

# Notes (To Be Deleted)

- Be able to speak to glasso over neighborhood selection, how sparsity guarantees MLE existence
- Finish missing data results

# Graphical Models

With a focus towards interrimly missing data

Dominic DiSanto

Department of Biostatistics, Harvard University

December 3, 2023

Downloadable Slides

# Outline

- 1 Introduction to Graphical Models
- 2 Estimation for Complete Data
  - Neighborhood Selection
  - Graphical Lasso
  - Further Notes
- 3 Estimation with Missingness
  - MissGLasso
  - *Ero*se Data and GI-JOE

# Disclaimers

- Historical coverage is to the best of my ability and time constraint, please correct me with additional information
- Interrupt with any questions, clarification, confusion, etc.
- This is far from a comprehensive treatment, but I attempt to be holistic in my coverage

# Outline (Redux)

## 1 Introduction to Graphical Models

## 2 Estimation for Complete Data

- Neighborhood Selection
- Graphical Lasso
- Further Notes

## 3 Estimation with Missingness

- MissGLasso
- *Erode* Data and GI-JOE

# Graph Theory Origins [5, 8]

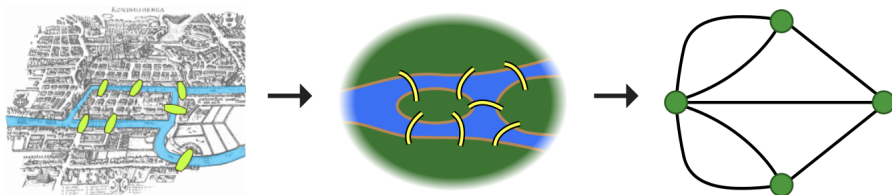


Figure: Euler's Bridges Conceptualization (Recreation)

1

<sup>1</sup>Image taken from Wikipedia ([https://en.wikipedia.org/wiki/Seven\\_Bridges\\_of\\_Konigsberg](https://en.wikipedia.org/wiki/Seven_Bridges_of_Konigsberg))

# Early Applications of Graphs in Mathematics

- Graph theory attributed to begin with Euler and the "Seven Bridges of Königsberg" ( $\sim 1736$ )
- Random graph theory began developing in  $\sim 1940$ 's (Moreno and Jennings) but most notably with the Erdős-Rényi random graph (1958)
- Ising model ( $\sim 1920$ 's) - proposed graphical model of interactions of atomic spin
- Statistical "beginnings"
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [2]
- Judea Pearl  $\sim 1980$ 's for causal interpretation of Bayesian networks
- Modern interest in related regularized M-estimation problems and graphical neural networks

[Downloadable Slides](#)


# Early Applications of Graphs in Mathematics

- Graph theory attributed to begin with Euler and the "Seven Bridges of Königsberg" ( $\sim 1736$ )
- Random graph theory began developing in  $\sim 1940$ 's (Moreno and Jennings) but most notably with the Erdős-Rényi random graph (1958)
- Ising model ( $\sim 1920$ 's) - proposed a graphical model of interactions of atomic spin
- Statistical "beginnings"
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [2]
- Judea Pearl  $\sim 1980$ 's for causal interpretation of Bayesian networks
- Modern interest in related regularized M-estimation problems and graphical neural networks



# Early Applications of Graphs in Mathematics

- Graph theory attributed to begin with Euler and the "Seven Bridges of Königsberg" ( $\sim 1736$ )
- Random graph theory began developing in  $\sim 1940$ 's (Moreno and Jennings) but most notably with the Erdős-Rényi random graph (1958)
- Ising model ( $\sim 1920$ 's) - proposed a graphical model of interactions of atomic spin
- Statistical "beginnings"<sup>2</sup> as a subset of methods for contingency tables and log-linear models ( $\sim 1970$ 's)
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [2]
- Judea Pearl  $\sim 1980$ 's for causal interpretation of Bayesian networks
- Modern interest in related regularized M-estimation problems and graphical neural networks

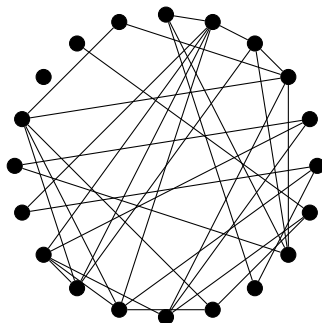
<sup>2</sup>Early use in physics were probabilistic, but this may be seen as an early "pursue statistics" application 

# Graphical Model Motivation

Suppose you have 20 random variables\*,  
how do you model their interrelationship?

