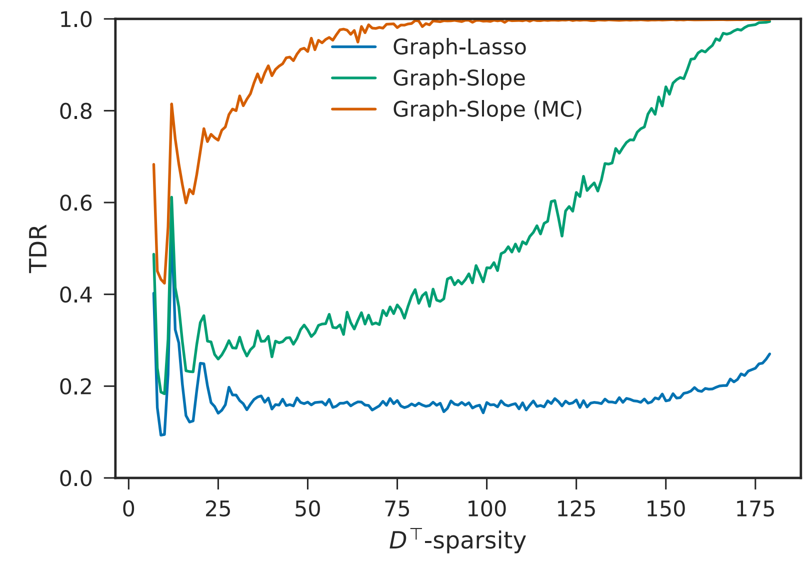
Some Results

Synthetic Results

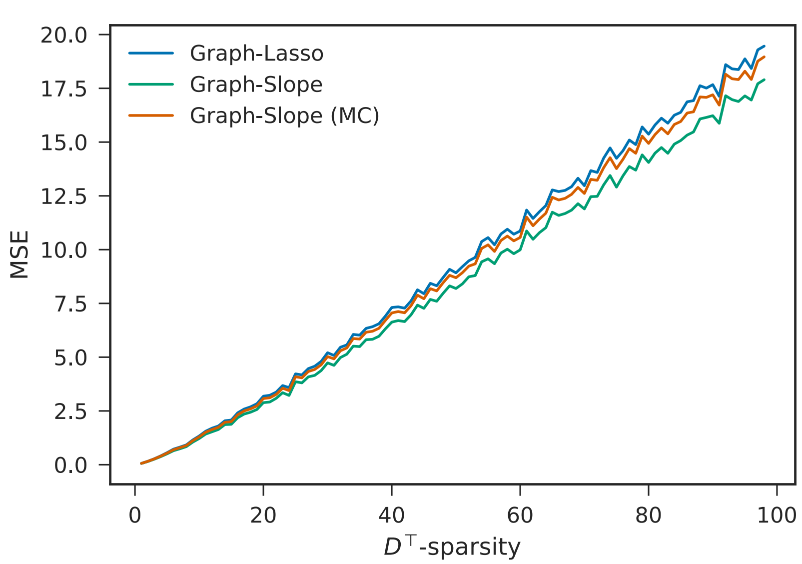
Caveman

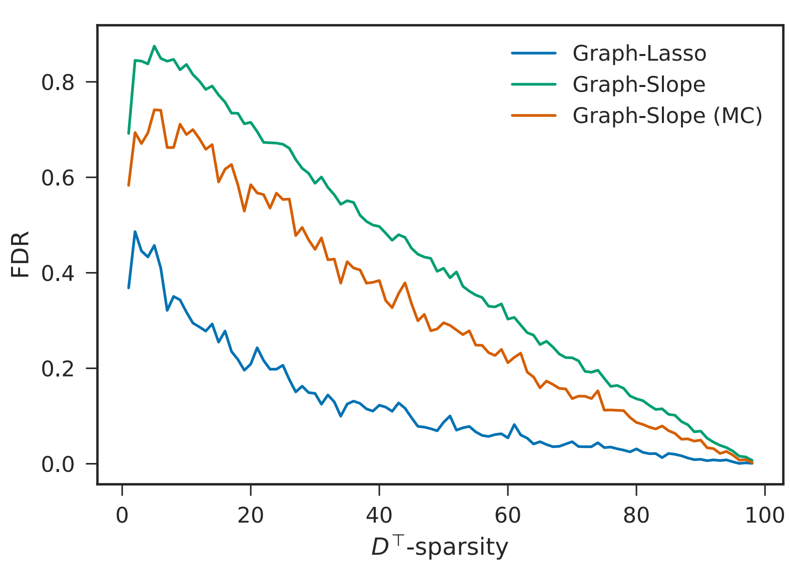


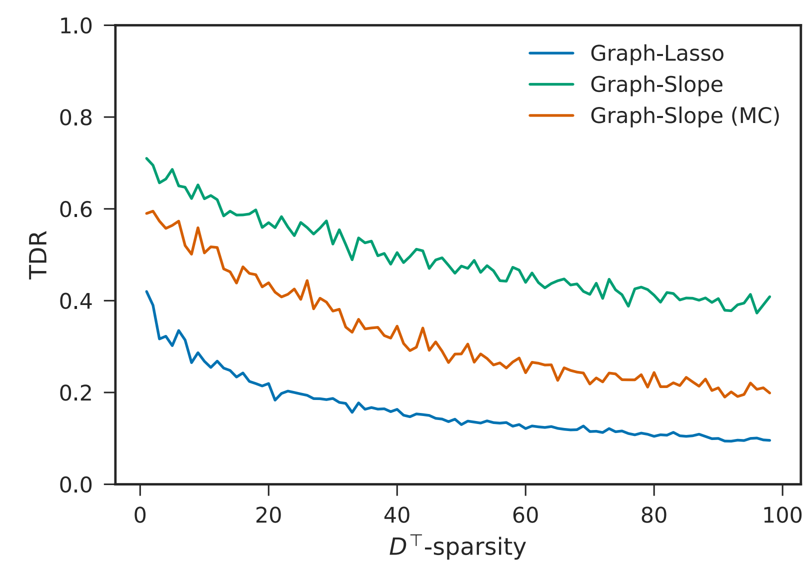




