Prior information and variational Bayes in high dimensional statistical network inference

Kpogbezan, G.B.

Citation

Kpogbezan, G. B. (2018, December 10). *Prior information and variational Bayes in high dimensional statistical network inference*. Retrieved from https://hdl.handle.net/1887/67526

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden]

Downloaded from: [https://hdl.handle.net/1887/67526]

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

Leiden University Repository

The handle http://hdl.handle.net/1887/67526 holds various files of this Leiden University dissertation.

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

Prior information and variational Bayes in high dimensional statistical network inference

PROEFSCHRIFT

ter verkrijging van
de graad van Doctor aan de Universiteit Leiden,
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op maandag 10 december 2018
klokke 10.00 uur

door

Gino Bertrand Kpogbezan

geboren te Cotonou in 1981

Promotoren:	prof.dr. A.W. van der Vaart	Universiteit Leiden
	prof.dr.ir. M.A. van de Wiel	Vrije Universiteit Amsterdam
Copromotoren:	dr. W.N. van Wieringen	Vrije Universiteit Amsterdam
Promotiecommissie:	prof.dr. J. Meulman	Universiteit Leiden/Stanford University (secretaris)
	prof.dr. J.J. Goeman	Universiteit Leiden
	dr. S.L. van der Pas	Universiteit Leiden
	dr. G.G.R. Leday	University of Cambridge

Research supported by the European Research Council under ERC Grant Agreement 320637.

Contents

Contents **iii**

Introduction **1**

0.1 Graphical models 1

0.2 Prior information 4

0.3 Variational Bayes approximation 5

0.4 Outline of this thesis 8

Chapter 1: Gene network reconstruction using global-local shrinkage priors **11**

1.1 Introduction 12

1.2 Methods 14

1.2.1 The Bayesian SEM 14

1.2.2 Variational approximation to posteriors 16

1.2.3 Empirical Bayes and prior calibration 19

1.2.4 Edge selection 21

1.2.5 Computational considerations 25

1.3 Model-based simulation 26

1.4 Data-based simulation 27

1.4.1 Reproducibility 29

1.4.2 Stability 31

1.5 Real data application 34

1.6 Conclusion 36

Chapter 2: An empirical Bayes approach to network recovery using external knowledge **39**

2.1 Introduction 40

2.2 Model 44

2.3 Variational Bayes method and Gibbs sampling 45
2.3.1 Variational Bayes inference. 46
2.3.2 Variational Bayes vs Gibbs sampling. 48
2.4 Global empirical Bayes for BSEMed 49
2.5 Numerical investigation 50
2.6 Illustration 54
2.7 Discussion 57

Chapter 3: Incorporating prior information and borrowing information in high-dimensional sparse regression using the horseshoe and variational Bayes **61**
3.1 Introduction 62
3.2 Network reconstruction 66
3.3 eQTL mapping 68
3.4 Posterior inference 69
3.4.1 Variational Bayes approximation 69
3.4.2 Global Empirical Bayes 72
3.4.3 Variable selection 74
3.5 Simulations 76
3.5.1 Model-based simulation 76
3.5.2 Variational Bayes vs MCMC 80
3.6 Applications 82
3.6.1 Reconstruction of the apoptosis pathway 82
3.6.2 eQTL mapping of the p38MAPK pathway 83
3.7 Conclusion 88
3.8 Appendix 91

Chapter 4: Borrow network information between observational and time-course studies: explorations **99**
4.1 Introduction 100
4.1.1 Related work 101
4.2 Shared precision information 102
4.2.1 Joint precisions estimation 103
4.2.2 Borrowing information from the time-course study 104
4.2.3 Illustration 105
4.3 Shared parameter information 107
4.3.1 Symmetric or triangular $\mathbf{A}^{(h)}$ and known $\boldsymbol{\Sigma}_{\varepsilon}^{(h)}$ 108

4.3.2 $\mathbf{A}^{(h)} = \delta\mathbf{A}^{(c)}$ and a diagonal $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ 108
4.3.3 Sparse $\mathbf{A}^{(h)}$, shared support and a diagonal $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ 110
4.3.4 Evaluation of the diagonal $\mathbf{\Sigma}_{\varepsilon}$ assumption 113
4.3.5 Illustration 114
4.4 Recovering the precision matrix from the human VAR(1) model parameters 116
4.5 Conclusion 117
4.6 Appendix I 118

Bibliography **121**

Samenvatting **133**

Summary **135**

Acknowledgements **137**

Biography **139**

Introduction

The rapid evolution of data acquisition technologies in the last 25 years has enabled a massive production of high-dimensional and highly complex datasets in many scientific domains, including genomics, finance and statistical pattern recognition, to name but a few. In genomics, high-throughput platforms such as microarrays provide measurements of many thousands of molecular aspects (e.g. gene expression) of the cell. While between 20,000 and 25,000 genes of a single patient are easily characterized simultaneously, the number n of patients runs typically in the tens or hundreds. This typically gives rise to data characterized by 'large p, small n'. The analysis of such high-dimensional data ($n \ll p$) is very challenging as the traditional statistical methods become useless. For instance, the sample covariance matrix becomes rank deficient and can not be inverted. Our contribution in this thesis consists of incorporating prior knowledge in the analysis of these data, which we model using mainly graphical models.

0.1 Graphical models

A graphical model is a way to 'marry' probability theory with graph theory. A graph $\mathcal{G}$, as used in this thesis, is a pair $(\mathcal{I}, \mathcal{E})$, where $\mathcal{I}$ is a set of indices (or vertices) and the set of edges $\mathcal{E}$ is a subset of the set $\mathcal{I} \times \mathcal{I}$ of ordered pairs of distinct vertices . An edge between vertices r and s is *undirected* if both (r, s) and (s, r) are in $\mathcal{E}$, whereas an edge $(r, s) \in \mathcal{E}$ whose opposite $(s, r) \notin \mathcal{E}$ is called *directed.* In the diagram of $\mathcal{G}$ an *undirected* edge is usually represented by a line between the corresponding vertices whereas a *directed* edge is represented by an arrow. A graph is called undirected if it possesses only undirected edges, and it is called directed if all edges are directed. Let $Y = (Y_1, Y_2, \cdots, Y_p)$ denote a random field with index set $\mathcal{I} = \{1, 2, \cdots, p\}$ taking values in probability spaces $\mathcal{Y}_i$, $i \in \mathcal{I}$ and $\mathcal{Y} = \times_{i \in \mathcal{I}} \mathcal{Y}_i$ being the product space. Furthermore, let $\mathcal{D}$ denote the set of all probability distributions on $\mathcal{Y}$. A graphical model consists of a graph $\mathcal{G} = (\mathcal{I}, \mathcal{E})$ and a set of properties (called *Markov properties*) that together determine a sub-family of probability distributions in $\mathcal{D}$. According to both the type of the allowed graphs $\mathcal{G}$ and the set of properties we distinguish between several graphical models - e.g. Markov networks.

Markov networks (or Markov Random Fields) arise when only undirected graphs over the p vertices are allowed and the family of distributions in $\mathcal{D}$ consists of probability distributions on $\mathcal{Y}$ obeying the *local Markov property*. The latter states that: conditional on its adjacent variables, any variable is independent of all the remaining variables. If furthermore, the distribution admits a strictly positive density with respect to some product measure μ on $\mathcal{Y}$ then, the local Markov property is equivalent to the *pairwise Markov property* [108]: any two non-adjacent variables are conditionally independent given all other variables. In the latter case the pairwise Markov property is in turn equivalent to the *global Markov property*. An undirected graph satisfiying the local Markov property is also referred to as a Conditional Independence Graph (CIG). Conditional independence graphs are of prime interest in this thesis.

Other type of graphical models are Bayesian networks which are based on directed acyclic graphs (DAG) [78, 79] and independence chain graphs based on chain graphs [4, 78]. Chain graphs contain both directed and undirected edges.

Gaussian Graphical Models.

A Gaussian Graphical Model (GGM) assumes data are drawn from a multivariate normal distribution:

$$Y \sim \mathrm{N}(0, \Omega_p^{-1}) \tag{1}$$

where Y is a p-dimensional random vector comprising the p random variables $Y_1, \ldots, Y_p$ corresponding to the nodes of $\mathcal{I}$ and Ω_p^{-1} is a non-singular $(p \times p)$-dimensional covariance matrix. The matrix Ω_p is referred to as the *precision matrix*. For a GGM the edge set $\mathcal{E}$ of the underlying conditional independence graph corresponds to the nonzero elements of Ω_p [78]. Hence, reconstructing the conditional independence graph is equivalent to determining the non-zero elements of this matrix.

Both frequentist and Bayesian approaches are used in the literature to estimate the conditional independence graph. Standard frequentist approaches rely on penalized likelihood estimation. The augmented penalty to the likelihood aims at resolving the high-dimensionality issue of the data. The commonly employed lasso and ridge penalties amount to adding the ℓ_1- and ℓ_2-norm, respectively, of the precision matrix to the likelihood [43, 139]. Both penalties shrink the elements of the precision matrix towards zero. The lasso penalty may shrink these to exactly zero, thus performing variable selection. The ridge penalization requires a post-hoc step to sparsify its precision matrix estimate. The usual Bayesian approach is to put a prior π on the structure of the conditional independence graph $\mathcal{G}$ and given $\mathcal{G}$ a prior $p(\Omega_p|\mathcal{G})$ on the precision matrix [33, 50, 68]. The joint density is given by

$$p(\mathcal{G}, \Omega_p, \mathbf{Y}) = \pi(\mathcal{G})p(\Omega_p|\mathcal{G})p(\mathbf{Y}|\mathcal{G}, \Omega_p) \qquad \text{where} \qquad \mathbf{Y} = (Y^1, \cdots, Y^n)$$

and a joint structural and quantitative learning is performed by computing the posterior $p(\mathcal{G}, \Omega_p|\mathbf{Y}) \propto p(\mathcal{G}, \Omega_p, \mathbf{Y})$. Except in very small problems, the space of graphs to consider is typically restricted to - e.g. decomposable graphs, forests, or trees.

In a multivariate Gaussian distribution, all conditional distributions are Gaussian linear regressions. Hence, to Gaussian model determination (with non-decomposable graphs), [33] propose estimating these conditional regressions from data using (Bayesian) sparse regression techniques (often called Simultaneous Equations Models).

Simultaneous Equations Models.

Simultaneous Equations Models (SEMs) are a framework for modeling and coding path diagrams. We will use the *very basic* SEMs consisting in modeling the full conditional distribution of each univariate random variable Y_i, $i \in \mathcal{I}$ and thus resulting in a system of regressions

$$(2) \qquad Y_i = \sum_{t \neq i} \beta_{i,t} Y_t + \epsilon_i, \quad \epsilon_i \perp\!\!\!\perp \{Y_t; t \neq i\}, \quad i \in \mathcal{I}.$$

SEMs are flexible tools and computationally very attractive. They account for experimental or biological covariates in the regressions and are appropriate for many types of data distribution [3, 25, 115]. They allow the integration of multiple data sets and at the same time are scalable to large datasets in their computational complexity. Moreover, there is an equivalence between GGM and SEMs, namely, the regression coefficients $\beta_i = (\beta_{i,t} : t \neq i)$ can be expressed in the precision matrix of Y as [97]

$$\beta_{i,t} = -\frac{(\Omega_p)_{it}}{(\Omega_p)ii},$$

in which case the residuals in (2) when regressing a single coordinate Y_i of a multivariate Gaussian vector linearly on the other coordinates Y_t, for $t \neq i$, are Gaussian. That means, the (non)zero entries in the ith row vector of the precision matrix Ω_p

correspond to the (non)zero coordinates of β_i. Consequently, the problem of identifying the Gaussian graphical model can be cast as a variable selection problem in the p regression models (2). This approach of recasting the estimation of the (support of the) precision matrix as a collection of regression problems was first suggested by [33] and latter introduced by [97], who employed Lasso regression [43, 130] to estimate the parameters. Other variable selection methods can be employed as well [73].

In this thesis, we introduce a Bayesian approach of the SEMs. In Chapter 1 we develop a Bayesian formulation of the SEMs with Gaussian, ridge-type priors on the regression coefficients. In Chapter 2, we extent the latter model to incorporate prior knowledge on the conditional independence graph. A disadvantage of the Gaussian priors employed in these papers is that they are not able to selectively shrink parameters, but shrink them jointly towards zero (although prior information used in Chapter 2 alleviates this by making this dependent on prior group). Chapter 3 proposes a general framework for analysing large-scale data sets with complex dependence structures using a collection of linear regression models corresponding to p characteristics (e.g. genes). The *horseshoe* prior [19, 20] has been introduced in order to better model the sparsity of the explanatory variables, thus being able to selectively shrink parameters towards zero. Reconstruction of conditional independence graphs by incorporating prior information is a special case of the proposed framework in Chapter 3.

0.2 Prior information

High-dimensional modeling is important in many scientific areas but is also a challenging task. In genomics, the identification of gene regulatory networks is crucial for understanding gene function, and hence important for both treatment and prediction of diseases. This challenge of analysing data consisting of few replicate measurements against large number of covariates "$n \ll p$" can be alleviated by incorporating external (or "prior") information in the analysis. In gene regulatory networks reconstruction, prior knowledge on the topology on the to-be-reconstructed network is readily available. For instance, the current beliefs on interactions among genes is condensed in repositories like KEGG and Reactome. The Bayesian framework provides a natural architecture to incorporate and accommodate such prior information. It may be believed that such priors can affect the integrity of the current study results and can even lead to conclusions that are driven not by the data but by a prior resulting from some non-relevant previous studies. However, the incorporation in a soft manner, so that it informs the analysis if correct, but can be overruled if completely incompatible with the data, helps overcoming this situation.

Many works have already been devoted to incorporating prior knowledge into network reconstruction. These sudies include [64, 65, 87, 102, 127, 149] for the incorporation of many types of different prior knowledge, including literature-based knowledge in Bayesian network learning and dynamic Bayesian network learning. However, none of these proposed methods explicitly estimate the agreement of the prior knowledge with the data at hand.

Our approach in this thesis is based on prior modelling of the regression parameters of the SEMs in a soft manner using respectively the Gaussian, ridge-type prior in Chapter 2 with a prior on the regularization parameter that depends on external information, and the horseshoe prior in Chapter 3 with a prior on the sparsity index that also depends on external information. Multiple sources of information are incorporated simultaneously. The proposed scheme attaches a latent variable to each source of information independently across sources. These latent variables enter the prior distributions of the coordinates of β_i, which marginally given the latent variable are scale mixtures of the normal distribution. Our soft borrowing of prior information is based on the estimation of these prior hyperparameters by an appealing empirical Bayes procedure (called *global empirical Bayes*).

In Chapter 4, we investigate how gene regulatory networks (GRNs) can be reconstructed from combining observational and time-course gene expression (cell line) data. We present strategies to borrow information respectively in a soft and hard manner from either study type in reconstructing both the CIG-based gene regulatory network and the *human* independence (or time-series) chain graph. The hard borrowing of prior information here means that the prior information is hard-wired in our analysis, because we intend to steer the results for reasons of interpretation or because we have a strong belief in the prior information.

0.3 Variational Bayes approximation

In Bayesian statistics, a prior is assigned to the parameter of interest. The prior belief is subsequently updated by means of current data and inference is based on the posterior distribution. Traditional Bayesian computation methods rely on Markov Chain Monte Carlo (MCMC). However, modern datasets (e.g. gene expression data) are extremely high-dimensional and the use of MCMC is often a computational bottleneck due to high-dimensional integral computations. Approximate Bayesian methods have emerged in recent years as fast alternatives methods to MCMC to overcome these shortcomings. Among the proposed methods *variational Bayes approximations* seem very promising.

Variational approximations are a set of deterministic methods used to make approximate inference for parameters in complex statistical models. The name *variational approximations* originates from the mathematical topic known as *variational calculus*. The latter is concerned with the problem of optimizing a functional over a class of functions. The problem becomes usually feasible when the domain of the functional is restricted to some sub-class of functions. The variational Bayes approximation to a distribution is the closest element q^* in a given target set $\mathcal{Q}$ of distributions, usually with "distance" measured by Kullback-Leibler divergence [141]. The set $\mathcal{Q}$ is chosen as a compromise between computational tractability and accuracy of approximation. If θ denote the parameter of interest in a generic Bayesian model and $\mathbf{Y}$ the observed data, the Kullback-Leibler divergence is defined as

$$KL\Big(q||p(\cdot|\,\mathbf{Y})\Big) = \mathbf{E}_q \log \frac{q(\theta)}{p(\theta|\,\mathbf{Y})} = \log p(\mathbf{Y}) - \mathbf{E}_q \log \frac{p(\mathbf{Y},\theta)}{q(\theta)}, \tag{3}$$

where $\theta \mapsto p(\theta|\,\mathbf{Y})$ is the posterior density, the expectation is taken with respect to θ having the density $q \in \mathcal{Q}$, and $(y,\theta) \mapsto p(y,\theta) = p(y|\,\theta)\,\pi(\theta)$ and $y \mapsto p(y) = \int p(y,\theta)\,d\theta$ are the joint density of $(\mathbf{Y},\theta)$ and the marginal density of $\mathbf{Y}$, respectively, in the model with prior density π on θ. Minimization of (3) is equivalent to the maximization of the expression on the far right hand side of (3) which is usually referred to as "the evidence lower bound", or "elbo". By the non-negativity of the Kullback-Leibler divergence it holds

$$\log p(\mathbf{Y}) \geq \mathbf{E}_q \log \frac{p(\mathbf{Y},\theta)}{q(\theta)} =: \mathrm{elbo}(q;\mathbf{Y}). \tag{4}$$

Early applications involved standard distributions such as Gaussian, Dirichlet, Laplace and extreme value models [5–7, 96, 142]. In the present thesis we use nonparametric approximations, restricted only by the assumption that the various parameters are (block) independent. (This may be referred to as *mean-field* variational Bayes, although this term appears to be used more often for independence of all univariate marginals, whereas we use block independence.) The restriction of $\mathcal{Q}$ to a subclass of product densities gives rise to explicit solutions for each product component in terms of the others, leading to iterative scheme for obtaining the solutions. Precisely, the assumption $q(\theta) = \prod_{i=1}^{M} q_i(\theta_i)$ yields

$$
\begin{aligned}
\text{elbo}(q;\mathbf{Y}) &= \int q(\theta)\log\frac{p(\mathbf{Y},\theta)}{q(\theta)}d\theta \\
&= \int \prod_{i=1}^{M} q_i(\theta_i)\Big[\log p(\mathbf{Y},\theta) - \sum_{i=1}^{M}\log q_i(\theta_i)\Big]d\theta_1\cdots d\theta_M \\
&= \int q_1(\theta_1)\Big[\int \log p(\mathbf{Y},\theta)\prod_{i=2}^{M} q_i(\theta_i)d\theta_2\cdots d\theta_M\Big]d\theta_1 \\
&\quad - \int q_1(\theta_1)\log q_1(\theta_1)d\theta_1 + \text{terms not involving } q_1
\end{aligned}
$$

Define

$$
G_1(\theta_1) = \int \log p(\mathbf{Y},\theta)\prod_{i=2}^{M} q_i(\theta_i)d\theta_2\cdots d\theta_M
$$

Then,

$$
\begin{aligned}
\text{elbo}(q;Y) &= \int q_1(\theta_1)\log\Big(\frac{\exp(G_1(\theta_1))}{q_1(\theta_1)}\Big)d\theta_1 + \text{terms not involving } q_1 \\
&= \int q_1(\theta_1)\log\left(\frac{\exp(G_1(\theta_1))/\int\exp(G_1(\theta_1))d\theta_1}{q_1(\theta_1)}\right)d\theta_1 + \text{terms not involving } q_1 \\
&= -KL\left(q_1 || \frac{\exp(G_1(\theta_1))}{\int\exp(G_1(\theta_1))d\theta_1}\right) + \text{terms not involving } q_1.
\end{aligned}
$$

Hence by the non-negativity of the Kullback-Leibler divergence, the optimal q_1^* satisfies

$$
\begin{aligned}
q_1^*(\theta_1) &= \frac{\exp(G_1(\theta_1))}{\int\exp(G_1(\theta_1))d\theta_1} \\
&\propto \exp\Big[\int \log p(\mathbf{Y},\theta)\prod_{i=2}^{M} q_i(\theta_i)d\theta_2\cdots d\theta_M\Big] \\
&= \exp\Big[E_{q_{-1}}\log p(\mathbf{Y},\theta)\Big]
\end{aligned}
$$

where $E_{q_{-1}}$ indicates the expectation over $(\theta_2,\cdots,\theta_M)$ with respect to $q_2\times\cdots\times q_M$. Unfortunately, this expression depends on $q_2,\cdots,q_M$. However, analog expressions for $q_2^*,\cdots,q_M^*$ can be derived, and it is hoped that repeatedly updating a density q_i^* using the current values of $q_1^*,\cdots,q_{i-1}^*,q_{i+1}^*\cdots,q_M^*$ will in the limit yield the maximizer of (4).

Variational Bayes typically produces accurate approximations to posterior means, but have been observed to underestimate posterior spread [12, 18, 48, 94, 131, 143, 145, 151], even for the marginal distributions. We find that in our setting the approximations agree reasonably well to MCMC approximations of the marginals, although the latter take much longer to compute.

High-dimensional Bayesian regressions

In high-dimensional linear regression, a regularization is required to guarantee the existence and accuracy of estimates. This is done in the Bayesian case by introducing a latent variable in the parameter vector θ_i, and the priors on the regression coefficients β_i are referred to as *regularization priors*. *Scale mixtures* of normal distributions are a well-known class of regularization priors giving rise to different priors for different choices of the mixing densities. In Chapter 1 and 2 we used an inverse-gamma mixing density which results in a ridge-type prior for the regression coefficients, whereas in Chapter 3 we employ a half-Cauchy mixing density. The latter is known as *horseshoe prior* [19, 20]. We fix the hyperparameters to the same values across regressions, thus allowing their estimation by our *global empirical Bayes* procedure. The classical empirical Bayes procedure estimates prior hyperparameters by maximizing the marginal likelihood of the data. Our *global empirical Bayes* procedure maximizes a sum of marginal likelihoods which is enabled by our global-local type prior for modeling multiple related high-dimensional and complex datasets. The procedure has been shown to be very efficient, specially in very high-dimensional settings [135]. The global empirical Bayes enables the borrowing of information across regressions.

0.4 Outline of this thesis

The thesis consists of four chapters organized as follows.

Chapter 1: *Gene network reconstruction using global-local shrinkage priors*

This chapter introduces a new global-local shrinkage ridge-type prior for undirected networks reconstruction based on SEMs with posterior edge selection. The proposed approach is computationally fast and outperforms known competitors such as the *graphical lasso*.

Chapter 2: *An empirical Bayes approach to network recovery using external knowledge*

Chapter 2 extends Chapter 1 to include prior information in reconstructing undirected networks. The incorporation of the prior knowledge is done in a soft manner allowing the data at hand to overrule the prior information if not relevant. Furthermore, the proposed method is able to explicitly estimate the agreement of the prior knowledge with the data at hand which is a novelty in incorporating prior information in network inference.

Chapter 3: *Incorporating prior information and borrowing information in high-dimensional sparse regression using the horseshoe and variational Bayes*

Chapter 3 introduces a framework for simultaneously analysing multiple related high-dimensional and complex datasets. Such analyses include gene regulatory network reconstruction, genetic association studies (e.g. eQTL mapping) and data integration in genomics, to name but a few. To enable the analysis for small n relative to large p, we introduce the *horseshoe* prior which allows for sparsity; a desired property for the analysis of such data. We illustrate the approach by two applications, namely: to the reconstruction of gene regulatory networks and to eQTL mapping.

Chapter 4: *Borrow network information between observational and time-course studies: explorations*

This chapter explores several approaches to reconstruct gene regulatory networks from combining observational (*in vivo*) and time-course cell line (*in vitro*) gene expression data. The dynamics of the human cell are assumed to obey a first-order vector autoregression VAR(1) model and it is investigated how the underlying model parameters can be efficiently learned using the two types of datasets. We saw in an application to real data that reconstruction of the conditional independence graph by borrowing information from the cell line data improved significantly. Moreover, our newly proposed strategies to learn the VAR(1) model parameters are able to indicate preserved transcriptional dynamics between the *in vitro* and *in vivo* environments.

