Prior information and variational Bayes in high dimensional statistical network inference

Kpogbezan, G.B.

Author: Kpogbezan, G.B.
Title: Prior information and variational Bayes in high dimensional statistical network inference
Issue Date: 2018-12-10

Chapter 1

Gene network reconstruction using global-local shrinkage priors

Reconstructing a gene network from high-throughput molecular data is an important but challenging task, as the number of parameters to estimate easily is much larger than the sample size. A conventional remedy is to regularize or penalize the model likelihood. In network models, this is often done locally *in the neighbourhood of each node or gene. However, estimation of the many regularization parameters is often difficult and can result in large statistical uncertainties. In this paper we propose to combine local regularization with* global *shrinkage of the regularization parameters to borrow strength between genes and improve inference. We employ a simple Bayesian model with non-sparse, conjugate priors to facilitate the use of fast variational approximations to posteriors. We discuss empirical Bayes estimation of hyper-parameters of the priors, and propose a novel approach to rank-based posterior thresholding. Using extensive model- and data-based simulations, we demonstrate that the proposed inference strategy outperforms popular (sparse) methods, yields more stable edges, and is more reproducible. The proposed method, termed* **`ShrinkNet`**, *is then applied to Glioblastoma to investigate the interactions between genes associated with patient survival.*

This chapter was published as: Gwenaël G.R. Leday, Mathisca C.M. de Gunst, Gino. B. Kpogbezan, Aad W. van der Vaart, Wessel N. van Wieringen and Mark A. van de Wiel (2017). Gene network reconstruction using global-local shrinkage priors. *The Annals of Applied Statistics*, 11(1), 41–68. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

1.1 Introduction

Gaussian Graphical Models (GGMs) are a popular tool in genomics to describe functional dependencies between biological units of interest, such as genes or proteins. These models provide means to apprehend the complexity of molecular processes using high-throughput experimental data, and shed light on key regulatory genes or proteins that may be interesting for further follow-up studies. Among the many approaches that have been advanced, simultaneous-equation models (SEMs), which express each gene or protein expression profile as a function of other ones, have been found particularly valuable owing to their flexibility and simplicity. Notably, SEMs facilitate *local* regularization, where for each gene the set of parameters that model its dependence on the other genes is penalized separately and possibly to a different amount. However this comes at the price of having many regularization parameters, which may be difficult to tune. Motivated by works in the field of differential expression analysis, in this paper we combine local regularization with *global* shrinkage of the regularizing parameters to stabilize and improve estimation. Adopting a Bayesian approach, we demonstrate, using extensive model- and data-based simulations, that such global shrinkage may substantially improve statistical inference.

High-throughput technologies such as microarrays provide the opportunity to study the interplay between molecular entities, which is central to the understanding of disease biology. The statistical description and analysis of this interplay is naturally carried out with GGMs in which nodes represent genes and edges between them represent interactions. The set of edges, which determines the network structure or topology, is often used to generate valuable hypotheses about the disease pathologies. Inferring this set from experimental data is, however, a challenging task as the number of parameter to estimate easily is much larger than the sample size. In this context statistical regularization techniques become necessary.

GGMs characterize the dependence structure between molecular variables using partial correlations. It is well known that two coordinates Y_i and Y_j of a multivariate normal random vector $Y = (Y_1, \dots, Y_p)^T$ are conditionally independent given the set of all other coordinates if and only if the partial correlation $\mathrm{corr}(Y_i, Y_j | Y_{\mathcal{J}\setminus\{i,j\}})$ is zero, where $\mathcal{J} = \{1, \dots, p\}$. Furthermore, if $Y \sim \mathcal{N}_p(0, \mathbf{\Omega}^{-1})$ with positive-definite *precision matrix* $\mathbf{\Omega} = (\omega_{ij})$, then these partial correlations can be expressed as $\mathrm{corr}(Y_i, Y_j | Y_{\mathcal{J}\setminus\{i,j\}}) = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$, for $i \neq j$. Thus the conditional dependence structure is fully coded in the precision matrix, and a network structure may be defined by discriminating the zero and non-zero entries of the precision matrix. It is convenient to represent this structure by an undirected graph $\mathcal{G} = \{\mathcal{J}, \mathcal{E}\}$, with the nodes $\mathcal{J}$ corresponding to the variables, and the edge set $\mathcal{E}$ consisting of all $\{i, j\}$ such that $\omega_{ij} \neq 0$.

Most modern inference techniques for GGMs focus on estimating $\mathbf{\Omega}$ or this underlying graph. For brevity we only discuss the most popular methods, which will also be used as benchmarks in our simulations.

Penalized likelihood estimation amounts to maximizing $\ell(\mathbf{\Omega}) = \log|\mathbf{\Omega}| - tr(S\mathbf{\Omega}) - \lambda J(\mathbf{\Omega})$, where S is the sample covariance estimate, J a penalty function, and λ a scalar tuning parameter. The penalty J may serve two purposes: (1) to ensure identifiability and improve the quality of estimation; (2) to discriminate zero from non-zero entries in $\mathbf{\Omega}$. The ℓ_1-norm (or versions thereof) is a popular choice [43], because it simultaneously achieves (1) and (2). Alternatively, a ridge-type penalty [81, 140, 147] may be used in combination with a thresholding procedure [93, 122]. Appropriate tuning of the penalty through the parameter λ is crucial for good performance. Various solutions, usually based on resampling or cross-validation, have been proposed [41, 46, 49, 89, 98, 158].

Simultaneous-equation modelling estimates $\mathbf{\Omega}$ by regressing each molecular variable Y_j against all others. The coefficients $\beta_{j,k}$ in the equations

$$Y_j = \sum_{k \in \mathcal{J} \backslash j} Y_k \beta_{j,k} + \epsilon_j, \quad j \in \mathcal{J}, \tag{1.1}$$

where $\epsilon_j \sim \mathcal{N}(0, \sigma_j^2)$ is independent of $(Y_k : k \neq j)$, can be shown to be given by $\beta_{j,k} = -\omega_{jj}^{-1}\omega_{jk}$. Also $\sigma_j^2 = \omega_{jj}^{-1}$. Consequently, identifying the nonzero entries of $\mathbf{\Omega}$ can be recast as a variable selection problem in p Gaussian regression models. This approach to graphical modeling was popularized by Meinshausen and Bühlmann [97]. They dealt with high-dimensionality by adding an ℓ_1-penalty to each regression problem, but other penalties are also used [74]. Because the model (1.1) misses the symmetry $\omega_{ij} = \omega_{ji}$ in $\mathbf{\Omega}$, estimation may lack efficiency. This may be overcome by working directly on partial correlations, as shown by Peng et al. [110]. Alternatively, Meinshausen and Bühlmann [97] proposed a *post-symmetrization* step with an 'AND' rule: edge $(i, j) \in \mathcal{E}$ if $\beta_{i,j} \neq 0$ and $\beta_{j,i} \neq 0$. Despite the symmetry issue, network reconstruction using (1.1) performs well and is widely used in practice.

Simultaneous-equation models are quite flexible. Experimental or biological covariates can easily be accounted for in the regression, and extensions to non-Gaussian data were suggested by [2, 25, 115, 156]. Also SEMs arise naturally from the differential equations of a general dynamical system model of gene regulation [103] and are often used to model directed graphs [155].

In this paper we develop a Bayesian approach to Gaussian graphical modeling using SEMs. Our contribution is three-fold: (1) we employ (1.1) in combination with (non-sparse) priors that induce both *local* and *global* shrinkage and provide evidence that global shrinkage may substantially improve inference; (2) we present a new approach to posterior thresholding using a concept similar to the local false discovery rate [37] and show that non-sparse priors coupled with a posteriori edge selection are a simple and attractive alternative to sparse priors; and (3) we provide a computationally attractive software tool called `ShrinkNet` (available at http://github.com/gleday/ShrinkNet), which is based on a coherent and complete estimation procedure that does not rely on resampling or cross-validation schemes to tune parameter(s).

The paper is organized as follows. Section 1.2 presents the Bayesian SEM, the variational approximation to posteriors and a novel posterior thresholding procedure to reconstruct the network. In this section we also describe estimation of the global shrinkage prior and discuss the important role of the proposed empirical Bayes procedure, along with its connection to existing literature. In Sections 1.3 and 1.4 we compare the performance of the new method with state-of-the-art sparse and non-sparse approaches, using both model- and data-based simulations. Notably in Section 1.4 we employ two mRNA expression data sets from The Cancer Genome Atlas (TCGA) and a random-splitting strategy to compare the reproducibility and stability of the various methods. Finally, in Section 1.5 the proposed method is applied to TCGA Glioblastoma data to investigate the interactions between genes associated with patient survival.

1.2 Methods

In this section we introduce the Bayesian SEM with global and local shrinkage priors along with a variational approximation of the resulting posterior distribution(s). Next we present empirical Bayes estimation of prior hyper-parameters. We conclude with a selection procedure for inferring the edge set $\mathcal{E}$.

1.2.1 The Bayesian SEM

Consider mRNA expression data on p genes from n sample tissues. Denote by $\mathbf{y}_j$ the $n \times 1$ vector of mRNA expression ($\log_2$) values for gene $j \in \mathcal{J} = \{1, \ldots, p\}$. The

Bayesian SEM is defined by equation (1.1) together with a hierarchical specification of prior distributions:

$$
\begin{aligned}
\mathbf{y}_j &= \sum_{k \in \mathcal{J} \setminus j} \mathbf{y}_k \beta_{jk} + \boldsymbol{\epsilon}_j, \qquad j = 1, \dots, p \\
\boldsymbol{\epsilon}_j &\sim \mathcal{N}_n(0, \sigma_j^2 \mathbf{I}_n), \\
\beta_{jk} &\sim \mathcal{N}(0, \sigma_j^2 \tau_j^2), \\
\tau_j^{-2} &\sim \mathcal{G}(a, b), \\
\sigma_j^{-2} &\sim \mathcal{G}(c, d).
\end{aligned}
\tag{1.2}
$$

Here every line is understood to be conditional on the lines below it and variables within a line are assumed independent, as are variables referring to different genes j. Furthermore, $\mathcal{G}\,(s, r)$ denotes a gamma distribution with shape and rate parameters s and r, and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Throughout the paper the hyper-parameters c and d are fixed to small values, e.g. 0.001, in contrast to a and b, which we will estimate (see Section 1.2.3). Although c and d could also be estimated, we prefer a non-informative prior for the parameters σ_j, as there seems no reason to connect the error variances across the equations.