\*Consider any of the following:

- General -omic data
- Spatial data
- Computational neuroscience data
- Clinical language (see: EHR LLM<sup>a</sup>)
- Time-series data



---

<sup>a</sup>Electronic Healthcare Record Large Language Model

# Graphs

- Graphs are a natural way to represent interrelationships among our data!
- Present nice properties for estimation of joint distributions
  - Can avail existing graphical algorithms
  - Ability to characterize conditional (in)dependencies
- Probabilistic graphical modelling provide a general formalism of many existing methods in statistics (e.g. Bayesian hierarchical modelling, Hidden Markov Models, Kalman filter)
- Wainwright, Jordan "*Graphical Models, Exponential Families, and Variational Inference*" (2007) is an excellent reference for further applications (and theory) behind graphical models [9]<sup>3</sup>

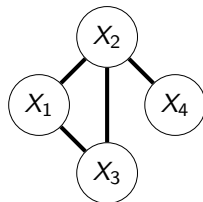
---

<sup>3</sup>See 2.4 specifically for applications

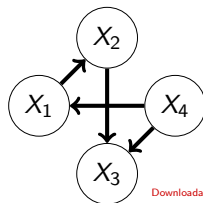
# Graphs

- Consider random vector  $X \sim N(\mu, \Sigma)$  and precision matrix  $\Theta \equiv \Sigma^{-1}$ 
  - Interested in estimating  $\Sigma$  to characterize joint distribution  $f_X$
- Can construct a resulting graph  $\mathcal{G} = (V, E)$ ,  $V = X, E \subseteq V \times V$ 
  - Let  $\text{ne}(x)$  represent the neighborhood of  $x$ , or  $\text{ne}(x) = \{b \in V \mid (x, b) \in E\}$
- Can construct adjacency matrix  $A \in \mathbb{R}^{|V| \times |V|}$  describing edge set  $E$ 
  - $A_{ij} = \mathbb{I}\{(i, j) \in E\}$
  - Let  $D_{\max}$  represent the maximum degree

Undirected Graph



Directed Graph



Downloadable Slides

# Notation/Nomenclature

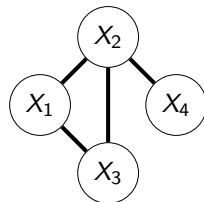
*Omitting some philosophical discrepancies*

- Directed (Acyclic) Graph  $\Leftrightarrow$  Bayesian network
- Undirected graph  $\Leftrightarrow$  Markov network / Markov random field

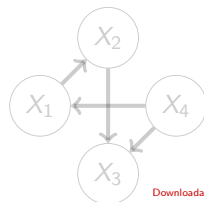
# Graphs

- Consider random vector  $X \sim N(\mu, \Sigma)$  and precision matrix  $\Theta \equiv \Sigma^{-1}$ 
  - Interested in estimating  $\Sigma$  to characterize joint distribution  $f_X$
- Can construct a resulting graph  $\mathcal{G} = (V, E)$ ,  $V = X, E \subseteq V \times V$ 
  - Let  $\text{ne}(x)$  represent the neighborhood of  $x$ , or  $\text{ne}(x) = \{b \in V \mid (x, b) \in E\}$
- Can construct adjacency matrix  $A \in \mathbb{R}^{|V| \times |V|}$  describing edge set  $E$ 
  - $A_{ij} = \mathbb{I}\{(i, j) \in E\}$
  - Let  $D_{\max}$  represent the maximum degree

Undirected Graph

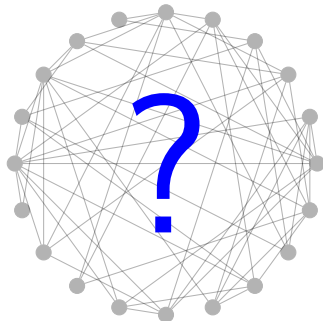


Directed Graph



Downloadable Slides

# How do we estimate graph structure?



# Gaussian Graphical Models

Recall the form and properties of a multivariate Gaussian random vector:

$$f(x; \mu, \Theta) = \frac{1}{(2\pi)^{d/2} |\Theta|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Theta (x - \mu) \right)$$

$$\mathbb{E}[X_i \mid X_{(-i)}] = \mu_i + (X_{(-i)} - \mu_{(-i)})^T \Theta_{j \neq i} \sigma_{i,j \neq i}$$

$$\text{Var}[X_i \mid X_{(-i)}] = \Sigma_{ii} - \sigma_{i,j \neq i}^T \Theta_{j \neq i} \sigma_{j \neq i, i}$$



# Gaussian Graphical Models

**Gaussianity gives us the nice property that**  $\Theta_{ij} = 0 \Leftrightarrow X_i \perp X_j | X_{-\{i,j\}}$

$$\mathbb{E}[X_i | X_{(-i)}] = \mu_i + (X_{(-i)} - \mu_{(-i)})^T \Theta_{j \neq i} \sigma_{i,j \neq i}$$

$$\text{Var}[X_i | X_{(-i)}] = \Sigma_{ii} - \sigma_{i,j \neq i}^T \Theta_{j \neq i} \sigma_{j \neq i, i}$$

# Outline (Redux)

## 1 Introduction to Graphical Models

## 2 Estimation for Complete Data

- Neighborhood Selection
- Graphical Lasso
- Further Notes

## 3 Estimation with Missingness

- MissGLasso
- *Erode* Data and GI-JOE

# Preface

- Consider the setting of graph estimation for  $X = (X_1, \dots, X_d)$ 
  - Generally your data may or may not contain a "response variable" of interest
- Identifying conditional relationships  $\Leftrightarrow$  Estimating/Identifying 0's in  $\Sigma$
- Enforce sparsity for  $\hat{\Theta}$  to account for possible rank-degeneracy of  $S$  if  $d \gg N$