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 $Y = (Y_1, \ldots, Y_p)^T$ 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 $\mathcal{J} = \{1, \ldots, p\}$. 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

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

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 $j \in \mathcal{J} = \{1, \dots, p\}$. The

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

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

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 $\mathbf{\Sigma}_j^{*(t)}$, $\boldsymbol{\beta}_j^{*(t)}$, $d_j^{*(t)}$, $b_j^{*(t)}$ 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

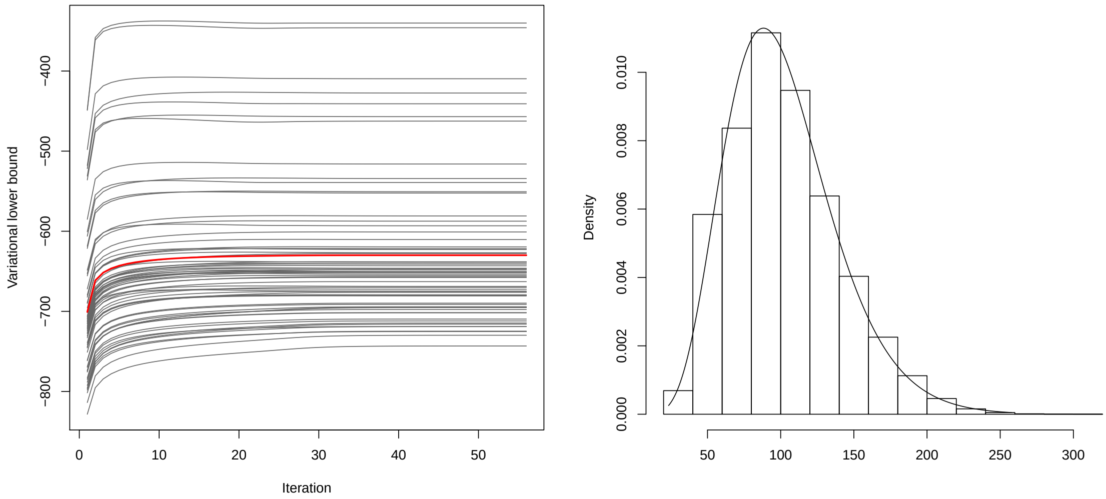
In this section we describe a separate procedure for edge selection. This consists of first ranking the edges based on summary statistics from the (marginal) posterior distributions under the model (1.2) obtained in the preceding sections, and next performing forward selection along this ordering. For 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 $\gamma = (1-\alpha)p_0/(\alpha(1-p_0))$,

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

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 $\mathrm{BF}(\bar{\kappa}_r)$

4. Only if $\mathrm{BF}(\bar{\kappa}_r) > \gamma$ 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 $\mathrm{BF}(\bar{\kappa}_r)$ 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.

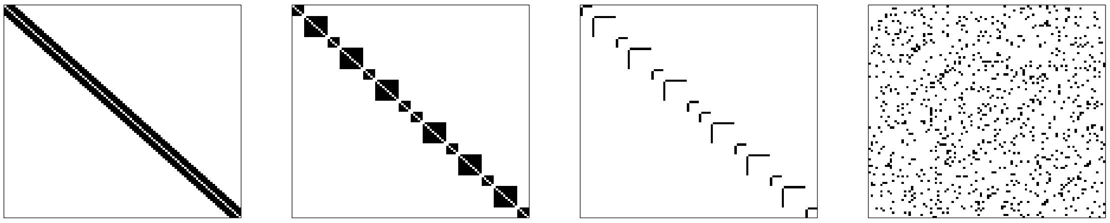

(a) Band (b) Cluster (c) Hub (d) Random

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 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=$2 \times$ (precision $\times$ 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. 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 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

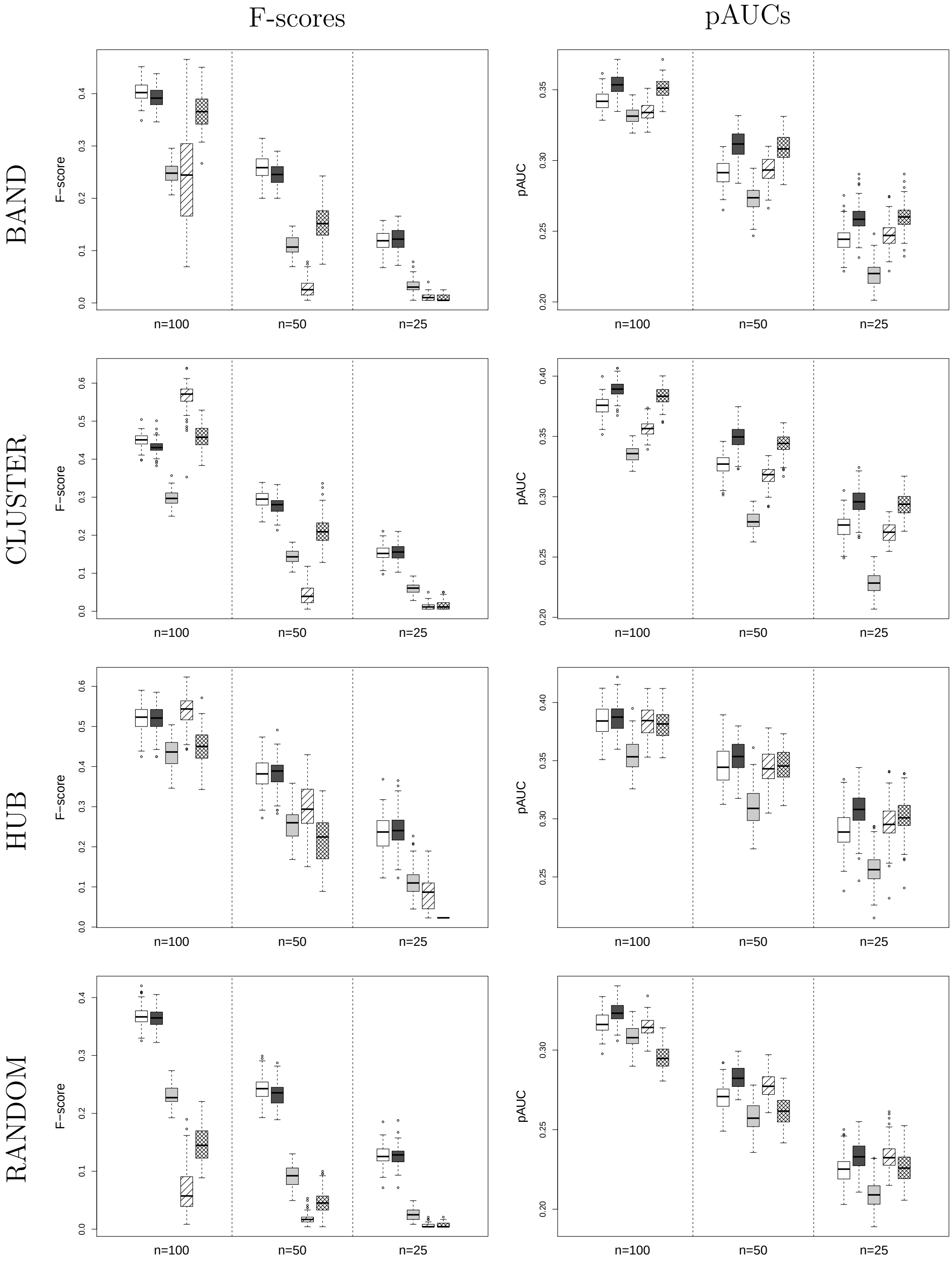


Figure 1.3: *Boxplots of F-scores (left column) and pAUCs (right column) over 100 repetitions as a function of the method, n and the true graph structure. The five methods under comparison are from left to right: NoShrink (white), ShrinkNet (dark grey), SEM$_L$ (light grey), GL$_\lambda$ (diagonal pattern)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

.

with 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 ($n_\text{brca} = 526$), which comprise $p^\text{p53} = 67$ genes, and the apoptosis pathway in the Ovarian data set ($n_\text{ov} = 537$) that comprises $p^\text{apopt} = 79$ 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 $n^\text{p53}_\text{small} \in \{134, 67, 34\}$ and $n^\text{apopt}_\text{small} \in \{158, 79, 40\}$ to achieve low-, moderate- and high-dimensional situations, and a large data set where $n^\text{p53}_\text{large} \in \{392, 459, 492\}$ and $n^\text{apopt}_\text{large} \in \{379, 458, 497\}$ (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 $\mathcal{S}_\text{b}$ 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 $\mathcal{S}_\text{b}$ too small, we only compare two methods at a time.

To assess performance in recovering $\mathcal{S}_\text{b}$ 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, 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.

	$n^\text{apopt}_\text{small} = 158$	$n^\text{apopt}_\text{large} = 379$	$n^\text{apopt}_\text{small} = 79$	$n^\text{apopt}_\text{large} = 458$	$n^\text{apopt}_\text{small} = 40$	$n^\text{apopt}_\text{large} = 497$
ShrinkNet	62.5 (5.7)	138.6 (5.9)	31.4 (5.1)	166.9 (6)	18.2 (4.8)	179.6 (5.6)
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.*

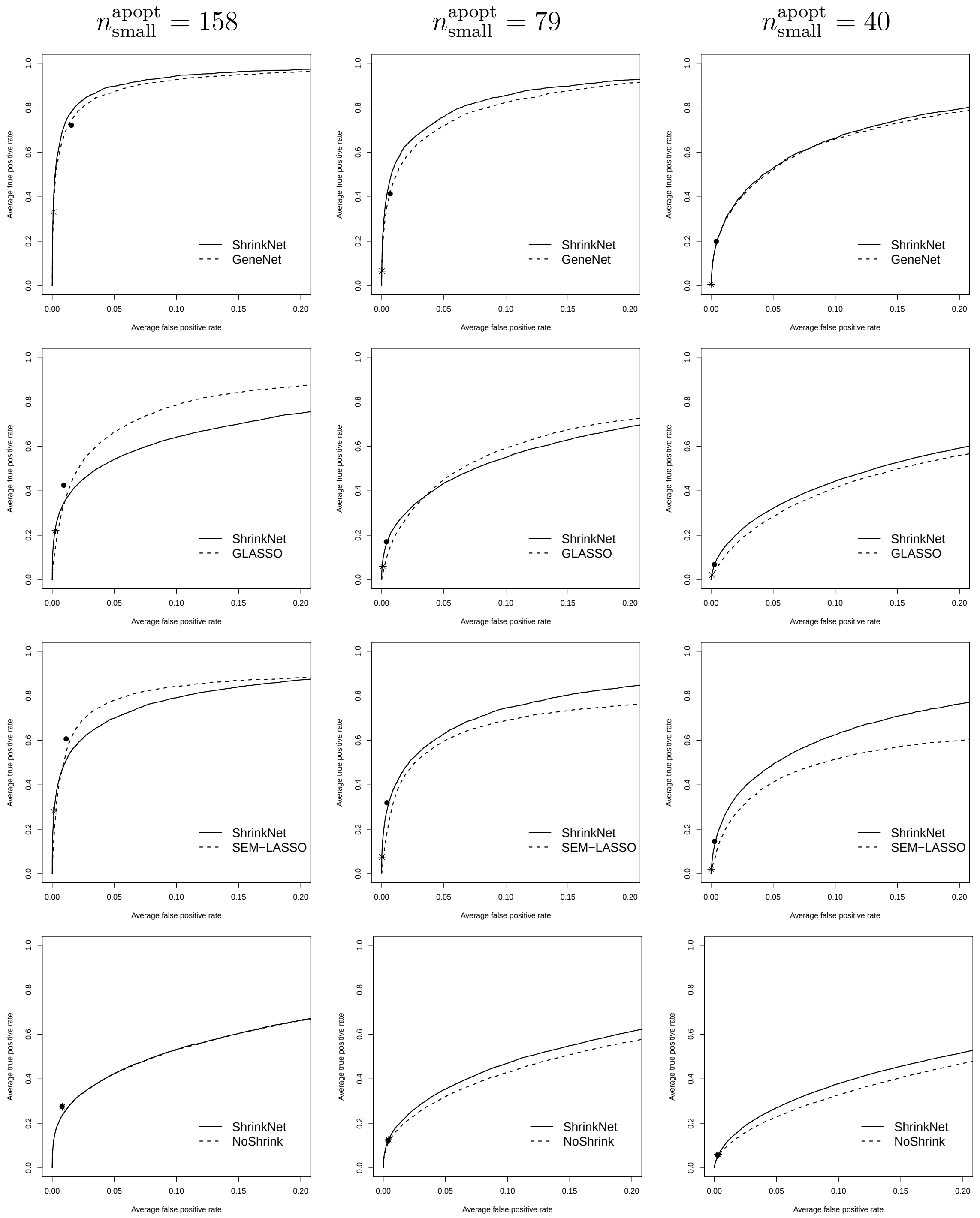


Figure 1.4: *Average partial ROC-curves corresponding to all pairwise comparisons of GeneNet, GL_λ, SEM_L and NoShrink with ShrinkNet when the apoptosis data are randomly split into a small data set of size $n^{p53}_{small} \in \{134, 67, 34\}$ and a large validation one of size $n^{p53}_{large} \in \{392, 459, 492\}$. 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.*

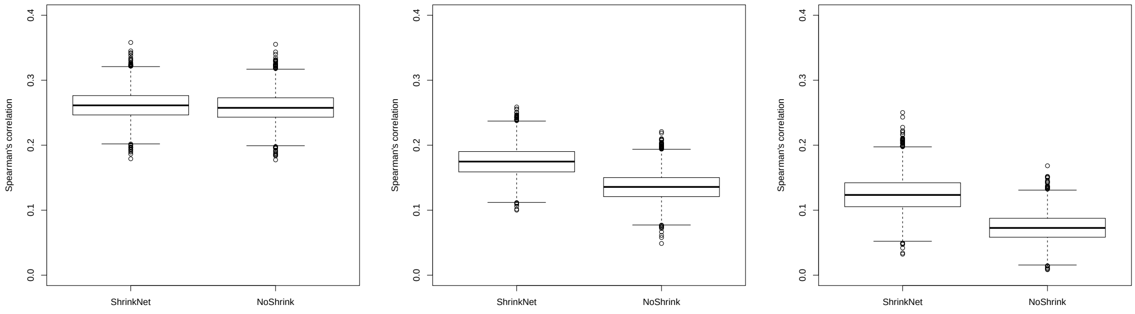
The number of selected edges differs a lot between GeneNet, $\mathrm{SEM}_{\mathrm{L}}$, GL_{λ} and ShrinkNet (Table 1.2). GeneNet is the most conservative approach whereas ShrinkNet selects more edges than others in the small data sets. However, when the sample size is large GL_{λ} selects more than 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 $\mathcal{S}_{\mathrm{b}}$. 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 $n_{\mathrm{small}}^{\mathrm{apopt}} = 40$, GeneNet, $\mathrm{SEM}_{\mathrm{L}}$ 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 $n_{\mathrm{small}}^{\mathrm{apopt}} = 79$ and $n_{\mathrm{small}}^{\mathrm{apopt}} = 40$. In case $n_{\mathrm{small}}^{\mathrm{apopt}} = 158$, $\mathrm{SEM}_{\mathrm{L}}$ 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 $n_{\mathrm{small}}^{\mathrm{apopt}}$ for ShrinkNet and NoShrink. The correlations are clearly higher for ShrinkNet than for NoShrink when $n_{\mathrm{small}}^{\mathrm{apopt}} \in \{79, 40\}$, which indicates that the global shrinkage improves the stability and, hence, reproducibility of edge ranking when the sample size $n_{\mathrm{small}}^{\mathrm{apopt}}$ 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 $n_{\mathrm{small}}^{\mathrm{apopt}}$. We define the set of stable edges by $S_{\mathrm{stable}} = \{(i, j) : \hat{\pi}_{ij} \geq \pi_{\mathrm{thr}}\}$ where $0.5 < \pi_{\mathrm{thr}} \leq 1$. To determine an appropriate cut-off π_{thr}, which is comparable between methods, we use



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

Figure 1.5: *Correlations of edge ranking as provided by ShinkNet and NoShrink across the 100 generated small data sets of size* n^{apopt}_{small}. *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^{apopt}_{small} *for each of the two methods. Note that one does not expect high rank correlation when considering all edges.*

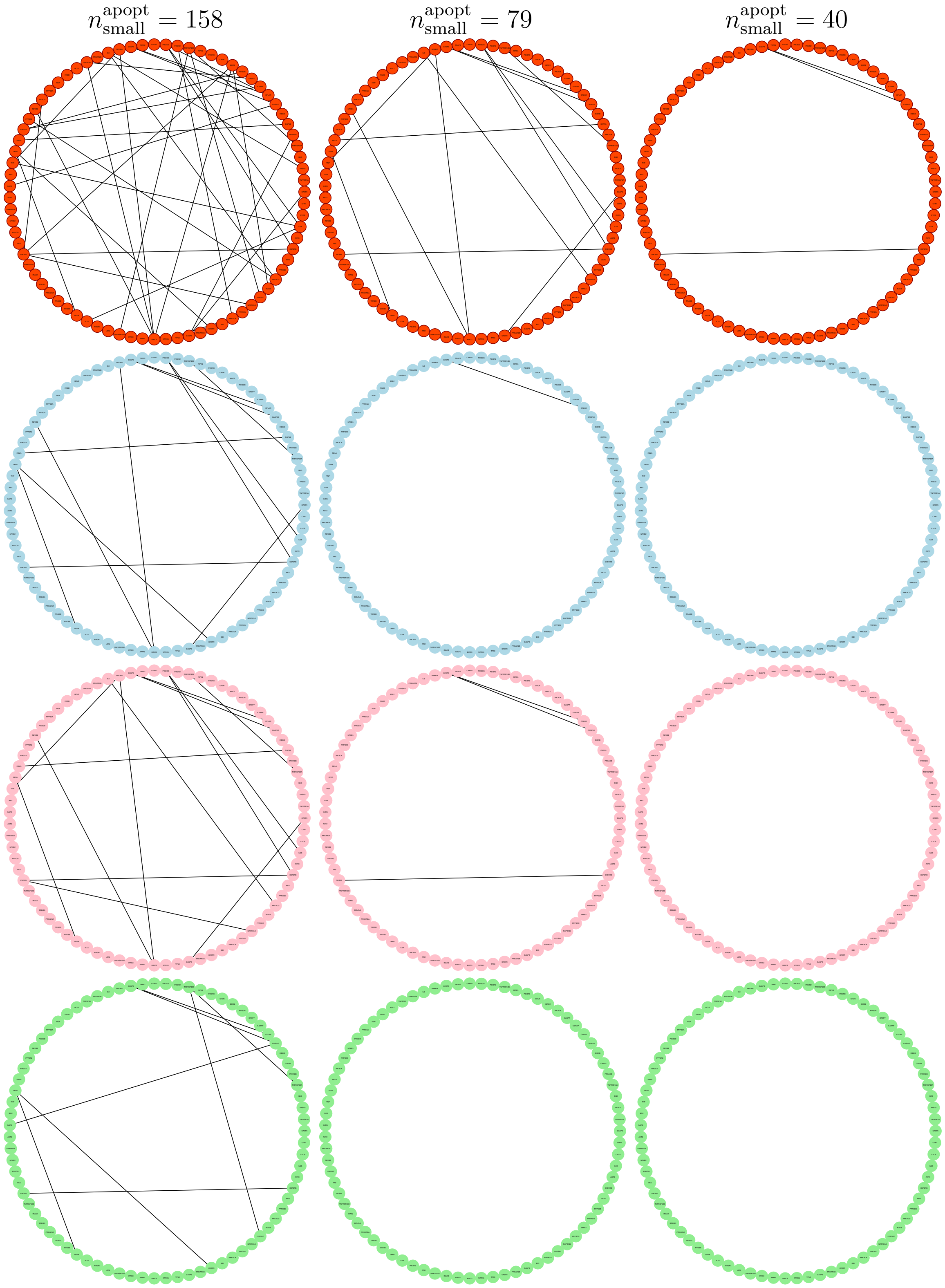
the stability criterion proposed by [98]. This is based on the following upper bound on the expected number $\mathbb{E}(V)$ of falsely selected edges:

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

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 $n^{\text{apopt}}_{\text{small}} = 158$ for the apoptosis data we obtain that $\pi^{\text{ShrinkNet}}_{\text{thr}} = 0.623$, $\pi^{\text{SEM}_\text{L}}_{\text{thr}} = 0.508$, $\pi^{\text{GL}_\lambda}_{\text{thr}} = 0.522$ and $\pi^{\text{GeneNet}}_{\text{thr}} = 0.503$, 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 $n^{\text{apopt}}_{\text{small}}$ and $n^{\text{p53}}_{\text{small}}$, 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 $n^{\text{apopt}}_{\text{small}} = 79$ ShrinkNet identifies a nearly identical network to GL_λ and SEM_L when $n^{\text{apopt}}_{\text{small}} = 158$. For the p53 pathway (see SM Figure 10), GL_λ detects more stable edges than ShrinkNet when $n^{\text{p53}}_{\text{small}} = 134$, as many as when $n^{\text{p53}}_{\text{small}} = 67$, and less when $n^{\text{p53}}_{\text{small}} = 40$. 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

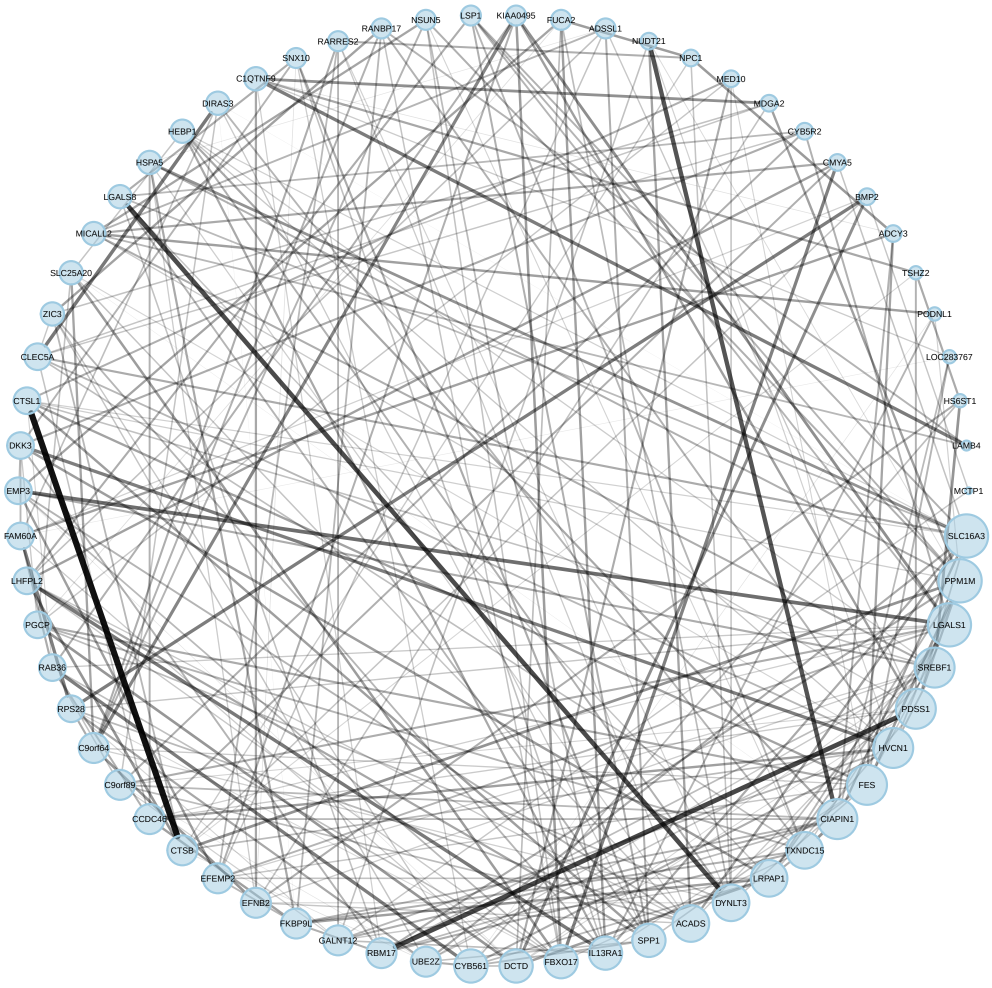


Figure 1.7: *Reconstructed network for the 66 genes associated with patient survival in GBM. Node size is proportional to the node degree and edge width/opacity is proportional to* $\bar{\kappa}_{j,k}$.

[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.

Chapter 2

An empirical Bayes approach to network recovery using external knowledge

Reconstruction of a high-dimensional network may benefit substantially from the inclusion of prior knowledge on the network topology. In the case of gene interaction networks such knowledge may come for instance from pathway repositories like KEGG, or be inferred from data of a pilot study. The Bayesian framework provides a natural means of including such prior knowledge. Based on a Bayesian Simultaneous Equation Model, we develop an appealing Empirical Bayes (EB) procedure which automatically assesses the agreement of the used prior knowledge with the data at hand. We use variational Bayes method for posterior densities approximation and compare its accuracy with that of Gibbs sampling strategy. Our method is computationally fast, and can outperform known competitors. In a simulation study we show that accurate prior data can greatly improve the reconstruction of the network, but need not harm the reconstruction if wrong. We demonstrate the benefits of the method in an analysis of gene expression data from GEO. In particular, the edges of the recovered network have superior reproducibility (compared to that of competitors) over resampled versions of the data.

This chapter was published as: G.B. Kpogbezan, A. W. van der Vaart, W. N. van Wieringen, Gwenaël G. R. Leday and M.A. van de Wiel (2017). An empirical Bayes approach to network recovery using external knowledge. *Biometrical Journal*, 59(5), 932–947. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

2.1 Introduction

Many areas of the quantitative sciences have witnessed a data deluge in recent years. This is due to an increased capacity of measuring and storing data in combination with a reduction in costs of acquiring this data. For instance, in the medical field high-throughput platforms yield measurements of many molecular aspects (e.g. gene expression) of the cell. As many as $20,000$ genes of a single patient can be characterized simultaneously. However, although the costs of such techniques have gone down over the years, the number of patients n in a typical clinical study is still small compared to the number of variables p measured. Reliable analysis of data of such a "$n \ll p$" study is difficult. In this paper we try to solve the problem of few replicate measurements by incorporating external (or "prior") data in the analysis. To allow interpretation, we restrict ourselves to predefined subsets of genes (e.g. pathways) for which p is usually moderately larger than n.

