

# Distributed Sparse Inverse Covariance Estimation from fMRI Data for Segmenting the Human Brain

Alnur Ali

*alnurali@cmu.edu*

Just FYI: this work is based on our paper, *Communication-Avoiding Optimization Methods for Distributed Massive-Scale Sparse Inverse Covariance Estimation*, which grew out of my practicum

# Outline

- What is “sparse inverse covariance estimation”?
- Computational approach
- Application to understanding the human brain
- Wrap-up

What is “sparse inverse covariance estimation”?

Let's break it down, piece-by-piece ...

“covariance estimation”



# “covariance estimation”

- Suppose we observe  $n$   $p$ -dim. Gaussian r.v.'s  $\sim \mathcal{N}(\mathbf{0}, \Sigma_0)$
- And we're interested in estimating  $\Sigma_0$  ; what do we do?

# “covariance estimation”

- Suppose we observe  $n$   $p$ -dim. Gaussian r.v.'s  $\sim \mathcal{N}(\mathbf{0}, \Sigma_0)$
- And we're interested in estimating  $\Sigma_0$  ; what do we do?
  
- Try *maximum likelihood estimation* (Fisher, 1912)
- First, we write down the log-likelihood (up to constants):

$$\underset{\Sigma \in \mathbf{S}_{++}^p}{\text{minimize}} \quad \log \det \Sigma + \mathbf{tr}(S \Sigma^{-1})$$

(space of  $p \times p$  positive definite matrices)   (sample covariance matrix)

# “covariance estimation”

- Suppose we observe  $n$   $p$ -dim. Gaussian r.v.'s  $\sim \mathcal{N}(0, \Sigma_0)$
- And we're interested in estimating  $\Sigma_0$ ; what do we do?
- Try *maximum likelihood estimation* (Fisher, 1912)
- First, we write down the log-likelihood (up to constants):

$$\underset{\Sigma \in \mathbf{S}_{++}^p}{\text{minimize}} \quad \log \det \Sigma + \mathbf{tr}(S \Sigma^{-1})$$

(space of  $p \times p$  positive definite matrices) (sample covariance matrix)

Problem # 1: the log-likelihood is **not** convex in  $\Sigma \in \mathbf{S}_{++}^p$   
(but it **is** convex in  $\Omega = \Sigma^{-1} \in \mathbf{S}_{++}^p$ )

“covariance estimation”



# “covariance estimation”

- But we can still compute the sample cov. matrix; it's just:

$$\hat{\Sigma} = \frac{1}{n} X^T X \quad \leftarrow \text{.....} \quad (X = \text{the “design matrix”} = n \times p)$$

Problem #2: the sample cov. matrix is **singular** if  $p > n$   
(can be written as the sum of  $n$  rank-one matrices)

# “covariance estimation”

- But we can still compute the sample cov. matrix; it's just:

$$\hat{\Sigma} = \frac{1}{n} X^T X \quad \leftarrow \text{.....} \quad (X = \text{the “design matrix”} = n \times p)$$

Problem #2: the sample cov. matrix is **singular** if  $p > n$   
(can be written as the sum of  $n$  rank-one matrices)

Problem #3: the sample covariance matrix is a “**bad**”  
estimate of  $\Sigma_0$  if  $p > n$

# “(sparse) inverse covariance estimation”

- Simple fix for problem #1 (nonconvexity): **change variables** and minimize over  $\Omega = \Sigma^{-1}$ , i.e., we now solve

$$\underset{\Omega \in \mathbf{S}_{++}^p}{\text{minimize}} \quad -\log \det \Omega + \mathbf{tr}(S\Omega) \quad (\text{convex})$$

- Subtle fix for problems #2,3 (bad estimate): **add regularization**, i.e., we now solve

$$\underset{\Omega \in \mathbf{S}_{++}^p}{\text{minimize}} \quad -\log \det \Omega + \mathbf{tr}(S\Omega) + \lambda \|\Omega\|_1 \quad (\text{still convex})$$

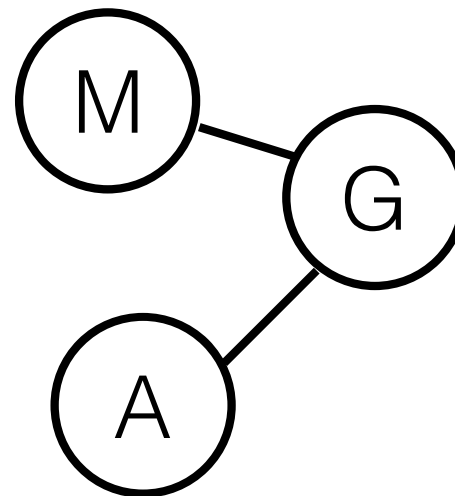
(elementwise L1 norm = sum up the absolute values of the entries of the argument)