The regression parameters β_{jk} are endowed with gene-specific, Gaussian priors for *local* shrinkage. A small value of the prior variance τ_j^2 encourages the posterior distributions of the β_{jk} (including their expectations $\mathbb{E}(\beta_{jk}|\mathbf{y}_j)$) to be shrunken towards zero. The stabilizing effect of this ridge-type shrinkage has been observed to be useful for ranking regression parameters as a first step in variable selection [14]. In Section 1.2.4 we show how similarly the marginal posterior distributions of the β_{jk} can be used for rank-based edge selection in a GGM. The prior variances of the β_{jk} are also defined proportional to the error variances σ_j^2 to bring the variances τ_j^2, and the induced shrinkage, on a comparable scale [107].

The equations for different genes j are connected through the gamma priors placed on the precisions τ_j^{-2} and the error variances σ_j^2, for $j \in \mathcal{J}$. The prior on the error variances has no structural role, and, as mentioned, we prefer a fixed non-informative prior. In contrast, the $\mathcal{G}(a, b)$-prior on the precisions τ_j^{-2} induces *global* shrinkage by borrowing strength across the regression equations. The *exchangeability* of the precisions expressed through this prior acknowledges the fact that the equations for the different genes are similar in a broad sense, which is plausible given that they share many common elements. When informative (i.e. small or moderate value of a/b^2), this prior shrinks the posterior distributions of τ_j^{-2} towards the prior mean a/b, which stabilizes estimation. This type of shrinkage is different from the shrinkage of the regression coefficients β_{jk}, which through their centered priors are always shrunken to zero. Of course, the "informed" shrinkage of the precisions τ_j^{-2} will be beneficial only if the hyper parameters a and b are chosen appropriately. We propose to set their values based on the data, using an empirical Bayes approach, discussed in Section 1.2.3.

The conjugacy of the Gaussian and gamma priors in model (1.2) confers the method a computational advantage over complex sparse priors. Fast approximations to the posteriors are readily available [106, 114, 118], whereas sparse, non-conjugate priors often require MCMC. The Gaussian priors allow to reparameterize the problem employing an SVD decomposition of the design matrix [150], and back-transform the posteriors to the original space (at least in our setting with approximately Gaussian posteriors; see Section 1.2.2), which is computationally advantageous.

A disadvantage of these priors is that they do not have an intrinsic variable selection property, whence the posterior does not automatically recover the graph structure. We solve this by a separate procedure for variable selection, which essentially consists of thresholding the scaled posterior means of the regression coefficients β_{jk}. In Section 1.2.4 we present an approach based on Bayes factors and a local false discovery rate.

1.2.2 Variational approximation to posteriors

Because intractable integrals make it difficult to obtain the exact marginal posterior distribution of the parameters, we use a variational approximation. Variational inference is a fast deterministic alternative to MCMC methods, and consists of computing a best approximation to the posterior distribution from a prescribed family of distributions. In our situation it provides an analytic expression for a lower bound on the log-marginal likelihood, which is useful for monitoring convergence of the algorithm and to assess model fit (Section 1.2.3).

For given hyper-parameters (a, b) and with the variables $\mathbf{y}_k$ in the right side of (1.2) considered fixed covariates, the prior and posterior distributions factorize (i.e. are independent) across the genes j. For simplicity of notation we shall omit the index j from τ_j^{-2}, σ_j^{-2}, $\mathbf{y}_j$ and $\boldsymbol{\beta}_j$ in the remainder of this section. Hence the formulas for $\boldsymbol{\lambda} := (\boldsymbol{\beta}, \tau^{-2}, \sigma^{-2})$ below apply to the joint posterior distribution of $(\boldsymbol{\beta}_j, \tau_j^{-2}, \sigma_j^{-2})$, for (any) given $j \in \mathcal{J}$.

We shall seek a variational approximation to the posterior distribution of $\boldsymbol{\lambda}$ within the class of all distributions with independent marginals over $\boldsymbol{\beta}$, τ^{-2} and σ^{-2}, where we measure the discrepancy by the Kullback-Leibler (KL) divergence. Thus letting $p(\boldsymbol{\lambda}|\mathbf{y})$ denote the posterior density in model (1.2), we seek to find a density $q(\boldsymbol{\lambda})$ of the form

$$q(\boldsymbol{\lambda}) = q_1(\boldsymbol{\beta})q_2(\tau^{-2})q_3(\sigma^{-2}), \tag{1.3}$$

for some marginal densities q_1, q_2, q_3, that minimizes the Kullback-Leibler divergence

$$\begin{aligned} \mathrm{KL}(q||p) &= \int q(\boldsymbol{\lambda}) \log \frac{q(\boldsymbol{\lambda})}{p(\boldsymbol{\lambda}|\mathbf{y})}\, d\boldsymbol{\lambda} \\ &= \mathbb{E}_q \log q(\boldsymbol{\lambda}) - \mathbb{E}_q \log p(\boldsymbol{\lambda}, \mathbf{y}) + \log p(\mathbf{y}), \end{aligned} \tag{1.4}$$

over all densities q of product form. Here $p(\mathbf{y})$ denotes the marginal density of the observation in model (1.2). Because the Kullback-Leibler divergence is nonnegative we have that

$$\mathbb{E}_q \log p(\boldsymbol{\lambda}, \mathbf{y}) - \mathbb{E}_q \log q(\boldsymbol{\lambda}) \leq \log p(\mathbf{y}). \tag{1.5}$$

Furthermore, minimization of the Kullback-Leibler divergence is equivalent to maximization of the left side of this inequality. Thus we may think of the procedure as maximizing a lower bound on the log marginal likelihood.

The solution q^* of this maximization problem, with the marginal densities q_1, q_2, q_3 left completely free, can be seen to be given by densities q_1^*, q_2^*, q_3^* satisying (see [13, 106])

$$q_m^*(\boldsymbol{\lambda}_m) \propto \exp\left\{ \mathbb{E}_{\prod_{m' \neq m} q_{m'}} \log p(\boldsymbol{\lambda}, \mathbf{y}) \right\}, \qquad m = 1, 2, 3. \tag{1.6}$$

In the context of our model this yields $q^*(\boldsymbol{\lambda}) = q_1^*(\boldsymbol{\beta})q_2^*(\tau^{-2})q_3^*(\sigma^{-2})$, with the marginal densities (see Section 1 of Supplementary Material (SM)) given by standard distributions,

$$\begin{aligned} q_1^*(\boldsymbol{\beta}) &=^d \mathcal{N}_{p-1}\left(\boldsymbol{\beta}^*, \boldsymbol{\Sigma}^*\right) \\ q_2^*(\tau^{-2}) &=^d \mathcal{G}\left(a^*, b^*\right), \\ q_3^*(\sigma^{-2}) &=^d \mathcal{G}\left(c^*, d^*\right), \end{aligned} \tag{1.7}$$

where the parameters on the right side satisfy

$$\begin{aligned}
\boldsymbol{\beta}^* &= \left(\mathbf{X}^T\mathbf{X} + \mathbb{E}_{q_2^*}\left[\tau^{-2}\right]\mathbf{I}_{p-1}\right)^{-1}\mathbf{X}^T\mathbf{y}\\
\boldsymbol{\Sigma}^* &= \left[\mathbb{E}_{q_3^*}\left[\sigma^{-2}\right]\left(\mathbf{X}^T\mathbf{X} + \mathbb{E}_{q_2^*}\left[\tau^{-2}\right]\mathbf{I}_{p-1}\right)\right]^{-1}\\
a^* &= a + \frac{p-1}{2},\\
b^* &= b + \frac{1}{2}\mathbb{E}_{q_3^*}\left[\sigma^{-2}\right]\mathbb{E}_{q_1^*}\left[\boldsymbol{\beta}^T\boldsymbol{\beta}\right],\\
c^* &= c + \frac{n+p-1}{2},\\
d^* &= d + \frac{1}{2}\mathbb{E}_{q_1^*}\left[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right] + \frac{1}{2}\mathbb{E}_{q_2^*}\left[\tau^{-2}\right]\mathbb{E}_{q_1^*}\left[\boldsymbol{\beta}^T\boldsymbol{\beta}\right].
\end{aligned}$$

Here $\mathbf{X}$ represents the n by $p-1$ fixed design matrix of (1.2). For the j^{th} equation in (1.2) this is equal to $\mathbf{y}_{-j} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_{j-1}^T, \mathbf{y}_{j+1}^T, \ldots, \mathbf{y}_p^T)^T$.

Furthermore, the variational lower bound on the log-marginal likelihood $\log p(\mathbf{y})$ (the left side of (1.5)) evaluated at $q = q^*$ simplifies to:

$$\begin{aligned}
\mathcal{L} = &-\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}^*| + \frac{1}{2}(p-1) + a\log b - \log\Gamma(a) - \\
&a^*\log b^* + \log\Gamma(a^*) + c\log d - \log\Gamma(c) - c^*\log d^* + \\
&\log\Gamma(c^*) + \frac{1}{2}\mathbb{E}_{q_3^*}\left[\sigma^{-2}\right]\mathbb{E}_{q_2^*}\left[\tau^{-2}\right]\mathbb{E}_{q_1^*}\left[\boldsymbol{\beta}^T\boldsymbol{\beta}\right].
\end{aligned} \tag{1.8}$$

See SM Section 1 for the details.

The equations (1.7) express the optimal densities q_1^*, q_2^* and q_3^* (or equivalently the parameters in the right side of (1.7)) in terms of each other. This motivates a coordinate ascent algorithm [13, 106] (depicted in Algorithm 1), which proceeds by updating the parameters in turn, replacing the variational densities on the right hand sides of the equations by their current estimates, at every iteration.

Upon convergence the marginal posteriors $p(\boldsymbol{\beta}|\mathbf{y})$, $p(\tau^{-2}|\mathbf{y})$ and $p(\sigma^{-2}|\mathbf{y})$ are approximated by $q_1^*(\boldsymbol{\beta})$, $q_2^*(\tau^{-2})$ and $q_3^*(\sigma^{-2})$. Although the algorithm needs to be repeated for each regression equation in (1.2), the overall computational cost of the procedure is low.

