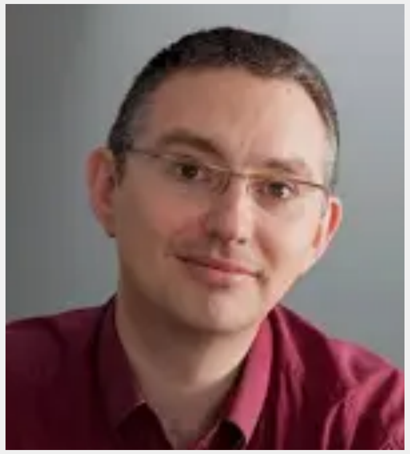
Learning graphs with precision matrices: statistical estimators, compressive approches, unrolled neural networks

Titouan Vayer

Can Pouliquen | Etienne Lasalle | Mathurin Massias | Rémi Gribonval | Paulo Gonçalves

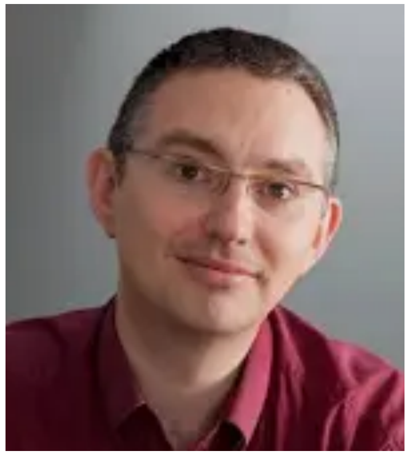



Overview of the talk

- **Part I: Finding graphs from unstructured data**
- **Part II: Schur's Positive-Definite Network**

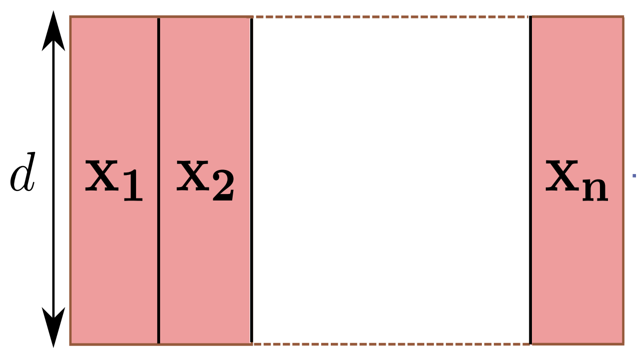
- **Part III: The sketching approach**



Graph Learning

- Input: a dataset

$$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$$

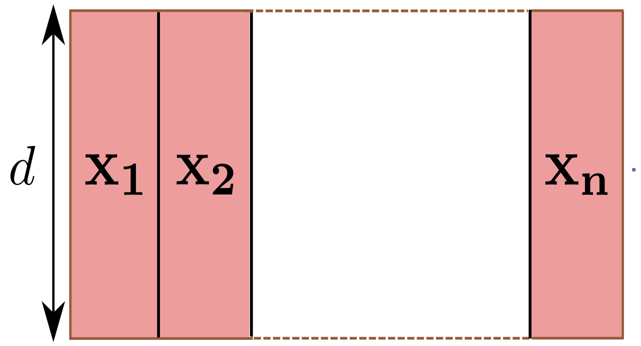


$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

Graph Learning

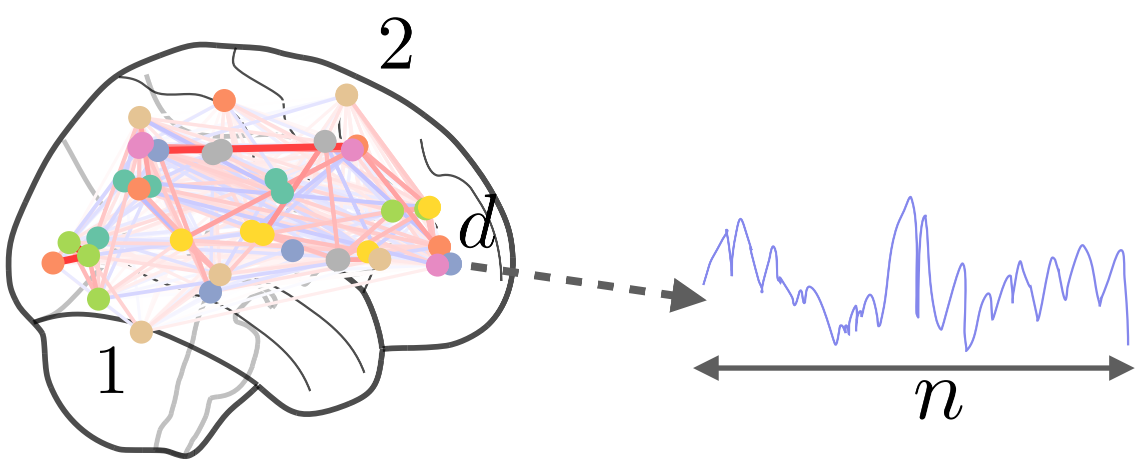
- Input: a dataset

$$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$$



$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

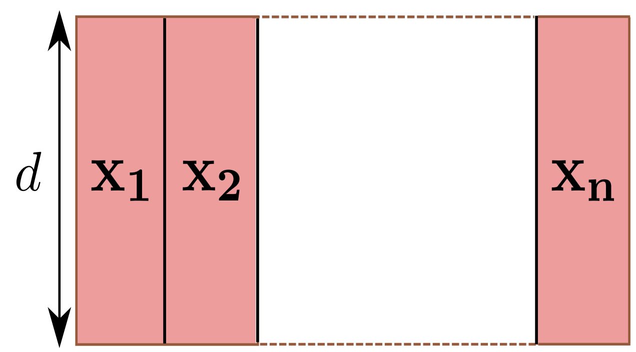
- Output: graph of **relations between the d variables**



Graph Learning

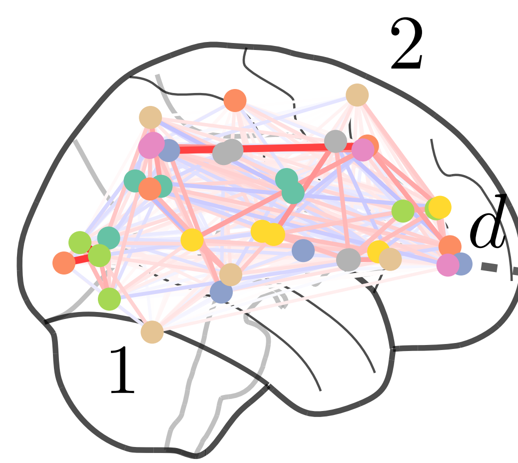
Input: a dataset

$$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$$

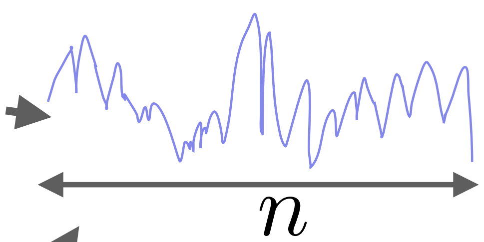


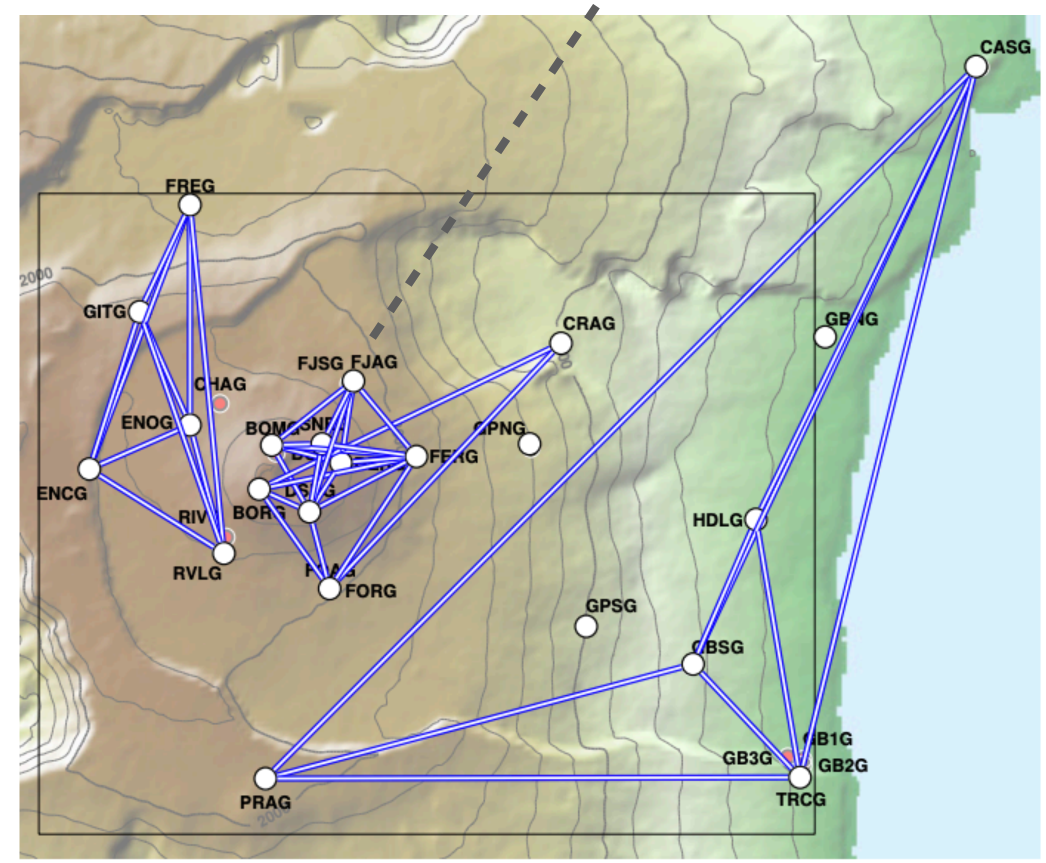
$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

