Graphical Models

With extensions for missing data

Dominic DiSanto

Department of Biostatistics, Harvard University

December 5, 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
   - Additional Missing Data Methods

Downloadable Slides

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
   - Additional Missing Data Methods

Downloadable Slides

Graph Theory Origins [8, 11]

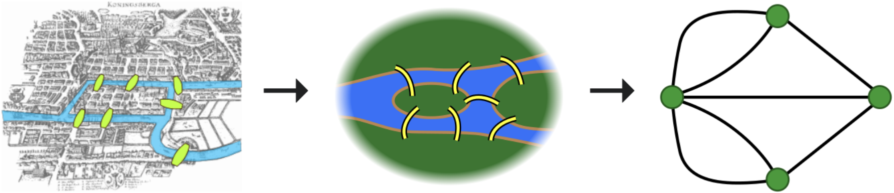


Figure: Euler's Bridges Conceptualization (Recreation)

1

Downloadable Slides

[1] Image taken from Wikipedia (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" (∼1736)
- Random graph theory began developing in ∼1940's (Moreno and Jennings) but most notably with the Erdös-Rényi random graph (1958)
- Ising model (∼1920's) - proposed graphical model of interactions of phase transitions (atomic spin)
- Statistical "beginnings"
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [5]
- Judea Pearl ∼1980's for causal intepretation 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" (∼1736)
- Random graph theory began developing in ∼1940's (Moreno and Jennings) but most notably with the Erdös-Rényi random graph (1958)
- Ising model (∼1920's) - proposed a graphical model of interactions of phase transitions (atomic spin)
- Statistical "beginnings"
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [5]
- Judea Pearl ∼1980's for causal intepretation 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" (∼1736)
- Random graph theory began developing in ∼1940's (Moreno and Jennings) but most notably with the Erdös-Rényi random graph (1958)
- Ising model (∼1920's) - proposed a graphical model of interactions of phase transitions (atomic spin)
- Statistical "beginnings"[2] as a subset of methods for contingency tables and log-linear models (∼1970's)
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [5]
- Judea Pearl ∼1980's for causal intepretation of Bayesian networks
- Modern interest in related regularized M-estimation problems and graphical neural networks

[2] Early use in physics were probabilistic, but this may be seen as an early "purse statistics" application

Downloadable Slides

Graphical Model Motivation

- 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 [12][3]

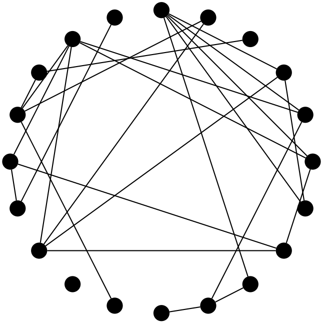
[3] See 2.4 specifically for applications

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[a])
- Time-series data



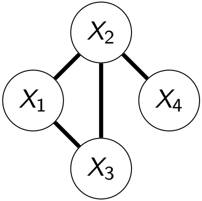
[a] Electronic Healthcare Record Large Language Model

Downloadable Slides

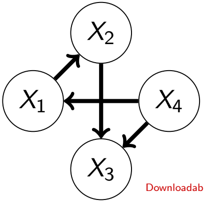
Graphs

- Consider random vector $X \sim N(\boldsymbol{\mu}, \Sigma)$ and precision matrix $\Theta \equiv \Sigma^{-1}$
  - Interested in estimating Σ 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

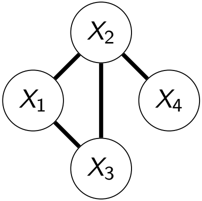


Downloadable Slides

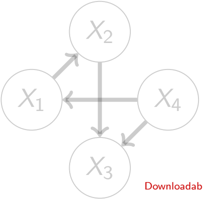
Graphs

- Consider random vector $X \sim N(\boldsymbol{\mu}, \Sigma)$ and precision matrix $\Theta \equiv \Sigma^{-1}$
  - Interested in estimating Σ 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

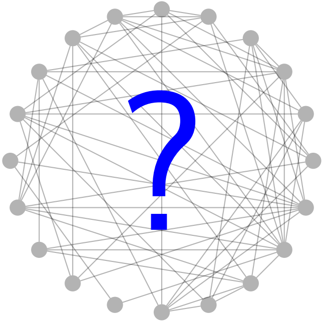


Directed Graph



Downloadable Slides

How do we estimate graph structure?



