Estimation of Metabolomic Networks with Gaussian Graphical Models

Katherine H. Shutta[1], Subhajit Naskar[2], Kathryn M. Rexrode[3,4], Denise M. Scholtens[5], Raji Balasubramanian[1]

[1]Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, MA, USA [2]College of Information and Computer Sciences, University of Massachusetts Amherst, Amherst, MA, USA
[3]Division of Women's Health, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA, [4]Harvard Medical School, Boston, MA, USA, [5]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
Correspondence: kshutta@umass.edu

Background and Motivation

- Network-based omics analyses have high potential to capture signatures of complex biological processes by providing a perspective for understanding how genes, proteins, or metabolites are associated in a particular system.
- Gaussian graphical model (GGM) estimation, also known as partial correlation network estimation, is one approach to such analyses.
- In the last decade, many open-source R packages have been published for GGM estimation, each containing one or more methods for this purpose. Choosing a method typically involves making several choices with regard to scoring criteria and estimation algorithms.
- The estimated GGM may be highly sensitive to these choices, and the relative effectiveness of each method may depend on structural characteristics of the underlying network.
- Here, we compare the performance of several methods within these packages across a variety of simulated network data, capturing a range of network topologies.

Interpreting GGMs

- GGMs begin with the assumption that observed data are a p-dimensional random vector following a multivariate normal distribution with some covariance matrix Σ. The inverse covariance matrix, also referred to as the precision matrix, is typically denoted as $\Theta = \Sigma^{-1}$.
- In a GGM for metabolomics data, an edge between two nodes (metabolites) corresponds to conditional dependence between the two metabolites conditioned on the rest of the metabolites in the network. An edge therefore indicates that two metabolites have an association that cannot be explained through other metabolites in the network.
- It can be shown that conditional dependence corresponds to nonzero entries of the inverse covariance, or precision, matrix [1]. Estimating a GGM is therefore equivalent to estimating the inverse covariance matrix Θ.

R Packages Applied

- The table below contains the methods used in our simulation study. Regularized methods can be used in the $n \geq p$ case; others cannot.

Package	Methods	Regularized
huge[2]	glasso - eBIC[3]	Yes
huge	glasso - RIC[4]	Yes
huge	glasso - StARS[5]	Yes
hglasso[6]	hglasso	Yes
bootnet[7]	eBIC	Yes
bootnet	pcor[8, 9, 10]	No
bootnet	ggmModSelect	No

References

[1] Caroline Uhler. Gaussian graphical models: An algebraic and geometric perspective. 2017. URL https://arxiv.org/pdf/1707.04345.pdf.

[2] Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, and Larry Wasserman. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, 13(Apr):1059–1062, 2012.

[3] Rina Foygel and Mathias Drton. Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pages 604–612, 2010.

[4] Shaun Lysen. Permuted inclusion criterion: a variable selection technique. *Publicly accessible Penn Dissertations*, page 28, 2009.

[5] Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Advances in neural information processing systems*, pages 1432–1440, 2010.

[6] KM Tan, P London, K Mohan, SI Lee, M Fazel, and D Witten. Learning graphical models with hubs. *Journal of machine learning research: JMLR*, 15:3297–3331, 2014.

[7] Sacha Epskamp, Denny Borsboom, and Eiko I Fried. Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1):195–212, 2018.

[8] Juliane Schaefer, Rainer Opgen-Rhein, V Zuber, M Ahdesmäki, AP Duarte Silva, and K Strimmer. corpcor: Efficient estimation of covariance and (partial) correlation. *R package version*, 1(6), 2013.

[9] Juliane Schäfer and Korbinian Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[10] Rainer Opgen-Rhein and Korbinian Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[11] Gabor Csardi, Tamas Nepusz, et al. The igraph software package for complex network research. *InterJournal, complex systems*, 1695 (5):1–9, 2006.

[12] Paul Erdős and Alfréd Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[13] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440, 1998.

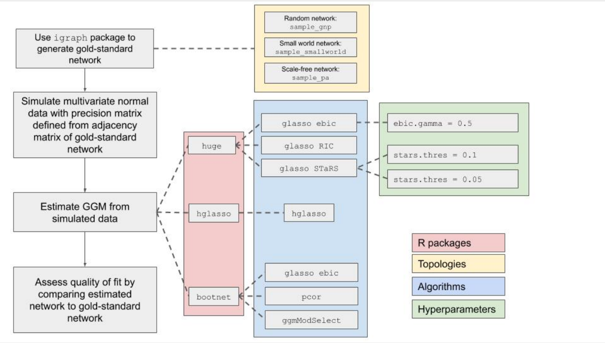
[14] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[15] William E Kraus, Christopher B Granger, Michael H Sketch, Mark P Donahue, Geoffrey S Ginsburg, Elizabeth R Hauser, Carol Haynes, L Kristin Newby, Melissa Hurdle, Z Elaine Dowdy, et al. A guide for a cardiovascular genomics biorepository: the cathgen experience. *Journal of cardiovascular translational research*, 8(8):449–457, 2015.

