

Understanding Mechanisms of Fetal Inflammatory Response through Graphical Modeling with External Network Data

Antonia George, Irene Villalonga Domínguez, Joule Voelz

Data Science Methodology Master Thesis

Barcelona School of Economics

Advisors: Robert Castelo, David Rossell, Jack Jewson

June 11, 2023

Overview

1. Introduction

2. Background

2.1 Fetal Inflammatory Response (FIR)

3. Data

4. Methods

4.1 Graphical Modeling

4.2 Graphical LASSO

4.3 Network Graphical LASSO

4.4 Network graphical spike-and-slab LASSO

5. Results

6. Conclusions

7. References

Motivation

- Comprehend the underlying mechanisms of fetal inflammatory response.
- Augment a graphical model with external network data.
- Introduce both frequentist and Bayesian approaches from Jewson et al. 2022, to improve understanding of dependencies in gene expressions.

Fetal Inflammatory Response (FIR)

Description: Systemic inflammatory response of neonates due to intrauterine infection

Impact: Increases the risk of death and serious complications.

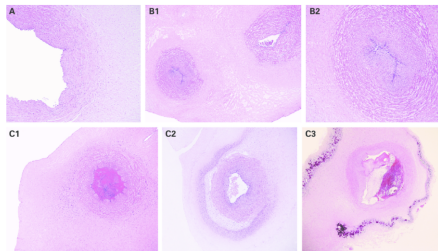


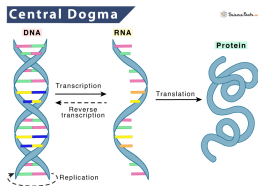
Figure: Different stages of Fetal Inflammatory Response

DATASETS

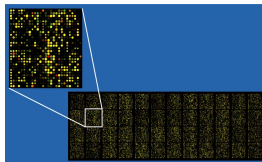
Datasets

We focus on 3 datasets providing gene expression values for small cohorts of neonates with low gestational age, some with and without FIR.¹

Dataset	Method	n	Columns	FIR-affected
Microarray	microarray technology	43	12,106	18
RNA	RNA sequencing	21	11,325	10
Protein	mass spectrometry proteomics	20	245	9



(a) The central dogma of molecular biology.

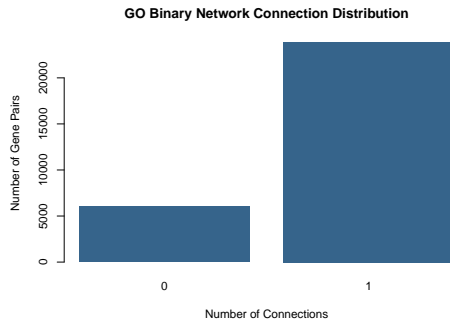


(b) Microarray technology.

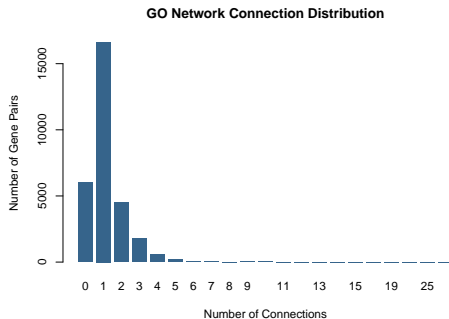
¹from Costa et al. 2020 and Costa and Castelo 2015

Network Data

- External network data from the **Gene Ontology (GO)** database.
- Extracted adjacency matrix, A , in which $a_{jk} = a_{kj}$ indicates number of connections between gene j and k .



(a) GO binary network, A_0 , connections.



(b) GO network, A_1 , connections.

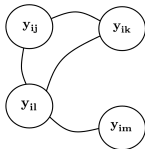
METHODS

Graphical Modeling

Goal: Want to estimate the precision matrix Θ (inverse of covariance matrix).

In a Gaussian Graphical model, all covariates are assumed to be $y_i \sim \mathcal{N}(\mu, \Theta^{-1})$.
Often normalize Θ such that its diagonal elements are 1 and off-diagonal elements are:

$$\rho_{jk} := \frac{\Theta_{jk}}{\sqrt{\Theta_{jj}\Theta_{kk}}} = -\text{corr}(y_{ij}, y_{ik} | y_{i\{1,\dots,p\}\setminus\{j,k\}})$$



	y_{ij}	y_{ik}	y_{il}	y_{im}
y_{ij}	1	-0.5	-0.7	0
y_{ik}	-0.5	1	-0.3	0
y_{il}	-0.7	-0.3	1	0.2
y_{im}	0	0	0.2	1

Graphical LASSO

Goal: Estimate $y_i \sim \mathcal{N}(\mu, \Theta^{-1})$. Estimating Θ requires fitting $p(p+1)/2$ parameters.