Downloadable Slides

Gaussian Graphical Models

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

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

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

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

Downloadable Slides

Gaussian Graphical Models

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

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

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

Downloadable Slides

Outline (Redux)

1. Introduction to Graphical Models
2. Estimation for Complete Data
   - Neighborhood Selection
   - Graphical Lasso
   - Further Notes
3. Estimation with Missingness
   - MissGLasso
   - Additional Missing Data Methods

Downloadable Slides

Preface

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

Downloadable Slides

Neighborhood Selection

- Note that $E\backslash\{\text{ne}(X_i)\}$ includes all nodes independent of X_i conditional upon $\text{ne}(X_i)$
- Proposed by Meinshausen & Bühlmann (2006) [10], for $X \in \mathbb{R}^d$ concern yourself only with $(\Theta)_{ij} = 0$, or $(\Theta)_{ij} \neq 0$
- Assume sparsity of Θ and fit d, element-wise lasso models
  - Regress $X_i \overset{\text{Lasso}}{\sim} X_1 + ...X_{i-1} + X_{i+1} + ... + X_d$ for all $i \in [d]$
  - Take $\hat{\beta}_{(-i)} \in \mathbb{R}^{d-1}$ from each model
  - Conclude[4] $(\Theta)_{ij} = 0 \Leftrightarrow \hat{\beta}_{(-i)j} = 0 \wedge \hat{\beta}_{(-j)i} = 0$
- Admits asymptotic consistency for "zero-selection" of Θ

[4] 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 [4]
- $\hat{\Theta}$ is *not* guaranteed to be positive semi-definite
- *Requires* sparsity assumptions for theoretical guarantees:
  - $\exists \kappa \in (0,1),\ \max_{a \in V} |\text{ne}(a)| = O\left(n^{\kappa}\right)$
  - For any conected nodes a, b (i.e.$\forall (a,b) \in E$), $\left|\left|\theta^{a,\text{ne}(b) \setminus \{a\}}\right|\right|_1 \leq \vartheta < \infty$

Graphical Lasso

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

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

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! [4]

So we assume sparsity and apply the ℓ_1 penalty

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

Downloadable Slides

Graphical Lasso - Algorithm[5]

```
Initialize $\mathbf{W} \leftarrow S + \lambda \mathbf{I}$
for $j = 1, 2, ..., d, 1, 2...$ until convergence do:
    Partition $\mathbf{W} = \begin{bmatrix} \mathbf{W}_{kk} & \mathbf{w}_{kj} \\ \mathbf{w}_{jk} & w_{jj} \end{bmatrix}$
    Solve estimating equations $\mathbf{W}_{kk}\beta - \mathbf{s}_{jk} + \lambda \text{sign}\beta = 0$
    Update $\mathbf{w_{kj}} = \mathbf{W}_{kk}\hat{\beta}$
for $j = 1, 2, ..., d$ final update: do
    Solve for $\hat{\theta}_{kj} = -\hat{\beta} \cdot \hat{\theta}_{jj}$ where $1/\hat{\theta}jj = 2_{jj} - \mathbf{w}_{kj}^T\hat{\beta}$
```

[5] Pseudocode adopted from Elements of Statistical Learning, Chapter 17 [6] and Statistical Learning with Sparsity, Chapter 9 [7]

Graphical Lasso - Algorithm

- 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
- Replacing $\mathbf{W}_{kk}$ with $\mathbf{S}_{kk}$ iterates once and returns the neighborhood selection algorithm!

What does this give us?

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

Downloadable Slides

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

Downloadable Slides

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 (accuracacy by proportion of correct edge recovery)
- Operator Norm Distance $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_2 \underbrace{\lesssim \sqrt{\frac{D_{max}^2 \log d}{N}}}_{\text{for GLasso}}$

Downloadable Slides

Simulations (Set-Up)

- Generated multivariate normal data for $d = \{64, 128, 256\}$ with AR(n,ρ) adjacency structures

- Assessed edge-selection performance for N ranging from 10 to 2000
  - TPR = (# of true edges selected) / (# of true edges)
  - TNR = (# of true non-edges not selected) / (# of true non-edges)
  - Operator norm $||\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}||_2$

Downloadable Slides

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)

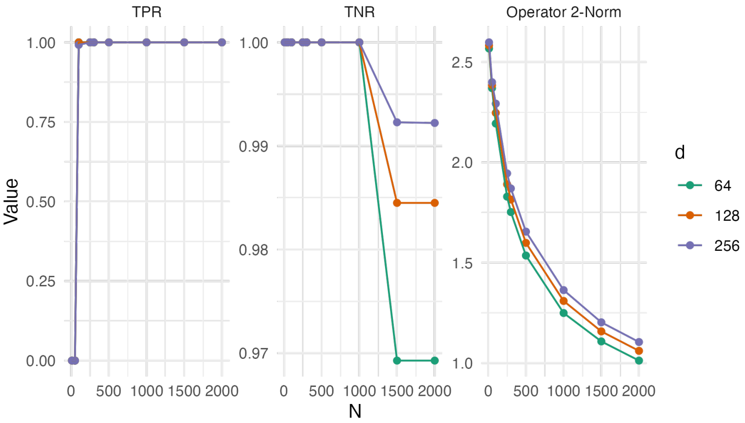


Figure: AR(1), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (GLASSO - Complete Data)

