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

Antonia George, Irene Villalonga Domínguez, Joule Voelz

Abstract

In this project, we augment a graphical model with external network data to improve inference on small samples of high-dimensional gene expression data from neonates with and without fetal inflammatory response (FIR). Using a penalized likelihood framework, we demonstrate that incorporating network data of gene interactions from the Gene Ontology (GO) database improves inference across three different sets of clinical data. Using a Bayesian spike-and-slab approach, we confirm that gene pairs with more connections in the GO database are more likely to have a non-zero partial correlation. Our results show agreement across models about the existence of edges between gene pairs known to be upregulated with FIR, demonstrating that our model captures biologically relevant relationships. In our analysis, we explore the limitations of our data and possible interpretations that could connect our graphical models to the understanding of the mechanisms behind FIR.

Data Science Methodology Master Thesis

June 2023

Supervisors: Robert Castelo, David Rossell, Jack Jewson

Contents

1 Introduction

Graphical modeling is a framework that can help describe dependencies among random variables in an interpretable and visually intelligible manner. However, existing graphical modeling techniques struggle to estimate covariance relationships when the number of observations is much smaller than the number of covariates.

Recently, the authors of Jewson et al. 2022 demonstrated the potential of incorporating external network data about covariates into a graphical model. They use both frequentist and Bayesian approaches to successfully improve in and out-of-sample estimation of dependencies among COVID-19 infection rates in US counties and stock market excess returns. In this project, we leverage their techniques to understand dependencies in gene expression data from neonates with and without the condition of fetal inflammatory response (FIR).

In particular, we incorporate external network data from the Gene Ontology (GO) database to inform the estimation of the precision (or inverse-covariance) matrix for three clinical data sets. Using a penalized likelihood framework that allows penalization parameters to depend on the network, we find that incorporating this external data improves on measures of likelihood both in and out-of-sample as compared to a graphical LASSO model with no external information. In order to better understand the proportion of non-zero edges in the data, we implement Jewson et al's spike-and-slab Bayesian framework and estimate it using Monte Carlo techniques. The results of our Bayesian estimation confirm that the network is relevant to estimating edges in the graphical model. Gene pairs with more connections in the GO network are more likely to have an edge between them. In our analysis, we identify several challenges with our data, particularly with regard to two data sets with very small sample sizes. Yet when taken together, the results of our project largely confirm that network data is indeed helpful for estimating covariance structures.

Though the contributions of this paper are mainly methodological and exploratory, we try to understand the possible connections of the gene expression data to FIR. We find that network GLASSO models of all three data sets consistently estimate positive partial correlations between genes that have been identified as upregulated in the presence of FIR. In particular, the gene pair known as *calprotectin*, notably associated with immune response and inflammation, emerges in our estimations with a large partial correlation across all three data sets.

This paper is organized as follows: Section 2 explains the phenomenon of fetal inflammatory response (FIR) and provides background on graphical modeling. Section 3 describes the data and preprocessing steps that were used. Section 4 details the statistical methods employed in this project, in particular the network GLASSO penalized likelihood approach and the Bayesian spike-and-slab framework. Section 5 summarizes the key results of the paper. Section 6 concludes and suggests avenues for future research.

2 Background

2.1 Fetal inflammatory response

Fetal inflammatory response (FIR) is a condition characterized by the systemic inflammatory response of neonates (newborn babies) due to intrauterine infection that often precedes premature birth. FIR increases the risk of death and serious complications, both in the short and long term. For example, neonates affected by FIR have higher rates of neonatal brain damage, neonatal sepsis, intraventricular hemorrhage, and periventricular leukomalacia [Jung et al. 2020]. Long-term complications include neurodevelopmental disorders such as cerebral palsy, retinopathy of prematurity, and sensorineuronal hearing loss [Jung et al. 2020]. Despite the severity of the condition, the underlying molecular events behind FIR are unknown. Moreover, it is difficult for doctors to diagnose FIR while the fetus is still in the womb, because the mother does not generally exhibit any symptoms.

To understand the molecular and genetic markers involved in the prenatal onset of FIR, researchers have utilized gene expression data generated by different sequencing techniques. In particular, Costa and Castelo 2015 found a broad and complex FIR gene expression signature in microarray expression data from the umbilical cord (UC) tissue of 43 neonates, changing up to 19% of all human genes. Significant changes affected the upregulation of many inflammatory pathways and molecules. A second paper Costa et al. 2020 studied a different cohort of 21 neonates with a post-natal dried blood spot sample, obtaining expression values by RNA sequencing and mass spectrometry proteomics. Comparing FIR-affected and unaffected neonates, the authors identified 782 gene and 27 protein expression changes of 50% magnitude or more.

Beyond identifying relevant genes, researchers may want to understand how such genes work together and particularly which pathways are most important in the onset of FIR, for instance with a graphical model. However, there are limitations associated with using gene sequencing data. Clinical studies face challenges due to the difficulty and high cost of collecting and sequencing RNA data. This type of data is also high-dimensional, encompassing thousands of genes, while the number of observations is typically low due to the complexities of collecting relevant clinical data.

The high p, low n nature of gene expression data means that creative statistical methods are necessary to improve inference and gain a better understanding of dependencies. In this paper, we focus on the data obtained from the two papers Costa et al. 2020 and Costa and Castelo 2015 and utilize external network data to improve inference of their covariance structures.

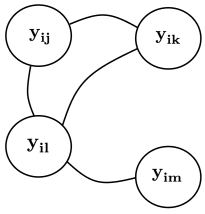
2.2 Graphical modeling

In statistical modeling, two random variables are said to be conditionally independent if they have independent distributions after accounting for a set of conditioning variables. By contrast, two variables are said to have a non-zero partial correlation if they are in fact dependent even after conditioning on the other variables. One way to express the conditional independence relationships between several different random variables is through an undirected graphical model, in which an edge between two nodes represents a partial correlation, and an absence of an edge represents conditional independence.

In a Gaussian graphical model, the covariates are assumed to be distributed joint Gaussian, $y_i \sim \mathcal{N}(\mu, \Theta^{-1})$, where $y_i, \mu \in \mathbb{R}^p$ and $\Theta^{-1} \in \mathbb{S}_+^p$, where $\mathbb{S}_+^p$ is the set of all positive definite $p \times p$ matrices. Θ, the inverse of the covariance matrix, is known as the precision matrix. Θ has the convenient property that if an element $\Theta_{jk} = 0$, then the jth and kth feature of y_i are conditionally independent. In order to work with comparable quantities, we often normalize Θ such that its diagonal elements are 1, and define the off-diagonal elements as:

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

where ρ_{jk} is the negative of the partial correlation of features y_{ij} and y_{ik}. We refer to this normalized precision matrix as R. Figure 1 gives a visual example of such a graphical model and the matrix R associated with it.



(a) Graphical representation

	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

(b) Corresponding R

Figure 1: Gaussian graphical model

For a model with many covariates, estimating Θ via maximum likelihood is bound to be expensive because there are $p(p+1)/2$ such parameters ρ_{jk} to estimate. When n is small relative to p, estimating Θ will be especially difficult. Techniques like the graphical LASSO (GLASSO) aim to address this problem by imposing sparsity [J. Friedman and Tibshirani 2008]. However, the regularization parameter set by GLASSO is one-size-fits-all for all ρ_{jk}, and there is the hope that we can improve upon this approach by allowing each ρ_{jk} to depend on external data that describes the network of connections between each pair of covariates j and k.

This is exactly what the authors of Jewson et al. achieve by using external network data to set an individual penalization parameter λ_{jk} for each element ρ_{jk}. For example, in order to model the dependence structure of COVID-19 infection rates in US counties, the authors incorporate two external networks: one that measures how geographically close the counties are, and one that measures the strength of Facebook connections between the two counties. They then allow λ_{jk} to depend on how closely counties j and k are connected in the network(s). This graphical model estimated with external network data significantly improves inference over standard GLASSO.

In this project, we follow the approach of Jewson et al. in order to see if the same strategy can be applied to gene expression data with positive results.

3 Data

In this paper, we focus our analysis on three data sets obtained from two previously cited papers, Costa et al. 2020 and Costa and Castelo 2015. These data sets provide gene expression values for small cohorts of neonates with low gestational age, some with and without FIR. By utilizing this genetic information, we aim to investigate potential biomarkers that can provide insights into the mechanisms behind FIR.

3.1 Microarray data

The first data set was collected from umbilical cord tissue of extremely low gestational age newborns (ELGANs). The gene expression information was obtained using a microarray sequencing technology. The data contains 43 rows corresponding to 43 different ELGANs, and 12,106 columns. Thirteen of these columns correspond to phenotypic traits including the presence of FIR, as well as other covariates including the sex of the baby and whether the baby was delivered via C-section. The other 12,093 columns record gene expression values. Each of the 12,093 gene columns provides an expression value in quantile-normalized log2 units of expression. The gender breakdown in this dataset is 27 male and 16 female neonates. In this dataset, 18 of the neonates were affected by FIR, and 25 were not affected. The fact that the genetic samples were obtained from umbilical cord tissue is significant in the sense that it records gene expression values from the environment to which the baby was exposed *before* birth. This is of significance when attempting to understand an inflammatory condition affecting the neonate in the womb.