Acknowledgements

Research reported in this poster was supported by the National Institutes of Health under award number 1R01HL122241-01A1.

Study Design



Use igraph package to generate gold-standard network

Simulate multivariate normal data with precision matrix defined from adjacency matrix of gold-standard network

Estimate GGM from simulated data

Assess quality of fit by comparing estimated network to gold-standard network

Random network: sample_gnp

Small world network: sample_smallworld

Scale-free network: sample_pa

huge

hglasso

bootnet

glasso ebic

glasso RIC

glasso STaRS

hglasso

glasso ebic

pcor

ggmModSelect

ebic.gamma = 0.5

stars.thres = 0.1

stars.thres = 0.05

R packages

Topologies

Algorithms

Hyperparameters

Creating Gold-Standard Network Structures

- We created six gold-standard network structures to simulate multivariate normal data (**Figure 1**).
- Three different network topologies (random, small world, and scale free) were considered along with two density levels (high: approximately 6% dense, low: approximately 2% dense).
- The igraph package in R was used to create these structures [11].
  - The sample_gnp function was used to generate Erdos-Renyi random networks [12]
  - The sample_smallworld function was used for small world networks [13]
  - The sample_pa function was used for Barabasi-Albert scale free networks [14].
- Networks have 100 nodes (metabolites).

Simulated Network Structures

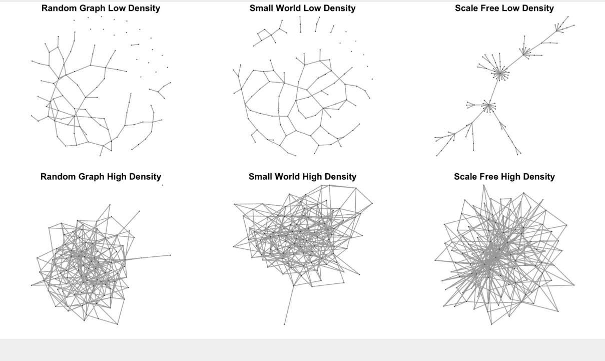


Figure 1: Six different gold-standard networks used to generate multivariate normal data for simulation studies.

Performance Metrics

- **Frobenius norm of error matrix**
  - The error matrix was computed as the difference between the adjacency matrix of the estimated GGM and the adjacency matrix of the true GGM.
  - The Frobenius norm (the square root of the sum of squared elements) of the error matrix was used as a measure of fit.
  - A small Frobenius norm indicates good model performance in the sense that the estimated network is close to the gold-standard network overall.
- **True positive rate (TPR) and false positive rate (FPR)**
  - Positive edges were defined as those edges corresponding to partial correlations greater in magnitude than the Fisher threshold at level $\alpha = 0.05$; negative edges were defined as those partial correlations less in magnitude than this threshold.
  - The Fisher threshold for the $n = 1000$, $p = 100$ case was 0.062. For the $n = 50$, $p = 100$ case, it was 0.28.
  - Good model performance is indicated by a low FPR with a high TPR, indicating that the method has good sensitivity and specificity.