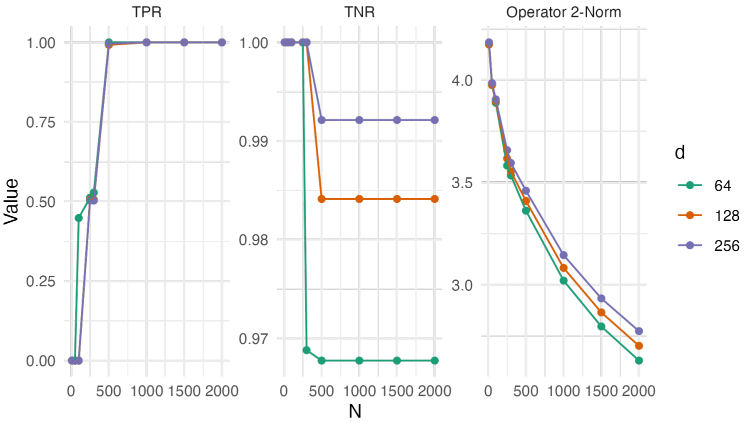


Figure: AR(2), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (GLASSO - Complete Data)

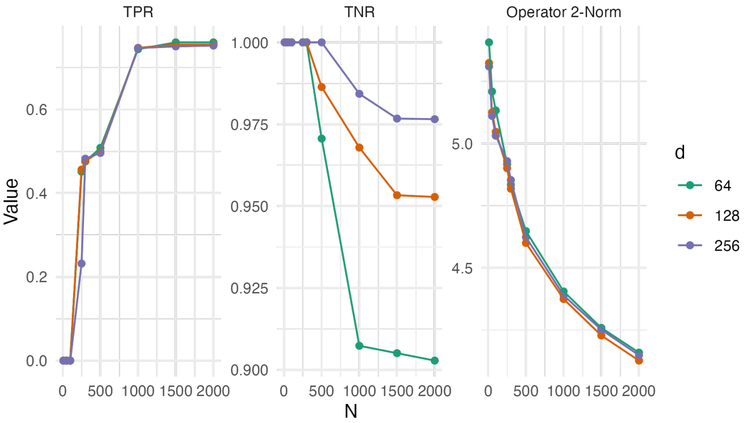


Figure: AR(4), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (GLASSO - Complete Data)

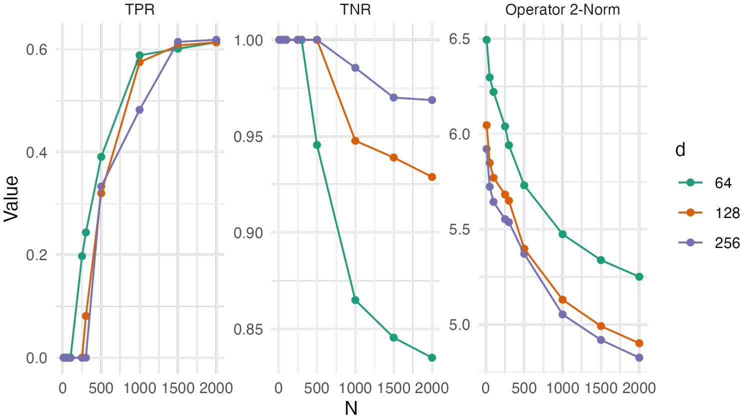


Figure: AR(8), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (Neighborhood Selection - Complete Data)

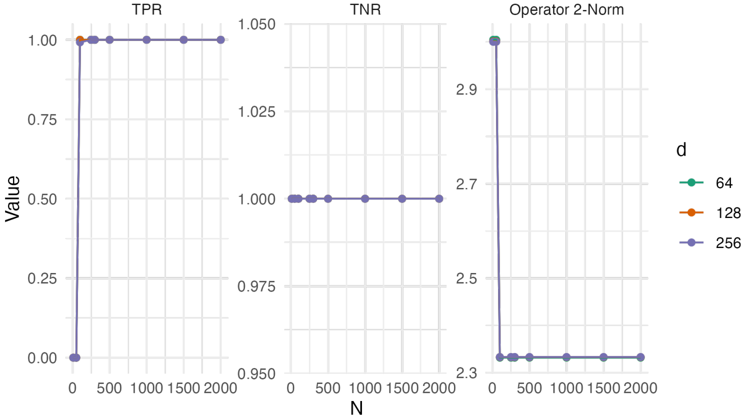


Figure: AR(1), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (Neighborhood Selection - Complete Data)