Output: graph of **relations between the d variables**



also: genomics,
biological networks, energy…

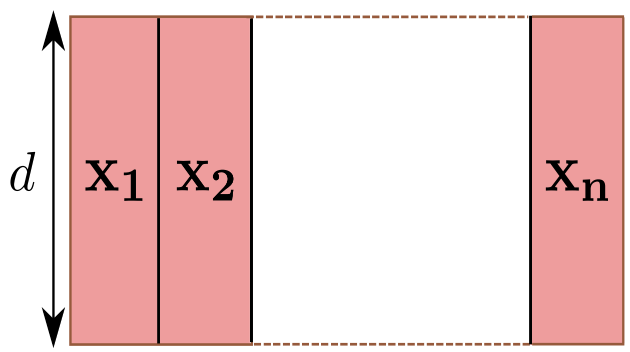




Graph Learning

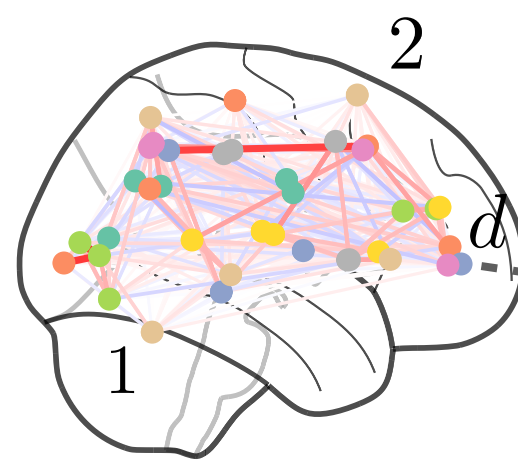
Input: a dataset

$$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$$

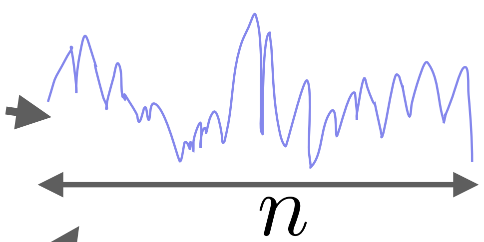


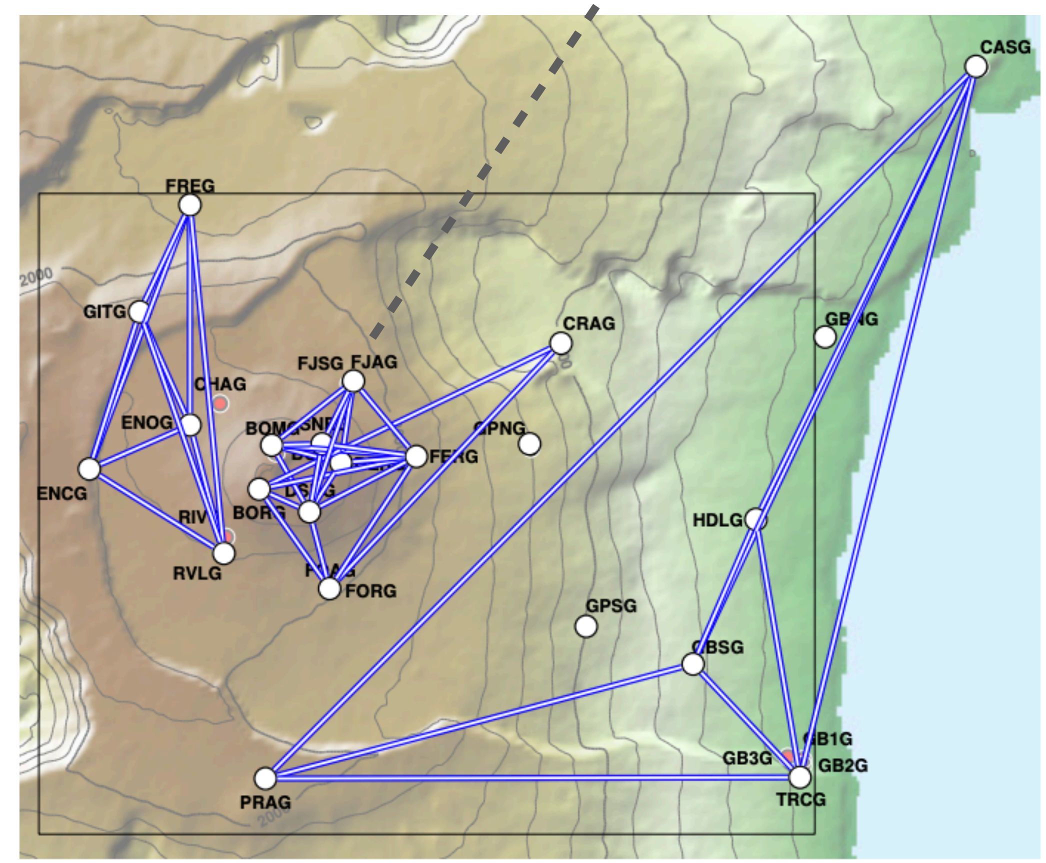
$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

Output: graph of **relations between the d variables**



also: genomics, biological networks, energy…





Graph modeled as a matrix:

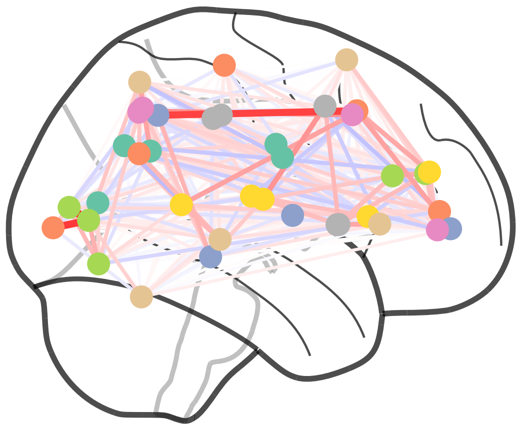
$$\mathbf{\Theta} \in \mathbb{R}^{d \times d}$$

$\mathbf{\Theta}_{ij}$: interaction between variable i and j

- statistical correlations
- statistical dependencies

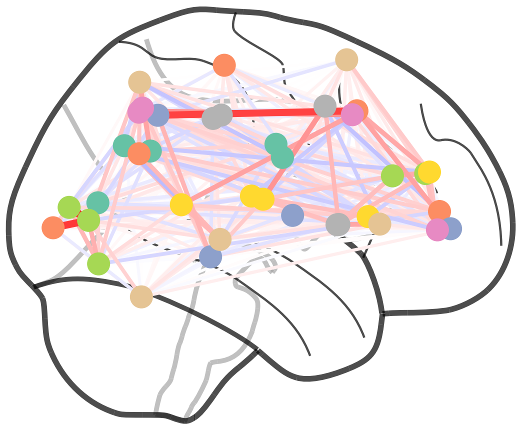
Epilepsy: the big picture

- One of the most common neurological disorder, affecting 1% of the global population
- Nearly 30% of patients are drug-resistant
- Surgical solution: remove the epileptic onset area (resection)

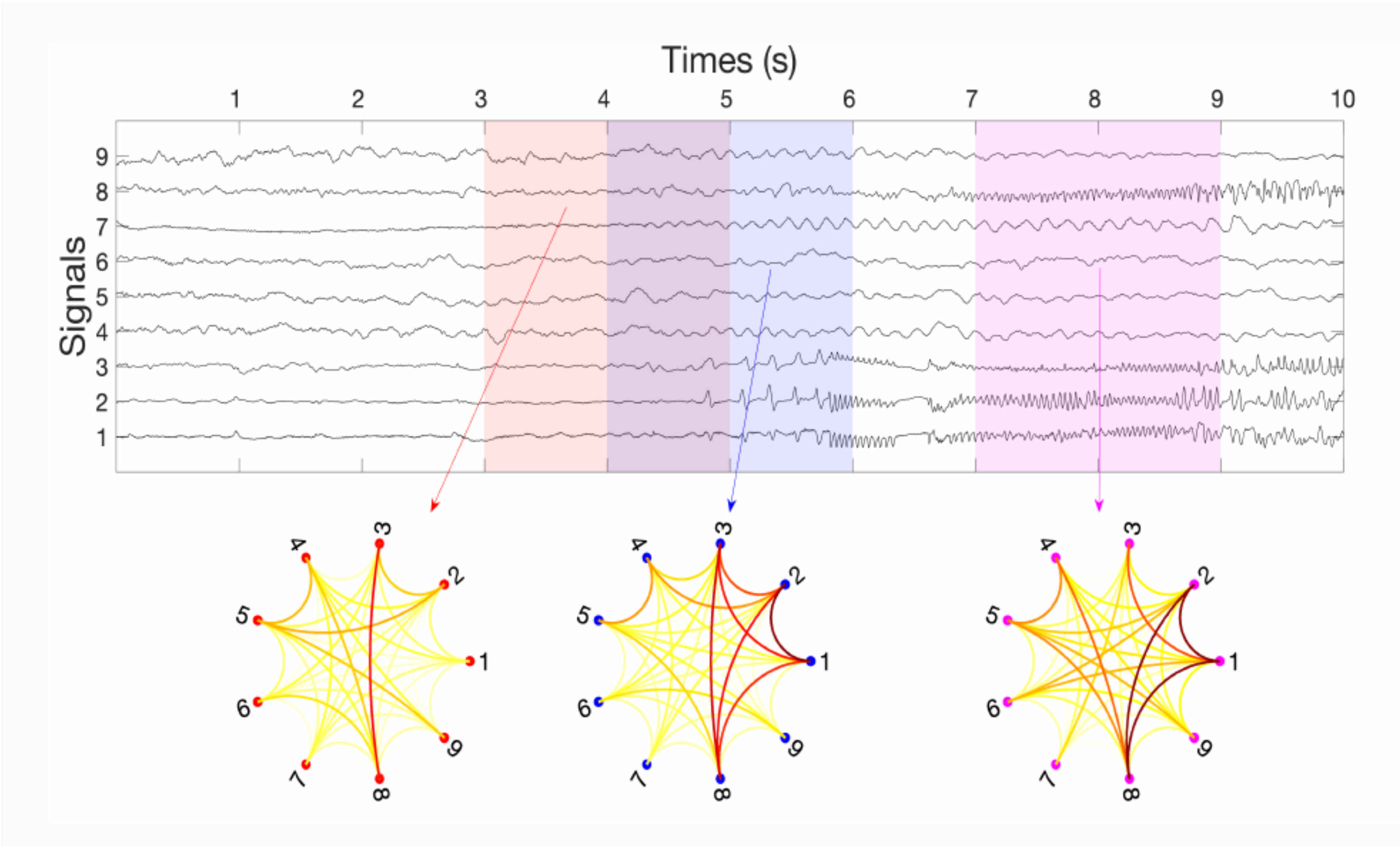


Epilepsy: the big picture

- One of the most common neurological disorder, affecting 1% of the global population
- Nearly 30% of patients are drug-resistant
- Surgical solution: remove the epileptic onset area (resection)



- Sparse dynamic functional connectivity graphs from EEG signals



Graphical LASSO

Side note

Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

Output: $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$

Graphical LASSO

Side note

Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

Output: $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \mathbf{\Sigma} = \mathbf{\Theta}^{-1})$

- $\mathbf{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Graphical LASSO

Side note

- Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$
  $\mathbf{x}_i \in \mathbb{R}^d \sim \mu$
- Output: $\mathbf{\Theta} \in \mathbb{R}^{d\times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \mathbf{\Sigma} = \mathbf{\Theta}^{-1})$

- $\mathbf{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Maximum Likehood estimator

Emp. cov. $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$

$$\mathbf{\Theta}_{\mathrm{MLE}} = \operatorname*{arg\,min}_{\mathbf{\Theta}\succ 0} - \operatorname{logdet}(\mathbf{\Theta}) + \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Theta}\rangle_F$$

- When $\widehat{\mathbf{\Sigma}}$ is invertible $\mathbf{\Theta}_{\mathrm{MLE}} = (\widehat{\mathbf{\Sigma}})^{-1}$

Graphical LASSO

Side note

- Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

  $\mathbf{x}_i \in \mathbb{R}^d \sim \mu$
- Output: $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$

- $\boldsymbol{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Maximum Likehood estimator

Emp. cov. $\widehat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$

$$\boxed{\boldsymbol{\Theta}_{\mathrm{MLE}} = \underset{\boldsymbol{\Theta} \succ 0}{\arg\min} - \operatorname{logdet}(\boldsymbol{\Theta}) + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle_F}$$

- When $\widehat{\boldsymbol{\Sigma}}$ is invertible $\boldsymbol{\Theta}_{\mathrm{MLE}} = (\widehat{\boldsymbol{\Sigma}})^{-1}$

May not be true

in high dim $n < d$

Graphical LASSO

Side note

Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

Output: $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \mathbf{\Sigma} = \mathbf{\Theta}^{-1})$

- $\mathbf{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

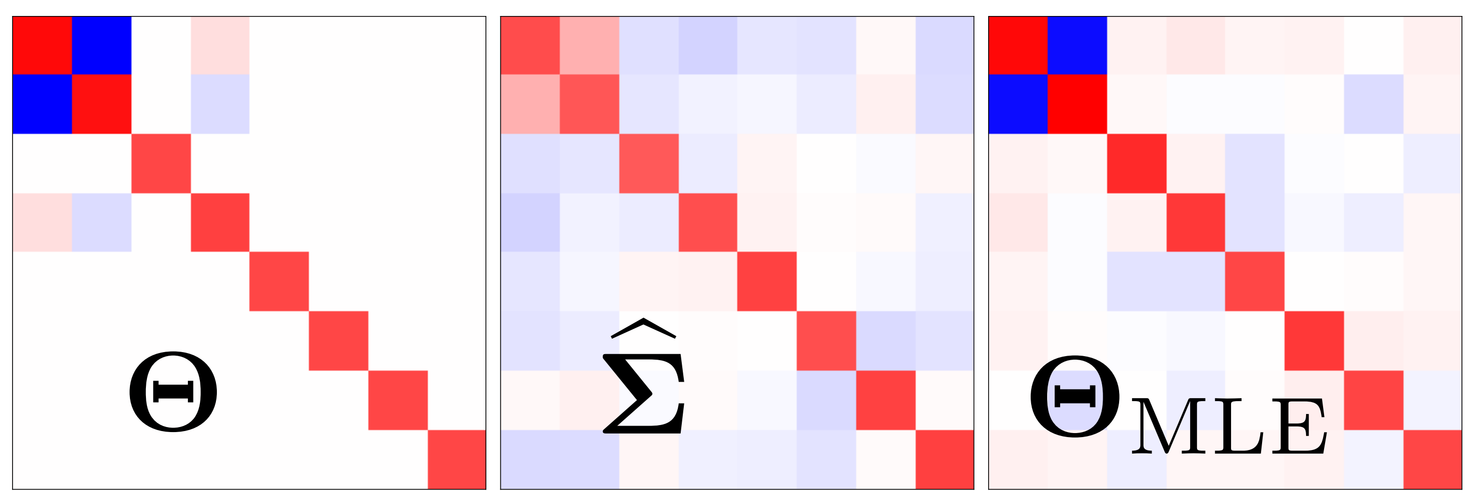
Maximum Likehood estimator

Emp. cov. $\widehat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top$

$$\mathbf{\Theta}_{\mathrm{MLE}} = \underset{\mathbf{\Theta} \succ 0}{\arg\min} - \operatorname{logdet}(\mathbf{\Theta}) + \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Theta} \rangle_F$$