Performance of Network Estimation Methods, $p < n$ case

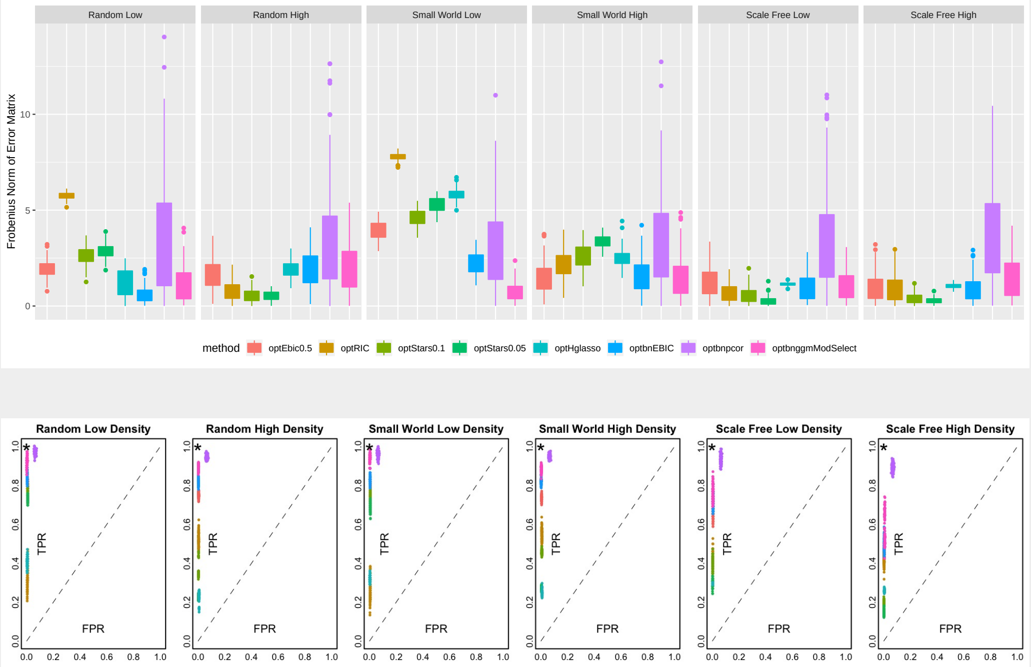


Figure 2: Simulation study results for $n = 1000$ and $p = 100$. Top: boxplots of the Frobenius norm of the difference between the true and estimated networks. Bottom: scatterplot of true positive rate vs. false positive rate, where "positive" is defined as a partial correlation greater in magnitude than 0.062, which is the critical value for significance level $\alpha = 0.05$. The star in the upper left corner indicates the performance of a perfect edge selection method. The dashed line is the performance expected from a "coin flip" edge selection method.

Performance of Network Estimation Methods, $p > n$ case

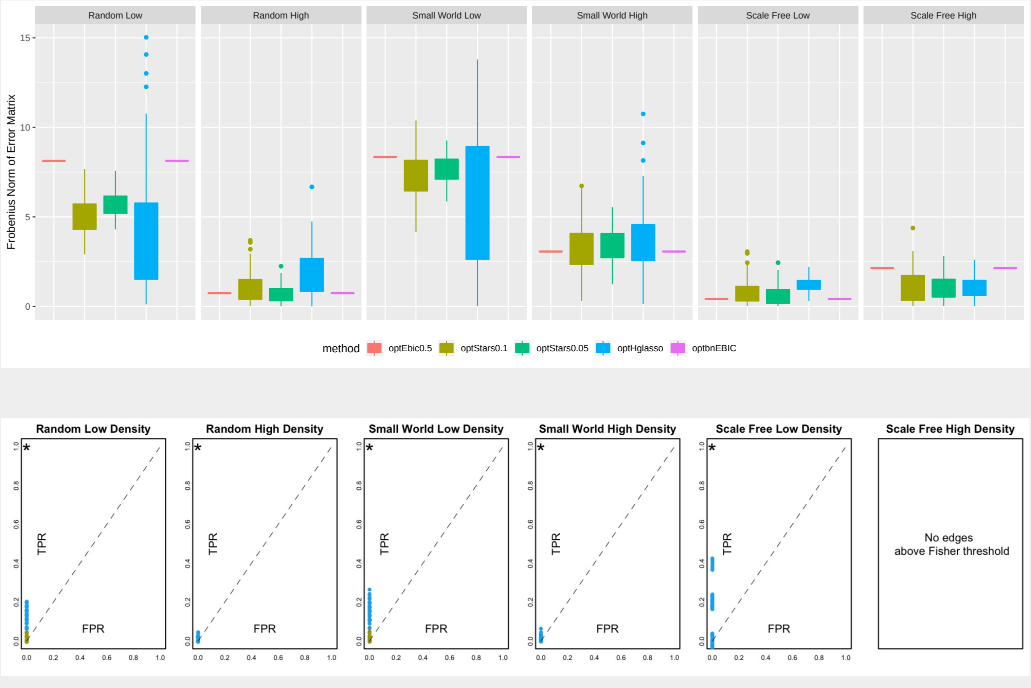


Figure 3: Simulation study results for $n = 50$ and $p = 100$. Top: boxplots of the Frobenius norm of the difference between the true and estimated networks. Trivial boxplots (flat lines) correspond to methods which selected an empty network in every iteration. Bottom: scatterplot of true positive rate vs. false positive rate, where "positive" is defined as a partial correlation greater in magnitude than 0.28, which is the critical value for significance level $\alpha = 0.05$. Only estimated networks with at least one positive by this definition are shown, i.e., only the StARS method with threshold 0.1 and the hub graphical lasso recovered edges with magnitude > 0.28. Note that no edge in the gold-standard network for the high-density scale free network was significant according to this threshold in either the gold-standard or estimated networks, so TPR and FPR are not shown.