3.2 RNA sequence data

The second data set was obtained by profiling the transcriptome using an RNA sequencing technique on neonatal dried blood spot specimens from another cohort of 21 ELGANs. The data set contains 21 rows, each corresponding to a different ELGAN, and 11,325 columns of phenotypic and genotypic information. Forty-five of the columns contain phenotypic characteristics. The remaining 11,280 columns contain gene expression values in TMM-normalized logCPM units of expression. This dataset includes 13 male and 8 female neonates. Out of 21 total neonates, 10 were affected by FIR and 11 were not affected. In contrast to the microarray data, this sample from dried blood spots captures gene expression that is present after birth, but not necessarily beforehand. One of the ancillary questions of this project is the extent to which these datasets from two different cohorts and from the pre- and post-natal environments will agree with one another.

3.3 Protein data

The third data set was collected from the same sample of neonates as the RNA-sequenced group. The difference was that this data set was obtained using a mass spectrometry proteomics method to profile proteins. The resulting data set contains 20 rows for the neonates, as there was one sample for which the data could not be generated for technical reasons. There are 245 columns in this data set that each correspond to normalized units of expression of specific proteins. There are 13 male and 7 female neonates in this data set. Furthermore, 9 of the neonates were affected by FIR and 11 were not affected.

3.4 Comparisons between sequencing technologies

The three different technologies that generated our three data sets – microarray, RNA sequencing, and mass spectrometry proteomics – share a common goal in detecting gene expression values, yet they have some key differences in how and what they measure. Understanding these differences is important to contextualize the results in this paper.

The central dogma of molecular biology is that genetic information flows from DNA, which is transcripted into RNA, which then plays a key role in synthesizing proteins. Proteins are the key functional units responsible for performing processes within a cell. Thus the expression of a gene found in DNA may be measured in terms of corresponding RNA *or* proteins found in a sample. Differences in measurements

arise both from the sequencing technology, as well as the fact that RNA and proteins come from different stages of the gene expression process.

The technique of RNA sequencing is usually considered to be most reliable, as it sequences the entire genome of the sample through reverse transcription, and thus it can be thought of as an unbiased and full picture of the genetic makeup of the specimen. Microarray and mass spectrometry proteomics test for the abundance of specific RNA molecules or proteins, so they show the expression of only the set of genes that the researchers deemed relevant. Despite this fact, in our microarray data set there was a comparable number of genes analyzed as in the RNA sequencing data.

One key advantage in the microarray data is that expression values are obtained by measuring the relative abundance of specific RNA molecules in a sample. This is done using a microarray chip containing probes in various locations on its surface that bind to and target specific RNA. This process of binding is referred to as hybridization. The relative abundance of hybridized RNA is then measured by the intensity level of the fluorescence. Although RNA sequencing data is usually considered to have a more accurate reading of all genetic information in a sample, the microarray technology produces better-behaved data because this measure of fluorescence intensity produces values within a distinct range and often measures at least some level of fluorescence for all RNA molecules.

A potential issue arising with the protein data set is that researchers chose a smaller sample of specific protein expression values to estimate, and thus this data has lower dimension than the other two data sets. These choices by the researchers could add an unwanted or unintentional bias to the data. Additionally, a key difference is that gene expression is indirectly measured using the mass spectrometry proteomics method. Protein expression values do not measure the underlying genetic makeup, but instead measure the expression of the functional proteins. RNA contains the genetic information that create protein molecules, which are the functional elements that perform the processes within the cell. Although it is generally understood that protein expression and RNA concentrations tend to have a positive correlation, they are still measuring slightly different things, which could lead to different outcomes or interpretations of the results.

Another potential issue with the RNA sequencing and protein data, which was mentioned in the paper by Costa et. al., is that the samples from which these data sets were obtained could have suffered from degradation due to the elapsed time between sample collection and RNA extraction. The samples were stored at room temperature for 1-7 years, which is a timeframe that can significantly affect the RNA integrity. In fact, the authors did indeed find evidence of degradation. This degradation affected the dimension of the protein data, because many protein expression values were dropped, and only the most reliably expressed ones were used.

3.5 External networks

3.5.1 Gene Ontology (GO) network data

The Gene Ontology (GO) data was obtained from Bioconductor, an open-source software for bioinformatics [Carlson 2019]. This source provided gene pair interactions that we utilized to create a network from which we pulled a pairwise adjacency matrix. A gene pair is said to be "connected" if the two genes in the pair share a molecular function, cellular location, or process. Bioconductor determines genes to have a certain connection based on different kinds of evidence, from theoretical to experimental. We specifically focused on six distinct groups of evidence codes. These codes, known as experimental evidence codes, provide support for gene annotations based on direct experimental findings. Each code represents a different type of experimental evidence, such as experimental results (EXP), direct assays (IDA), physical interactions (IPI), mutant phenotypes (IMP), genetic interactions (IGI), and expression patterns (IEP). We chose to utilize these specific groups because they had the strongest evidence of gene interactions.

3.5.2 BioPlex network data

In addition to the GO network data, we also compiled data from BioPlex [Geistlinger et al. 2023], another package available through Bioconductor that contains protein-protein interaction data. Bioplex used affinity-purification mass spectrometry methods to profile these interactions in the human cell line. Although they have created two proteome-scale, cell-line-specific protein-protein interaction networks, we used only the first: BioPlex 3.0. The network was generated by purifying 10,128 human proteins, representing approximately half of the proteome, from 293T cells. It encompasses a total of 118,162 interactions involving 14,586 distinct proteins.

3.6 Pre-processing of data

3.6.1 Pre-processing of clinical data

For reasons both theoretical and practical, we decided to exclude the phenotypic data and limit the columns of each dataset to only the genes found in the protein data, i.e. 245 genes maximum. Besides validating the network GLASSO approach, the goal of this project is to inform understanding and ultimately detection of FIR in-utero. For this reason, we do not want to include phenotypic data that is only observable after birth. We limited the genes to the ones found in the protein data set for two reasons. First of all, estimating a graphical model of over 12,000 genes is neither practical nor desirable. Secondly, we wanted to be able to compare the results of our GLASSO estimation across data sets. In the end, not every gene in the protein data was found in the other two data sets and thus the reduced form of the data had 171, 200, and 245 genes for the microarray, RNA sequence, and protein data respectively. As a final step, the columns of each data set were scaled to a standard normal distribution.

3.6.2 Pre-processing of network data

In order to pre-process the GO network data to be functional for our models, we first extracted the gene groups that were of interest to us. We then removed any gene groups that were composed of only a single gene, because we could not derive any additional information on pairwise associations from these groups of singletons. This left us with 410 distinct gene groups. From these groups, we extracted all pairwise connections between genes by counting how many groups each pair of genes had in common. We were left with 35,778 total gene pairs. From these pairs, we created a network structure and extracted the adjacency matrix, a square matrix A in which $a_{jk} = a_{kj}$ indicates how many connections exist between the pair of genes j and k. We also created a binary adjacency matrix in which the entries were 0 if the gene pair had no connections and 1 if the gene pair had at least one connection.

From the BioPlex package, we extracted all protein-protein interactions. Out of 245 genes in our data, only 210 were included in the database. We filtered to only keep pairs where both genes were in our data set, which left us with 149 gene pairs. Using these gene pairs, we again created a network structure and extracted the adjacency matrix. The BioPlex adjacency matrix was much more sparse than the GO network matrix, as there were only 149 connected gene pairs compared to 21,796 gene pairs with no connection.

3.7 Relevant genes

In the paper Costa et al. 2020, several groups of genes were identified as being biologically relevant to the incidence of FIR. By examining the gene expression values between FIR-affected and FIR-unaffected neonates, they found associations between FIR and the upregulation (an increase of cellular component quantities) of specific sets of genes. These genes were selected using a one-tailed Fisher's test and were provided in the paper. We compared this set of relevant genes to the results of our graphical model analyses to see whether the biologically relevant genes were more likely to have nonzero partial correlations in our models. In figures 2 and A.12, these relevant genes are plotted in orange.

4 Methods

In this work, we apply the two network graphical LASSO methodologies elaborated in Jewson et al. 2022 to the molecular and clinical data described in Section 3. The first approach is a penalized likelihood framework where regularization parameters are allowed to depend on an exogenous network. The second method is a spike-and-slab Bayesian model where the slab probability, location, and scale depend on the network connection strength.

4.1 Gaussian graphical model

In this project, we model each data set of gene expressions as a multivariate normal distribution. Index each data set $Y^{(l)}$ where $l \in \{1, 2, 3\}$. Then the ith observation in the lth data set $y_i^{(l)}$ is distributed:

$$y_i^{(l)} \sim \mathcal{N}_l(\mu^{(l)}, \Theta^{(l)-1})$$

where $\mu^{(l)} \in \mathbb{R}^{p_l}$ and $\Theta^{(l)} \in \mathbb{R}^{p_l \times p_l}$ where p_l is the number of covariates in data set l. $\Theta^{(l)}$ is a positive definite precision (or inverse-covariance) matrix. From here we drop the superscript l for convenience.

