# Towards Compressive Recovery of Sparse Precision Matrices

**Titouan Vayer** joint work with Rémi Gribonval Paulo Gonçalves

**INRIA Lyon - Ockham Team**

titouan.vayer@inria.fr

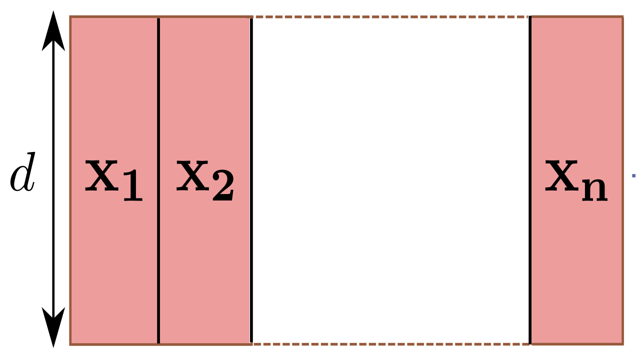
# Overview of the talk

- **Part I: Finding graphs from unstructured data**
- **Part II: The sketching approach**
- **Part III: Algorithmic solution**
- **Part IV: Limits, Open questions, partial answers**

# Graph Learning

- Input: a dataset

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

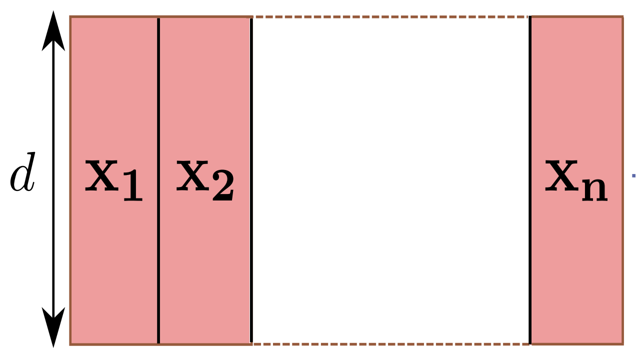



$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

# Graph Learning

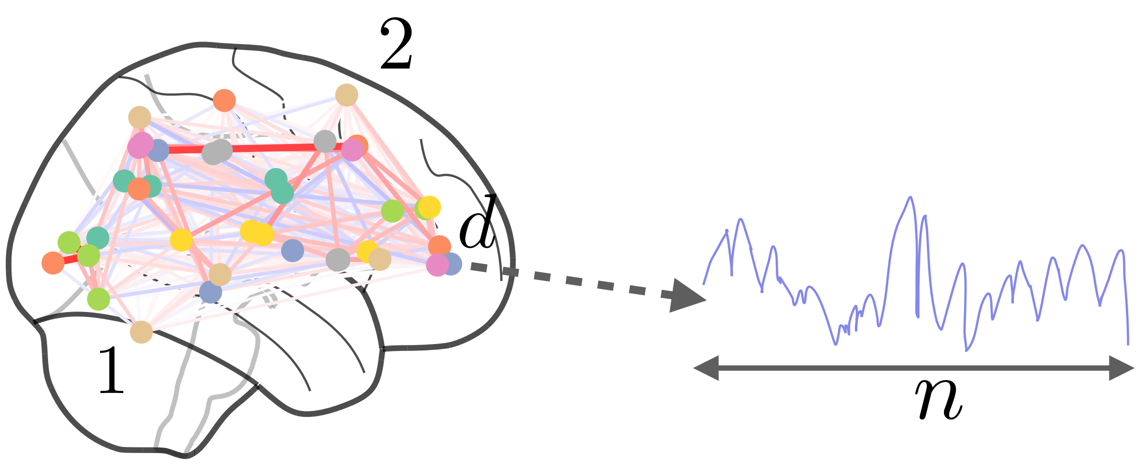
Input: a dataset

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




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

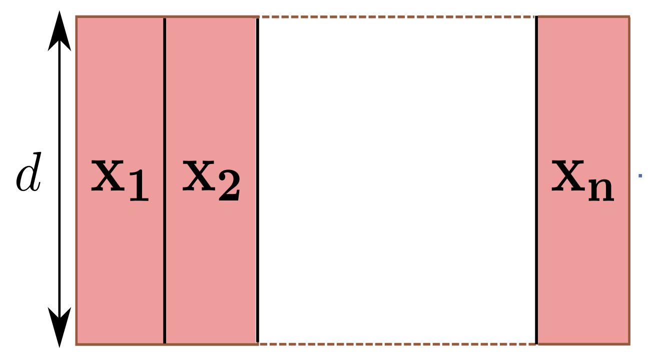
Output: graph of **relations between the d variables**



# Graph Learning

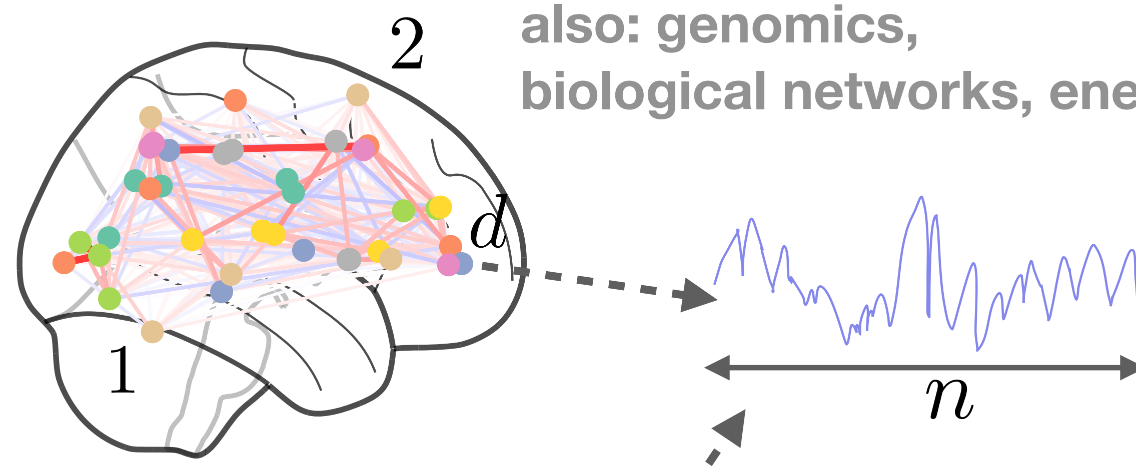
Input: a dataset

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

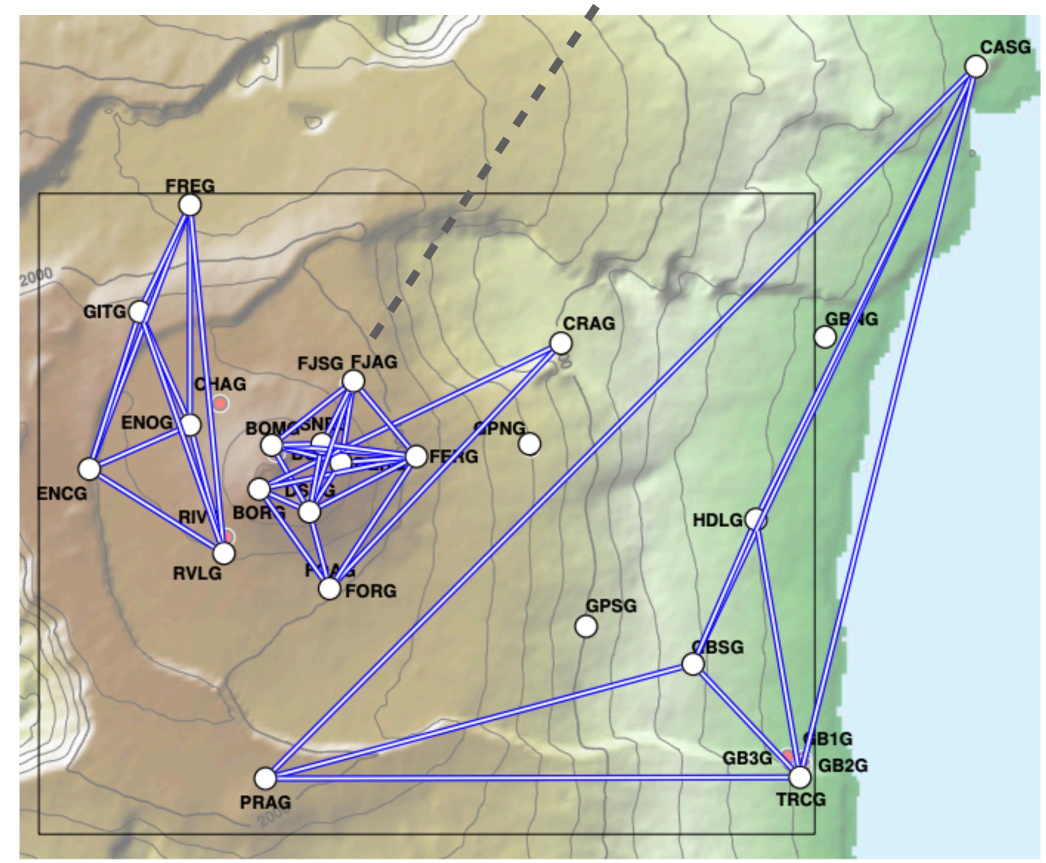



$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

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



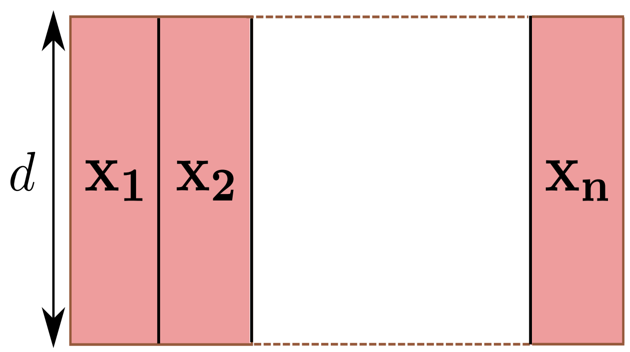

also: genomics, biological networks, energy…



# Graph Learning

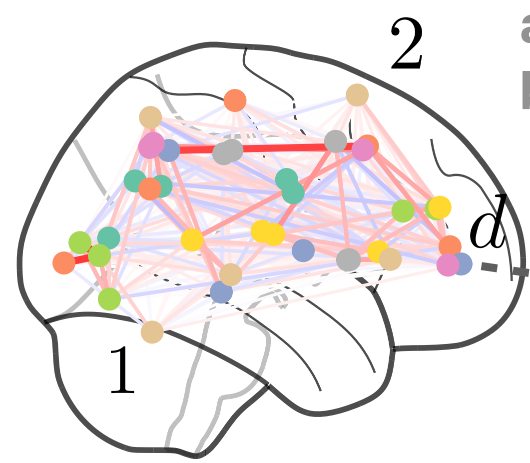
Input: a dataset

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

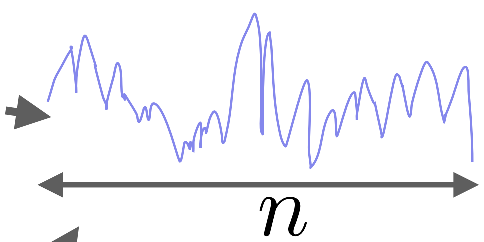


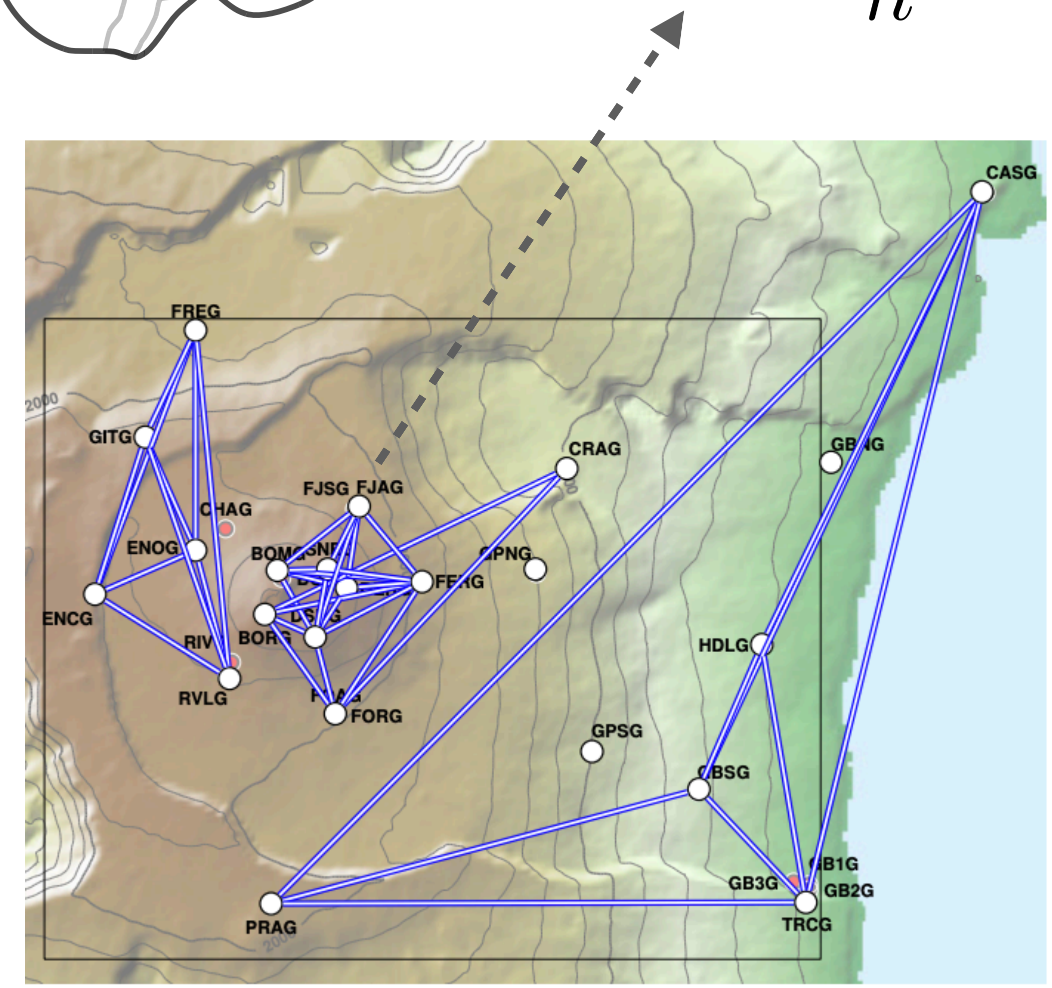

$$\mathbf{x}_i \in \mathbb{R}^d \sim \mu$$

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

# Graphical LASSO

Side note

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

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

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

# 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

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

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

$$\boldsymbol{\Theta}_{\text{MLE}} = \arg\min_{\boldsymbol{\Theta} \succ 0} - \text{logdet}(\boldsymbol{\Theta}) + \langle \widehat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle_F$$

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

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

## 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

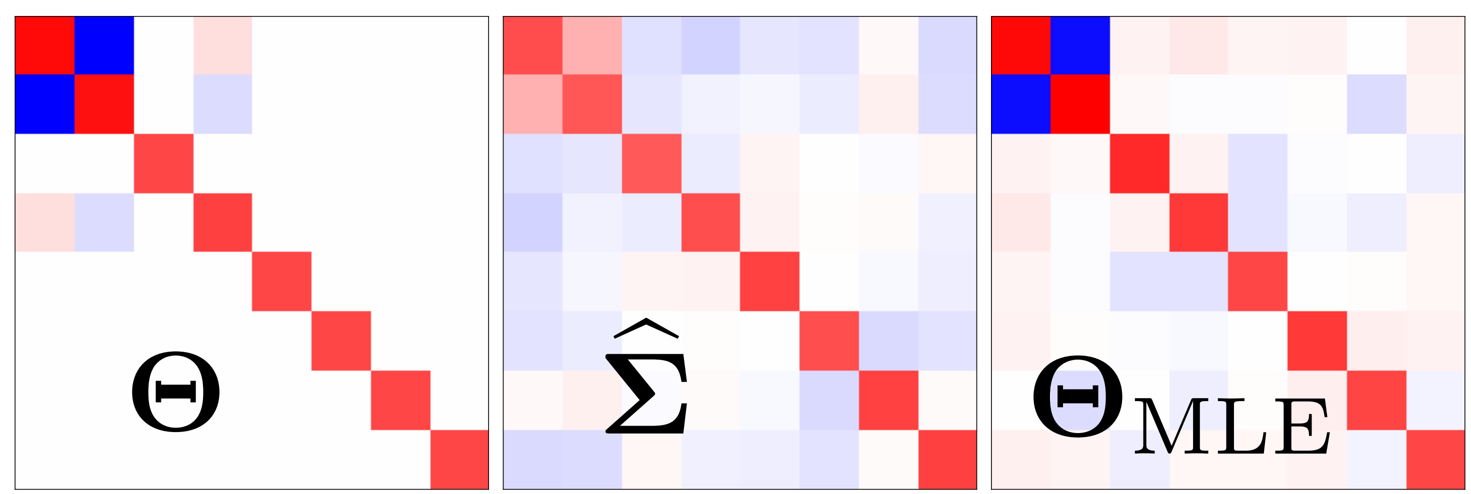
## Maximum Likehood estimator

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

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

- When $\widehat{\boldsymbol{\Sigma}}$ is invertible $\boldsymbol{\Theta}_{\mathrm{MLE}} = (\widehat{\boldsymbol{\Sigma}})^{-1}$ → 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: $\mathbf{\Theta} \in \mathbb{R}^{d \times d}$