Algorithm 1 Variational algorithm for local shrinkage

```
Initialize:
$b = d = b^{*(0)} = d^{*(0)} = 0.001$, $\xi = 10^{-3}$, $M = 1000$ and $t = 1$
while $|\mathcal{L}^{(t)} - \mathcal{L}^{(t-1)}| \geq \xi$ and $2 \leq t \leq M$ do
    update $\boldsymbol{\Sigma}^{*(t)} \leftarrow \left[\mathbb{E}_{q_3^{*(t-1)}}(\sigma^{-2})\left(\mathbf{X}^T\mathbf{X} + \mathbb{E}_{q_2^{*(t-1)}}(\tau^{-2})\mathbf{I}_{p'}\right)\right]^{-1}$
    update $\boldsymbol{\beta}^{*(t)} \leftarrow \mathbb{E}_{q_3^{*(t-1)}}(\sigma^{-2})\boldsymbol{\Sigma}^{*(t)}\mathbf{X}^T\mathbf{y}$
    update
        $d^{*(t)} \leftarrow d + \frac{1}{2}\left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{*(t)})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{*(t)}) + \mathrm{tr}\{\mathbf{X}^T\mathbf{X}\boldsymbol{\Sigma}^{*(t)}\}\right] +$
        $\frac{1}{2}\mathbb{E}_{q_2^{*(t-1)}}(\tau^{-2})\left[{\boldsymbol{\beta}^{*(t)}}^T\boldsymbol{\beta}^{*(t)} + \mathrm{tr}\{\boldsymbol{\Sigma}^{*(t)}\}\right]$
    update $b^{*(t)} \leftarrow b + \frac{1}{2}\mathbb{E}_{q_3^{*(t-1)}}(\sigma^{-2})\left[{\boldsymbol{\beta}^{*(t)}}^T\boldsymbol{\beta}^{*(t)} + \mathrm{tr}\{\boldsymbol{\Sigma}^{*(t)}\}\right]$
    update $\mathcal{L}^{(t)}$
    $t \leftarrow t + 1$
end while
```

1.2.3 Empirical Bayes and prior calibration

In the preceding discussion we have treated the vector of hyper-parameters $\boldsymbol{\alpha} = (a, b)$ as fixed. We now turn to its estimation and present a modified variational algorithm in which $\boldsymbol{\alpha}$ is updated along with the other parameters. The new algorithm is akin to an EM algorithm [15] in which the two steps are, respectively, replaced with a variational E-step, where the lower bound is optimized over the variational parameters via coordinate ascent updates, and a variational M-step, where the lower bound is optimized over $\boldsymbol{\alpha}$ with the variational parameters held fixed.

We now use the SEM for all genes together, and write the variational approximation for the posterior density of the parameters for the jth gene as q^j. (For each j this is given by a triple of three marginal densities.) The target is to maximize the sum over the genes of the lower bounds on the log-marginal likelihood, i.e. the sum over j of the left side of (1.5), which can be written as

$$\sum_{j=1}^{p} \mathbb{E}_{q^j} \log p(\mathbf{y}_j|\boldsymbol{\lambda}_j) + \sum_{j=1}^{p} \mathbb{E}_{q^j} \log \frac{p_{\boldsymbol{\alpha}}(\boldsymbol{\lambda}_j)}{q^j(\boldsymbol{\lambda}_j)} \leq \sum_{j=1}^{p} \log p_{\boldsymbol{\alpha}}(\mathbf{y}_j). \tag{1.9}$$

Maximization of the left side with respect to the densities q^j for a fixed hyper-parameter $\boldsymbol{\alpha}$ would lead to the variational estimates q^{j*} given by (1.7). However, rather than iterating (1.7) until convergence, we now alternate between ascending in q and in $\boldsymbol{\alpha}$. For the variational estimates q^j fixed at their current iterates, optimizing the left-hand side of (1.9) relative to the parameter $\boldsymbol{\alpha}$ amounts to maximizing, with the current iterate q^{j*} replacing q^j,

$$\sum_{j=1}^{p} \mathbb{E}_{q^{j*}} \log p_{\boldsymbol{\alpha}}(\tau_j^{-2}) = \sum_{j=1}^{p} \Big(a \log b - \log\Gamma(a) + (a-1)\mathbb{E}_{q^{j*}} \log \tau_j^{-2} - b\,\mathbb{E}_{q^{j*}} \tau_j^{-2}\Big). \tag{1.10}$$

The exact solution to this problem can be found using a fixed-point iteration method, as in [132]. Alternatively, the following approximate solution arises by analytical maximization after replacing the digamma function $\psi(x) = \frac{\partial}{\partial x} \log\Gamma(x)$ by the approximation $\log(x) - 0.5x^{-1}$:

$$\begin{cases} \hat{a} = \dfrac{1}{2}\left[\log\left(\displaystyle\sum_{j=1}^{p} \mathbb{E}_{q^{j*}} \tau_j^{-2}\right) - p^{-1}\displaystyle\sum_{j=1}^{p} \mathbb{E}_{q^{j*}} \log \tau_j^{-2} - \log p\right]^{-1} \\ \hat{b} = \hat{a} \cdot p \cdot \left[\displaystyle\sum_{j=1}^{p} \mathbb{E}_{q^{j*}} \tau_j^{-2}\right]^{-1} \end{cases} \tag{1.11}$$

Algorithm 2 outlines how the updates of the hyper-parameters are folded into the variational algorithm. At iteration t the hyper-parameters $a^{(t)}$ and $b^{(t)}$ are computed according to (1.11) with the expectations $\mathbb{E}_{q^{j*}} \tau_j^{-2}$ and $\mathbb{E}_{q^{j*}} \log \tau_j^{-2}$ computed under the current estimates q^{j*}. Next the variational parameters defining the densities q^{j*} are updated according to (1.7) using the values $a^{(t)}$ and $b^{(t)}$ for a and b. Figure 1.1(a) illustrates the convergence of the algorithm and shows that the lower bound on the sum of log-marginal likelihoods increases at each step of the algorithm (red line). Although this is not true for the lower bounds of each regression equation in the SEM, this does demonstrate that the estimation procedure yields a well-informed prior that is beneficial overall.

The second summand on the left-hand side of (1.9) is equal to minus $\sum_{j=1}^{p} \mathrm{KL}(q^{j*} || p_{\boldsymbol{\alpha}})$. This suggests that the procedure will seek to set the hyper parameters $\boldsymbol{\alpha}$ so that the prior density $p_{\boldsymbol{\alpha}}$ of the $\boldsymbol{\lambda}_j$ on the average most resembles their (approximate) posteriors q^{j*}, based on the different genes. This connects to the recent work of van de Wiel et al. [133] on shrinkage priors for differential gene expression analysis, whose empirical Bayes procedure consists in finding $\boldsymbol{\alpha}$ such that $p_{\boldsymbol{\alpha}}(\tau_j^{-2}) \approx n^{-1} \sum_j p_{\boldsymbol{\alpha}}(\tau_j^{-2} | \mathbf{y}_j)$. Figure 1.1(b) shows that our approach fulfills the same objective. It is natural for the empirical Bayes procedure to have this "averaging of marginal posteriors" property, as it attempts to calibrate the prior according to the data. The role of the global shrinkage prior $\mathcal{G}(a, b)$ is to encourage the posterior distributions of the τ_j^{-2}, for $j \in \mathcal{J}$, to shrink to a common distribution, centered around the (prior) mean a/b.

Algorithm 2 Variational EM algorithm with global-local shrinkage priors

```
Initialize:
$a^{(0)} = b^{(0)} = a^{*(0)} = 0.001, \forall j \in \mathcal{J},\ b_j^{*(0)} = d_j^{*(0)} = 0.001,\ \xi = 10^{-3},\ M = 1000$ and $t = 1$
while $\max|\mathcal{L}_j^{(t)} - \mathcal{L}_j^{(t-1)}| \geq \xi$ and $2 \leq t \leq M$ do
    E-step: Update variational parameters:
    for $j = 1$ to $p$ do
        update $a^{*(t)} \leftarrow a^{(t-1)} + \frac{p-1}{2}$
        update $\boldsymbol{\Sigma}_j^{*(t)}$, $\boldsymbol{\beta}_j^{*(t)}$, $d_j^{*(t)}$, $b_j^{*(t)}$ and $\mathcal{L}_j^{(t)}$ in that order (as in Algorithm 1)
    end for
    M-step: Update hyper-parameters:
    $a^{(t)} \leftarrow 0.5 \left( p^{-1} \sum_{j=1}^{p} \left( \log(b_j^{*(t)}) - \psi(a^{*(t)}) \right) - \log p + \log \sum_{j=1}^{p} \frac{a^{*(t)}}{b_j^{*(t)}} \right)^{-1}$
    $b^{(t)} \leftarrow a^{(t)} \cdot p \left( \sum_{j=1}^{p} \frac{a^{*(t)}}{b_j^{*(t)}} \right)^{-1}$
    $t \leftarrow t + 1$
end while
```

1.2.4 Edge selection

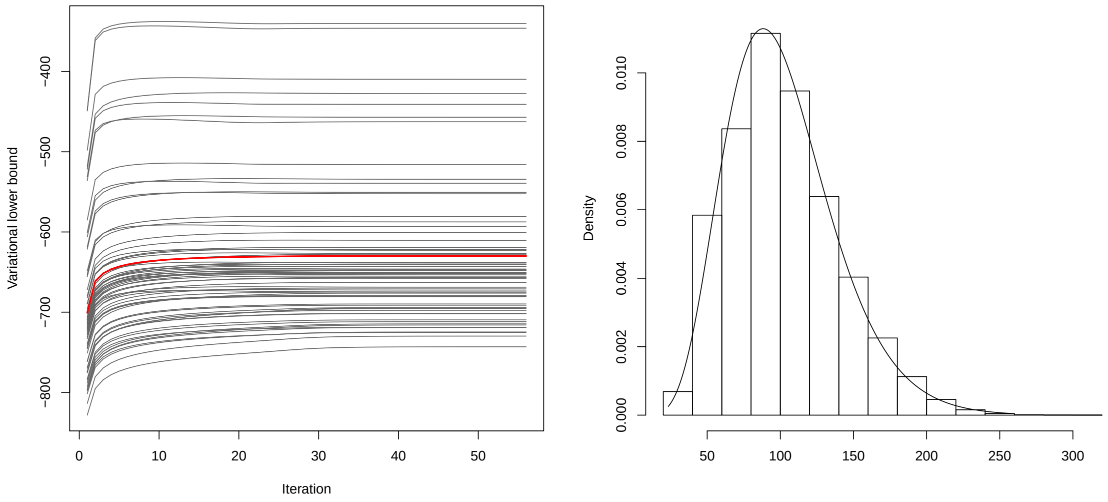
In this section we describe a separate procedure for edge selection. This consists of first ranking the edges based on summary statistics from the (marginal) posterior distributions under the model (1.2) obtained in the preceding sections, and next performing forward selection along this ordering. For the latter we use Bayes factors and their relation to a Bayesian version of the local false discovery rate [37, lfdr].