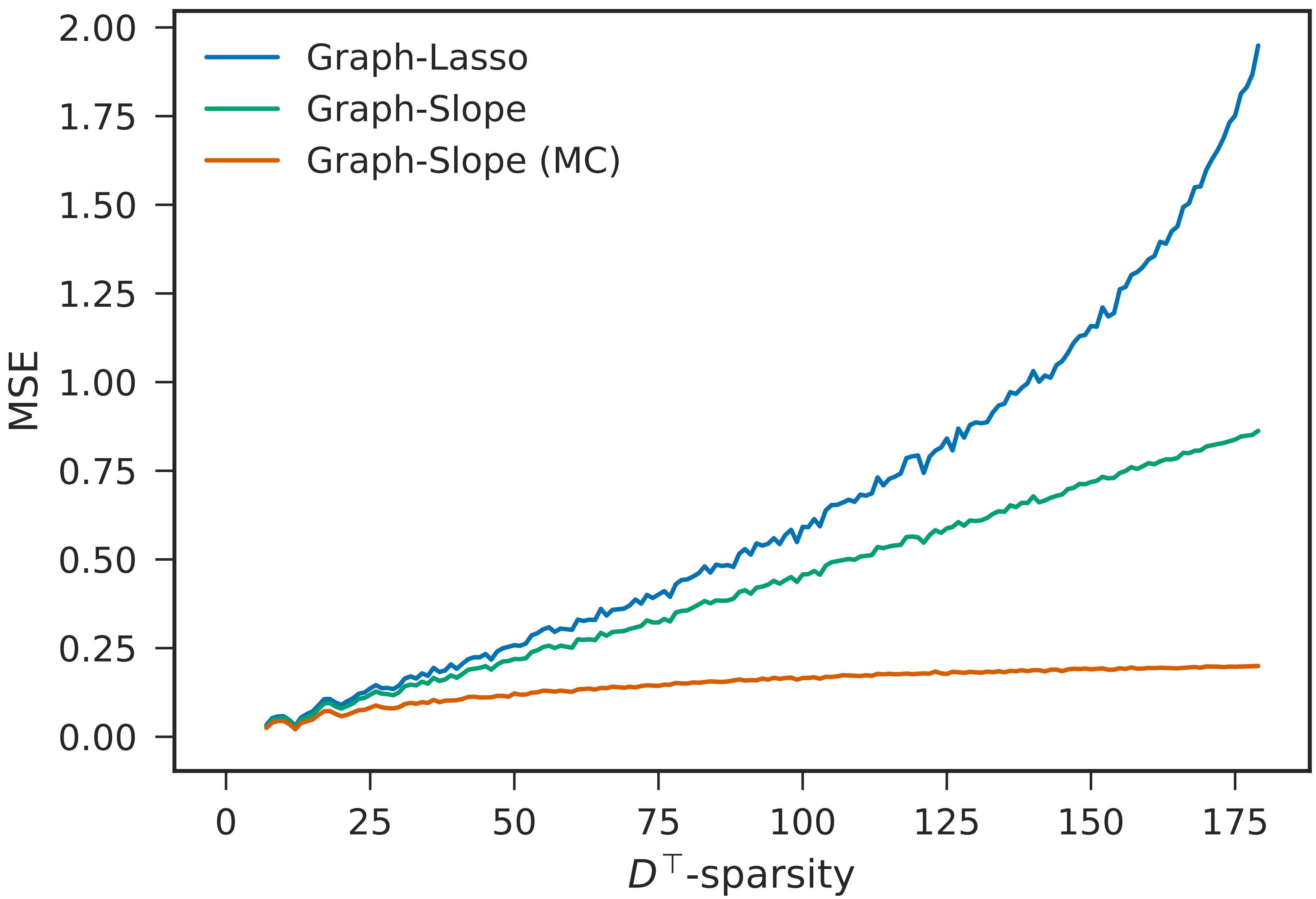
TV-1D (path graph)





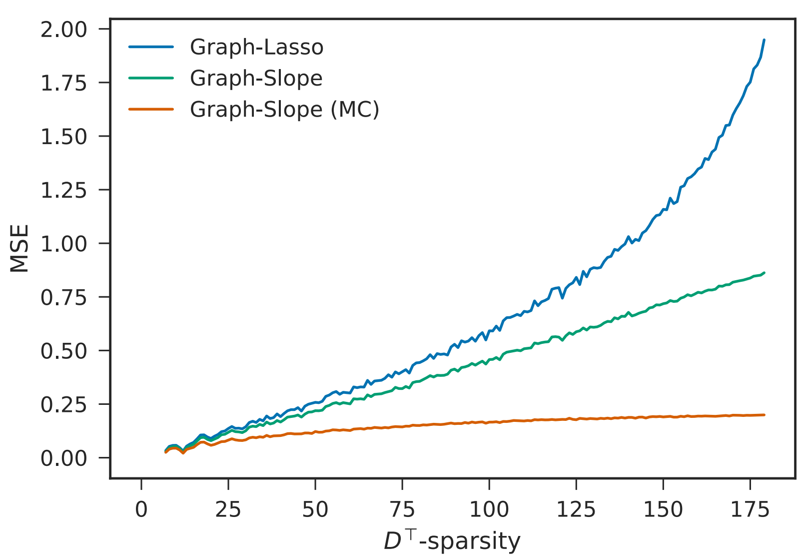


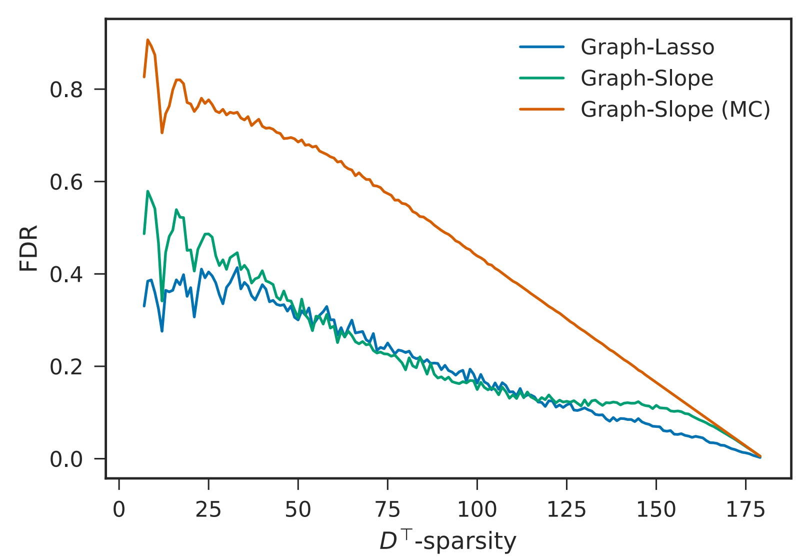
$$\frac{1}{n}\|\beta^\star - \beta^\star\|^2$$