High-dimensional modelling based on a small data set is particularly challenging in studies of relationships between variables. The number of potential pairwise relationships between even a modest number of genes is $p(p-1)/2$. However, some of these relationships may be known from the vast body of medical literature available. For instance, the current beliefs on interactions among genes is condensed in repositories like KEGG and Reactome. Although such information may not be reliable, or be only partially relevant for the case at hand, its flexible inclusion may help the analysis of high-dimensional data. Methodology that exploits such prior information may accelerate our understanding of complex systems like the cell.

The cohesion of variables constituting a complex system is often represented by a network, also referred to as a *graph*. A graph $\mathcal{G}$ consists of a pair $(\mathcal{I}, \mathcal{E})$ where $\mathcal{I} = \{1, ..., p\}$ is a set of indices representing nodes (the variables of the system) and $\mathcal{E}$ is the set of edges (relations between the variables) in $\mathcal{I} \times \mathcal{I}$. An edge can be characterized in many ways, we concentrate on it representing conditional independence between the node pair it connects. More formally, a pair $(i_1, i_2) \in \mathcal{E}$ if and only if random variables represented by nodes i_1 and i_2 are conditionally dependent, given all remaining nodes in $\mathcal{I}$. All pairs of nodes of $\mathcal{I}$ not in $\mathcal{E}$ are conditionally independent given the remaining nodes. Graphs endowed with this operationalization of the edges are referred to as conditional independence graphs (Whittaker, 1990).

Conditional independence graphs are learned from data by graphical models. Graphical models specify how data are generated obeying the relations among the variables as specified by a conditional independence graph. A Gaussian Graphical

Model (GGM) assumes data are drawn from a multivariate normal distribution:

$$Y^j \sim^{\text{iid}} \mathrm{N}(0, \Omega_p^{-1}), \qquad j \in \{1, ..., n\}. \tag{2.1}$$

Here Y^j is a p-dimensional random vector comprising the p random variables $Y_1^j, \ldots, Y_p^j$ corresponding to the nodes of $\mathcal{I}$ and Ω_p^{-1} is a non-singular $(p \times p)$-dimensional covariance matrix. The matrix Ω_p, as opposed to its inverse, is referred to as the *precision matrix*. For a GGM the edge set $\mathcal{E}$ of the underlying conditional independence graph corresponds to the nonzero elements of Ω_p (Lauritzen, 1996). Hence, to reconstruct the conditional independence graph it suffices to determine the non-zeros elements of this matrix.

Reconstruction of the conditional independence graph may concentrate on the direct estimation of the precision matrix. Here we choose a different estimation strategy. This exploits an equivalence between Gaussian graphical models and Simultaneous Equations Models (SEMs), which we introduce first before pointing out the equivalence. Our choice for SEM is mainly motivated by its flexibility and its performance. It can account for experimental or biological covariates in the regression, and extensions to non-Gaussian data are available (Chen et al., 2015; Allen and Liu, 2013; Yang et al., 2012; Ravikumar et al., 2010). Its Bayesian counterpart is appealing for including prior knowledge, which likely is more complicated in many other frameworks. Its good performance in comparison with alternatives including (sparse) graphical models was demonstrated by Leday et al. (2017). In addition, SEM is also computational efficient (Meinshausen and Bühlmann, 2006). We treat SEMs as a system of regression equations, with each equation modelling the conditional distribution of a node given the other nodes. If we collect all observations on node $i \in \mathcal{I}$ in a vector $Y_i := (Y_i^1, \ldots, Y_i^n)^T$, then we can write:

$$Y_i = X_i \beta_i + \epsilon_i, \quad i \in \mathcal{I}, \tag{2.2}$$

where X_i is the $n \times (p-1)$-matrix with columns the observations of the $p-1$ nodes different from i, i.e. $X_i = [Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_p]$ (where the square brackets mean “combine the vectors in a matrix”). The error vector ϵ_i is defined by the equation, and possesses a multivariate Gaussian distribution $\mathrm{N}(0, \sigma_i^2 \mathbf{I_n})$ under the GGM. (The covariances between the errors of different equations are in general non-zero, but are left unspecified.) The equivalence between the thus formulated SEM and the GGM as specified above stems from the one-to-one relationship between the regression parameters of the SEM and the elements of the GGM’s precision matrix (Lauritzen (1996)):

$\beta_{i,r} = -\omega_{ii}^{-1}\omega_{ir}$. In particular, (non)zero entries in the i-th row vector of the precision matrix Ω_p correspond to the (non)zero coefficients of β_i. The problem of identifying (non)zero entries in Ω_p can therefore be cast as a variable selection problem in the p regression models (2.2). Lasso regression (Tibshirani, 1996) may be used for this purpose (as in Meinshausen and Bühlmann (2006)), but other variable selection methods have also been employed. The problem that every partial correlation appears in two regression equations is usually resolved by post-symmetrization through application of the 'AND'-rule: an edge $(i, j) \in \mathcal{E}$ if and only if $\beta_{i,j} \neq 0$ and $\beta_{j,i} \neq 0$ (Meinshausen and Bühlmann, 2006). Graph structures recovery based on model (2.2) performs well and is widely used in practice.

Previously, we proposed a Bayesian formulation of the SEM (Leday et al., 2017). In this Bayesian SEM (henceforth BSEM) the structural model (2.2) is endowed with the following prior:

$$\begin{aligned}\epsilon_i|\,\sigma_i^2, \tau_i^2 &\sim \mathrm{N}(0_n, \sigma_i^2\mathbf{I}_n),\\ \beta_i|\,\sigma_i^2, \tau_i^2 &\sim \mathrm{N}(0_s, \sigma_i^2\tau_i^{-2}\mathbf{I}_s),\\ \tau_i^2 &\sim \mathrm{Gamma}(a_1, b_1),\\ \sigma_i^{-2} &\sim \mathrm{Gamma}(a_2, b_2),\end{aligned} \tag{2.3}$$

where $\mathbf{I}$ is an identity matrix, $s = p-1$, and Gamma(a, b) denotes a gamma distribution with shape parameter a and rate parameter b, and τ_i^2 and σ_i^{-2} are independent. The normal-gamma-gamma (NGG) prior of model (2.3) regularizes the parameter estimates (e.g. estimated as the posterior mean) in two distinct ways. First, due to the normal prior on the regression coefficients $\beta_{i,r}$ (corresponding to a ridge penalty), the estimates of these parameters are shrunken *locally* (i.e. within each equation) to zero. Second, the estimates are simultaneously shrunken *globally* (i.e. across equations), due to the fact that the hyperparameters $\alpha = \{a_1, b_1, a_2, b_2\}$ do not depend on the index i. There seems to be no reason to connect the error variances (which reflect the noise levels of the genes) across the equations, and hence we use a vague prior (e.g. $a_2 = b_2 = 0.001$). In contrast, estimating the parameters a_1, b_1 in EB fashion is advantageous, as it further "borrows information" across the regression equations. The resulting global shrinkage improves inference in particular for large networks (see also Section 2.5). Note that assuming a Gaussian distribution for the regression coefficients is also done in ridge regression and random effects models. The BSEM model can be fit computationally efficiently by a variational method, and generally outperforms the aforementioned lasso regression approach to the estimation of model (2.2). Furthermore, variables can be accurately selected based on the marginal posterior distributions of the regression coefficients (Leday et al., 2017).

The problem of network reconstruction is challenging due to the vast space of possible graphs for even a moderate number of variables. This endeavour is further complicated by the inherent noise in the measurements used for the reconstruction. Fortunately, network reconstruction need not start from scratch, as often similar networks have been studied previously. Prior information on the network may be available in the literature, repositories, or simply as pilot data. It is natural to take such information along in network reconstruction. Many works have already been devoted to incorporating prior knowledge into network reconstruction. Among these studies, Imoto et al. (2003) use energy functions to incorporate prior knowledge sources into Bayesian gene regulatory network models and propose the incorporation of many types of different prior knowledge, including literature-based knowledge. The approach of Imoto et al. has been extended by Werhli and Husmeier which proposed a framework to incorporate multiple sources of prior knowledge into dynamic Bayesian network using MCMC sampling (Werhli and Husmeier, 2007). In the same line, Steele et al. proposed an advanced text-mining technique to incorporate literature-based prior knowledge into Bayesian network learning of gene networks. Similarly, Li et al. developed an approach that combines literature mining and microarray analysis in constructing biological networks (Li et al., 2006). Murkherjee and Speed (2008) proposed a method to incorporate network features including edges, classes of edges, degree distributions, and sparsity using MCMC sampling in Bayesian network learning. Still in Bayesian network learning, Isci et al. (2013) proposed also a framework to incorporate multiple sources of external knowledge where the incorporation of external knowledge uses Bayesian network infrastructure itself. However, none of these proposed methods explicitly estimate the agreement of the prior knowledge with the data at hand.

In this paper we develop a method for incorporating external data or prior information into the reconstruction of a conditional independence network. To this aim we extend in Section 2.2 the Bayesian SEM framework (2.2)-(2.3). The extension incorporates prior knowledge in a flexible manner. Next in Section 2.3 we develop a variational Bayes approach to approximate the posterior distributions of the regression parameters for given hyperparameters, and show this to be comparable in accuracy to Gibbs sampling, although computationally much more efficient. In Section 2.4 this is complemented by a derivation of an empirical Bayes approach to estimate the hyperparameters. Using simulations we show in Section 2.5 that the method performs better, in terms of ROC curves, than BSEM when the prior knowledge agrees with the data, and is as accurate when it is not. In Section 2.6 we show the full potential of our approach on real data. We conclude the paper with a discussion.

2.2 Model

The BSEM approach, comprising model (2.2) with priors (2.3), is modified to incorporate external information on the to-be-reconstructed network. The resulting model is referred to as BSEMed (BSEM with *e*xternal *d*ata).

Prior knowledge on the network is assumed to be available as a “prior network”, which specifies which edges (conditional independencies) are present and absent. This is coded in an adjacency matrix P, which contains only zeros and ones corresponding to the absence and presence of an edge in the prior network. That is, $\mathrm{P}_{i,r} = 1$ if node i is connected with node r and $\mathrm{P}_{i,r} = 0$ otherwise. Note that the adjacency matrix P is symmetric (for the purpose of undirected network reconstruction).

The BSEMed approach keeps equation (2.2), but replaces the priors (2.3) of BSEM by:

$$\begin{aligned}
\epsilon_i |\, \sigma_i^2, \tau_{i,0}^2, \tau_{i,1}^2 &\sim \mathrm{N}(0_n, \sigma_i^2 \mathbf{I}_n), \\
\beta_i |\, \sigma_i^2, \tau_{i,0}^2, \tau_{i,1}^2 &\sim \mathrm{N}(0_s, \sigma_i^2 \mathbf{D}_{\tau_i^{-2}}), \\
\mathbf{D}_{\tau_i^{-2}} &= \mathrm{diag}(\tau_{i,1}^{-2}, ..., \tau_{i,s}^{-2}), \\
\tau_{i,r}^2 &= \begin{cases} \tau_{i,0}^2 \sim \mathrm{Gamma}(a_0, b_0), & \text{if} \quad \mathrm{P}_{i,r} = 0, \\ \tau_{i,1}^2 \sim \mathrm{Gamma}(a_1, b_1), & \text{if} \quad \mathrm{P}_{i,r} = 1, \end{cases} \\
\sigma_i^{-2} &\sim \mathrm{Gamma}(a_2, b_2).
\end{aligned} \tag{2.4}$$

where $\beta_i = \beta_{i,1}, ..., \beta_{i,i-1}, \beta_{i,i+1}, ..., \beta_{i,p}$.
The normal-gamma-gamma-gamma (NGGG) prior (2.4) retains the ridge-type regularization of the regression parameters $\beta_{i,r}$ of (2.3), through Gaussian priors on these coefficients. The crucial difference between the two priors reveals itself in the variances of the latter priors. For each regression equation i there are two possible variances:

$$\beta_{i,r} \sim \begin{cases} \mathrm{N}(0, \sigma_i^2 \tau_{i,0}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 0, \\ \mathrm{N}(0, \sigma_i^2 \tau_{i,1}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 1. \end{cases}$$

Hence, the regression coefficients corresponding to edges that are present according to the prior information share the same variance, and similarly for the other set of regression coefficients. Both variances can be both small and large, as they are themselves modelled through Gamma priors, where small values lead to small regression coefficients. If the prior information on the network were correct, then naturally a small value of $\tau_{i,0}^{-2}$ would be desirable, smaller than the value of $\tau_{i,1}^{-2}$. However, the construction is flexible in that the two values, and even their priors, are not fixed a-priori. In (2.4) the two parameters $\tau_{i,0}^{-2}$ and $\tau_{i,1}^{-2}$ are assumed to have gamma priors, with different hyperparameters (a_0, b_0) and (a_1, b_1). For further flexibility these hyperparameters will be estimated from the data with an empirical Bayes method. Then, if the absence of an edge in the prior network is supported by the current data, the corresponding regression coefficient $\beta_{i,r}$ may stem from a prior with a small variance, and will tend to be small; a similar, but opposite, situation will occur for edges that are present in the prior network. Indeed in Section 2.5 we shall see that the EB approach will tend to find similar values of $\tau_{i,0}^2$ and $\tau_{i,1}^2$ when the prior knowledge is non-informative, and rather different values otherwise.

The fact that model (2.4) contains the model (2.3) as a submodel, provides robustness against the misspecification of the prior information. Although the number of latent variables in (2.4) is considerably higher (namely $p-1$ additional variances, one for each regression equation), the actual number of extra parameters is only two (the pair (a_1, b_1)). This suggests that if the prior information doesn't agree with the data at hand, then the cost in terms of precision of the estimators is minor. It is amply compensated by the gains if the prior information is correct. We corroborate this in our simulation study in Section 2.5. In this connection it is also of interest to note the flexible roles of $\tau_{i,0}^2$ and $\tau_{i,1}^2$, $\tau_{i,0}^2$ (resp. $\tau_{i,1}^2$) is freely estimated from the data using the absent (resp. present) prior connections. We allow $\tau_{i,0}^2 < \tau_{i,1}^2$ which accommodates (rare) situations in which a prior is complementary to the data.

2.3 Variational Bayes method and Gibbs sampling

In this section we develop a variational Bayes approach to approximate the (marginal) posterior distributions of the parameters $\beta_{i,r}, \tau_{i,0}^2, \tau_{i,1}^2, \sigma_i^2$ in model (2.4). The algorithm is similar, but still significantly different, from the algorithm developed in Leday et al. (2017) for the model (2.3). In the following we can see that, due to (2.4), the variational parameters have a form which renders the implementation of (2.4) much more challenging. We also verify that these approximations are accurate by comparing them to the results obtained using a Gibbs sampling strategy, which is much slower. Computational efficiency is an important characteristic, especially for fitting large networks.

In this section we work on a single regression equation, i.e. for a fixed index i, and given hyperparameters a_k, b_k, for $k = 0, 1, 2$. In the next section we combine the regression equations to estimate the hyperparameters.

2.3.1 Variational Bayes inference.

In general a "variational approximation" to a distribution is simply the closest element in a given target set $\mathcal{Q}$ of distributions, usually with "distance" measured by Kullback-Leibler divergence. The set $\mathcal{Q}$ is chosen both for its computational tractability and accuracy of approximation. Distributions Q with stochastically independent marginals (i.e. product laws) are popular, and then the "accuracy" of approximation is naturally restricted to the marginal distributions.

In our situation we wish to approximate the posterior distribution of the parameter $\theta := (\beta_i, \tau_{i,0}^2, \tau_{i,1}^2, \sigma_i^2)$ given the prior (2.4) and the observation Y_i given in (2.2), for a fixed i. Here in (2.2) we take X_i (which depends on Y_j for $j \neq i$) as given, as in a fixed-effects linear regression model. For $p(\cdot|\, Y_i)$ the posterior density in this model, the variational Bayes approximation is given as

$$q^* = \operatorname*{argmin}_{q \in \mathcal{Q}} \mathbf{E}_q \log \frac{q(\theta)}{p(\theta|\, Y_i)},$$

where the expectation is taken with respect to the density $q \in \mathcal{Q}$. For $p(Y_i, \theta)$ the joint density of (Y_i, θ), this is equivalent to finding the maximizer of

$$\mathbf{E}_q \log \frac{p(Y_i, \theta)}{q(\theta)}. \tag{2.5}$$

By the nonnegativity of the Kullback-Leibler divergence, the latter expression is a lower bound on the marginal density $p(Y_i) = \int p(Y_i, \theta)\, d\theta$ of the observation, and it is usually referred to as "the lower bound". Solving the variational problem is equivalent to maximizing this lower bound (over $\mathcal{Q}$).

We choose the collection $\mathcal{Q}$ equal to the set of distributions of θ for which the components β_i, $\tau_{i,0}^2$, $\tau_{i,1}^2$ and σ_i^2 are stochastically independent, i.e. $q(\theta) = \prod_{l=1}^4 q_l(\theta_l)$, where the marginal densities q_l are arbitrary. Given such a factorization of q it can be shown in general (see e.g. Ormerod and Wand (2010)), that the optimal marginal densities q_l^* satisfy:

$$q_l^*(\theta_l) \propto \exp(\mathbf{E}_{q_{-l}} \log p(Y_i, \theta)), \qquad \text{where } \mathbf{E}_{q_{-l}} = \mathbf{E}_{q_1} \ldots \mathbf{E}_{q_{l-1}} \mathbf{E}_{q_{l+1}} \ldots \mathbf{E}_{q_4}.$$

It can be shown (see the Supplementary Material) that in model (2.4) for regression equation i, with $\theta = (\beta_i, \tau_{i,0}^2, \tau_{i,1}^2, \sigma_i^{-2})$, this identity can be written in the “conjugate” closed-form

$$\begin{aligned} \beta_i |\, Y_i &\sim \mathrm{N}\Big(\beta_i^*, \Sigma_i^*\Big), \\ \tau_{i,0}^2 |\, Y_i &\sim \mathrm{Gamma}\Big(a_{i,0}^*, b_{i,0}^*\Big), \\ \tau_{i,1}^2 |\, Y_i &\sim \mathrm{Gamma}\Big(a_{i,1}^*, b_{i,1}^*\Big), \\ \sigma_i^{-2} |\, Y_i &\sim \mathrm{Gamma}\Big(a_{i,2}^*, b_{i,2}^*\Big), \end{aligned} \tag{2.6}$$

where

$$\begin{aligned} \Sigma_i^* &= \Big[\mathbf{E}_{q_4^*}(\sigma_i^{-2})\Big(X_i^T X_i + \mathbf{D}_{\mathbf{E}_{q_2^* \cdot q_3^*}(\tau_i^2)}\Big)\Big]^{-1}, \\ \beta_i^* &= \Big[X_i^T X_i + \mathbf{D}_{\mathbf{E}_{q_2^* \cdot q_3^*}(\tau_i^2)}\Big]^{-1} X_i^T Y_i, \end{aligned}$$

$$\begin{aligned} a_{i,0}^* &= a_0 + \tfrac{1}{2} s^0, & b_{i,0}^* &= b_0 + \tfrac{1}{2}\mathbf{E}_{q_4^*}(\sigma_i^{-2})\mathbf{E}_{q_1^*}(\beta_i^{0^T} \beta_i^0), \\ a_{i,1}^* &= a_1 + \tfrac{1}{2} s^1, & b_{i,1}^* &= b_1 + \tfrac{1}{2}\mathbf{E}_{q_4^*}(\sigma_i^{-2})\mathbf{E}_{q_1^*}(\beta_i^{1^T} \beta_i^1), \\ a_{i,2}^* &= a_2 + \tfrac{1}{2} n + \tfrac{1}{2} s, & b_{i,2}^* &= b_2 + \tfrac{1}{2}\mathbf{E}_{q_{-4}^*}\Big(\beta_i^T \mathbf{D}_{\tau_i^2} \beta_i\Big) \\ & & &\quad + \tfrac{1}{2}\mathbf{E}_{q_1^*}(Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i), \end{aligned}$$

where s^0 and s^1 are the number of 0’s and 1’s in the i-th row of the adjacency matrix P, not counting the diagonal element; and $\beta_i^0 = \{\beta_{i,r} : r \in \mathcal{I}\backslash i, \mathrm{P}_{i,r} = 0\}$ and $\beta_i^1 = \{\beta_{i,r} : r \in \mathcal{I}\backslash i, \mathrm{P}_{i,r} = 1\}$ are the coordinates of the vector of regression parameters corresponding to these 0’s and 1’s. Furthermore

$$\mathbf{D}_{\mathbf{E}_{q_2^* \cdot q_3^*}(\tau_i^2)} = \mathrm{diag}\Big(\mathbf{E}_{q_2^*}\mathbf{E}_{q_3^*}(\tau_{i,1}^2), ..., \mathbf{E}_{q_2^*}\mathbf{E}_{q_3^*}(\tau_{i,s}^2)\Big).$$

In these identities the optimal densities q_l^* appear both on the left and the right of the equations and hence the identities describe the optimal densities only as a fixed point. In practice the identities are iterated “until convergence” from suitable starting values.

The iterations also depend on the hyperparameters a_k, b_k. In the next section we describe how these parameters can be estimated from the data by incorporating updates of these parameters in the iterations.

2.3.2 Variational Bayes vs Gibbs sampling.

Under the true posterior distribution the coordinates $\beta_i, \tau_{i,0}^2, \tau_{i,1}^2, \sigma_i^2$ are not independent. This raises the question how close the variational approximation is to the true posterior distribution. As the latter is not available in closed form, we investigate this question in this section by comparing the variational approximation to the distribution obtained by running a Gibbs sampling algorithm. As for the network reconstruction we only use the marginal posterior distributions of the regression parameters, we restrict ourselves to these marginal distributions.

The full conditional densities of BSEMed can be seen to take the explicit form:

$$\begin{aligned}
\beta_i |\, Y_i, \tau_{i,0}^2, \tau_{i,1}^2, \sigma_i^{-2} &\sim \mathrm{N}(\beta_i^*, \Sigma_i^*),\\
\tau_{i,0}^2 |\, Y_i, \beta_i, \tau_{i,1}^2, \sigma_i^{-2} &\sim \mathrm{Gamma}(a_{i,0}^*, b_{i,0}^*),\\
\tau_{i,1}^2 |\, Y_i, \beta_i, \tau_{i,0}^2, \sigma_i^{-2} &\sim \mathrm{Gamma}(a_{i,1}^*, b_{i,1}^*),\\
\sigma_i^{-2} |\, Y_i, \beta_i, \tau_{i,0}^2, \tau_{i,1}^2 &\sim \mathrm{Gamma}(a_{i,2}^*, b_{i,2}^*),
\end{aligned}$$

where the parameters Σ_i^*, β_i^*, $a_{i,k}^*$ and $b_{i,k}^*$ satisfy the same system of equations as in the variational algorithm, except that all expectations $\mathbf{E}_{q^*}$ must be replaced by the "current" values taken from the conditioning (see Supplementary Material). Thus Gibbs sampling of the full posterior $(\beta_i, \tau_{i,0}^2, \tau_{i,1}^2, \sigma_i^{-2}) |\, Y_i$ is easy to implement, although slow.

We ran a simulation study with a single regression equation (say $i = 1$) with $n = p = 50$, and compared the variational Bayes estimates of the marginal densities with the corresponding Gibbs sampling-based estimates. Thus we sampled $n = 50$ independent replicates from a $p = 50$-dimensional normal distribution with mean zero and $(p \times p)$-precision matrix Ω, and formed the vector Y_1 and matrix X_1 as indicated in (2.2). The precision matrix was chosen to be a *band matrix* with a lower bandwidth b_l equal to the upper bandwith b_u. It is $b_l = b_u = 4$, thus a total number of 9 band elements including the diagonal. For both the variational approximation and the Gibbs sampler we used prior hyperparameters $a_2 = b_2 = 0.001$ and prior hyperparameters $\hat{a}_0, \hat{b}_0, \hat{a}_1, \hat{b}_1$ fixed to the values set by the *global* empirical Bayes method described in Section 2.4. The Gibbs iterations were run $nIter = 100,000$ times, after which the first $nBurnin = 1000$ iterates were discarded. Histograms based on subsampling every 10th value of the iterations are compared with the variational Bayes approximation to the marginal posterior densities. The correspondence between the two methods is remarkably good (see the Supplementary

Material).

We conclude that the variational Bayes method gives reliable estimates of the posterior marginal distributions. The computing times in seconds are 40 for BSEMed and $2542 \times 50 =$ 35h 18min 20sec for the Gibbs sampling (in R). The variational method clearly outperforms the Gibbs sampling method, which would hardly be feasible even for $n = p = 50$.