## Gaussian Graphical Model

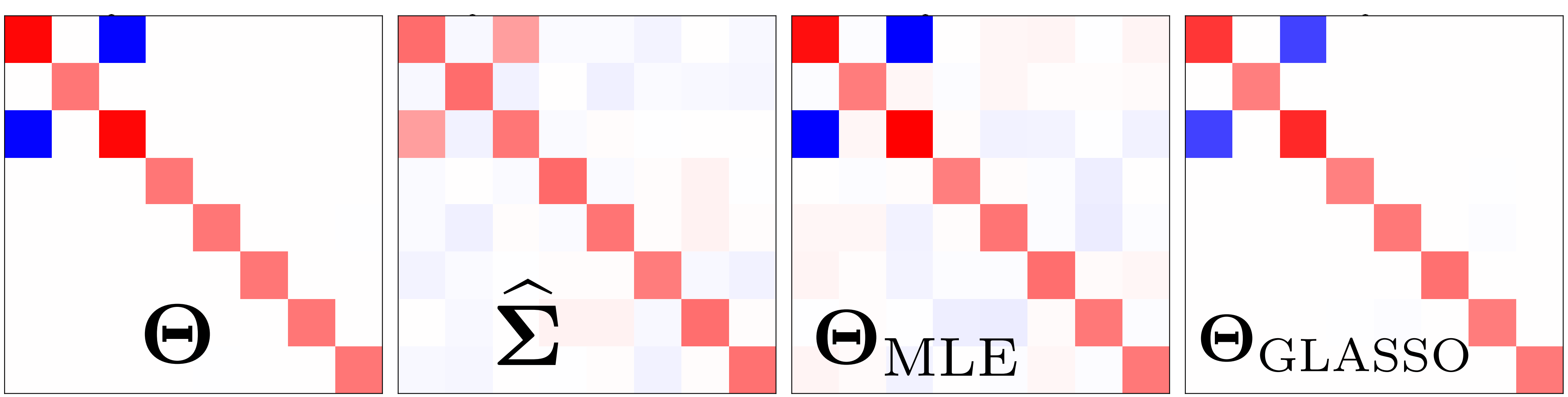
Gaussian assumption $\mu = \mathcal{N}(0, \mathbf{\Sigma} = \mathbf{\Theta}^{-1})$

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

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

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

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



# 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

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

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

## 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

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

- Optimization: convex problem

Coordinate descent

Involves LASSO steps (on the rows)

- Many 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]

# Graphical LASSO

Side note

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

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

- 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:

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

- Complexity of GLASSO:

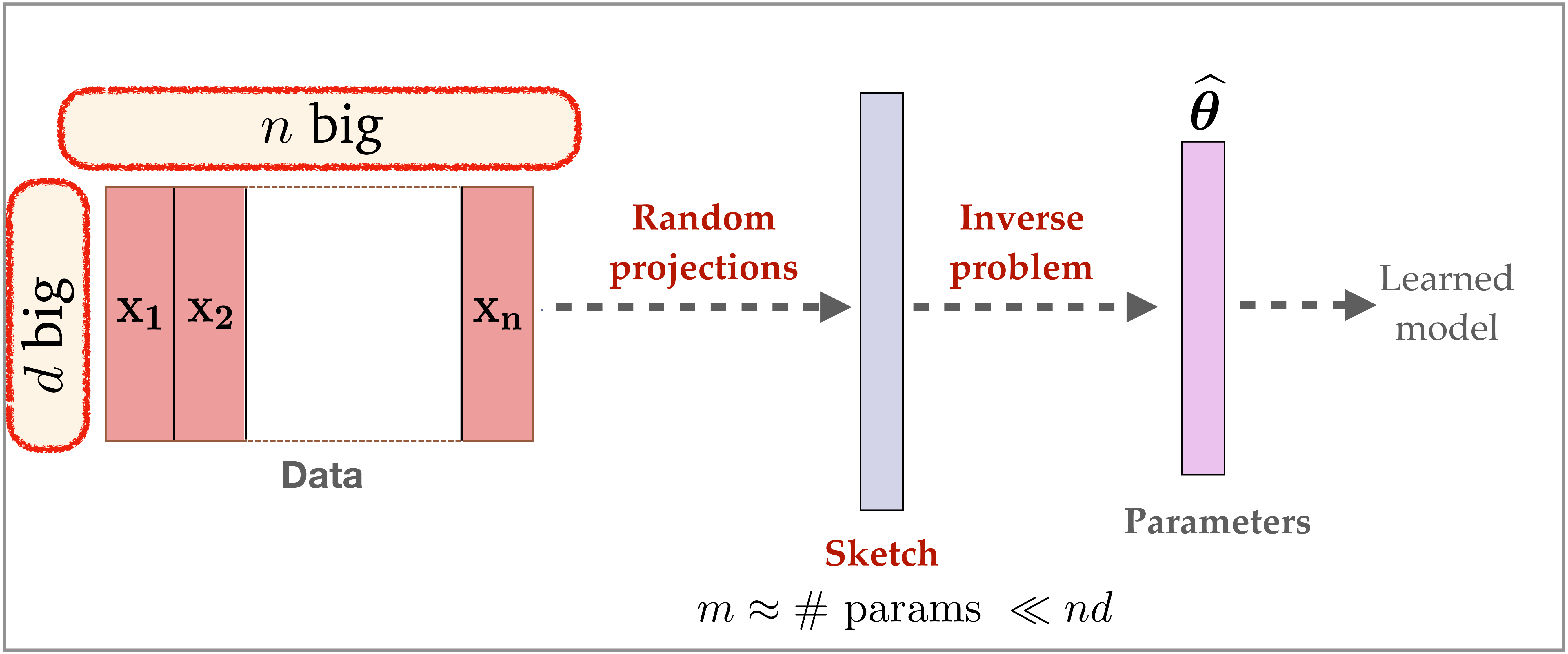
| | In memory | In time |
|---|---|---|
| $\widehat{\mathbf{\Sigma}}$ → | $\mathcal{O}(d^2)$ | $\mathcal{O}(d^3)$ |

# Overview of the talk

- Part I: Finding graphs from unstructured data
- Part II: The sketching approach
- Part III: Algorithmic solution
- Part IV: Limits, Open questions, partial answers

# The sketching approach

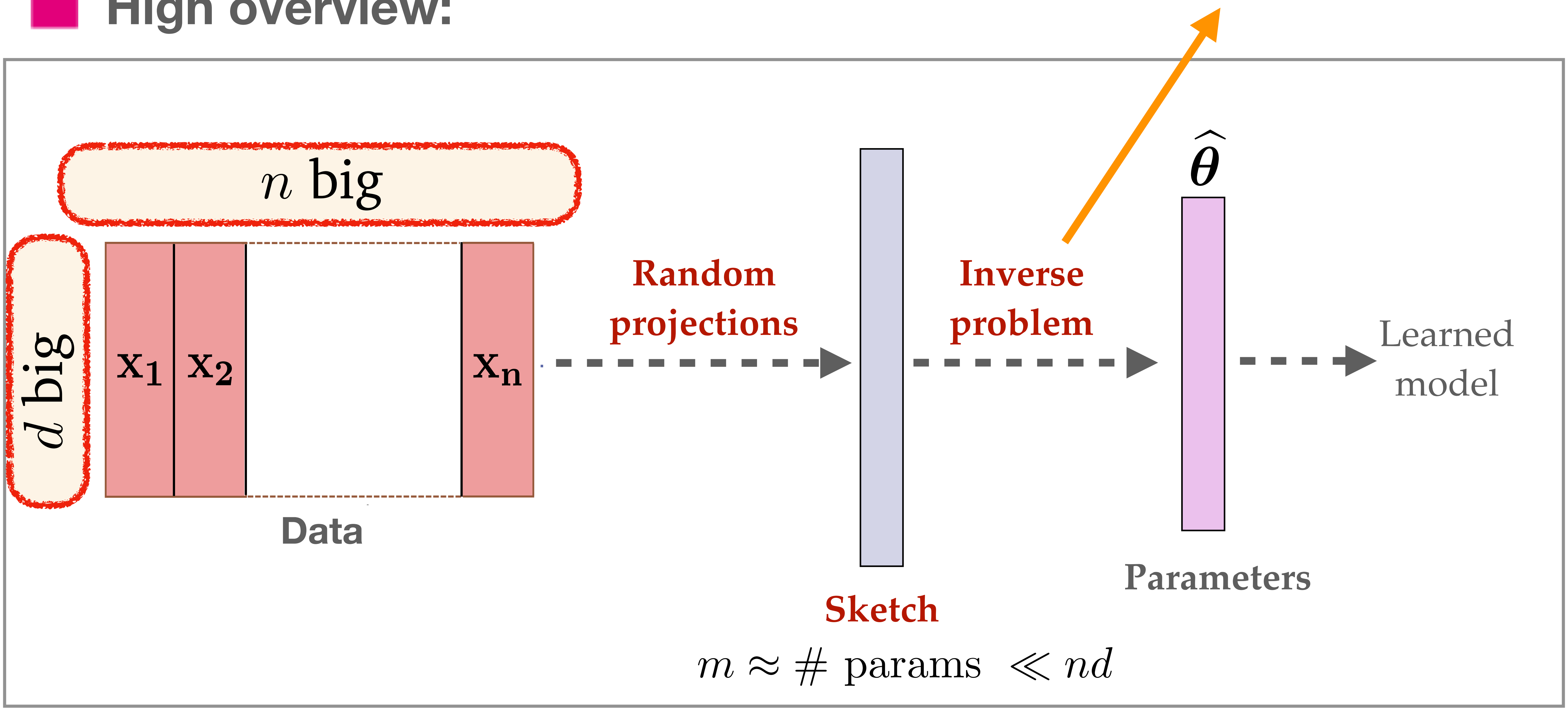
## High overview:



# The sketching approach

## High overview:

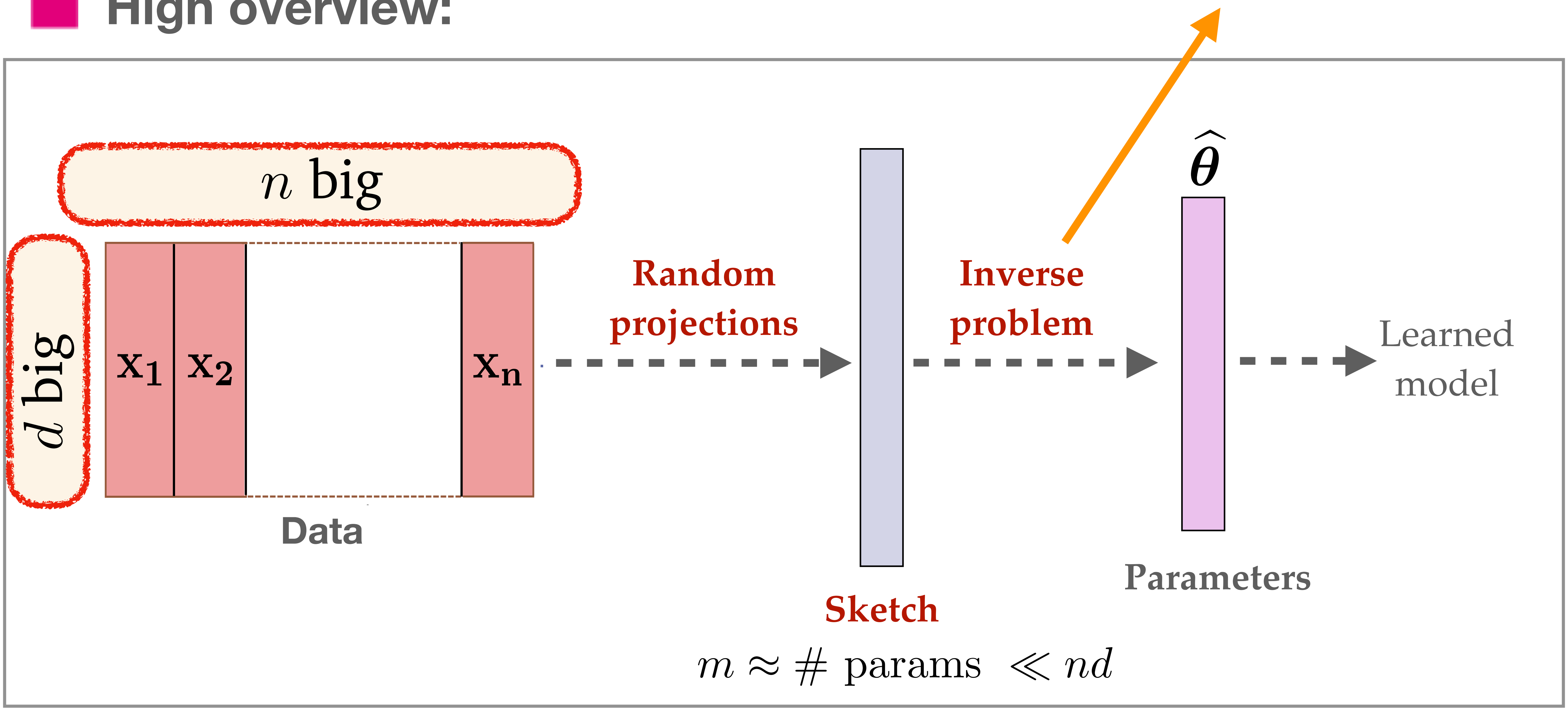
generalizes the principles of compressed sensing