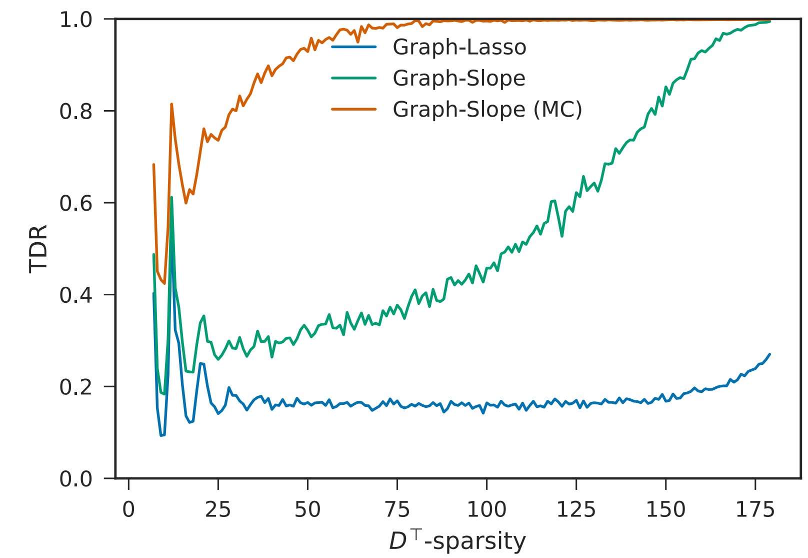


Synthetic Results

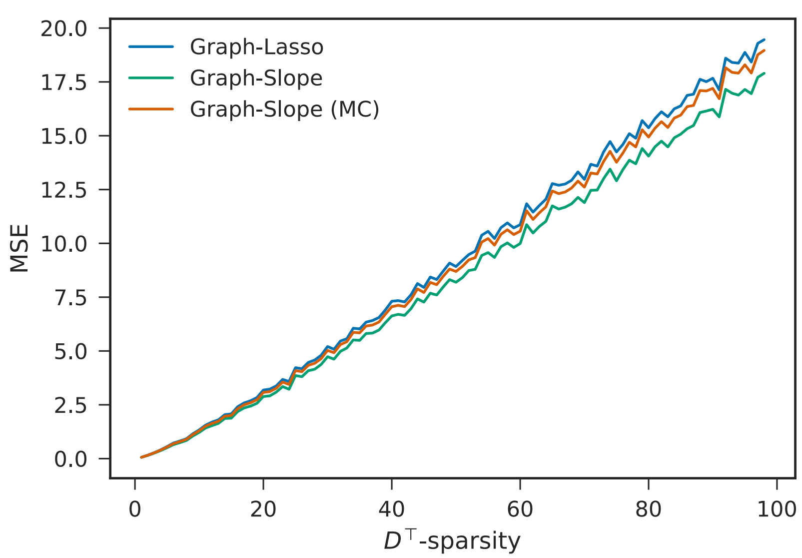
Caveman

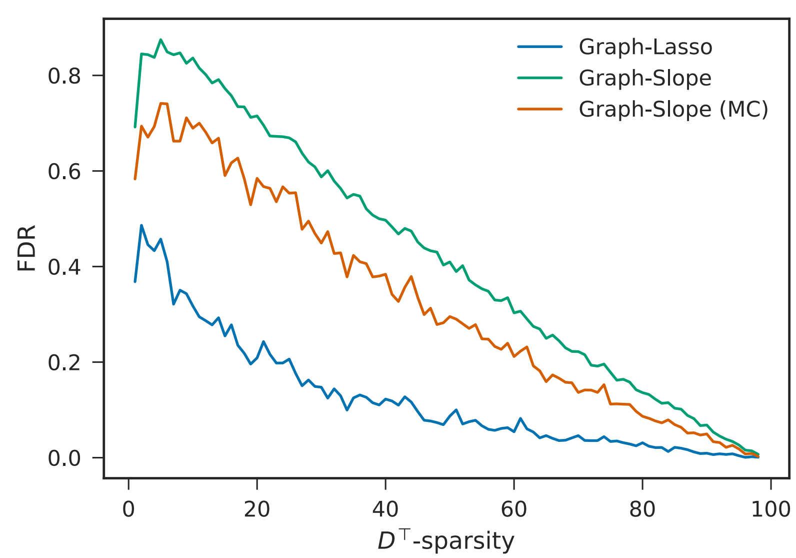


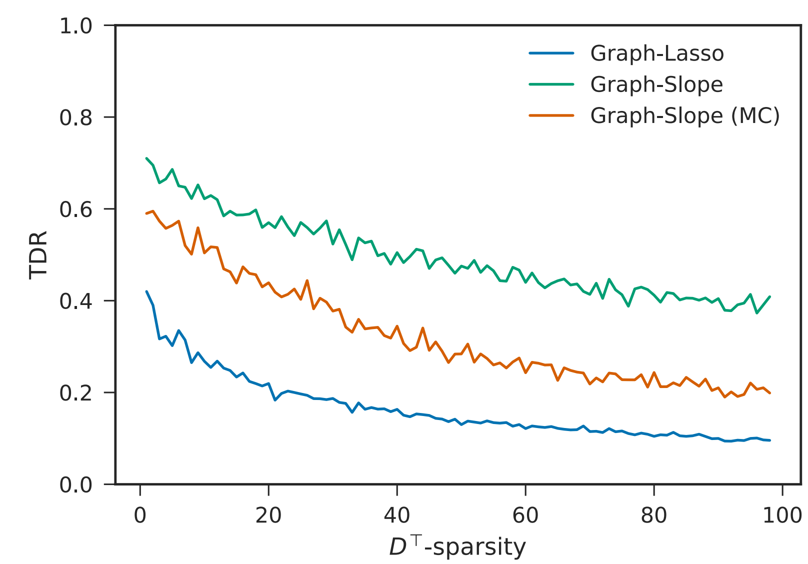




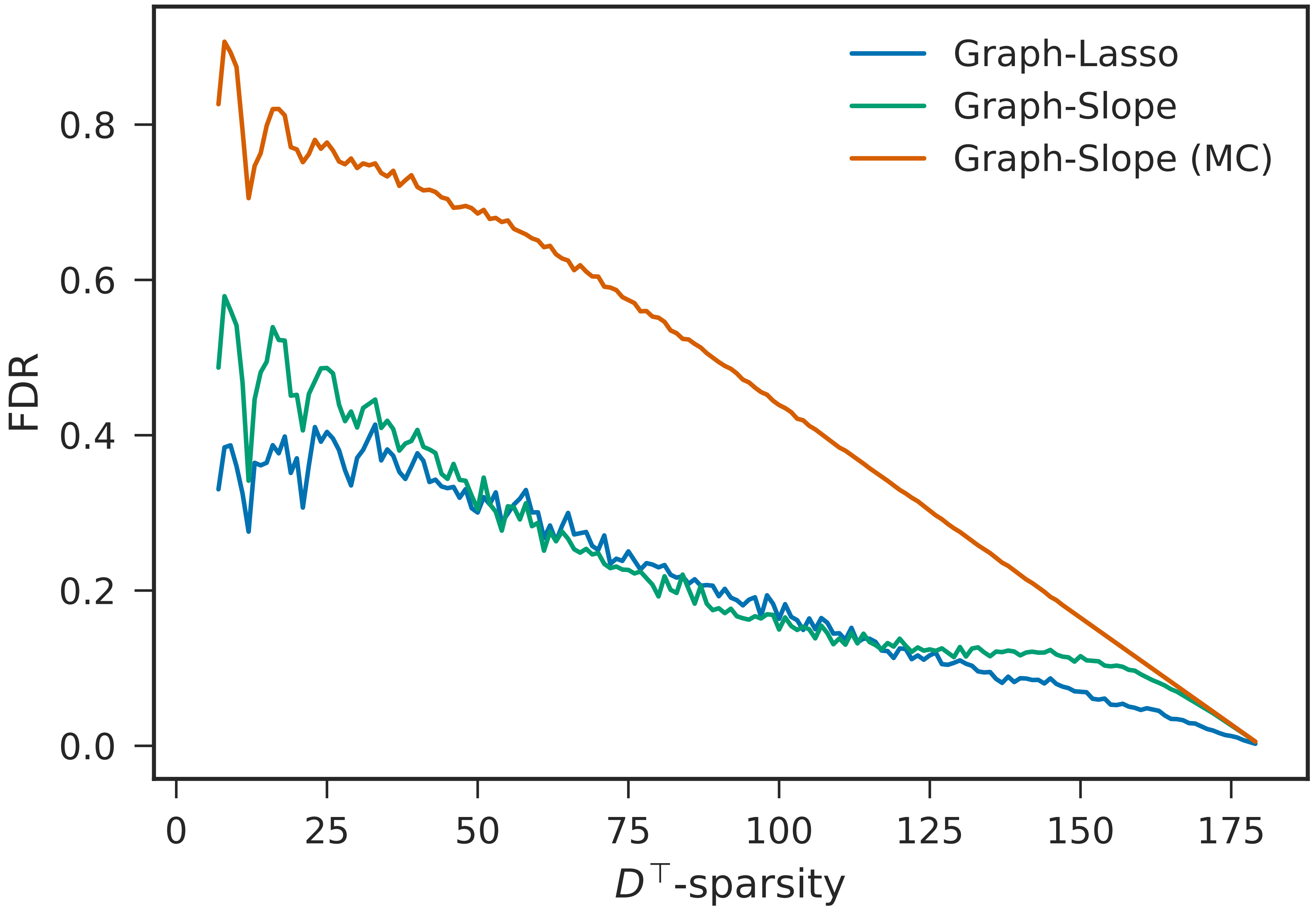
TV-1D (path graph)





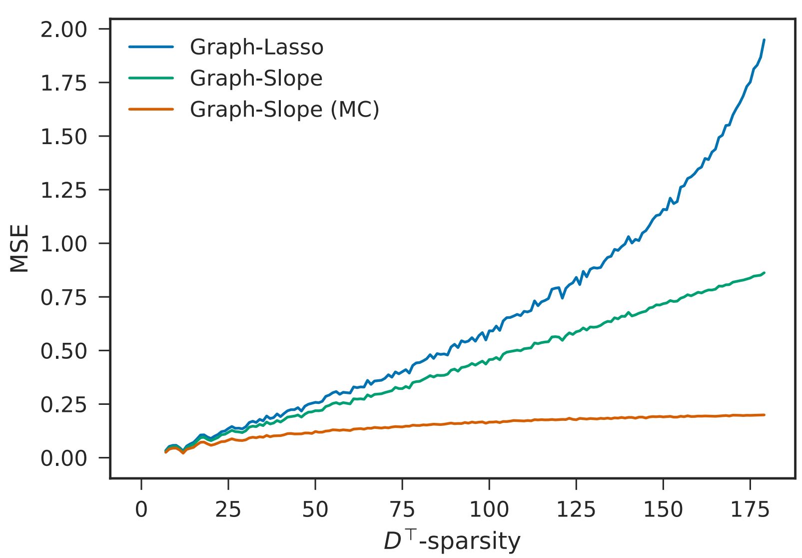


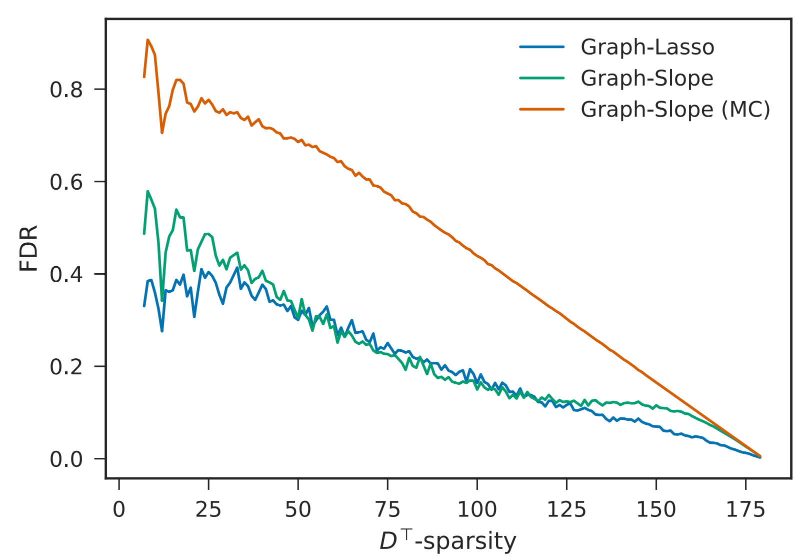
$$\mathrm{FDR}(\hat{\beta}, \beta^\star) = \begin{cases} \dfrac{\left|\left\{j \in [p] : j \in \mathrm{supp}(\mathbf{D}^\top \hat{\beta}) \text{ and } j \notin \mathrm{supp}(\mathbf{D}^\top \beta^\star)\right\}\right|}{\left|\mathrm{supp}(\mathbf{D}^\top \hat{\beta})\right|} & \text{if } \mathbf{D}^\top \hat{\beta} \neq 0 \\ 0 & \text{if } \mathbf{D}^\top \hat{\beta} = 0 \end{cases}$$