Edge ordering

Denote the approximate posterior expectation and variance of $\beta_{j,k}$ obtained in Sections 1.2.2 and 1.2.3 for SEM (1.2) by $\mathbb{E}_{q^{j*}}\left[\beta_{j,k}|\mathbf{y}_j\right]$ and $\mathbb{V}_{q^{j*}}\left[\beta_{j,k}|\mathbf{y}_j\right]$, and define

$$\kappa_{j,k} = \frac{\left|\mathbb{E}_{q^{j*}}\left[\beta_{j,k}|\mathbf{y}_j\right]\right|}{\sqrt{\mathbb{V}_{q^{j*}}\left[\beta_{j,k}|\mathbf{y}_j\right]}}, \qquad j,k \in \mathcal{J} \text{ with } j \neq k. \tag{1.12}$$

Next for a given edge (j,k) (between genes j and k) define the quantity $\bar{\kappa}_{j,k} = (\kappa_{j,k} + \kappa_{k,j})/2$, and order the set of $P = p(p-1)/2$ edges according to their associated values $\bar{\kappa}_{j,k}$, from large to small. Let $(j(r), k(r))$ denote the rth edge in this ordering,



(a) Convergence

(b) Global shrinkage prior

Figure 1.1: *Illustration of (a) the convergence of the variational algorithm and (b) the estimated global shrinkage prior on the breast cancer data set (P53 pathway). Figure (a) displays the variational lower bounds $\mathcal{L}_j$ of each regression equation in the SEM as a function of iterations. The red continuous line represents the average lower bound. Figure (b) displays an empirical mixture of marginal posteriors of τ_j^{-2} obtained by drawing 1000 samples from $q_2^j(\tau_j^{-2}; \mathbf{y}_j)$, $\forall j$. The continuous line represents the density of the estimated global shrinkage prior on τ_j^{-2}, which correspond to $\mathcal{G}(7.404, 0.073)$.*

and abbreviate its associated value to $\bar{\kappa}_r = \bar{\kappa}_{j(r),k(r)}$. This ordering is retained in all of the following. However, we do not necessarily select all edges below a certain threshold, but proceed by forward selection, for $r = 1, \ldots, P$.

Bayes factors

Selection at stage r (see Section 1.2.4) will be based on Bayes factors $\mathrm{BF}(j(r), k(r))$ and $\mathrm{BF}(k(r), j(r))$ for the two regression parameters $\beta_{j(r),k(r)}$ and $\beta_{k(r),j(r)}$ associated with the rth edge.

Denote by $m_{j(r),k(r),1}$ the model in SEM (1.2) for the response variable $\mathbf{y}_{j(r)}$, with the covariates (or nonzero $\beta_{j(r),k}$) restricted to the edge $(j(r), k(r))$ and any *previously selected* edge (involving node $j(r)$) with rank lower or equal to $r - 1$. Likewise, define $m_{j(r),k(r),0}$, but with the restriction $\beta_{j(r),k(r)} = 0$, which is equivalent to the exclusion of edge $(j(r), k(r))$. The Bayes factor associated with this model is

$$\mathrm{BF}(j(r), k(r)) = \frac{p(\mathbf{y}_{j(r)} | m_{j(r),k(r),1})}{p(\mathbf{y}_{j(r)} | m_{j(r),k(r),0})}, \qquad r = 1, \ldots, P. \tag{1.13}$$

The Bayes factor $\mathrm{BF}(k(r), j(r))$ is defined analogously from the regression models $m_{k(r),j(r),1}$ and $m_{k(r),j(r),0}$ for response variable $\mathbf{y}_{j(k)}$.

Prior for Bayesian variable selection

The global shrinkage prior for the precision parameters τ_j^{-2} estimated from the data in Section 1.2.3 is not appropriate for computing the Bayes factors (1.13). Because it has been calibrated (by the variational Bayes method outlined in Algorithm 2) for the network comprised of all edges, it is likely to be located away from zero, which will induce strong regularization on the regression parameters, making it difficult for the Bayes factors to discriminate between the subsequent models (in particular when n is small). A non-informative prior runs into the same problem (perhaps even in a more sever manner).

Motivated by the Zellner-Siow prior [90, 160] we propose to employ instead the "default prior" $\tau_j^{-2} \sim \mathcal{G}(1/2, n/2)$. This concentrates near its prior expectation n^{-1} (i.e. the fixed unit information prior of Kass and Wasserman [70]), and hence is concentrated near 0 for moderate and large values of n, while less stringent for small n (see illustration in SM Section 4).

Bayesian analogue of lfdr

Since both Bayes factors $\mathrm{BF}(j(r), k(r))$ and $\mathrm{BF}(k(r), j(r))$ are informative for the relevance of edge $(j(r), k(r))$, we need to combine these and find a suitable threshold. For that purpose, we link the Bayes factors to the posterior null-probability $\mathrm{P}_0(\bar{\kappa}_r) = P(\beta_{j(r),k(r)} = 0, \beta_{k(r),j(r)} = 0|\mathbf{y})$, where $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_p^T)^T$. The absence of edge $(j(r), k(r))$ is reflected by $\beta_{j(r),k(r)} = \beta_{k(r),j(r)} = 0$, which, in the spirit of forward selection, implies the null models $m_{j(r),k(r),0}$ and $m_{k(r),j(r),0}$. The posterior null-probability is linked to the local false discovery rate [37, lfdr]. However, as in van de Wiel et al. [133], we condition on the data $\mathbf{y}$ rather than on a test statistic. Then, we have

$$\begin{aligned} \mathrm{P}_0(\bar{\kappa}_r) &= P(\beta_{j(r),k(r)} = 0, \beta_{k(r),j(r)} = 0|\mathbf{y}) \\ &\leq \min\{P(\beta_{j(r),k(r)} = 0|\mathbf{y}), P(\beta_{k(r),j(r)} = 0|\mathbf{y})\}. \end{aligned} \tag{1.14}$$

Here, the bound is used because the SEM may not provide accurate joint probabilities on regression coefficients from different regression models. Now, assume the prior null probability $P(\beta_{j,k} = 0|\mathbf{y}_{-j}) = p_0, \forall j \in \mathcal{J}$, where $\mathbf{y}_{-j} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_{j-1}^T, \mathbf{y}_{j+1}^T, \ldots, \mathbf{y}_p^T)^T$. Note that a constant value of p_0 is reasonable, because it simply reflects the prior probability that response $\mathbf{y}_j$ does not respond to covariate $\mathbf{y}_k$ (which is a member of $\mathbf{y}_{-j}$). Then,

$$
\begin{aligned}
P(\beta_{j,k}=0|\mathbf{y}) &= P(\beta_{j,k}=0|\mathbf{y}_j,\mathbf{y}_{-j}) \\
&= \frac{P(\mathbf{y}_j|\beta_{j,k}=0,\mathbf{y}_{-j})P(\beta_{j,k}=0|\mathbf{y}_{-j})}{P(\mathbf{y}_j|\mathbf{y}_{-j})} \\
&= \frac{P(\mathbf{y}_j|m_{j,k,0})p_0}{P(\mathbf{y}_j|m_{j,k,0})p_0+(1-p_0)P(\mathbf{y}_j|m_{j,k,1})} \\
&= \frac{p_0}{p_0+(1-p_0)\mathrm{BF}(j,k)}.
\end{aligned}
\tag{1.15}
$$

Define the max Bayes factor: $\mathrm{BF}(\bar{\kappa}_r)=\max\{\mathrm{BF}(j(r),k(r)),\mathrm{BF}(k(r),j(r)\}$. Then, after substituting (1.15) into (1.14) we have, for threshold $\gamma=(1-\alpha)p_0/(\alpha(1-p_0))$,

$$
\mathrm{BF}(\bar{\kappa}_r)\geq\gamma \Longleftrightarrow \mathrm{P}_0(\bar{\kappa}_r)\leq\alpha. \tag{1.16}
$$

Equation (1.16) suggests that edges in the graph can be selected using a thresholding rule on the Bayes factors that controls the posterior null-probability. For example, when we have $p_0=0.9$, then $\mathrm{BF}(\bar{\kappa}_r)>81$ implies $\mathrm{P}_0(\bar{\kappa}_r)<0.1$. However, to use this approach an estimate of p_0 is required. We simply propose

$$
\hat{p}_0=\frac{1}{2P}\left(\sum_{r=1}^{P}(I_{\{\mathrm{BF}'(j(r),k(r))\leq 1\}}+I_{\{\mathrm{BF}'(k(r),j(r))\leq 1\}})\right). \tag{1.17}
$$

where $\mathrm{BF}'(j(r),k(r))$ is defined analogously to $\mathrm{BF}(j(r),k(r))$, but *without* forward selection (so all covariates corresponding to edge ranks $\leq r$ are included), because the forward selection procedure requires knowing $\hat{p}_0$.

Forward selection procedure

We introduce the following sequential procedure to update the set E of selected edges and the models $m_{j(r),k(r),0}$, $m_{j(r),k(r),1}$, $m_{k(r),j(r),0}$, $m_{k(r),j(r),1}$ when increasing r:

1. Initiate α, $r=1$, $\ell=0$ and $\mathsf{E}^0=\emptyset$. Compute γ from α and $\hat{p}_0$.

2. Determine the models $m_{j(r),k(r),0}$ and $m_{k(r),j(r),0}$ which are the current models for $\mathbf{y}_{j(r)}$ and $\mathbf{y}_{k(r)}$ that correspond to E^{r-1}. Augment those models with covariates $\mathbf{y}_{k(r)}$ and $\mathbf{y}_{j(r)}$, respectively, and fit these models to obtain $m_{j(r),k(r),1}$ and $m_{k(r),j(r),1}$.