In this project, we are interested in estimating Θ, and specifically its standardized form R, whose entries ρ_{jk} are the negative of the partial correlations between covariates:

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

Crucially, ρ_{jk} is equal to zero when y_{ij} and y_{ik} are conditionally independent.

4.2 Graphical LASSO

The graphical LASSO (GLASSO) estimates Θ by selecting Θ to maximize the Gaussian log-likelihood of the data y_i for $i = 1, 2, ..., n$ using a LASSO penalty:

$$\arg\max_{\Theta \in \mathbf{S}_+^p} \log\det(\Theta) - tr(S\Theta) - \lambda \sum_{j \neq k} |\Theta_{jk}| \tag{1}$$

where S is the empirical covariance matrix of Y [J. Friedman and Tibshirani 2008]. The larger the penalty λ is, the more the off-diagonal elements of Θ are pushed towards zero. λ is generally set by cross-validation or through an information criterion. Because cross-validation is not model-selection consistent, we set λ by minimizing the Bayesian information criterion (BIC):

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

Here, $l_n(\hat{\Theta}(\lambda))$ is the Gaussian log-likelihood, and $|\mathbb{E}(\hat{\Theta}(\lambda))|$ is the number of edges (non-zero partial correlations) in the graph associated with $\hat{\Theta}(\lambda)$.

4.3 Network graphical LASSO

The innovation of Jewson et al. is to extend the GLASSO approach and replace the one λ with many λ_{jk}. These λ_{jk} are allowed to depend on external network data. Thus, we fit this model:

$$\arg\max_{\Theta \in \mathbb{S}_+^p} \log\det(\Theta) - tr(S\Theta) - \sum_{j \neq k} \lambda_{jk} |\Theta_{jk}| \tag{3}$$

To set the penalization parameter λ_{jk} for the edge between gene j and gene k, we model it as a function of the gene ontology (GO) network adjacency matrix A (as we discuss later in Section 5, the BioPlex matrix did not prove useful):

$$\lambda_{jk} = \lambda_{jk}(A)$$

We model the penalization parameters as:

$$\lambda_{jk} = \exp\{\beta_0 + \beta_1 a_{jk}\} \tag{4}$$

$\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$ are regularization hyperparameters. β_0 captures how much sparsity there is in general in the model. The more positive β_0, the more sparse the model. The more negative β_0, the more edges there will be. β_1 captures the way in which penalization depends on the connections in A. If β_1 is zero, then the network A is irrelevant. A negative β_1 indicates that the more connected genes j and k are in the network, the more likely they are to have a nonzero partial correlation. The more negative β_1 is, the more relevant the GO network is in informing dependencies in the gene expression data.

Following Jewson et al, we estimate this model using an R package that implements optimization for the GOLAZO class of models outlined in Lauritzen and Zwiernik 2020.[1]

4.4 Network graphical spike-and-slab LASSO

We are also interested in clarifying how the proportion of edges and the mean of the corresponding non-zero partial correlations depend on the network. To answer this question, we follow Jewson et al. and consider a Bayesian framework that sets a prior $\pi(\Theta)$ on the values of the precision matrix Θ. We parameterize Θ in terms of the partial correlations ρ in order to avoid issues with scale invariance. We treat the diagonal and off-diagonal values independently so that $\pi(\Theta) = \pi(diag(\Theta), \rho) = \pi(diag(\Theta))\pi(\rho)$.

[1] https://github.com/pzwiernik/golazo

We set an inverse gamma prior on the diagonal values, so $\sqrt{\Theta_{ii}} \sim \mathcal{IG}(a, b)$ where $a = 0.01$ and $b = 0.01$, reflecting an uninformative prior on the diagonal elements.

Consider the following model:

$$\pi(\rho) \propto \mathbb{I}(\rho \succ 0) \prod_{j>k} (1 - \omega_{jk}) DE(\rho_{jk}; 0, s_0) + \omega_{jk} DE(\rho_{jk}; \eta_0^\top a_{jk}, s_{jk}) \tag{5}$$

$$\omega_{jk} = (1 + e^{-\eta_2^\top a_{jk}})^{-1}, \quad s_{jk} = s_0(1 + e^{-\eta_1^\top a_{jk}}) \tag{6}$$

$\pi(\rho)$ is the prior density of the off-diagonal elements of Θ. $\mathbb{I}(\rho \succ 0)$ is an indicator representing whether Θ is a positive definite matrix, since we cannot have a non-zero prior probability for values of ρ that do not generate valid precision matrices. This joint marginal distribution is the product of the densities of all ρ_{jk}.

With probability $1 - \omega_{jk}$, ρ_{jk} is drawn from a spike, a double exponential distribution centered at 0 with very very small dispersion s_0. With probability ω_{jk}, ρ_{jk} is drawn from the slab represented by a double exponential centered at $\eta_0^\top a_{jk}$ with dispersion s_{jk}.

Crucially, the parameters of the slab as well as ω_{jk} all depend on the network A, via hyperparameters $\eta = (\eta_0, \eta_1, \eta_2)$ Each $\eta_m \in \mathbb{R}^2$ is a two-dimensional vector with an intercept term and a second term that is a coefficient for a_{jk}. η_{02} measures the extent to which a connection a_{jk} pushes the center of the slab away from zero. The more negative η_{02}, the more negative the location of the slab and thus the more positive the mean of the partial correlation (remember that we are estimating ρ_{jk}, the *negative* of the partial correlation). η_{12} measures the extent to which a connection a_{jk} increases the dispersion of the slab. The more negative η_{12}, the more a connection a_{jk} makes the dispersion of a nonzero partial correlation. Finally, η_{22} measures the extent to which a connection a_{jk} makes it more likely that ρ_{jk} will be in the slab – i.e. that there will be an edge. The more positive η_{22}, the greater a connection a_{jk} makes the probability of the corresponding ρ_{jk} being non-zero.

We sample from the posterior distribution $\pi(diag(\Theta), \rho, \eta | Y)$, which we can do using Hamiltonian Monte Carlo via the statistical software Stan. We adapt our code from Jewson et al's implementation of Stan in R, also using their elicited priors for η.[2]

5 Results

In this section, we summarize the key findings of our study and compare results from the different techniques employed. To contextualize our main findings, we begin by detailing our exploratory data analysis (Section 5.1). We then present the results of the two network GLASSO methods and one extension with demeaned and stratified data (Sections 5.2, 5.3, and 5.4). Finally, we analyze the results of the models as they relate to the genes that are upregulated in the presence of FIR (Section 5.5).

5.1 Exploratory data analysis

5.1.1 Benchmark LASSO regression

A motivating question of this project is whether external information can improve inference about the incidence of FIR. Therefore, though prediction is not our primary goal, we would like to understand how well we can predict FIR with no external information, and what genes are most relevant in such a prediction. For a benchmark model, we fit a LASSO logistic regression. We assume that the FIR outcome $z_i \in \{0, 1\}$ for neonate i is a Bernoulli trial with parameter π_i. The logit of π_i depends in a linear fashion on the p-dimensional vector of gene expressions y_i, plus an intercept term.

$$\begin{aligned} z_i &\sim Bernoulli(\pi_i) \\ logit(\pi_i) &= \frac{\pi_i}{1 - \pi_i} \\ logit(\pi_i) &= \beta_0 + \sum_{j=1}^{p} \beta_j y_{ij} \end{aligned}$$

[2] https://github.com/llaurabat91/graphical-models-external-networks

We then run a LASSO regression to obtain $\hat{\beta}$ such that it satisfies:

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} ||Z - Y\beta||_2^2 + \lambda ||\beta||_1$$

Here, Z is an n-dimensional vector of FIR outcomes for n neonates, Y is an $n \times (p+1)$ matrix of gene expressions plus an intercept term, β is a $p+1$-dimensional vector of coefficients, and λ is a penalization parameter. We experiment by setting λ in two ways: first by selecting λ that minimizes prediction error as measured by leave-one-out cross-validation (LOOCV), and second by minimizing the BIC.

Data	**Method**	**n**	**p**	**Error**	**Nonzero genes**
marray	CV	43	171	0.070	ALDH9A1 APC APOC1 CFB PSMC2 **S100A8 S100A9** STMN1 **UBE2L3**
marray	BIC	43	171	0.233	A2M SCYL2
rnaseq	CV	21	200	0.048	KPNB1 LTF ORM1 S100A6 **S100A9** TKT
rnaseq	BIC	21	200	0.095	ACLY LCP1 LXN ORM2 **S100A8** TLN1
protein	CV	20	245	0.100	ALDH9A1 HIST1H1C UBA52
protein	BIC	20	245	0.300	A1BG HIST1H2AJ **UBE2L3**

Table 1: LASSO Regression Results

The results of LASSO regression are summarized in Table 1. As we can see, a simple LASSO model predicts relatively well out-of-sample, with the protein data performing the weakest. Across the selected genes, there is some overlap in that genes *S100A8*, *S100A9*, and *UBE2L3* are selected in the two different cohorts. As discussed in depth in Section 5.5, the gene pair *(S100A8, S100A9)* is an important biomarker for immune response and inflammation. The importance of these genes in the LASSO indicates that the two different cohorts of neonates share some similar patterns of gene expression, and that these patterns are relevant to the incidence of FIR.