Simulation Results

- **Low density vs. high density networks**
  - Density does not appear to affect estimation as much in the $p < n$ case as it does in the $p > n$ case. In the $p > n$ case, estimation appears slightly worse in the low-density case than the high-density case for the random and small world settings; in the scale-free setting, results are similar for both densities.
- **Random vs. small world vs. scale-free networks**
  - Permutation and subsampling selection criteria (RIC, STaRS) perform well relative to other methods in random and scale-free networks, but perform worse in small world networks.
  - This may be due to permutation or subsampling approaches failing to preserve the "shortcuts" that are characteristic in creating the short average path length that is characteristic of a small-world topology.
  - The hub graphical lasso (hglasso [6]) is designed to perform well in networks expected to have hubs. We see it is far less variable than other methods in the scale-free setting, and has comparable performance to other algorithms in the low-dimensional case (**Figure 2**). In the high-dimensional case, variability of hglasso results is larger, but the method generally outperforms all other methods in terms of sensitivity and specificity (**Figure 3**).

Application

- To demonstrate the variation in estimated networks in a real metabolomics setting, we estimated GGMs on metabolomics profiles from a cardiovascular disease metabolomics study nested within the CATHGEN biorepository. The CATHGEN biorepository consists of samples collected from 9334 consenting individuals who underwent cardiac catheterization at Duke University Hospital between 2001 and 2010, with annual follow-up visits [15]. One goal of assembling the biorepository was to collect molecular data that could be used to identify biomarkers of cardiovascular disease.
- 136 participants from CATHGEN were selected for this metabolomics study.
- 407 metabolites were measured for each sample. For an illustrative low-dimensional example, we randomly selected 20 of these metabolites.
- The estimated networks shown below highlight the variation that can be observed from different methods.

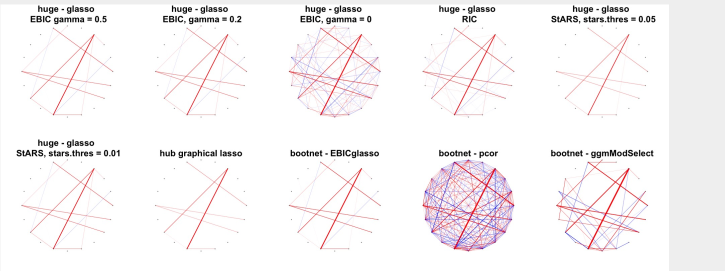


Figure 4: Estimated GGMs for the CATHGEN data, $n = 136$, $p = 20$. Red edges correspond to positive partial correlations; blue, negative. Edge width is proportional to magnitude of partial correlation.

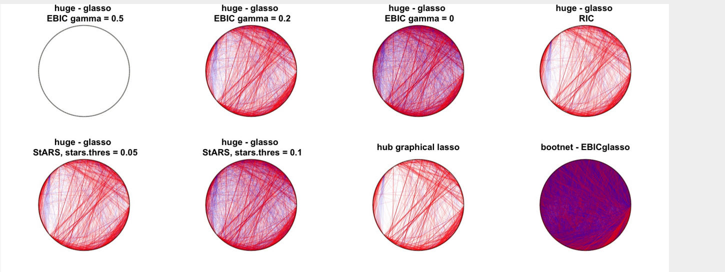


Figure 5: Estimated GGMs for CATHGEN data, $n = 136$, $p = 407$. Edges are interpreted as in Figure 4.

Guidance for Researchers Using GGMs

- These results demonstrate that GGM estimation results vary based on choices such as method and tuning parameters (which the user can control) and true underlying topology (which is *a priori* unknown).
- Exploring multiple methods and using approaches such as bootstrapping should therefore be an important part of GGM estimation. The bootnet package provides a convenient framework for such bootstrap analysis [7].
- There are many diverse packages available for GGM estimation in R. Here, we have explored a subset of the available options that were easiest to access. All packages used in this analysis were installed directly from CRAN.
- The igraph package provides an easy-to-use interface for constructing networks of specified topology for the purpose of simulation studies like the work shown here [11].