3. Calculate the max Bayes factor $\text{BF}(\bar{\kappa}_r)$

4. Only if $\text{BF}(\bar{\kappa}_r) > \gamma$ update $\mathsf{E}^r = \mathsf{E}^{r-1} \cup \{(j(r), k(r))\}$

5. Update $r = r + 1$ and go back to step 2

For the purpose of variable selection we include intercepts in the SEM. Finally, we estimate $\mathcal{E}$ by the last update of E.

The selection procedure respects the initial ranking of the edges in terms of the order in which they are considered for inclusion in the forward selection. However, the procedure is set up to proceed when a given edge is not selected, because in the light of the current model subsequent edges may (slightly) increase the marginal likelihood. As in practice we observed that the Bayes factor decreases with r (see Supplementary Figure 2), a stopping criterion may be practical if P is large; e.g. stop if r reaches $r_{\max} = (1 - \hat{p}_0)P$, or if $\text{BF}(\bar{\kappa}_r)$ has not exceeded γ for, say, 100 consecutive values of r.

1.2.5 Computational considerations

In Algorithm 1 and 2 it is generally preferable to reparameterize the model relative to the principal components of $\mathbf{X}^T\mathbf{X}$. This way the variational updates and lower bound can be modified to achieve important computational savings (see SM Section 2). For edge selection, when the number of edges is large it is preferable to approximate (1.17) using a random subset of, say, 1000 edges. With these considerations the proposed methodology is shown to be computationally attractive (see Table 1.1 and SM Section 13).

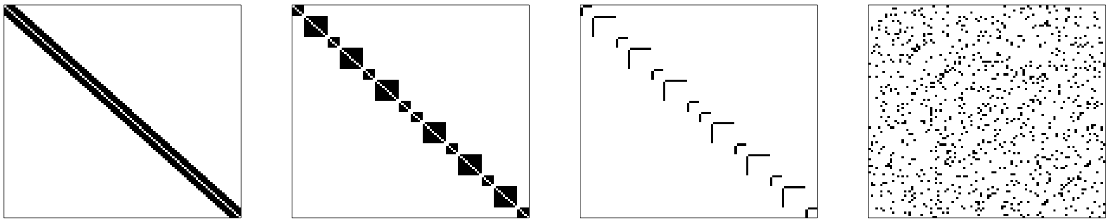
	$p = 50$	$p = 100$	$p = 200$	$p = 500$	$p = 1000$
$n = 50$	0:00:01	0:00:10	0:00:08	0:00:52	0:08:51
$n = 100$	0:00:01	0:00:21	0:00:31	0:01:50	0:12:02
$n = 200$	0:00:02	0:00:40	0:01:20	0:05:25	0:21:14
$n = 500$	0:00:07	0:01:12	0:02:14	0:23:42	1:51:21

Table 1.1: *Average elapsed time (H:MM:SS) as a function of the number of samples n and variables p. For n and p fixed, 10 random data sets were generated from the complete Breast cancer data set (Section 1.4.1). When $p > 100$ we approximated (1.17) using a random subset of 1000 edges. Computations were made on 2.60GHz CPU without parallelization strategy.*

For very large p, `ShrinkNet` contains an option to restrict the number of reported edges, e.g. to 1000, which may be practical from both a computational and interpretational point of view. Then, when $n = 200$, computing times drop to 5 and 21 minutes for $p = 500$ and $p = 1000$, respectively. For the curated Breast cancer data used by Schäfer and Strimmer [121, 49 samples and 3,883 genes], `ShrinkNet` takes 2 hours and 15 minutes when the forward selection is limited to the top 10,000 edges.

1.3 Model-based simulation

In this section we investigate the performance of our approach, termed ShrinkNet, in recovering the structure of an undirected network and compare it to popular approaches. We generate $n \in \{25, 50, 100\}$ samples from a multivariate normal distribution with mean vector $\mathbf{0}$ and 100×100 precision matrix $\boldsymbol{\Omega}$, corresponding to four different graph structures: *band*, *cluster*, *hub* and *random* [163] (see Figure 1.2 for illustration), every of them sparse, with graph density ranging from 0.017 to 0.096. We generated the inverse covariance matrix $\boldsymbol{\Omega}$ corresponding to each graph structure from a G-Wishart distribution [100] with scale matrix equal to the identity and $b = 4$ degrees of freedom. In SM Section 2 we provide statistical summaries on the magnitude of the generated partial correlations.



(a) Band (b) Cluster (c) Hub (d) Random

Figure 1.2: *Graph structures considered for the precision matrix* $\boldsymbol{\Omega}$ *in our simulation. Black and white dots represent non-zero and zero entries, respectively. Only off-diagonal elements are displayed. For precision matrices with block-diagonal structures (clusters and hubs), block sizes were set to 5 and 10. In (a) the bandwidth is equal to four. The graph density* δ *is (a)* $\delta = 0.079$*, (b)* $\delta = 0.071$*, (c)* $\delta = 0.017$ *and (d)* $\delta = 0.096$.

We compared our approach ShrinkNet to the popular frequentist SEM with the Lasso penalty (SEM_L) [97], the Graphical Lasso (GL_λ) [43], and GeneNet [119]. The latter combines a non-sparse linear shrinkage estimator with an *a posteriori* edge selection procedure. For the purpose of comparison with ShrinkNet, we also consider the Bayesian SEM (1.2) with the non-informative global shrinkage prior $\mathcal{G}(0.001, 0.001)$, which we subsequently refer to as 'NoShrink'.

Briefly, graph selection is as follows. For $\mathrm{SEM_L}$ and GL_λ we use the EBIC criterion [23, 41] for selecting the optimal regularization parameter(s), whereas for GeneNet and ShrinkNet a threshold of 0.1 on the local false discovery rate and the posterior null probability P_0 is employed. In SM Section 3 we provide more details as to how an edge ranking is obtained for each method.

To evaluate the performance of the methods in recovering the graph structures we report partial ROC curves (SM Section 5), which depict the true positive rate (TPR) as a function of the false positive rate (FPR) for FPR$<$ 0.2), and various performance measures on selected graphs. Figure 1.3 below displays boxplots of F-scores and partial AUCs (pAUC) [35] as a function of the method, n and the true graph structure. The F-score=2 × (precision × TPR)/(precision + TPR) is a popular performance measure, defined as the harmonic mean between the TPR=TP/(TP+FN) (also called *recall*) and the precision=TP/(TP+FP), where TP, FP, and FN are the number of true positives, false positives, and false negatives, respectively.

Figure 1.3 shows that ShrinkNet achieves the highest partial AUCs in almost all situations. The results also indicate that NoShrink is often outperformed by GeneNet, and comparable to GL_λ, which suggests that the global shrinkage carried out by ShrinkNet considerably improves edge ranking. $\mathrm{SEM_L}$ has the lowest pAUC in almost all situations.

The performance of each method in recovering the true graph structure can also be evaluated by the F-score. According to this metric the best performance is achieved by NoShrink and ShrinkNet in all but two cases. In moderate- ($n = 50$) and high-dimensional cases ($n = 25$), NoShrink and ShrinkNet show a much larger F-score than others. This is particularly pronounced when $n = 25$, in which case GL_λ and GeneNet have an F-score (and TPR) very close to zero. In this context $\mathrm{SEM_L}$ is performing better than GL_λ and GeneNet, but worse than NoShrink and ShrinkNet.

All in all, the simulation study demonstrates that global shrinkage considerably improves edge ranking. For network reconstruction, the small discrepancy between ShrinkNet and NoShrink indicates that the selection procedure of Section 1.2.4 is relatively robust to edge ranking. The proposed selection procedure is also shown to outperform contenders in the most high-dimensional cases.

1.4 Data-based simulation

In this section we employ gene expression data from The Cancer Genome Atlas (TCGA) to compare the performance of our approach in reconstructing networks

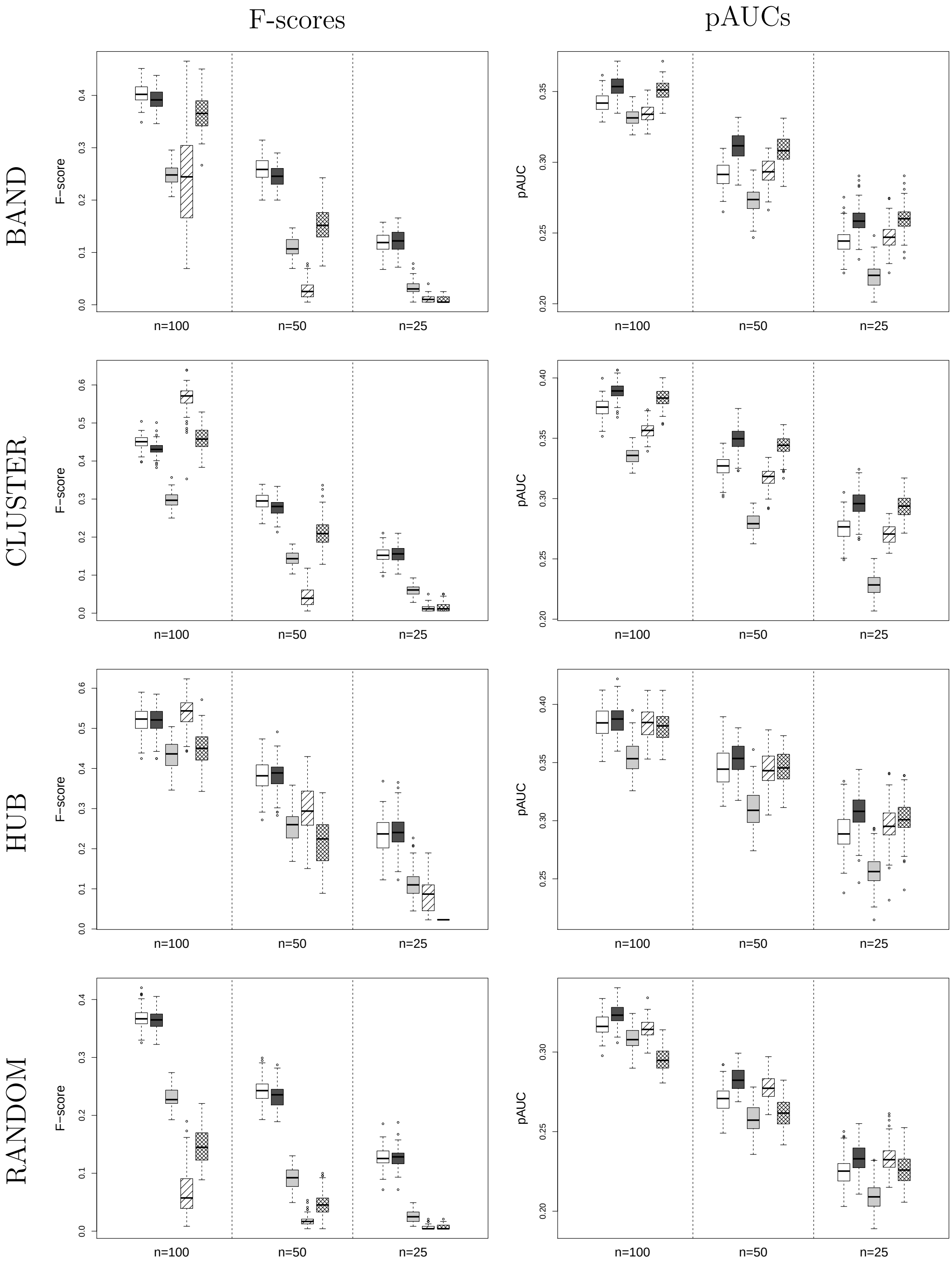


Figure 1.3: *Boxplots of F-scores (left column) and pAUCs (right column) over 100 repetitions as a function of the method,* n *and the true graph structure. The five methods under comparison are from left to right: NoShrink (white), ShrinkNet (dark grey),* SEM_L *(light grey),* GL_λ *(diagonal pattern) and GeneNet (mesh pattern)*