2.4 Global empirical Bayes for BSEMed

Model (2.4) possesses three pairs of hyperparameters (a_k, b_k), for $k \in \{0, 1, 2\}$. The pair (a_2, b_2) controls the prior of the error variances σ_i^2; we fix this to numerical values that render a vague prior, e.g. to $(0.001, 0.001)$. In contrast, we let the values of the parameters $\alpha = (a_0, b_0, a_1, b_1)$ be determined by the data. As these hyperparameters are the same in every regression model i, this allows information to be borrowed across the regression equations, leading to *global shrinkage* of the regression parameters.

A natural method to estimate the parameter α is to apply maximum likelihood to the marginal likelihood of the observations in the Bayesian BSEMed model determined by (2.2) and (2.4). Here "marginal" means that all parameters except α are integrated out of the likelihood according to their prior. The approach is similar to the one in van de Wiel et al. (2012). As a first simplification of this procedure we treat the vectors $Y_1, \ldots, Y_p$ as independent, thus leading to a likelihood of product form. As the exact marginal likelihoods of the Y_i are intractable, we make a second simplication and replace these likelihoods by the lower bound (2.5) to the variational Bayes criterion (see Supplementary Material).

Recall that in model (2.4) each regression parameter $\beta_{i,r}$ corresponds to one of two normal priors, that is:

$$\beta_{i,r} \sim \begin{cases} \mathrm{N}(0, \sigma_i^2 \tau_{i,0}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 0, \\ \mathrm{N}(0, \sigma_i^2 \tau_{i,1}^{-2}), & \text{if} \quad \mathrm{P}_{i,r} = 1. \end{cases}$$

It is the regression coefficients corresponding to edges that are not present according to the prior information share the same precision $\tau_{i,0}^2$, and similarly the coefficients corresponding to the edges that are present obtain the precision $\tau_{i,1}^2$. Both precisions are assumed to have gamma priors with different hyperparameters that are adapted by the current data by the means of the global EB procedure described above. Then, if the absence of an edge in the prior network is supported by the current data, the corresponding regression coefficient $\beta_{i,r}$ will have a small variance, and will tend to be small; a similar, but opposite, situation will occur for edges that are present in the prior network. In next Section we shall see that the EB approach will tend to find similar values of $\tau^2_{i,0}$ and $\tau^2_{i,1}$ when the prior knowledge is non-informative, and rather different values otherwise.

We developed a dedicated edge selection algorithm for BSEM model in Leday et al. (2017). It is based on summarizing $\beta_{i,r}$ and $\beta_{r,i}$ by $\bar{\kappa}_{i,r}$,

$$\bar{\kappa}_{i,r} = (\kappa_{i,r} + \kappa_{r,i})/2 \quad \text{with} \quad \kappa_{i,r} = \frac{\left| \mathbf{E}_{q^{i*}} \left[\beta_{i,r} | \mathbf{y}_i \right] \right|}{\sqrt{\mathbf{V}_{q^{i*}} \left[\beta_{i,r} | \mathbf{y}_i \right]}} \tag{2.7}$$

where $\mathbf{E}_{q^{i*}} \left[\beta_{i,r} | \mathbf{y}_i \right]$ and $\mathbf{V}_{q^{i*}} \left[\beta_{i,r} | \mathbf{y}_i \right]$ denote the approximate posterior expectation and variance of $\beta_{i,r}$ obtained in Section 2.3. The $\bar{\kappa}_{i,r}$ values are ranked and corresponding edges are consecutively included according to a local false discovery rate (lfdr) criterion, which explores the relationship between lfdr and Bayes factors. Details are given in the Supplementary material.

2.5 Numerical investigation

To study the effect of including a prior network in the model framework we compare BSEMed with BSEM. Hereto, we generated data $Y^1, \ldots, Y^n$ according to (3.3), for $p = 100$ and $n \in \{50, 200\}$, which reflect a high- and a low-dimensional situation, respectively. We considered precision matrices Ω_p, which imply *band*, *cluster* and *hub* network topologies (Zhao et al., 2012) (See Supplementary Material).

For BSEMed we vary the quality of the prior network information: ‘perfect’ prior information, i.e. the generating model; ‘75%’ true edges; ‘50%’ true edges; ‘0%’ true edges. To generate 75% (or 50%, or 0%) true information, we swapped 25% (or 50%, or 100%) of the true edges with the same number of absent edges, i.e. in the adjacency matrix P that describes the prior network we swapped these percentages of 1s with 0s. It may be noted that in the last case the prior network is completely wrong for the true edges, but not for the absent edges due to over-sampling of the 0’s, which seems realistic. Each simulation is repeated 50 times. We display the performances of BSEM and BSEMed by ROC curves, as based on ranking $\bar{\kappa}_{i,r}$, which summarizes $\beta_{i,r}$ and $\beta_{r,i}$ (2.7) (see Figure 2.1). We observe from Figure 2.1 that BSEMed performs better than BSEM when the prior information agrees the data and as good as BSEM when the prior doesn't. The latter reflects the adaptive nature of the EB procedure.

We also consider the EB estimates. We summarize the precisions by their prior means, as estimated by the EB procedure: $E(\tau^2_{i,k}) = \hat{a}_k/\hat{b}_k$, for $k \in \{0, 1\}$. When there is some agreement of the prior knowledge with the data, we expect $\hat{a}_0/\hat{b}_0 > \hat{a}_1/\hat{b}_1$. In the case with 0% true edges, the prior is partly wrong: none of the truly present edges are in the prior network while some of the truly absent edges are part of the prior network. Hence, we expect the EB procedure to produce $\hat{a}_1/\hat{b}_1$ that are slightly larger than $\hat{a}_0/\hat{b}_0$. As discussed in Section 2.2 for the complementary case, reversal of the roles of the two priors can still improve performance of BSEMed, or at least not deteriorate it.

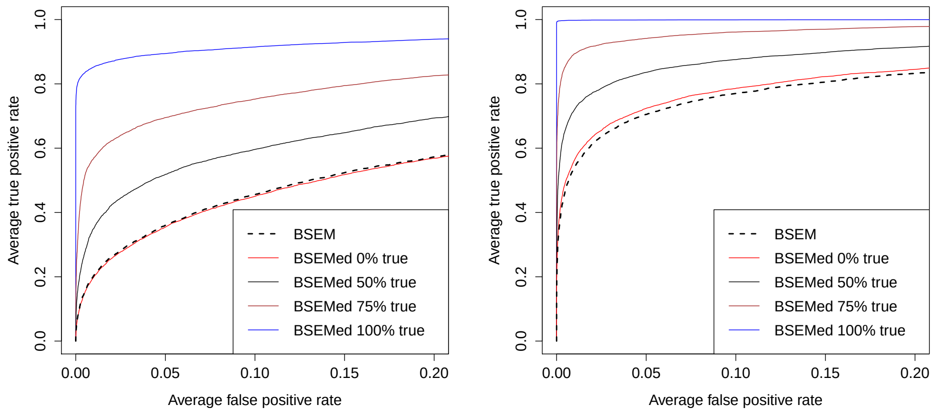
The EB estimates of the prior means are presented in Table 2.1 for the case corresponding to Figure 2.1(a): *band* structure, $n = 50$.

	$\hat{a}_0/\hat{b}_0$	$\hat{a}_1/\hat{b}_1$	ratio
true	366.10	8.08	45.30
0.75% true edges	272.97	14,36	19.00
0.50% true edges	216.10	27.56	7.84
0% true edges	142.59	152.95	1.07

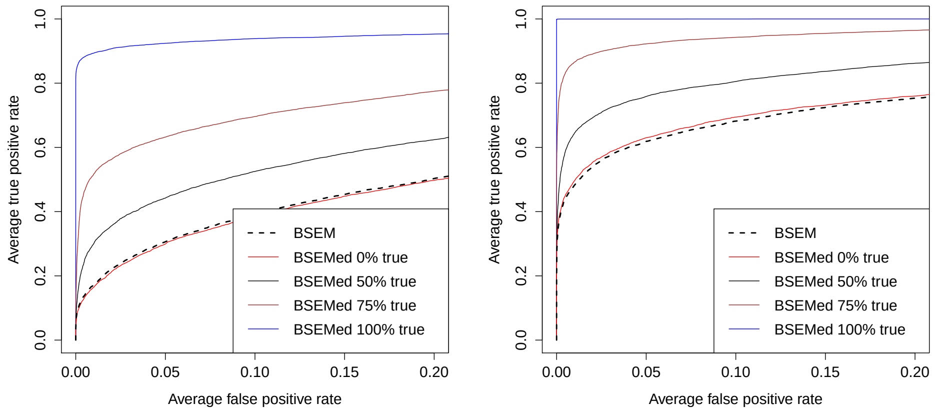
Table 2.1: *EB estimates of the prior means of precisions* $\tau^2_{i,0}$ *and* $\tau^2_{i,1}$ *in case of the* band *structure and* $n = 50$ *for various qualities of prior information*

Table 2.1 displays the prior means of precision, as estimated by EB, for BSEMed models with various qualities of prior information. It is clear that the better the quality of the prior information is, the larger the ratio of mean prior precisions is. Tables for other simulation settings are available in the Supplementary material. These generally show the same pattern.

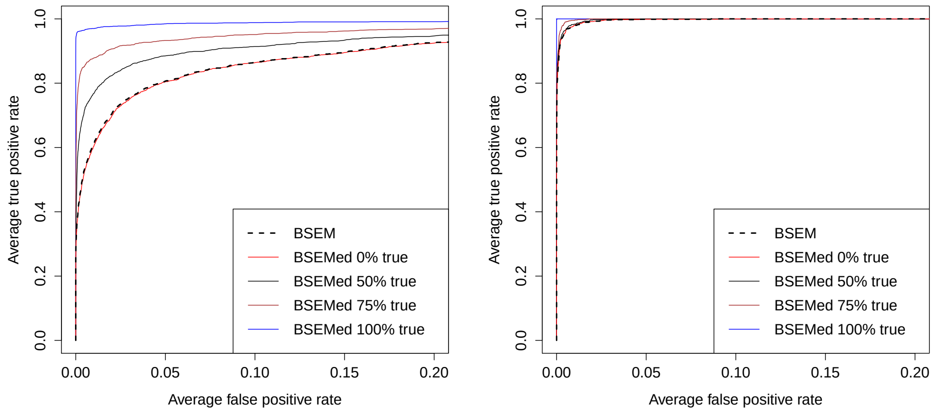
Figure 2.2 displays BSEM and BSEMed estimates of $\beta_{i,r}$ (2.3) and (2.4) for the *band* structure when $n = 50$ and $p = 100$ using the R package *rags2ridges* (Peeters and van Wieringen, 2014; van Wieringen and Peeters, 2014). Figures 2.1 & 2.2 show that BSEMed estimates become more accurate when prior knowledge quality increases and are as good as BSEM estimates when using 0% true edges information. It is also easy to see (Figure 2.2) a convergence of the BSEMed estimates to the true graph when the prior knowledge quality increases.



(a) n = 50 (b) n = 200

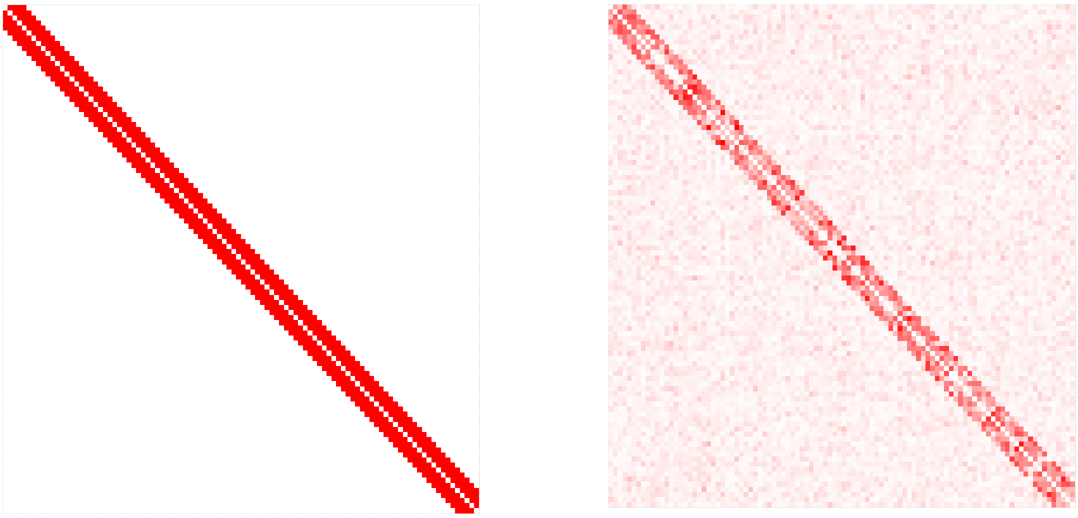


(c) Cluster: n =50 (d)

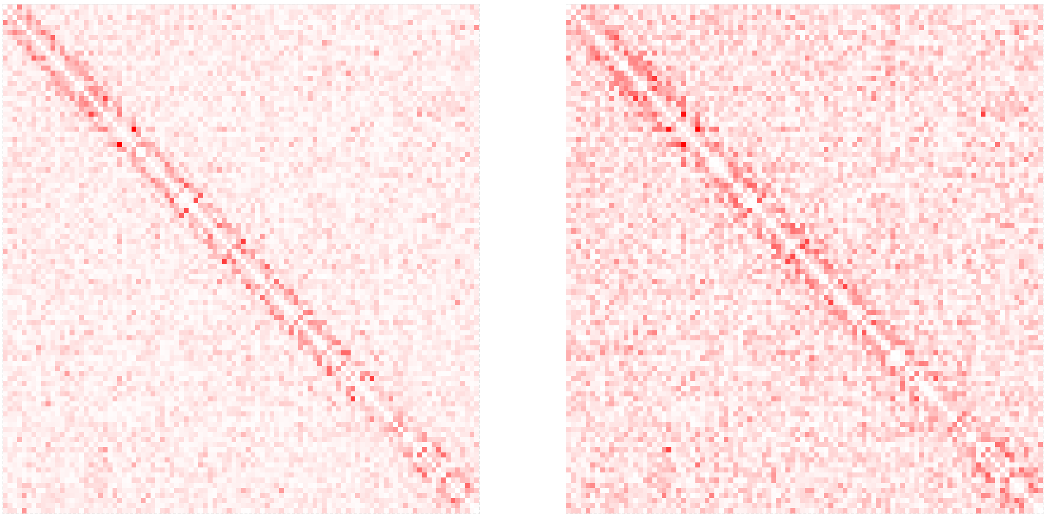


(e) Hub: n = 50 (f) Hub: n = 200

Figure 2.1: *ROC curves for BSEM (dashed) and BSEMed using perfect prior information (blue), BSEMed using 75% true edges present in the prior (brown), BSEMed using 50% true edges present in the prior (black) and BSEMed using 0% true edges present in the prior (red). Here,* $p = 100$ *and* $n \in \{50, 200\}$.



(a) True graph (b) BSEMed: perfect prior



(c) BSEMed: 50 % true Info (d) BSEM

Figure 2.2: *Visualization of BSEMed* $'\bar{\kappa}_{i,r}'$ *using perfect prior (b), BSEMed* $'\bar{\kappa}_{i,r}'$ *using 50% true edges information (c), BSEM* $'\bar{\kappa}_{i,r}'$ *(d) and the true graph (a) in case* $n = 50$ *and* $p = 100$*.*

2.6 Illustration

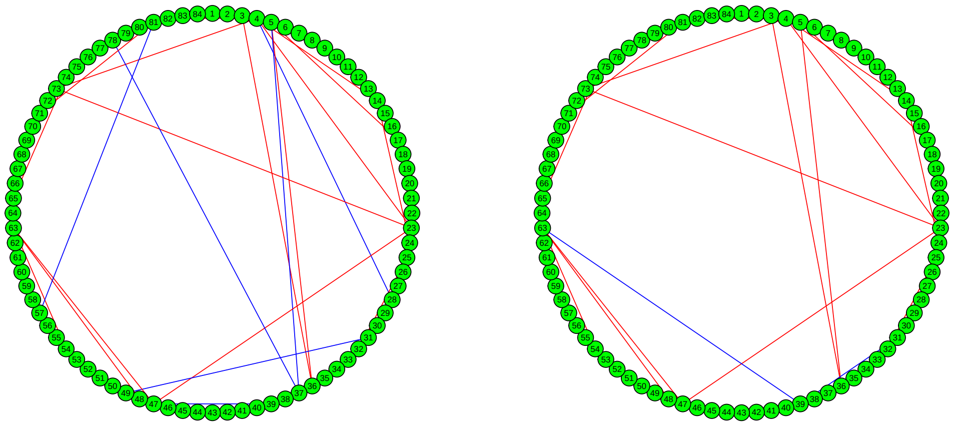
We turn to real data in this section. We use gene expression data from the Gene Expression Omnibus (GEO) to illustrate and evaluate methods for reconstructing gene networks. We consider two types of cancer and cancer-related pathways. First, we focus on the Apoptosis pathway with $p = 84$ genes in a lung data set (Landi et al., 2008), consisting of $n_1^{\text{lung}} = 49$ observations from normal tissue and $n_2^{\text{lung}} = 58$ observations from tumor tissue, so $n^{\text{lung}} = 107$ in total. Secondly, we considered the p53 pathway in a pancreas data set (Badea et al., 2008) with $p = 68$ genes, consisting of $n_1^{\text{pancreas}} = 39$ observations from normal tissue and $n_2^{\text{pancreas}} = 39$ observations from tumor tissue, hence $n^{\text{pancreas}} = 78$ in total. Note that the data were scaled per gene prior to the computations.

BSEMed, BSEM, Graphical Lasso (GL_λ) (Friedman et al., 2008), SEM with the Lasso penalty (SEM_L) (Meinshausen and Bühlmann, 2006) and GeneNet (Schäfer et al., 2006) were applied on the tumor data parts of the data sets. For BSEMed, the corresponding data parts from normal tissue were used as prior knowledge by fitting genes networks on the normal data using BSEM. The idea is that, while tumors and normal tissue may differ quite strongly in terms of mean gene expression, the gene-gene interaction network may be relatively more stable.

We first illustrate the results from BSEM and BSEMed. Before considering the edge selection, we compare the total log-marginal likehood, as estimated by the variational lower bound, across the regression models for BSEM (2.3) and BSEMed (2.4) as a measure for goodness-of-fit. For the lung data set (resp. pancreas data set) we obtained -7082.93 for BSEM and -7071.99 for BSEMed (resp. -3807.58 for BSEM and -3798.91 for BSEMed). These improvements are clearly larger than what may be expected under random prior information of the same size, as shown in Supplementary Material in Section 7.

Figure 2.3 (Figure 2.4) displays the estimated gene-gene network interaction in lung cancer (pancreas cancer) and their overlaps using the described selection procedure with estimated lfdr ≤ 0.1. Considerable overlap (red edges), but also notable differences can be seen.

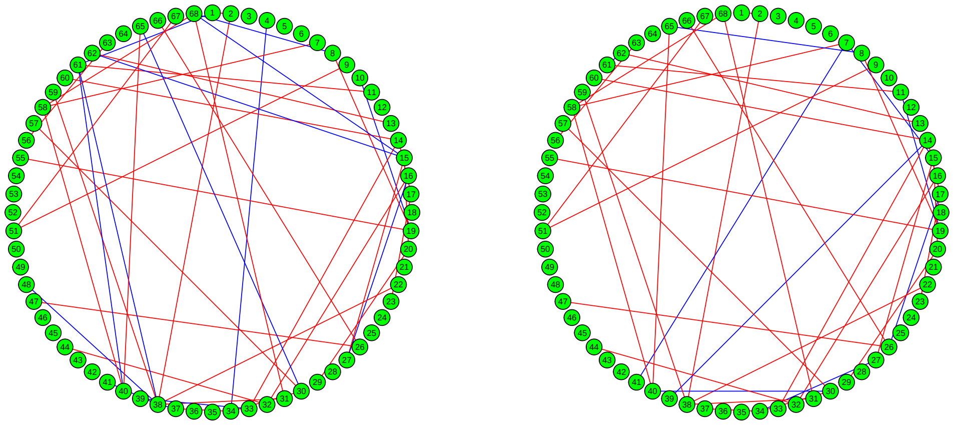
Table 2.2 displays the prior means of precision, as estimated by EB. The prior network is clearly of use: the mean prior precision for regression parameters corresponding to the edges absent in the prior network is relatively large, which effectuates stronger shrinkage towards zero than for parameters corresponding to edges present in the prior network.



(a) BSEM network estimate

(b) BSEMed network estimate

Figure 2.3: *BSEM vs BSEMed network estimates in lung cancer. Red edges are the overlap edges.*



(a) BSEM network estimate

(b) BSEMed network estimate

Figure 2.4: *BSEM vs BSEMed network estimates in pancreas cancer. Red edges are the overlap edges.*

	$\hat{a}_0/\hat{b}_0$	$\hat{a}_1/\hat{b}_1$	ratio
Lung	27.32	1.71	15.97
Pancreas	20.03	1.21	12.97

Table 2.2: *EB estimates of precisions* $\tau^2_{i,0}$ *and* $\tau^2_{i,1}$ *of prior distributions in lung data (resp. pancreas data) set.*

In the following, we argue that BSEMed network estimates may be more reliable in this setting than those of BSEM, Graphical Lasso (GL_λ) (Friedman et al., 2008), SEM with the Lasso penalty (SEM_L) (Meinshausen and Bühlmann, 2006) and GeneNet (Schäfer et al., 2006) (see the Supplementary Material for methodological details). For that, we assess performance of all methods by studying reproducibility of edges. We randomly split the tumor data part of the lung data set (pancreas data set) into two equal and independent parts: $n_{2,1}^{\text{lung}}$ and $n_{2,2}^{\text{lung}}$ (resp. $n_{2,1}^{\text{pancreas}}$ and $n_{2,2}^{\text{pancreas}}$). BSEM, BSEMed, GL_λ, GeneNet and SEM_L were applied on each subset of the tumor data. We repeated the procedure 50 times. We report in Table 2.3 (Table 2.4) the average number of overlapping edges between the two subsets for each method when the total number of edges selected by each method on each subset is set to 50, 100 and 200.

# edges	BSEM overlap	GeneNet overlap	SEM_L overlap	GL_λ overlap	BSEMed overlap	# prior edges in BSEMed
50	4.56	1.88	1.32	3.42	29.58	13.4
100	10.68	5.7	5.64	7.86	37.88	22.14
200	24.16	17.2	16.46	18.14	51.54	33.7

Table 2.3

Lung data, reproducibility study: Average number of overlapping edges among the top 50 (100, 200) strongest ones in two equally-sized splits of the tumor data for BSEMed, BSEM, GL_λ, GeneNet and SEM_L.

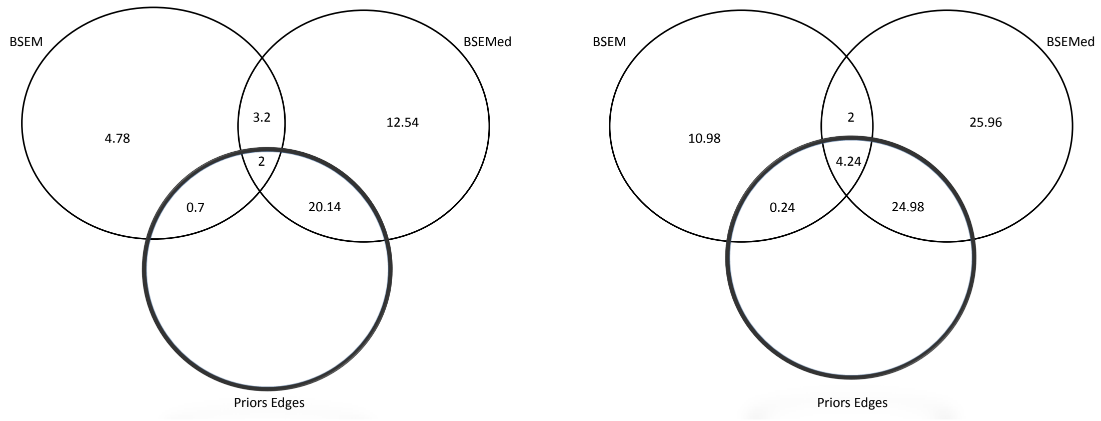
We observe from Tables 2.3 & 2.4 that the results from the BSEMed networks are much more reproducible than that of BSEM, which is on its turn more reproducible than the other ones. Clearly, the improvement can partly be explained by overlapping edges that were also part of the prior network. However, it is clear from Figure 2.5 that the BSEMed network estimate in tumor tissue is not just a 'finger print' of

# edges	BSEM overlap	GeneNet overlap	SEM_L overlap	GL_λ overlap	BSEMed overlap	# prior edges in BSEMed
50	7.42	3.32	2.8	4.52	27.82	11.92
100	17.46	10.34	9.08	11.4	57.18	29.22
200	44.14	30.94	28.54	33.66	81.66	54.1

Table 2.4

Pancreas data, reproducibility study: Average number of overlapping edges among the top 50 (100, 200) strongest ones in two equally-sized splits of the tumor data for BSEMed, BSEM, GL_λ, GeneNet and SEM_L.

the prior network (normal tissue network): BSEMed can even reveal edges that are neither in prior network nor in BSEM network estimate.