- When $\widehat{\mathbf{\Sigma}}$ is invertible $\mathbf{\Theta}_{\mathrm{MLE}} = (\widehat{\mathbf{\Sigma}})^{-1}$ usually **not sparse**

May not be true
in high dim $n < d$



Graphical LASSO

Side note

Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

Output: $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

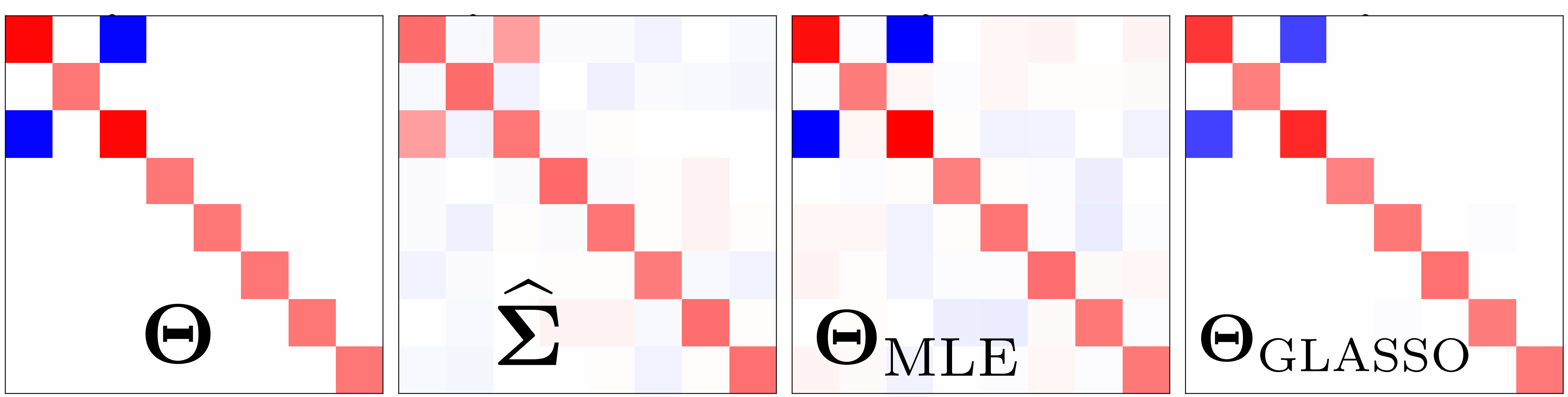
Gaussian assumption $\mu = \mathcal{N}(0, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$

- $\boldsymbol{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Penalized Maximum Likehood estimator [Friedman-Hastie-Tibshirani, 2007]

$$\boldsymbol{\Theta}_{\mathrm{GLASSO}} = \underset{\boldsymbol{\Theta} \succ 0}{\arg\min} - \operatorname{logdet}(\boldsymbol{\Theta}) + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle_F + \lambda \|\boldsymbol{\Theta}\|_{1,\mathrm{off}}$$

$\|\boldsymbol{\Theta}\|_{1,\mathrm{off}} = \sum_{i<j} |\boldsymbol{\Theta}_{ij}|$ promotes **sparsity** for the output graph

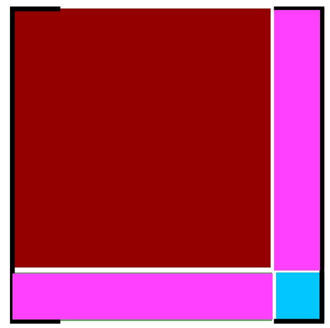


Graphical LASSO

Solving for GLASSO

$$\underset{\Theta \succ 0}{\operatorname{argmin}} \; F(\Theta) = -\operatorname{logdet}(\Theta) + \langle \hat{\Sigma}, \Theta \rangle + \lambda \|\Theta\|_{1,\text{off}}$$

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^{\top} & \theta_{22} \end{pmatrix}$$



- Schur complement

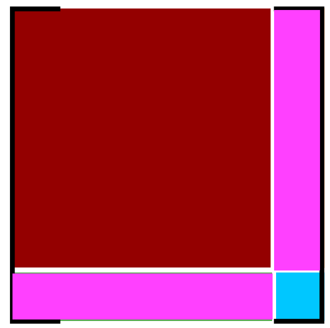
$$\det(\Theta) = \det(\Theta_{11}) \times \det(\theta_{22} - \theta_{12}{}^{\top} \Theta_{11}^{-1} \theta_{12})$$

Graphical LASSO

Solving for GLASSO

$$\underset{\Theta \succ 0}{\operatorname{argmin}} \; F(\Theta) = -\operatorname{logdet}(\Theta) + \langle \hat{\Sigma}, \Theta \rangle + \lambda \|\Theta\|_{1,\text{off}}$$

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix}$$



- Schur complement

$$\det(\Theta) = \det(\Theta_{11}) \times \det(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12})$$

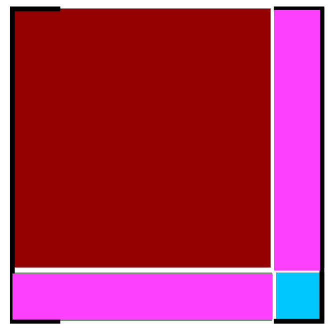
$$F(\Theta) = -\operatorname{logdet}(\Theta_{11}) - \log(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12}) + \hat{\sigma}_{22}\theta_{22} + \langle \hat{\Sigma}_{11}, \Theta_{11} \rangle + 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\Theta_{11}\|_{1,\text{off}} + \lambda \|\theta_{12}\|_1$$

Graphical LASSO

Solving for GLASSO

$$\underset{\Theta \succ 0}{\operatorname{argmin}} \; F(\Theta) = -\operatorname{logdet}(\Theta) + \langle \hat{\Sigma}, \Theta \rangle + \lambda \|\Theta\|_{1,\text{off}}$$

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix}$$



Schur complement

$$\det(\Theta) = \det(\Theta_{11}) \times \det(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12})$$

$$F(\Theta) = -\operatorname{logdet}(\Theta_{11}) - \log(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12}) + \hat{\sigma}_{22}\theta_{22} + \langle \hat{\Sigma}_{11}, \Theta_{11} \rangle + 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\Theta_{11}\|_{1,\text{off}} + \lambda \|\theta_{12}\|_1$$

BCD algorithm:

$$\underset{\theta_{22}}{\operatorname{argmin}} \; F(\Theta_{11}, \theta_{12}, \theta_{22}) = \frac{1}{\hat{\sigma}_{22}} + {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12} := \theta_{22}^\star$$

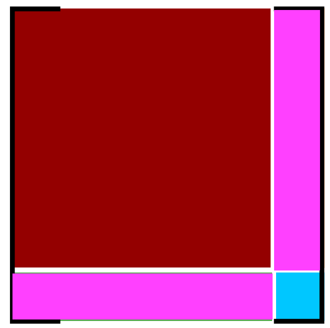
$$\underset{\theta_{12}}{\operatorname{argmin}} \; F(\Theta_{11}, \theta_{12}, \theta_{22}^\star) = 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\theta_{12}\|_1 + \hat{\sigma}_{22} {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12}$$

Graphical LASSO

Solving for GLASSO

$$\underset{\Theta \succ 0}{\operatorname{argmin}} \; F(\Theta) = -\operatorname{logdet}(\Theta) + \langle \hat{\Sigma}, \Theta \rangle + \lambda \|\Theta\|_{1,\text{off}}$$

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^{\top} & \theta_{22} \end{pmatrix}$$



Schur complement

$$\det(\Theta) = \det(\Theta_{11}) \times \det(\theta_{22} - {\theta_{12}}^{\top} \Theta_{11}^{-1} \theta_{12})$$

$$F(\Theta) = -\operatorname{logdet}(\Theta_{11}) - \log(\theta_{22} - {\theta_{12}}^{\top} \Theta_{11}^{-1} \theta_{12}) + \hat{\sigma}_{22}\theta_{22} + \langle \hat{\Sigma}_{11}, \Theta_{11} \rangle + 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\Theta_{11}\|_{1,\text{off}} + \lambda \|\theta_{12}\|_1$$

BCD algorithm: $\underset{\theta_{22}}{\operatorname{argmin}} \; F(\Theta_{11}, \theta_{12}, \theta_{22}) = \dfrac{1}{\hat{\sigma}_{22}} + {\theta_{12}}^{\top} \Theta_{11}^{-1} \theta_{12} := \theta_{22}^{\star}$

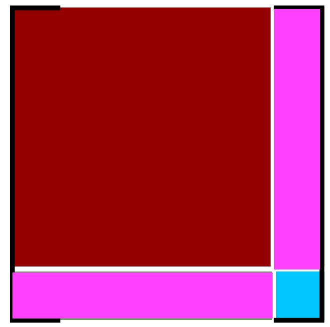
LASSO $\sim$ $\underset{\theta_{12}}{\operatorname{argmin}} \; F(\Theta_{11}, \theta_{12}, \theta_{22}^{\star}) = 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\theta_{12}\|_1 + \hat{\sigma}_{22} {\theta_{12}}^{\top} \Theta_{11}^{-1} \theta_{12}$

Graphical LASSO

Solving for GLASSO

$$\underset{\Theta \succ 0}{\operatorname{argmin}} \; F(\Theta) = -\operatorname{logdet}(\Theta) + \langle \hat{\Sigma}, \Theta \rangle + \lambda \|\Theta\|_{1,\text{off}}$$

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix}$$



Schur complement

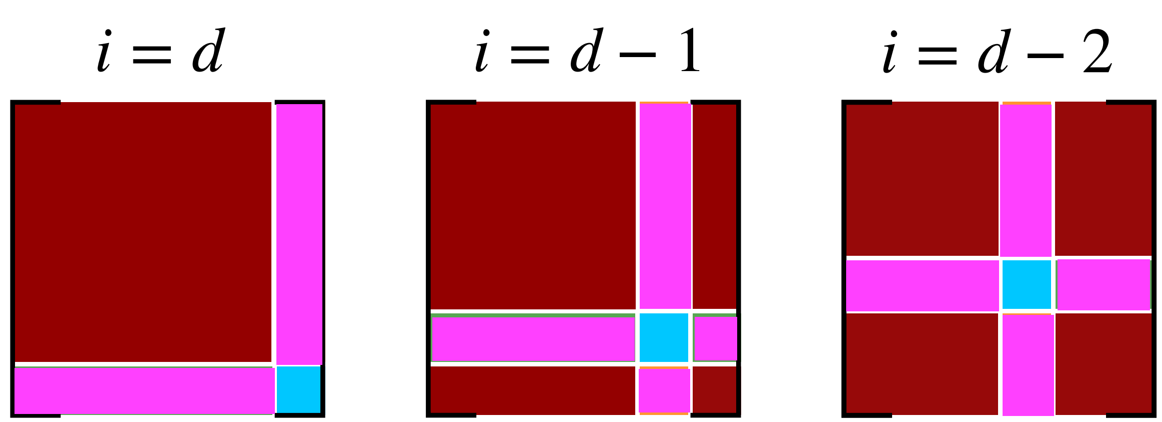
$$\det(\Theta) = \det(\Theta_{11}) \times \det(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12})$$

$$F(\Theta) = -\operatorname{logdet}(\Theta_{11}) - \log(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12}) + \hat{\sigma}_{22}\theta_{22} + \langle \hat{\Sigma}_{11}, \Theta_{11} \rangle + 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\Theta_{11}\|_{1,\text{off}} + \lambda \|\theta_{12}\|_1$$