# “*sparse* inverse covariance estimation”

- Another (nonobvious) benefit of regularization:
  - The regularized estimate gives rise to a *sparse graph*, where...
  - Vertices = variables
  - Edges = two variables are (conditionally) independent given all the others
- So, the regularized estimate has useful interpretability properties

	M	G	A
M	0.12	-0.01	0
G	-0.01	0.11	-0.02
A	0	-0.02	0.11



# Computational approach

# Computational approach

- We want to solve (same optimization problem from two slides ago):

$$\underset{\Omega \in \mathbf{S}_{++}^p}{\text{minimize}} \quad -\log \det \Omega + \mathbf{tr}(S\Omega) + \lambda \|\Omega\|_1$$

- A popular choice: use something like the backward Euler discretization (actually: a *proximal gradient method*); see Parikh & Boyd (2014)

# Computational approach

- We want to solve (same optimization problem from two slides ago):

$$\underset{\Omega \in \mathbf{S}_{++}^p}{\text{minimize}} \quad -\log \det \Omega + \mathbf{tr}(S\Omega) + \lambda \|\Omega\|_1$$

- A popular choice: use something like the backward Euler discretization (actually: a *proximal gradient method*); see Parikh & Boyd (2014)
- The main computational bottlenecks turn out to be:
  - Computing the **dense-dense** product  $S = \frac{1}{n} X^T X$ :  $O(p^2 n)$
  - Computing the **dense-sparse** product  $S\Omega$ :  $O(p^3)$

• We use recent *communication-avoiding* algorithms (Ballard et al., 2014) to compute these quantities in a distributed environment (**Edison**, **Eos**); toy example on the next slide

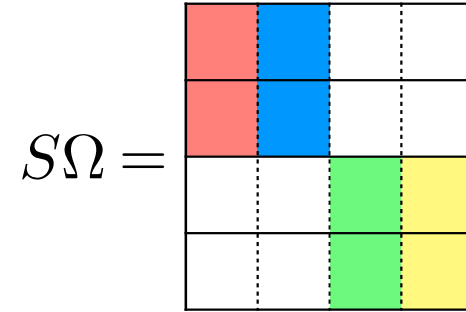
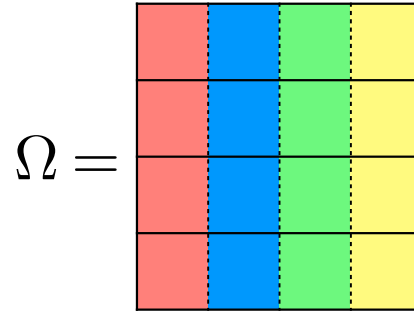
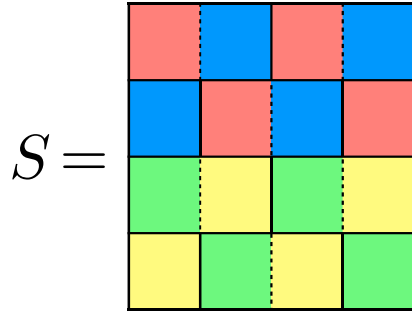
$$S, \Omega \in \mathbf{R}^{4 \times 4}$$

Proc. 0 (red), 1 (blue), 2 (green), 3 (yellow)

*On round 0 ...*

State:

Compute:





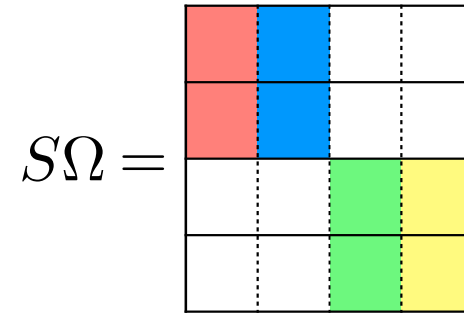
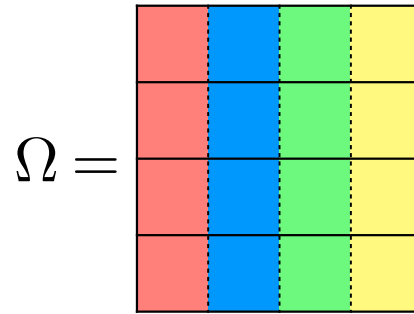
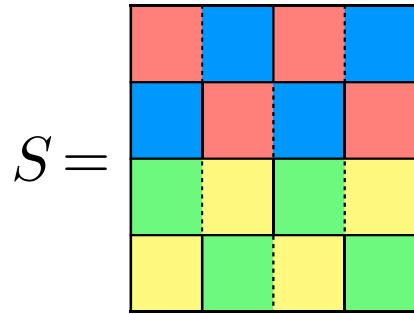
$$S, \Omega \in \mathbf{R}^{4 \times 4}$$