# The sketching approach

generalizes the principles of compressed sensing

## High overview:



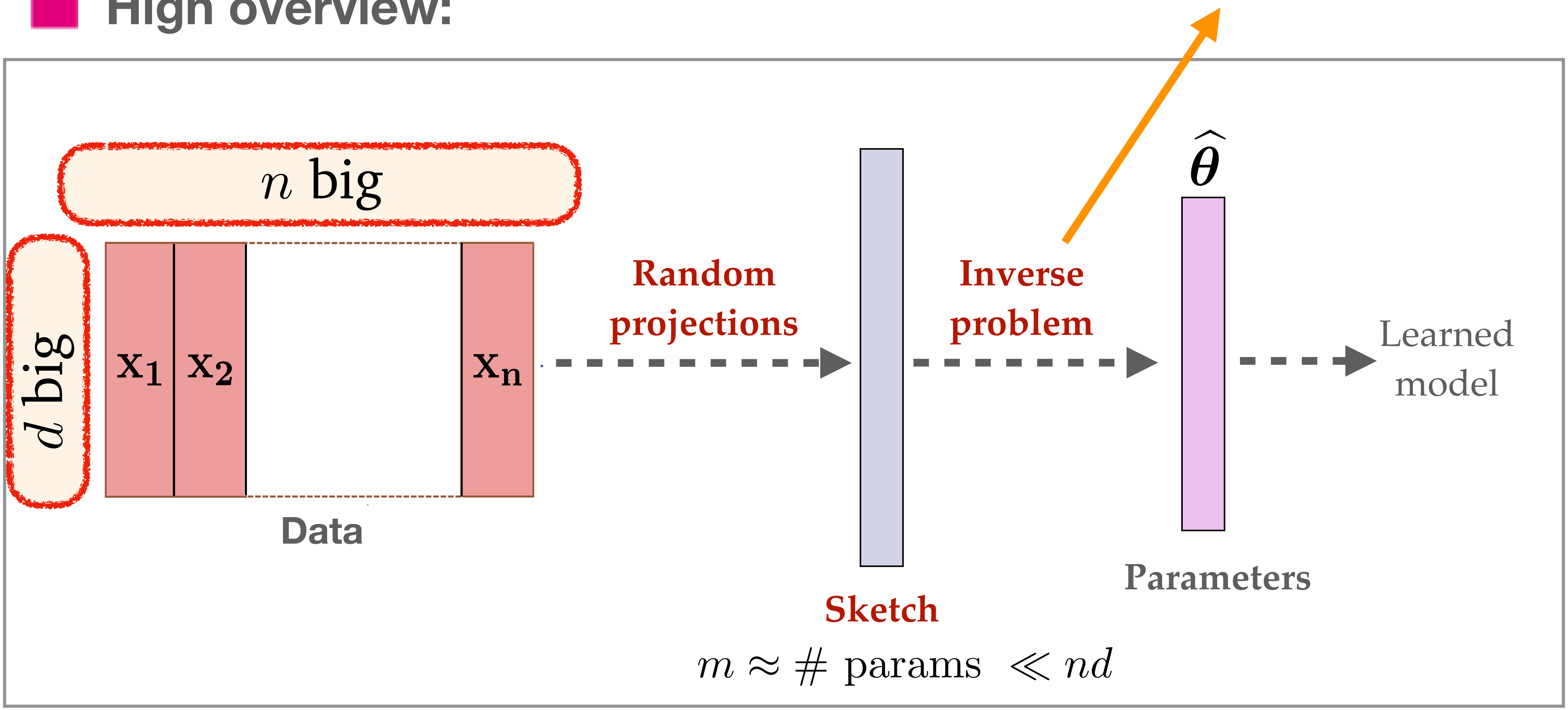

- Advantages for storage and transfer
- Many statistical problems

How to choose $\Phi$ ? -> connection to optimal transport

# The sketching approach

## High overview:

generalizes the principles of compressed sensing



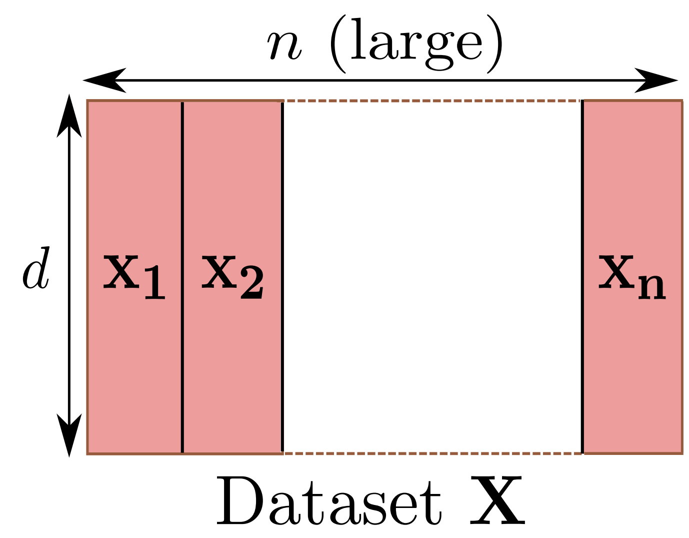

- Advantages for storage and transfer
- Many statistical problems

How to choose $\Phi$ ? -> connection to **optimal transport**

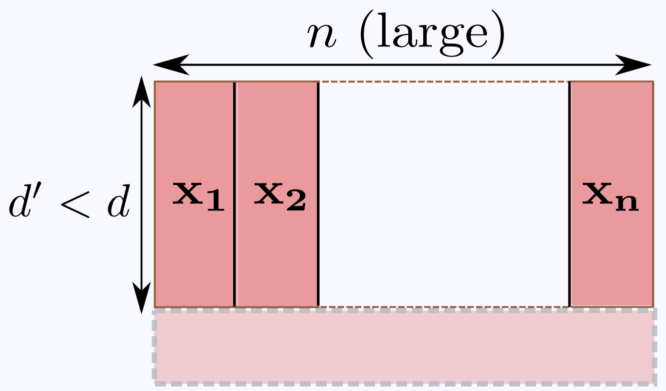
**Rémi Gribonval,** Anthony Bourrier, **Nicolas Keriven**, **Antoine Chatalic**, Gilles Puy, Nicolas Tremblay, Yann Traonmilin, Clément Elvira, Patrick Perez, Mike Davies, Gilles Blanchard, Laurent Jacques, **Vincent Schellekens**, Florimond Houssiau, Phil Schniter, Evan Byrne, ...

# The sketching approach

- « Dimension » reduction
  - « Low-dim » representation of a dataset

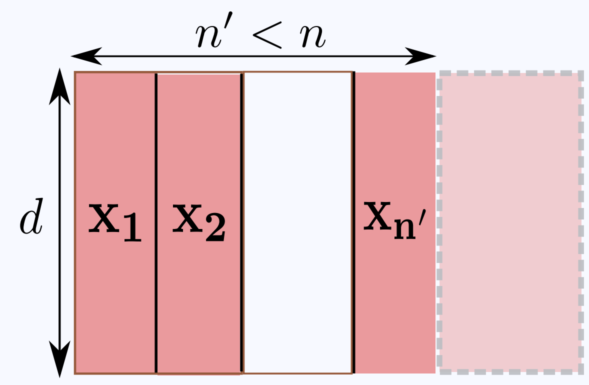



### Dimension reduction



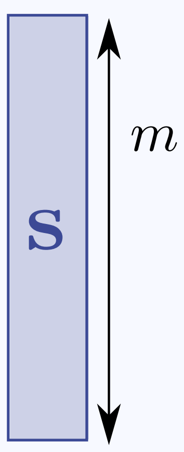

- Random projections (JL lemma)
- Feature selection
- Minimum distortion embedding, PCA

### Subsampling




- Coresets
- Importance sampling

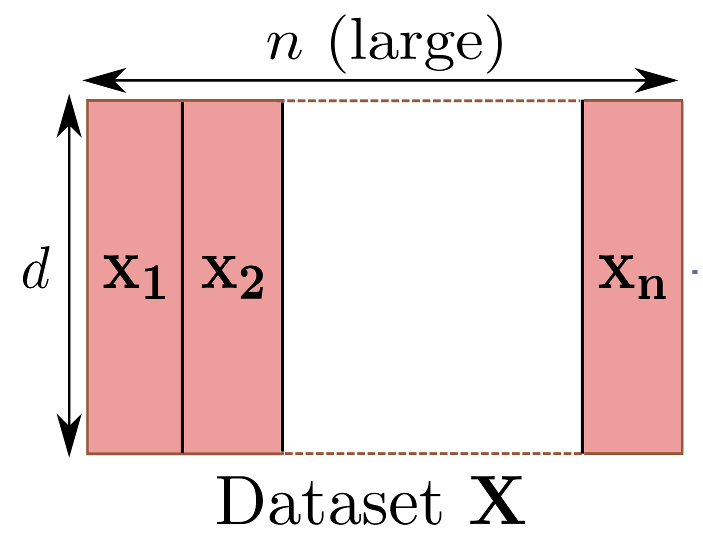
### Here: linear « sketch »




- Only one vector

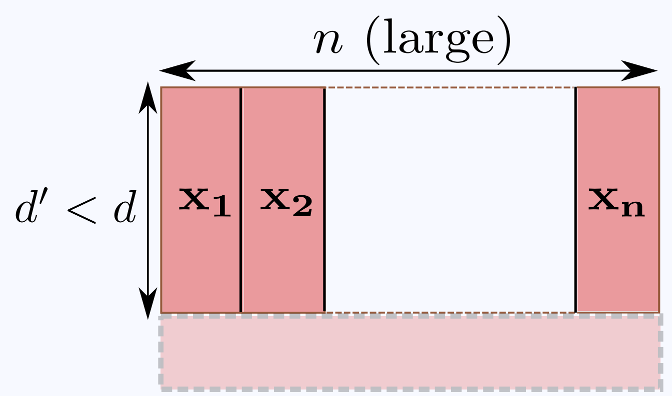
# The sketching approach

- « Dimension » reduction
  - « Low-dim » representation of a dataset



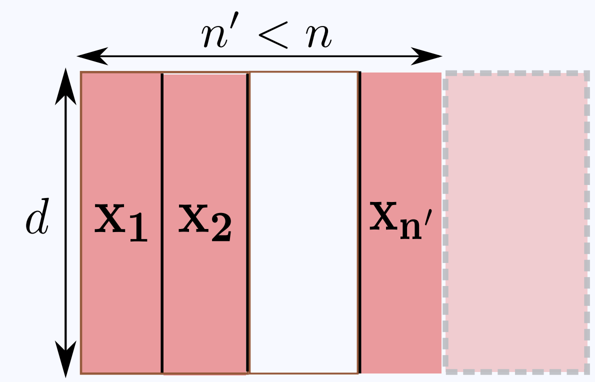

How do we sketch ? How do we learn from sketch ?

### Dimension reduction



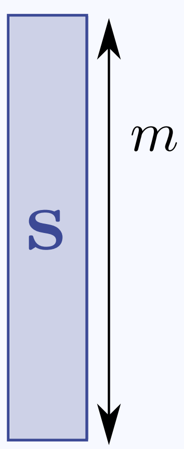

- Random projections (JL lemma)
- Feature selection
- Minimum distortion embedding, PCA

### Subsampling




- Coresets
- Importance sampling

### Here: linear « sketch »

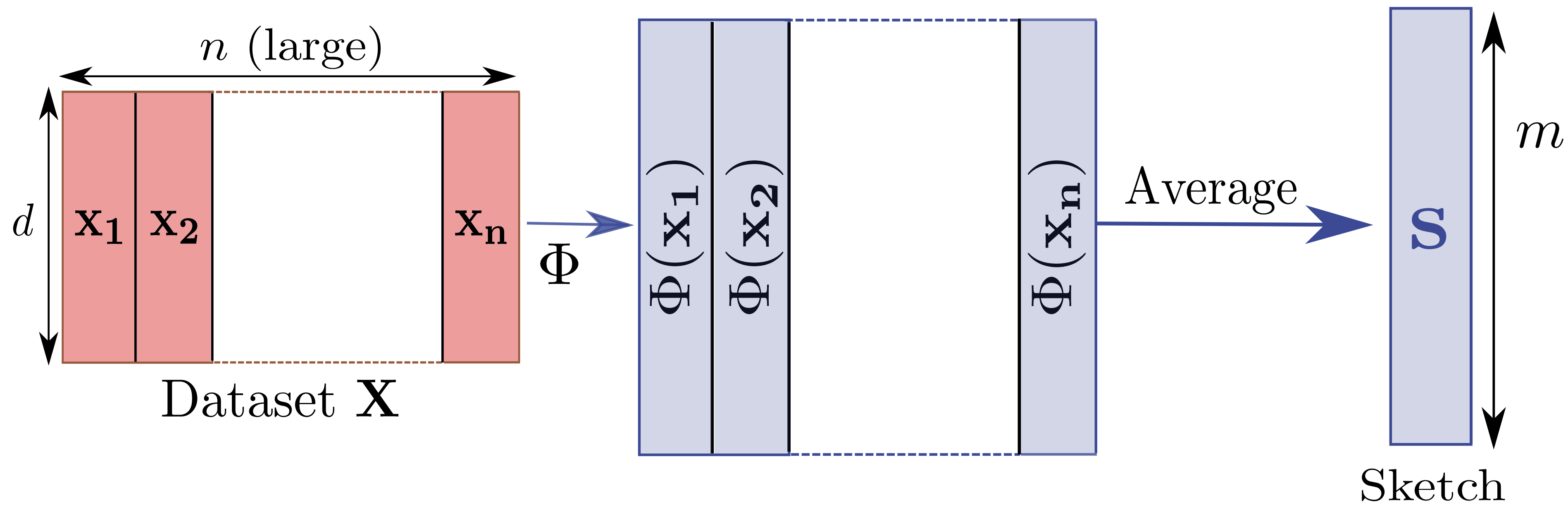



- Only one vector

# The sketching approach

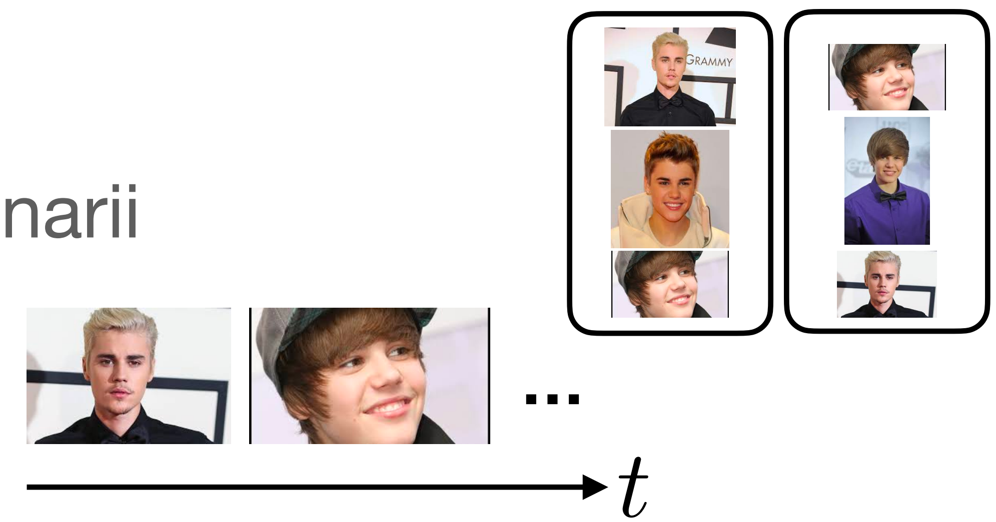
## Obtaining the sketch

- A function called feature operator $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^m$
- 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

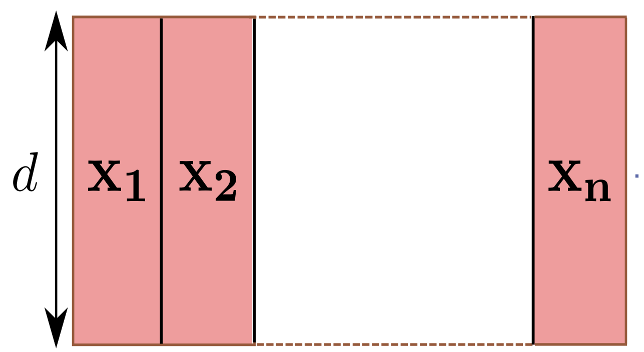


# Goal of this talk

- Input: a dataset

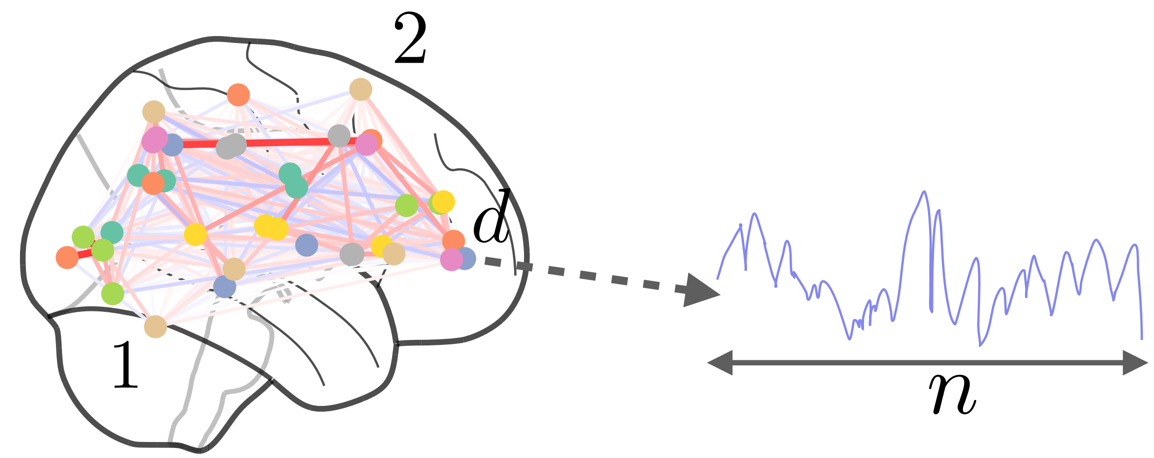
GLASSO

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




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

- 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

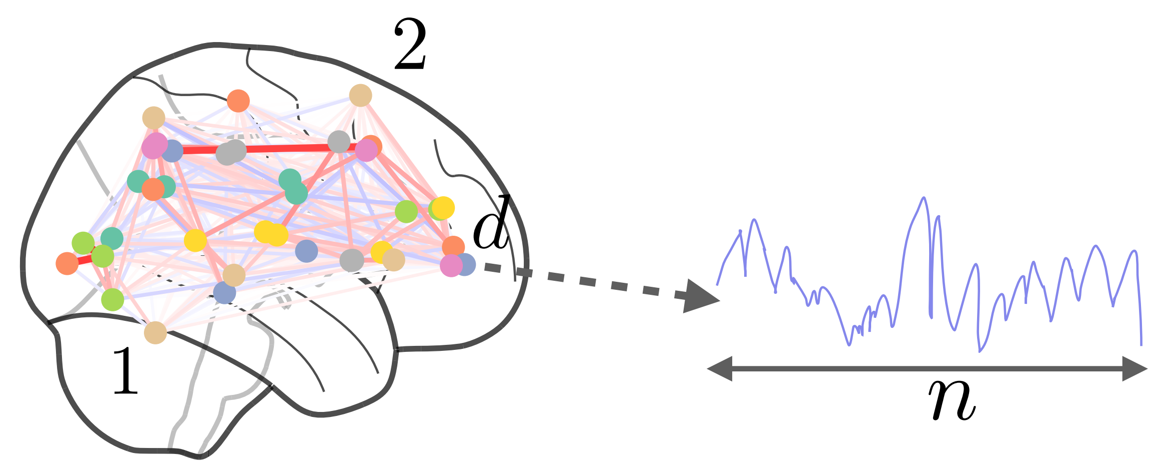
Input: a dataset

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

$m \approx \#$ edges

**Sketch**

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




GLASSO

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

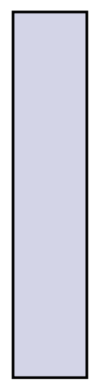
Sketching

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

# Goal of this talk

- Input: a dataset

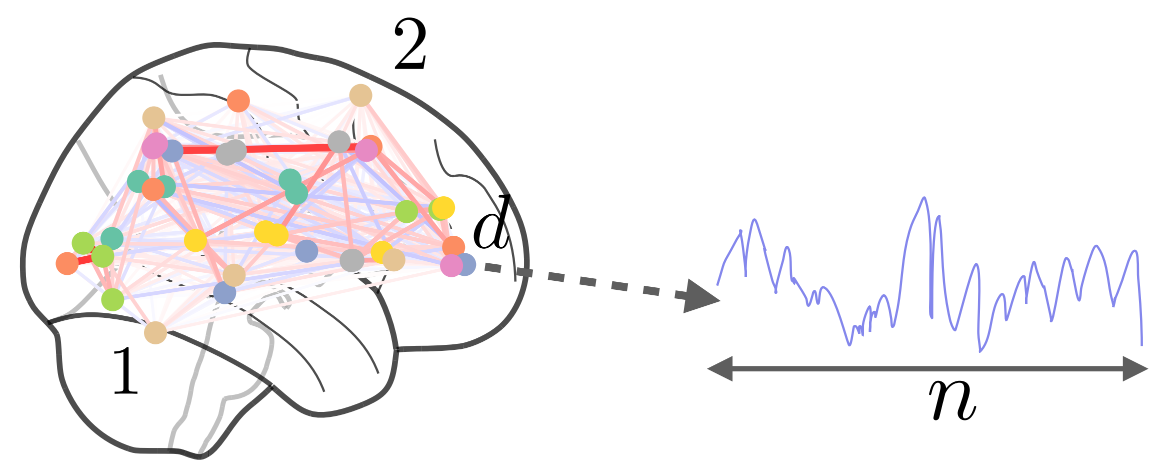
$$\mathbf{s} = \mathrm{Sketch}(\mathbf{X}) \in \mathbb{R}^m$$