# Neighborhood Selection

- Note that  $E \setminus \{\text{ne}(X_i)\}$  includes all nodes independent of  $X_i$  conditional upon  $\text{ne}(X_i)$
- Proposed by Meinshausen & Bühlmann (2006) [7], for  $X \in \mathbb{R}^d$  concern yourself only with  $(\Theta)_{ij} = 0$ , or  $(\Theta)_{ij} \neq 0$
- Assume sparsity of  $\Theta$  and fit  $d$ , element-wise lasso models
  - Regress  $X_i \stackrel{\text{Lasso}}{\sim} X_1 + \dots X_{i-1} + X_{i+1} + \dots + X_d$  for all  $i \in [d]$
  - Take  $\hat{\beta}_{(-i)} \in \mathbb{R}^{d-1}$  from each model
  - Conclude<sup>4</sup>  $(\Theta)_{ij} = 0 \Leftrightarrow \hat{\beta}_{(-i)j} = 0 \wedge \hat{\beta}_{(-j)i} = 0$
- Admits asymptotic consistency for "zero-selection" of  $\Theta$

<sup>4</sup>Authors both AND or OR rule for final step with similar performance

# Neighborhood Selection

## Potential Drawbacks:

- Fitting  $d$  regression models is almost assuredly redundant
- Although consistent, does not exactly compute but approximates the joint likelihood over  $X$ , and thus does not necessarily produce MLE [1]
- $\hat{\Theta}$  is *not* guaranteed to be positive semi-definite
- *Requires* sparsity assumptions for theoretical guarantees:
  - $\exists \kappa, \max_{a \in V} |\text{ne}(a)| = O(n^\kappa)$
  - For any connected nodes  $a, b$  (i.e.  $\forall (a, b) \in E$ ),  $\|\theta^{a, \text{ne}(b) \setminus \{a\}}\|_1 \leq \vartheta < \infty$

# Graphical Lasso

Natural extension, why not just maximize the log-likelihood?

$$\hat{\Theta}_{MLE} = \operatorname{argmax}_{\Theta} \{ \log \det \Theta - \operatorname{trace}(S\Theta) \}$$

For  $N < d$ , we have the empirical covariance matrix  $S = n^{-1} \sum X_i X_i^T$  is rank-degenerate, and the MLE does not exist!

# Graphical Lasso

So we assume sparsity and apply the  $\ell_1$  penalty

$$\hat{\Theta}_{\lambda, MLE} = \operatorname{argmax}_{\Theta} \left\{ \log \det \Theta - \operatorname{trace}(S\Theta) - \lambda \sum_{i \neq j} |\Theta_{ij}| \right\}$$

What does this give us?

- True graph recovery guaranteed for  $N = \Omega(D_{max}^3 \log p)$
- Convex program, quickly optimizable

# Graphical Lasso - Algorithm<sup>5</sup>

- 1: Initialize  $\mathbf{W} \leftarrow \mathbf{S}$
- 2: **for**  $j = 1, 2, \dots, d, 1, 2, \dots$  until convergence **do**:
- 3:     Partition  $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{kk} & \mathbf{w}_{kj} \\ \mathbf{w}_{jk} & w_{jj} \end{bmatrix}$
- 4:     Solve estimating equations  $\mathbf{W}_{kk}\beta - \mathbf{s}_{jk} + \lambda \text{sign}\beta = 0$
- 5:     Update  $\mathbf{w}_{kj} = \mathbf{W}_{kk}\hat{\beta}$
- 6: **for**  $j = 1, 2, \dots, d$  final update: **do**
- 7:     Solve for  $\hat{\theta}_{kj} = -\hat{\beta} \cdot \hat{\theta}_{jj}$  where  $1/\hat{\theta}_{jj} = 2_{jj} - \mathbf{w}_{kj}^T \hat{\beta}$ 
  - The estimating equations themselves are solved using cyclical coordinate-descent algorithm
  - Additional structural checks on  $\mathbf{S}, \mathbf{W}$  at initialization and interim steps have since been implemented

<sup>5</sup>Pseudocode adopted from Elements of Statistical Learning, Chapter 17 [3] and Statistical Learning with Sparsity, Chapter 9 [4]



# Simulations (Complete Data)

- `glasso` package in R can fit Graphical Lasso as well as neighborhood-selection approximation
- `huge` is a very nice extension of `glasso` with algorithmic/convergence fixes, computation in C, additional flexibility, graph generating functions
- `sklearn` has similar `sklearn.covariance.graphicallasso` command
- `skggm` extends Gaussian Graphical Model methods

# Simulations (Complete Data)

- Theory-suggested penalty  $\lambda = 2\sqrt{\frac{\log d}{N}}$ , but implementations often supply a range similar to `glmnet` default behavior
- Graph Recovery (accuracy by proportion of correct edge recovery)
- Operator Norm Distance  $\|\hat{\Theta} - \Theta\|_2 \lesssim \sqrt{\frac{D_{\max}^2 \log d}{N}}$

# Simulations (Set-Up)

- Generated multivariate normal data for  $d = \{64, 128, 256\}$  with  $\text{AR}(n, \rho)$  adjacency structures
- Assessed edge-selection performance for  $N$  ranging from 10 to 2000
  - $\text{TPR} = (\# \text{ of true edges selected}) / (\# \text{ of true edges})$
  - $\text{TNR} = (\# \text{ of true non-edges not selected}) / (\# \text{ of true non-edges})$
  - Operator norm  $\|\hat{\Theta} - \Theta\|_2$