.

with $\mathrm{SEM_L}$, GL_λ, GeneNet and NoShrink (see previous Section). Data were retrieved from the TCGA cBioPortal using the R package 'cgdsr' [21, 67]. In particular, we focus on the p53 pathway in the Breast cancer data set ($n_{\mathrm{brca}} = 526$), which comprise $p^{\mathrm{p53}} = 67$ genes, and the apoptosis pathway in the Ovarian data set ($n_{\mathrm{ov}} = 537$) that comprises $p^{\mathrm{apopt}} = 79$ genes. Since the true molecular network is not exactly known, we employ a random splitting strategy for the two data sets to assess discoveries.

1.4.1 Reproducibility

To compare reproducibility, we randomly split the data into a small data set where $n^{\mathrm{p53}}_{\mathrm{small}} \in \{134, 67, 34\}$ and $n^{\mathrm{apopt}}_{\mathrm{small}} \in \{158, 79, 40\}$ to achieve low-, moderate- and high-dimensional situations, and a large data set where $n^{\mathrm{p53}}_{\mathrm{large}} \in \{392, 459, 492\}$ and $n^{\mathrm{apopt}}_{\mathrm{large}} \in \{379, 458, 497\}$ (representing the complement). The large data set is then used to validate discoveries made from the small one. As a benchmark for validation we employ the edge set $\mathcal{S}_{\mathrm{b}}$ defined by edges that are simultaneously selected by the different methods based on the large data set. Because the lack of consensus between the different methods may render $\mathcal{S}_{\mathrm{b}}$ too small, we only compare two methods at a time.

To assess performance in recovering $\mathcal{S}_{\mathrm{b}}$ from the small data set we generate 100 random data splits and report average partial ROC curves and average TPR and FPR from the selected graphs. Figure 1.4 summarizes results for the four pairwise comparisons of GeneNet, $\mathrm{SEM_L}$, GL_λ and NoShrink with ShrinkNet for the apotosis pathway in the Ovarian cancer data set. Simulation results for the p53 pathway for the Breast cancer data are provided in SM Section 7. Table 1.2 and Supplementary Table 2 summarize the number of selected edges in the small and large data sets for each method.

	$n^{\mathrm{apopt}}_{\mathrm{small}} =$ 158	$n^{\mathrm{apopt}}_{\mathrm{large}} =$ 379	$n^{\mathrm{apopt}}_{\mathrm{small}} =$ 79	$n^{\mathrm{apopt}}_{\mathrm{large}} =$ 458	$n^{\mathrm{apopt}}_{\mathrm{small}} =$ 40	$n^{\mathrm{apopt}}_{\mathrm{large}} =$ 497
ShrinkNet	62.5 (5.7)	138.6 (5.9)	31.4 (5.1)	166.9 (6)	18.2 (4.8)	179.6 (5.6)
$\mathrm{SEM_L}$	16.0 (3.9)	54.0 (5.3)	4.7 (2.3)	65.1 (4.7)	1.6 (1.2)	69.2 (4)
GL	25.8 (10.6)	145.7 (35.5)	9.6 (4.7)	224.1 (56)	5.3 (3.2)	282.2 (58.1)
GeneNet	10.2 (4.6)	22.9 (4.6)	2.2 (2.3)	25.8 (3.5)	0.3 (1.5)	26.1 (2.4)

Table 1.2: *Average number of selected edges (and standard deviations in parentheses) for each method in the small and large data sets over 100 random partitioning of the Ovarian cancer data set.*

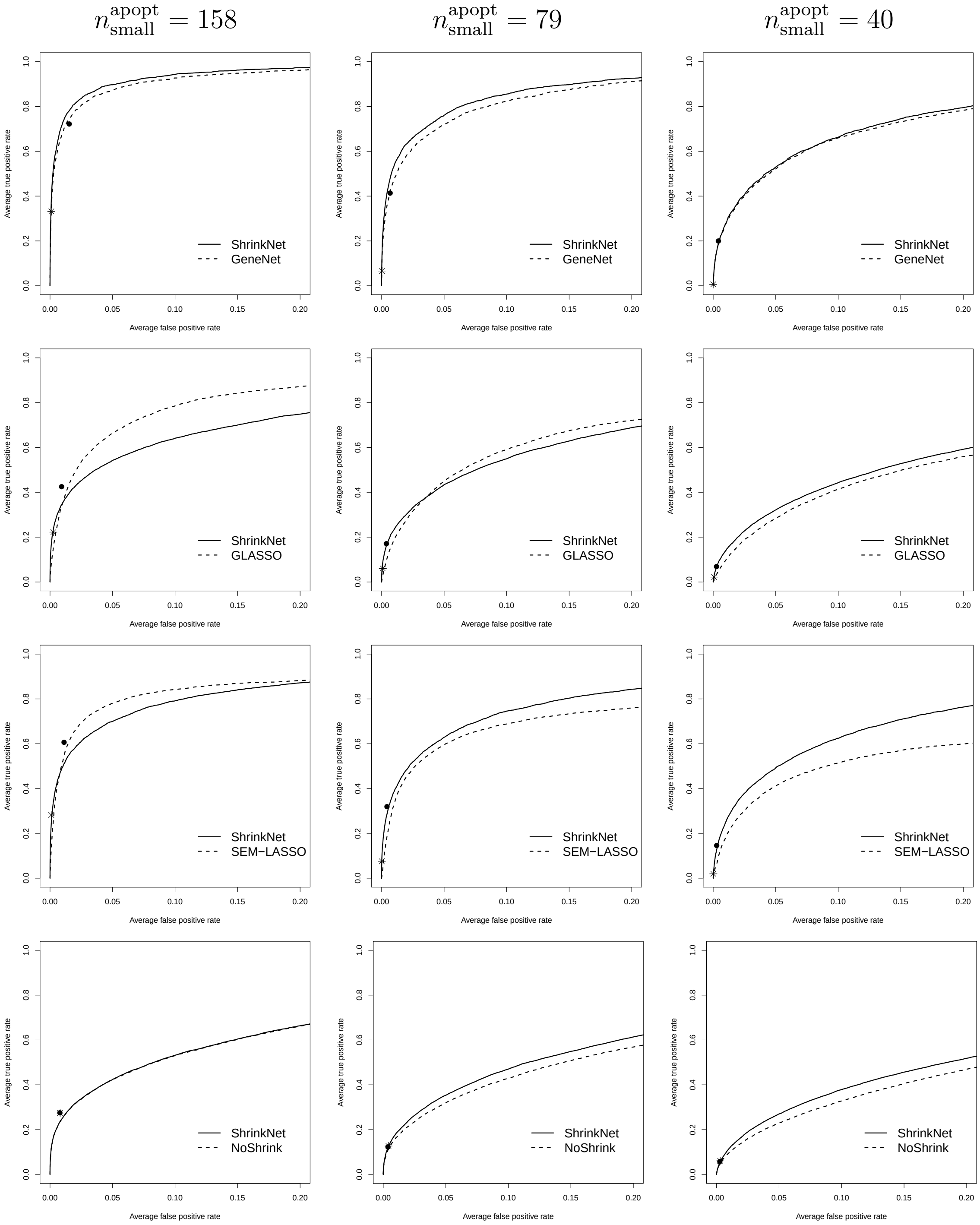

$n_{\text{small}}^{\text{apopt}} = 158$ $n_{\text{small}}^{\text{apopt}} = 79$ $n_{\text{small}}^{\text{apopt}} = 40$

Figure 1.4: *Average partial ROC-curves corresponding to all pairwise comparisons of GeneNet, GL$_\lambda$, SEM$_L$ and NoShrink with ShrinkNet when the apoptosis data are randomly split into a small data set of size $n_{small}^{p53} \in \{134, 67, 34\}$ and a large validation one of size $n_{large}^{p53} \in \{392, 459, 492\}$. Each plot depicts the performance of ShrinkNet (black continuous line) versus one of the contenders (black discontinuous line). Circle (ShrinkNet) and star (contender) points correspond to average TPR and FPR of selected graph structures as obtained by the two inference methods under comparison. Note that the circle point is not expected to be located on the curve.*

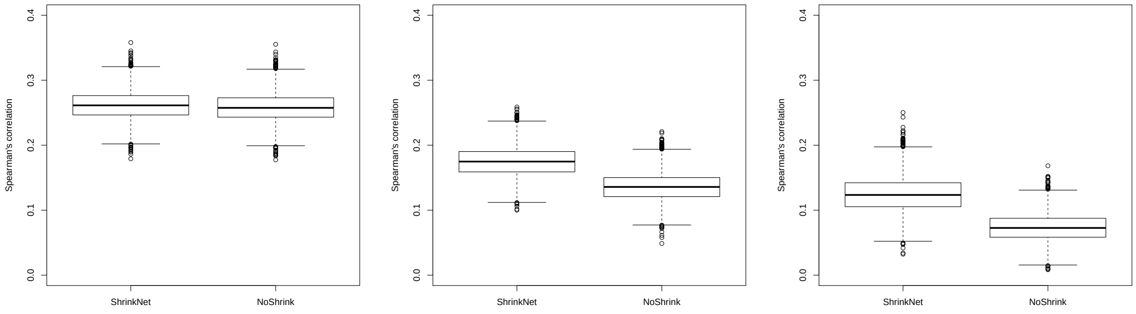
The number of selected edges differs a lot between GeneNet, SEM_L, GL_λ and ShrinkNet (Table 1.2). GeneNet is the most conservative approach whereas ShrinkNet selects more edges than others in the small data sets. However, when the sample size is large GL_λ selects more than ShrinkNet, as illustrated by the number of discoveries in the large data sets. It is interesting to see in Table 1.2 that ShrinkNet is remarkably stable in selection. The variability (as measured by the standard deviations) of the number of selected edges is relatively low, and in fact surprisingly constant in the small and large data sets, regardless of the number of selected edges. Conversely, GL_λ exhibits relatively larger variability and also large differences in number of edges.