$m \approx \#$ 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

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

# 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

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

**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

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

**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

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

**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.

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

# Towards theoretical compressive recovery

## Objective/setting

$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \;\; \mathbf{x}_i \in \mathbb{R}^d \sim \mu \quad \text{with} \quad \mathbb{E}_{\mathbf{x}\sim\mu}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}_\star = \mathbf{\Theta}_\star^{-1}$

$\mathbf{\Theta}_\star \in \mathfrak{S}$ in some low-dim. space (e.g. sparse p.d matrices = sparse graph)

- The objective is to find $\mathbf{\Theta}_\star$ from the sketch $\mathbf{s} \in \mathbb{R}^m$

# Towards theoretical compressive recovery

## Objective/setting

$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ $\mathbf{x}_i \in \mathbb{R}^d \sim \mu$ with $\underset{\mathbf{x}\sim\mu}{\mathbb{E}}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}_\star = \mathbf{\Theta}_\star^{-1}$

$\mathbf{\Theta}_\star \in \mathfrak{S}$ in some low-dim. space (e.g. sparse p.d matrices = sparse graph)

- The objective is to find $\mathbf{\Theta}_\star$ from the sketch $\mathbf{s} \in \mathbb{R}^m$
- Can be framed as a compressed sensing problem

$$\text{Find } \mathbf{\Theta}_\star \text{ from } \mathbf{s} = \mathcal{A}\widehat{\mathbf{\Sigma}} = \mathcal{A}\mathbf{\Sigma}_\star + \mathbf{e}$$

# Towards theoretical compressive recovery

## Objective/setting

$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \;\; \mathbf{x}_i \in \mathbb{R}^d \sim \mu$ with $\underset{\mathbf{x}\sim\mu}{\mathbb{E}}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}_\star = \mathbf{\Theta}_\star^{-1}$

$\mathbf{\Theta}_\star \in \mathfrak{S}$ in some low-dim. space (e.g. sparse p.d matrices = sparse graph)

- The objective is to find $\mathbf{\Theta}_\star$ from the sketch $\mathbf{s} \in \mathbb{R}^m$
- Can be framed as a compressed sensing problem

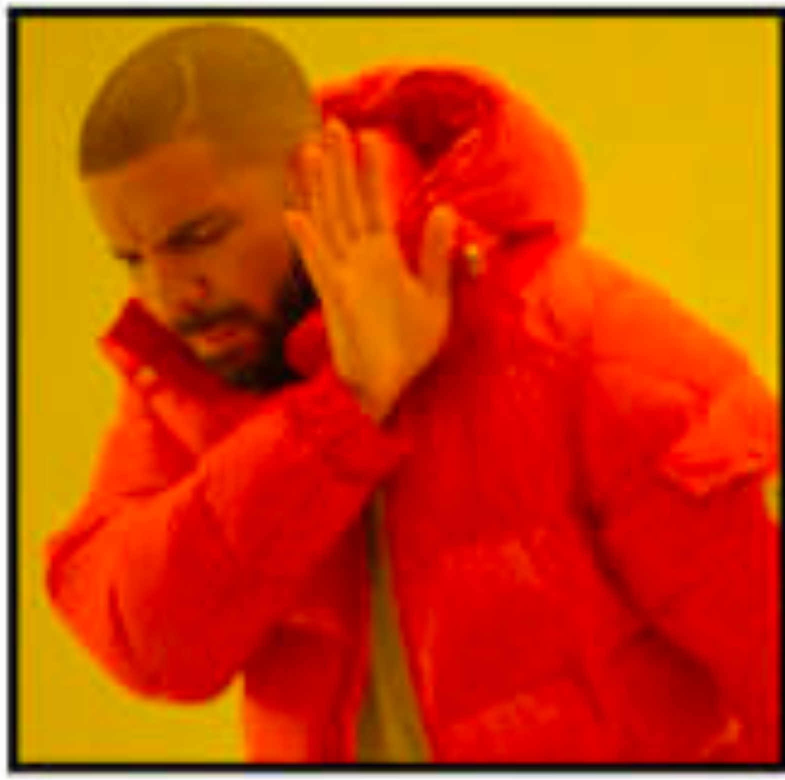
$$\text{Find } \mathbf{\Theta}_\star \text{ from } \mathbf{s} = \mathcal{A}\widehat{\mathbf{\Sigma}} = \mathcal{A}\mathbf{\Sigma}_\star + \mathbf{e}$$

- Notable difference

We want **the inverse of the matrix** that is measured

**not** find $\mathbf{\Sigma}_\star$ given $\mathcal{A}\mathbf{\Sigma}_\star + \mathbf{e}$

# Towards theoretical compressive recovery



find $\boldsymbol{\Sigma}_\star$ given $\mathcal{A}\boldsymbol{\Sigma}_\star + \mathbf{e}$

find $\boldsymbol{\Sigma}_\star^{-1}$ given $\mathcal{A}\boldsymbol{\Sigma}_\star + \mathbf{e}$

# Towards theoretical compressive recovery

## Objective/setting

$\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ $\mathbf{x}_i \in \mathbb{R}^d \sim \mu$ with $\mathbb{E}_{\mathbf{x}\sim\mu}[\mathbf{x}\mathbf{x}^\top] = \mathbf{\Sigma}_\star = \mathbf{\Theta}_\star^{-1}$

$\mathbf{\Theta}_\star \in \mathfrak{S}$ in some low-dim. space (e.g. sparse p.d matrices = sparse graph)

- The objective is to find $\mathbf{\Theta}_\star$ from the sketch $\mathbf{s} \in \mathbb{R}^m$
- Can be framed as a compressed sensing problem

$$\text{Find } \mathbf{\Theta}_\star \text{ from } \mathbf{s} = \mathcal{A}\widehat{\mathbf{\Sigma}} = \mathcal{A}\mathbf{\Sigma}_\star + \mathbf{e}$$

- Notable difference

We want **the inverse of the matrix** that is measured

**not** find $\mathbf{\Sigma}_\star$ given $\mathcal{A}\mathbf{\Sigma}_\star + \mathbf{e}$

- Ill-posed problem: $m \ll d^2$

Key assumption:

$\mathbf{\Theta}_\star$ lies in some low-dim. space

# Towards theoretical compressive recovery

## Example of low-dim space

- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

$$\mathfrak{S}_{k,a,b} = \{\mathbf{\Theta} \succ 0\ ;\ \|\mathbf{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\mathbf{\Theta}) \subseteq [a,b]\}$$

sym. positive definite matrices

localized spectrum

with at most 2k non-zero coeff
outside the diagonal

# Towards theoretical compressive recovery

## Example of low-dim space

- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

$$\mathfrak{S}_{k,a,b} = \{\boldsymbol{\Theta} \succ 0 \; ; \|\boldsymbol{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\boldsymbol{\Theta}) \subseteq [a, b]\}$$

sym. positive definite matrices

localized spectrum

with at most 2k non-zero coeff
outside the diagonal

- Sparse precision (= sparse graph) matrices with known spectrum
- Called the model set

# The Restricted Isometric Property

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

$$\exists \delta \in [0, 1[, \forall \mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \underline{\mathfrak{S}}$$

$$(1-\delta)\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|^2 \leq \|\mathcal{A}(\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1})\|_2^2 \leq (1+\delta)\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|^2$$

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

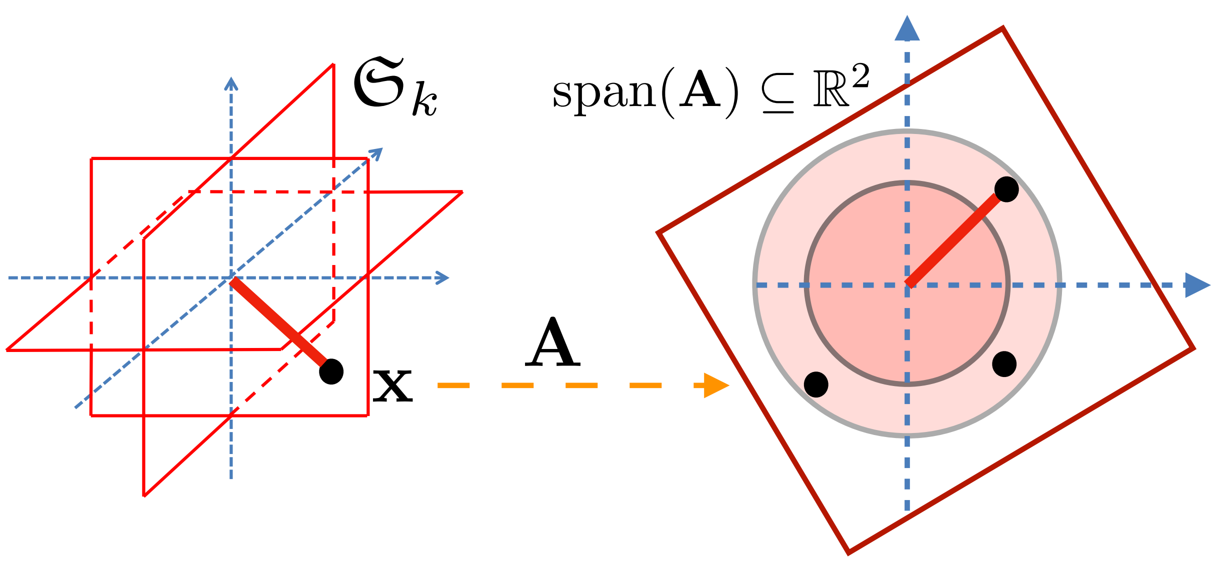
$$\exists \delta \in [0,1[, \forall \mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \underline{\mathfrak{S}}$$

$$(1-\delta)\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|^2 \leq \|\mathcal{A}(\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1})\|_2^2 \leq (1+\delta)\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|^2$$

## The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\exists \delta_k \in [0,1[ \quad \forall \mathbf{x} \ k\text{-sparse}$$

$$(1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2$$



# The Restricted Isometric Property

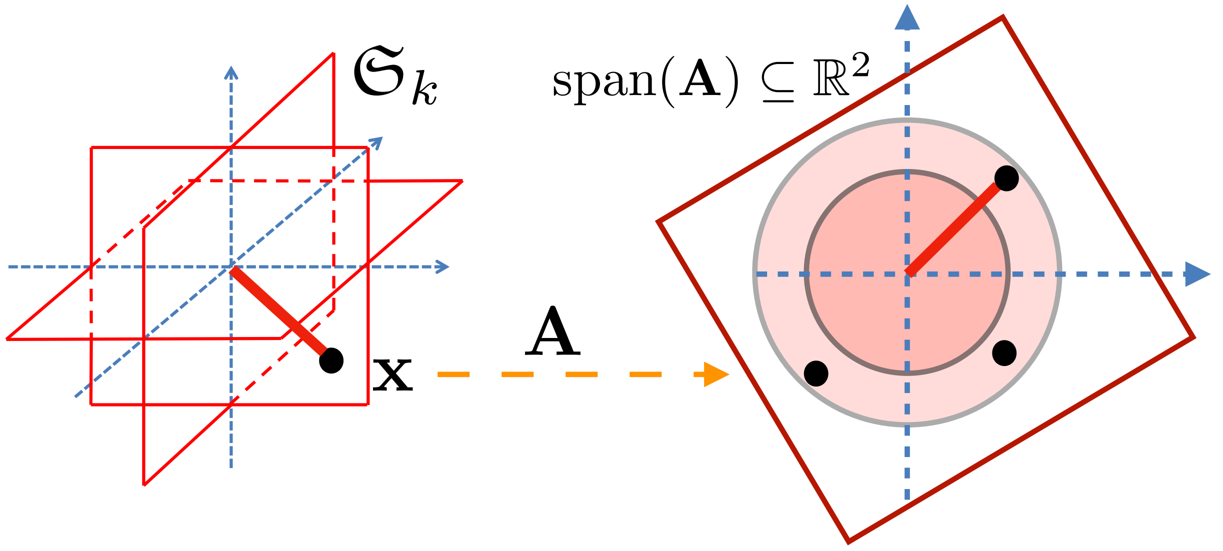
- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

$$\exists \delta \in [0, 1[, \forall \mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \underline{\mathfrak{S}}$$
$$(1-\delta)\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|^2 \leq \|\mathcal{A}(\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1})\|_2^2 \leq (1+\delta)\|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|^2$$

## The restricted isometric property (RIP) [Candes & Tao, 2005]

$$\exists \delta_k \in [0, 1[ \quad \forall \mathbf{x} \ k\text{-sparse}$$
$$(1-\delta_k)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\|_2^2 \leq (1+\delta_k)\|\mathbf{x}\|_2^2$$




- E.g. **Gaussian matrices**

$$m \gtrsim \delta^{-2} \, k \ln(e\tfrac{d}{k})$$

… also Johnson-Lindenstrauss Lemma

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

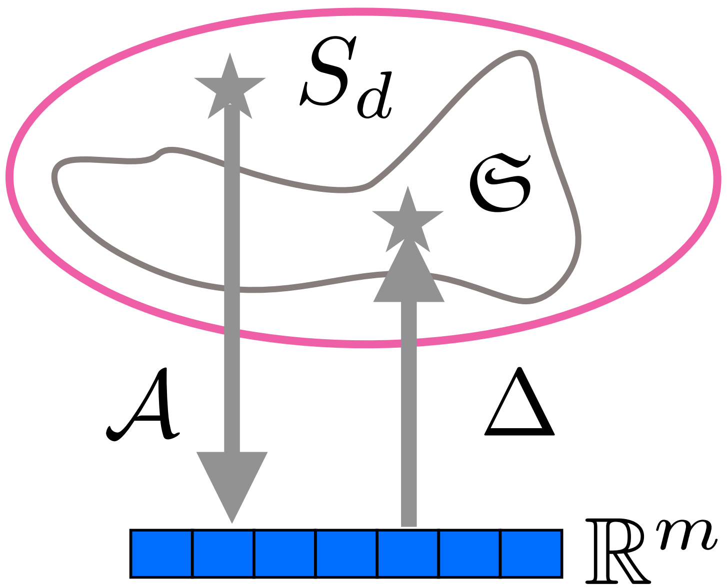
# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

## The optimal decoder

- If $\mathcal{A} : S_d \to \mathbb{R}^m$ satisfies the **invRIP** on $\mathfrak{S}$
- There exists a **optimal decoder** $\Delta : \mathbb{R}^m \to \mathfrak{S}$