BCD algorithm: $\underset{\theta_{22}}{\operatorname{argmin}} \; F(\Theta_{11}, \theta_{12}, \theta_{22}) = \dfrac{1}{\hat{\sigma}_{22}} + {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12} := \theta_{22}^\star$

$$\text{LASSO} \sim \underset{\theta_{12}}{\operatorname{argmin}} \; F(\Theta_{11}, \theta_{12}, \theta_{22}^\star) = 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\theta_{12}\|_1 + \hat{\sigma}_{22} {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12}$$

Then iterates on the columns

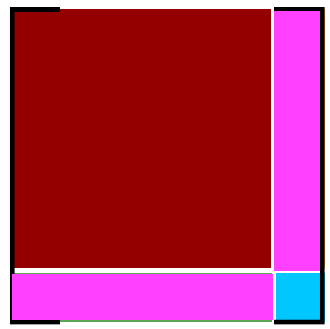


Graphical LASSO

Solving for GLASSO

$$\underset{\Theta \succ 0}{\operatorname{argmin}} \ F(\Theta) = -\operatorname{logdet}(\Theta) + \langle \hat{\Sigma}, \Theta \rangle + \lambda \|\Theta\|_{1,\text{off}}$$

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix}$$



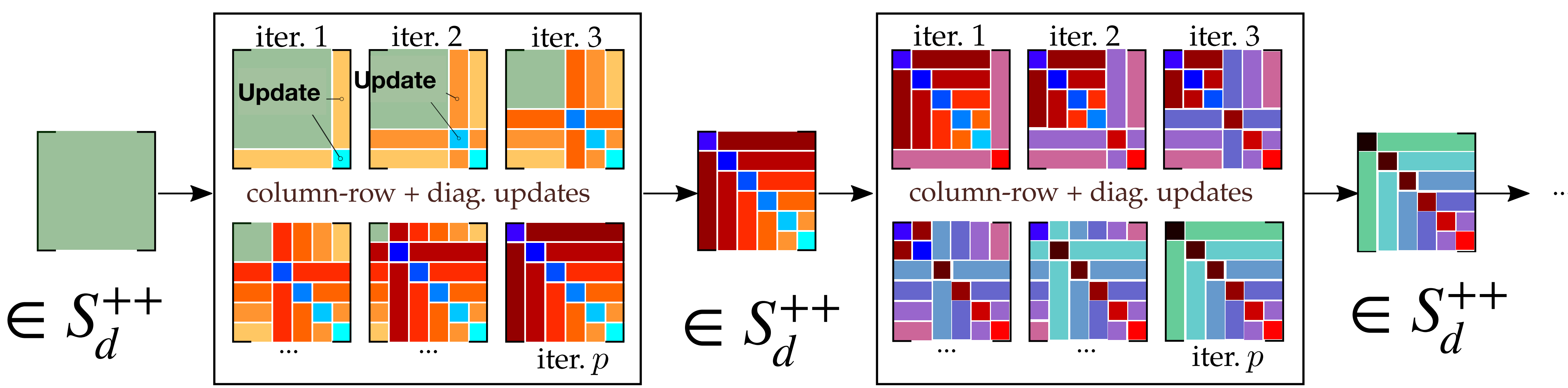
Schur complement

$$\det(\Theta) = \det(\Theta_{11}) \times \det(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12})$$

$$F(\Theta) = -\operatorname{logdet}(\Theta_{11}) - \log(\theta_{22} - {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12}) + \hat{\sigma}_{22}\theta_{22} + \langle \hat{\Sigma}_{11}, \Theta_{11} \rangle + 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\Theta_{11}\|_{1,\text{off}} + \lambda \|\theta_{12}\|_1$$

BCD algorithm: $\underset{\theta_{22}}{\operatorname{argmin}} \ F(\Theta_{11}, \theta_{12}, \theta_{22}) = \dfrac{1}{\hat{\sigma}_{22}} + {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12} := \theta_{22}^\star$

$$\text{LASSO} \sim \underset{\theta_{12}}{\operatorname{argmin}} \ F(\Theta_{11}, \theta_{12}, \theta_{22}^\star) = 2\langle \hat{\sigma}_{12}, \theta_{12} \rangle + \lambda \|\theta_{12}\|_1 + \hat{\sigma}_{22} {\theta_{12}}^\top \Theta_{11}^{-1} \theta_{12}$$



Graphical LASSO

Side note

Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

Output: $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$

- $\boldsymbol{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Penalized Maximum Likehood estimator

$$\boldsymbol{\Theta}_{\text{GLASSO}} = \underset{\boldsymbol{\Theta} \succ 0}{\arg\min} - \operatorname{logdet}(\boldsymbol{\Theta}) + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle_F + \lambda \|\boldsymbol{\Theta}\|_{1,\text{off}}$$

- Optimization: convex problem

Coordinate descent

Involves LASSO steps (on the rows)

Graphical LASSO

Side note

- Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

  $\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

- Output: $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$

- $\boldsymbol{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Penalized Maximum Likehood estimator

$$\boldsymbol{\Theta}_{\mathrm{GLASSO}} = \operatorname*{arg\,min}_{\boldsymbol{\Theta} \succ 0} - \operatorname{logdet}(\boldsymbol{\Theta}) + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle_F + \lambda \|\boldsymbol{\Theta}\|_{1,\mathrm{off}}$$

- Optimization: convex problem

  Coordinate descent

  Involves LASSO steps (on the rows)

- Many large scale variants:

  QUIC, Big & QUIC [Hsieh & al, 2013-2014]

  SQUIC [Bollhöfer, 2019] + other estimators…

Graphical LASSO

Side note

Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

Output: $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \mathbf{\Sigma} = \mathbf{\Theta}^{-1})$

- $\mathbf{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Penalized Maximum Likehood estimator

$$\mathbf{\Theta}_{\mathrm{GLASSO}} = \underset{\mathbf{\Theta} \succ 0}{\arg\min} - \operatorname{logdet}(\mathbf{\Theta}) + \langle \widehat{\mathbf{\Sigma}}, \mathbf{\Theta} \rangle_F + \lambda \|\mathbf{\Theta}\|_{1,\mathrm{off}}$$

- Optimization: convex problem

  Coordinate descent

  Involves LASSO steps (on the rows)

- Many modelisation variants:

  $\mathbf{\Theta} = \mathcal{L}(\mathcal{G})$ is a Laplacian matrix of a graph [Kumar, 2020]

- Many large scale variants:

  QUIC, Big & QUIC [Hsieh & al, 2013-2014]

  SQUIC [Bollhöfer, 2019] + other estimators…

Graphical LASSO

Side note

Input: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$

$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$

Output: $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$

Gaussian Graphical Model

Gaussian assumption $\mu = \mathcal{N}(0, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1})$

- $\boldsymbol{\Theta}_{ij} = 0 \iff$ variable i is independent of j conditionally to the others

Penalized Maximum Likehood estimator

$$\boldsymbol{\Theta}_{\mathrm{GLASSO}} = \underset{\boldsymbol{\Theta} \succ 0}{\arg\min} - \operatorname{logdet}(\boldsymbol{\Theta}) + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle_F + \lambda \|\boldsymbol{\Theta}\|_{1,\mathrm{off}}$$

- Optimization: convex problem

  Coordinate descent

  Involves LASSO steps (on the rows)

- Many large scale variants:

  QUIC, Big & QUIC [Hsieh & al, 2013-2014]

  SQUIC [Bollhöfer, 2019] + other estimators…

- Many modelisation variants:

  $\boldsymbol{\Theta} = \mathcal{L}(\mathcal{G})$ is a Laplacian matrix of a graph [Kumar, 2020]

- Complexity of GLASSO:

	In memory	In time
$\widehat{\boldsymbol{\Sigma}}$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^3)$

Overview of the talk

- **Part I: Finding graphs from unstructured data**
- **Part II: Schur’s Positive-Definite Network**
- **Part III: The sketching approach**

SpodNet: Schur's Positive-Definite Network

Motivations

- Model-based vs learning-based approach

SpodNet: Schur's Positive-Definite Network

Motivations

- Model-based vs learning-based approach
- Neural network architecture for SDP learning **with structure**

$$\mathrm{NN} : S_d^{++} \to S_d^{++}$$ **focus on element-wise sparsity of output**

- Inspired by the GLASSO solver: **unrolled architecture**

SpodNet: Schur's Positive-Definite Network

Motivations

- Model-based vs learning-based approach
- Neural network architecture for SDP learning **with structure**

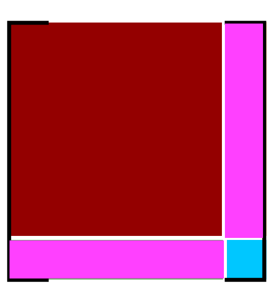
$$\mathrm{NN} : S_d^{++} \to S_d^{++}$$

focus on element-wise sparsity of output

- Inspired by the GLASSO solver: **unrolled architecture**

Key ingredient: Schur's condition for PSDness

$$\Theta = \begin{pmatrix} \Theta_{11} & \theta_{12} \\ \theta_{12}^\top & \theta_{22} \end{pmatrix} \succ 0 \iff \begin{array}{c} \Theta_{11} \succ 0 \\ \& \\ \theta_{22} - \theta_{12}^\top \Theta_{11}^{-1} \theta_{12} > 0 \end{array}$$

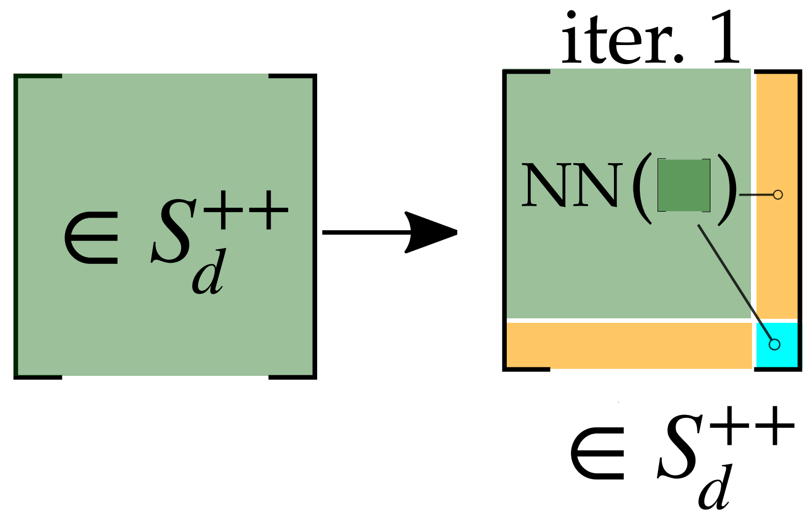


- Holds for any value of the column !

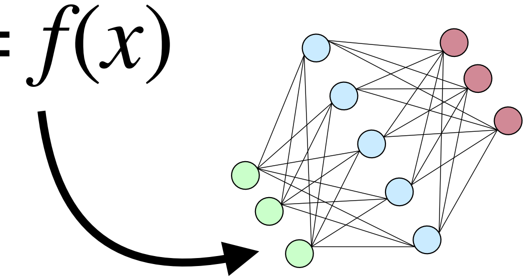
SpodNet: Schur's Positive-Definite Network

The architecture

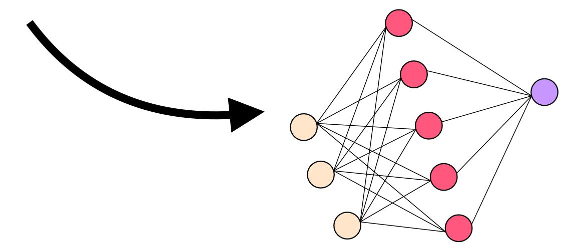
One layer



$$\theta_{12}^{+} = f(x)$$



$$\theta_{22}^{+} = g(y) + {\theta_{12}^{+}}^{\top} \Theta_{11}^{-1} \theta_{12}^{+}$$