Proc. 0 (red), 1 (blue), 2 (green), 3 (yellow)

*On round 0 ...*

State:

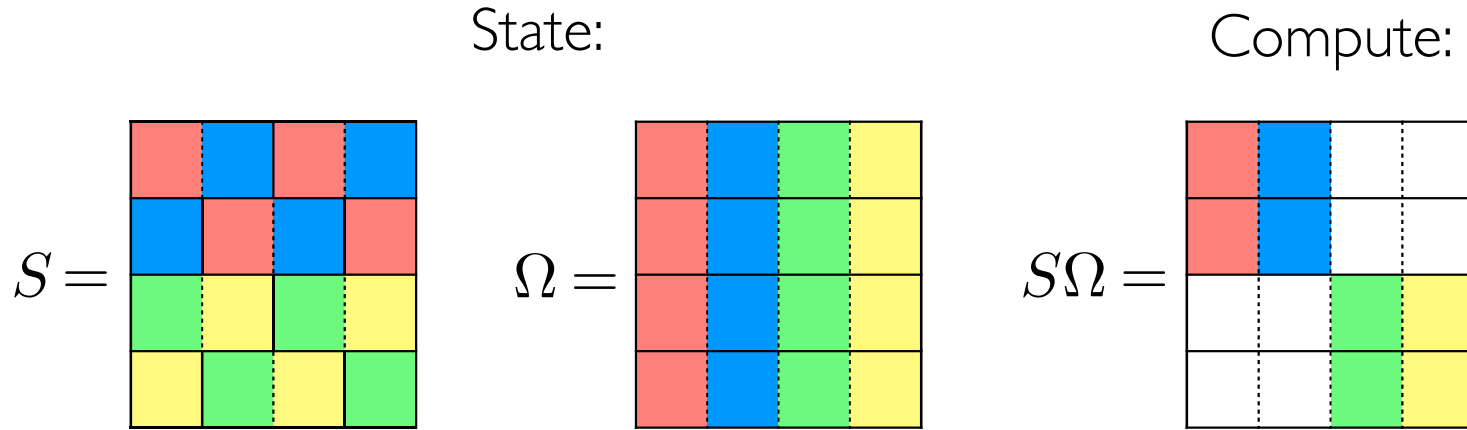
Compute:



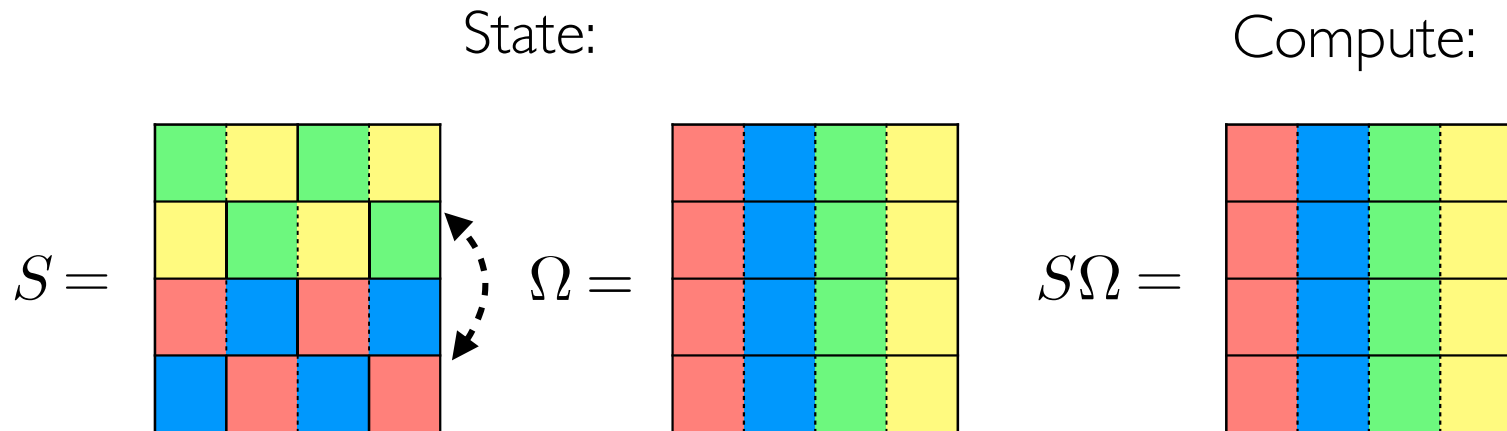
$$S, \Omega \in \mathbf{R}^{4 \times 4}$$

Proc. 0 (red), 1 (blue), 2 (green), 3 (yellow)

On round 0 ...