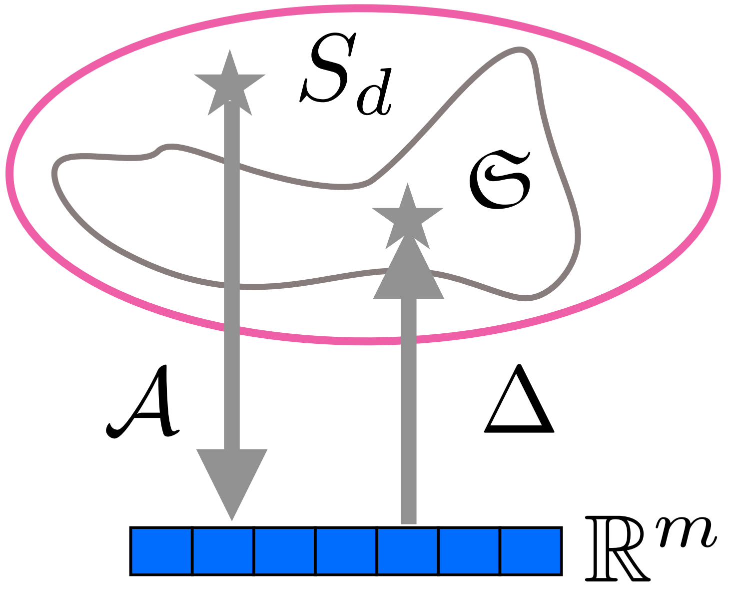


# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

## The optimal decoder




- If $\mathcal{A} : S_d \to \mathbb{R}^m$ satisfies the **invRIP** on $\mathfrak{S}$
- There exists a **optimal decoder** $\Delta : \mathbb{R}^m \to \mathfrak{S}$

sketch

$\forall \boldsymbol{\Theta}_\star \in \mathfrak{S}$ , the estimator $\widehat{\boldsymbol{\Theta}} = \Delta[\mathbf{s}]$ **satisfies**

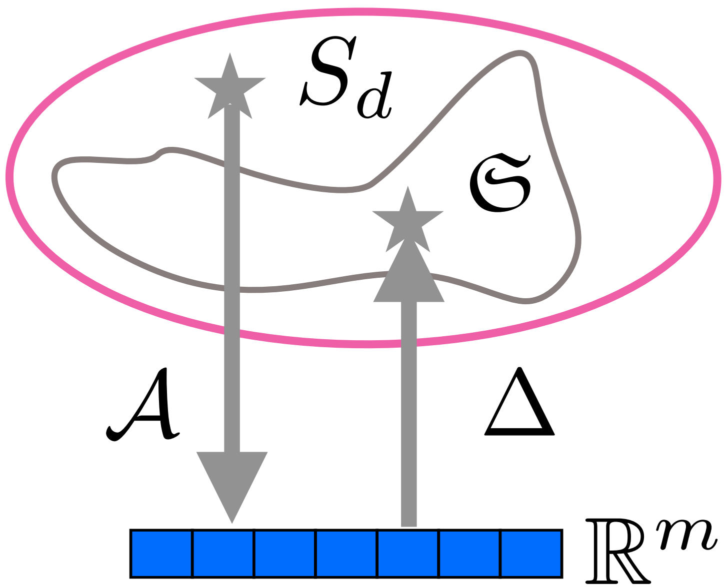
$$\|\boldsymbol{\Theta}_\star - \widehat{\boldsymbol{\Theta}}\| \leq \frac{2}{1-\delta} \|\mathcal{A}(\widehat{\boldsymbol{\Sigma}}) - \mathcal{A}(\boldsymbol{\Sigma}_\star)\|_2$$

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

## The optimal decoder




- If $\mathcal{A} : S_d \to \mathbb{R}^m$ satisfies the **invRIP** on $\mathfrak{S}$
- There exists a **optimal decoder** $\Delta : \mathbb{R}^m \to \mathfrak{S}$

$\forall \mathbf{\Theta}_\star \in \mathfrak{S}$ , the estimator $\widehat{\mathbf{\Theta}} = \Delta[\mathbf{s}]$ **satisfies**

$$\|\mathbf{\Theta}_\star - \widehat{\mathbf{\Theta}}\| \leq \tfrac{2}{1-\delta}\|\mathcal{A}(\widehat{\mathbf{\Sigma}}) - \mathcal{A}(\mathbf{\Sigma}_\star)\|_2 = \mathcal{O}(n^{-1/2})$$

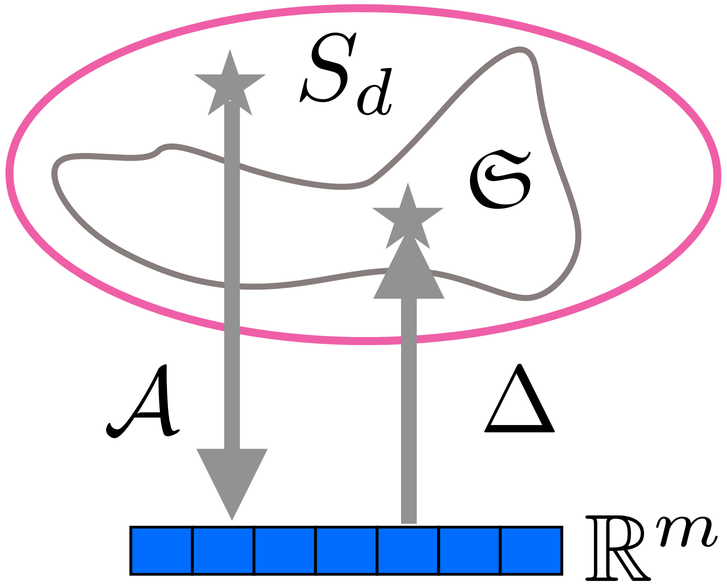
- + robust to noise !

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

## The optimal decoder




- If $\mathcal{A} : S_d \to \mathbb{R}^m$ satisfies the **invRIP** on $\mathfrak{S}$
- There exists a **optimal decoder** $\Delta : \mathbb{R}^m \to \mathfrak{S}$

$\forall \mathbf{\Theta}_\star \in \mathfrak{S}$ , the estimator $\widehat{\mathbf{\Theta}} = \Delta[\mathbf{s}]$ **satisfies**

$$\|\mathbf{\Theta}_\star - \widehat{\mathbf{\Theta}}\| \leq \frac{2}{1-\delta} \|\mathcal{A}(\widehat{\mathbf{\Sigma}}) - \mathcal{A}(\mathbf{\Sigma}_\star)\|_2 = \mathcal{O}(n^{-1/2})$$

- + robust to noise !

$$\Delta[\mathbf{s}] \in \arg\min_{\mathbf{\Theta} \in \mathfrak{S}} \|\mathcal{A}(\mathbf{\Theta}^{-1}) - \mathbf{s}\|_2$$

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

↓

**Optimal decoder**

↓

$$\Delta[\mathbf{s}] \in \arg\min_{\mathbf{\Theta} \in \mathfrak{S}} \|\mathcal{A}(\mathbf{\Theta}^{-1}) - \mathbf{s}\|_2$$

**Estimator from the sketch with $\mathcal{O}(n^{-1/2})$ error**

# The Restricted Isometric Property

- Consider a **linear op.** on **symmetric matrices** $\mathcal{A} : S_d \to \mathbb{R}^m$
- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

**invRIP**

↓

**Optimal decoder**

↓

$$\Delta[\mathbf{s}] \in \arg\min_{\boldsymbol{\Theta} \in \mathfrak{S}} \|\mathcal{A}(\boldsymbol{\Theta}^{-1}) - \mathbf{s}\|_2$$

**Estimator from the sketch with $\mathcal{O}(n^{-1/2})$ error**

↓

**Happiness and joy**

# Proving the Restricted Isometric Property

$\mathcal{A}\mathbf{S} = \frac{1}{\sqrt{m}} \left( \langle \mathbf{A}_j, \mathbf{S} \rangle_F \right)_{j \in [\![m]\!]}$ **is random** - - - ▶ **invRIP** with high prob.

# Proving the Restricted Isometric Property

$\mathcal{A}\mathbf{S} = \frac{1}{\sqrt{m}} \left( \langle \mathbf{A}_j, \mathbf{S} \rangle_F \right)_{j \in [\![m]\!]}$ **is random** - - - ▶ **invRIP** with high prob.

- **Two ingredients**
  - The pointwise concentration of $\mathcal{A}$ i.e. $\forall \mathbf{\Theta}_1, \mathbf{\Theta}_2 \in \underline{\mathfrak{S}}$

$$\mathbb{P}\left( \left| \|\mathcal{A}(\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1})\|_2^2 - \|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|_F^2 \right| > t \|\mathbf{\Theta}_1^{-1} - \mathbf{\Theta}_2^{-1}\|_F^2 \right) \leq C_1(t; m)$$

# Proving the Restricted Isometric Property

$\mathcal{A}\mathbf{S} = \frac{1}{\sqrt{m}} \left( \langle \mathbf{A}_j, \mathbf{S} \rangle_F \right)_{j \in [\![m]\!]}$ **is random** - - -▶ **invRIP** with high prob.

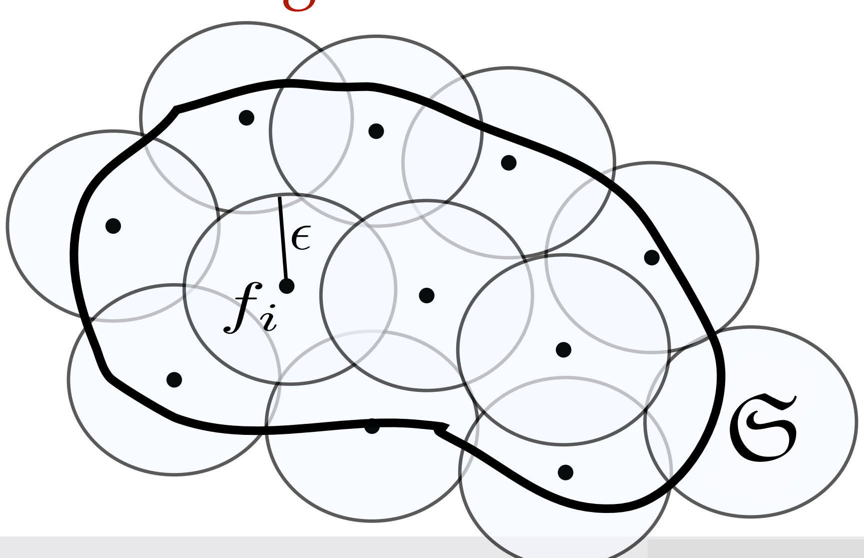
## Two ingredients

- The pointwise concentration of $\mathcal{A}$ i.e. $\forall \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \underline{\mathfrak{S}}$

$$\mathbb{P}\left( \left| \|\mathcal{A}(\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1})\|_2^2 - \|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_F^2 \right| > t\|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_F^2 \right) \leq C_1(t; m)$$

- The set $\mathfrak{S}$ is sufficiently « low-dim », has « low complexity »

Covering numbers ◀- - - $\mathcal{N}(\mathfrak{S}, \varepsilon)$ is sufficiently small



# Proving the Restricted Isometric Property

$\mathcal{A}\mathbf{S} = \frac{1}{\sqrt{m}} \left(\langle \mathbf{A}_j , \mathbf{S} \rangle_F \right)_{j \in [\![m]\!]}$ **is random** - - - ▶ **invRIP** with high prob.

## Two ingredients

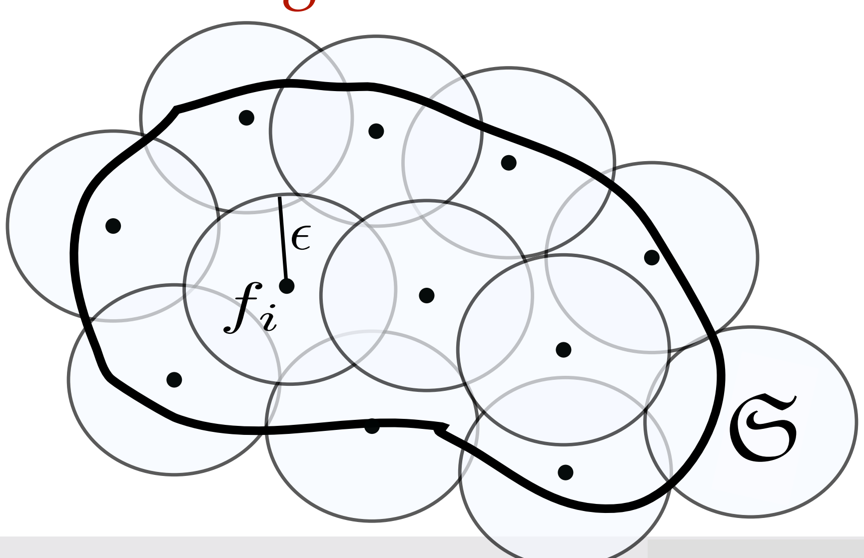
- The pointwise concentration of $\mathcal{A}$ i.e. $\forall \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \underline{\mathfrak{S}}$

$$\mathbb{P}\left( \left| \|\mathcal{A}(\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1})\|_2^2 - \|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_F^2 \right| > t \|\boldsymbol{\Theta}_1^{-1} - \boldsymbol{\Theta}_2^{-1}\|_F^2 \right) \leq C_1(t;m)$$

- The set $\mathfrak{S}$ is sufficiently « low-dim », has « low complexity »

$\mathcal{N}(\mathfrak{S}, \varepsilon)$ is sufficiently small

Covering numbers




e.g. k-sparse vectors in dimension d

$$\ln \mathcal{N}(\mathfrak{S}_k, \varepsilon) \leq k \left( \ln(3\varepsilon^{-1}) + \ln(e\tfrac{d}{k}) \right)$$

- In our case ok with:

$$\mathfrak{S}_{k,a,b} = \{ \boldsymbol{\Theta} \succ 0 \, ; \|\boldsymbol{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\boldsymbol{\Theta}) \subseteq [a,b] \}$$

# Main results

- Sketch with

**Gaussian measurements**

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

- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

$$\mathfrak{S}_{k,a,b} = \{\mathbf{\Theta} \succ 0 \,;\, \|\mathbf{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\mathbf{\Theta}) \subseteq [a, b]\}$$

# Main results

- Sketch with

**Gaussian measurements**

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

- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

$$\mathfrak{S}_{k,a,b} = \{\boldsymbol{\Theta} \succ 0 \,;\, \|\boldsymbol{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\boldsymbol{\Theta}) \subseteq [a,b]\}$$

$$m \gtrsim (d+2k)\ln(d)$$ → **invRIP** with high prob.

Hides $a, b, \delta, \rho$

# Main results

- Sketch with

**Gaussian measurements**

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

- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

$$\mathfrak{S}_{k,a,b} = \{\boldsymbol{\Theta} \succ 0 \; ; \|\boldsymbol{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\boldsymbol{\Theta}) \subseteq [a, b]\}$$

$$m \gtrsim (d+2k)\ln(d)$$

**invRIP** with high prob.

Hides $a, b, \delta, \rho$

**Optimal decoder**

$$\Delta[\mathbf{s}] \in \underset{\boldsymbol{\Theta}\in\mathfrak{S}}{\arg\min} \, \|\mathcal{A}(\boldsymbol{\Theta}^{-1}) - \mathbf{s}\|_2$$

**Happiness and joy**

# Main results

- Sketch with

**Gaussian measurements**

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

- $\mathfrak{S}$ a subspace of $S_d$ of idealized precision matrices (low-dim)

$$\mathfrak{S}_{k,a,b} = \{\boldsymbol{\Theta} \succ 0 \, ; \|\boldsymbol{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\boldsymbol{\Theta}) \subseteq [a,b]\}$$

$$m \gtrsim (d+2k)\ln(d)$$

→ invRIP with high prob.