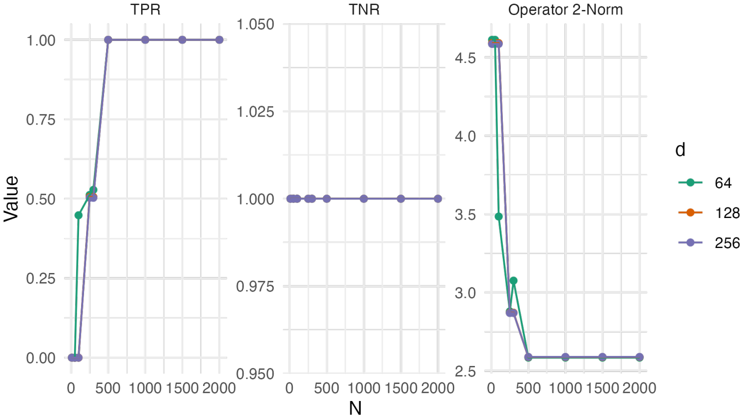


Figure: AR(2), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (Neighborhood Selection - Complete Data)

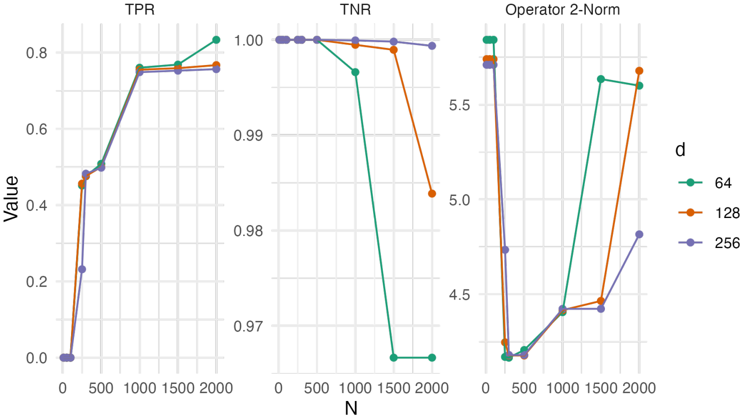


Figure: AR(4), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (Neighborhood Selection - Complete Data)

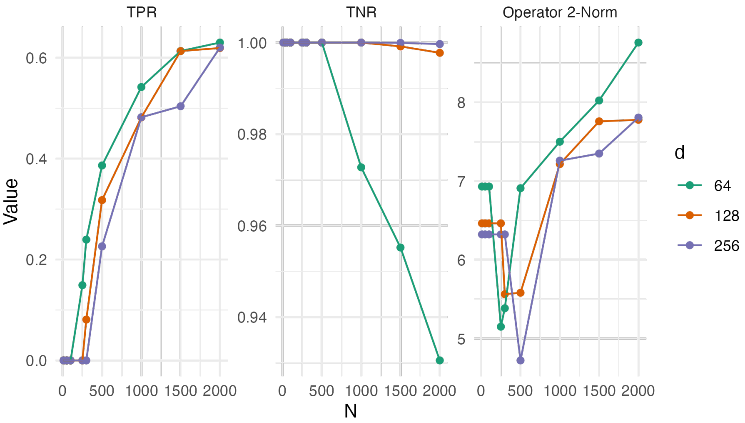


Figure: AR(8), $\rho = 0.4$ adjacency structure

Downloadable Slides

Further Notes

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

[6] These results are also slightly altered by choice of ρ, our value $\rho = 0.3$ was chosen arbitrarily and kept constant only for brevity

Downloadable Slides

Outline (Redux)

1. Introduction to Graphical Models
2. Estimation for Complete Data
   - Neighborhood Selection
   - Graphical Lasso
   - Further Notes
3. Estimation with Missingness
   - MissGLasso
   - Additional Missing Data Methods