5.1.2 GLASSO without external data

Before proceeding with our network-dependent model, we look for motivating evidence that adding external information may help in our understanding of partial correlations.

Using the method outlined in Section 4, we fit a standard graphical LASSO on all three data sets. The resulting Θ matrices are sparse: only around 14%, 10%, and 3% of partial correlations are non-zero in the microarray, RNA sequence, and protein data sets respectively (Figure 4). This fraction is not simply lower for the protein data because it has more gene measurements ($p = 245$ versus 200 and 171 for RNA sequence and microarray, respectively). In fact the protein GLASSO estimates only 961 edges, while the microarray and RNA sequence estimate 2,037 and 1,944, respectively. Especially when compared to the RNA sequence data, which is from the same cohort of neonates, this result suggests that the protein data may only weakly capture the relationships that are present between the covariates, because of the different in measurement technology.

When we plot the GLASSO-estimated partial correlations against each other, we find an encouraging amount of agreement among the data sets, most prominently between the microarray and RNA sequence data (Figure 2). In other words, when the partial correlation between two genes is non-zero across both data sets, they are likely to agree on the sign. This relationship is shown clearly in Figure 3. At least 68% of gene pairs with non-zero partial correlations in two data sets agree on sign, and between the microarray and RNA sequence data sets, at least 89% agree on their sign. This is encouraging news that these two data sets, though they come from two different cohorts and are measured using different technologies, are capturing the same biological mechanisms and can be compared against each other.

Of particular interest are the gene pairs marked in orange in Figure 2. These are genes pairs identified by the authors of Costa et al. 2020 as genes that are upregulated in the presence of FIR. When non-zero in both data sets, these partial correlations are only found in the first quadrant, meaning that both data sets agree on the sign of the partial correlation. We can be more confident, then, that our estimation techniques are capturing some relevant biological relationships.

We now turn our attention to the two network data sets and whether they may be helpful in estimating dependencies among gene pairs. Figure A.10 illustrates the density of the number of connections in the GO network among gene pairs in all three data sets. We note that a large majority of gene pairs have

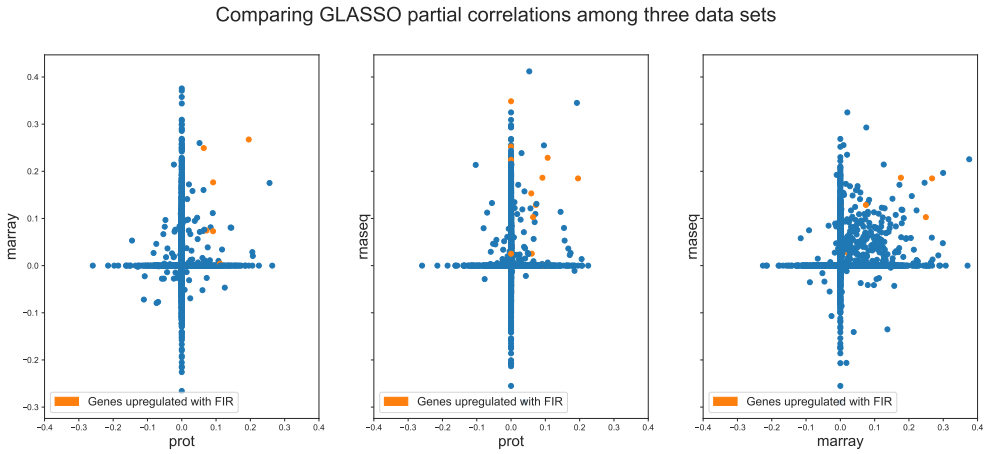


Figure 2: Comparison of GLASSO partial correlations among three data sets

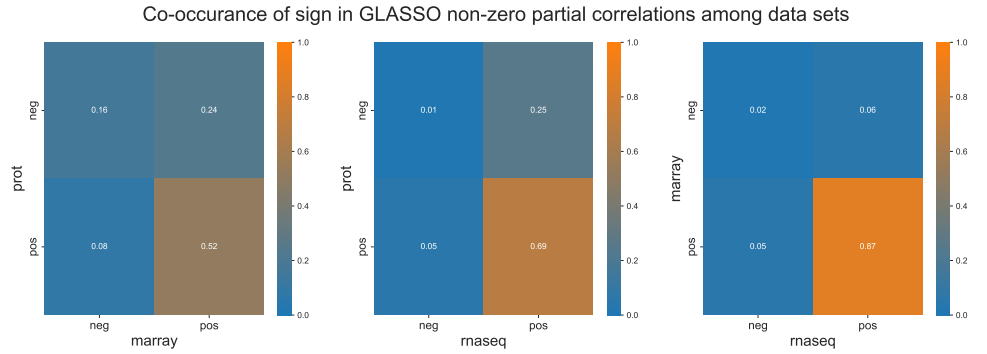


Figure 3: Co-occurance of sign in GLASSO non-zero partial correlations

exactly one connection in the GO network, while significant minorities have either zero or more than one connection.

When we look at the percentage of edges for genes both in and out of the network, we see that there is no overall agreement across data sets as to whether gene pairs that are connected at least once in GO are more likely to have an edge between them (Figure 4). In the RNA sequence and protein data (i.e. data from the same cohort), gene pairs connected in GO seem slightly more likely to have a non-zero partial correlation. This relationship does not appear in the microarray data. This preliminary result indicates that simply being connected in the GO network may not be informative about whether an edge exists between a gene pair. However, it is still possible that the *number* of connections is informative.

When we consider how *many* connections exist between gene pairs in the GO network, we find that the magnitude of partial correlation seems to increase with the number of GO connections, across all three data sets (Figure 5). These relationships provide support to the idea that if a gene pair has more connections in the GO network, it is more likely to have a non-zero partial correlation in our model. This evidence gives us hope that the GO network may be informative in conducting inference on the precision matrix that generates our data sets. However, we note that the partial correlations estimated in the protein data are quite small, and the corresponding graph quite sparse. Thus we expect the network GLASSO to perform best on the microarray and RNA sequence data.

Unfortunately, when we compare gene pairs from all three data sets to the BioPlex network, we find that less than 1% of all gene combinations are connected (Figure A.11). Thus we conclude that the BioPlex network will not be helpful external information for our project, and we proceed only with the GO network.

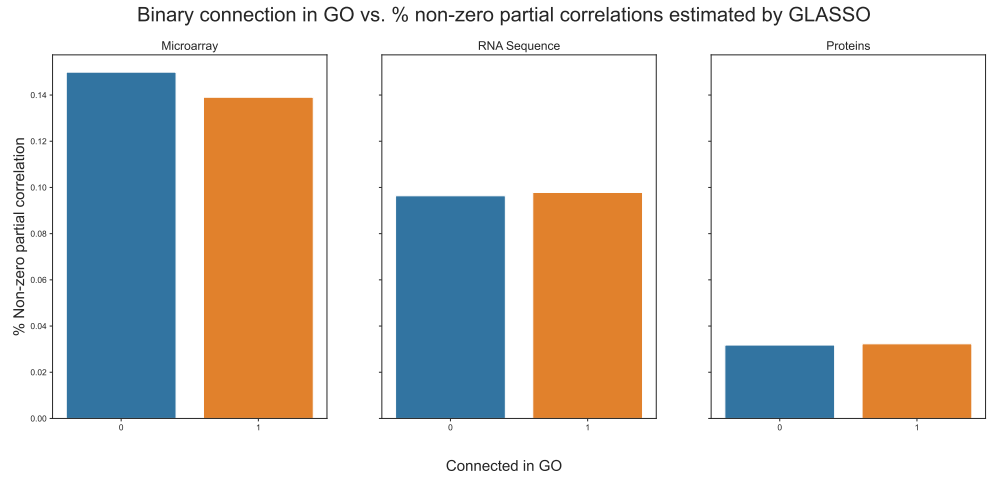


Figure 4: Binary GO connection vs. % non-zero partial correlations estimated by GLASSO

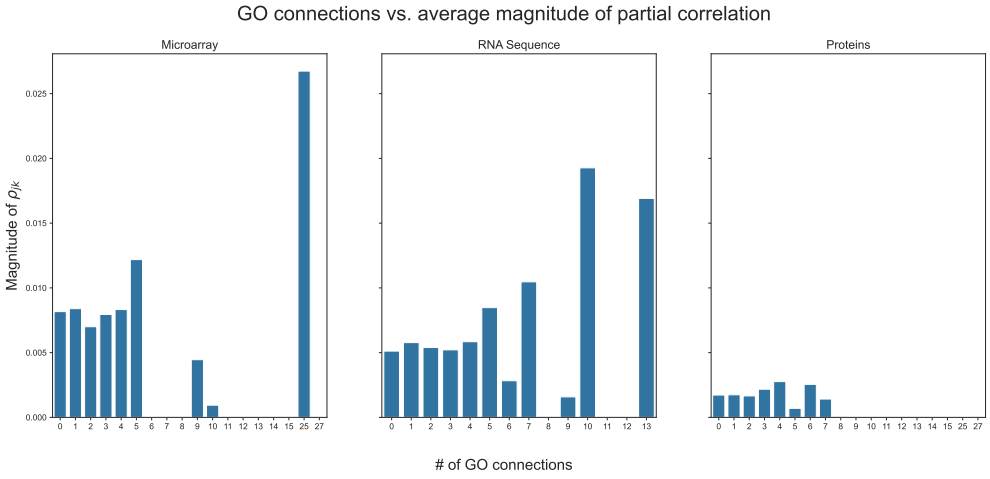


Figure 5: GO connections vs. average magnitude of partial correlations estimated by GLASSO

5.2 Network GLASSO

In the following section, we analyze the results of our frequentist network GLASSO model with external data for all three datasets, summarized in tables 2 and 3.