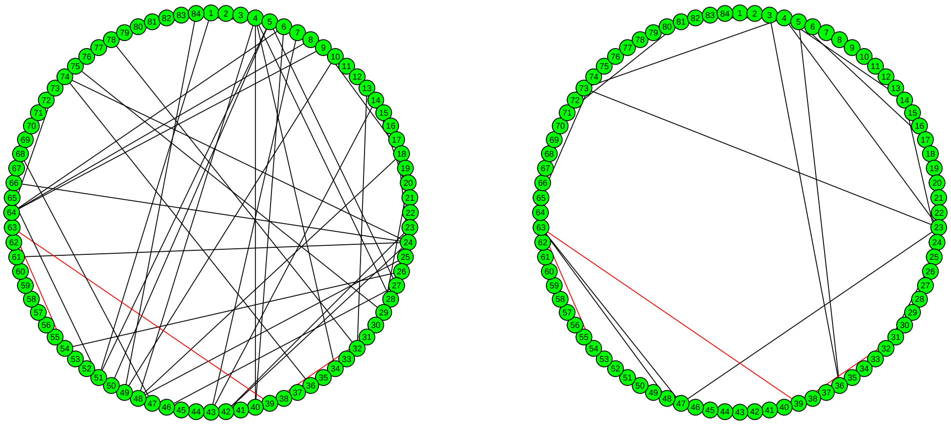
(a) Lung data (b) Pancreas data

Figure 2.5: *Venn diagrams displaying the mean overlap of reproduced top-ranking edges, corresponding to the second row of Table 2.3 (Figure 2.5.a) and Table 2.4 (Figure 2.5.b).*

Figure 2.6 (resp. Figure 2.7) displays the network in normal tissue against the network in tumor tissue in the lung data (resp. in the pancreas data). The purpose of displaying Figure 2.6 and 2.7 is to emphasize the dysregulation of gene-gene interactions in cancer (Vogelstein and Kinzler, 2004; van Wieringen and van der Vaart, 2015) which may be caused by the heterogeneity of cancer (Nowell, 1976). Heterogeneity of tumor samples makes it more difficult to pinpoint reliable links, hence our selection algorithm which is based on local fdr ≤ 0.1 is likely to select fewer links in cancer samples.

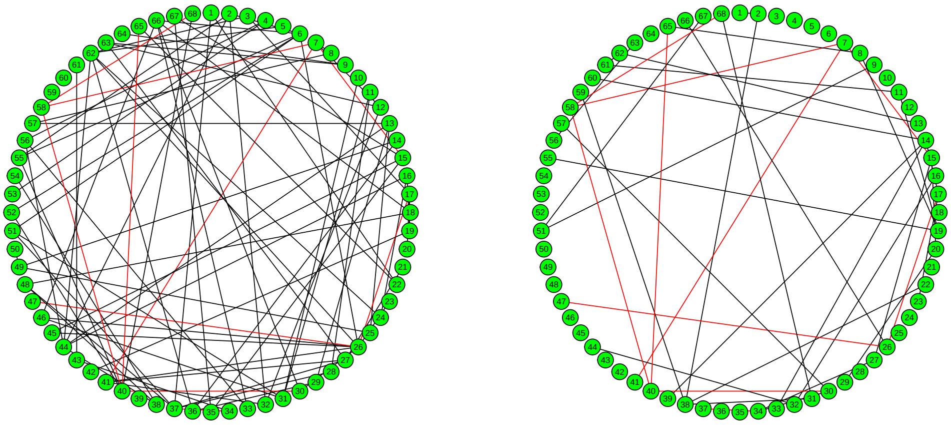
2.7 Discussion

We have presented a new method for incorporating prior information in undirected network reconstrustion based on Bayesian SEM. Our approach allows the use of two central Gaussian distributions per regression equation for coefficients $\beta_{i,r}$'s of our SEMs, where the prior information determines which of the two applies to a specific $\beta_{i,r}$. Empirical Bayes estimation of the parameters of the two hyper priors of the



(a) Network estimate in Normal tissue (b) BSEMed network estimate in tumor tissue

Figure 2.6: *Network in a normal cell vs BSEMed network in lung cancer. Red edges are the overlap edges between prior and posterior networks.*



(a) Network estimate in Normal tissue (b) BSEMed network estimate in tumor tissue

Figure 2.7: *Network in a normal cell vs BSEMed network in pancreas cancer. Red edges are the overlap edges between prior and posterior networks.*

precisions introduces shrinkage and accommodates the situation where there would not be an agreement of the prior information with the data at hand. We showed in simulation with different graph structures that BSEMed outperforms BSEM when the used prior knowledge (partially) agrees with the data and as good as when not.

In addition, for two real data sets we showed better reproducibility of top ranking edges with respect to other methods .

In some cases, it may be desirable to give more weight only to some important edges of the prior graph rather than the whole graph. In gene regulatory networks reconstruction particularly, this may be edges that are known to characterise the disease biology. Assuming one is able to express such prior information as prior probabilities on edges, our software is able to incorporate such information via the Bayes factors used in the post-hoc selection procedure (Leday et al., 2017). Likewise, a user can also increase the weight of the entire prior graph uniformly. (See Supplementary Material for details).

Instead of assigning Gaussian distributions to the coefficients, other (e.g. sparse) priors can be used. The complement property (Section 2.2) is preserved whenever the same functional forms of the priors are used for both classes. However, a combination of e.g. a Gaussian and a sparse prior ruins this property, which renders such a combination less attractive.

Future research also focuses on extending our method to situations with more than two classes. For example, when considering integrative networks for two sets of molecular markers or two (related) pathways, the three class setting is relevant: two classes represent the connections within the two sets and a third one between the two sets. Finally, multiple sources of external data may be available for incorporation in BSEMed. This requires to model the parameter(s) of the priors in terms of contibutions of those external sources, and weigh those sources in a data-driven manner, as it is unlikely that the sources are equally informative.

Chapter 3

Incorporating prior information and borrowing information in high-dimensional sparse regression using the horseshoe and variational Bayes

We introduce a sparse high-dimensional regression approach that can incorporate prior information on the regression parameters and can borrow information across a set of similar datasets. Prior information may for instance come from previous studies or genomic databases, and information borrowed across a set of genes or genomic networks. The approach is based on prior modelling of the regression parameters using the horseshoe prior, with a prior on the sparsity index that depends on external information. Multiple datasets are integrated by applying an empirical Bayes strategy on hyperparameters. For computational efficiency we approximate the posterior distribution using a variational Bayes method. The proposed framework is useful for analysing large-scale data sets with complex dependence structures. We illustrate this by applications to the reconstruction of gene regulatory networks and to eQTL mapping.

This chapter is submitted as: Gino B. Kpogbezan, Mark A. van de Wiel, Wessel N. van Wieringen, and Aad W. van der Vaart. Incorporating prior information and borrowing information in high-dimensional sparse regression using the horseshoe and variational Bayes. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

3.1 Introduction

The analysis of high-dimensional data is important in many scientific areas, and often poses the challenge of the availability of a relatively small number of cases versus a large number of unknown parameters. It has been documented both practically and theoretically that under the assumption of sparsity of the underlying model, larger effects or dependencies can be inferred even in the very high-dimensional case [53, 57]. Still in many cases conclusions can be much improved by incorporating prior knowledge in the analysis, or by "borrowing information" by simultaneously analysing multiple related datasets. In this paper we introduce a methodology that achieves both, and that is at the same time scalable to large datasets in its computational complexity. It is based on an empirical Bayesian setup, where external information is incorporated through the prior, and information is borrowed across similar analyses by empirical Bayes estimation of hyperparameters. Sparsity is induced through utilisation of the horseshoe prior, and computational efficiency through novel variational Bayes approximations to the posterior distribution. We illustrate the methodology by two applications in genomics: network reconstruction and eQTL mapping, but the proposed framework should be useful also for analysing other large-scale data sets with complex dependence structures.

Our working model is a collection of linear regression models, indexed by $i = 1, \ldots, p$, corresponding to p characteristics (e.g. genes). For each characteristic we have measurements on n individuals, labelled $j = 1, \ldots, n$, consisting of a univariate response Y_i^j and a vector X_i^j of s_i explanatory variables. We collect the n responses on characteristic i in the n-vector $Y_i = (Y_i^1, \ldots, Y_i^n)^T$ and similarly collect the explanatory variables in the $n \times s_i$-matrix X_i, having rows X_i^j, and adopt the regression models

$$Y_i = X_i\beta_i + \epsilon_i, \qquad i = 1, \ldots, p. \tag{3.1}$$

Here the regression coefficients β_i form a vector in $\mathbb{R}^{s_i}$, and the error vectors ϵ_i's are unobserved. The dimension s_i of the regression parameter β_i may be different for different characteristics i.

Our full set of observations consists of the pairs $(Y_1, X_1), \ldots, (Y_p, X_p)$, whose stochastic dependence will not be used and hence need not be modelled. In addition to these regression pairs we assume available prior information on the vectors β_i in the form of a 2-dimensional array P, whose ith row presents a grouping of the coordinates of β_i into G groups, indexed by $g = 1, \dots, G$: the value $P_{i,t}$ is the index of the group to which the tth coordinate of β_i belongs. (Because the β_i may have different lengths, P is a possibly "ragged array" and not a matrix.) The information in P is considered to be soft in that coordinates of β_i that are assigned to the same group are thought to be similar in size, but not necessarily equal. The information may for instance come from a previous analysis of similar data, or be taken from a genomic database.

We wish to analyse this data, satisfying four aims:

- Borrow information across the characteristics $i = 1, \dots, p$ by linking the analyses of the models (3.1) for different i.
- Incorporate the prior information P in a soft manner so that it informs the analysis if correct, but can be overruled if completely incompatible with the data.
- Allow for sparsity of the explanatory models, i.e. focus the estimation towards parameter vectors β_i with only a small number of significant coefficients, enabling analysis for small n relative to s_i and/or p.
- Achieve computational efficiency, enabling analysis with large s_i and/or p.

To this purpose we model the parameters β_i and the scales σ_i of the error vectors through a prior, and next perform empirical Bayesian inference. This analysis is informed by the model (3.1) and the following hierarchy of a generating model (referred to as *pInc* later on) for the errors and a prior model for (β_i, σ_i):

$$
\begin{aligned}
\epsilon_i |\, \sigma_i &\sim \mathrm{N}(0_n, \sigma_i^2 \mathbf{I}_n), \\
\beta_{i,t} |\, \sigma_i, \tau_{i,P_{i,t}}, \lambda_{i,t} &\sim \mathrm{N}\Big(0, \sigma_i^2 \tau_{i,P_{i,t}}^2 \lambda_{i,t}^2\Big), \qquad t = 1, \dots, s_i, \\
\sigma_i^{-2} &\sim \Gamma(c, d), \\
\lambda_{i,t} &\sim C^+(0,1), \qquad t = 1, \dots, s_i, \\
\tau_{i,g}^{-2} &\sim \Gamma(a_g, b_g), \qquad g = 1, \dots, G.
\end{aligned}
\tag{3.2}
$$

Here N is a (multivariate) normal distribution, $\mathbf{I}_n$ is the $(n \times n)$-identity matrix, $C^+(0,1)$ denotes the standard Cauchy distribution restricted to the positive real axis, and $\Gamma(u, v)$ denotes the gamma distribution with shape and rate parameters u and v. As usual the hierarchy should be read from bottom to top, where dependencies of distributions on variables at lower levels are indicated by conditioning, and absence of these variables in the conditioning should be understood as the assumption of conditional independence on variables at lower levels of the hierarchy. The specification (3.2) gives the model for the ith characteristic. The models for different i are linked by assuming the same values of the hyperparameters $a_1, b_1, \dots, a_G, b_G, c, d$ for all $i = 1, \dots, p$. These hyperparameters will be estimated from the combined data $(Y_1, X_1), \dots, (Y_p, X_p)$ by the empirical Bayes method, thus borrowing strength across responses and achieving the first of the four aims, as listed previously.

We also consider a variant of the model (later referred to as *pInc2*) in which the last line of the hierarchy is dropped and the parameters $\tau_{i,g}$ are pooled into a single parameter $\tau_{i,g} = \tau_g$ per group $(i = 1, \dots, s_i)$. The parameters τ_g are then estimated by empirical Bayes on the data pooled over i. In some of the simulations this model outperformed (3.2).

The ith row of P gives a grouping of the s_i coordinates $\beta_{i,t}$ of β_i into G groups. The scheme (3.2) attaches a latent variable $\tau_{i,g}$ to each group, for $g = 1, \dots, G$, whose squares possess inverse gamma distributions, independently across groups. These latent variables enter the prior distributions of the coordinates of β_i, which marginally given $\tau_{i,g}$ are scale mixtures of the normal distribution. Choosing the scale parameters $\lambda_{i,t}$ from the half-Cauchy distribution gives the so-called *horseshoe prior* [19, 20]. This may be viewed as a continuous alternative to the traditional *spike-and-slab* prior, which is a mixture of a Dirac measure at zero and a widely spread second component, and is widely used as a prior that induces sparsity.

The horseshoe density with scale τ is the mixture of the univariate normal distributions $N(0, \tau\lambda)$ relative to the parameter $\lambda \sim C^+(0, 1)$. It combines an infinite peak at zero with heavy tails, and is able to either shrink parameters to near zero or estimate them unbiasedly, much as an improper flat prior. The relative weights of the two effects are moderated by the value of τ. In the model (3.2) the coordinates of β_i corresponding to the same group g receive a common parameter $\tau_{i,g}$, and are thus either jointly shrunk to zero or left free, depending on the value of $\tau_{i,g}$. This allows to achieve the aims two and three as listed previously. Theoretical work in [20, 31, 136–138] (in a simpler model) suggests an interpretation of $\tau_{i,g}$ as approximately the fraction of nonzero coordinates in the gth group, and corroborates the interpretation of $\tau_{i,g}$ as a sparsity parameter. In model (3.2) this number is implicitly set by the data, based on the inverse gamma prior on $\tau^2_{i,g}$. Requiring the hyperparameters of these gamma distributions to be the same across the characteristics i induces the borrowing of information between the characteristics i, in particular with respect to the sparsity of the vectors β_i.

Model (3.2) chooses the squares of the scales σ_i of the error variables from an inverse gamma distribution, which is the usual conjugate prior. The priors on the regression parameters β_i are also scaled by σ_i, thus giving them a priori the same order of magnitude. This seems generally preferable.

The Bayesian model described by (3.1) and (3.2) leads to a posterior distribution of (β_i, σ_i) in the usual way, but this depends on the hyperparameters $a_1, b_1, \ldots, a_G, b_G, c, d$. In Section 3.4.2 we introduce a method to estimate these hyperparameters from the full data $(Y_1, X_1), \ldots, (Y_p, X_p)$, and next base further inference on the posterior distributions of the parameters (β_i, σ_i) evaluated at the plugged-in estimates of the hyperparameters. Because the prior on the coefficients β_i is continuous, the posterior distribution does not provide automatic model (or variable) selection, which is a disadvantage of the horseshoe prior relative to the spike-and-slab priors. To overcome this, we develop a way of testing for nonzero regression coefficients based on the marginal posterior distributions of the $\beta_{i,t}$ in Section 3.4.3.

The horseshoe prior has gained popularity, mainly due to its computational advantage over spike-and-slab priors. However, in our high-dimensional setting the approximation of the posterior distribution by an MCMC scheme turns out to be still a computational bottleneck. The algorithm studied by [9], which can be applied in the special case of a single group $(G = 1)$ has complexity $O(n^2 s_i)$ for a single regression (i.e. $p = 1$) per MCMC iteration. We show in Section 3.5.2 that this is too slow to be feasible in our setting. For this reason we develop in Section 3.4.1 a variational Bayesian (VB) scheme to approximate the posterior distribution, in order to satisfy the fourth aim in our list.

The variational Bayesian method consists of approximating the posterior distribution by a distribution of simpler form, which is chosen as a compromise between computational tractability and accuracy of approximation. The quality of the approximation is typically measured by the Kullback-Leibler divergence [141]. Early applications involved standard distributions such as Gaussian, Dirichlet, Laplace and extreme value models [5–7, 96, 142]. In the present paper we use nonparametric approximations, restricted only by the assumption that the various parameters are (block) independent. (This may be referred to as *mean-field* variational Bayes, although this term appears to be used more often for independence of all univariate marginals, whereas we use block independence.) In this case the variational posterior approximation can be calculated by iteratively updating the marginal distributions [11, 104]. Variational Bayes typically produces accurate approximations to posterior means, but have been observed to underestimate posterior spread [12, 18, 48, 94, 131, 143, 145, 151]. We find that in our setting the approximations agree reasonably well to MCMC approximations of the marginals, although the latter take much longer to compute.

The model (3.1)-(3.2) may be useful for data integration in a variety of scientific setups, and for data sources as diverse as gene expression, copy number variations, single nucleotide polymorphisms, functional magnetic resonance imaging, or social media data. The external information incorporated in the array P may reflect data of a different type, and/or of a different stage of research, and the simultaneous analysis of different characteristics allows further data integration. For example, in genetic association studies data from multiple stages can help the identification of true associations [54, 58, 116]. In this paper we consider applications to gene regulation networks and to eQLT mapping, which we describe in the next two sections, before developing the general algorithms for models (3.1) and (3.2).

The remainder of the paper is organised as follows. In Section 3.4.1 we develop a variational Bayes approach to approximate the posterior distributions of the regression parameters for given hyperparameters, and show this to be comparable in accuracy to Gibbs sampling in Section 3.5.2, although computationally much more efficient. In Section 3.4.2 we develop the Empirical Bayes (EB) approach for estimating the hyperparameters, and in Section 3.4.3 we present a threshold based-procedure for selecting nonzero regression coefficients based on the marginal posterior distributions of the $\beta_{i,t}$. We show in Section 3.5 by means of model-based simulations that the proposed approach performs better, in terms of both average ℓ_1-error and average ROC curves, than its ridge counterpart in the framework of network reconstruction. The potential of our approach is shown on real data in Section 3.6 both in gene regulatory network reconstruction and in eQTL mapping. Section 3.7 concludes the paper.

3.2 Network reconstruction

The identification of gene regulatory networks is crucial for understanding gene function, and hence important for both treatment and prediction of diseases. Prior knowledge on a given network is often available in the literature, from repositories or pilot studies, and combining this with the data at hand can significantly improve the accuracy of reconstruction [72].

A *Gaussian graphical model* readily gives rise to a special case of the model (3.1)-(3.2). In such a model the data concerning p genes measured in a single individual (e.g. tissue) is assumed to form a multivariate Gaussian p-vector, and the network of interest is the corresponding *conditional independence graph* [152]. The nodes of this graph are the genes and correspond to the p coordinates of the Gaussian vector. Two nodes/genes are connected by an edge in the graph if the corresponding coordinates are *not* conditionally independent given the other coordinates. It is well known that this is equivalent to the corresponding element in the precision matrix of the Gaussian vector being nonzero [78].

Assume that we observe a gene vector for n individuals, giving rise to n independent copies $Y^1, \dots, Y^n$ of p-vectors satisfying

$$Y^j \sim^{\text{iid}} \mathrm{N}(0_p, \Omega_p^{-1}), \qquad j = 1, \dots, n. \tag{3.3}$$

Here Ω_p is the *precision matrix*; its inverse is the covariance matrix of the vector Y^j and is assumed to be positive-definite. The Gaussian graphical model consists of a graph with nodes $1, 2, \dots, p$ and with edges (i, j) given by the nonzero elements $(\Omega_p)_{i,j}$ of the precision matrix. Hence to reconstruct the conditional independence graph it suffices to determine the non-zero elements of the latter matrix.

We relate this to the notation used in the introduction by writing $Y^j = (Y_1^j, \dots, Y_p^j)^T$, and next collecting the observations Y_i^j per gene i, giving the n-vector $Y_i = (Y_i^1, \dots, Y_i^n)^T$, for $i = 1, \dots, p$. We next define

$$X_i = [Y_1, Y_2, ..., Y_{i-1}, Y_{i+1}, ..., Y_p]$$

as the $(n \times (p-1))$-matrix with columns Y_t, for $t \neq i$. It is well known that the residual when regressing a single coordinate Y_i^j of a multivariate Gaussian vector linearly on the other coordinates Y_t^j, for $t \neq i$, is Gaussian. Furthermore, the regression coefficients $\beta_i = (\beta_{i,t} : t \neq i)$ can be expressed in the precision matrix of Y^j as

$$\beta_{i,t} = -\frac{(\Omega_p)_{it}}{(\Omega_p)ii}.$$

This shows that (3.1) holds with $s_i = p - 1$ and a multivariate normal error vector ϵ_i with variance σ_i^2 equal to the residual variance. Moreover, the (non)zero entries in the ith row vector of the precision matrix Ω_p correspond to the (non)zero coordinates of β_i. Consequently, the problem of identifying the Gaussian graphical model can be cast as a variable selection problem in the p regression models (3.1).

This approach of recasting the estimation of the (support of the) precision matrix as a collection of regression problems was introduced by [97], who employed Lasso regression [43, 130] to estimate the parameters. Other variable selection methods can be employed as well [73]. A Bayesian approach with Gaussian, ridge-type priors on the regression coefficients was developed in [80], and extended in [72] to incorporate prior knowledge on the conditional independence graph. A disadvantage of the Gaussian priors employed in these papers is that they are not able to selectively shrink parameters, but shrink them jointly towards zero (although prior information used in [72] alleviates this by making this dependent on prior group). This is similar to the shrinkage effect of the ridge penalty [139] relative to the Lasso, which can shrink some of the precision matrix elements to exactly zero, and hence possesses intrinsic model selection properties. The novelty of the present paper is to introduce the horseshoe prior in order to better model the sparsity of the network.

We assume that the prior knowledge on the to-be-reconstructed network is available as a "prior network", which specifies which edges (conditional independencies) are likely present or absent. This information can be coded in an adjacency matrix P, whose entries take the values 0 or 1 corresponding to the absence and presence of an edge: $\mathrm{P}_{i,t} = 1$ if variable i is connected with variable t and $\mathrm{P}_{i,t} = 0$ otherwise. Thus in this example we only have two groups, i.e. $G = 2$.

The advantage of reducing the network model to structural equation models of the type (3.1) is computational efficiency. An alternative would be to model the precision matrix directly through a prior. This would typically consist of a prior on the graph structure, followed by a specification of the numerical values of the precision matrix given its set of nonzero coefficients. The space of graphs is typically restricted to e.g. decomposable graphs, forests, or trees [33, 50, 68]. The posterior distribution of the graph structure can then be used as the basis of inference on the network topology. However, except in very small problems, the computational burden is prohibitive.

3.3 eQTL mapping

In eQTL mapping the expression of a gene is taken as a quantitative trait, and it is desired to identify the genomic loci that influence it, much as in a classical study of quantitative trait loci (QTL) of a general phenotype. Typically one measures the expression of many genes simultaneously and tries to map these to their QTL. Since gene expression levels are related to disease susceptibility, elucidating these eQLT (expression QTL) may give important insights into the genetic underpinnings of complex traits. We shall identify genetic loci here with single nucleotide polymorphisms (SNPs), but other biomarkers can be substituted.

Early work by [26, 128, 165] considered every gene separately for association. However, many genes are believed to be co-regulated and to share a common genetic basis [113, 162]. In addition, SNPs with pleiotropic effects may be more easily identified by considering multiple genes together. Therefore following [71, 83, 125], we focus on a joint analysis, borrowing information across genes. We regress the expression of a given gene on SNPs both within and around the gene, where our model is informed about the SNP location. The sparse parametrization offered by our model is suitable, as most genetic variants are thought to have a negligible (if any) differential effect on expression.

Suppose we collect the (standardized) expression levels of p genes over n individuals, and identify for each gene i a collection of s_i SNPs to be investigated for association. For instance, the latter collections might contain all SNPs in a relatively large window around the gene, some of which falling inside the gene and some outside. For each individual and SNP we ascertain the number of minor alleles (0, 1 or 2), and change all 2's to 1's. Because there are not many 2's in the data this does not reduce the information while it simplifies the modelling. We use these numbers to form the $n \times s_i$-matrix X_i. Let Y_i be the n-vector of expression levels for gene i, and assume the linear model (3.1).

It is believed that SNPs that occur within a gene may play a more direct role in the gene's function than SNPs at other genomic locations [84, 123]. Therefore, it is natural to treat SNPs falling within a given gene differently than the ones not falling within that gene. This gives rise to two groups of SNPs for a given gene, which we can encode as prior knowledge in a 2-dimensional array P with values 0 and 1.

Thus we have another instance of model (3.1)-(3.2) with two groups, i.e. $G = 2$.

3.4 Posterior inference

In this section we discuss statistical inference for the model (3.1)-(3.2). This consists of three steps: the approximation to the posterior distribution of the model for given hyperparameters (and given i), the estimation of the hyperparameters (across i), and finally a method of variable selection.

3.4.1 Variational Bayes approximation

The *variational Bayes approximation* to a distribution is simply the closest element in a given target set $\mathcal{Q}$ of distributions, usually with "distance" measured by Kullback-Leibler divergence [141]. In our situation we wish to approximate the posterior distribution of the parameter $\theta_i := (\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}, \cdots, \tau_{i,G}, \sigma_i)$ given Y_i in the model (3.1)-(3.2), for a fixed i. Here we take the regression matrix X_i as given.