Downloadable Slides

Motivation

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

Downloadable Slides

MissGLasso

- Suppose for $X^{(i)} \sim N(\mu, \Sigma)$, we can partition $X = (X_o^{(i)}, X_m^{(i)})$
- Consider all observed data $\mathbf{X}_o = (X_o^{(1)}, X_o^{(2)}, ..., X_o^{(N)})$ and maximize the resulting likelihood (with penalty):

$$\hat{\Theta} = argmax_{\Theta \succ 0} \ell_\lambda(\Sigma; \mathbf{X}_o) = argmin_{\Theta \succ 0} - \ell_\lambda(\Sigma; \mathbf{X}_o)$$

$$-\ell_\lambda(\Sigma; \mathbf{X}_o) = -\sum_{i=1}^{N} \frac{1}{2} \log\det\Theta_o + \frac{1}{2}(X_o^{(i)} - \mu_o)^T \Theta_o (X_o^{(i)} - \mu_o) + \lambda ||\Theta||_1$$

- Despite similar structure to our typical Gaussian log-likelihood, because of missingness, **this is not necessarily convex**!

Downloadable Slides

MissGLasso - Algorithm

- We can consider our unobserved $\mathbf{X}_m$ latent, and implement EM!

$$-\ell_\lambda(\Sigma; \mathbf{X}) = -\frac{n}{2}\log\det\Theta + \frac{n}{2}\mu^T\Theta\mu - \mu^T\Theta\underbrace{X^T\mathbf{1}}_{\mathbf{T}_1} + \frac{1}{2}\mathrm{tr}(\Theta\underbrace{X^TX}_{\mathbf{T}_2}) + \lambda||\Theta||_1$$

- $\mathbb{E}(\ell_\lambda|X_o, \mu, \Theta)$ is only random in $\mathbf{T_1}, \mathbf{T_2}$
- The E-step reduces to $\mathbb{E}(\mathbf{T}_1|X_o, \mu, \Theta)$ and $\mathbb{E}(\mathbf{T}_2|X_o, \mu, \Theta)$
  - More simply, $\mathbb{E}(x_k^{(i)}|X_o^{(i)}, \mu, \Theta)$ and $\mathbb{E}(x_j^{(i)}x_k^{(i)}|X_o^{(i)}, \mu, \Theta)$, $j, k = 1, ..., d; i = 1, ..., N$

Downloadable Slides

MissGLasso - Algorithm

Moreover, by conditional properties of the Gaussian, for the γth iteration, we have simple forms for these expectations!

$$\text{Let } c = \mu_m^{(\gamma)} - (\Theta_{mm}^{(-1)})^{(\gamma)} \Theta_{mo}^{(\gamma)} (X_o - \mu_o)$$

$$\mathbf{T}_1^{(\gamma+1)} = \mathbb{E}(x_j | X_o, \mu, \Theta) = \begin{cases} x_j, & \text{if } x_j \text{ observed} \\ c_j & \textit{otherwise} \end{cases}$$

$$\mathbf{T}_2^{(\gamma+1)} = \mathbb{E}(x_j x_k | X_o, \mu, \Theta) = \begin{cases} x_j x_k, & \text{if } x_j, x_k \text{ observed} \\ x_j c_k, & \text{if only } x_j \text{ observed} \\ \left(\Theta_{mm}^{(\gamma)}\right)^{-1}_{jk} + c_j c_k, & \text{if } x_j, x_k \text{ missing} \end{cases}$$

Downloadable Slides

MissGLasso - Algorithm

M-step also follows from first-order optimality conditions on the log-likelihood with these nice expectations:

$$\mu^{(\gamma+1)} = \frac{1}{n}\mathbf{T}_1^{(\gamma+1)}$$

$$\Theta^{(\gamma+1)} = argmin_{\Theta \succ 0}\Big\{ -\log\det\Theta + \mathrm{tr}\left(\Theta\left[\frac{1}{n}\mathbf{T}_2^{(\gamma+1)} - \mu^{(\gamma+1)}(\mu^{(\gamma+1)})^T\right]\right) + \frac{2\lambda}{n}||\Theta||_1\Big\}$$

Downloadable Slides

MissGLasso - Algorithm

1. **Estimate the expectations of sufficient statistics over complete data**
    - **Conditional Gaussian properties give simple, closed forms**
2. **Maximize likelihood (GLasso) and update parameter estimates by optimality conditions**

Downloadable Slides

MissGlasso - Notes

- Fairly intuitive implementation which allows us to use all data, including rarely observed nodes
- Outperforms complete case analysis and mean-imputation in graph selection
- Assumes MAR and empirically underperforms in simulations with MNAR data

What is the state of methods for partially-observed, MNAR graphs?