5.2.1 Estimations with full and binary GO Networks

We estimate our network GLASSO using two versions of the GO network: the binary network A_0, which simply measures whether gene pairs are connected at all in the GO database, and the full network A_1, which measures how many pairwise connections exist between each pair of genes in the GO database. Table 2 summarizes the in-sample estimates of network GLASSO with both the binary network (Network GLASSO + A_0) and full network (Network GLASSO + A_1) as compared to the standard GLASSO.

Based on these results, it appears that both networks of pairwise gene connections improved the in-sample performance of the graphical LASSO model for all three data sets. In particular, the network LASSO estimated with the full network, A_1, consistently exhibits the best BIC value. These results reinforce the findings of Jewson et al. and suggest the value of adding the external network to the graphical LASSO model. Figure 6 demonstrates that across all data sets, gene pairs with more connections have a larger average magnitude of partial correlation when the model is estimated using the full network, A_1. Moreover, the coefficient for the A_1 network ($\hat{\beta}_1$) for all data sets is negative, meaning that highly connected genes according to the GO network are less regularized by the model.

For all three data sets, the network-regularized solutions are more sparse for the binary network (A_0) GLASSO than the standard GLASSO. For microarray and protein, the most sparse solution was the

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

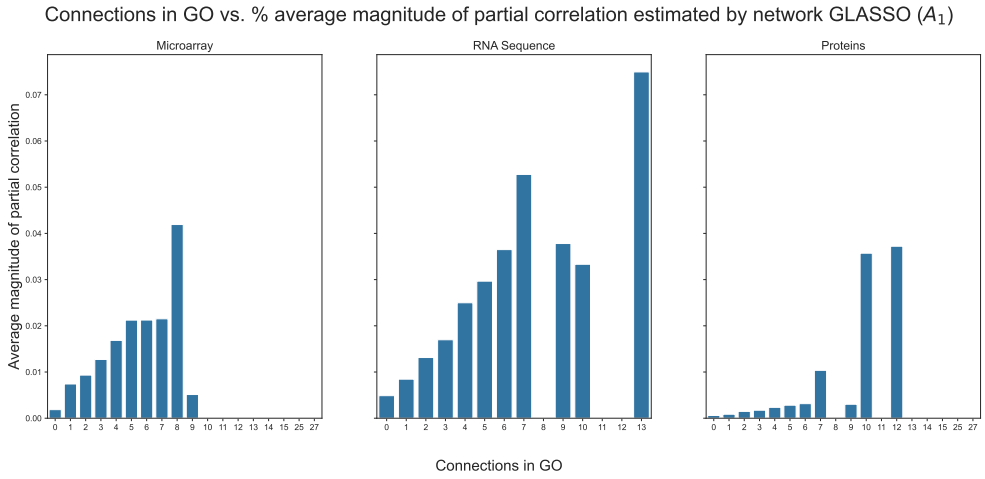


Figure 6: GO connections vs. avg. magnitude of partial correlation estimated by network GLASSO (A_1)

full network (A_1) GLASSO. Despite the full network solutions being more sparse, there were 754 and 152 non-zero edges that were not included in the standard GLASSO for microarray and protein data, respectively. Thus the network data is pushing the GLASSO to select not only fewer gene pairs, but some different gene pairs as well.

Interestingly, this pattern of increasingly sparse solution with a more informative network did not hold true for the RNA sequence data set. While the binary network solution was more sparse than the standard GLASSO, the full network GLASSO had more non-zero edges than any other estimated model across all data sets. It is difficult to interpret the meaning of this finding, and one wonders if this is due to an error or possible over-fitting on such a small data set ($n = 20$). Thus we treat this result with some caution.

The positive sign $\hat{\beta}_1$ for network GLASSO with A_0 for the RNA sequence and protein data is similarly puzzling. However, there may be a simple explanation in that while the binary GO network is not very informative (as suggested by Figure 4 in the exploratory analysis), increased degrees of freedom in regularization naturally lead to a better in-sample fit. The protein data in particular seems to have very weak conditional independence relationships. It is possible that all the external network is doing for the protein data is allowing for over-fitting.

For this reason, we attempt to validate our results with a cross-validation exercise. (Although unfortunately this proved to be impossible for the RNA sequence and protein data, for reasons that may be related to the strange coefficients in the first place.)

5.2.2 Cross-validated estimation

Table 3 summarizes the out-of-sample estimation of the graphical LASSO model without the network (GLASSO) and with full network (Network GLASSO + A_1) for the microarray data set. This analysis was performed using 5-fold cross-validation, where BIC and log-likelihood were assessed as a measure of out-of-sample predictive accuracy. The log-likelihood is listed under the column "5-fold" in the table.

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

We observe that incorporating the network data into the graphical model leads to a lower out-of-sample BIC, indicating a better fit. However, the model with the full network did not perform quite as well in terms of log-likelihood compared to the model without network incorporation.

While this outcome does not entirely invalidate the approach, it does introduce some uncertainty regarding the model's ability to generalize to unseen data, thus limiting the conclusiveness of our findings on out-of-sample performance. One possible explanation for these results could be due to the small sample size. With such a limited number of observations ($n = 43$), the cross-validated results are relying on very small training and testing sets, which could possibly lead to less conclusive results when trying to measure out-of-sample model performance.

While the cross-validation exercise worked well to assess the microarray data set, the RNA sequence and protein data sets were not compatible with the approach. During our attempts to run the cross-validated method for RNA sequence and proteins, the model would diverge and produce inconclusive or incomplete results, prompting us to question whether the model was misaligned or if the sample size was simply incompatible for this approach. With the low number of observations ($n = 21$ and $n = 20$, respectively), each train group in the 5-fold cross-validated method contained only sixteen samples. One potential way to figure out if the sample size itself is the problem would be to run the cross-validation exercise on a smaller sample of the microarray data.

Another potential issue with the RNA sequence and protein data could be that the technology used to obtain the data does not always provide an expression value for all genes, resulting in numerous zero values. In contrast, expression values obtained using microarray technology are, by design, measured within a definitive and positive range for all genes. This distinction arises from the fact that microarray technology employs a measure of fluorescence, which is almost always present to some extent. This discrepancy in measurement methods could lead to a wider and more diverse range of values observed in the RNA sequence and protein data, introducing complications when attempting to fit the model on such a small subset of the data. Additionally, as discussed earlier, there may have been issues with the quality of the data for the RNA sequencing and protein data sets due to degradation of the samples, which could be further contributing to the issues observed.

These issues highlight the need for further exploration to understand why the cross-validated network GLASSO method is failing for these data sets and understand what can be done to better measure out-of-sample performance by considering alternative approaches that could be better suited to accommodating these specific nuances.

For now, the main takeaway seems to be that incorporating increasingly relevant external network data consistently improves estimation of the microarray data, which is the data with the largest sample size.

5.3 Bayesian spike-and-slab network GLASSO

Next, we applied the Bayesian spike-and-slab framework to obtain further insights into how the probability, location, and dispersion of partial correlations depend on the GO network. Because of time constraints, we only ran the model on the microarray and RNA sequence data with the full GO network, A_1. The total runtime of these two models was two and four days, respectively. As a sanity check, we note that the estimated partial correlations from these two Bayesian models generally agree in sign with their counterparts estimated by network GLASSO (Figure A.16).

Table 4 presents the empirical hyperparameter estimates for the microarray data. The slab location parameter was estimated to increase with connection in the GO network (negative η_{01}). The slab dispersion parameter was also estimated to increase with the GO network (negative η_{11}), though the credibility interval included zero. The coefficient η_{21} parameterizing the slab probability was positive, meaning that a greater connection in GO was associated with a larger probability of the corresponding ρ_{jk} being nonzero and thus being an edge. However, the credibility interval included zero.