Thus the variational Bayes approximation is given as the density $q \in \mathcal{Q}$ that minimizes over $\mathcal{Q}$,

$$KL\Big(q||p(\cdot|\,Y_i)\Big) = \mathbf{E}_q \log \frac{q(\theta_i)}{p(\theta_i|\,Y_i)} = \log p(Y_i) - \mathbf{E}_q \log \frac{p(Y_i,\theta_i)}{q(\theta_i)},$$

where $\theta_i \mapsto p(\theta_i|\,Y_i)$ is the posterior density, the expectation is taken with respect to θ_i having the density $q \in \mathcal{Q}$, and $(y,\theta_i) \mapsto p(y,\theta_i) = p(y|\,\theta_i)\,\pi_i(\theta_i)$ and $y \mapsto p(y) = \int p(y,\theta_i)\,d\theta_i$ are the joint density of (Y_i,θ_i) and the marginal density of Y_i, respectively, in the model (3.1)-(3.2), with prior density π_i on θ_i. As the marginal density is free of q, minimization of this expression is equivalent to maximization of the second term

$$\mathbf{E}_q \log \frac{p(Y_i,\theta_i)}{q(\theta_i)}. \tag{3.4}$$

By the non-negativity of the Kullback-Leibler divergence, this expression is a lower bound on the logarithm of the marginal density $p(Y_i)$ of the observation. For this reason it is usually referred to as “the lower bound”, or “ELBO”, and solving the variational problem is equivalent to maximizing this lower bound.

The set $\mathcal{Q}$ is chosen as a compromise between computational tractability and accuracy of approximation. Restricting $\mathcal{Q}$ to distributions for which all marginals of θ_i are independent is known as *mean-field* variational Bayes, or also as the “naïve factorization” [141]. Here we shall use the larger set of distributions under which the blocks of β, λ, τ and σ-parameters are independent. Thus we optimize over probability densities q of the form

$$q(\theta_i) = q_\beta(\beta_i)\cdot q_\lambda(\lambda_{i,1},\cdots,\lambda_{i,s_i})\cdot q_\tau(\tau_{i,1},\cdots,\tau_{i,G})\cdot q_\sigma(\sigma_i).$$

There is no explicit solution to this optimization problem. However, if all marginal factors but a single one in the factorization are fixed, then the latter factor can be characterised easily, using the non-negativity of the Kullback-Leibler divergence. This leads to an iterative algorithm, in which the factors are updated in turn.

In the Appendix Section we show that in our case the iterations take the form:

$$\begin{aligned}
\beta_i|\,Y_i &\sim \mathrm{N}\Big(\beta_i^*,\Sigma_i^*\Big), && \\
\lambda_{i,t}|\,Y_i &\sim \Lambda_{\lambda_{it}}, && t = 1,\cdots,s_i,\\
\tau_{i,g}^{-2}|\,Y_i &\sim \Gamma(a_{i,g}^*,b_{i,g}^*), && g = 1,\cdots,G,\\
\sigma_i^{-2}|\,Y_i &\sim \Gamma\Big(c_i^*,d_i^*\Big), &&
\end{aligned} \tag{3.5}$$

where Λ_l is the distribution with probability density function proportional to

$$\lambda \mapsto \frac{1}{\lambda(1+\lambda^2)} e^{-l\lambda^{-2}}, \qquad (\lambda > 0),$$

and the parameters on the right hand side satisfy

$$\begin{aligned}
\Sigma_i^* &= \left[\mathbf{E}_{q_\sigma^*}(\sigma_i^{-2})\Big(X_i^T X_i + \mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}}\Big)\right]^{-1}, \\
\beta_i^* &= \Big(X_i^T X_i + \mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}}\Big)^{-1} X_i^T Y_i, \\
a_{i,g}^* &= a_g + 0.5 \cdot \frac{s_i^g}{2}, \\
b_{i,g}^* &= b_g + 0.5 \cdot \mathbf{E}_{q_\sigma^*}(\sigma_i^{-2}) \mathbf{E}_{q^*_{-\tau_g}} \left(\beta_i^{gT} \mathbf{D}_{\lambda_i}^{-1} \beta_i^g\right), \qquad g = 1, \cdots, G, \\
c_i^* &= c + \frac{n}{2} + \frac{s_i}{2}, \\
d_i^* &= d + 0.5 \cdot \mathbf{E}_{q^*_{-\sigma}} \left(\beta_i^T \mathbf{D}^{-1}_{\tau_i\lambda_i} \beta_i\right) + 0.5 \cdot \mathbf{E}_{q^*_\beta}(Y_i - X_i\beta_i)^T (Y_i - X_i\beta_i), \\
\mathbf{D}_{\lambda_i} &= \operatorname{diag}(\lambda^2_{i,1}, \ldots, \lambda^2_{i,s_i}), \\
\mathbf{D}_{\tau_i\lambda_i} &= \operatorname{diag}(\tau^2_{i,P_{i,1}} \lambda^2_{i,1}, \ldots, \tau^2_{i,P_{i,s_i}} \lambda^2_{i,s_i}), \\
\mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}} &= \operatorname{diag}\left(\mathbf{E}_{q^*_{\tau_i}}(\tau^{-2}_{i,P_{i,1}}) \mathbf{E}_{q^*_{\lambda_{i1}}}(\lambda^{-2}_{i,1}), \ldots, \mathbf{E}_{q^*_{\tau_i}}(\tau^{-2}_{i,P_{i,s_i}}) \mathbf{E}_{q^*_{\lambda_{is_i}}}(\lambda^{-2}_{i,s_i})\right), \\
l_{it} &= \frac{1}{2} \mathbf{E}_{q^*_\sigma}(\sigma_i^{-2}) \mathbf{E}_{q^*_\tau}(\tau^{-2}_{i,P_{i,t}}) \mathbf{E}_{q^*_\beta}(\beta^2_{i,t}).
\end{aligned}$$

In these expressions, s_i^g is the number of g's in the i-th row of the 2-dimensional array P encoding the G groups, $g = 1, \cdots, G$; and $\beta_i^g = \{\delta_{\{\mathrm{P}_{i,r}=g\}}\beta_{i,r} : r \in \{1, \cdots, s_i\}\}$ is the vector obtained from β_i by replacing the coordinates not corresponding to group g by 0.

The expected value of $z_{it} := (\lambda_{it})^{-2}$, which appears in the expression of β_i^*, Σ_i^*, $b_{i,g}^*$ and d_i^* above, is given in the following lemma.

Lemma 1. *The norming constant for* Λ_l *is* $2\exp(-l)/E_1(l)$ *and the expectation of* $z_{it} = (\lambda_{it})^{-2}$ *if* $\lambda_{it} \sim \Lambda_{\lambda_{it}}$ *is given by*

$$\mathbf{E}(z_{it}) = \frac{1}{l_{it} \cdot \exp(l_{it}) \cdot E_1(l_{it})} - 1,$$

where E_1 *is the* exponential integral function of order 1, *defined by*

$$E_1(x) \equiv \int_x^\infty \frac{e^{-t}}{t} dt, \qquad x \in \mathbb{R}^+.$$

Proof. This follows by easy manipulation and the standard density transform formula. ☐

The function E_1 can be evaluated effectively by the function `expint_E1()` in the R package **gsl** [56]. The latter uses the GNU Scientific Library [45].

In addition, the variational lower bound (3.8) on the log marginal likelihood at $q = q^*$ takes the form (See Appendix for details)

$$
\begin{aligned}
\mathcal{L}_i = &-\frac{n}{2}\log(2\pi) - s_i\log(\pi) + \frac{1}{2}\log|\Sigma_i^*| + \frac{1}{2}s_i\\
&+\sum_{g=1}^{G}(a_g\log b_g - \log\Gamma(a_g) - a_{i,g}^*\log b_{i,g}^* + \log\Gamma(a_{i,g}^*))\\
&+c\log d - \log\Gamma(c) - c_i^*\log d_i^* + \log\Gamma(c_i^*)\\
&+\sum_{g=1}^{G}\Big(\frac{1}{2}\mathbf{E}_{q_\sigma^*}(\sigma_i^{-2})\mathbf{E}_{q_\tau^*}(\tau_{i,g}^{-2})\mathbf{E}_{q^*}(\beta_i^{g^T}\mathbf{D}_{\lambda_i}^{-1}\beta_i^g)\Big)\\
&+\sum_{t=1}^{s_i}\Big(\log E_1(l_{it}) + \frac{1}{\exp(l_{it})E_1(l_{it})}\Big).
\end{aligned}
\tag{3.6}
$$

3.4.2 Global Empirical Bayes

Model (3.2) possesses the $G+1$ pairs of hyperparameters $(a_1, b_1), \cdots, (a_G, b_G), (c, d)$. The pair (c, d) controls the prior of the error variances σ_i^2; we fix this to numerical values that render a vague prior, e.g. to $(0.001, 0.001)$. In contrast, we let the values of the parameters $\alpha = (a_1, b_1, \cdots, a_G, b_G)$ be determined by the data. As these hyperparameters are the same in every regression model i, this allows information to be borrowed across the regression equations, leading to *global shrinkage* of the regression parameters. The approach is similar to the one in [134].

Precisely, we consider the criterion

$$
\begin{aligned}
\alpha = (a_1, b_1, \cdots, a_G, b_G) \mapsto &\sum_{i=1}^{p}\mathbf{E}_q\log\frac{p_\alpha(Y_i, \theta_i)}{q(\theta_i)}\\
&= \sum_{i=1}^{p}\mathbf{E}_q\log\frac{p(Y_i|\,\theta_i)}{q(\theta_i)} + \sum_{i=1}^{p}\mathbf{E}_q\log\pi_\alpha(\theta_i).
\end{aligned}
\tag{3.7}
$$

The maximization of the function on the right with respect to $q \in \mathcal{Q}$ for fixed α leads to the variational estimator q^* considered in Section 3.4.1 (which depends on $\alpha = (a_1, b_1, \cdots, a_G, b_G)$). Rather than running the iterations (3.5) for computing this estimator to “convergence”, next inserting $q = q_\alpha^*$ in the preceding display (3.15), and finally maximizing the resulting expression with respect to α, we blend iterations to find q^* and α^* as follows. Given an *iterate* q^* of (3.5) we set q in (3.15) equal to q^* and find its maximizer α^* with respect to α. Next given α^* we set α (in the display following (3.5) equal to α^* and use (3.5) to find a next iterate of q^*. We repeat these alternations to "convergence".

For fixed $q = q^*$ the far right side in the second row of the preceding display depends on α only through

$$\sum_{i=1}^{p} \mathbf{E}_{q^*}\Big(\log \pi_\alpha(\theta_i)\Big).$$

Using the approximation $\log(x) - \frac{1}{2x} \approx \Psi(x) = \frac{\partial}{\partial x}\log\Gamma(x)$, where Ψ is the digamma function, the maximization yields (see Appendix for details)

$$\begin{aligned}
\hat{a}_g &\approx \tfrac{1}{2}\Bigg[\log\Bigg(\sum_{i=1}^{p}\mathbf{E}_{q^*}\tau_{i,g}^{-2}\Bigg) - p^{-1}\Bigg(\sum_{i=1}^{p}\mathbf{E}_{q^*}\log\tau_{i,g}^{-2}\Bigg) - \log p\Bigg]^{-1}\\
\hat{b}_g &= \hat{a}_g \cdot p \cdot \Bigg[\sum_{i=1}^{p}\mathbf{E}_{q^*}\tau_{i,g}^{-2}\Bigg]^{-1}
\end{aligned}$$

where $g \in \{1, \cdots, G\}$. The following algorithm summarizes the above described procedure.

Variational algorithm with sparse local-global shrinkage priors

```
Initialize
$a_g^{(0)} = b_g^{(0)} = 10^{-3}$, $g \in \{1,\cdots,G\}$ and $\forall i \in \mathcal{I}$, $b_{i,g}^* = d_i^* = 10^{-3}$, $\epsilon = 10^{-3}$,
M $= 10^3$ and $k = 1$
while max $|\mathcal{L}_i^{(k)} - \mathcal{L}_i^{(k-1)}| \geq \epsilon$ and $2 \leq k \leq$M do
    E-step: Update variational parameters
        for $i = 1$ to $p$ update
            $a_{i,g}^{*(k)}$, $c_i^{*(k)}$,
            $\Sigma_i^{*(k)}$, $\beta_i^{*(k)}$, $b_{i,g}^{*(k)}$, $d_i^{*(k)}$, $l_{it}^{(k)}$ and $\mathcal{L}_i^{(k)}$;    $\forall g$ and $\forall t$ in that order
        end for

    M-step: Update hyperparameters
    $a_g^{(k)}$, $b_g^{(k)}$;   $\forall g$
        $k \leftarrow k + 1$
end while
```

3.4.3 Variable selection

Because the horseshoe prior is continuous, the resulting posterior distribution does not set parameters exactly equal to zero, and hence variable selection requires an additional step. We investigated two schemes that both take the marginal posterior distributions of the parameters as input.

Thresholding

A natural method is to set a parameter $\beta_{i,r}$ equal to zero (i.e. remove the corresponding independent variable from the regression model) if the point 0 is in the tails of its marginal posterior distribution, or more precisely, if 0 does not belong to a central marginal credible interval for the parameter. Given that our variational Bayes scheme produces conditional Gaussian distributions, this is also equivalent to the absolute ratio of posterior mean and standard deviation

$$\kappa_{i,r} = \frac{\left|\mathbf{E}_{q^{i*}}\left[\beta_{i,r}|\mathbf{Y}_i\right]\right|}{\mathbf{sd}_{q^{i*}}\left[\beta_{i,r}|\mathbf{Y}_i\right]} \tag{3.8}$$

exceeding some threshold. (In the network setup of Section 3.2 we use the symmetrized quantity $(\kappa_{i,r}+\kappa_{r,i})/2$, as the two constituents of the average refer to the same parameter.)

To determine a suitable cutoff or credible level we applied the variational Bayes procedure of Section 3.4.1 with all credible levels η on a grid with step size 5% within the range $[10\%, 99.99\%]$, resulting in a model, or set of 'nonzero' parameters $\beta_{i,r}$, for every η. We allow rather lenient credible levels because the model might benefit from the inclusion of fewer variables, in particular when strong collinearity is present. We next refitted the model (3.1)-(3.2) with the non-selected parameters $\beta_{i,r}$ set equal to 0, evaluated the variational Bayes lower bound on the likelihood (3.8) (equivalently (3.6)), and chose the value of η and the corresponding model that maximized this likelihood. When refitting we did not re-estimate the hyperparameters (a's and b's for *pInc*, τ's for *pInc2*, as explained in Section 3.4.2), but used the values resulting from the entire data set. Even though this procedure sounds involved, it is computationally fast, because it is free of the empirical Bayes step and typically needs to evaluate only models with few predictors.

An alternative selection scheme

As an alternative selection scheme we investigated the *decoupled shrinkage and selection* (DSS) criterion proposed by [53]. For each regression model i, given the posterior mean vector $\bar{\beta}_i = \mathbf{E}_{q^{i*}}\left[\beta_{i,\cdot}|\mathbf{Y}_i\right]$ determined by the pooled procedure of Sections 3.4.1-3.4.2, this calculates the adaptive lasso type estimate

$$\hat{\boldsymbol{\gamma}}_i(\lambda_i) = \underset{\boldsymbol{\gamma}_i}{\operatorname{argmin}} \left[\frac{1}{n} \|\mathbf{X}_i \bar{\beta}_i - \mathbf{X}_i \boldsymbol{\gamma}_i\|_2^2 + \lambda_i \sum_{t=1}^{p} \frac{|\gamma_{i,t}|}{|\bar{\beta}_{i,t}|} \right], \tag{3.9}$$

and next chooses the model corresponding to the nonzero coordinates of $\boldsymbol{\gamma}_i$. The authors [53] advocate this method over thresholding, in particular because it may better handle multi-collinearity. In genomics applications, such as the eQTL Example (Section 3.6.2), multi-collinearity is likely strong, in particular between neighbouring genomic locations. Another attractive aspect of (3.9) is that it only relies on the posterior means, which we have shown to be accurately estimated by the variational Bayes approximation.

In the DSS approach the thresholding in order to obtain models of different sizes is performed through the smoothing parameters λ_i. The authors [53] propose a heuristic to choose λ_i based on the credible interval of the explained variation. An alternative is to apply K-fold cross-validation based on the squared prediction error:

$$\text{MSE}(\lambda_i) = \frac{1}{n} \sum_{k=1}^{K} \|\mathbf{Y}_i^k - \mathbf{X}_i^k \hat{\boldsymbol{\gamma}}_i^{-k}(\lambda_i)\|_2^2, \tag{3.10}$$

where superscript k refers to the observations used as test sample in fold $k = 1, \ldots, K$, and $-k$ to the complementary training sample used to calculate $\hat{\boldsymbol{\gamma}}_i^{-k}(\lambda_i)$, by (3.9) with $\mathbf{X}_i^{-k}$ and $\bar{\beta}_i^{-k}$ replacing $\mathbf{X}_i$ and $\bar{\beta}_i$. Again we throughout fix the hyperparameters of the priors to the ones resulting from the variational Bayes algorithm on the entire data set. We have found that the function $\lambda_i \mapsto \text{MSE}(\lambda_i)$ can be flat, which, to some extent, is a 'by-product' of the strong shrinkage properties of the horseshoe prior. (Given a sparse true vector, many posterior means $\bar{\beta}_{i,r}$ will be close to zero, which renders the DSS solution (3.9) less dependent on λ_i.) To overcome this, and because we prefer sparser models, we used the maximum value of λ_i for which the MSE is within 1 standard error of the minimum of the mean square errors.

In the next sections, if not specified, selection should be understood as the first scheme based on thresholding.

3.5 Simulations

We performed model-based simulations to compare model (3.2), referred to as *pInc*, with the alternative method *pInc2*, in which there is only one parameter τ_g per group, and their ridge counterpart *ShrinkNet* ([80]). We refer to the latter paper for comparisons of *ShrinkNet* to other competing methods. *ShrinkNet* was indeed shown in [80] to outperform the *graphical lasso* [43], the *SEM Lasso* [97] and the *GeneNet* [120] using exactly the same data used below in this simulation. As *ShrinkNet* was developed for network reconstruction only and does not incorporate prior knowledge, we initially considered the setup of network reconstruction in Section 3.2 and set $G = 1$ in (3.2). Next we compared *pInc* and *pInc2* in the same network recovery context, but incorporating prior information. Finally, we compared the accuracy and computing time of our variational Bayes approximation approach with Gibbs sampling-based strategies [9].

3.5.1 Model-based simulation

We generated data $Y^1, \dots, Y^n$ according to (3.3), for $p = 100$ and $n \in \{10, 100, 200, 500\}$ to reflect high and low-dimensional designs. We generated precision matrices Ω_p corresponding to *band*, *cluster* and *hub* network topologies [80, 163] from a G-Wishart distribution [101] with scale matrix equal to the identity and $b = 4$ degrees of freedom.

The performance of the methods was investigated using average ℓ_1 errors $\|\hat{\beta}_0 - \beta_0\|_1$ and $\|\hat{\beta}_1 - \beta_1\|_1$ across 50 replicates of the experiment. Here β_1 (or β_0) is the vector consisting of all nonzero (or zero) values of the partial correlation matrix $-(\Omega_p)_{it}/(\Omega_p)ii$ except the diagonal elements, and $\hat{\beta}_1$ (or $\hat{\beta}_0$) is the vector consisting of the corresponding posterior means.

The results are displayed in Tables 3.1 and 3.2. Both methods *pInc* and *pInc2* outperform *ShrinkNet* in all simulation setups. For the nonzero parameters ('signals') *pInc* and *pInc2* are on par, but for the zero parameters *pInc* outperforms *pInc2* for small n in the Band and Cluster topologies, but when n increases and in the Hub topology this turns around.

Somewhat worrisome is that the performance of all methods on the zero parameters initially seems to suffer from increasing sample size n. The empirical Bayes choice of shrinkage level clearly favours strong shrinkage for small n, giving good performance on the zero parameters, but relaxes this when the sample size increases. Thus the better performance for increasing n on the nonzero parameters is partly

	Sample size	*ShrinkNet*	*pInc2*	*pInc*
Band	$n = 10$	25.26	1.77	0.66
	$n = 100$	265.89	180.42	78.46
	$n = 200$	291.33	113.12	121.29
	$n = 500$	251.47	81.38	150.62
Cluster	$n = 10$	15.74	0.71	0.51
	$n = 100$	224.89	186.88	39.97
	$n = 200$	259.94	130.70	98.77
	$n = 500$	231.33	82.82	107.58
Hub	$n = 10$	7.44	0.28	0.34
	$n = 100$	155.87	8.70	47.85
	$n = 200$	154.63	12.65	84.46
	$n = 500$	132.50	21.51	106.31

Table 3.1: *Average* l_1 *error,* $\|\hat{\beta}_0 - \beta_0\|_1$ *across* 50 *simulation replicates with sample size* $n \in \{10, 100, 200, 500\}$ *and* $p = 100$*. The precision matrices used correspond respectively to Band, Cluster and Hub structure.*

	Sample size	*ShrinkNet*	*pInc2*	*pInc*
Band	$n = 10$	220.15	220.55	221.92
	$n = 100$	162.58	112.01	134.82
	$n = 200$	124.01	66.08	65.66
	$n = 500$	72.51	29.08	29.25
Cluster	$n = 10$	288.86	288.64	289.44
	$n = 100$	254.03	160.05	217.48
	$n = 200$	215.88	75.24	86.54
	$n = 500$	133.22	27.99	29.95
Hub	$n = 10$	40.25	39.34	40.52
	$n = 100$	24.14	15.39	13.99
	$n = 200$	17.58	9.42	8.65
	$n = 500$	12.54	5.42	5.26

Table 3.2: *Average* l_1 *error,* $\|\hat{\beta}_1 - \beta_1\|_1$ *across* 50 *simulation replicates with sample size* $n \in \{10, 100, 200, 500\}$ *and* $p = 100$*. The precision matrices used correspond respectively to Band, Cluster and Hub structure.*

	Quality of prior Info	*pInc2*	*pInc*
Band	True model	6.90	0.68
	50% true edge info	6.66	5.30
Cluster	True model	4.96	0.60
	50% true edge info	3.25	3.28
Hub	True model	0.22	0.27
	50% true edge info	0.46	5.88

Table 3.3: *Average* l_1 *error,* $\|\hat{\beta}_0 - \beta_0\|_1$ *across* 50 *simulation replicates with sample size* $n = 10$ *and* $p = 100$*. Qualities of prior information correspond to true model and* 50% *true edge information.*

	Quality of prior Info	*pInc2*	*pInc*
Band	True model	216.25	209.48
	50% true edge info	219.57	217.39
Cluster	True model	285.72	281.21
	50% true edge info	286.98	286.73
Hub	True model	29.40	27.55
	50% true edge info	37.79	34.60

Table 3.4: *Average* l_1 *error,* $\|\hat{\beta}_1 - \beta_1\|_1$ *across* 50 *simulation replicates with sample size* $n = 10$ *and* $p = 100$*. Qualities of prior information correspond to true model and* 50% *true edge information.*

offset by a decline in performance on the zero parameters. This balance between zero and nonzero parameters is restored only for relatively large sample sizes. A similar phenomenon was observed in [135].

Tables 3.3 and 3.4 compare the performance of *pInc* and *pInc2* when prior information is available (both with sample size $n = 10$). The prior information consists either of the correct adjacency matrix P for the network (i.e. $P_{i,t} = 1$ if $\Omega_{i,t} \neq 0$ and $P_{i,t} = 0$ otherwise), or an adjacency matrix in which 50 % of the positive entries are correct. The latter matrix was obtained by swapping a random selection of half the 1s in the correct adjacency matrix with a random selection of equally many 0s. The tables shows that *pInc* usually outperforms *pInc2*, the zero parameters in the *Hub* case with 50% true edge prior knowledge being the only significant exception.