Approach: Select Θ to maximize the log-likelihood of the data using a LASSO penalty:

$$\arg \max_{\Theta \in \mathbf{S}_+^p} \log \det(\Theta) - \text{tr}(S\Theta) - \lambda \sum_{j \neq k} |\Theta_{jk}|$$

The larger the penalty λ is, the more the off-diagonal elements of Θ are pushed towards zero. We set λ by minimizing the BIC:

$$BIC(\lambda) = -2l_n(\hat{\Theta}(\lambda)) + |\mathbb{E}(\hat{\Theta}(\lambda))| \times \log n$$

Network Graphical LASSO

Innovation from Jewson et al: Replace the one λ with many λ_{jk} :

$$\arg \max_{\Theta \in \mathbb{S}_+^p} \log \det(\Theta) - \text{tr}(S\Theta) - \sum_{j \neq k} \lambda_{jk} |\Theta_{jk}|$$

where we model the λ_{jk} as a function of the network adjacency matrix A

$$\begin{aligned}\lambda_{jk} &= \lambda_{jk}(A) \\ \lambda_{jk} &= \exp\{\beta_0 + \beta_1 a_{jk}\}\end{aligned}$$

β_0 captures how much sparsity there is in general in the model while β_1 captures of the way in which penalization depends on the connections in A .

Network graphical spike-and-slab LASSO

Goal: understand how a_{jk} affects the probability of an edge between variables j and k .

Approach: Set a prior $\pi(\Theta) = \pi(\text{diag}(\Theta), \rho) = \pi(\text{diag}(\Theta))\pi(\rho)$, where ρ are the normalized off-diagonal elements. We sample from the posterior distribution $\pi(\text{diag}(\Theta), \rho, \eta | Y)$.

$$\pi(\rho) \propto \mathbb{I}(\rho \succ 0) \prod_{j>k} (1 - \omega_{jk}) DE(\rho_{jk}; 0, s_0) + \omega_{jk} DE(\rho_{jk}; \eta_0^\top a_{jk}, s_{jk})$$

$$\omega_{jk} = (1 + e^{-\eta_2^\top a_{jk}})^{-1}, \quad s_{jk} = s_0(1 + e^{-\eta_1^\top a_{jk}})$$

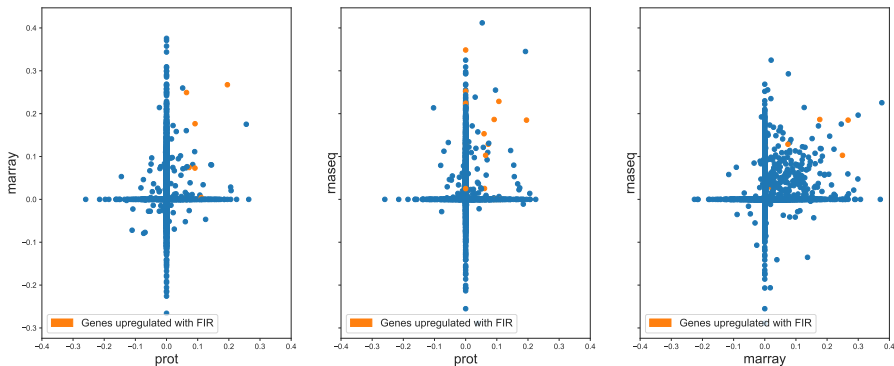
The slab probability (η_2), location (η_0), and dispersion (η_1) all depend on the network A .

RESULTS

GLASSO without external data

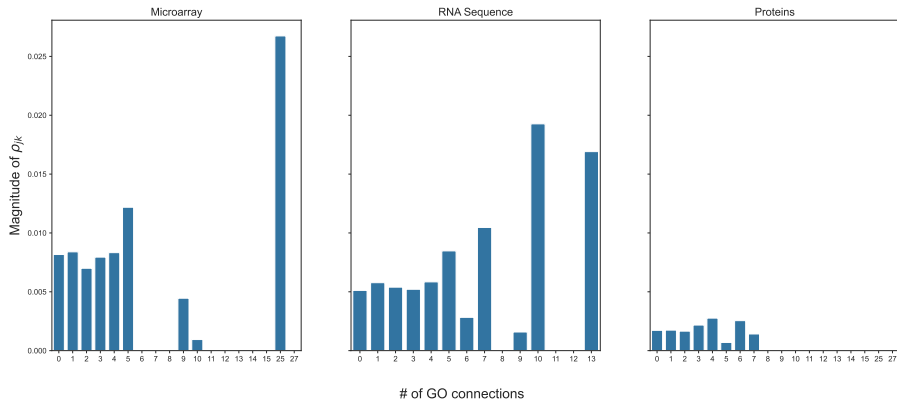
Motivating evidence: Will external information help in our understanding of partial correlations?