Table 5 presents the same estimates for the RNA sequence data. The slab location parameter was estimated to increase with the GO network (negative η_{01}), but the credibility interval included zero. The slab dispersion parameter was estimated to decrease with the GO network (positive η_{11}). The coefficient

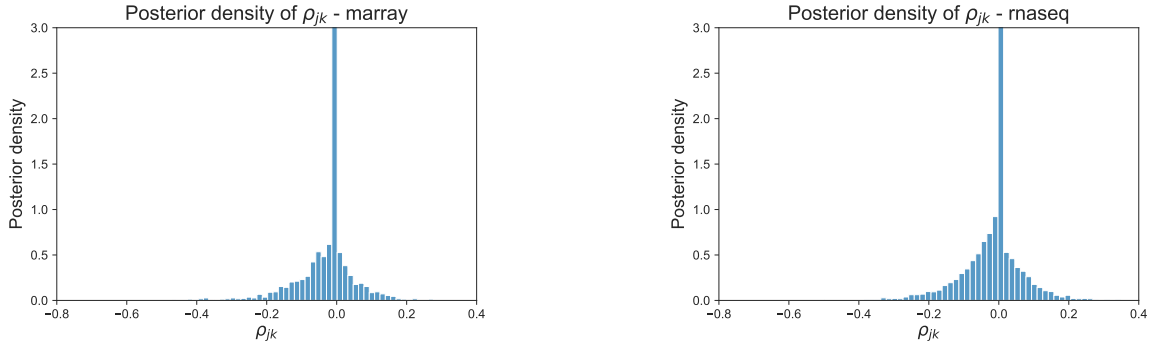
	Intercept	A_1
η_0 slab location	-0.023	-0.015
95% interval	(-0.031, -0.016)	(-0.024, -0.008)
η_1 slab dispersion	-3.237	-0.010
95% interval	(-3.331, -0.081)	(-0.081, 0.057)
η_2 slab probability	-1.879	0.139
95% interval	(-2.042, -1.709)	(-0.065, 0.238)

Table 4: Network spike-and-slab estimates and 95% posterior intervals for microarray data

for slab probability η_{21} was positive, meaning that a greater connection in GO was associated with a larger probability of an edge.

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



(a) Microarray

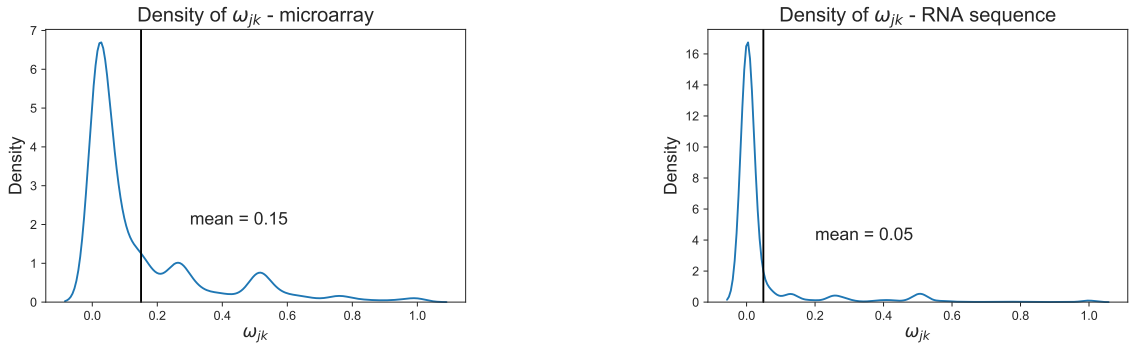
(b) RNA sequence

Figure 7: Posterior density of ρ_{jk}

Taken together, these results agree that the probability of a non-zero partial correlation depends positively on the network, and even more so for the RNA sequence data. This result is in agreement with the findings from the original GLASSO and network GLASSO, in which the RNA sequence data showed the strongest relationship between having a connection in the GO network and having an edge (Figure A.13).

However, the plot of GO connections vs. average magnitude of partial correlations for the spike-and-slab (Figure 9) does not exactly mirror the results obtained by the network GLASSO (Figure 5). The magnitude of partial correlations in the network GLASSO increases almost linearly with GO connections. However, the magnitude of the partial correlations in the network spike-and-slab GLASSO do not increase linearly, though gene pairs with many connections (around 10 to 13) are likelier to have much larger partial correlations. This observation reveals one of the advantages of the spike-and-slab approach, in that it allows us to separate the effects of the GO network on slab probability from the effects of the GO network on slab mean. And indeed while the spike-and-slab results agree that the location of the slab depends on the network, the relationship seems to be much weaker. In other words, connections in GO increase the probability of an edge, but have less of an effect on the size of the corresponding partial correlation. Figure 7 shows the estimated posterior densities of ρ_{jk} from 1,000 iterations.

Figure 8 shows the estimated density of ω_{jk} – the probability of being drawn from the slab – over 1,000 iterations. Interestingly, around 15% of gene pairs in the microarray data fell in the slab, while



(a) Microarray

(b) RNA sequence

Figure 8: Distribution of ω_{jk} - probability of being in the slab

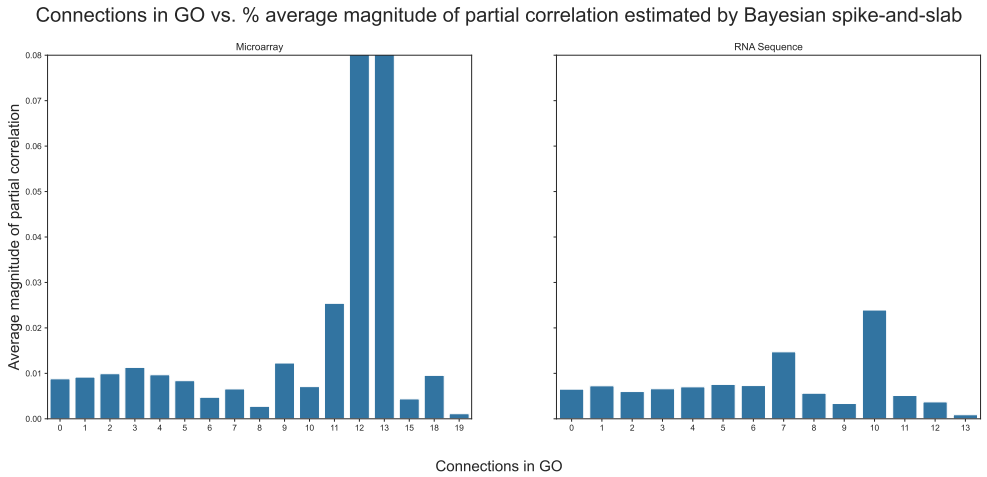


Figure 9: GO connections vs. avg. magnitude of partial correlation estimated by spike-and-slab

only 5% of the gene pairs in RNA data did so. This 15% for the microarray data is not so far away from the percentage of non-zero partial correlations estimated by the A_1 network GLASSO (Figure A.13), and in fact it is less sparse. However, the 5% for RNA sequence data is much sparser than the edges estimated by the network GLASSO. In addition, the estimates of η suggests that the RNA data relies even more on the GO network to determine whether a partial correlation should be non-zero or not. These results again call into question the network GLASSO results obtained for the RNA sequence data. The microarray data, on the other hand, remains consistant and more trustworthy.

Overall, however, the results of our Bayesian model do reinforce the idea that the GO network data is relevant to the estimation of a graphical model of gene sequencing data.

5.4 Network GLASSO with demeaned and stratified data

5.4.1 Estimation with demeaned data and full GO network

Since one motivation of this research is to understand biological mechanisms that underly FIR, we wanted to understand how much the partial correlations estimated by the previous models were related to FIR. As an additional exercise, we ran the network GLASSO model on a modified version of the three data sets, each demeaned by FIR status. In other words, we centered each data point around the mean of the group to which the observation belonged: neonates with FIR and without FIR. Table 6 shows the in-sample estimation of the network GLASSO model with the full network A_1, using demeaned data for all three data sets.

These results are hard to interpret because we cannot compare the BIC directly to the estimates on non-demeaned data. However, the coefficients for the network ($\hat{\beta}_1$) for all three models are negative, indicating that the genes that are highly connected in the network are regularized less by the model. However, the number of edges is only 1 for the microarray and proteins, and 0 for RNA sequence.

Dataset	Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	Edges
marray	Network GLASSO + A_1	4320.249	-1.395	-0.895	1
rnaseq	Network GLASSO + A_1	1021.159	-3.079	-1.289	0
protein	Network GLASSO + A_1	4860.624	0.895	-0.579	1

Table 6: In-sample performance for Network GLASSO + A_1 model using demeaned data.

This result implies that almost all of the partial correlations estimated in the original data sets are related to the presence of FIR or lack thereof. Put another way, all the gene expressions in the data are "conditionally independent" when conditioned on FIR status. Is that possible?

There are several possible issues that could be causing this strange result. First, there could be an issue with how we demeaned the data, or with the code. In the time we had, we were not able to identify one. Another issue may be that we did not grid search over the best values, and our estimates are an overly sparse local minimum of the BIC. One other possible problem is that the microarray data and RNA sequence data have been greatly reduced to only include genes in the protein data set. The subset of 245 genes measured in the protein data set may be particularly relevant to FIR, especially since the nature of proteomics technology means they were chosen specifically to be measured by the researcher. When the effect of FIR is removed from the data, they do not have strong covariance relationships. If we were to include the other 12,000+ genes measured in the microarray and RNA sequence data, demeaned by FIR status, we may not achieve the same result. Then again, it would not be advisable to run a graphical model on that many features.