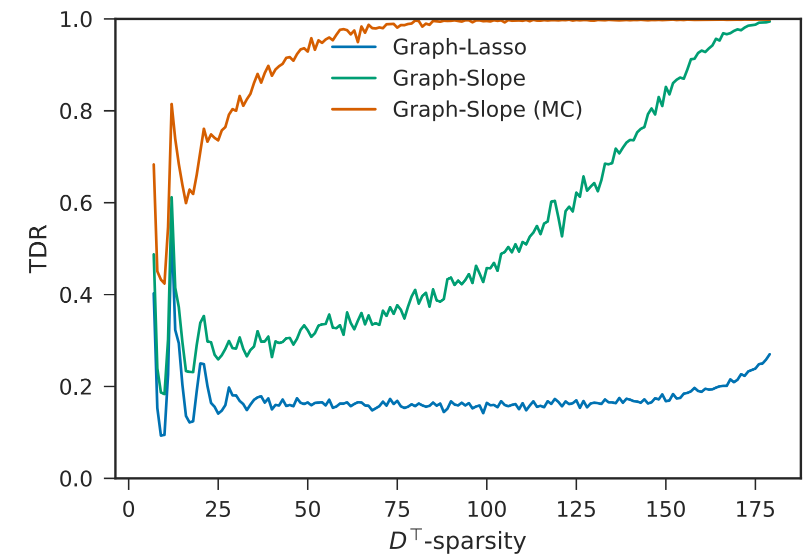


Synthetic Results

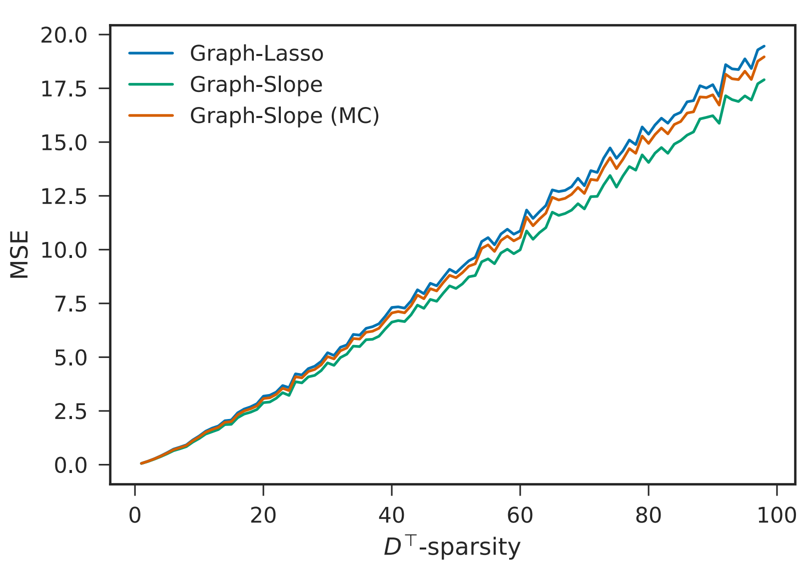
Caveman

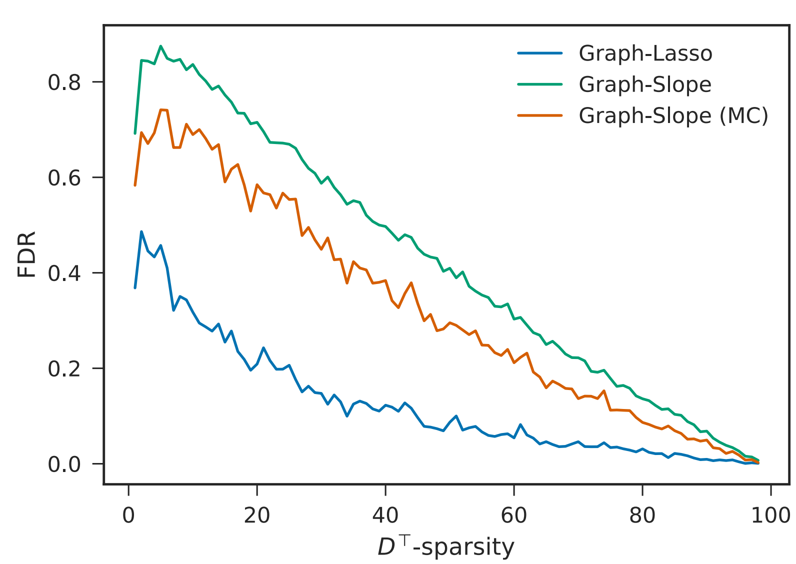


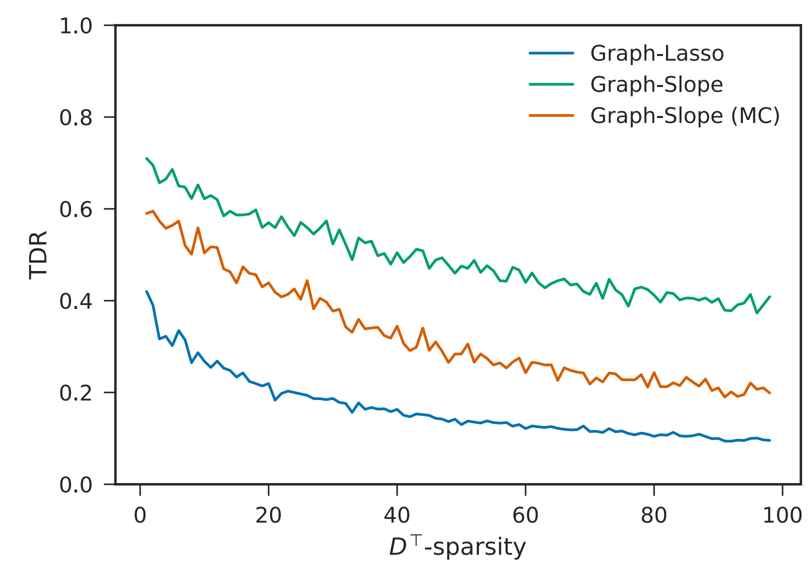




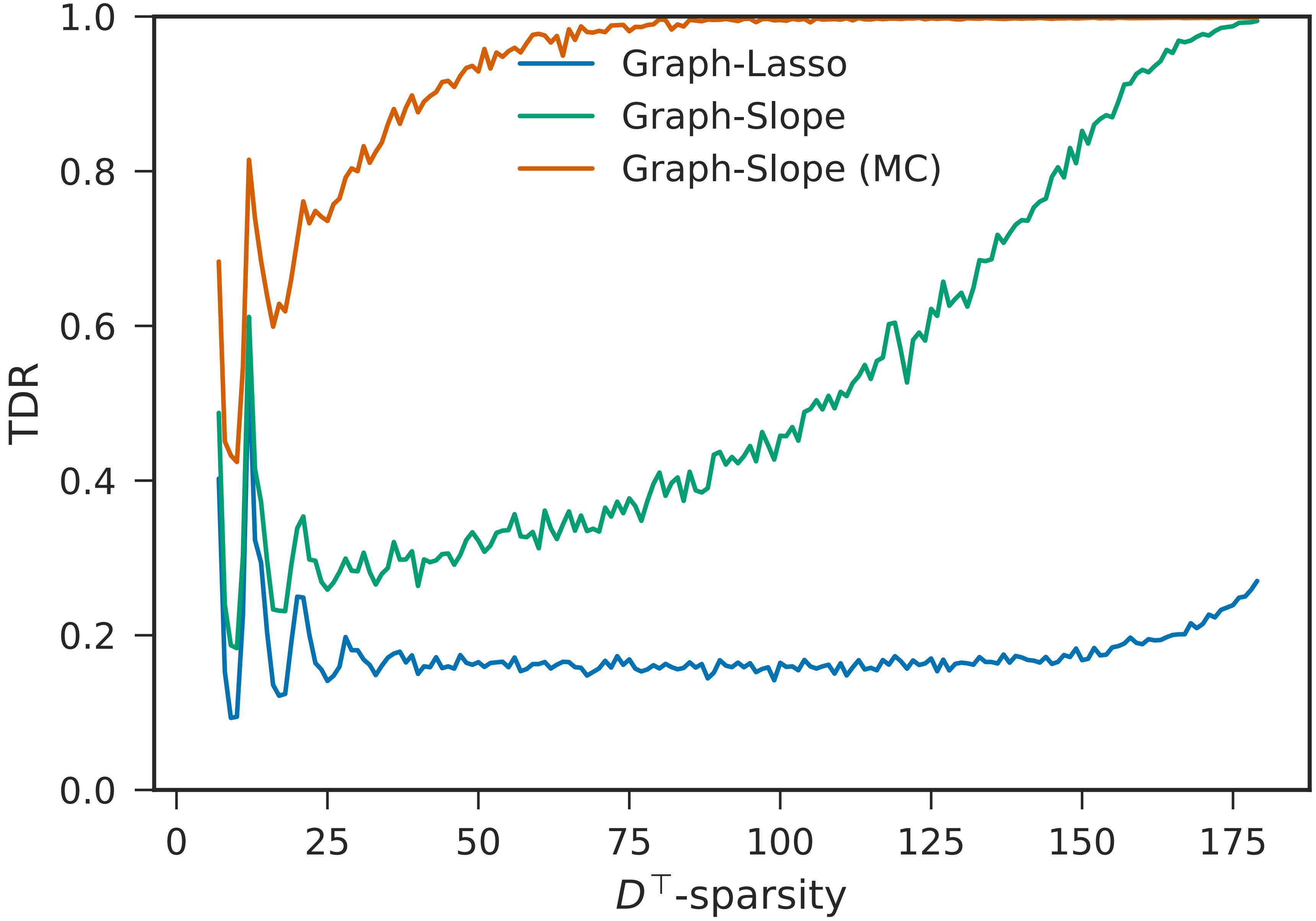
TV-1D (path graph)