The results in Figure 1.4 suggest that ShrinkNet compares very favourably to the other methods in recovering the benchmark edge set $\mathcal{S}_\text{b}$. In particular, edge selection (as represented by dots in the ROC plots) is shown to outperform the other methods clearly in all situations. In the most high-dimensional case $n_\text{small}^\text{apopt} = 40$, GeneNet, SEM_L and GL_λ detect almost no edges in the small data set (see Table 1.2), whereas ShrinkNet still detects a non-negligible number of edges, which translates into a higher TPR (with negligible FPR). Partial ROC curves in Figure 1.4 also indicate that edge ranking as provided by ShrinkNet is often superior to others. This is particularly true when $n_\text{small}^\text{apopt} = 79$ and $n_\text{small}^\text{apopt} = 40$. In case $n_\text{small}^\text{apopt} = 158$, SEM_L and GL_λ outperform ShrinkNet for edge ranking, but not for edge selection. This suggests that the selection procedure proposed in Section 1.2.4 is robust to the edge ranking on which it is based. This is confirmed by comparing ShrinkNet with NoShrink, where there is no difference in selection performance, whereas edge ranking appears to be improved by the global shrinkage prior.

Finally Figure 1.5 displays rank correlation of edges between all pairs of data sets of size $n_\text{small}^\text{apopt}$ for ShrinkNet and NoShrink. The correlations are clearly higher for ShrinkNet than for NoShrink when $n_\text{small}^\text{apopt} \in \{79, 40\}$, which indicates that the global shrinkage improves the stability and, hence, reproducibility of edge ranking when the sample size $n_\text{small}^\text{apopt}$ is not large.

1.4.2 Stability

In this section, the random splitting strategy is used to study the stability of edges selected by each method. Let $\hat{\pi}_{ij}$ be the empirical selection probability of edge (i, j) for a given method over the 100 generated small data sets of size $n_\text{small}^\text{apopt}$. We define the set of stable edges by $S_\text{stable} = \{(i, j) : \hat{\pi}_{ij} \geq \pi_\text{thr}\}$ where $0.5 < \pi_\text{thr} \leq 1$. To determine an appropriate cut-off π_thr, which is comparable between methods, we use



(a) $n_{\text{small}}^{\text{apopt}} = 158$ (b) $n_{\text{small}}^{\text{apopt}} = 79$ (c) $n_{\text{small}}^{\text{apopt}} = 40$

Figure 1.5: *Correlations of edge ranking as provided by ShinkNet and NoShrink across the 100 generated small data sets of size n_{small}^{apopt}. Each boxplot displays Spearman rank correlations between the values of $\bar{\kappa}_r$, $r = 1, \ldots, P$, obtained from all the* $(100 \times 99)/2 = 4950$ *pairs of data sets of size n_{small}^{apopt} for each of the two methods. Note that one does not expect high rank correlation when considering all edges.*

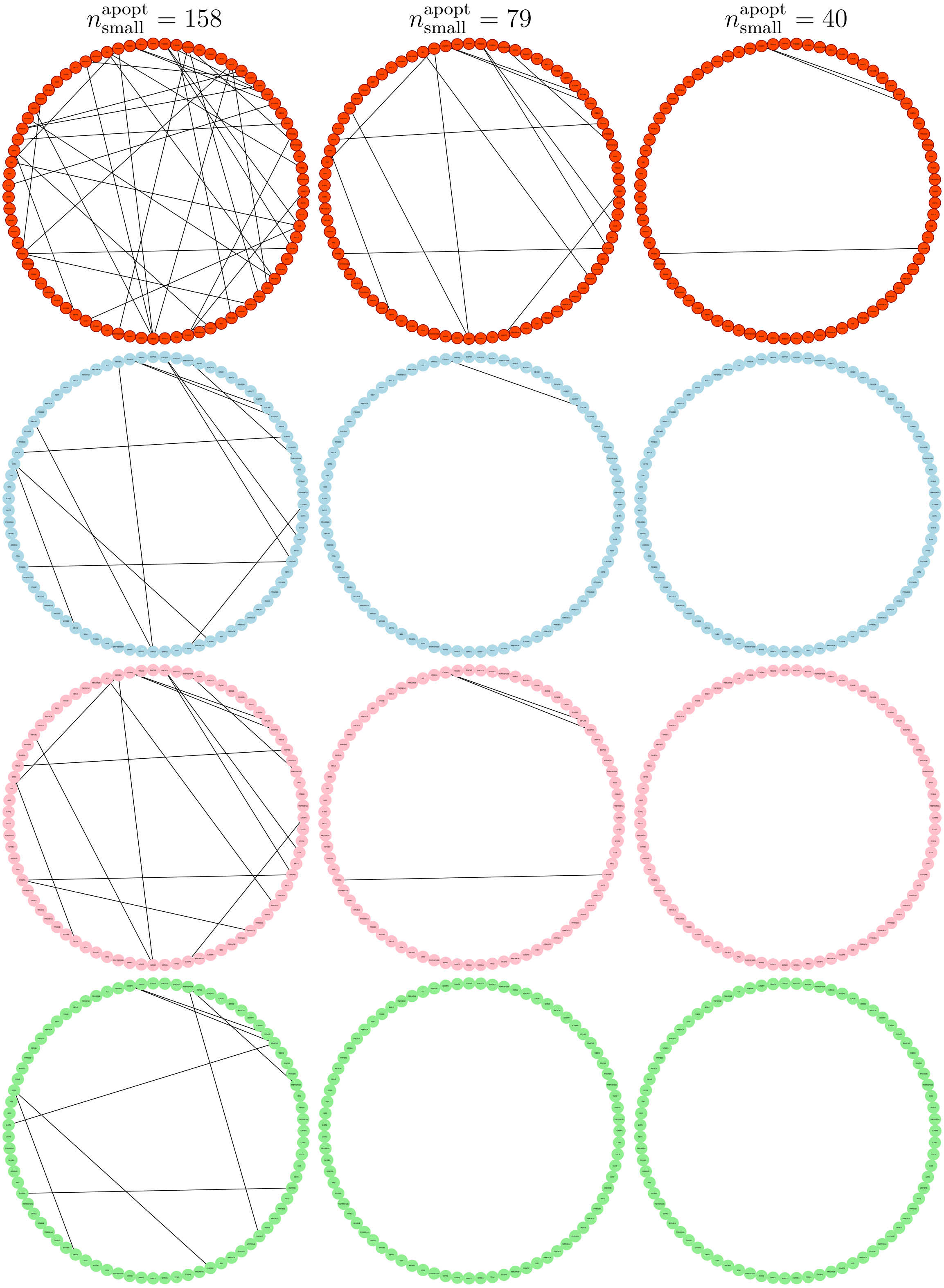
the stability criterion proposed by [98]. This is based on the following upper bound on the expected number $\mathbb{E}(V)$ of falsely selected edges:

$$\mathbb{E}(V) \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)P}, \tag{1.18}$$

where q is the expected number of edges selected by the given method and P is the total number of edges ($P_{\text{apopt}} = 3081$ and $P_{\text{p53}} = 2211$). To compare the set of stable edges between the different methods, we set $\mathbb{E}(V) = 30$ as in Meinshausen and Bühlmann [98]. Then, π_{thr} (and hence S_{stable}) is determined using an empirical estimate of q (see Table 1.2 and SM Table 2). Because the type I error is controlled in the same way for all methods, comparison can reasonably be based on the number of stable edges.

To illustrate, when $n_{\text{small}}^{\text{apopt}} = 158$ for the apoptosis data we obtain that $\pi_{\text{thr}}^{\text{ShrinkNet}} = 0.623$, $\pi_{\text{thr}}^{\text{SEM}_\text{L}} = 0.508$, $\pi_{\text{thr}}^{\text{GL}_\lambda} = 0.522$ and $\pi_{\text{thr}}^{\text{GeneNet}} = 0.503$, which result in 27, 12, 12 and 8 stables edges, respectively. These are illustrated in the left column of Figure 1.6. As $\mathbb{E}(V)$ is fixed, the value of π_{thr} only varies between methods because estimates of q differ. This is intuitive: if the method selects a lot of (few) edges we expect π_{thr} to be large (small).

Figure 1.6 and SM Figure 10 display stables edges obtained with each method as a function of $n_{\text{small}}^{\text{apopt}}$ and $n_{\text{small}}^{\text{p53}}$, respectively. For the two data sets ShrinkNet selects an important number of stable edges. This is particularly true for the apoptosis pathway



$n_{\text{small}}^{\text{apopt}} = 158$ $n_{\text{small}}^{\text{apopt}} = 79$ $n_{\text{small}}^{\text{apopt}} = 40$

Figure 1.6: *Stable edges for the apoptosis pathway obtained with ShrinkNet (red), SEM$_L$ (blue), GL (pink) and GeneNet (green) when $\mathbb{E}(V) = 30$ as a function of n_{small}^{apopt}. Plots were generated using the R CRAN package rags2ridges.*

where the method clearly yields more stable edges than SEM_L, GL_λ and GeneNet in all situations. Specifically, when $n_{\text{small}}^{\text{apopt}} = 79$ ShrinkNet identifies a nearly identical network to GL_λ and SEM_L when $n_{\text{small}}^{\text{apopt}} = 158$. For the p53 pathway (see SM Figure 10), GL_λ detects more stable edges than ShrinkNet when $n_{\text{small}}^{\text{p53}} = 134$, as many as when $n_{\text{small}}^{\text{p53}} = 67$, and less when $n_{\text{small}}^{\text{p53}} = 40$. This suggests that when the sample size is small ShrinkNet tends to select more stable edges than GL_λ. Finally, for the two data sets ShrinkNet detects more stable edges than SEM_L and GeneNet.