Hides $a, b, \delta, \rho$

→ Optimal decoder → Happiness and joy

$$\Delta[\mathbf{s}] \in \underset{\boldsymbol{\Theta}\in\mathfrak{S}}{\arg\min} \, \|\mathcal{A}(\boldsymbol{\Theta}^{-1}) - \mathbf{s}\|_2$$

## Remarks

- The number of measurements low compared to $d^2$
- Almost optimal up to a logarithmic factor
- Now all we need is to compute the decoder …

# Overview of the talk

- Part I: Finding graphs from unstructured data
- Part II: The sketching approach
- Part III: Algorithmic solution
- Part IV: Limits, Open questions, partial answers

# Towards practical recovery

## Recover the precision matrix from the sketch

- We need to compute: $\Delta[\mathbf{s}] \in \arg\min_{\boldsymbol{\Theta} \in \mathfrak{S}} \|\mathcal{A}(\boldsymbol{\Theta}^{-1}) - \mathbf{s}\|_2$
- Where: $\mathfrak{S}_{k,a,b} = \{\boldsymbol{\Theta} \succ 0 \, ; \|\boldsymbol{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\boldsymbol{\Theta}) \subseteq [a,b]\}$
- Quite difficult because of $\ell_0$ constraints! (probably NP-hard)

# Towards practical recovery

## Recover the precision matrix from the sketch

- We need to compute: $\Delta[\mathbf{s}] \in \arg\min_{\mathbf{\Theta} \in \mathfrak{S}} \|\mathcal{A}(\mathbf{\Theta}^{-1}) - \mathbf{s}\|_2$
- Where: $\mathfrak{S}_{k,a,b} = \{\mathbf{\Theta} \succ 0 \; ; \|\mathbf{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\mathbf{\Theta}) \subseteq [a, b]\}$
- Quite difficult because of $\ell_0$ constraints! (probably NP-hard)
- Approximate decoder:

$$\widetilde{\mathbf{\Theta}} \in \arg\min_{\substack{\mathbf{\Theta} \in S_d(\mathbb{R}) \\ a\mathbf{I}_d \preceq \mathbf{\Theta} \preceq b\mathbf{I}_d}} \frac{1}{2}\|\mathcal{A}(\mathrm{inv}(\mathbf{\Theta})) - \mathbf{s}\|_2^2 + \lambda \|\mathbf{\Theta}\|_{1,\mathrm{off}}$$

# Towards practical recovery

## Recover the precision matrix from the sketch

- We need to compute: $\Delta[\mathbf{s}] \in \arg\min_{\boldsymbol{\Theta}\in\mathfrak{S}} \|\mathcal{A}(\boldsymbol{\Theta}^{-1}) - \mathbf{s}\|_2$
- Where: $\mathfrak{S}_{k,a,b} = \{\boldsymbol{\Theta} \succ 0\ ; \|\boldsymbol{\Theta}\|_0 \leq d + 2k, \mathrm{spec}(\boldsymbol{\Theta}) \subseteq [a,b]\}$
- Quite difficult because of $\ell_0$ constraints! (probably NP-hard)
- Approximate decoder:

$$\widetilde{\boldsymbol{\Theta}} \in \arg\min_{\substack{\boldsymbol{\Theta}\in S_d(\mathbb{R}) \\ a\mathbf{I}_d \preceq \boldsymbol{\Theta} \preceq b\mathbf{I}_d}} \frac{1}{2}\|\mathcal{A}(\mathrm{inv}(\boldsymbol{\Theta})) - \mathbf{s}\|_2^2 + \lambda\|\boldsymbol{\Theta}\|_{1,\mathrm{off}}$$

## Optimisation problem:

$$\arg\min_{\boldsymbol{\Theta}\in\mathbb{R}^{d\times d}} f(\boldsymbol{\Theta}) + \lambda g(\boldsymbol{\Theta}) + h(\boldsymbol{\Theta})$$

$$f(\boldsymbol{\Theta}) \triangleq \frac{1}{2}\|\mathcal{A}(\mathrm{inv}(\boldsymbol{\Theta})) - \mathbf{s}\|_2^2$$

$$g(\boldsymbol{\Theta}) \triangleq \|\boldsymbol{\Theta}\|_{1,\mathrm{off}}$$

$$h(\boldsymbol{\Theta}) \triangleq \imath\left(\{a\mathbf{I} \preceq \boldsymbol{\Theta} \preceq b\mathbf{I}\} \cap \{\boldsymbol{\Theta} \in S_d(\mathbb{R})\}\right)$$

# Towards practical recovery

## Optimisation problem:

$$\underset{\boldsymbol{\Theta}\in\mathbb{R}^{d\times d}}{\arg\min}\; f(\boldsymbol{\Theta}) + \lambda g(\boldsymbol{\Theta}) + h(\boldsymbol{\Theta})$$

$$f(\boldsymbol{\Theta}) \triangleq \frac{1}{2}\|\mathcal{A}(\mathrm{inv}(\boldsymbol{\Theta})) - \mathbf{s}\|_2^2$$

$$g(\boldsymbol{\Theta}) \triangleq \|\boldsymbol{\Theta}\|_{1,\mathrm{off}}$$

$$h(\boldsymbol{\Theta}) \triangleq \imath\left(\{a\mathbf{I} \preceq \boldsymbol{\Theta} \preceq b\mathbf{I}\} \cap \{\boldsymbol{\Theta} \in S_d(\mathbb{R})\}\right)$$

- Non-convex with non-smooth penalty + smooth fidelity term -> $\nabla f$

# Towards practical recovery

## Optimisation problem:

$$\arg\min_{\boldsymbol{\Theta}\in\mathbb{R}^{d\times d}} \; f(\boldsymbol{\Theta}) + \lambda g(\boldsymbol{\Theta}) + h(\boldsymbol{\Theta})$$

$$f(\boldsymbol{\Theta}) \triangleq \frac{1}{2}\|\mathcal{A}(\mathrm{inv}(\boldsymbol{\Theta})) - \mathbf{s}\|_2^2$$

$$g(\boldsymbol{\Theta}) \triangleq \|\boldsymbol{\Theta}\|_{1,\text{off}}$$

$$h(\boldsymbol{\Theta}) \triangleq \imath\left(\{a\mathbf{I} \preceq \boldsymbol{\Theta} \preceq b\mathbf{I}\} \cap \{\boldsymbol{\Theta} \in S_d(\mathbb{R})\}\right)$$

- Non-convex with non-smooth penalty + smooth fidelity term -> $\nabla f$
- We use Davis & Yin three-operator splitting scheme

$$\mathrm{prox}_{\lambda\varphi}(\boldsymbol{\Theta}) \triangleq \arg\min_{\mathbf{A}\in\mathbb{R}^{d\times d}} \varphi(\mathbf{A}) + \frac{1}{2\lambda}\|\mathbf{A} - \boldsymbol{\Theta}\|_F^2$$

[Davis & Yin, 2017]

**Algorithm 1** Three-operator splitting algorithm

1: Input: initial guess $\boldsymbol{\Theta}_{\text{init}}$ and step size $\gamma > 0$.
2: Initialize $\mathbf{Z}_0 = \mathrm{prox}_{\gamma h}(\boldsymbol{\Theta}_{\text{init}})$ and $\mathbf{U}_0 = 0$.
3: **for** $t \in 0, 1 \cdots,$ **do**
4: $\quad \boldsymbol{\Theta}_{t+1} = \mathrm{prox}_{\gamma g}(\mathbf{Z}_t - \lambda_t \mathbf{U}_t - \gamma\nabla f(\mathbf{Z}_t))$
5: $\quad \mathbf{Z}_{t+1} = \mathrm{prox}_{\gamma h}(\boldsymbol{\Theta}_{t+1} + \gamma\mathbf{U}_t)$
6: $\quad \mathbf{U}_{t+1} = \mathbf{U}_t + (\boldsymbol{\Theta}_{t+1} - \mathbf{Z}_{t+1})/\gamma$
7: **end for**

# Towards practical recovery

## Optimisation problem:

$$\underset{\boldsymbol{\Theta}\in\mathbb{R}^{d\times d}}{\arg\min}\; f(\boldsymbol{\Theta}) + \lambda g(\boldsymbol{\Theta}) + h(\boldsymbol{\Theta})$$

$$f(\boldsymbol{\Theta}) \triangleq \frac{1}{2}\|\mathcal{A}(\mathrm{inv}(\boldsymbol{\Theta})) - \mathbf{s}\|_2^2$$

$$g(\boldsymbol{\Theta}) \triangleq \|\boldsymbol{\Theta}\|_{1,\mathrm{off}}$$

$$h(\boldsymbol{\Theta}) \triangleq \imath\left(\{a\mathbf{I} \preceq \boldsymbol{\Theta} \preceq b\mathbf{I}\} \cap \{\boldsymbol{\Theta} \in S_d(\mathbb{R})\}\right)$$

- Non-convex with non-smooth penalty + smooth fidelity term -> $\nabla f$
- We use Davis & Yin three-operator splitting scheme

$$\mathrm{prox}_{\lambda\varphi}(\boldsymbol{\Theta}) \triangleq \underset{\mathbf{A}\in\mathbb{R}^{d\times d}}{\arg\min}\; \varphi(\mathbf{A}) + \frac{1}{2\lambda}\|\mathbf{A} - \boldsymbol{\Theta}\|_F^2$$

[Davis & Yin, 2017]

**Algorithm 1** Three-operator splitting algorithm

1: Input: initial guess $\boldsymbol{\Theta}_{\mathrm{init}}$ and step size $\gamma > 0$.
2: Initialize $\mathbf{Z}_0 = \mathrm{prox}_{\gamma h}(\boldsymbol{\Theta}_{\mathrm{init}})$ and $\mathbf{U}_0 = 0$.
3: **for** $t \in 0, 1 \cdots,$ **do**
4: $\quad \boldsymbol{\Theta}_{t+1} = \mathrm{prox}_{\gamma g}(\mathbf{Z}_t - \lambda_t \mathbf{U}_t - \gamma \nabla f(\mathbf{Z}_t))$
5: $\quad \mathbf{Z}_{t+1} = \mathrm{prox}_{\gamma h}(\boldsymbol{\Theta}_{t+1} + \gamma \mathbf{U}_t)$
6: $\quad \mathbf{U}_{t+1} = \mathbf{U}_t + (\boldsymbol{\Theta}_{t+1} - \mathbf{Z}_{t+1})/\gamma$
7: **end for**

- We know the prox

$$\mathrm{prox}_{\lambda g}(\boldsymbol{\Theta}) = \mathbf{D} + \mathbf{D}^\top + \mathrm{diag}(\boldsymbol{\Theta})$$

where $\begin{cases} D_{ij} = \max\{\Theta_{ij} - \lambda, 0\} & i < j \\ D_{ij} = 0 & i \geq j \end{cases}$

$$\mathrm{prox}_h(\boldsymbol{\Theta}) = \mathbf{U}\mathrm{diag}(\mathbf{v})\mathbf{U}^\top$$

where $v_i = \min\{\max\{\lambda_i(\boldsymbol{\Theta}), a\}, b\}$

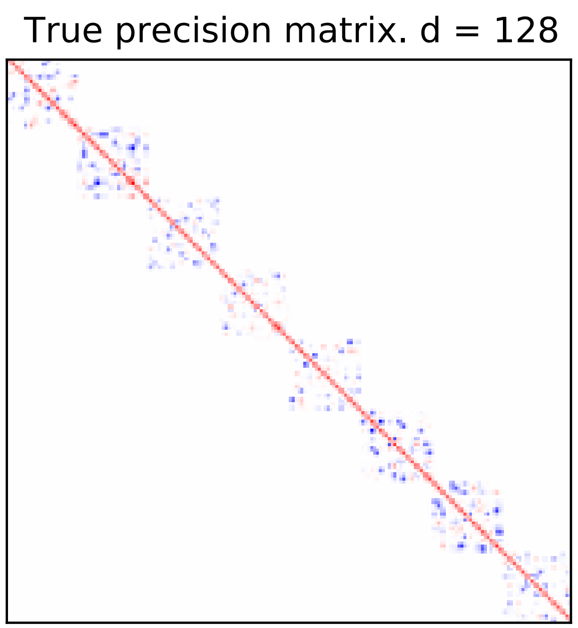
$\mathcal{O}(d^3)$ **complexity**

# Results

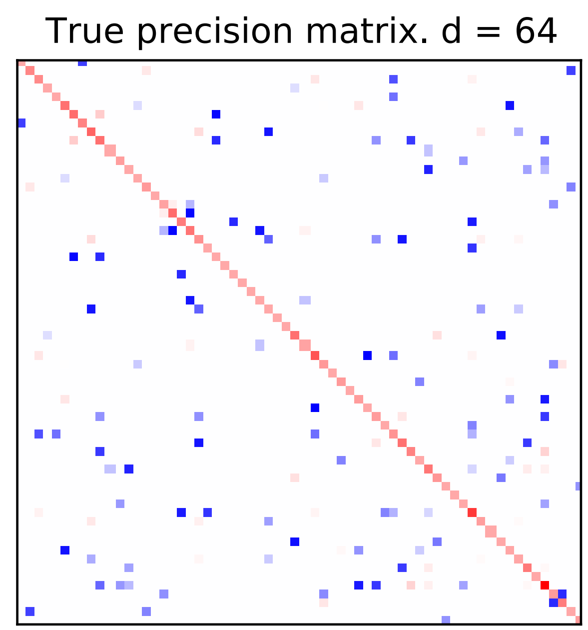
## Setting:

- We generate sparse $\boldsymbol{\Theta}_\star$ according to some laws

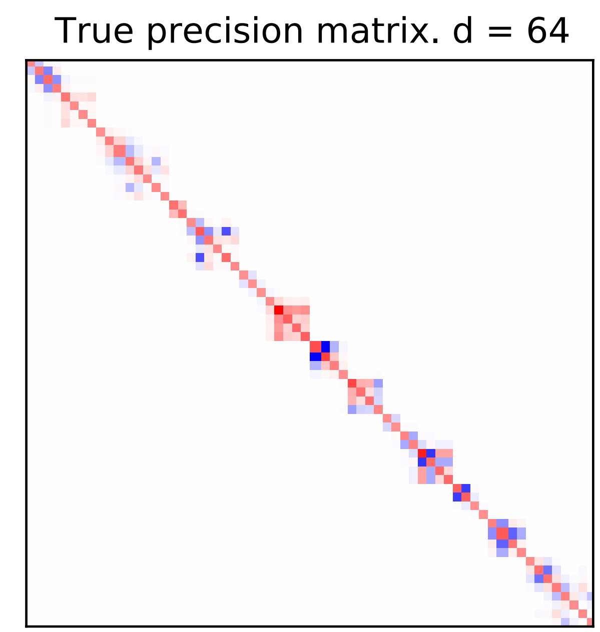
Erdös-Réyni




Sklearn




Powerlaw




- Data $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ $\mathbf{x}_i \sim \mathcal{N}(0, \boldsymbol{\Theta}_\star^{-1})$ and sketch
- Compare with the approximate decoder $\widetilde{\boldsymbol{\Theta}}$