Comparing GLASSO partial correlations among three data sets



GLASSO without external data

GO connections vs. average magnitude of partial correlation



Network GLASSO

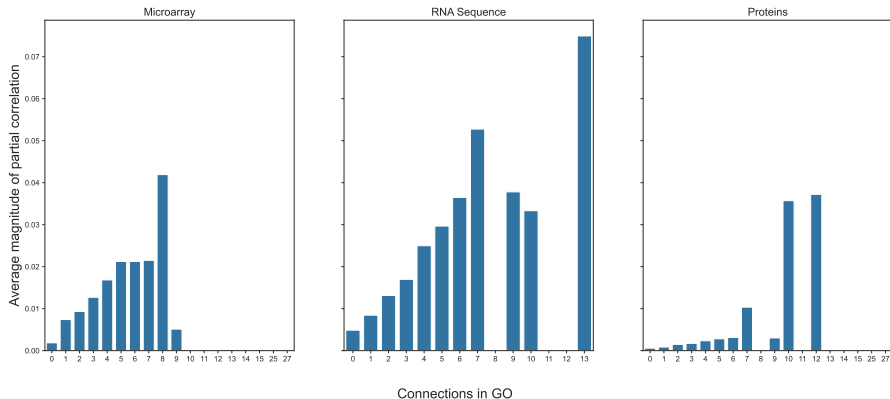
We estimate GLASSO using two versions of the GO network: the binary network A_0 and full network A_1 .

Dataset	Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	Edges
marray	GLASSO	4891.010	-	-	2037
marray	Network GLASSO + A_0	4516.021	-1.868	-0.684	1937
marray	Network GLASSO + A_1	4320.249	-1.395	-0.895	1483
rnaseq	GLASSO	3678.174	-	-	1944
rnaseq	Network GLASSO + A_0	2285.975	-0.368	0.789	1026
rnaseq	Network GLASSO + A_1	1021.159	-3.079	-1.289	3278
protein	GLASSO	6196.498	-	-	961
protein	Network GLASSO + A_0	4900.000	0.474	0.579	506
protein	Network GLASSO + A_1	4860.624	0.895	-0.579	191

Table: In-sample performance of network GLASSO with both networks (A_0 , A_1) vs. standard GLASSO

Network GLASSO

Connections in GO vs. % average magnitude of partial correlation estimated by network GLASSO (A_1)



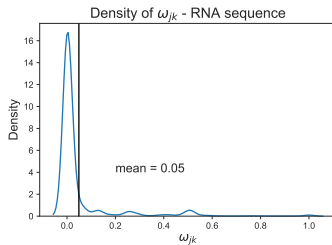
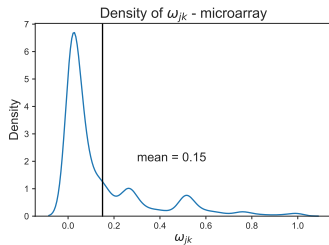
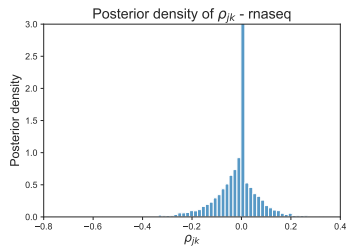
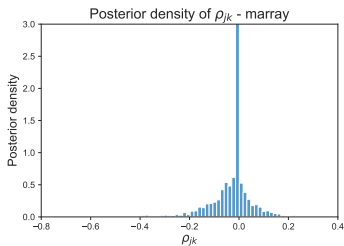
Network GLASSO: Cross-validated estimation

- Incorporating the network data into the graphical model leads to a lower out-of-sample BIC, indicating a better fit.
- The model with the full network did not perform quite as well in terms of log-likelihood compared to the model without network incorporation.
- This limits the conclusiveness of our findings on out-of-sample performance. One possible explanation could be the small sample size ($n = 43$) for the 5-fold cross validated method.
- For RNA Sequence ($n = 21$) and Protein ($n = 20$) datasets the model diverges and produces inconclusive or incomplete results.