1.5 Real data application

Glioblastoma multiform (GBM) is a common and aggressive form of brain tumor in adults which, unfortunately, is also one of the most malignant of glial tumors. Patients with GBM have a poor prognosis and usually survive less than 15 months following diagnosis. GBM mRNA expression and clinical data (level 3 normalized; Agilent 244K platform) were obtained from the TCGA data portal (tcga-data.nci.nih.gov). The data contained measurements of 17,814 genes in tumor tissue samples from 532 GBM patients, of whom 505 had available survival information. Missing expression values were imputed using the R function impute.knn (using default parameters) from the Bioconductor package `impute`. Instead of characterizing globally the interactions between all genes, we focused on the subset of the 66 genes with the strongest association with patient survival (FDR$\leq$0.01). These genes are expected to be related via the different biological processes that promote cancer and thereby impact survival. ShrinkNet was then used to identify the potential relationships between these genes, which may help to further prioritize them (e.g. by node degree) and their potential interactions (e.g. by edge strength). Indeed, highly connected 'hub' genes are thought to play an important role into the disease biology.

Figure 1.7 displays the undirected gene network reconstructed by ShrinkNet using $\alpha = 0.10$ (Bayesian analogue of lfdr; see Section 1.2.4). The graph comprises 260 edges which corresponds to a density of 0.12. Node degrees vary from 2 to 13. Among the genes with highest degree (see SM Section 12), known regulators are found. For example, LGALS1 (degree equal to 13) encodes the Galectin-1 protein which is a multifaceted promoter of glioma malignancy [17]. This protein instigates increased glioma invasiveness and its expression correlates directly with tumour grade [40]. SLC16A3 (also with degree equal to 13) encodes for the MCT4 protein whose overexpression has been reported in several solid tumors, including metastases of breast cancer to the brain, which suggests its association with aggressive tumor behavior

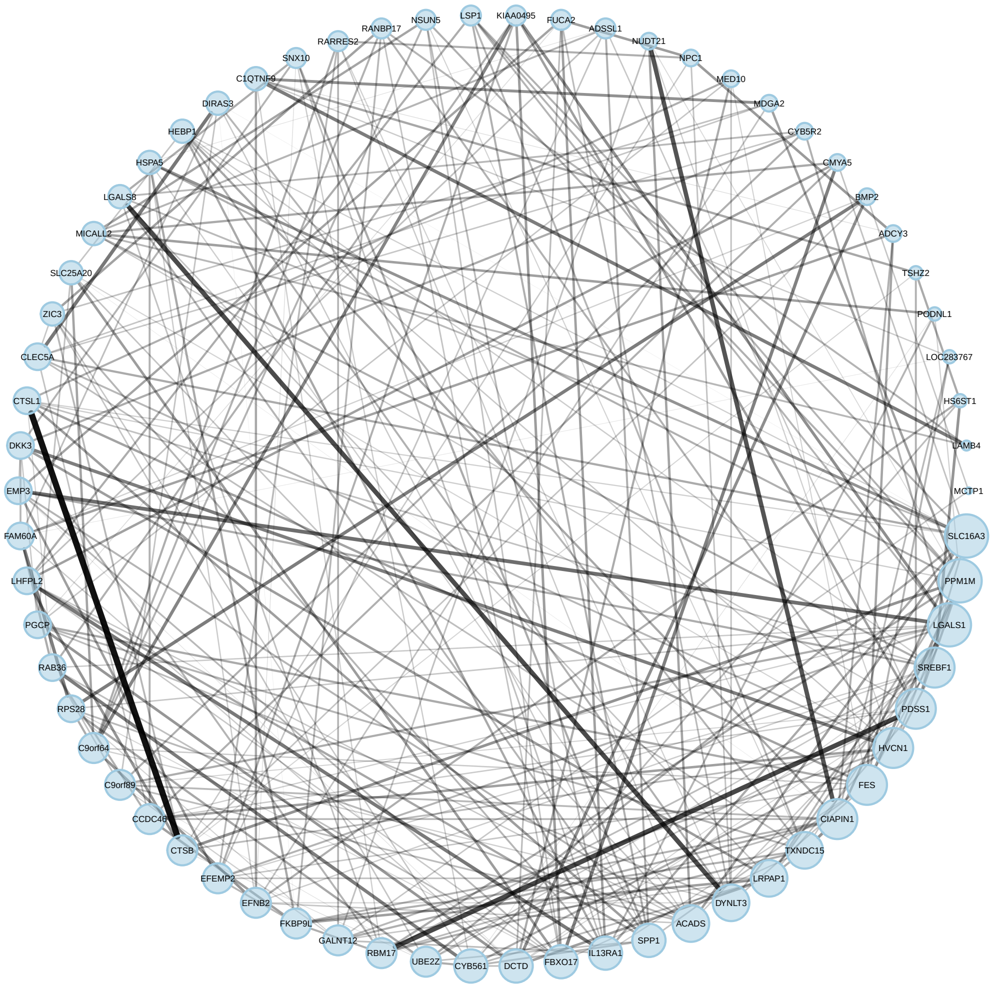


Figure 1.7: *Reconstructed network for the 66 genes associated with patient survival in GBM. Node size is proportional to the node degree and edge width/opacity is proportional to* $\bar{\kappa}_{j,k}$.

[91]. SREBF1 (degree equal to 12), also known as SREBP1, is a protein regulating lipid composition that has been associated with the proliferation of cancer cells. SREBP1 activity is known to be regulated by the Akt/mTORC1 signaling axis that is responsible for the growth and survival of cancer cells by sustaining lipid biosynthesis [85, 112]. As a final example, IL13RA1 (degree equal to 10) encodes for a protein belonging to the interleukin-13 (IL-13) receptor that elicits both proinflammatory and anti-inflammatory immune responses, and is strongly associated with Glioblastoma [95]. IL-13 has been widely suggested for cancer drug therapy.

Multiple links that are identified by ShinkNet were also previously reported in relation to Glioblastoma. Using the complete human protein interaction network from Pathway Commons (pathwaycommons.org; Cerami et al. [22]) we could validate several edges identified by ShrinkNet (see SM Section 12). This is true in particular for the most significant edge (as measured by $\bar{\kappa}_{j,k}$; see Section 1.2.4), which links genes CTSB and CTSL1. These genes participate in protein degradation and turnover [27]. This finding hence supports the idea that cathepsins participate in enhancing invasion and metastasis [51, 69], both so descriptive of GBM. Besides, the database also confirmed the following interactions found by ShrinkNet: LGALS1 $\leftrightarrow$ RPS28, HSPA5 $\leftrightarrow$ SLC16A3, ACADS $\leftrightarrow$SLC16A3, and ACADS $\leftrightarrow$ HSPA5.

1.6 Conclusion

In this paper we proposed a Bayesian SEM with global-local shrinkage priors for gene network reconstruction. The model employs simple conjugate priors to impose regularization. Because these are not sparse, a novel method for a posteriori edge selection was introduced to infer the graph structure. Computational efficiency was achieved by SVD decompositions and fast variational approximations. We discussed empirical Bayes estimation of prior hyper-parameters and embedded this in a variational EM-type algorithm. The simulations showed that the proposed approach is often superior to popular (sparse) methods in low-, moderate- and high-dimensional cases. In particular, on real data the method yielded more stable and reproducible discoveries. Network analysis of genes associated with patient survival in Glioblastoma confirmed the method's ability to discover biologically meaningful interactions and hub genes. Our method, termed ShrinkNet, is implemented as an R package and available at http://github.com/gleday/ShrinkNet.

A novelty of our work is the use of *global* shrinkage priors, which allow the borrowing of information across regression equations. We are not aware of any previous works combining global and local shrinkage priors. In the frequentist setting Yuan et al. [159] borrows information across the regularizing parameters corresponding to ℓ_1-penalties by combining local and global searches. In the Bayesian setting the focus is often on studying the equivalence between the SEM and a proper joint distribution [33, 47]. In this paper we have shown that the combined use of global and local shrinkage priors improves statistical inference, in particular edge ranking.

Our variable selection method performs simultaneous selection of the two parameters that are associated with each edge, but unlike sparsity-based methods performs separate estimation and selection steps. However, separating estimation and selection may also come as an advantage in terms of optimizing performance with respect to either of these criteria. In fact, "The idea of pre-ranking covariates and then selecting models has become a well established technique in the literature" [66, Remark 6].

An important practical advantage of our approach is that the estimation procedure is coherent and complete, and does not rely on tuning, resampling, or cross-validation to set regularization parameter(s). This is particularly encouraging for extending the method to settings with multiple types of high-dimensional covariates, which would require different amounts of shrinkage. For methods based on resampling or cross-validation this may become overly computationally burdensome.

The proposed method is particularly suitable for gene network reconstruction using expression data. This type of network aims at providing a picture of regulatory mechanisms that act between genes. In practice, the interest often lies in a relatively small subset of genes that are known to be functionally linked (e.g. a pathway). In this context the Bayesian SEM may be more appropriate than others, because such a gene set is usually of moderate dimension and, hence, due to the functional link, the corresponding network is likely to be relatively less sparse. Therefore strong dependencies between genes are more likely to occur and this may favor Normal-Gamma (ridge-type) regularization. In addition, the coherence in functionality may render shrinkage beneficial for parameter estimation in the SEM.

We have focused on recovering the support of the precision matrix, but it is also possible to obtain an estimate of it. An immediate approach is to use the graph structure provided by ShrinkNet as a prior for precision estimation (sometimes referred to as *parameter learning* [124]). Versions of the Wishart distribution, such as the G-Wishart [34, 144], are computationally attractive. Other estimation strategies have been proposed outside the Bayesian paradigm. See, for example, Zhou et al. [164] and Yuan [157].

We foresee several extensions. SEMs are appropriate to describe directed networks and it would be interesting to investigate different types of shrinkage priors suitable in this context, for example to shrink in- and outgoing edges differently. Extension to non-Gaussian data is possible, where it may be desirable to adopt a flexible likelihood model and other types of posterior approximations may be considered [118]. Finally the model suits construction of integrative networks when allowing different priors for different types of interactions.