Censored Data

- Methods for censored data included in `cglasso` using recent papers from the package authors [1, 2]
- Include an extension of the MissGLasso EM algorithm for truncated Gaussian distribution
- Clearly extends beyond the MAR framework but requires strict, known missing data mechanism (e.g. detection limits)

Erose Data

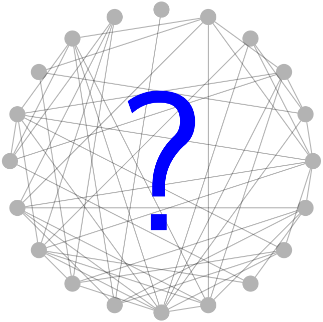
- *Erose* data is a term coined by Zheng, Allen (2023) for data with irregular availability [13]
    - Leads to "drastically different" sample size for a small subset of nodes
    - Erose data almost certainly violate MAR/MCAR assumptions of existing methods
    - Motivated by neuroscience but with applications in genetic expression data,
- Authors propose the **Graph Inference when Joint Observations are Erose** (GI-JOE) method for edge-wise inference in erosely observed edges
    - Adaptation of debiased neighborhood selection algorithm
    - Does not require imputation of missing values

Concluding Notes

- Graphs are a powerful representation of your multivariate data **(intuitively and algorithmically)**
- These extensions tend to distill to regularized M-estimation problems, an area with great theoretical contributions and guarantees
- Important to consider sparsity/degree scaling of your graph
  - Early methods quickly recover graph structure, even when $N \ll d$
- `huge, cglasso` packages are poweful implementations in R

Downloadable Slides

Graphs, how and why (revisited)?



$\longrightarrow$ Regularized M-estimation(+)!

→ Consistency

→ Fast computation even for $p \gg N$

→ MLE via GLasso

→ EM Extensions for MAR

Downloadable Slides

References I

- Some diagrams generated in conjunction with ChatGPT 3.5

[1] Luigi Augugliaro, Antonino Abbruzzo, and Veronica Vinciotti. "L1 -Penalized censored Gaussian graphical model". en. In: *Biostatistics* (Sept. 2018).

[2] Luigi Augugliaro, Gianluca Sottile, and Veronica Vinciotti. "The conditional censored graphical lasso estimator". en. In: *Statistics and Computing* 30.5 (Sept. 2020), pp. 1273–1289.

[3] Luigi Augugliaro et al. *cglasso: Conditional Graphical LASSO for Gaussian Graphical Models with Censored and Missing Values*. Jan. 2023.

Downloadable Slides

References II

[4] 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.

[5] A. P. Dempster. "Covariance Selection". In: *Biometrics* 28.1 (1972). Publisher: [Wiley, International Biometric Society], pp. 157–175.

[6] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer, 2009.

[7] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: the Lasso and Generalizations*. Dec. 2016.

References III

[8] Imperatorskaia akademia nauk (Russia). *Commentarii Academiae scientiarum imperialis Petropolitanae*. lat. Petropolis, Typis Academiae, 1726.

[9] Rahul Mazumder and Trevor Hastie. *The Graphical Lasso: New Insights and Alternatives*. arXiv:1111.5479 [cs, stat]. Aug. 2012.

[10] 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.

[11] 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 IV

[12] 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.

[13] Lili Zheng. *GI-JOE: Graph Inference when Joint Observations are Erose*. Mar. 2023.

Appendix Slides

Downloadable Slides

Erdös Rényi Graph Results (Complete Data)

Downloadable Slides

Simulations (GLASSO - Complete Data)

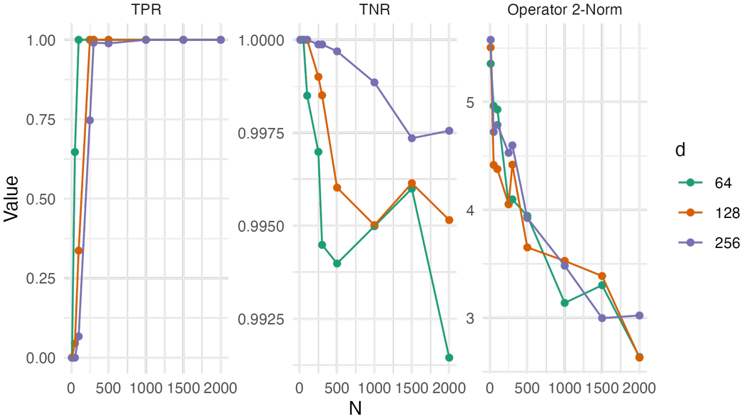


Figure: ER(p=0.01), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (GLASSO - Complete Data)