5.4.2 Estimation with stratified data and full GO network

As a final exercise, we created stratified data sets in which we split each of our three data sets into two groups, samples with FIR and samples without FIR. By doing this, we were able to run the same network GLASSO model, but on the FIR-affected and FIR-unaffected groups separately. Immediately we noticed that the results were suspect, as the RNA sequence and protein data generated strange values of the BIC as well as coefficients (Table 7). Similarly to the inconclusiveness of several previous results, this could be due to various reasons including small sample sizes or issues with the data extraction methods used in these two data sets.

5.5 Relevant gene analysis

Throughout our paper we have referenced certain relevant genes that are upregulated in FIR-affected neonates, as identified by the authors of Costa et al. 2020. Cross-referencing these genes with the results of the network GLASSO, we found only two such FIR-upregulated gene pairs shared an edge in all three GLASSO models with full network (A_1) incorporation. These two gene pairs are (*LTF* and *S100A12*) and (*S100A8* and *S100A9*). Additionally, in the microarray demeaned model results, the only gene pair with non-zero partial correlation was (*S100A8* and *S100A9*).

Interestingly, (*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 [Pathirana et al. 2018]. Calprotectin has been used as a biomarker for inflammation in the gastrointestinal tract, helping to indicate conditions such as inflammatory bowel disease (IBD) from irritable bowel syndrome (IBS) [Pathirana et al. 2018]. It has emerged as an important protein complex in understanding acute and chronic inflammation, and has been used as a biomarker for certain types of cancer [Gebhardt et al. 2006]. Finding this gene pair as relevant in all of our models is an indication that the results we are obtaining are not only biologically sound but could potentially provide vital information for understanding FIR.

In order to better assess the importance of calprotectin in understanding FIR, we also examined the difference in partial correlation between *S100A8* and *S100A9* in the stratified model framework, where FIR-affected and FIR-unaffected datasets were assessed in separate models. When examining the difference between calprotectin expression values between the two groups, we found that the partial correlation between *S100A8* and *S100A9* was stronger within the FIR-affected group. Specifically, we found a correlation of 0.93 within the FIR-affected microarray group, whereas the correlation within the FIR-unaffected group was 0.75. This suggests a potential association between calprotectin expression and the presence of FIR, providing even further evidence supporting the significance of calprotectin as an important biomarker related to FIR.

6 Conclusion

In this study, 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. Using gene expression values from two different cohorts of neonates with and without fetal inflammatory response (FIR), we applied the methods developed in Jewson et al. 2022, utilizing both a frequentist and Bayesian approach. We focused on the inclusion of network data from the Gene Ontology (GO) database in the hopes of improving the estimation of the precision matrix that describes conditional dependencies among genes.

In our network GLASSO model, we showed that including an external network of relevant connections among covariates is a worthwhile approach that can improve inference on the covariance of gene expression values. As compared to a standard LASSO, the network GLASSO demonstrated improved inference when estimating in-sample, but somewhat inconclusive results in an out-of-sample cross-validation exercise. The estimates of our Bayesian network graphical spike-and-slab LASSO also supported the conclusion that the probability, location, and dispersion of partial correlations between genes depend on their connection in the GO network.

Across all models, the microarray data ($n = 43$, $p = 245$) showed consistency in both the sign of the coefficients and proportion of estimated edges. However, the RNA sequence and protein data showed some inconsistencies and computational difficulties that make us cautious about the penalized likelihood approach applied to these particular data. These difficulties may be due to the extremely small sample size ($n = 21$ and $n = 20$, respectively), as well as differences in the gene expression measurement technology, among other things. Nevertheless, the results of the network GLASSO highlight the potential of incorporating external network information into this analysis.

Furthermore, we identified additional evidence of high dependency in specific relevant gene pairs previously identified as upregulated in FIR-affected neonates. The confirmation of these findings using different techniques on different cohorts of neonates strengthens the validity of the previous results and further supports their significance in relation to FIR. The protein complex calprotectin was consistently identified in our analysis as having a large partial correlation in all of the models obtained using the standard GLASSO and network GLASSO approaches. There is compelling biological evidence supporting the crucial role of calprotectin in immune response and inflammatory conditions. Therefore, the confirmation of the significance of this protein complex in our study 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.

Further work could consider how to deal with the nuances and challenges of our clinical data, how to reduce the dimension of such data in an appropriate way, and how to biologically interpret the distribution of edges that we observe in this context. On the statistical side, future work could extend the Bayesian spike-and-slab model to estimate how gene pair dependencies differ between neonates with and without FIR. We hope that our work contributes to the understanding of FIR and demonstrates the value of incorporating external data in many possible applications.

References

Carlson, M (2019). *Org.Hs.eg.db*. URL: https://www.bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html.

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.

Gebhardt et al. (2006). "S100A8 and S100A9 in inflammation and cancer". In: *Biochemical pharmacology* 72, pp. 1622–1631.

Geistlinger, L et al. (2023). *BioPlex*. URL: https://bioconductor.org/packages/release/data/experiment/html/BioPlex.html.

J. Friedman, T. Hastie and R. Tibshirani (2008). "Sparse inverse covariance estimation with the graphical lasso". In: *Biostatistics*, 432–441.

Jewson, Jack et al. (2022). *Graphical model inference with external network data*. arXiv: 2210.11107 [stat.AP]. URL: https://arxiv.org/abs/2210.11107.

Jung, Eunjung et al. (2020). "The fetal inflammatory response syndrome: The origins of a concept, pathophysiology, diagnosis, and obstetrical implications". In: *Seminars in Fetal and Neonatal Medicine* 25.4, p. 101146. DOI: 10.1016/j.siny.2020.101146. URL: https://pubmed.ncbi.nlm.nih.gov/33164775/.

Lauritzen, Steffen and Piotr Zwiernik (2020). "Locally associated graphical models and mixed convex exponential families". In: *arXiv* 2008.04688:1–34, to appear in Annals of Statistics.

Pathirana et al. (2018). "Faecal Calprotectin. " The Clinical biochemist". In: *The Clinical biochemist. Reviews* 39, pp. 77–90.