# Simulations (Set-Up)

$$AR(3, \rho) = \begin{bmatrix} 1 & \rho^1 & \rho^2 & \rho^3 & 0 & \dots & \dots & \dots & 0 \\ \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 & 0 & \dots & \dots & 0 \\ \rho^2 & & \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 & 0 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & \rho^3 & \rho^2 & \rho^1 & 1 & \rho^1 & \rho^2 & \rho^3 \end{bmatrix}$$

# Simulations (GLASSO - Complete Data)

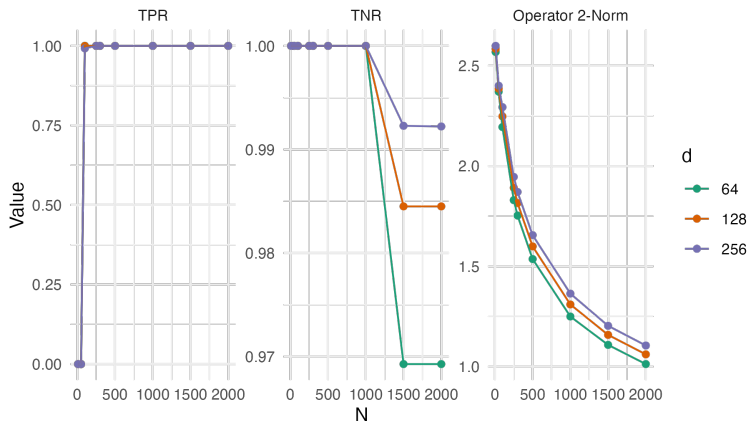


Figure: AR(1),  $\rho = 0.4$  adjacency structure

# Simulations (GLASSO - Complete Data)

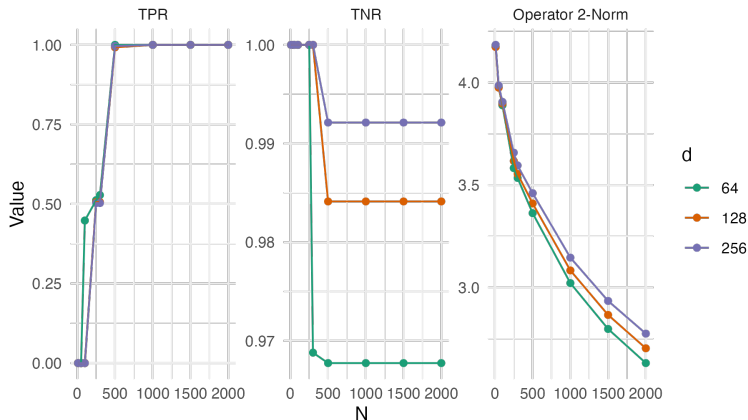


Figure: AR(2),  $\rho = 0.4$  adjacency structure

# Simulations (GLASSO - Complete Data)

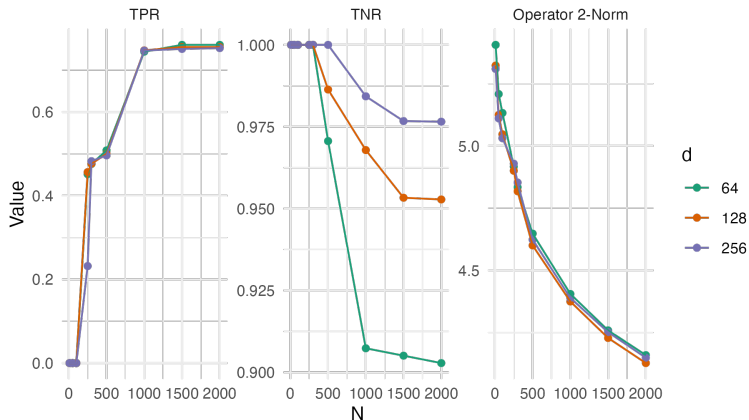


Figure: AR(4),  $\rho = 0.4$  adjacency structure

# Simulations (GLASSO - Complete Data)

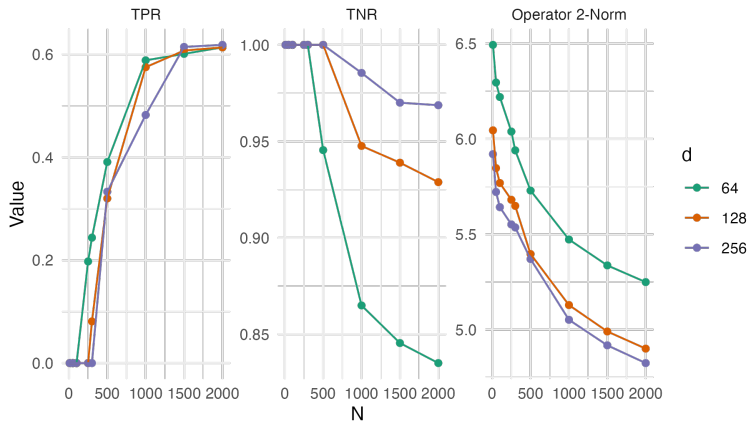


Figure: AR(8),  $\rho = 0.4$  adjacency structure



# Simulations (Neighborhood Selection - Complete Data)

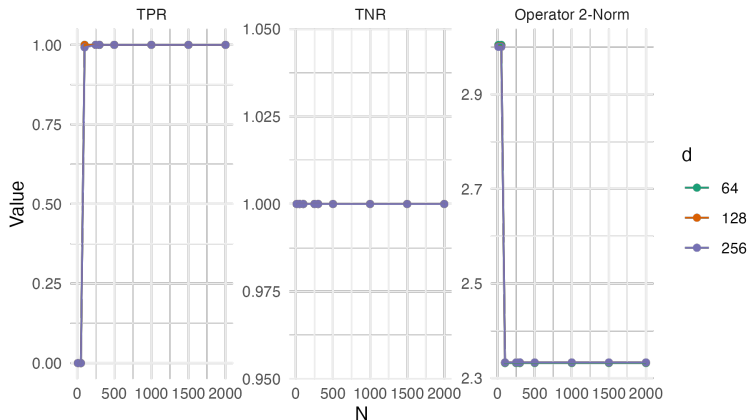


Figure: AR(1),  $\rho = 0.4$  adjacency structure