On round 1 ...



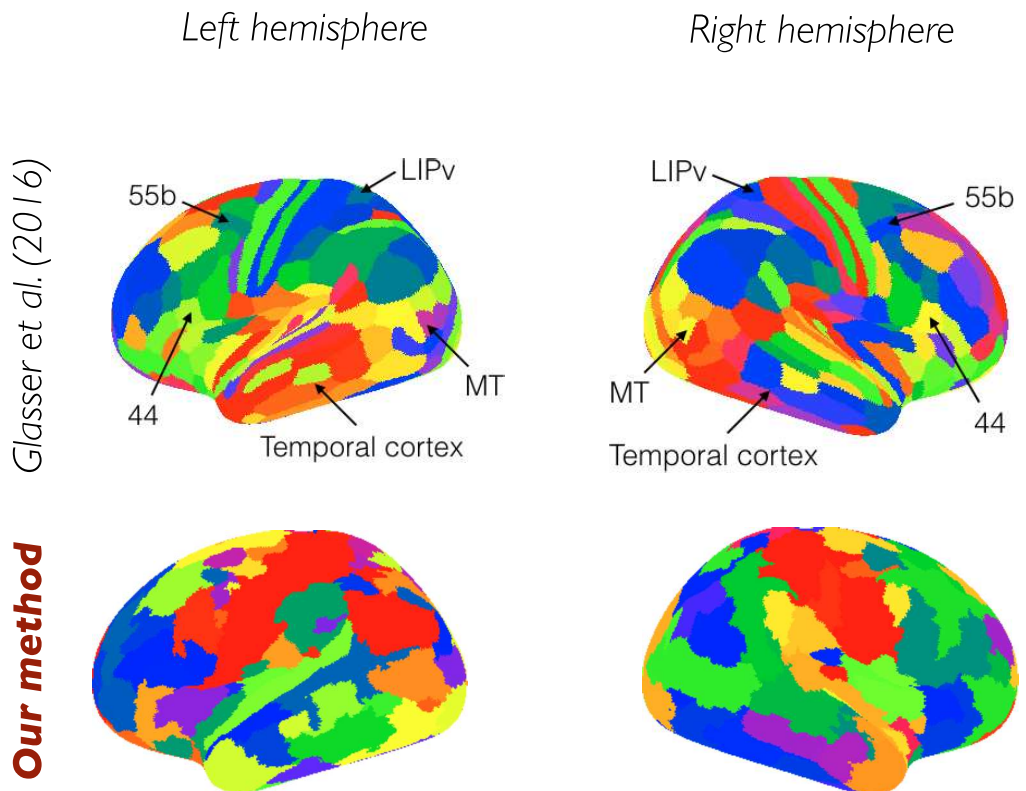
# Empirical evaluation

- Used our method to make progress on a challenging problem in neuroscience: “which parts of the human brain work together?”
- Starting point: functional magnetic resonance imaging (fMRI) data set, from the **Human Connectome Project**, where  $n = \#$  of patients,  $p = \#$  of voxels = 91,282 (hard)
- Our approach:
  - Run our method on the data, get a graph
  - Segment the graph into connected components (vertices = voxels), get a clustering
  - Compare to baseline from the neuroscience literature



# Numerical results

- Baseline = Glasser et al. (2016), generated by hand
- Our method **gets**:
  - Area 55b (hearing)
  - Lateral intraparietal cortex (eye movement)
  - Temporal cortex (information processing)
  - Other methods miss these (overly smooth)
- Our method **misses**:
  - Brodmann's area 44 (hearing + speaking)
  - Middle temporal visual area (seeing moving objects)



# Discussion

- Presented a method for sparse inverse covariance estimation, from *very* large-scale data
  - Our method is much more scalable than other methods in the literature (didn't have time to get into this)
- Applied the method to generate a segmentation of the cerebral cortex, from fMRI data
- The method recovered the structure in the data without being told how, performed comparably to strong baseline

*Thanks for listening!*