# Estimation of Metabolomic Networks with Gaussian Graphical Models

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

*Joint [1]University of Massachusetts, Amherst [2]Brigham and Women's Hospital, Harvard Medical School

[3]Northwestern University Feinberg School of Medicine.

UMASS AMHERST | School of Public Health and Health Sciences

## Abstract

- Network-based metabolomic analyses have high potential to capture signatures of complex biological processes [1].
- Gaussian graphical model (GGM) estimation is one approach to network estimation. Recently, several open-source R packages have been developed for this purpose [2, 3].
- GGM estimation involves several choices with regard to scoring criteria, precision matrix estimation algorithms, and data transformations.
- We present results from a simulation study designed to investigate these choices, with the goal of providing practical guidance to researchers applying GGM approaches to metabolomic data.

## GGM vs Correlation Network

A Gaussian graphical model begins with the assumption an $p$-dimensional random vector of metabolite measurements that follows the multivariate normal distribution [4].

$$\mathbf{X} = (X_1, \ldots, X_p) \sim MVN(\underline{\mu}, \mathbf{\Sigma}) \qquad (1)$$

In this setting, $\underline{\mu} = \underline{0}$ and $\mathbf{\Sigma}$ represents the between-metabolite covariance matrix. Under the MVN assumption, this framework allows us to estimate two different types of networks:

**Correlation network** (Edges from $\mathbf{\Sigma}$)

- $X_i \perp X_j \iff \mathbf{\Sigma}_{i,j} = 0$
- Edges correspond to pairwise dependence
- This marginal dependence may be able to be explained by other metabolites in the network

**GGM network** (Edges from $\mathbf{\Sigma}^{-1}$)

- $X_i \perp X_j | \{X_{k \neq i,j}\} \iff \mathbf{\Sigma^{-1}}_{i,j} = 0$
- Edges correspond to conditional dependence
- This dependence is conditioned on the state of the rest of the network metabolites
- The observed relationship between two metabolites cannot be explained through any of the other metabolites in the network

## Algorithms

- Meinshausen-Bühlmann (mb): uses penalized regression to model each individual metabolite on the others in the network [5]
- Correlation Thresholding (ct): applies a threshold to the correlation matrix
- Graphical LASSO (glasso): uses penalized regression to estimate a sparse inverse covariance matrix [2]

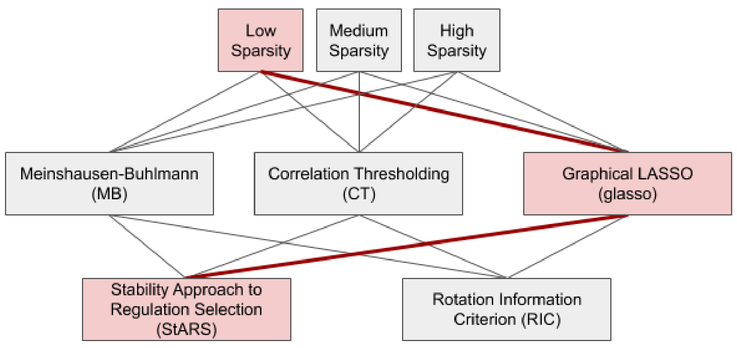
## Scoring Criteria

- Rotation Information Criterion (ric): estimates optimal tuning parameter by permutation-based approach [6]
- Stability Approach to Regularization Selection (StARS): estimates optimal tuning parameter by subsampling approach [7]