# Simulations (Neighborhood Selection - Complete Data)

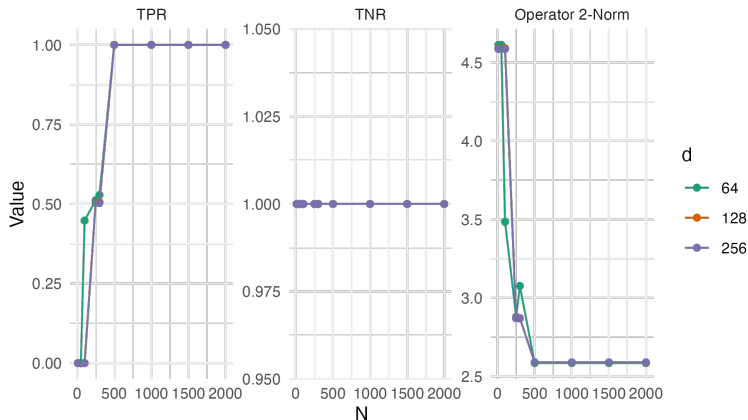


Figure: AR(2),  $\rho = 0.4$  adjacency structure

# Simulations (Neighborhood Selection - Complete Data)

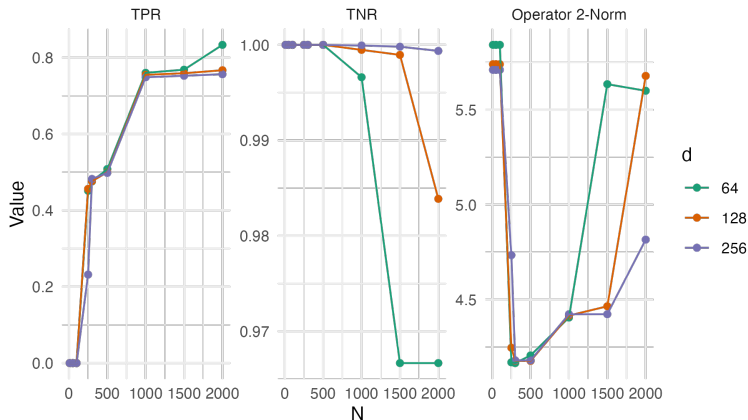


Figure: AR(4),  $\rho = 0.4$  adjacency structure

# Simulations (Neighborhood Selection - Complete Data)

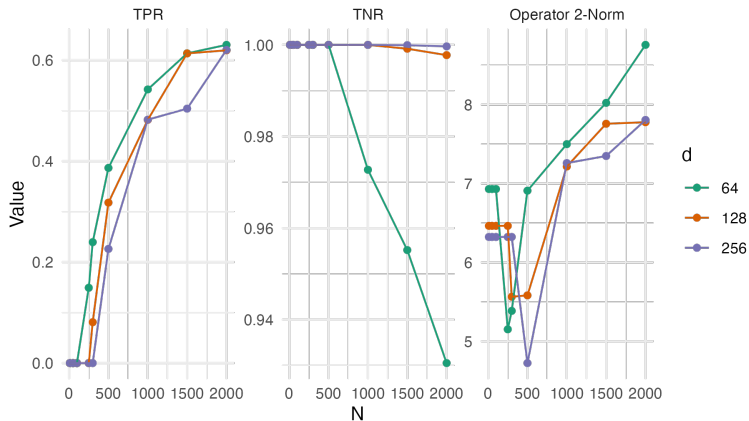


Figure: AR(8),  $\rho = 0.4$  adjacency structure

## Further Notes

- Neighborhood selection (slightly) seems to outperform GLasso in edge-selection, more notably in our less-sparse settings<sup>6</sup>
- GLasso "better approximate"  $\Theta$  with asymptotic guarantees not provided by neighborhood selection
- Error scaling more vulnerable to  $D_{max}$  than  $d$ -dimensionality of random vector
- Omitted results for random Erdős Rényi graphs yield similar results, conclusions

---

<sup>6</sup>These results are also slightly altered by choice of  $\rho$ , our value  $\rho = 0.3$  was chosen arbitrarily and kept constant only for brevity