A Supplementary Materials

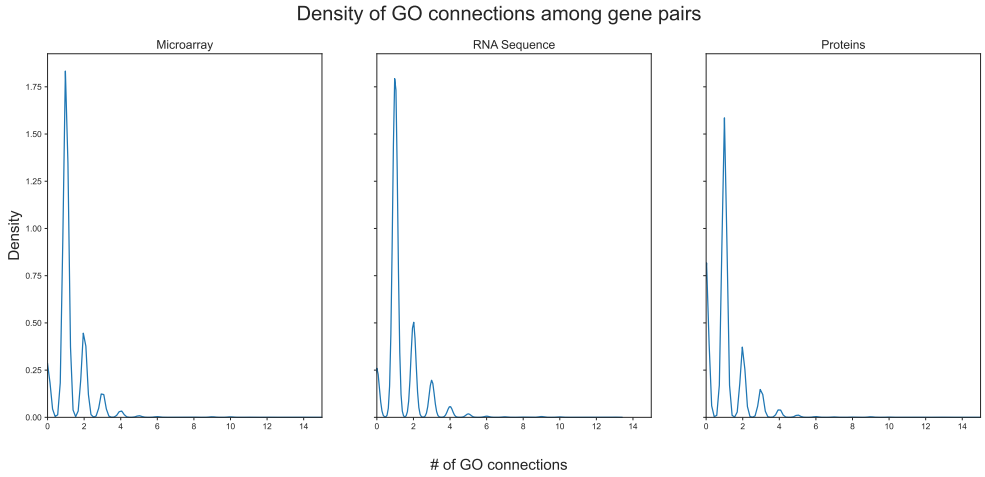


Figure A.10: Density of GO connections among gene pairs

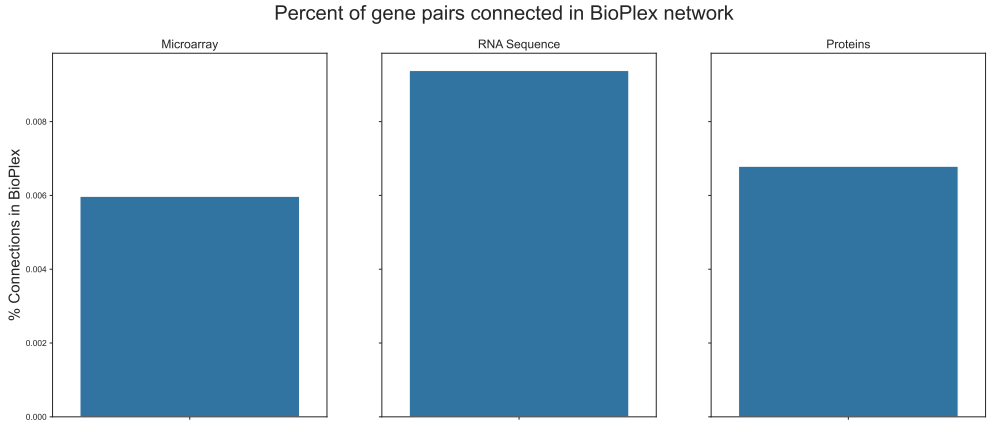


Figure A.11: Percent of gene pairs connected in BioPlex network

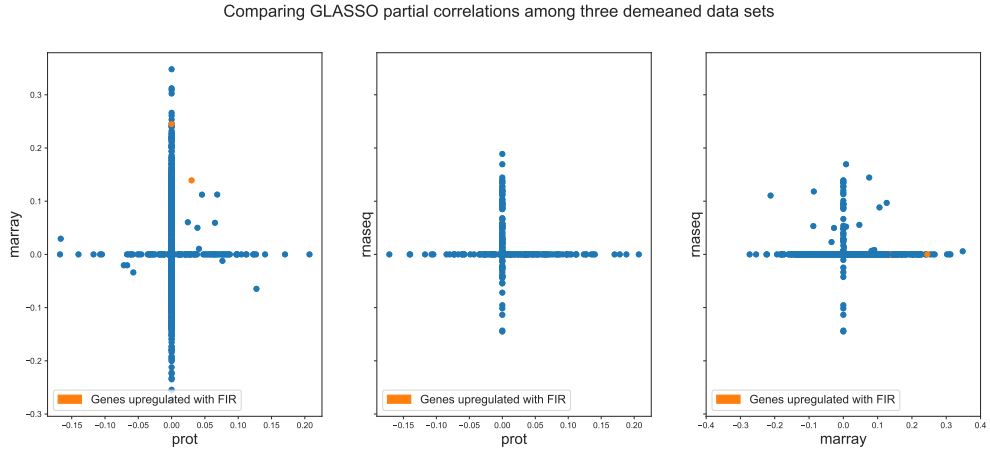


Figure A.12: Comparison of GLASSO partial correlations in demeaned data

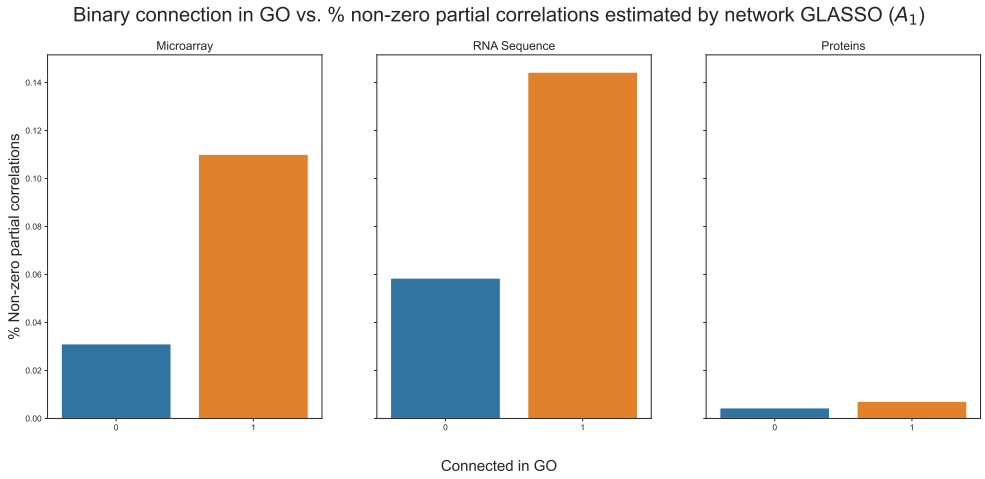


Figure A.13: Binary GO connection vs. % non-zero PC estimated by network GLASSO with A_1

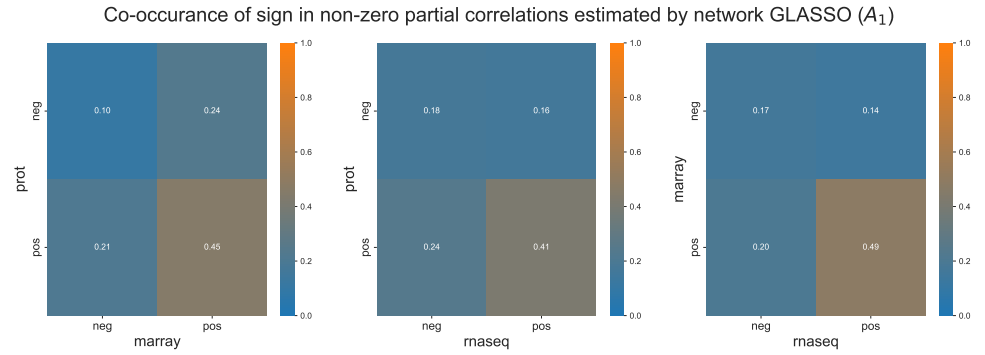


Figure A.14: Co-occurance of sign in non-zero partial correlations estimated by network GLASSO (A_1)

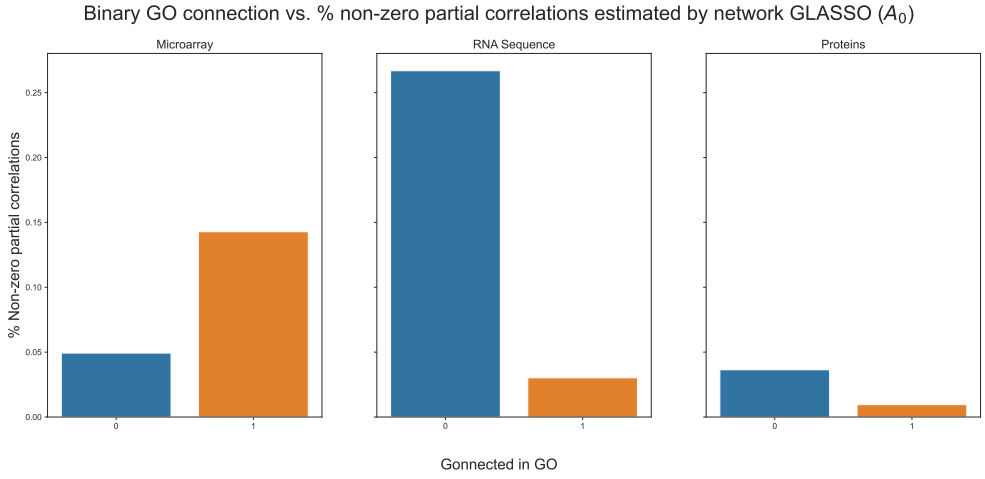
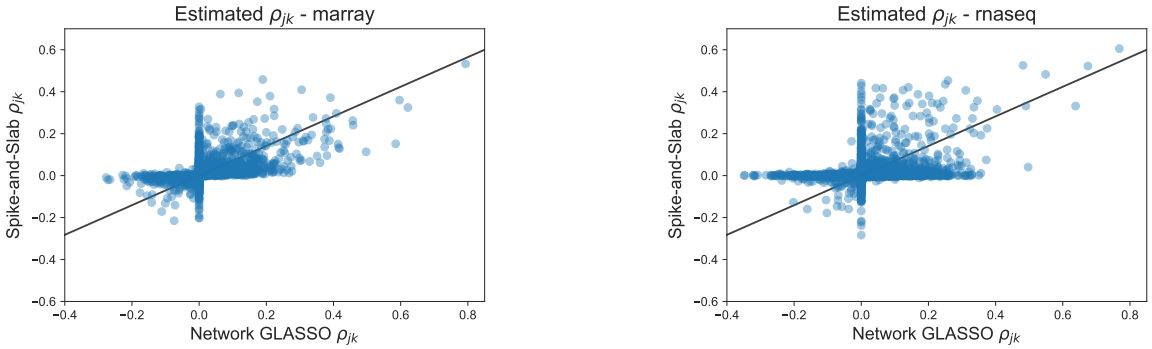


Figure A.15: Binary GO connection vs. % non-zero ρ_{jk} estimated by network GLASSO (A_0)

Dataset	Method	BIC	$\hat{\beta}_0$	$\hat{\beta}_1$	Edges
marray: FIR = 1	Network GLASSO + A_1	2712.659	0.22	-0.67	323
marray: FIR = 0	Network GLASSO + A_1	3478.418	-0.67	-0.67	862
rnaseq: FIR = 1	Network GLASSO + A_1	-427.920	-2.00	2.00	2038
rnaseq: FIR = 0	Network GLASSO + A_1	-298.550	-2.00	2.00	2144
protein: FIR = 1	Network GLASSO + A_1	-633.029	-2.00	2.00	2391
protein: FIR = 0	Network GLASSO + A_1	666.255	-2.00	2.00	2930

Table 7: Problematic results for network GLASSO + A_1 model using stratified data.



(a) Microarray

(b) RNA sequence

Figure A.16: Estimated ρ_{jk} - network GLASSO vs. spike-and-slab