To study the performance of the different methods on model selection we computed ROC curves, showing the true positive rate (TPR) and false positive rate (FPR) as a function of the threshold on the test statistic (3.8) for inclusion of a parameter in the model. Figure 3.1 shows that in the absence of prior information *pInc2* performs

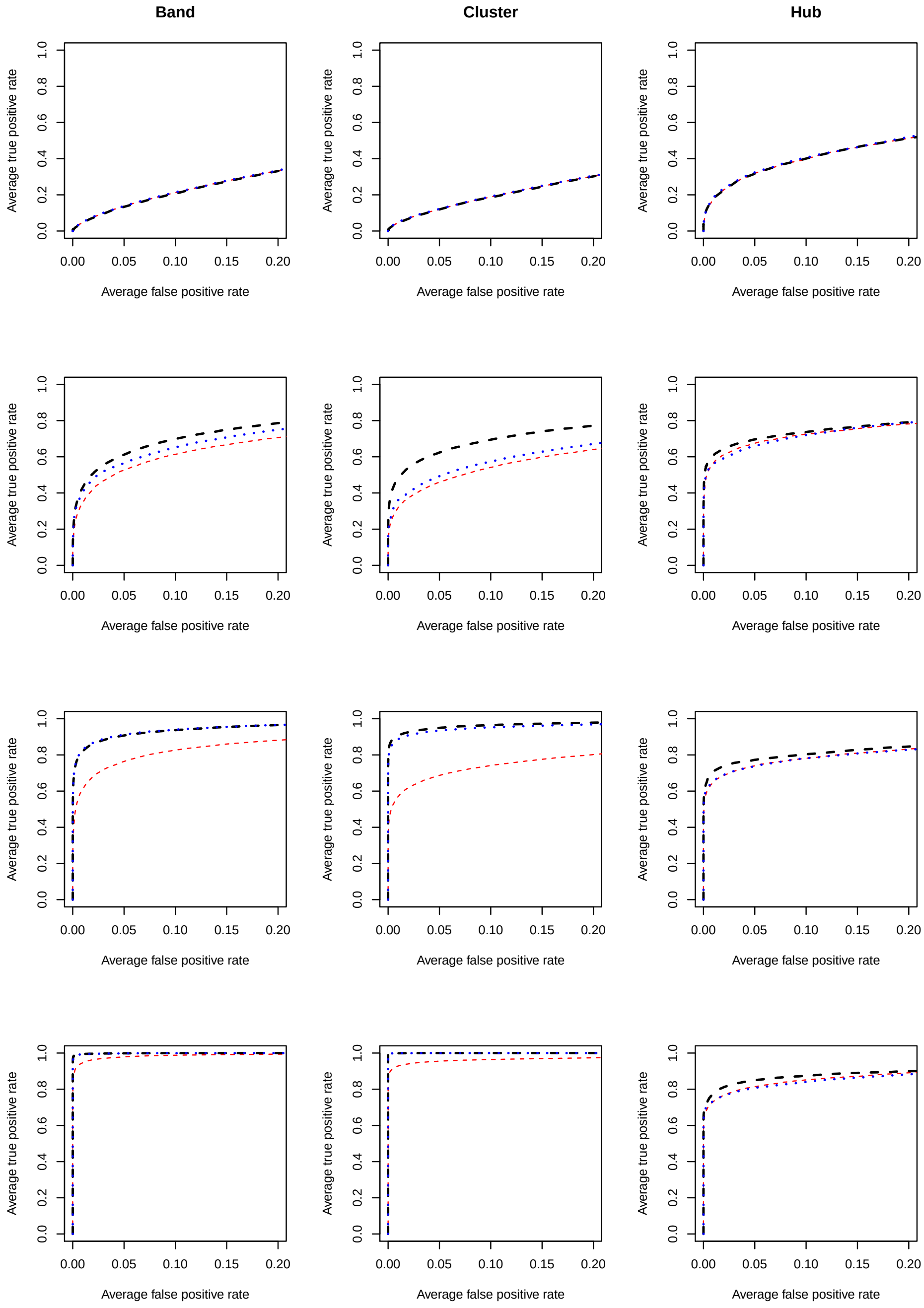


Figure 3.1: *Average partial-ROC curves comparing performance of ShrinkNet (dashed red), pInc2 (dashed black) and pInc (dashed blue) where* $n \in \{10, 100, 200, 500\}$ *and* $p = 100$*. First, second, third and fourth rows correspond respectively to the performances of* $n = 10$*,* $n = 100$*,* $n = 200$ *and* $n = 500$*.*

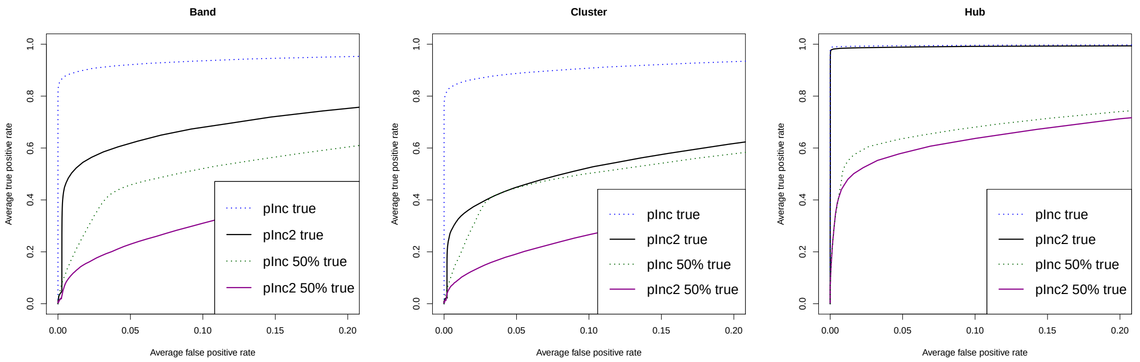


Figure 3.2: *Average partial-ROC curves comparing performance of pInc using perfect prior information (dashed blue), pInc2 using perfect prior information (black), pInc using* 50% *true edge information (dashed dark green) and pInc2 using* 50% *true edge information (darkmagenta). Sample size and network dimension were* $n = 10$ *and* $p = 100$.

best, closely followed by *pInc*, and both methods outperform *ShrinkNet*. Given either correct or 50% correct information *pInc* is the winner, as seen in Figure 3.2, which also shows the usefulness of incorporating prior information. These findings are consistent with the results on estimation presented in Tables 3.1–3.4 in their ordering of *pInc* above *pInc2* in the case of availability of external information.

Figure 3.3 displays histograms of the EB estimates of prior parameter/hyperparameter τ^2's by pInc (TauSq) and pInc2 (TauSq2) across the 50 simulation replicates. The initial hyperparameter value for *pInc2* was set to 0.05. The figure shows that the estimated parameters are bigger (hence less shrinkage) when the sample size is larger. Furthermore, for a fixed sample size the estimates are reasonably stable, the quotient of the largest and smallest across the 50 replicates being below a small constant.

3.5.2 Variational Bayes vs MCMC

We investigated the quality of the variational approximation by comparing it to the output of a long MCMC run. As we only use the univariate marginal posterior distributions of the regression parameters for model selection, we focused on these. We ran a simulation study with a single regression equation (say $i = 1$) with $n = p = 100$, and compared the variational Bayes estimates of the marginal densities with the corresponding MCMC-based estimates. We sampled $n = 100$ independent replicates from a $p = 100$-dimensional normal distribution with mean zero and $(p \times p)$-precision matrix Ω_p, and formed the vector Y_1 and matrix X_1 as indicated in Section 3.2. The precision matrix was chosen to be a *band matrix* with lower and upper bandwidths

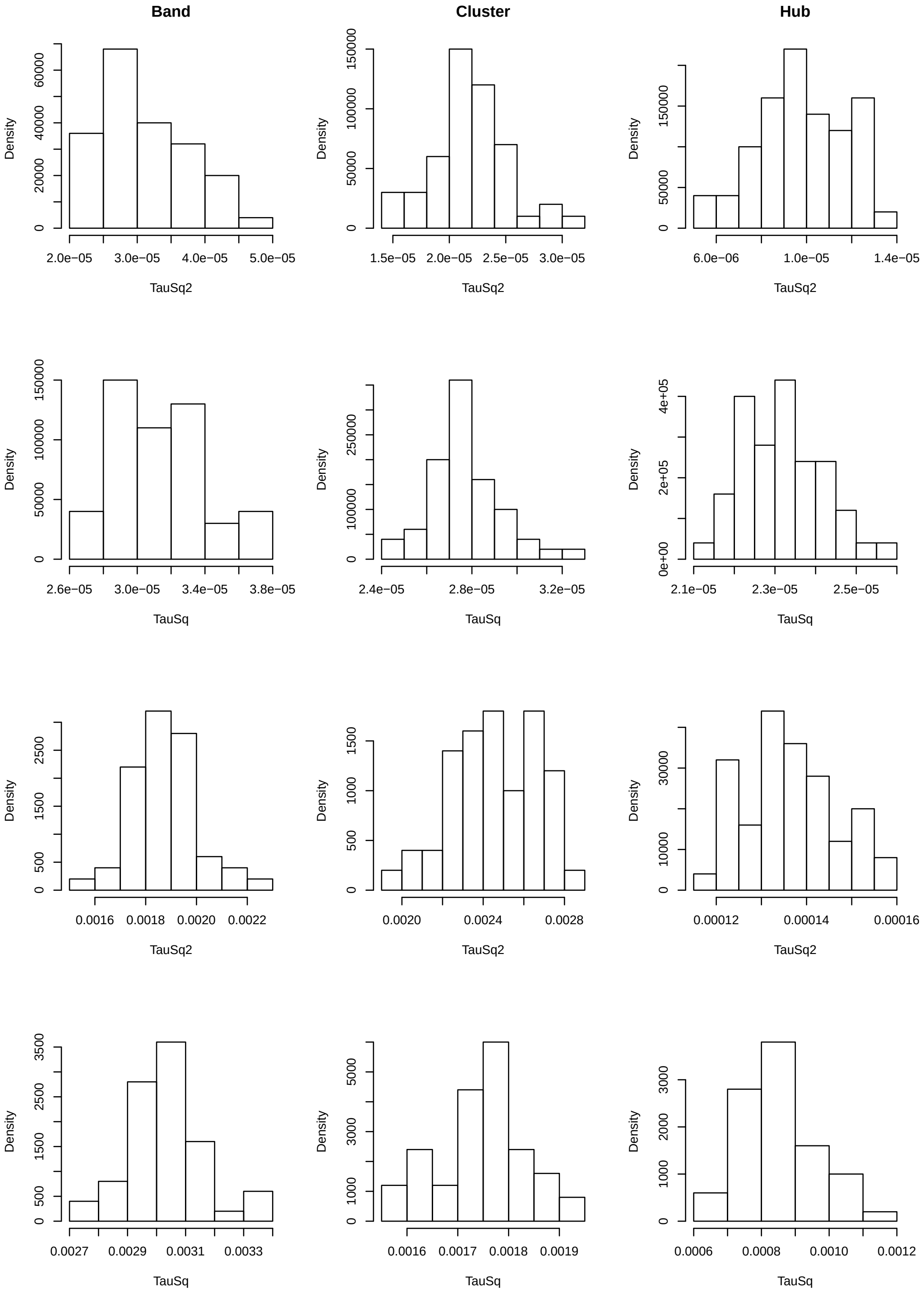


Figure 3.3: *Histograms of the global variance parameter* τ^2 *estimates by EB by pInc (TauSq) and by pInc2 (TauSq2) across* 50 *simulation replicates. First, second and third columns correspond respectively to Band, Cluster and Hub structures for the precision matrix. First row* ($n = 10$) *and third row* ($n = 200$) *display* τ^2 *estimates by pInc2 whereas second row* ($n = 10$) *and fourth row* ($n = 200$) *display* τ^2 *estimates by pInc. We used* $p = 100$.

	average l_1 loss $\|\|\hat{\beta}_1 - \beta_1\|\|_1$ in 20 replications ($i = 1$)	computing time needed for all the 100 regressions
pInc	1.41	58 sec
MCMC method	2.22	13h 15 min

Table 3.5: *Performance comparison between pInc and the MCMC method.*

	average l_1 loss $\|\|\hat{\beta}_1 - \beta_1\|\|_1$ in 20 replications ($i = 1$)	computing time needed for all the 100 regressions
pInc2	2.25	1min 48 sec
MCMC method	3.03	13h 19 min

Table 3.6: *Performance comparison between pInc2 and the MCMC method.*

equal to 4, thus a band of total width 9. For both the variational approximations and the MCMC method we used prior hyperparameters $c = d = 0.001$ and prior hyperparameters $(\hat{a}, \hat{b})$ (resp. $\hat{\tau}^2$ for *pInc2*) fixed to the values set by the *global* empirical Bayes method described in Section 3.4.2. The MCMC iterations were run $nIter = 4 \times 10^4$ times without thinning, after which the first $nBurnin = 2 \times 10^4$ iterates were discarded [111]. Tables 3.5 and 3.6 summarize the comparison.

The correspondence between the two methods is remarkably good. The posterior means obtained from the variational method are even slightly better as estimates of the true parameters than the ones from the MCMC method, in terms of ℓ_1-loss. With respect to computing time the variational method was vastly superior to the MCMC method, which would hardly be feasible even for $n = p = 100$.

3.6 Applications

We applied the methods to two real datasets, both as illustration.

3.6.1 Reconstruction of the apoptosis pathway

The cells of multicellular organisms possess the ability to die by a process called programmed cell death or *apoptosis*, which contributes to maintaining tissue homeostasis. Defects in the apoptosis-inducing pathways can eventually lead to expansion of a population of neoplastic cells and cancer [55, 63, 75]. Resistance to apoptosis may increase the escape of tumour cells from surveillance by the immune system. Since chemotherapy and irradiation act primarily by inducing apoptosis, defects in the apoptotic pathway can make cancer cells resistant to therapy. For this reason resistance to apoptosis remains an important clinical problem.

In this section we illustrate the power of our method in reconstructing the apoptosis network from lung cancer data [76] from the Gene Expression Omnibus (GEO). The data comprises $p = 84$ genes, consisting of $n_1 = 49$ observations from normal tissue and $n_2 = 58$ observations from tumor tissue, hence $n = 107$ observations in total. We fitted *pInc* on the tumor data, using the data on normal tissue as prior knowledge. To the latter aim we fitted *pInc* to the normal data with a single group $G = 1$, and applied the model selection procedure of Section 3.4.3 to create an array P of incidences, which served as input when fitting *pInc* on the tumor data. The idea is that, while tumors and normal tissue may differ strongly in terms of mean gene expression, the gene-gene interaction network may be relatively more stable.

When fitting the *pInc* model with the two groups (gene interaction absent or present in normal tissue), we observed a huge difference in the empirical Bayes estimates of the hyperparameters governing the priors of the parameters τ^{-2} of the two groups, namely prior mean $\hat{a}_0/\hat{b}_0 = 8476.97$ for absent and $\hat{a}_1/\hat{b}_1 = 3.70$ for present in the prior network. This strongly indicates the relevance of the prior knowledge [72], so that superior performance of *pInc* in the reconstruction can be expected.

Figure 3.4 displays the reconstructed undirected network by *pInc*. A total number of 27 edges were found with various edge strengths. The ten most significant edges in decreasing order were: PRKACG ↔ FASLG, MYD88 ↔ CSF2RB, PIK3R2 ↔ CHUK, TNFRSF10B ↔ CHP1, PRKAR1B ↔ AKT2, PIK3R2 ↔ NGF, TRAF2 ↔ BAX, TNF ↔ IL1B, PRKAR2B ↔ AKT3, and TRAF2 ↔ PIK3R2.

Node degrees varied from 0 to 4 with PIK3R2 and PRKAR1A yielding the highest degree 4, followed by TRAF2 having degree 3, and CHUK, CHP1, BIRC3, FAS, IL1B and NFKBIA having each degree 2.

3.6.2 eQTL mapping of the p38MAPK pathway

The p38MAPK pathway is activated *in vivo* by environmental stress and inflammatory cytokines, and plays a key role in the regulation of inflammatory cytokines biosynthesis. Evidence indicates that p38MAPK activity is critical for normal immune and inflammatory response [8, 62, 82]. The pathway also plays an important role in cell differentiation. Its key role in the conversion of myoblasts to differentiated myotubes during myogenic progression has been established by [88, 154, 161]. More recently, *in vivo* studies demonstrated that p38MAPK signalling is a crucial determinant of myogenic differentiation during early embryonic myotome develop-

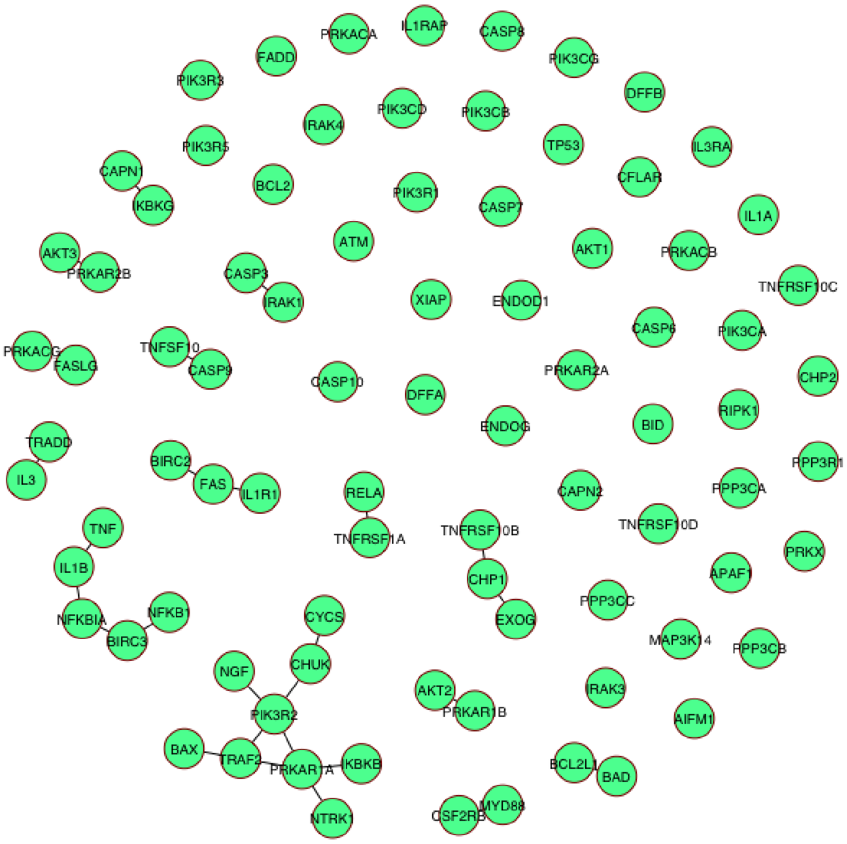


Figure 3.4: *Apoptosis network reconstructed for the 84 genes by pInc.*

ment [32]. Finally, the pathway is involved in chemotactic cell migration [59, 60]. Lack of p38MAPK function may lead to cell cycle deficiency and tumorigenesis, and genetic variants of some genes in the p38MAPK pathway are associated with lung cancer risk [39]. Studying the pathway in healthy cells may enhance understanding the underlying biological mechanism, but has received less attention.

We investigated the association between single nucleotide polymorphisms (SNPs) and the genes in the P38MAPK pathway, using GEUVADIS data. In the GEUVADIS project [77], 462 RNA-Seq samples from lymphoblastoid cell lines were obtained, while the genome sequence of the same individuals is provided by the 1000 Genomes Project.

The samples in this project come from five populations: CEPH (CEU), Finns (FIN), British (GBR), Toscani (TSI) and Yoruba (YRI). In our analysis we excluded the YRI population samples and samples without expression and genotype data, which resulted in a remaining sample size of 373. We also excluded SNPs with minor allele frequency (MAF) $< 5\%$. Using a window of 10^5 bases upstream and 10^5 downstream of every gene, we obtained a total number of 42,054 SNPs for the 99 genes of the pathway belonging to the 22 autosomes. This resulted in a system of 99 regression models, with dimensions varying from 56 to 1169. We scaled (per gene) the gene expression data prior to the computations.

Following Section 3.3 we classified the SNPs connected to each gene as located either within the gene range or outside, and applied *pInc* with two groups ($G = 2$). We observed a big difference in the empirical Bayes estimates of the hyperparameters of the priors of τ^{-2}: mean value $\hat{a}_0/\hat{b}_0 = 27,568.76$ for SNPs outside the gene ranges versus $\hat{a}_1/\hat{b}_1 = 4102.46$ for SNPs inside. The prior information is thus clearly relevant, and hence an improved mapping by *pInc* can be expected.

We found using Selection procedure 3.4.3 (Thresholding) the expression levels of 13 out of the 99 genes (genes 15, 40, 48, 50, 51, 61, 75, 78, 85, 86, 93, 96, 98) to be associated with a total number of 50 SNPs from the 42,054 SNPs under consideration. Gene 50 yielded the highest number 9 of associated SNPs, followed by gene 40 with 6 SNPs and genes 86, 93 and 96 with 5 SNPs each. Figures 3.5 and 3.6 display the estimates of the effect sizes of the SNPs (posterior means $\mathbf{E}_{q^*}(\beta_{i,r}|\, Y_i)$), green for SNPs outside the gene ranges and blue for SNPs within a gene, with 'red stars' indicating the SNPs that were selected. The 6 largest associations were observed within genes 93, 15, 96, 98 and 78 (red vertical lines in Figures 3.5 and 3.6). The active SNPs for all genes, except genes 40 and 50 (although for gene 50 only one of the selected SNPs is not within), are located inside the gene range. This confirms the belief that SNPs falling inside genes are more prone to influence these genes than SNPs outside. The SNP effects on the remainder 86 ($= 99 - 13$) genes are similar to the ones on gene 1 displayed in Figure 3.6. The selection obtained by using *pInc*-DSS is similar.

Comparison of *pInc*-DSS with lasso

From the many dedicated methods for eQTL analysis [16, 71, 83, 86, 125], we chose the lasso as a bench-mark to compare the model selection by *pInc* combined with *DSS* (Section 3.4.3). Our choice for DSS comes from the interest to investigate whether '*pInc* + lasso' indeed outperforms a direct lasso, as suggested for the basic horseshoe. As a criterion we used predictive performance when using a sparse model restricted

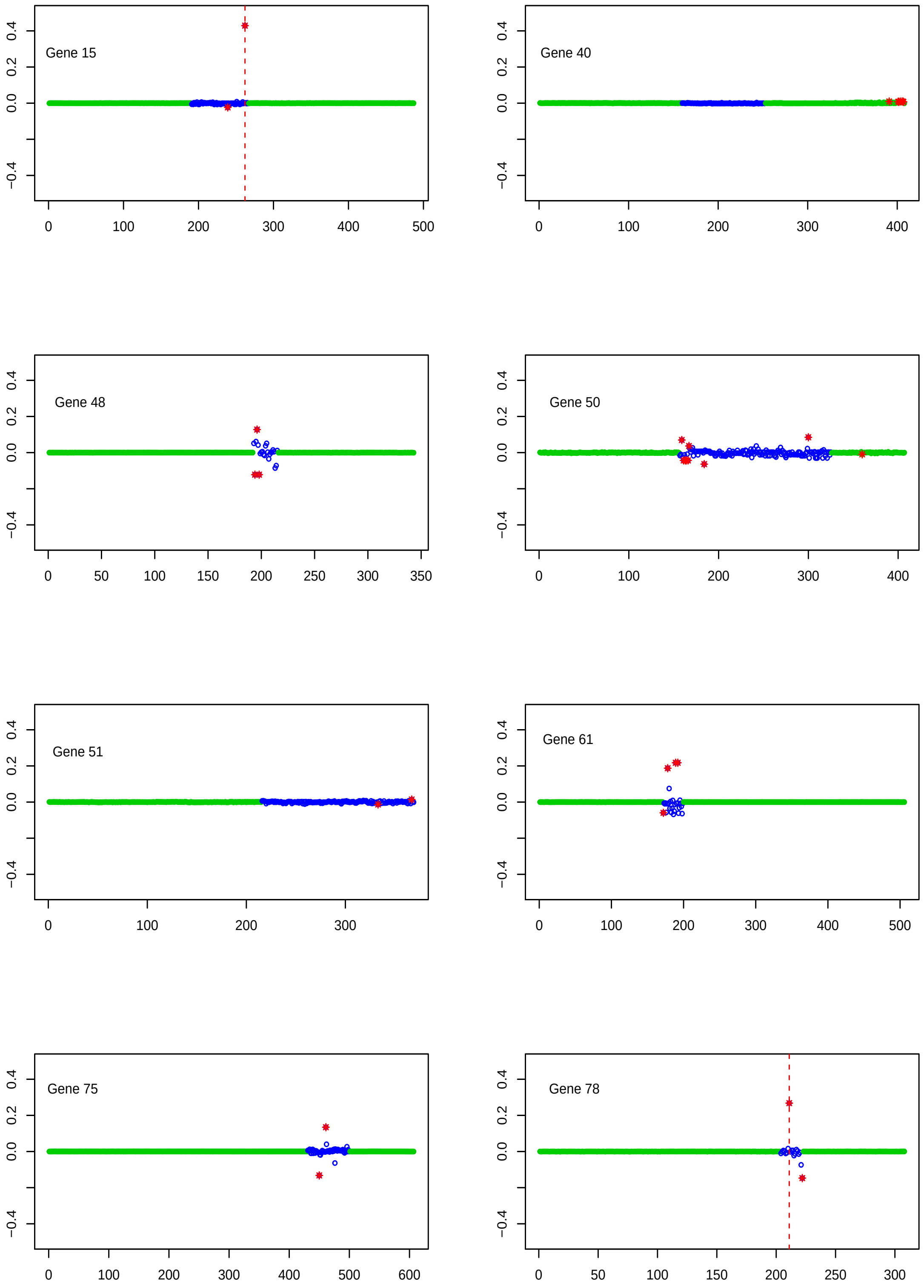


Figure 3.5: *Estimates of SNP effects on genes 15, 40, 48, 50, 51, 61, 75 and 78 using* pInc*. Green dots indicate effects estimates for SNPs outside the gene range and blue dots for SNPs inside the gene range. Red 'stars' indicate selected SNP effects. Dashed vertical lines indicate the 6 largest effects.*