# Outline (Redux)

- 1 Introduction to Graphical Models
- 2 Estimation for Complete Data
  - Neighborhood Selection
  - Graphical Lasso
  - Further Notes
- 3 Estimation with Missingness
  - MissGLasso
  - *Erode* Data and GI-JOE

# Motivation

- Methods above largely assume complete data
- Networks change, measurement availability (and quality) varies
- Measurement is also often differential between nodes
- As a result, complete case analysis can drastically reduce sample size if requiring complete data on all nodes

# Graphical Methods for Missingness

- Suppose for  $X \sim N(0, \Sigma)$ , we can partition  $X = (X_o, X_m)$
-

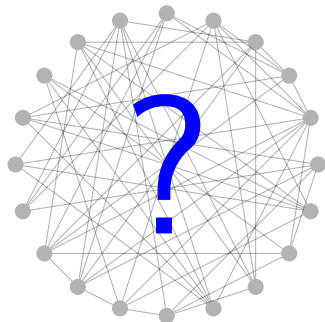


# Erose Data

- *Erose* data is a term coined by Zheng, Allen (2023) for data with irregular availability [10]
  - Leads to "drastically different" sample size for a small subset of nodes
  - *Erose* data almost certainly violate MAR/MCAR assumptions of existing methods
  - Motivated by neuroscience but with applications in genetic expression data,

# GI-JOE for Erore Data

# Graphs, how and why (revisited)?



————→ Regularized M-estimation(+)

# Conclusion

- Graphs are a powerful representation of your multivariate data (intuitively and algorithmically)
- Useful, theoretical extensions may follow more immediately under the graphical model formalism
- These extensions tend<sup>7</sup> to distill to regularized M-estimation problems, an area with great theoretical contributions and guarantees
- Extensions beyond Gaussianity substantially increase complexity

---

<sup>7</sup>Under a high-dimensional/assumed-sparsity regime

# References I

- Some diagrams generated in conjunction with ChatGPT 3.5

- [1] Onureena Banerjee and Laurent El Ghaoui. “Model Selection Through Sparse Maximum Likelihood Estimation for Multivariate Gaussian or Binary Data”. en. In: *Journal of Machine Learning Research* 9 (2008), pp. 485–516.
- [2] A. P. Dempster. “Covariance Selection”. In: *Biometrics* 28.1 (1972). Publisher: [Wiley, International Biometric Society], pp. 157–175.
- [3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer, 2009.
- [4] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. Dec. 2016.

# References II

- [5] Imperatorskaia akademiia nauk (Russia). *Commentarii Academiae scientiarum imperialis Petropolitanae*. lat. Petropolis, Typis Academiae, 1726.
- [6] Rahul Mazumder and Trevor Hastie. *The Graphical Lasso: New Insights and Alternatives*. arXiv:1111.5479 [cs, stat]. Aug. 2012.
- [7] Nicolai Meinshausen and Peter Bühlmann. “High-dimensional graphs and variable selection with the Lasso”. In: *The Annals of Statistics* 34.3 (June 2006). Publisher: Institute of Mathematical Statistics, pp. 1436–1462.
- [8] Rob Shields. “Cultural Topology: The Seven Bridges of Königsburg, 1736”. en. In: *Theory, Culture & Society* 29.4-5 (July 2012). Publisher: SAGE Publications Ltd, pp. 43–57.

# References III

- [9] Martin J. Wainwright and Michael I. Jordan. “Graphical Models, Exponential Families, and Variational Inference”. *en. In: Foundations and Trends® in Machine Learning* 1.1–2 (2007), pp. 1–305.
- [10] Lili Zheng. *GI-JOE: Graph Inference when Joint Observations are Erode*. Mar. 2023.