Dataset	Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	Edges	5-fold
marray	GLASSO	4615.082	-1.93	-	2000	-163.237
marray	Network GLASSO + A_1	4183.308	-1.67	-0.33	1674	-166.911

Table: Out-of-sample performance of GLASSO vs network GLASSO + A_1 models for microarray

Bayesian spike-and-slab network GLASSO



Bayesian spike-and-slab network GLASSO

Microarray

	Intercept	A_1
η_0 slab location	-0.023	-0.015
95% interval	(-0.031, -0.016)	(-0.024, -0.008)
η_1 slab dispersion	-3.237	-0.010
95% interval	(-3.331, -0.081)	(-0.081, 0.057)
η_2 slab probability	-1.879	0.139
95% interval	(-2.042, -1.709)	(-0.065, 0.238)

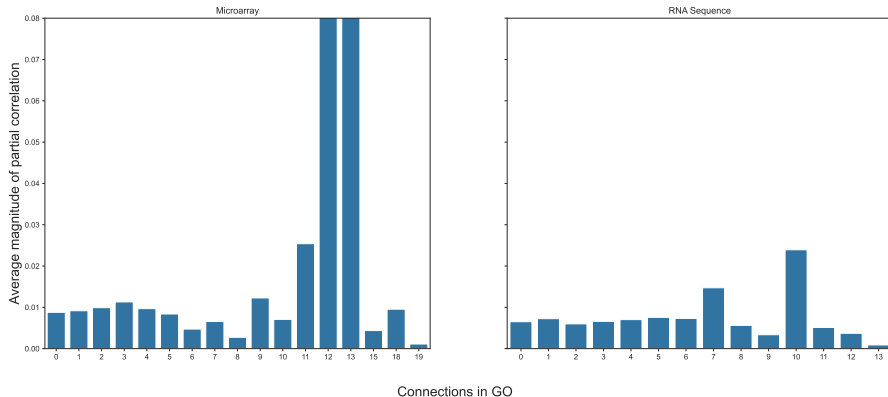
RNA Sequence

	Intercept	A_1
η_0 slab location	-0.106	-0.012
95% interval	(-0.148, -0.079)	(-0.024, 0.009)
η_1 slab dispersion	-3.951	0.068
95% interval	(-4.200, -3.806)	(0.011, 0.162)
η_2 slab probability	-3.366	0.266
95% interval	(-3.683, -3.163)	(0.182, 0.483)

Table: Network spike-and-slab estimates and 95% posterior intervals for microarray and rnaseq data

Bayesian spike-and-slab network GLASSO

Connections in GO vs. % average magnitude of partial correlation estimated by Bayesian spike-and-slab






Relevant Gene Analysis

- In Costa et al. 2020, certain genes were identified to be upregulated in FIR-affected neonates. Two of them (LTF and S100A12) and (S100A8 and S100A9), coincide with our findings as they have non-zero partial correlation across the board in all the GLASSO model results.
- The protein pair (S100A8 and S100A9) is referred to as *calprotectin* and has been known to play an important role in many physiological functions, especially immune response and inflammation. It is used as a biomarker for certain inflammatory conditions, such as inflammatory bowel disease (IBD).
- In the stratified results, the difference in partial correlation between S100A8 and S100A9 was stronger within the FIR-affected group than the FIR-unaffected group (partial correlation of 0.93 and 0.75, respectively).

Conclusions

- We explored and validated graphical modeling techniques on biological data, shedding light on the promises of these approaches and their limitations in the context of our data.
- We found evidence that partial correlations between genes depend on their connection in the GO network using both the frequentist and Bayesian approaches.
- Across all models, the microarray data showed the most consistency.
- The confirmation of the significance of *calprotectin* is both promising in terms of its potential as a biomarker and validates the effectiveness and rationale of incorporating the external GO network to inform our models.

References

-  Costa, Daniel and Robert Castelo (2015). “Umbilical cord gene expression reveals the molecular architecture of the fetal inflammatory response in extremely preterm newborns”. In: *Pediatric Research* 79.3, 473–481. DOI: 10.1038/pr.2015.233. URL: <https://doi.org/10.1111/febs.15578>.
-  Costa, Daniel et al. (Oct. 2020). “Genome-wide postnatal changes in immunity following fetal inflammatory response”. In: *The FEBS Journal* 288. DOI: 10.1111/febs.15578. URL: <https://dx.doi.org/10.1038/pr.2015.233>.
-  Jewson, Jack et al. (2022). *Graphical model inference with external network data*. arXiv: 2210.11107 [stat.AP]. URL: <https://arxiv.org/abs/2210.11107>.