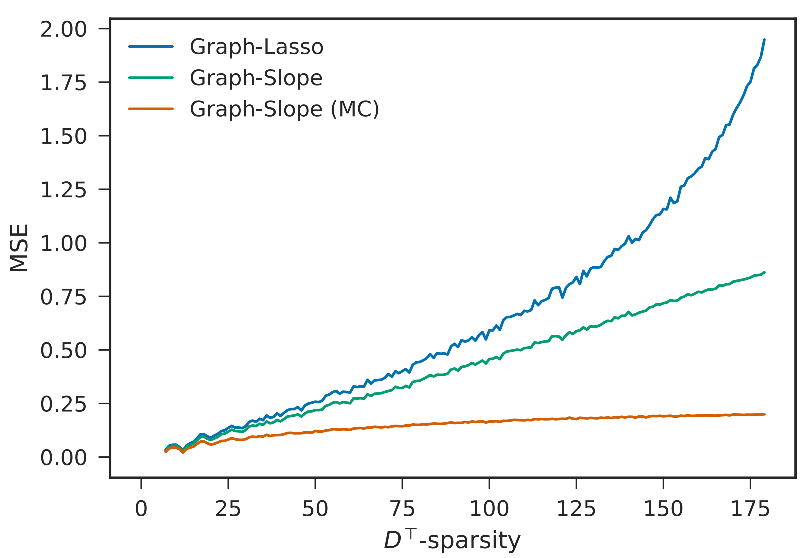


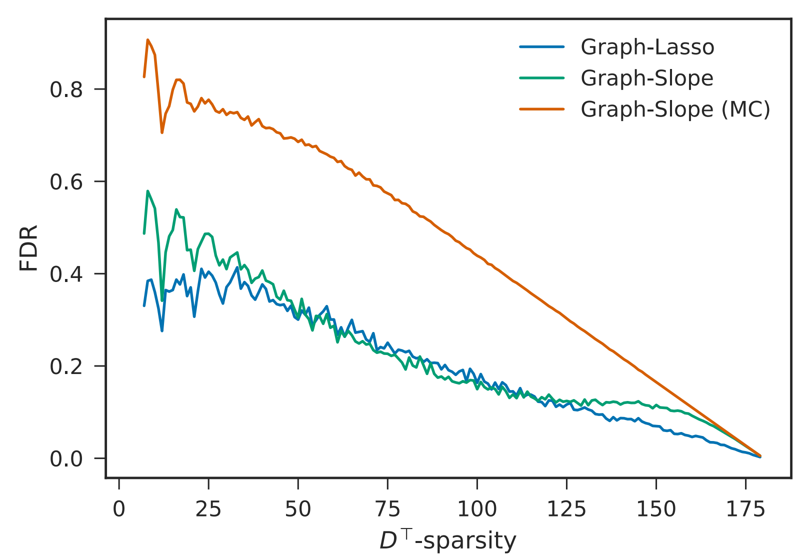
$$\mathrm{TDR}(\hat{\beta}, \beta^\star) = \begin{cases} \dfrac{\left|\left\{ j \in [p] : j \in \mathrm{supp}(\mathbf{D}^\top \hat{\beta}) \text{ and } j \in \mathrm{supp}(\mathbf{D}^\top \beta^\star) \right\}\right|}{\left|\mathrm{supp}(\mathbf{D}^\top \beta^\star)\right|}, & \text{if } \mathbf{D}^\top \beta^\star \neq 0 \\ 0, & \text{if } \mathbf{D}^\top \beta^\star = 0 \end{cases}$$