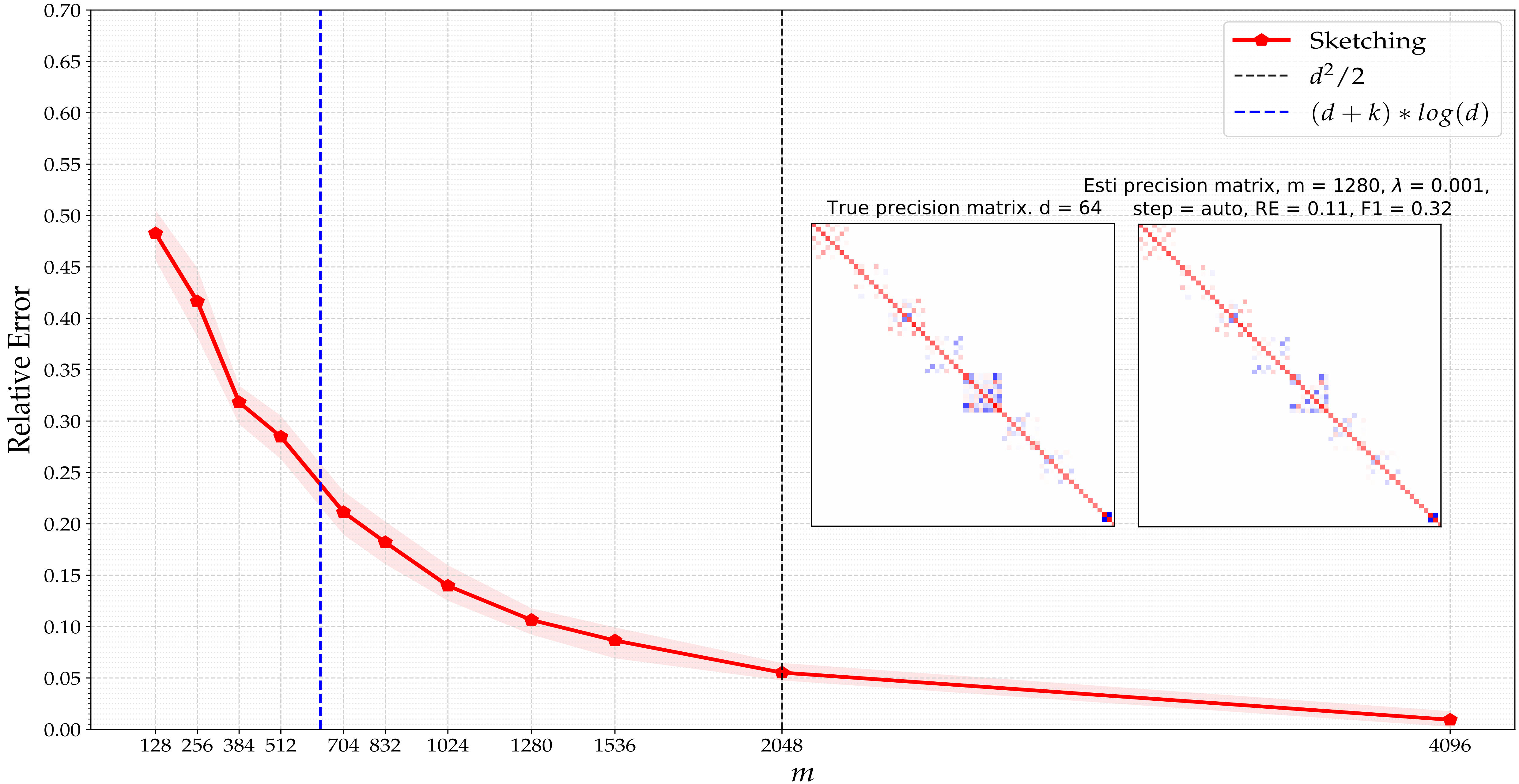
# Results

**Sketch of the true cov**

$$\mathbf{s} = \mathcal{A}\Sigma_\star \quad (n = +\infty)$$

## Sanity-check: $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Theta}_\star^{-1})$

Erdös-Réyni $d = 64$



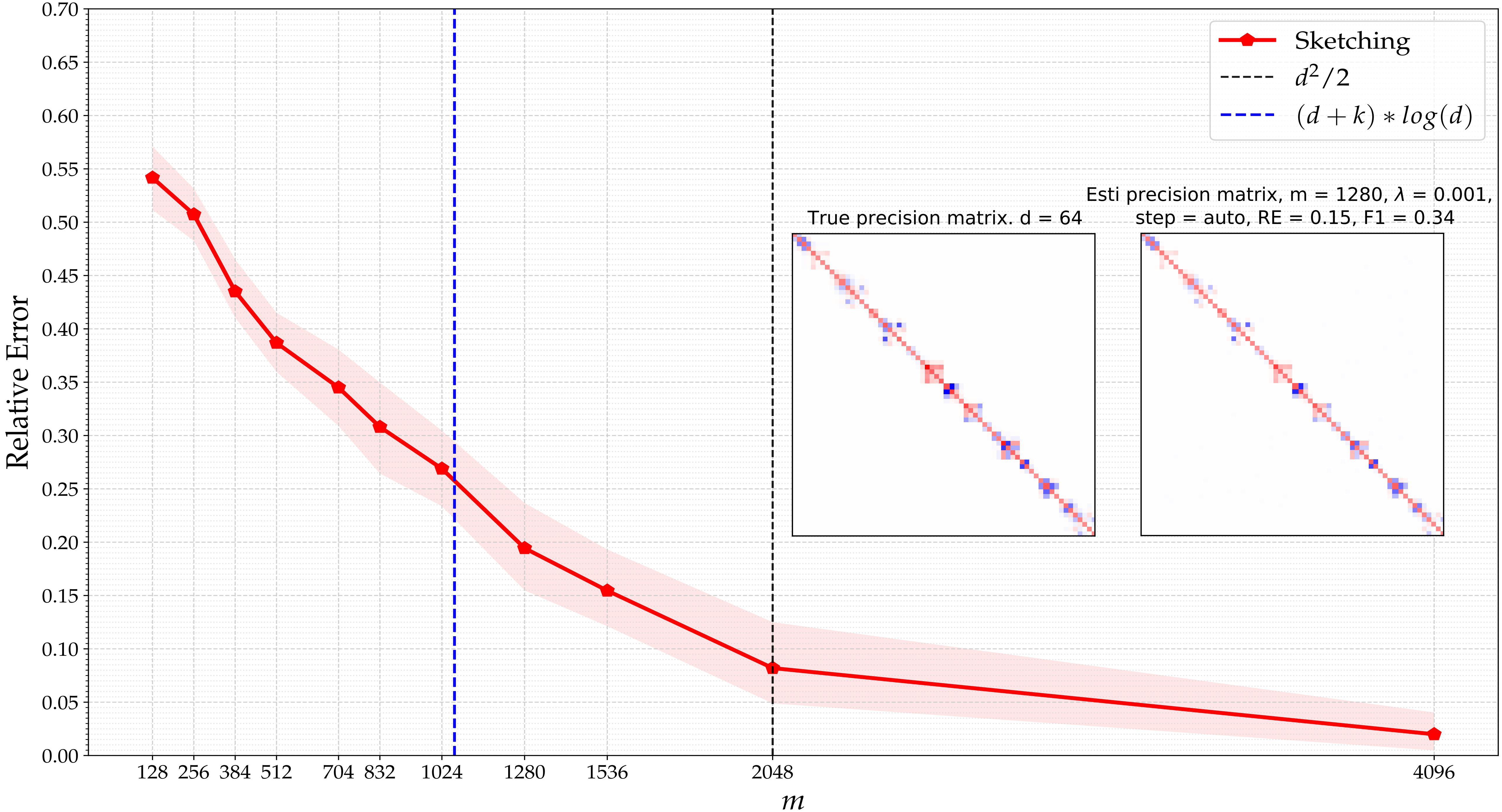
# Results

**Sketch of the true cov**

$$\mathbf{s} = \mathcal{A}\mathbf{\Sigma}_\star \quad (n = +\infty)$$

**Sanity-check:** $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \quad \mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Theta}_\star^{-1})$

Powerlaw $d = 64$



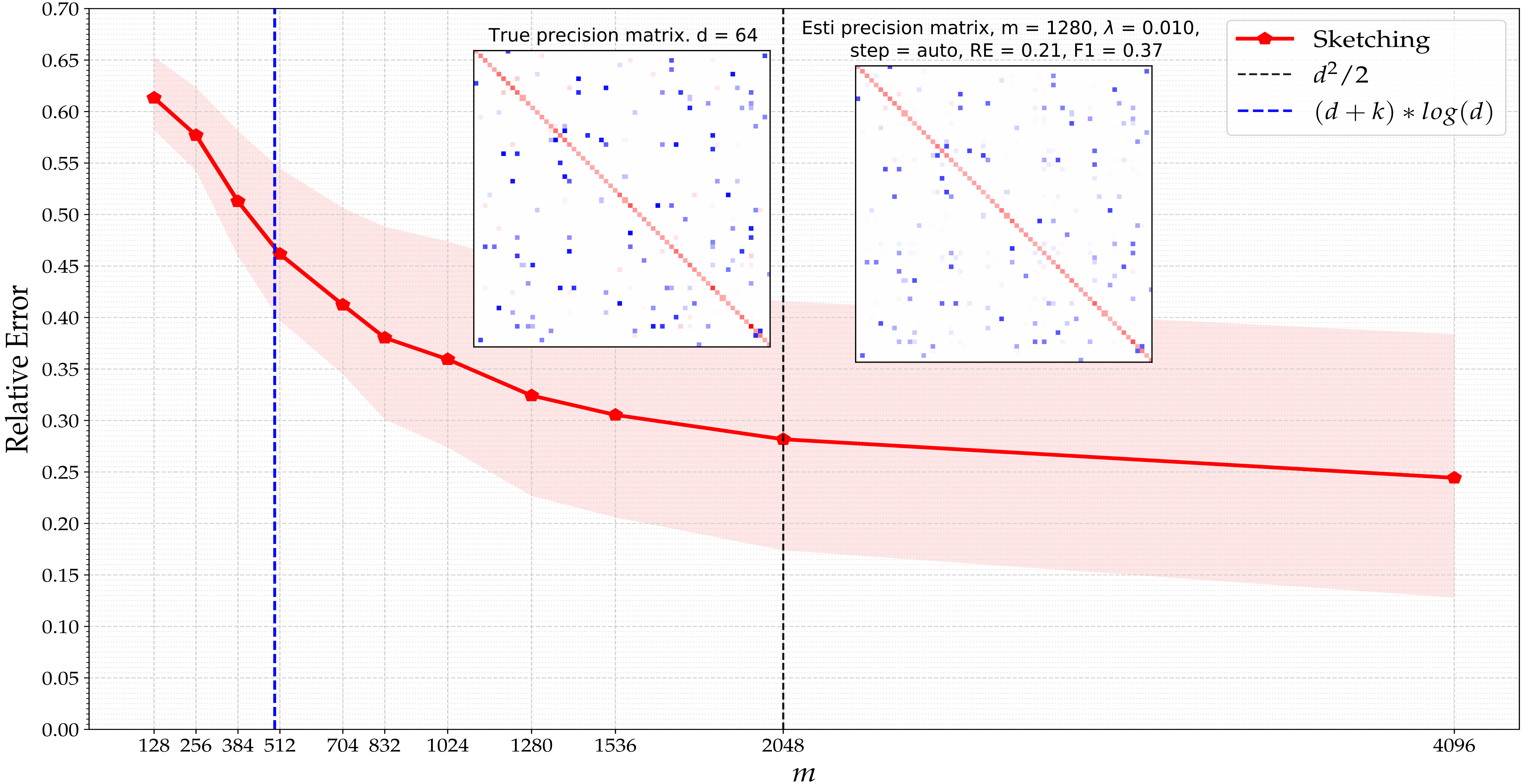
# Results

Sketch of the true cov

$$\mathbf{s} = \mathcal{A}\mathbf{\Sigma}_\star \quad (n = +\infty)$$

**Sanity-check:** $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n) \;\; \mathbf{x}_i \sim \mathcal{N}(0, \mathbf{\Theta}_\star^{-1})$

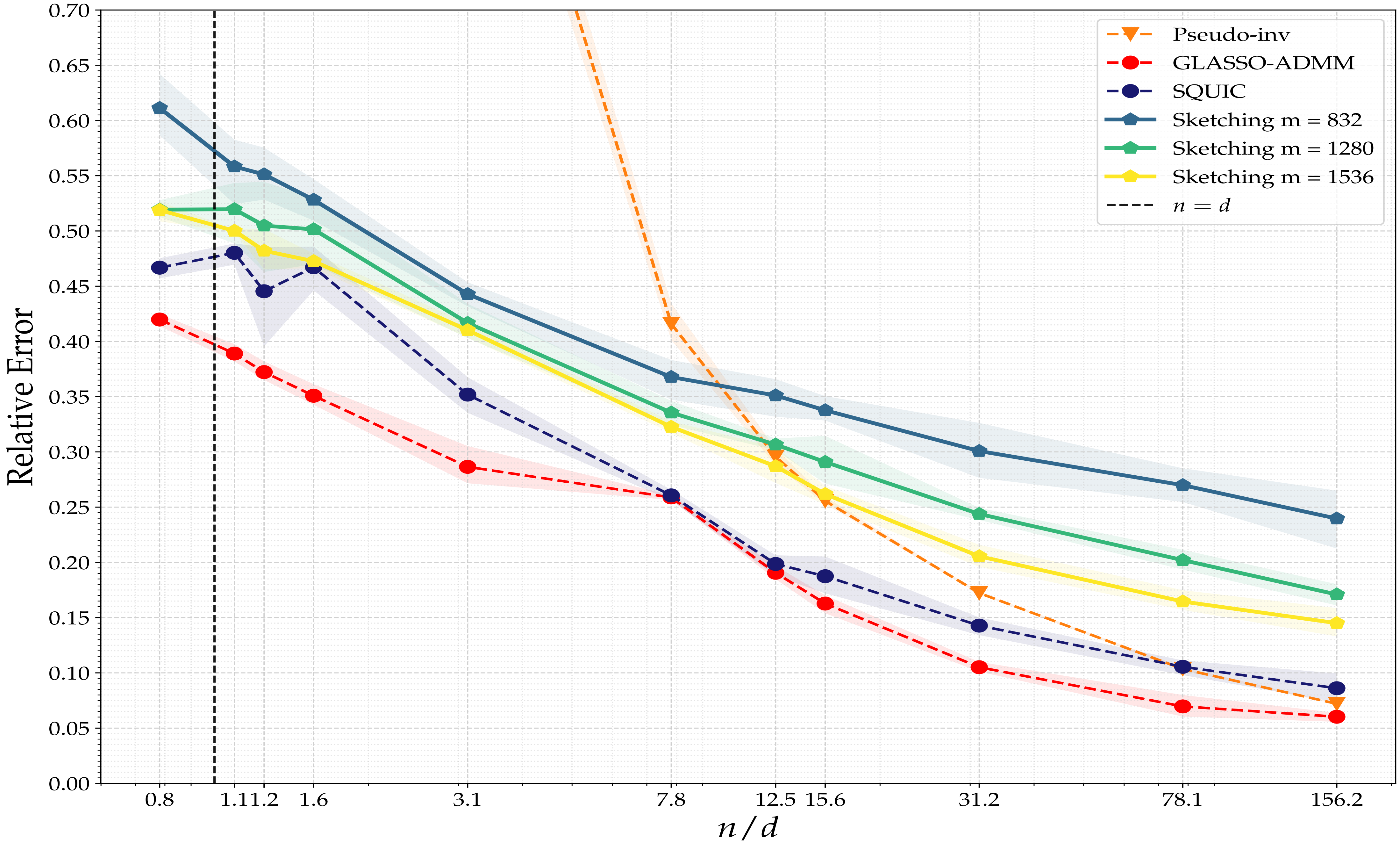
Sklearn $d = 64$



# Results

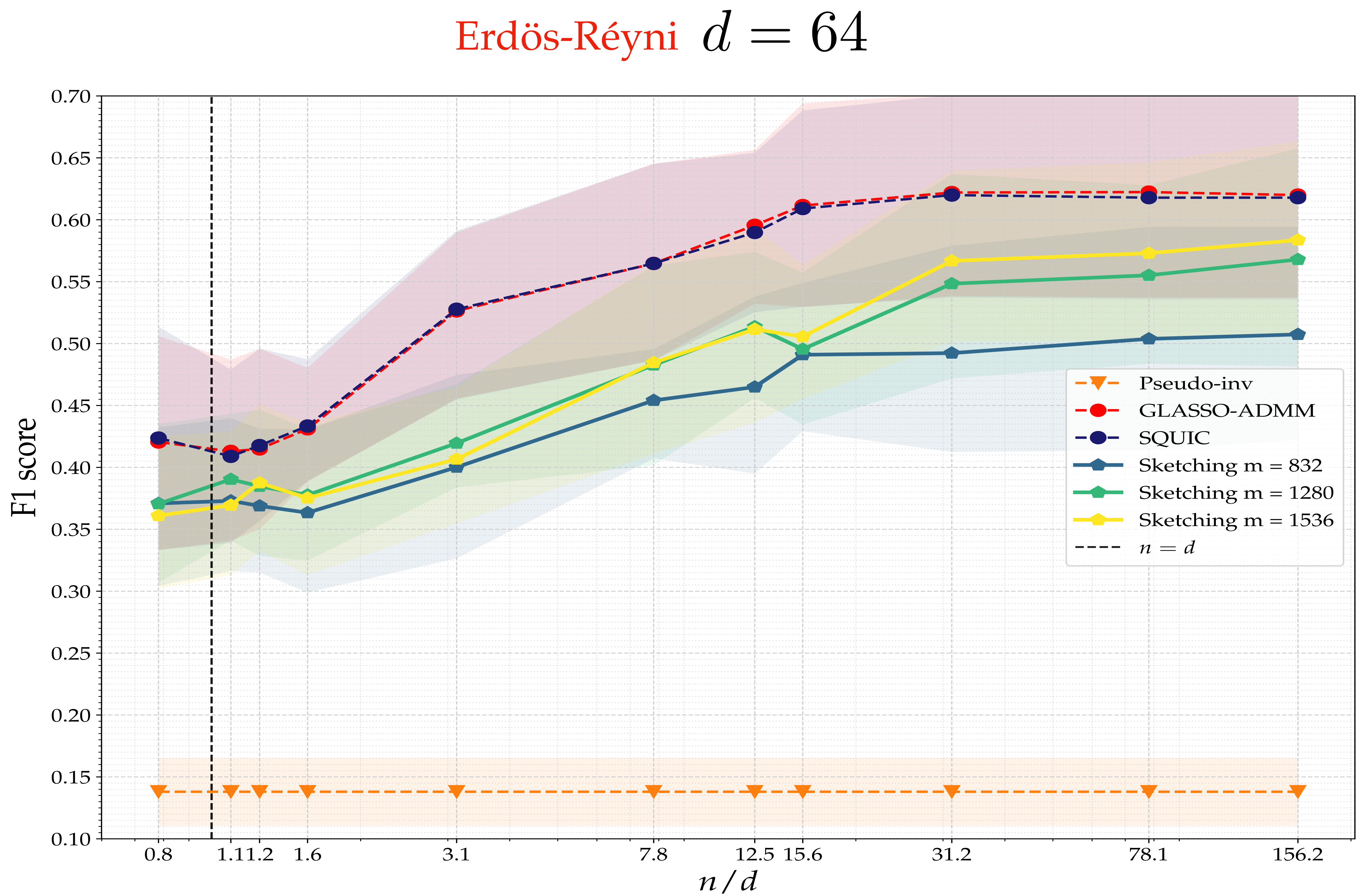
## Comparison with GLASSO:

Erdös-Réyni $d = 64$



# Results

## Comparison with GLASSO:



# Overview of the talk

- **Part I: Finding graphs from unstructured data**
- **Part II: The sketching approach**
- **Part III: Algorithmic solution**
- **Part IV: Limits, Open questions, partial answers**