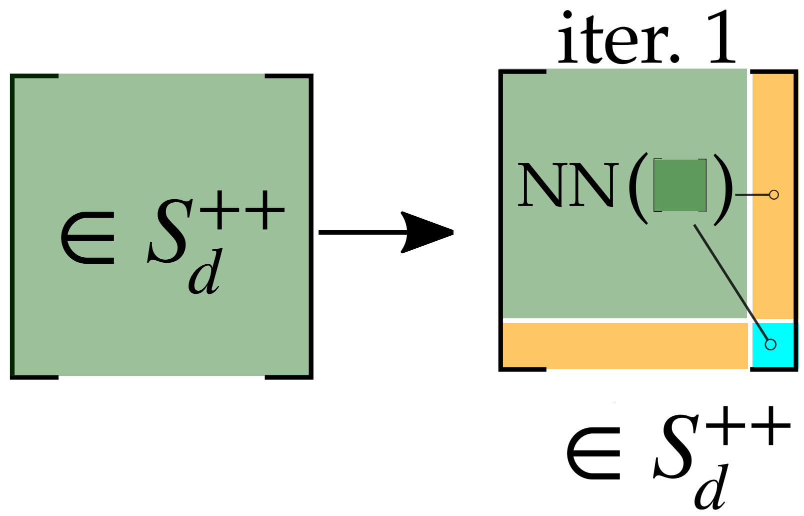


> 0

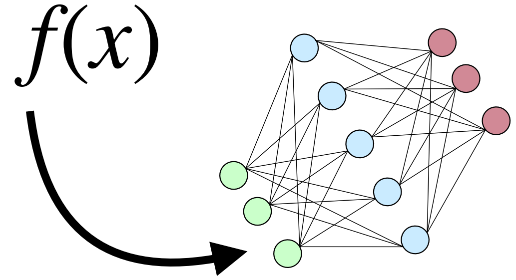
SpodNet: Schur's Positive-Definite Network

The architecture

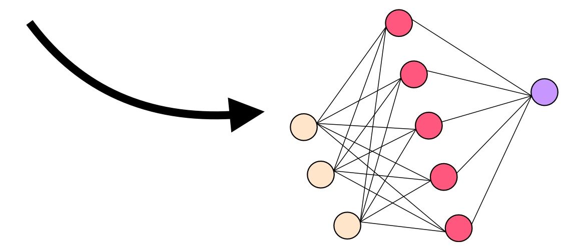
- One layer



$$\theta_{12}^{+} = f(x)$$

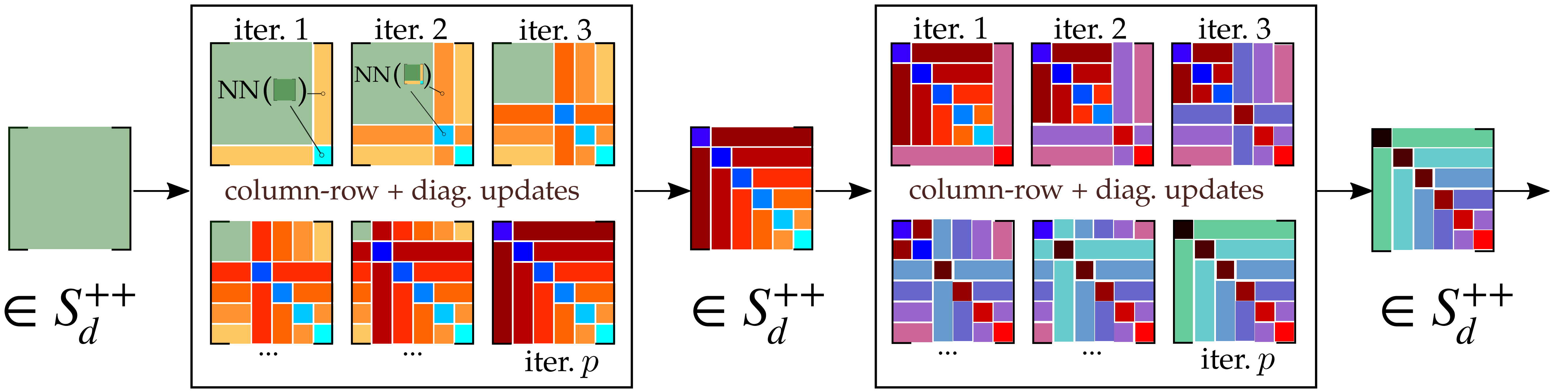


$$\theta_{22}^{+} = g(y) + {\theta_{12}^{+}}^{\top} \Theta_{11}^{-1} \theta_{12}^{+}$$



> 0

- Multiple layers:



SpodNet: Schur's Positive-Definite Network

Efficient updates

- Each iter. requires the computation of $\theta_{12}^{+\top} \Theta_{11}^{-1} \theta_{12}^{+} \longrightarrow \mathcal{O}(d^3)$

SpodNet: Schur's Positive-Definite Network

Efficient updates

- Each iter. requires the computation of $\theta_{12}^{+\top}\Theta_{11}^{-1}\theta_{12}^{+} \longrightarrow \mathcal{O}(d^3)$
- We can improve by keeping in memory $W = \Theta^{-1}$

$$\Theta_{11}^{-1} = W_{11} - \frac{1}{w_{22}} w_{12} w_{12}^\top$$

$$\theta_{12}^{+\top}\Theta_{11}^{-1}\theta_{12}^{+} \longrightarrow \mathcal{O}(d^2)$$

SpodNet: Schur's Positive-Definite Network

Efficient updates

- Each iter. requires the computation of $\theta_{12}^{+\top}\Theta_{11}^{-1}\theta_{12}^{+} \longrightarrow \mathcal{O}(d^3)$
- We can improve by keeping in memory $W = \Theta^{-1}$

$$\Theta_{11}^{-1} = W_{11} - \frac{1}{w_{22}} w_{12} w_{12}^{\top}$$

$$\theta_{12}^{+\top}\Theta_{11}^{-1}\theta_{12}^{+} \longrightarrow \mathcal{O}(d^2)$$

Maintaining the inverse

- Using Schur's inversion theorem if $\Theta^{+} = \begin{pmatrix} \Theta_{11} & \theta_{12}^{+} \\ \theta_{12}^{+\top} & \theta_{22}^{+} \end{pmatrix}$ is the update

$$W^{+} = \begin{pmatrix} [\Theta_{11}]^{-1} + \dfrac{[\Theta_{11}]^{-1}\theta_{12}^{+}\theta_{12}^{+\top}[\Theta_{11}]^{-1}}{g(y)} & -\dfrac{[\Theta_{11}]^{-1}\theta_{12}^{+}}{g(y)} \\ \left(-\dfrac{[\Theta_{11}]^{-1}\theta_{12}^{+}}{g(y)}\right)^{\top} & \dfrac{1}{g(y)} \end{pmatrix}$$

satisfies $W^{+} = [\Theta^{+}]^{-1}$

SpodNet: Schur's Positive-Definite Network

- **Examples of implementations**

SpodNet: Schur's Positive-Definite Network

Examples of implementations

Unrolled Block-Graphical ISTA (UBG)

$$f : (\theta_{12}, \hat{\sigma}_{12}, w_{12}) \to \mathrm{ST}_{\lambda^+} \left(\theta_{12} - \gamma^+ \times (\hat{\sigma}_{12} - w_{12})\right)$$

✦ λ^+, γ^+ are small perceptrons that learn regularization and step-size

SpodNet: Schur's Positive-Definite Network

Examples of implementations

- Unrolled Block-Graphical ISTA (UBG)

$$f : (\theta_{12}, \hat{\sigma}_{12}, w_{12}) \to \mathrm{ST}_{\lambda^+}\left(\theta_{12} - \gamma^+ \times (\hat{\sigma}_{12} - w_{12})\right)$$

✦ λ^+, γ^+ are small perceptrons that learn regularization and step-size

✦ Inspired by a proximal coordinate gradient descent step on the GLASSO objective

Expressivity +

Sparsity +++

Interpretability +++

SpodNet: Schur's Positive-Definite Network

Examples of implementations

- Unrolled Block-Graphical ISTA (UBG)

$$f : (\theta_{12}, \hat{\sigma}_{12}, w_{12}) \to \mathrm{ST}_{\lambda^+} \left(\theta_{12} - \gamma^+ \times (\hat{\sigma}_{12} - w_{12})\right)$$

- Plug and play

$$f : (\theta_{12}, \hat{\sigma}_{12}, w_{12}) \to \Psi \left(\theta_{12} - \gamma^+ \times (\hat{\sigma}_{12} - w_{12})\right)$$

✦ $\Psi : \mathbb{R}^{d-1} \to \mathbb{R}^{d-1}$ is a learned operator (proximal?)

Expressivity ++

Sparsity ++

Interpretability ++

SpodNet: Schur's Positive-Definite Network

Examples of implementations

- Unrolled Block-Graphical ISTA (UBG)

$$f : (\theta_{12}, \hat{\sigma}_{12}, w_{12}) \rightarrow \mathrm{ST}_{\lambda^+} \left(\theta_{12} - \gamma^+ \times (\hat{\sigma}_{12} - w_{12})\right)$$

- Plug and play

$$f : (\theta_{12}, \hat{\sigma}_{12}, w_{12}) \rightarrow \Psi \left(\theta_{12} - \gamma^+ \times (\hat{\sigma}_{12} - w_{12})\right)$$

- End to end

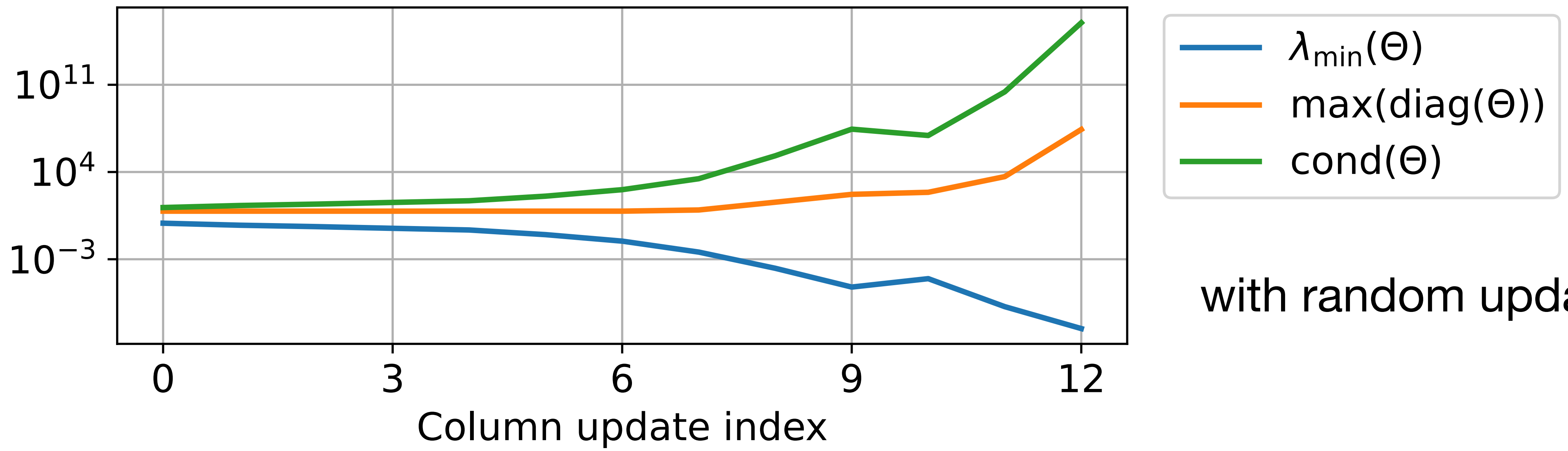
$$f : \theta_{12} \rightarrow \Phi \left(\theta_{12}\right)$$