# Appendix Slides



# Erdős Rényi Graph Results (Complete Data)

# Simulations (GLASSO - Complete Data)

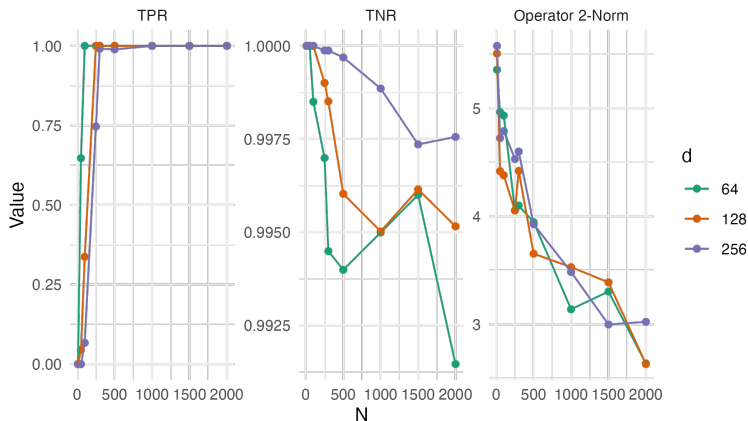


Figure: ER( $p=0.01$ ),  $\rho = 0.4$  adjacency structure

# Simulations (GLASSO - Complete Data)

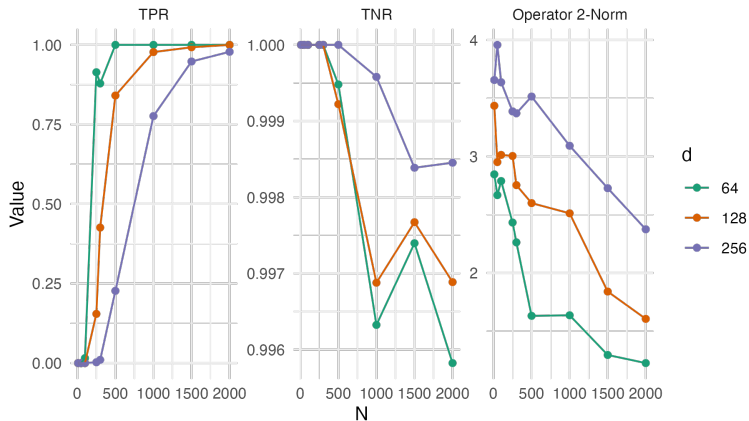


Figure: ER( $p=0.05$ ),  $\rho = 0.4$  adjacency structure

# Simulations (GLASSO - Complete Data)

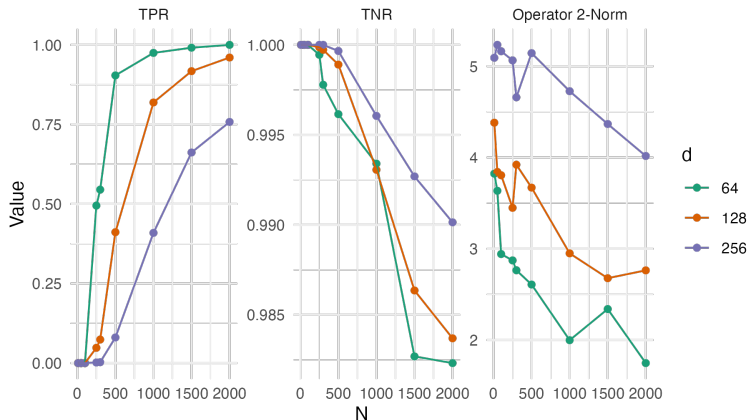


Figure: ER(p=0.1),  $\rho = 0.4$  adjacency structure

# Simulations (GLASSO - Complete Data)

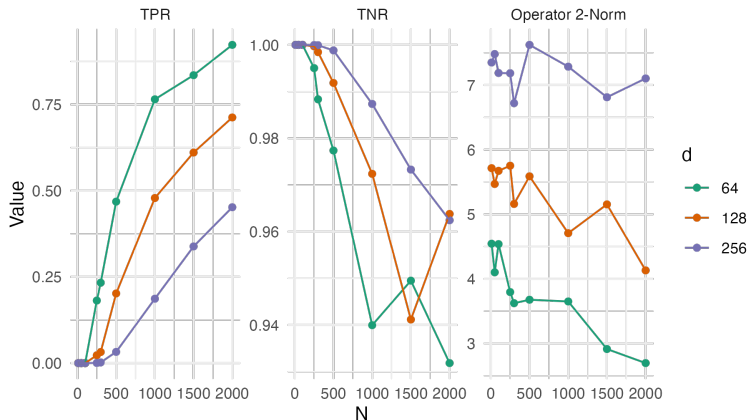


Figure: ER(p=0.2),  $\rho = 0.4$  adjacency structure

# ER-Simulations (Neighborhood Selection - Complete Data)

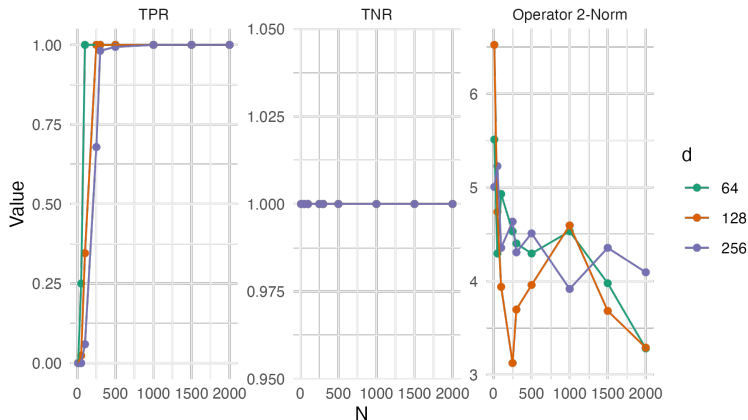


Figure: ER( $p=0.01$ ),  $\rho = 0.4$  adjacency structure

# ER-Simulations (Neighborhood Selection - Complete Data)

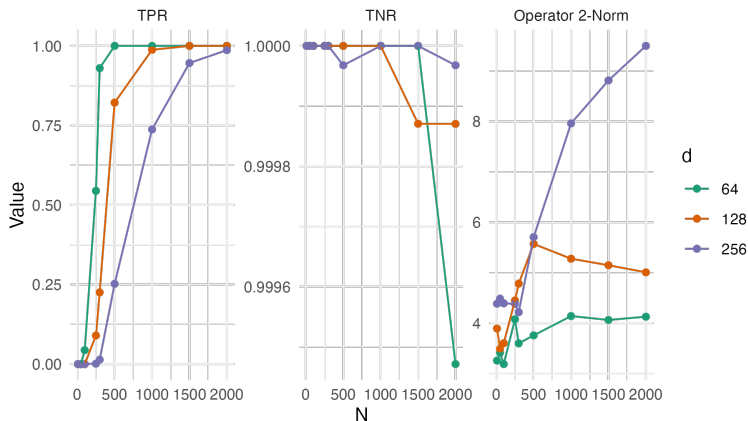


Figure: ER( $p=0.05$ ),  $\rho = 0.4$  adjacency structure

# ER-Simulations (Neighborhood Selection - Complete Data)

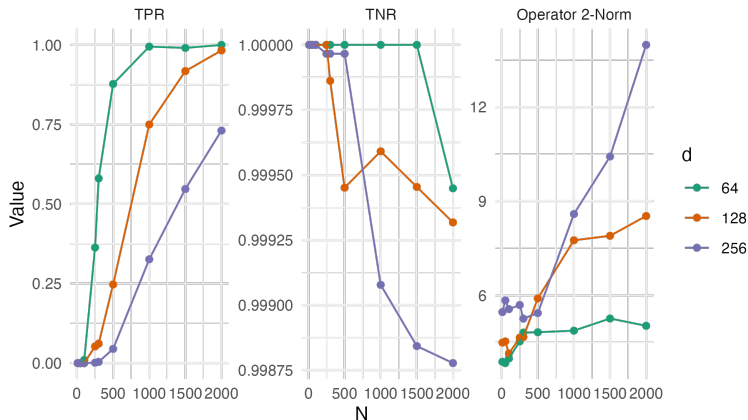


Figure: ER( $p=0.1$ ),  $\rho = 0.4$  adjacency structure



# ER-Simulations (Neighborhood Selection - Complete Data)

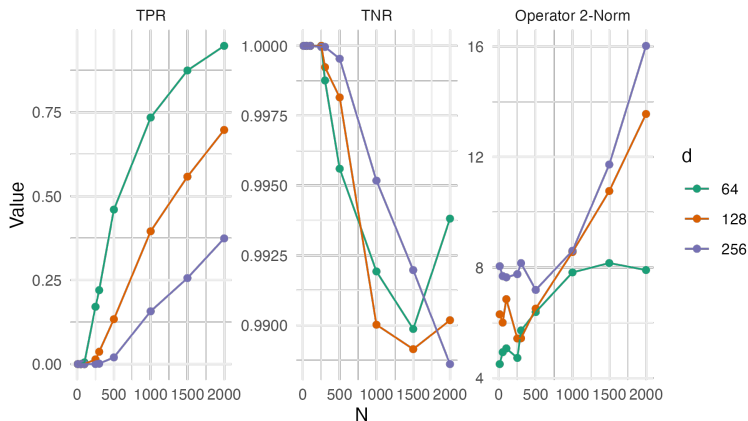


Figure: ER(0.2),  $\rho = 0.4$  adjacency structure

# Misc. Notes

# Conditional Graphical Lasso

- Cite `cglasso` package, MissGLasso paper for relevant conditional graphical lasso (i.e. predictor/response sets of interest)

# Missing Data Methods *Utilizing Graphs*

- We focus on estimating graphs when the underlying (graphical) data of interest includes missingness
- cite mohan/Pearl articles

# Forgoing Sparsity Assumptions

- In the above methods, we have almost uniformly assumed some sparsity and applied a penalty ( $\ell_1$ )
  - 1 How often is this a viable assumption?