# Limitations and perspectives

## About the complexity:

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

# Limitations and perspectives

- **About the complexity:**

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

| | In memory | In time |
|---|---|---|
| s - - ▶ | $m \approx (d+k)\ln(d)$ | |

# Limitations and perspectives

## About the complexity:

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

s $\dashrightarrow m \approx (d+k)\ln(d)$

+

$\mathcal{A}$ $\dashrightarrow$

| In memory | In time |
|---|---|
| $m \approx (d+k)\ln(d)$ | |
| **Gaussian** $\mathcal{O}(m \times d^2)$ | |
| **Rank-one** $\mathcal{O}(m \times d)$ | |

# Limitations and perspectives

## About the complexity:

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

| | In memory | In time |
|---|---|---|
| $\mathbf{s}$ → | $m \approx (d+k)\ln(d)$ | **Three op. splitting** |
| + | | $\mathcal{O}(d^3)$ |
| $\mathcal{A}$ → | **Gaussian** $\mathcal{O}(m \times d^2)$ | + |
| | **Rank-one** $\mathcal{O}(m \times d)$ | compute $\mathbf{s}$ $\mathcal{O}(n \times md)$ |

# Limitations and perspectives

## About the complexity:

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

$\mathbf{s} \dashrightarrow m \approx (d+k)\ln(d)$

$+$

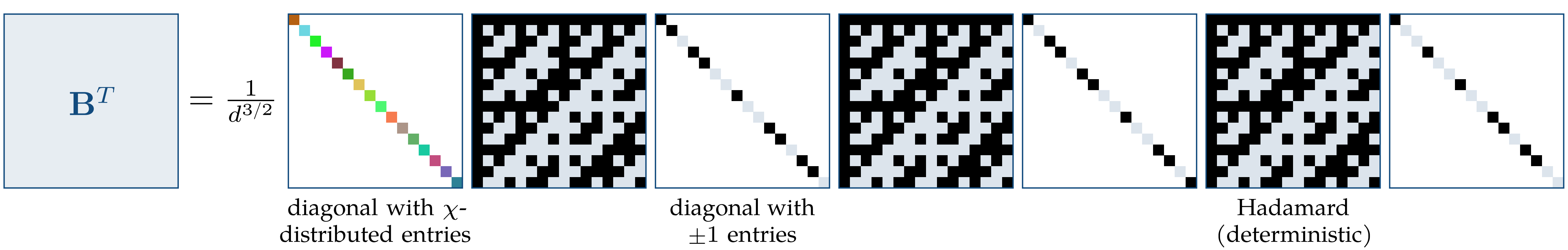
$\mathcal{A} \dashrightarrow$

| In memory | In time |
|---|---|
| $m \approx (d+k)\ln(d)$ | **Three op. splitting** |
| **Gaussian** $\mathcal{O}(m \times d^2)$ | $\mathcal{O}(d^3)$ + compute $\mathbf{s}$ |
| **Rank-one** $\mathcal{O}(m \times d)$ | $\mathcal{O}(n \times md)$ |

## Structured rank-one:

- Use random structured matrices:

$$\mathbf{W} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \cdots & \mathbf{B}_b \end{bmatrix}$$



# Limitations and perspectives

## About the complexity:

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

| | In memory | In time |
|---|---|---|
| $\mathbf{s} \dashrightarrow$ | $m \approx (d+k)\ln(d)$ | **Three op. splitting** $\mathcal{O}(d^3)$ |
| $+$ $\mathcal{A} \dashrightarrow$ | **Gaussian** $\mathcal{O}(m \times d^2)$ | $+$ |
| | **Rank-one** $\mathcal{O}(m \times d)$ | compute $\mathbf{s}$ $\mathcal{O}(n \times md)$ |

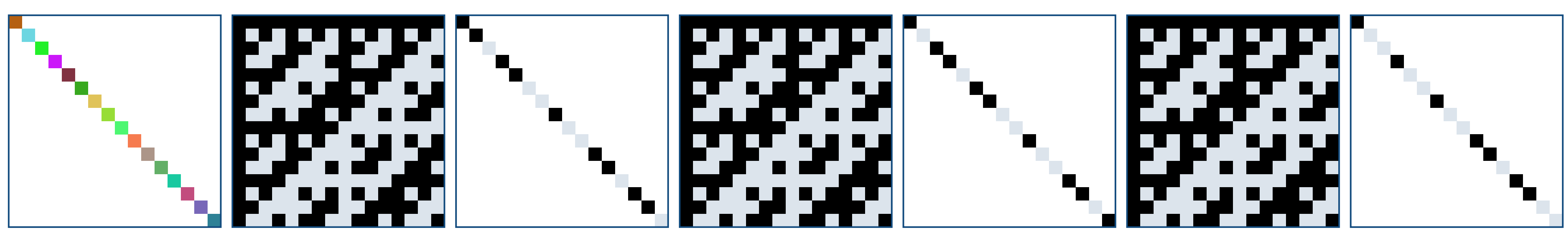
## Structured rank-one:

- Use random structured matrices:

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

$$\mathbf{a}_j \sim \mathrm{row}_j(\mathbf{W})$$ **NOT I.I.D !**

$$\mathbf{W} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \cdots & \mathbf{B}_b \end{bmatrix}$$

$\mathbf{B}^T = \frac{1}{d^{3/2}}$ 

# Limitations and perspectives

## About the complexity:

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

| | In memory | In time |
|---|---|---|
| $\mathbf{s}$ - - ▶ | $m \approx (d+k)\ln(d)$ | **Three op. splitting** |
| $+$ | | |
| $\mathcal{A}$ - - ▶ | **Gaussian** $\mathcal{O}(m \times d^2)$ | $\mathcal{O}(d^3)$ $+$ compute $\mathbf{s}$ |
| | **Rank-one** $\mathcal{O}(m \times d)$ | $\mathcal{O}(n \times md)$ |

## Structured rank-one:

- Use random structured matrices:

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

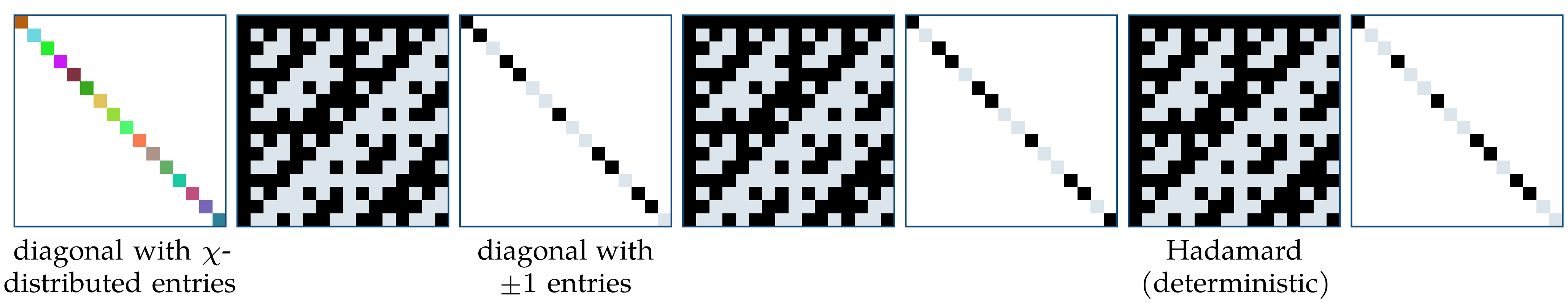
$$\mathbf{a}_j \sim \mathrm{row}_j(\mathbf{W})$$ **NOT I.I.D !**

$$\mathbf{W} = \begin{bmatrix} \mathbf{B}_1 & \mathbf{B}_2 & \mathbf{B}_3 & \cdots & \mathbf{B}_b \end{bmatrix}$$

**Structured rank-one**

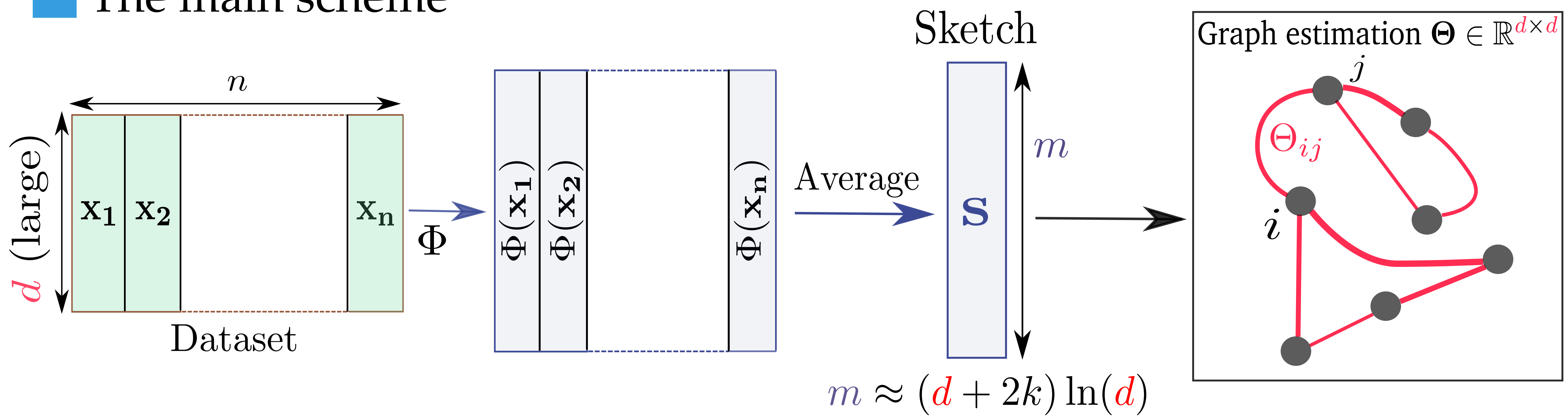
**Total in memory =** $\mathcal{O}(m)$

$$\mathbf{B}^T = \frac{1}{d^{3/2}}$$



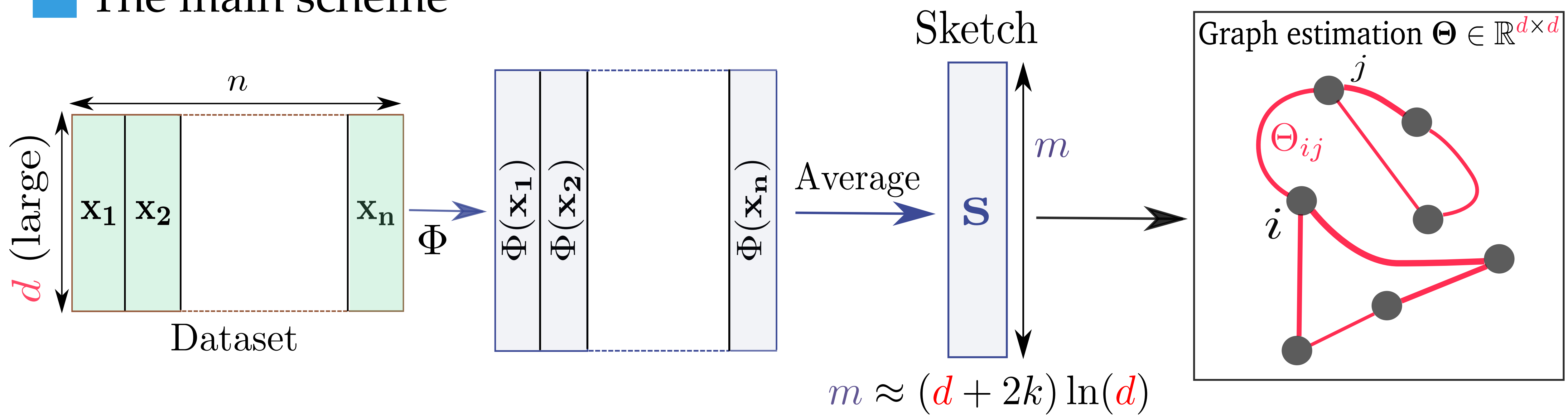
# Conclusion

## The main scheme



# Conclusion

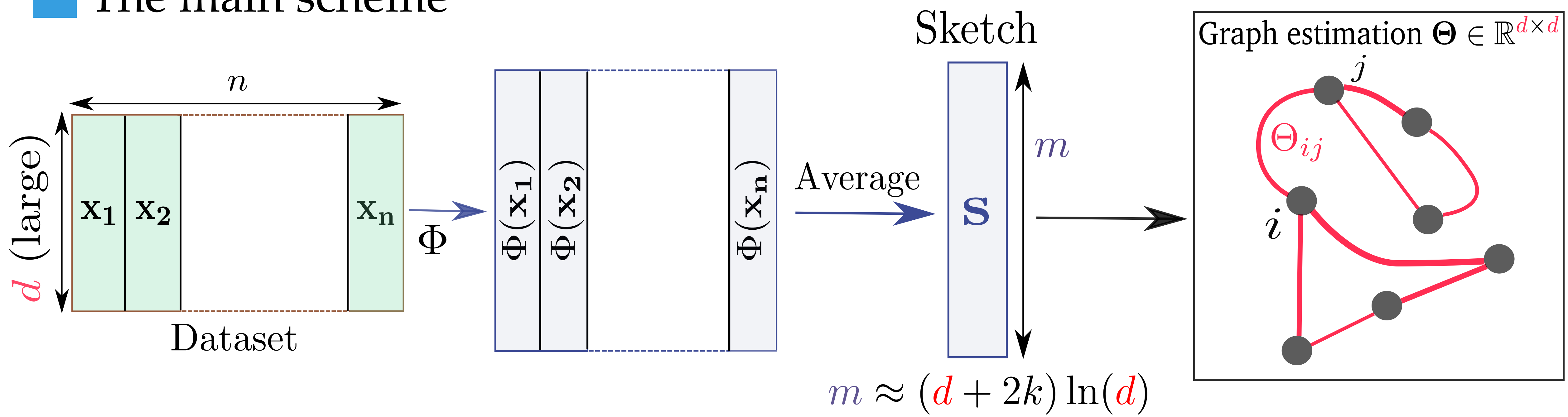
- The main scheme




- Theoretical guarantees: RIP, optimal decoders
- Recovery via Davis & Yin three operator spitting

# Conclusion

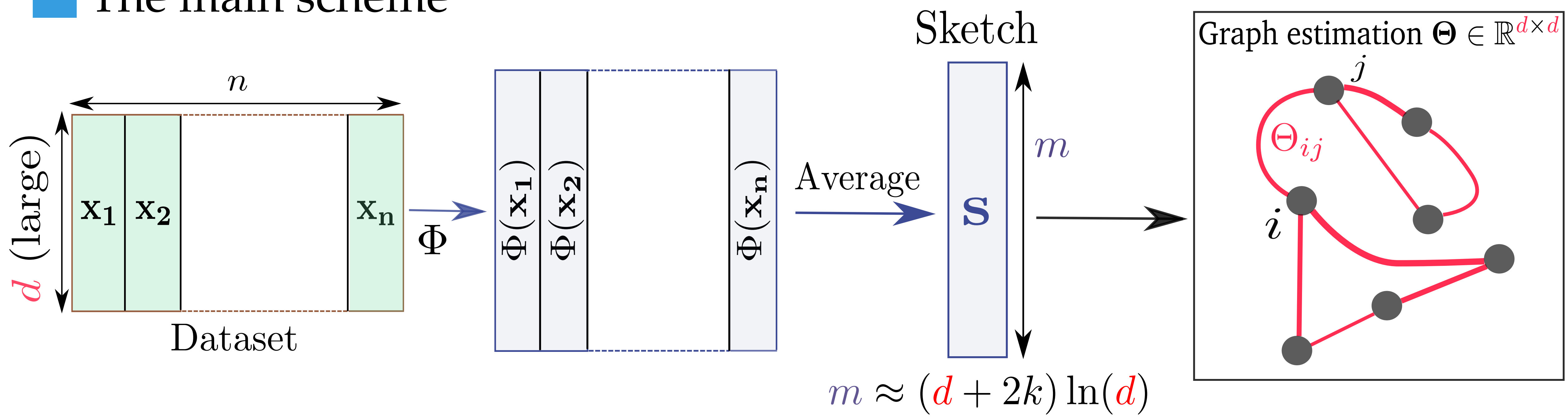
- The main scheme




- Theoretical guarantees: RIP, optimal decoders
- Recovery via Davis & Yin three operator spitting
- Limitations: we have to store $\mathcal{A}$ (Gaussian = $\mathcal{O}(d^3)$, rank one = $\mathcal{O}(d^2)$)
- Limitations: Algo not that efficient (Greedy approches ?)

# Conclusion

- The main scheme




- Theoretical guarantees: RIP, optimal decoders
- Recovery via Davis & Yin three operator spitting
- Limitations: we have to store $\mathcal{A}$ (Gaussian = $\mathcal{O}(d^3)$, rank one = $\mathcal{O}(d^2)$)
- Limitations: Algo not that efficient (Greedy approches ?)
- Perspectives: structured operators, different algo (greedy approaches ?)

theoretical guarantees ?

# Thank you!