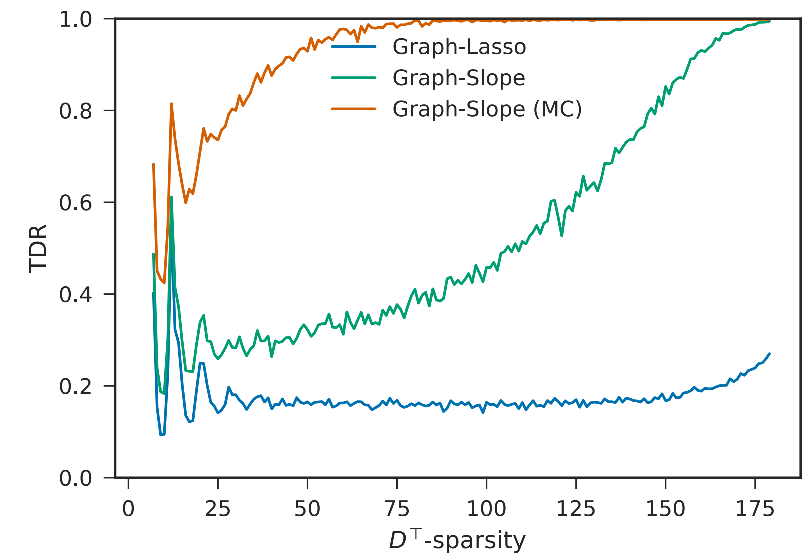


Synthetic Results

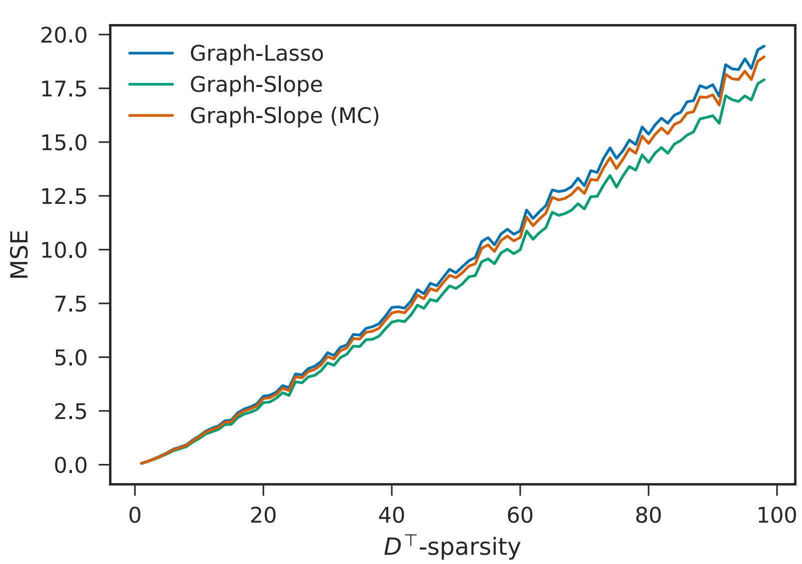
Caveman

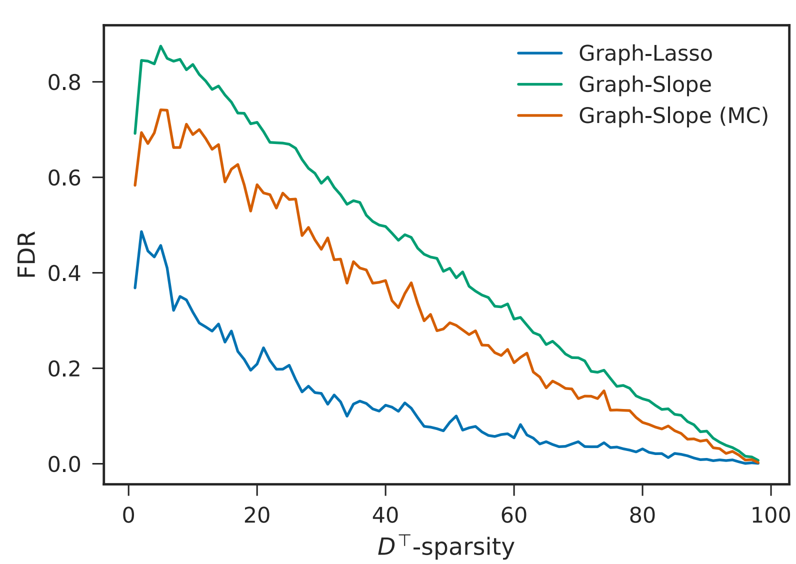


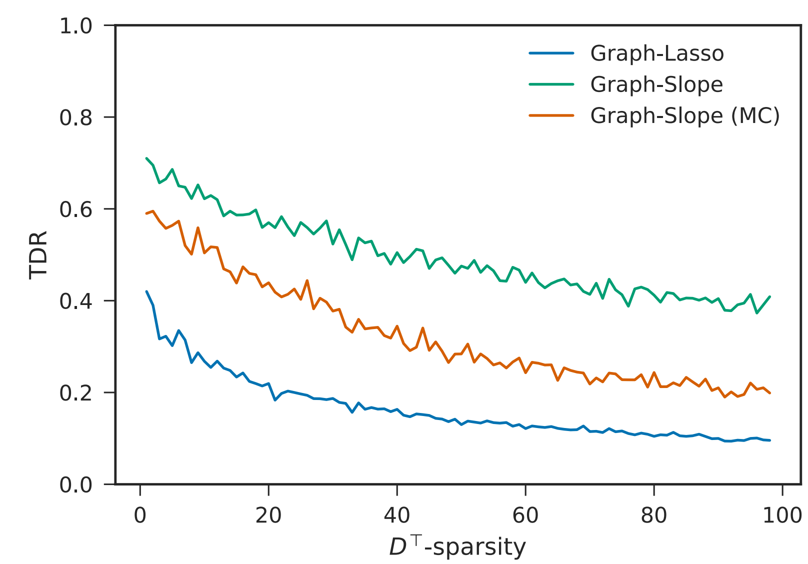




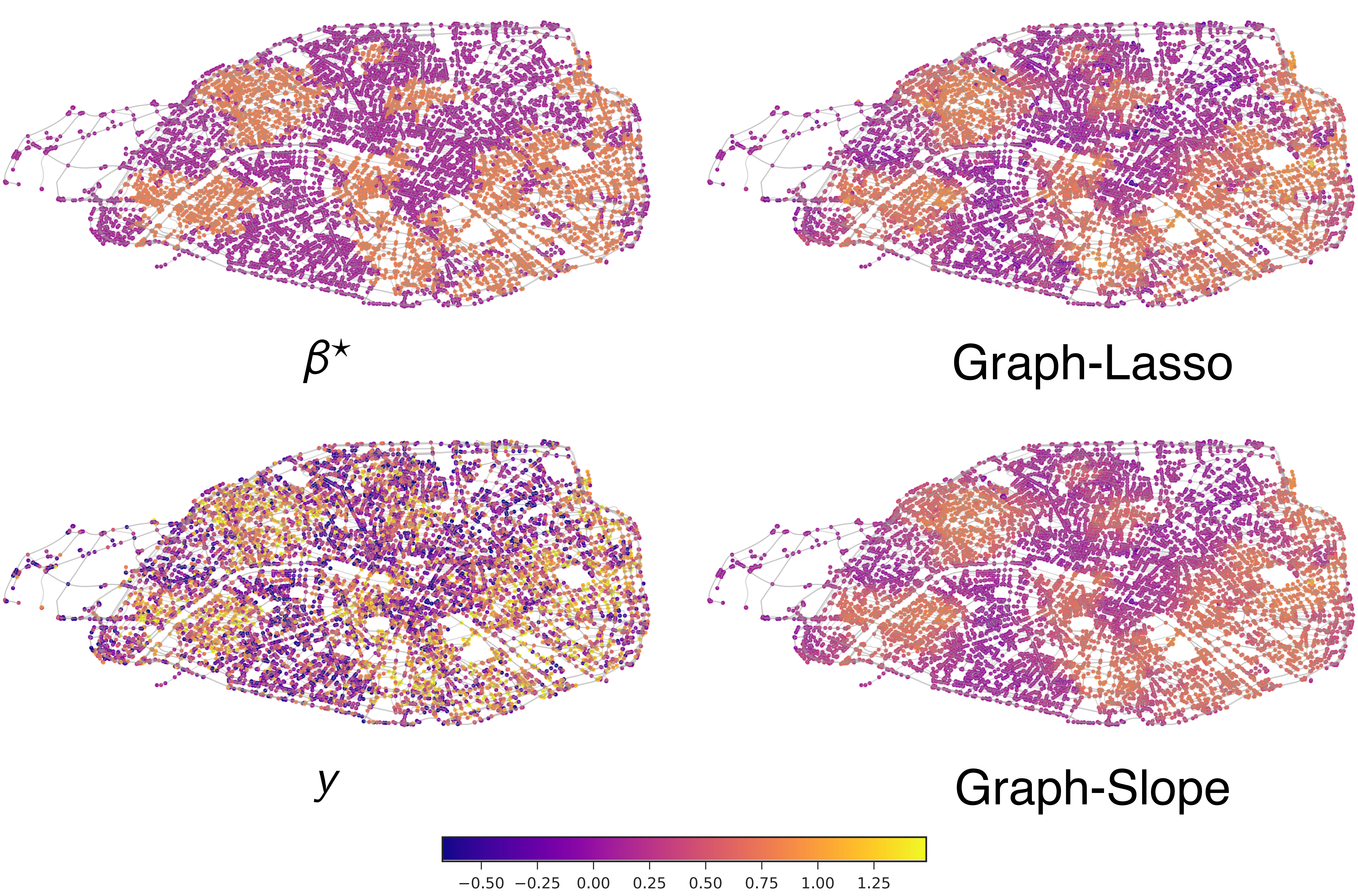
TV-1D (path graph)







Infect Paris!



Take-Away

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\text{argmin}} \; \frac{1}{2n}\|y - \beta\|^2 + \lambda J(\mathbf{D}^\top \beta)$$

Graph-Lasso ⟵ Lasso [Tibshirani ’95, Donoho '95]

new estimator: **Graph-Slope** ⟵ Slope [Bogdan et al. ’14]

- better statistical properties (oracle inequality rate)
- roughly the same computational complexity
- "better" (but similar) practical results

Perspectives

Regression

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n} \|y - \beta\|^2 + \lambda J(\mathbf{D}^\top \beta)$$

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{2n} \|y - X\beta\|^2 + \lambda J(\mathbf{D}^\top \beta)$$

Better algorithms: “safe-rules” & no-sorting dependency

Practical choice of weights for large graphs

Efficient debiasing strategy (CLEAR [Deledalle et al. ’16]?)

Real-life applications! (please help us)

Thanks for your attention!

Pierre C. Bellec, Joseph Salmon, SV,
A sharp oracle inequality for Graph-Slope,
Electron. J. Statist., 2017.

Jupyter notebook & source-code available at
http://github.com/svaiter/gslope_oracle_inequality