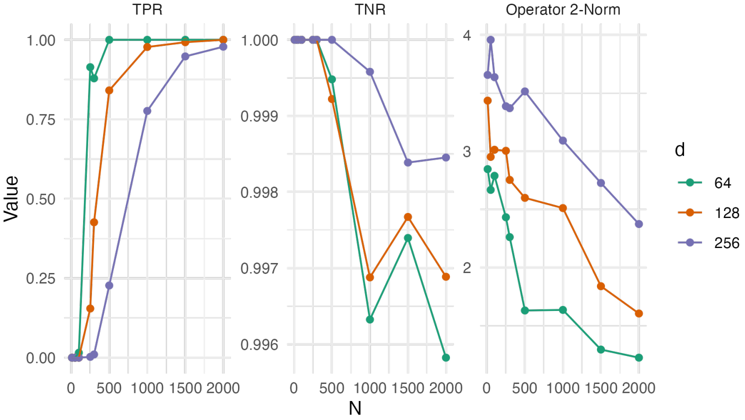


Figure: ER(p=0.05), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (GLASSO - Complete Data)

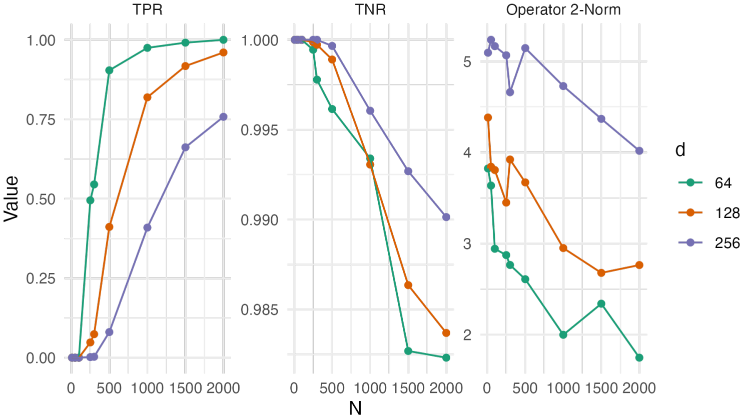


Figure: ER(p=0.1), $\rho = 0.4$ adjacency structure

Downloadable Slides

Simulations (GLASSO - Complete Data)

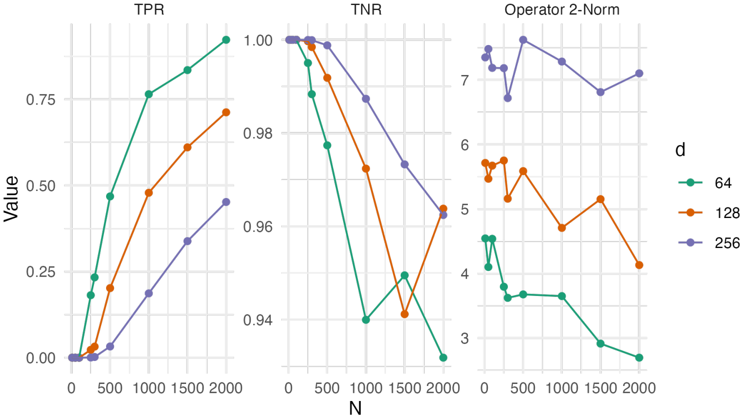


Figure: ER(p=0.2), $\rho = 0.4$ adjacency structure

Downloadable Slides

ER-Simulations (Neighborhood Selection - Complete Data)

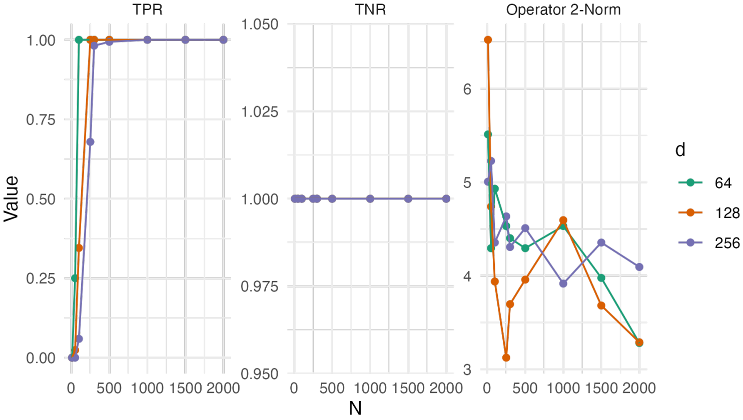


Figure: ER(p=0.01), $\rho = 0.4$ adjacency structure

Downloadable Slides

ER-Simulations (Neighborhood Selection - Complete Data)

