Notes (To Be Deleted)

- Maybe scrap nhood selection for time
- Be able to speak to glasso over neighborhood selection, how sparsity guarantees MLE existence
- Analyses/slides for missingness TBD

Graphical Models

With a focus towards interimly missing data

Dominic DiSanto

Department of Biostatistics, Harvard University

December 2, 2023

Downloadable Slides

Outline

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)

Graph Theory Origins [5, 8]

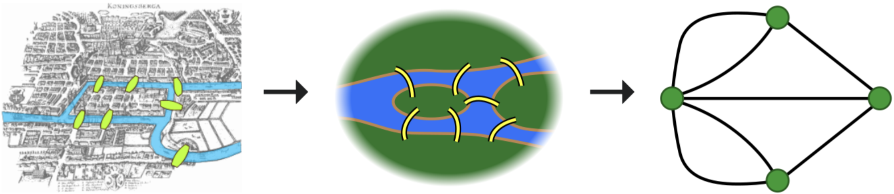


Figure: Euler's Bridges Conceptualization (Recreation)

[1]

[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 atomic spin
- Statistical "beginnings"
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [3]
- 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 atomic spin
- Statistical "beginnings"
- Arthur Dempster (founding Harvard Stats professor) introduced covariate selection by precision matrix estimation in 1972 [3]
- 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 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 [3]
- Judea Pearl ∼1980's for causal intepretation of Bayesian networks
- Modern interest in related regularized M-estimation problems and graphical neural networks

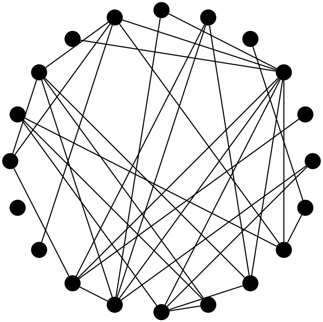
Downloadable Slides

[2] Early use in physics were probabilistic, but this may be seen as an early "purse 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[a])
- Time-series data



[a] Electronic Healthcare Record Large Language Model

Graphs

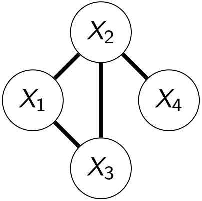
- Graphs are a natural way to represent interrelationships among our data!
- Present nice properties for estimation of joint distributions
  - Can avail existing graphical algorithms
  - Ability to characterize conditional (in)dependencies
- Probabilistic graphical modelling provide a general formalism 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][3]

[3] See 2.4 specifically for applications

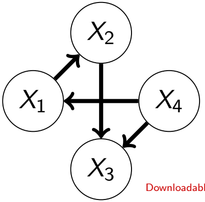
Graphs

- 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

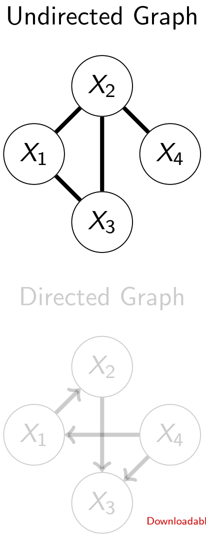
Notation/Nomenclature

Omitting some philosophical discrepancies

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

Graphs

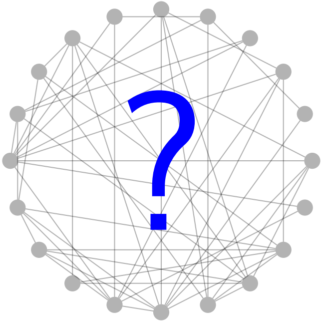
- Consider random vector $X \sim N(\boldsymbol{\mu}, \Sigma)$ and precision matrix $\Theta \equiv \Sigma^{-1}$
  - Interested in estimating Σ to characterize joint distribution f_X
- Can construct a resulting graph $\mathcal{G} = (V, E)$, $V = X, E \subseteq V \times V$
  - Let $\text{ne}(x)$ represent 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



Gaussian Graphical Models

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

How do we estimate graph structure?



Outline (Redux)

Downloadable Slides

Preface

- $X = (X_1, ..., X_d)$ may or may not contain a "response variable" of interest EDIT do you do anything differently if so?
- Identifying conditional relationships $\Leftrightarrow$ Estimating/Identifying 0's in Σ
- Enforse sparsity for $\hat{\Theta}$ to account for possible rank-degeneracy of S if $d \gg N$
- Exhaustive search is quickly infeasible, but can yield solutions with theoretical guarantees
  - Consistency
  - Nice statistical rates

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) [7], 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 [1]
- $\hat{\Theta}$ is *not* guaranteed to be positive semi-definite
- *Requires* sparsity assumptions for theoretical guarantees:
  - $\exists \kappa, \ \max_{a \in V} |\text{ne}(a)| = O\left(n^{\kappa}\right)$
  - 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?

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

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! Prove, 9.2c in SLS

Graphical Lasso

So we assume sparsity and apply the ℓ_1 penalty

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

What does this give us?

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

Graphical Lasso - Optimization

Block coordinate optimiaztion here?