- For all architectures

$$g = \mathrm{NN}(\theta_{22}, \hat{\sigma}_{22}, \mathrm{schur}) > 0$$

SpodNet: Schur's Positive-Definite Network

Nice but …



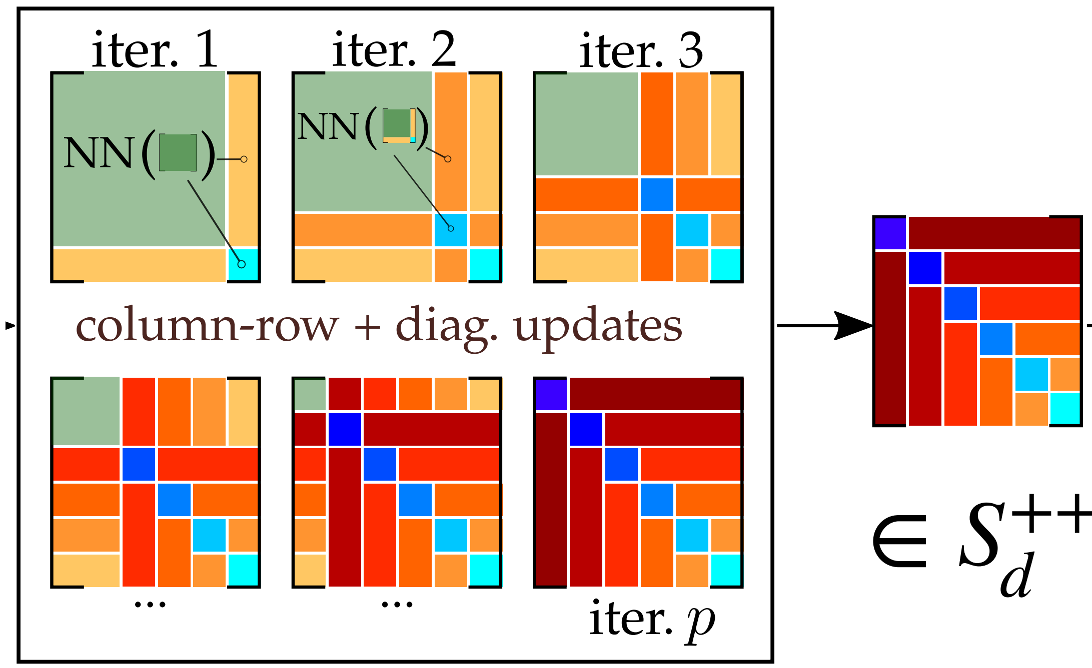
with random update of θ_{12}

- One workaround: « adaptive » normalization of the columns

SpodNet: Schur’s Positive-Definite Network

Learning with SpodNet

$\hat{\Sigma} + \epsilon I \longrightarrow$



$\in S_d^{++}$

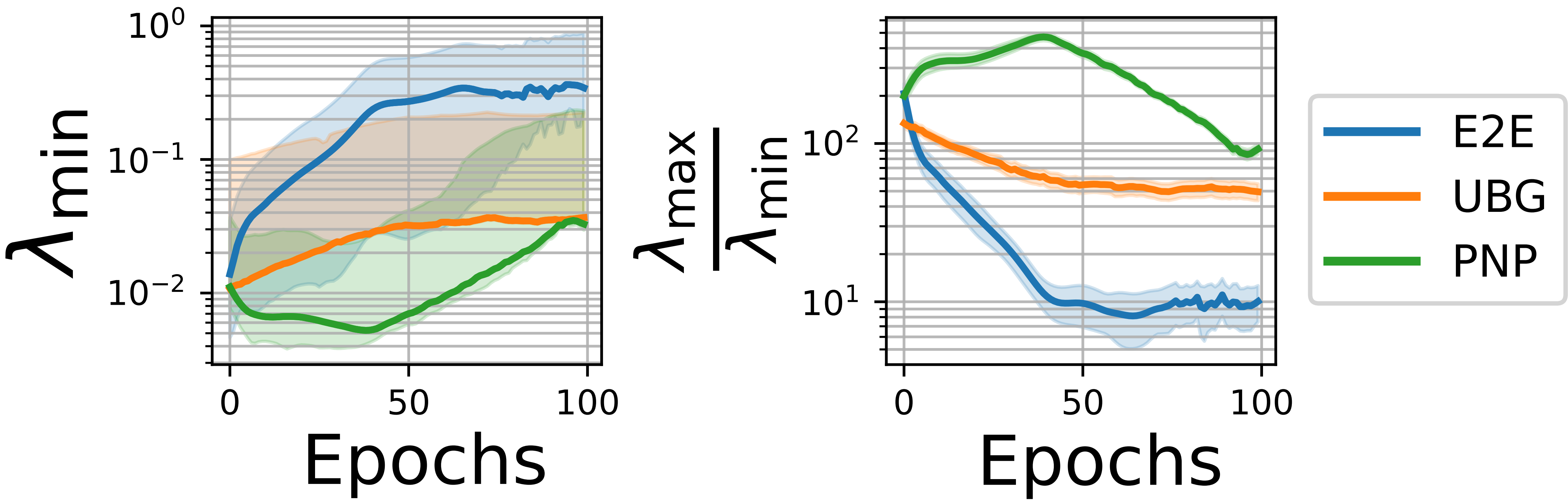
$\longrightarrow \bullet \bullet \bullet \longrightarrow \mathrm{SpodNet}(\hat{\Sigma}) = \hat{\Theta}$

Data driven approach

- Generate N sparse PSD matrices $(\Theta_i)_i$
- Minimize $L_{MSE} = \sum_{i=1}^{N} \|\Theta_i - \hat{\Theta}_i\|_F^2$

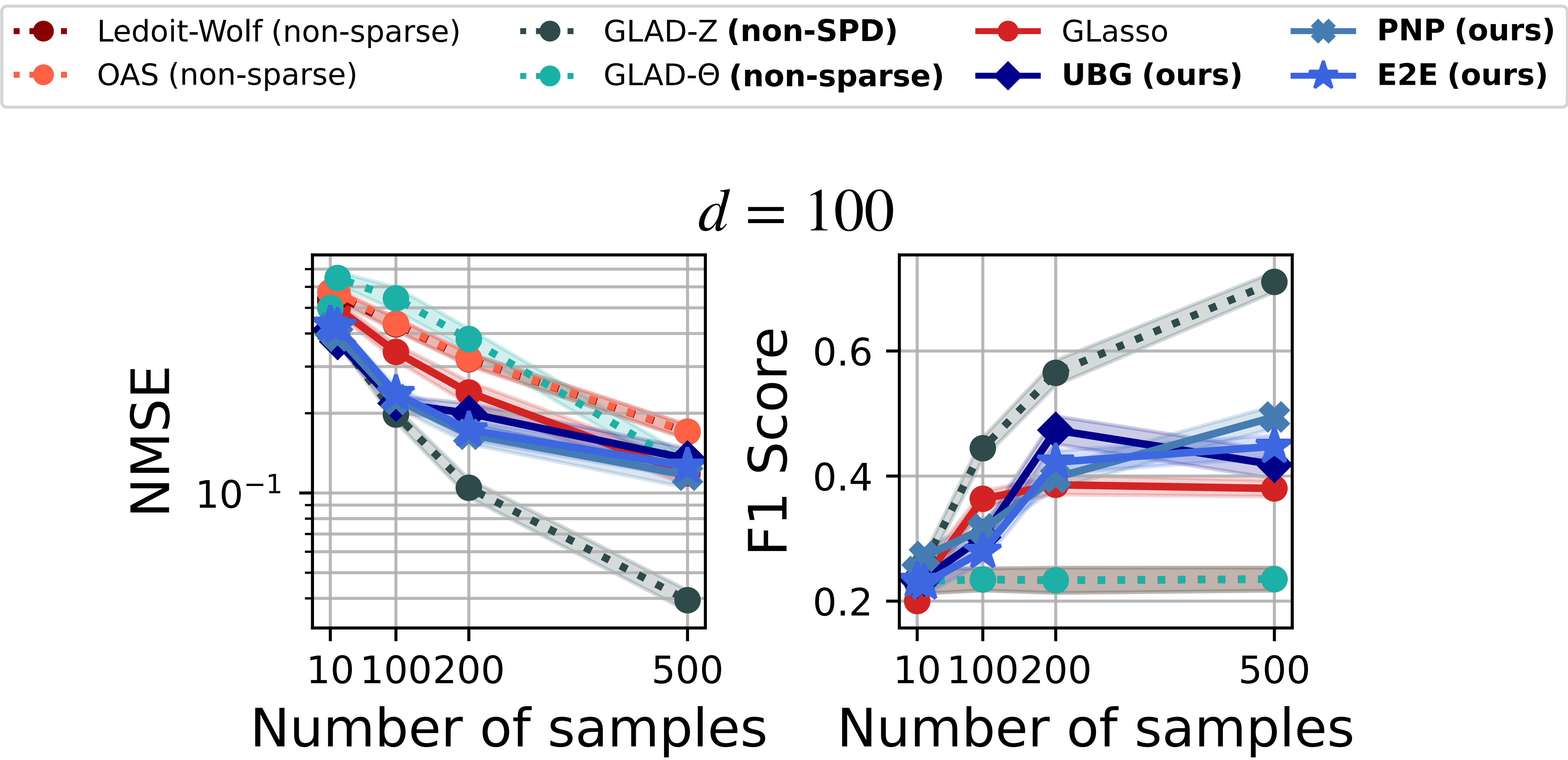
SpodNet: Schur's Positive-Definite Network

Stability



SpodNet: Schur's Positive-Definite Network

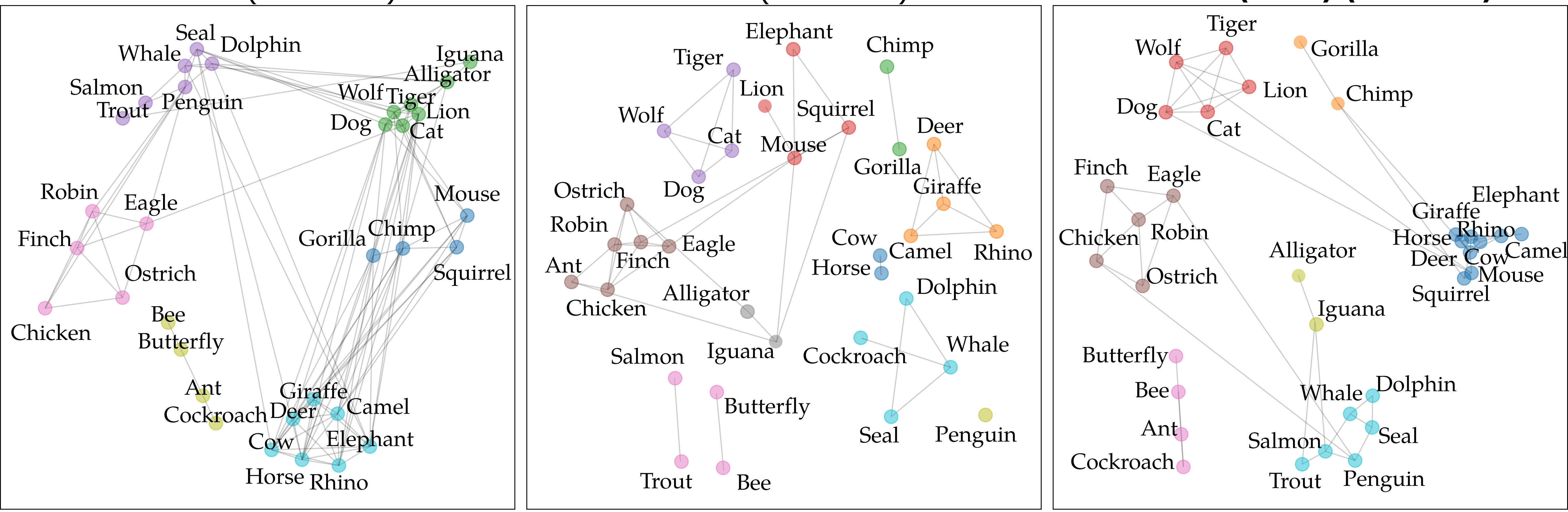
Performances on generated data



SpodNet: Schur's Positive-Definite Network

Performances on animals

GLasso (m = 0.61) | EGFM (m = 0.86) | **UBG (ours) (m=0.78)**



SpodNet: Schur's Positive-Definite Network

Takeaways

- Architecture for SPD + XXX is hard
- SpodNet: clever column/row update maintains SPD
- Can plug any column value: additional structure (e.g. sparse)
- Numerical stability is still a pain

Overview of the talk

- **Part I: Finding graphs from unstructured data**
- **Part II: Schur's Positive-Definite Network**
- **Part III: The sketching approach**

Graphical LASSO

Penalized Maximum Likehood estimator

$$\boldsymbol{\Theta}_{\text{GLASSO}} = \underset{\boldsymbol{\Theta} \succ 0}{\arg\min} - \operatorname{logdet}(\boldsymbol{\Theta}) + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle_F + \lambda \|\boldsymbol{\Theta}\|_{1,\text{off}}$$

- Optimization: convex problem

  Coordinate descent

  Involves LASSO steps (on the rows)

- Many large scale variants:

  QUIC, Big & QUIC [Hsieh & al, 2013-2014]

  SQUIC [Bollhöfer, 2019] + other estimators…

- Many modelisation variants:

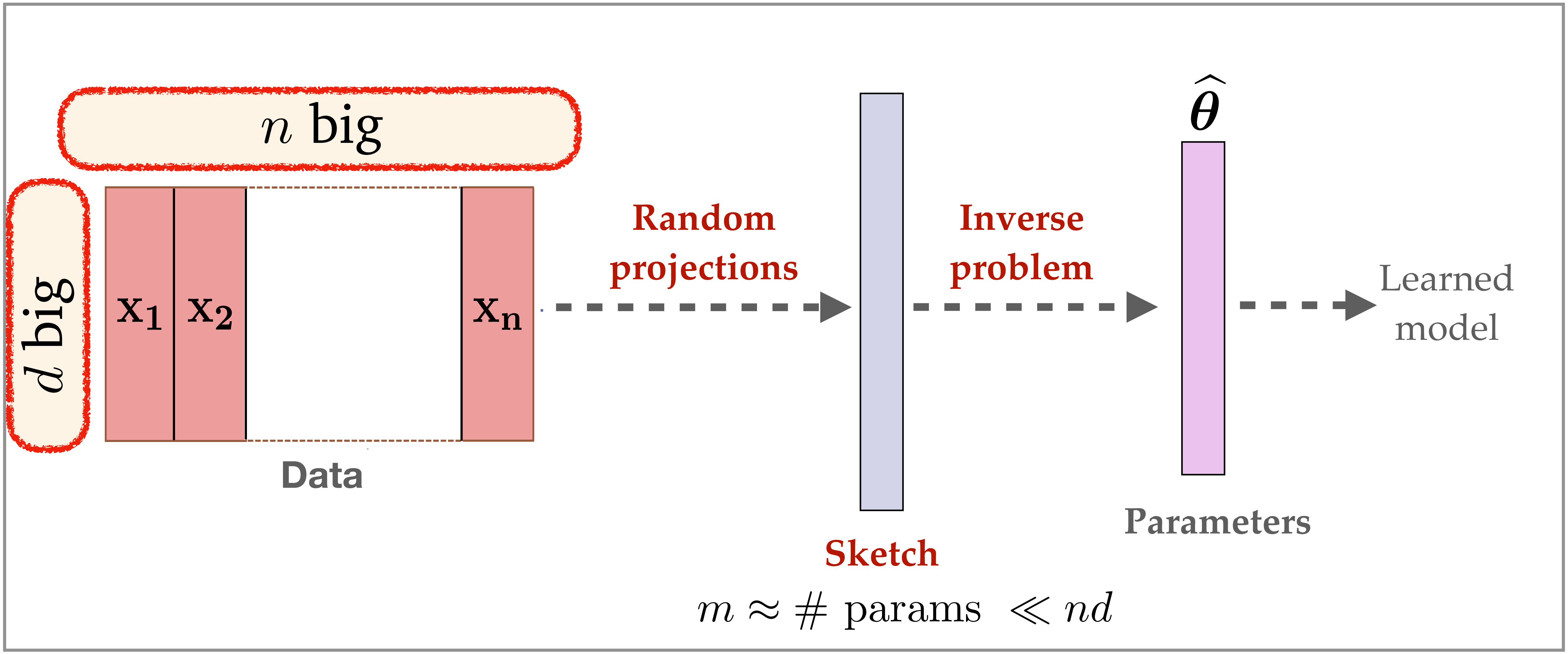
  $\boldsymbol{\Theta} = \mathcal{L}(\mathcal{G})$ is a Laplacian matrix of a graph [Kumar, 2020]

- Complexity of GLASSO:

	In memory	In time
$\widehat{\boldsymbol{\Sigma}} \dashrightarrow$	$\mathcal{O}(d^2)$	$\mathcal{O}(d^3)$

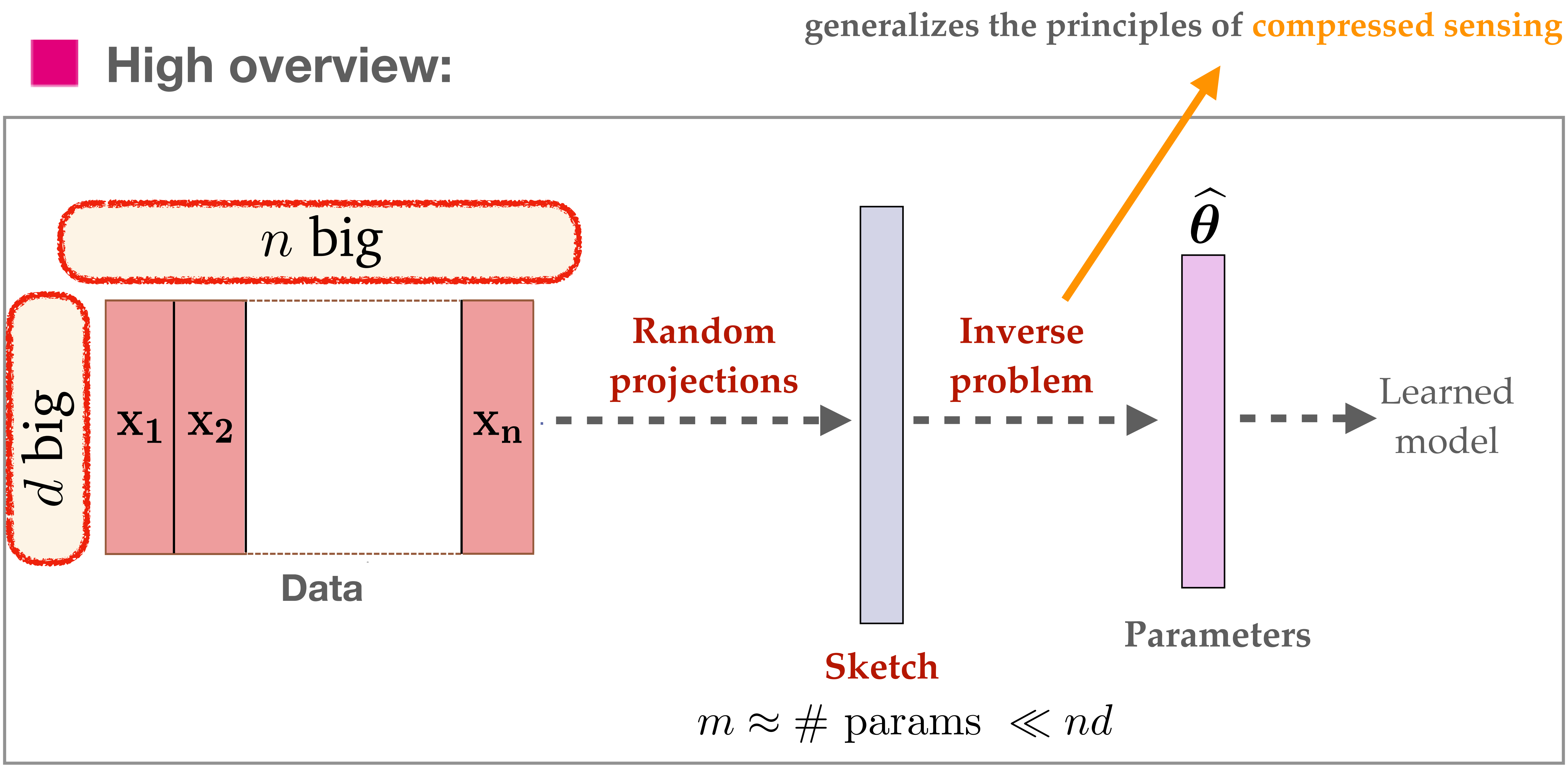
The sketching approach

- **High overview:**



The sketching approach

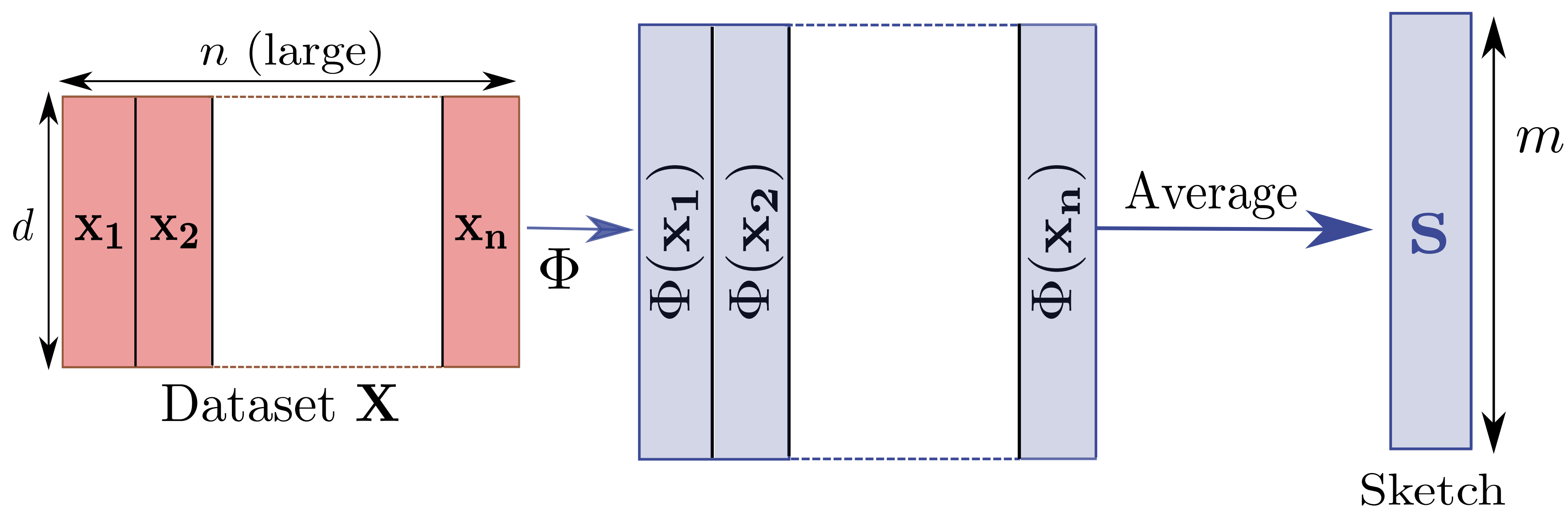
■ **High overview:**



The sketching approach

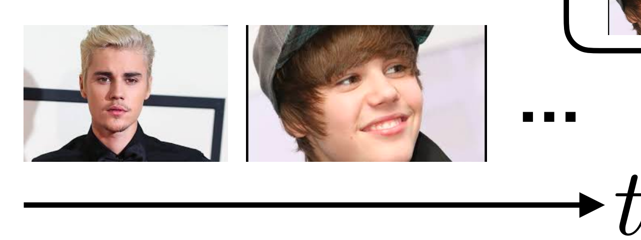
Obtaining the sketch

- A function called feature operator $\boxed{\Phi : \mathbb{R}^d \to \mathbb{R}^m}$
- Averaging **n points ->** $\mathbf{s} := \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i)$



Average is a simple idea but …

- Suitable for distributed /streaming scenarii
- It can be calculated in parallel

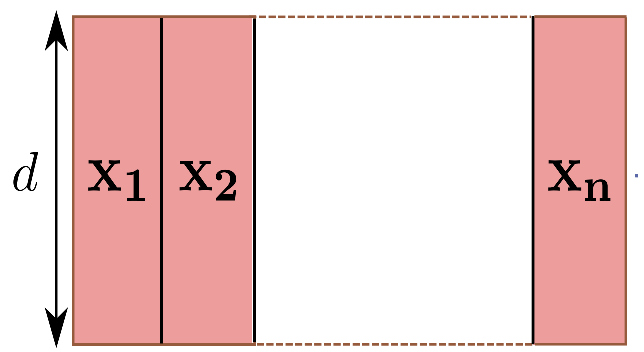


Goal of this talk

- Input: a dataset

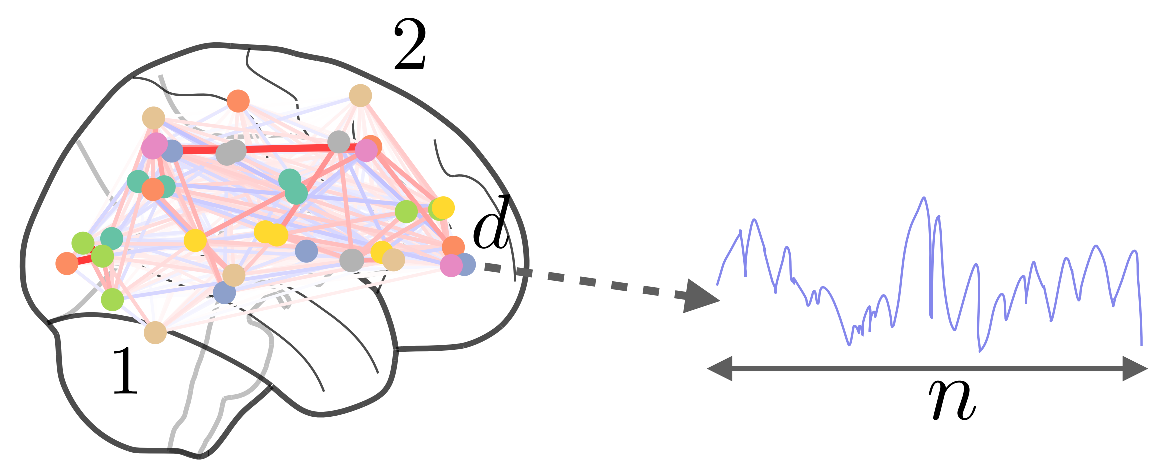
GLASSO

$$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$$



$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

- Output: graph of **relations between the d variables** $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$