## Funding

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

## Simulated Networks



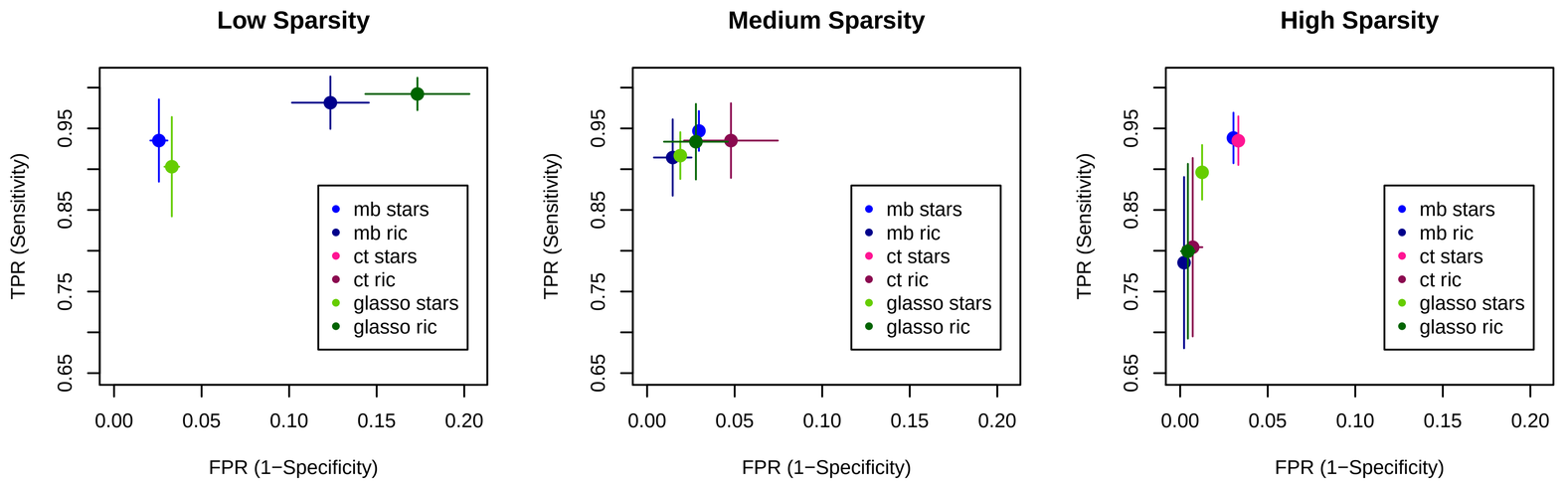

As a gold standard for reference, we generated 3 precision matrices corresponding to random graphs using the Erdos-Renyi random graph generation process in `igraph` [8]. The sparsities modeled were high (edge probability 0.01), medium(0.025), and low(0.1). For chosen simulation settings (e.g., in the highlighted example, a low sparsity matrix estimated with the `glasso` algorithm and scored with StARS), we repeated the following 100 times:

1. Draw 100 samples from the $MVN(\underline{0}, \mathbf{\Sigma})$ distribution
2. Obtain the 400 x 400 sample covariance matrix
3. Apply the chosen algorithm and scoring criterion to obtain an estimated adjacency matrix
4. Compare the estimated adjacency matrix to the gold-standard precision matrix $\mathbf{\Sigma}^{-1}$ from which the data were generated

## Edge Recovery Performance

With three estimation algorithms and two scoring criteria, we studied a total of six network estimation approaches for each sparsity level. To assess the sensitivity and specificity of each algorithm and criterion combination, the following definitions were used (where $\mathbf{\Sigma}^{-1}$ is the gold-standard precision matrix for the simulation):

**True Positive:** an edge in $\mathbf{\Sigma}^{-1}$ with magnitude of conditional correlation $> \rho^* \approx 0.2$ that was detected by the estimation. ($\rho^*$ is the threshold for significance testing of null hypothesis $\rho = 0$ at $\alpha = 0.05$ for a sample of size $n = 100$.)
**True Negative:** an edge in $\mathbf{\Sigma}^{-1}$ with magnitude of conditional correlation exactly 0 that was not detected by the estimation.
**False Positive:** an edge detected in the estimation which has weight exactly 0 in $\mathbf{\Sigma}^{-1}$.
**False Negative:** an edge not detected in the estimation that has absolute weight $> \rho^*$ in $\mathbf{\Sigma}^{-1}$.
Edges in the gold-standard precision matrix with absolute edge weight between 0 and $\rho^*$ were not considered in this analysis.