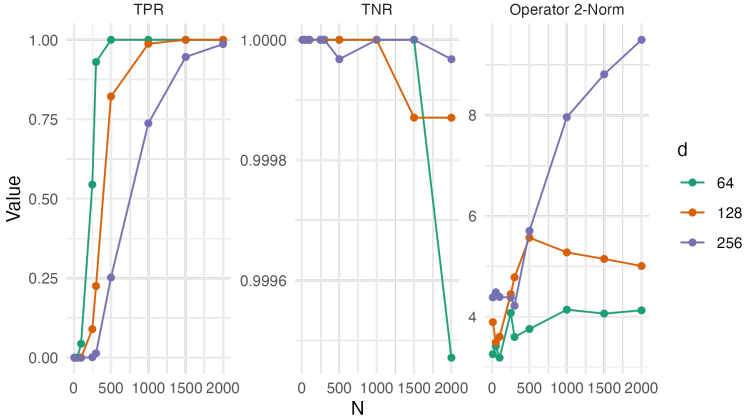


Figure: ER(p=0.05), $\rho = 0.4$ adjacency structure

Downloadable Slides

ER-Simulations (Neighborhood Selection - Complete Data)

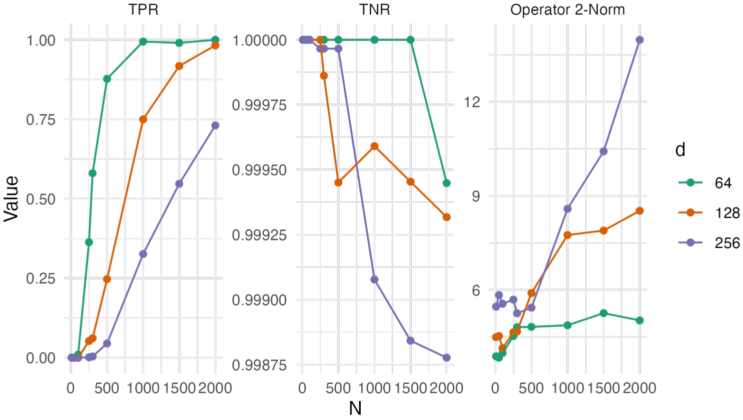


Figure: ER(p=0.1), $\rho = 0.4$ adjacency structure

Downloadable Slides

ER-Simulations (Neighborhood Selection - Complete Data)

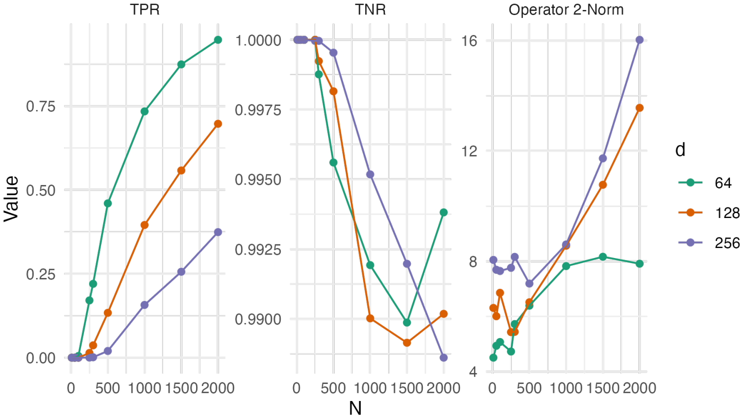


Figure: ER(0.2), $\rho = 0.4$ adjacency structure

Downloadable Slides

Misc. Notes

Conditional Graphical Lasso

- Can consider random vectors Y, X
  - Y is a random vector of interest (i.e. our graph structure)
  - X is a common set of covariates included in all conditioning
- See `cglasso` package [3] for implementation
- Package includes conditional MissGLasso implementation
- Methods are extended for censoring

Downloadable Slides

Forgoing Sparsity Assumptions

- In the above methods, we have almost uniformly assumed some sparsity and applied a penalty (ℓ_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 AR($\cdot$) extension sims?
- Mazumder (2012) [9] offers an updated algorithm and insight into performance for p close to but larger than N
- Interplay between d, N, and graph-connectedeness affect computation time and convergence

Downloadable Slides

Time-Series Data

- Consider that our repeated observations are time-indexed:
  - $\{X_j(t), t \in \mathcal{T}, j = 1, ..., N\}, X_j \in \mathbb{R}^d$
- 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

Downloadable Slides

Inference with Debiased Lasso

- The typical lasso estimator $\hat{\beta}_\lambda = \text{argmin}\beta||Y - X\beta||_2^2 + \lambda||\beta||_1$ 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 ℓ_1 penalization?
- Debiased neighborhood selection is partial motivation for Zheng, Allen's GI-JOE paper [13]

Downloadable Slides

"Nothing new under the sun"

My (likely useless and certaintly 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!

Downloadable Slides