GLASSO

	In memory	In time
$\widehat{\boldsymbol{\Sigma}}$ ▶	$\mathcal{O}(d^2)$	$\mathcal{O}(d^3)$

Goal of this talk

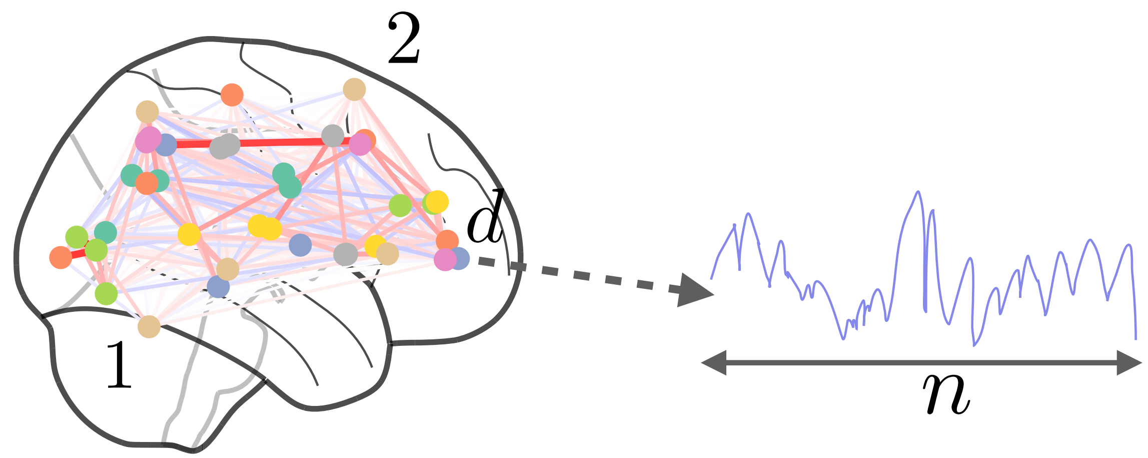
Input: a dataset

$$\mathbf{s} = \text{Sketch}(\mathbf{X}) \in \mathbb{R}^m$$

$m \approx \#$ edges

Sketch

Output: graph of **relations between the d variables** $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$



GLASSO

	In memory	In time
$\widehat{\boldsymbol{\Sigma}}$ ▶	$\mathcal{O}(d^2)$	$\mathcal{O}(d^3)$

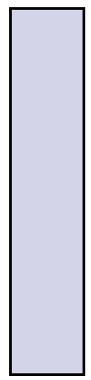
Sketching

	In memory	In time
$\mathbf{s}$ ▶	$\mathcal{O}(m) \ll d^2$	$\mathcal{O}(?)$

Goal of this talk

Input: a dataset

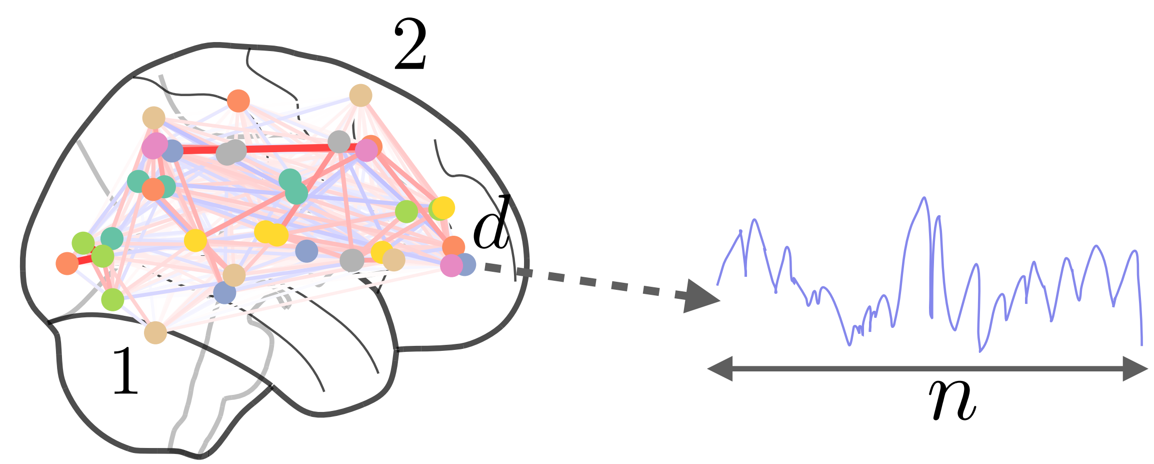
$$\mathbf{s} = \mathrm{Sketch}(\mathbf{X}) \in \mathbb{R}^m$$



$m \approx \#$ edges

Sketch

Output: graph of **relations between the d variables** $\boldsymbol{\Theta} \in \mathbb{R}^{d \times d}$



GLASSO

	In memory	In time
$\widehat{\boldsymbol{\Sigma}}$ ⟶	$\mathcal{O}(d^2)$	$\mathcal{O}(d^3)$

Sketching

	In memory	In time
$\mathbf{s}$ ⟶	$\mathcal{O}(m) \ll d^2$	$\mathcal{O}(?)$

Why should it work ?

- The underlying graph is sparse
- Keep only what we need through the sketch

Towards theoretical compressive recovery

- **The feature operator** $\Phi : \mathbb{R}^d \to \mathbb{R}^m$

Towards theoretical compressive recovery

- **The feature operator** $\Phi : \mathbb{R}^d \to \mathbb{R}^m$
  - In this talk: quadratic measurements $\mathbf{A}_j \sim \Lambda$ is a random matrix

$$\boxed{\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\mathbf{x}^\top \mathbf{A}_1 \mathbf{x}, \cdots, \mathbf{x}^\top \mathbf{A}_m \mathbf{x})^\top}$$

Towards theoretical compressive recovery

- **The feature operator** $\Phi : \mathbb{R}^d \to \mathbb{R}^m$
  - In this talk: quadratic measurements $\mathbf{A}_j \sim \Lambda$ is a random matrix

$$\boxed{\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\mathbf{x}^\top \mathbf{A}_1 \mathbf{x}, \cdots, \mathbf{x}^\top \mathbf{A}_m \mathbf{x})^\top}$$

Gaussian measurements

$$\mathbf{A}_j \underset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_{d \times d})$$

Towards theoretical compressive recovery

- **The feature operator** $\Phi : \mathbb{R}^d \to \mathbb{R}^m$
  - In this talk: quadratic measurements $\mathbf{A}_j \sim \Lambda$ is a random matrix

$$\boxed{\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\mathbf{x}^\top \mathbf{A}_1 \mathbf{x}, \cdots, \mathbf{x}^\top \mathbf{A}_m \mathbf{x})^\top}$$

Gaussian measurements

$$\mathbf{A}_j \underset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_{d\times d})$$

Structured rank-one

End of presentation

Rank-one measurements

$$\mathbf{A}_j = \mathbf{a}_j \mathbf{a}_j^\top \quad \mathbf{a}_j \underset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_d)$$

$$\Phi(\mathbf{x}) = \left(|\langle \mathbf{a}_j, \mathbf{x}\rangle|^2\right)_{j\in[\![m]\!]}$$

  - Inspired by works on low-rank matrix completion

Towards theoretical compressive recovery

- **The feature operator** $\Phi : \mathbb{R}^d \to \mathbb{R}^m$
  - In this talk: quadratic measurements $\mathbf{A}_j \sim \Lambda$ is a random matrix

$$\boxed{\Phi(\mathbf{x}) = \frac{1}{\sqrt{m}}(\mathbf{x}^\top \mathbf{A}_1 \mathbf{x}, \cdots, \mathbf{x}^\top \mathbf{A}_m \mathbf{x})^\top}$$

Gaussian measurements

$$\mathbf{A}_j \underset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_{d\times d})$$

Structured rank-one

End of presentation

Rank-one measurements

$$\mathbf{A}_j = \mathbf{a}_j \mathbf{a}_j^\top \quad \mathbf{a}_j \underset{i.i.d}{\sim} \mathcal{N}(0, \mathbf{I}_d)$$

$$\Phi(\mathbf{x}) = \left(|\langle \mathbf{a}_j, \mathbf{x}\rangle|^2\right)_{j\in[\![m]\!]}$$

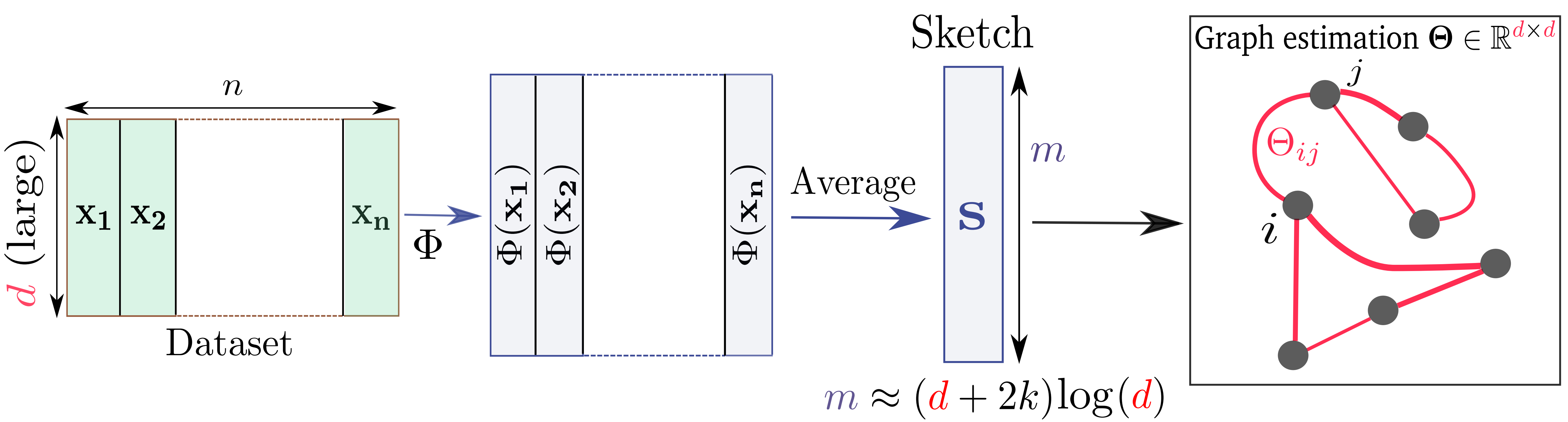
- Inspired by works on low-rank matrix completion
- Defined a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$

$$\mathcal{A}\mathbf{S} = \frac{1}{\sqrt{m}}\left(\langle \mathbf{A}_j, \mathbf{S}\rangle_F\right)_{j\in[\![m]\!]}$$

Emp. cov.

$$\boxed{\mathbf{s} = \frac{1}{n}\sum_{i=1}^{n} \Phi(\mathbf{x}_i) = \mathcal{A}\widehat{\mathbf{\Sigma}} \in \mathbb{R}^m}$$

Summary



n
d (large)
$\mathbf{x_1}$
$\mathbf{x_2}$
$\mathbf{x_n}$
Dataset
Φ
$\Phi(\mathbf{x_1})$
$\Phi(\mathbf{x_2})$
$\Phi(\mathbf{x_n})$
Average
Sketch
$\mathbf{s}$
m
$m \approx (d + 2k)\log(d)$
Graph estimation $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$
j
Θ_{ij}
i