  - 2 What do we do (or what happens) if we don't meet this sparsity requirement, more severely than our  $\text{AR}(\cdot)$  extension sims?
- Mazumder (2012) [6] offers an updated algorithm and insight into performance for  $p$  close to but larger than  $N$
- Interplay between  $d$ ,  $N$ , and graph-connectedness affect computation time and convergence

# Time-Series Data

- Consider that our repeated observations are time-indexed:
  - $\{X_j(t), t \in \mathcal{T}, j = 1, \dots, N\}, X_j \in \mathbb{R}^d$
- Graphical perspective of vector auto-regressive models
  - $X_d(t) = \varepsilon_d(t) + \sum_{j \neq d} \sum_{t \in \mathcal{T}} \alpha_t X_j(t)$
- Can infer "Granger causal" relationships
  - Causal relationships for some time-series using prior data from a *different time series*

See Michael Eichler's "*Granger-causality graphs for multivariate time series*" (2007) and Dahlhaus's and Eichler's (2003) "*Causality and graphical models in time series*" for further discussion

# Inference with Debiased Lasso

- The typical lasso estimator  $\hat{\beta}_\lambda = \operatorname{argmin}_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$  is biased for true  $\beta^*$
- Can construct debiased estimator  $\hat{\beta}_\lambda^d$  with asymptotic normality
- What inference does this permit in graphical models that use  $\ell_1$  penalization?

# "Nothing new under the sun"

My (likely useless and certainly non-falsifiable) conspiracy theory: Did Euler *really* originate graph theory? For how intuitive graphs seem to understanding interrelationships, this much have existed in some primitive form? Or for how financially relevant this seems, I'm sure some BCE gambler had an idea of "interconnectedness"

For our historical blinders, see Babylonian and Chinese origins of the Pythagorean Theorem

Thoughts, possible leads? Let me know!