Simulations (Complete Data)

- `glasso` package in R can fit Graphical Lasso as well as neighborhood-selection approximation
- `huge` is an extension of `glasso` with algorithmic/convergence fixes, computation in C, and extended methods
- sklearn has similar `sklearn.covariance.graphicallasso` command
- `skggm` extends Gaussian Graphical Model methods

Simulations (Complete Data)

- $\lambda = 2\sqrt{\frac{\log d}{N}}$
- Graph Recovery (accuracacy by proportion of correct edge recovery)
- Operator Norm Distance $||\hat{\mathbf{\Theta}} - \mathbf{\Theta}||_2 \lesssim \sqrt{\frac{D_{max}^2 \log d}{N}}$

Simulations (Complete Data)

Graph Structure:
banded/AR(.band), random/ER

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

Simulations (Complete Data)

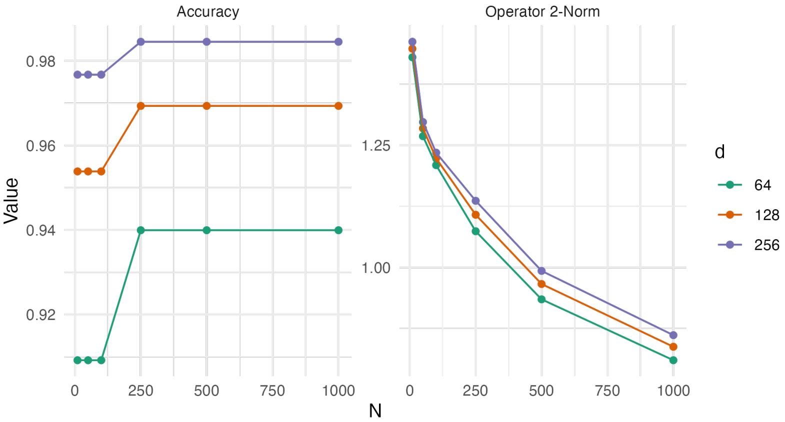


Figure: AR(3), $\rho = 0.3$ adjacency structure

Simulations (Complete Data)

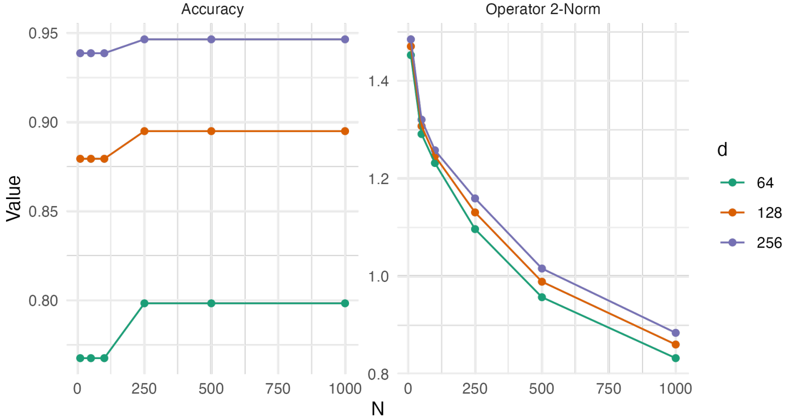


Figure: AR(8), $\rho = 0.3$ adjacency structure

Further Notes

- Neighborhood Selection maximizes d conditional likelihoods
- GLasso maximizes an ℓ_1 penalized, joint (Gaussian) likelihood
- (Highly non-trivial) extensions beyond Gaussianity exist:
  - REVIEW/ADD/EDIT be able to speak in some depth
  - Heterogenous graphs (specifically mixture of Gaussian, Exponential, Poisson, Binomial nodes) [2]
  - Non-parametric extensions [4]

Outline (Redux)

Motivation

- Methods above largely assume complete data[5]
- Networks change, measurement availability (and quality) varies
- Measurement is also often differential between nodes

Visualization here

[5] In the Gaussian setting, hidden/unobserved nodes or missingness can be mean-imputed, but this is fairly näive/ad-hoc procedure

Naïve Methods with Missingness

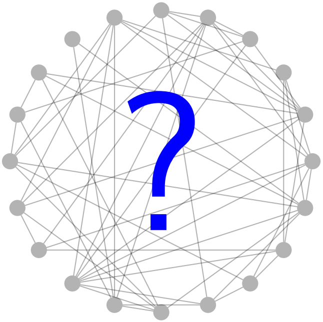
Results with nhood, GLasso in grahpicswith missingness

Graphical Methods for Missingness

Erose Data

- Cite [10]

Graphs, how and why (revisited)?



$\longrightarrow$ 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[6] to distill to regularized M-estimation problems, an area with great theoretical contributions and guarantees
- Extensions beyond Gaussianity substantially increase complexity

[6] Under a high-dimensiona/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] Shizhe Chen, Daniela M. Witten, and Ali Shojaie. "Selection and estimation for mixed graphical models". In: *Biometrika* 102.1 (Mar. 2015), pp. 47–64.

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

[4] Hao Dong and Yuedong Wang. "Nonparametric Neighborhood Selection in Graphical Models". en. In: *Journal of Machine Learning Research* 23 (2022), pp. 1–26.

References II

[5] Imperatorskaia akademia 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 Erose*. Mar. 2023.

Appendix 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?
- 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-connectedeness affect computation time and convergence

Time-Series Data

- Consider that our repeated observations are time-indexed:
  - $\{X_j(t), t \in \mathcal{T}, j = 1, ..., N\}, X_j \in \mathbb{R}^d$
- 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 $\hat{\beta}_\lambda = \text{argmin}\beta||Y - X\beta||_2^2 + \lambda||\beta||_1$ 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?

"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