**Low sparsity network**

- Both CT approaches detected a very small number of edges and are not shown (TPR, FPR $\approx 0$)
- StARS criterion has lower sensitivity and higher specificity than the RIC criterion for both the MB and glasso algorithms

**Medium sparsity network**

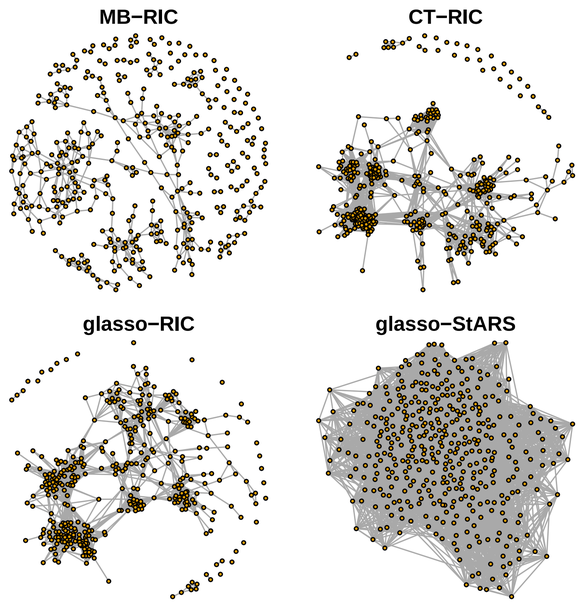
- Performance is comparable among all methods with the exception of CT-StARS
- CT-StARS detects a very small number of edges and is not shown

**High sparsity network**

- StARS criterion has higher sensitivity and lower specificity than RIC
- Difference in criterion has more impact than difference in algorithm

## Application: CATHGEN

We used the three algorithms and two criteria to fit six estimated networks for a dataset of targeted metabolomic data from the CATHGEN Biorepository [9]. The estimated topologies varied depending on choice of algorithm. Not shown are the MB-StARS and CT-StARS estimated networks; almost no edges were estimated for these approaches.




The table below shows the edge count for each approach.

| | MB | CT | glasso |
|---|---|---|---|
| RIC | 610 | 4530 | 2542 |
| StARS | 0 | 1 | 6045 |

## Conclusion

Estimated GGMs can vary broadly depending on method, and this variability may depend on network topology. Cross-validation and sensitivity analyses are recommended.

## References

[1] A Rosato, L Tenori, M Cascante, PR De Atauri Carulla, VAP Martins Dos Santos, and E. Saccenti. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics*, (14(4)):37, 2018.

[2] J Friedman, T Hastie, and R Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, (9(3)):432–41, 2018.

[3] Roeder K Lafferty J Wasserman L Zhao T, Liu H. The huge package for high-dimensional undirected graph estimation in r. *Journal of Machine Learning Research*, (13):1059–1062, 2012.

[4] Caroline Uhler. Gaussian graphical models: An algebraic and geometric perspective. 2017.

[5] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[6] Lysen Shaun. Permuted inclusion criterion: A variable selection technique. *Publicly Accessible Penn Dissertations*, 28, 2009.

[7] Roeder K Liu H and Wasserman L. Stability approach to regularization selection for high dimensional graphical models. *Advances in Neural Information Processing Systems*, 2010.

[8] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[9] WE Kraus, CB Granger, MH SketchJr, MP Donahue, GS Ginsburg, ER Hauser, C Haynes, LK Newby, M Hurdle, ZE Dowdy, and SH Shah. A guide for a cardiovascular genomics biorepository: the cathgen experience. *Journal of Cardiovascular Translational Research*, 8(8):449–57, 2015.