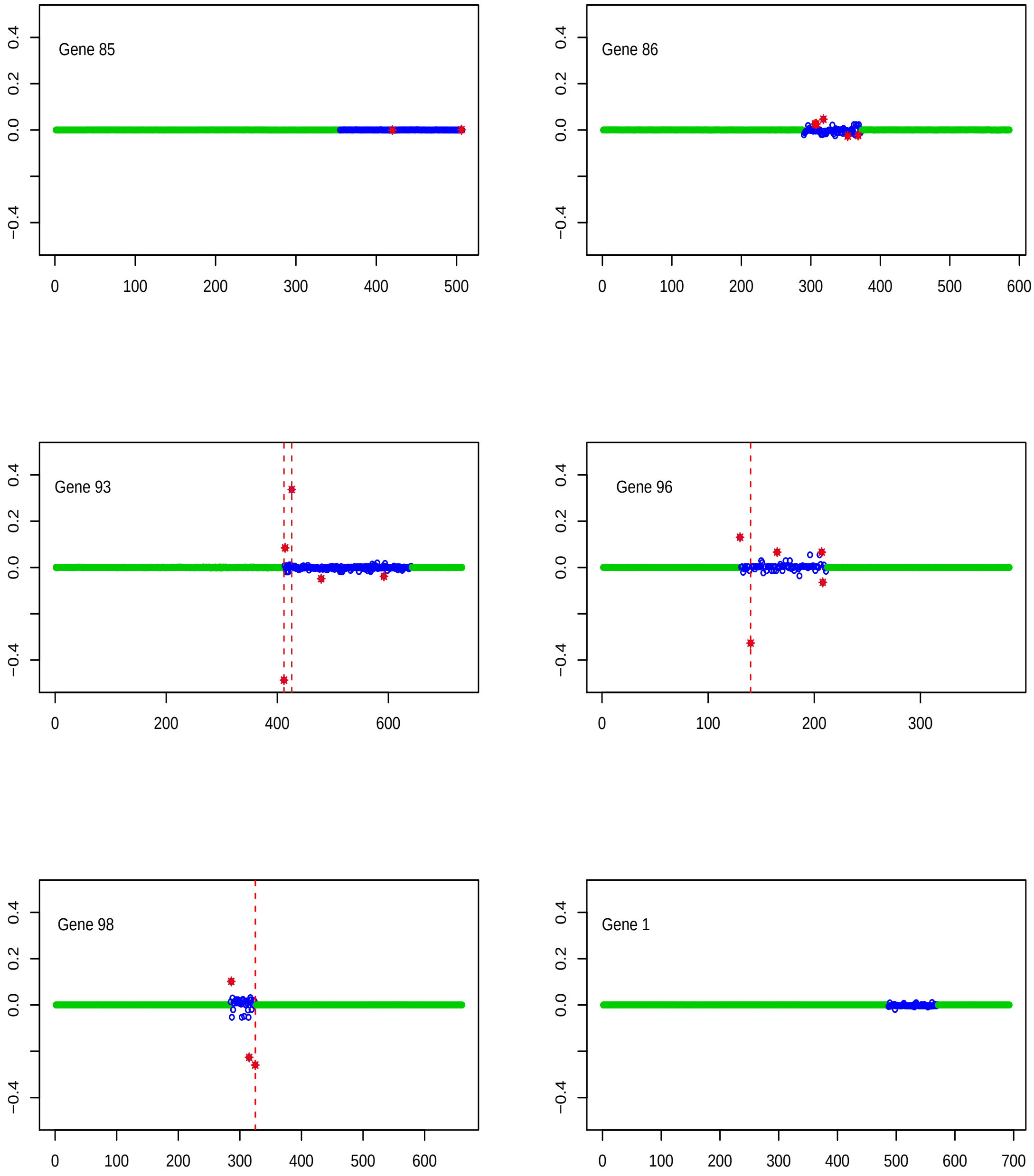


Figure 3.6: *Estimates of SNP effects on genes 75, 78, 85, 86, 93, 96, 98 and 1 using* pInc. *Green dots indicate effects estimates for SNPs outside the gene range and blue dots for SNPs inside the gene range. Red 'stars' indicate selected SNP effects. Dashed vertical lines indicate the 6 largest effects.*

to include a maximal number of predictor variables (SNPs). As for the lasso, the number of selected variables is easy to control by *pInc-DSS*, because the entire trace of the adaptive lasso (3.9) is available. To evaluate predictive performance, we used a single 2/3-1/3 split of the data, leading to training and test sets of 249 and 124 observations, respectively. The lasso was computed using `GLMnet` by [44], also (3.9).

The four panels of Figure 3.7 report the results for the maximal number of predictor variables set equal to 1, 3, 5, or 10. The vertical axis shows the relative reduction of the MSE on the test set as compared to the empty model (all $\beta_i = \mathbf{0}$), defined by

$$\frac{\mathrm{MSE}_0 - \mathrm{MSE}(m_i)}{\mathrm{MSE}_0}, \tag{3.11}$$

where MSE_0 is the MSE of the empty model and $\mathrm{MSE}(m_i)$ the MSE of linear model m_i. This quantity was calculated for all 99 genes in the pathway (horizontal axis), for both the lasso (displayed in black) and *pInc-DSS* (displayed in red), large values indicating accurate prediction. The results of the lasso are somewhat more 'noisy', likely due to less shrinkage of the (near-)zero parameter estimates, and the lasso regularly performs inferior to both the empty model (negative values) and *pInc-DSS*, with gene 13 an extreme case. For genes with considerable signal w.r.t. the empty model (e.g. genes 61, 93 and 98), *pInc-DSS* explains much more of the signal than the lasso. This could be explained by less shrinkage of the non-zero parameters by the horseshoe prior, which is designed to separate zero and nonzero values. This is illustrated in Figure 3.8 for gene 98. Gene 50 is the one exception, where lasso beats *pInc-DSS*, in the case of selecting 3 variables.

3.7 Conclusion

We have introduced a sparse high-dimensional regression approach that can incorporate prior information on the regression parameters and can borrow information across a set of similar datasets. It is based on an empirical Bayesian setup, where external information is incorporated through the prior, and information is borrowed across similar analyses by empirical Bayes estimation of hyperparameters. We have shown the power of the approach both in model-based simulations of Gaussian graphical models and in real data analyses in genomics. Incorporating the information was shown to enhance the analysis, even when the prior information was only partly correct (e.g. 50 % accurate). We explain this by the fact that the empirical Bayesian approach is able to incorporate prior information in a soft manner. Such a flexible

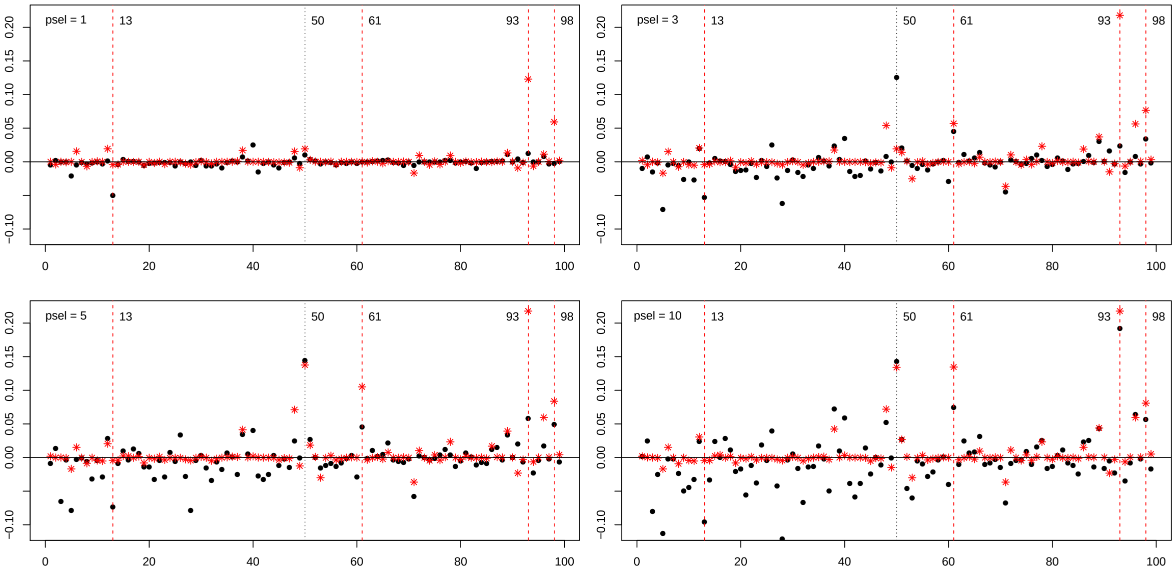


Figure 3.7: *Relative reduction of MSE (y-axis) for the lasso (black dots) and* pInc-DSS *(red stars) for all genes* $i = 1, \ldots, 99$ *(x-axis) when maximal number of variables is fixed to 1, 3, 5, or 10 (top-left, top-right, bottom-left, bottom-right). The genes with the large differences are highlighted by vertical lines*

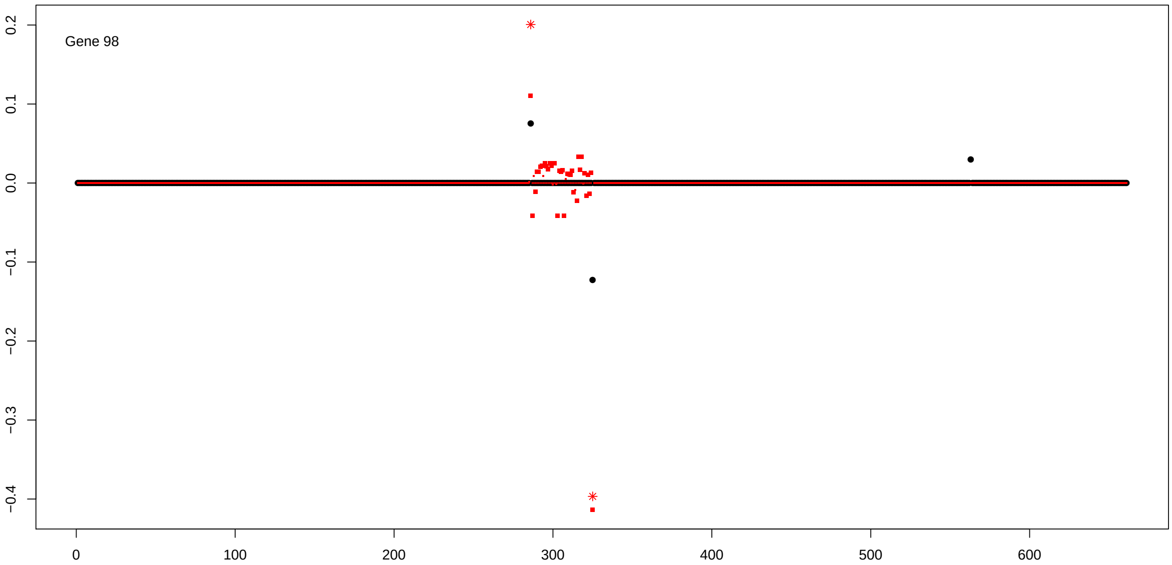


Figure 3.8: *Estimates of SNP effects on gene 98 using* pInc *(red squares), and* pInc-DSS *(red stars) and the lasso (black dots) with 3 predictor variables for the latter two. X-axis denotes SNP index.*

approach is particularly attractive in high-dimensional situations where the amount of data is small relative to the number of parameters and an increasing amount of prior information is available.

To make our approach scalable to large models and/or datasets we developed a variational Bayes approximation to the posterior distribution resulting from the horseshoe prior distribution. We showed the accuracy of the resulting approximation to the marginal posterior distributions of the regression parameters by comparison to state-of-the-art MCMC schemes for the horseshoe prior. The variational Bayes approach obtained the same (if not better) accuracy at a fraction of CPU time.

We studied two versions of the model, one with a gamma prior on the 'sparsity' parameters and one in which these parameters are estimated by the empirical Bayes method. We found that the gamma prior is preferable when relevant prior knowledge can be used, but in the absence of prior knowledge the alternative model may be preferable.

3.8 Appendix

1. Variational Bayes approximation

1.1. Variational marginal densities derivation.

We provide in this section the details of the variational approximation to the posterior distribution for given hyperparameters and for a fixed regression i. Let recall the likelihood and prior densities of the model.

Likelihood:

$$Y_i|X_i,\beta_i,\sigma_i^{-2} \sim \mathrm{N}(X_i\beta_i,\sigma_i^2\mathbf{I}_n).$$

Thus,

$$p(Y_i|X_i,\beta_i,\sigma_i^{-2}) = (2\pi)^{-\frac{n}{2}}(\sigma_i^{-2})^{\frac{n}{2}}\exp\Big(-\frac{1}{2}\sigma_i^{-2}(Y_i-X_i\beta_i)^T(Y_i-X_i\beta_i)\Big)$$

Priors:

$$\begin{aligned}
\epsilon_i|\,\sigma_i^{-2} &\sim \mathrm{N}(0_n,\sigma_i^2\mathbf{I}_n),\\
\beta_i|\,\sigma_i^{-2},\tau_{i,1}^{-2},\dots,\tau_{i,G}^{-2},\lambda_{i,1},\dots,\lambda_{i,s_i} &\sim \mathrm{N}(0_{s_i},\sigma_i^2\mathbf{D}_{\tau_i\lambda_i}),\\
\mathbf{D}_{\tau_i\lambda_i} &= \mathrm{diag}(\tau_{i,P_{i1}}^2\lambda_{i,1}^2,\dots,\tau_{i,P_{is_i}}^2\lambda_{i,s_i}^2),\\
\lambda_{i,t} &\sim C^+(0,1), \qquad t=1,\dots,s_i,\\
\tau_{i,g}^{-2} &\sim \Gamma(a_g,b_g), \qquad g=1,\dots,G,\\
\sigma_i^{-2} &\sim \Gamma(c,d).
\end{aligned}$$

Hence,

$$\begin{aligned}
p(\beta_i|\,\sigma_i^{-2},\tau_{i,1}^{-2},\dots,\tau_{i,G}^{-2},\lambda_{i,1},\dots,\lambda_{i,s_i}) &= (2\pi)^{-\frac{s_i}{2}}\Big(|\sigma_i^2\mathbf{D}_{\tau_i\lambda_i}|\Big)^{-\frac{1}{2}}\\
&\quad\cdot\exp\Big\{-\frac{1}{2}\sigma_i^{-2}\beta_i^T\mathbf{D}_{\tau_i\lambda_i}^{-1}\beta_i\Big\},\\
p(\lambda_{i,t}) &= \frac{2}{\pi(1+\lambda_{i,t}^2)}, \qquad t=1,\dots,s_i,\\
p(\tau_{i,g}^{-2}) &= \frac{b_g^{a_g}}{\Gamma(a_g)}(\tau_{i,g}^{-2})^{a_g-1}\exp\Big\{-b_g\tau_{i,g}^{-2}\Big\},\\
&\qquad g=1,\dots,G,\\
p(\sigma_i^{-2}) &= \frac{d^c}{\Gamma(c)}(\sigma_i^{-2})^{c-1}\exp\Big\{-d\sigma_i^{-2}\Big\}
\end{aligned}$$

We wish to approximate the posterior distribution of the parameter $\theta_i := (\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2})$ given Y_i, for a fixed i by minimizing the Kullback-Leibler (KL) divergence from $q \in \mathcal{Q}$ to the joint posterior $p(\theta_i|\, Y_i)$. Assuming the approximate posterior q factorizes into a product of densities:

$$q(\theta_i) = q_{\beta_i}(\beta_i) \cdot q_{\lambda_i}(\lambda_{i,1}, \cdots, \lambda_{i,s_i}) \cdot q_{\tau_i}(\tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}) \cdot q_{\sigma_i}(\sigma_i^{-2}),$$

the optimal $q^*_{l_r}$, $r = 1, \cdots, 4$; $l_r \in \{\beta_i, \lambda_i, \tau_i, \sigma_i\}$, satisfy [105] (See also Introduction chapter):

$$q^*_{l_r}(.) \propto \exp\left\{\mathbf{E}_{q^*_{-l_r}}\Big[\ln p(Y_i, \theta_i)\Big]\right\}$$

where $\mathbf{E}_{q^*_{-l_r}} = \mathbf{E}_{q^*_{l_1}} \ldots \mathbf{E}_{q^*_{l_{r-1}}} \mathbf{E}_{q^*_{l_{r+1}}} \ldots \mathbf{E}_{q^*_{l_4}}$.

The approximate marginal densities can now be derived. It is:

$$\begin{aligned}
q^*_{\beta_i}(\beta_i) &\propto \exp\left\{\mathbf{E}_{q^*_{-\beta_i}}\Big[\ln p(\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}, Y_i)\Big]\right\} \\
&\propto \exp\left\{\mathbf{E}_{q^*_{-\beta_i}}\Big[\ln p(Y_i|\beta_i, \sigma_i^{-2}) + \ln p(\beta_i|\lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2})\Big]\right\} \\
&\propto \exp\left\{\mathbf{E}_{q^*_{-\beta_i}}\left[-\frac{\sigma_i^{-2}}{2}\Big((Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i) + \beta_i^T \mathbf{D}^{-1}_{\tau_i\lambda_i}\beta_i\Big)\right]\right\} \\
&\propto \exp\left\{-\frac{\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})}{2}\Big[(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i) + \beta_i^T \mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}}\beta_i\Big]\right\} \\
&\propto \exp\left\{-\frac{\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})}{2}\Big[\beta_i^T\Big(X_i^T X_i + \mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}}\Big)\beta_i - 2\beta_i^T X_i^T Y_i\Big]\right\} \\
&\propto \exp\left\{-\frac{\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})}{2}\Big[\Big(\beta_i - \beta_i^*\Big)^T\Big(X_i^T X_i + \mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}}\Big)\Big(\beta_i - \beta_i^*\Big)\Big]\right\}
\end{aligned}$$

where the last line uses the matrix square completion formula

$$u^T A^{-1} u - 2u^T v = (u - Av)^T A^{-1}(u - Av) - v^T A v$$

and

$$\mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}} = \operatorname{diag}\Big(\mathbf{E}_{q^*_{\tau_i}}(\tau_{i,P_{i1}}^{-2})\mathbf{E}_{q^*_{\lambda_{i1}}}(\lambda_{i,1}^{-2}), \ldots, \mathbf{E}_{q^*_{\tau_i}}(\tau_{i,P_{is_i}}^{-2})\mathbf{E}_{q^*_{\lambda_{is_i}}}(\lambda_{i,s_i}^{-2})\Big).$$

Hence, $\beta_i|Y_i \sim \mathrm{N}(\beta_i^*, \Sigma_i^*)$ where

$$\begin{aligned}
\Sigma_i^* &= \left[\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})\Big(X_i^T X_i + \mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}}\Big)\right]^{-1},\\
\beta_i^* &= \Big(X_i^T X_i + \mathbf{D}^{-1}_{\mathbf{E}_{q^*_{\tau_i}\cdot q^*_{\lambda_i}}}\Big)^{-1} X_i^T Y_i.
\end{aligned}$$

$$\begin{aligned}
q^*_{\lambda_{it}}(\lambda_{i,t}) &\propto \exp\Big\{\mathbf{E}_{q^*_{-\lambda_{it}}}\Big[\ln p(\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}, Y_i)\Big]\Big\}\\
&\propto \exp\Big\{\mathbf{E}_{q^*_{-\lambda_{it}}}\Big[\ln p(\beta_i|\, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}) + \ln p(\lambda_{it})\Big]\Big\}\\
&\propto \exp\Big\{\mathbf{E}_{q^*_{-\lambda_{it}}}\Big[\ln p(\beta_i|\, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2})\Big]\Big\} \cdot \ln p(\lambda_{it})\\
&\propto \exp\left\{\mathbf{E}_{q^*_{-\lambda_{it}}}\left[\ln\Big(\prod_{v\neq t}(\lambda_{iv}^{-1})\exp\Big(-\frac{\sigma_i^{-2}}{2}\tau_{i,P_{iv}}^{-2}\beta_{iv}^2\lambda_{iv}^{-2}\Big)\Big)\right]\right\}\\
&\qquad \cdot \exp\left\{\mathbf{E}_{q^*_{-\lambda_{it}}}\left[\ln\Big((\lambda_{it}^{-1})\exp\Big(-\frac{\sigma_i^{-2}}{2}\tau_{i,P_{it}}^{-2}\beta_{it}^2\lambda_{it}^{-2}\Big)\Big)\right]\right\} \cdot \frac{1}{1+\lambda_{it}^2}\\
&\propto \frac{1}{\lambda_{it}\cdot(1+\lambda_{it}^2)} \cdot \exp\Big\{-\frac{1}{2}\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})\mathbf{E}_{q^*_{\tau_i}}(\tau_{i,P_{it}}^{-2})\mathbf{E}_{q^*_{\beta_i}}(\beta_{it}^2)\lambda_{it}^{-2}\Big\}\\
&\propto \frac{1}{\lambda_{it}\cdot(1+\lambda_{it}^2)} \cdot \exp\Big\{-l_{it}\lambda_{it}^{-2}\Big\}
\end{aligned}$$

where,

$$l_{it} = \frac{1}{2}\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})\mathbf{E}_{q^*_{\tau_i}}(\tau_{i,P_{it}}^{-2})\mathbf{E}_{q^*_{\beta_i}}(\beta_{it}^2)$$

Let's denote by K_{it} the normalizing factor for this kernel. It is

$$K_{it} = \int_0^\infty \frac{\exp\{-l_{it}\lambda_{it}^{-2}\}}{\lambda_{it}(1+\lambda_{it}^2)} d\lambda_{it}.$$

Variable transformation $z_{it} := \frac{1}{\lambda_{it}^2}$ and standard integration techniques yield

$$K_{it} = \frac{1}{2}\int_0^\infty \frac{\exp\{-l_{it}z_{it}\}}{1+z_{it}} dz_{it} = \frac{1}{2}\exp(l_{it})E_1(l_{it}), \tag{3.12}$$

where E_1 is the *exponential integral function of order 1*, defined by

$$E_1(x) \equiv \int_x^\infty \frac{e^{-t}}{t} dt, \qquad x \in \mathbb{R}, \quad x > 0.$$

(cf. 3.352(4) of Gradshteyn and Ryzhik (1994) [52]).

Hence, $\lambda_{i,t}| Y_i \sim \Lambda_{\lambda_{it}}, \quad t = 1, \cdots, s_i$ which has density function

$$\begin{aligned}
\Lambda'_{\lambda_{it}}(\lambda_{i,t}) &= \frac{2}{\exp(l_{it})E_1(l_{it}) \cdot \lambda_{it} \cdot (1+\lambda_{it}^2)} \cdot \exp\Big\{ - l_{it}\lambda_{it}^{-2}\Big\} \\
&= \frac{\pi}{\exp(l_{it})E_1(l_{it})} \cdot p(\lambda_{it}) \cdot \frac{1}{\lambda_{it}} \cdot \exp\Big\{ - l_{it}\lambda_{it}^{-2}\Big\}.
\end{aligned}$$

$$\begin{aligned}
q^*_{\tau_{ig}}(\tau_{i,g}^{-2}) &\propto \exp\Big\{\mathbf{E}_{q^*_{-\tau_{ig}}}\Big[\ln p(\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}, Y_i)\Big]\Big\} \\
&\propto \exp\Big\{\mathbf{E}_{q^*_{-\tau_{ig}}}\Big[\ln p(\beta_i| \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}) + \ln p(\tau_{i,g}^{-2})\Big]\Big\} \\
&\propto \exp\Big\{\mathbf{E}_{q^*_{-\tau_{ig}}}\Big[\ln p(\beta_i| \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2})\Big]\Big\} \cdot p(\tau_{i,g}^{-2}) \\
&\propto \Big(\tau_{i,g}^{-2}\Big)^{\frac{s_i^g}{2}} \exp\Big\{ - \frac{1}{2}\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})\mathbf{E}_{q^*_{-\tau_{ig}}}\Big(\beta_i^{gT}\mathbf{D}_{\lambda_i}^{-1}\beta_i^g\Big) \cdot \tau_{i,g}^{-2}\Big\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \cdot (\tau_{i,g}^{-2})^{a_g-1}\exp\{-b_g(\tau_{i,g}^{-2})\} \\
&\propto \Big(\tau_{i,g}^{-2}\Big)^{a_g+\frac{s_i^g}{2}-1} \exp\Big\{ - \Big[b_g + \frac{1}{2}\mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})\mathbf{E}_{q^*_{-\tau_{ig}}}\Big(\beta_i^{gT}\mathbf{D}_{\lambda_i}^{-1}\beta_i^g\Big)\Big] \cdot \tau_{i,g}^{-2}\Big\}
\end{aligned}$$

where s_i^g is the number of g's in the i-row of P encoding the G groups,

$$\mathbf{D}_{\lambda_i} = \mathrm{diag}(\lambda_{i,1}^2, \ldots, \lambda_{i,s_i}^2) \quad \text{and} \quad \beta_i^g = \{\delta_{\{\mathrm{P}_{i,t}=g\}}\beta_{i,t} : t \in \{1, \cdots, s_i\}\}$$

Hence, $\tau_{i,g}^{-2}|Y_i \sim \Gamma(a_{i,g}^*, b_{i,g}^*)$ where

$$\begin{aligned}
a_{i,g}^* &= a_g + 0.5 \cdot \frac{s_i^g}{2}, \\
b_{i,g}^* &= b_g + 0.5 \cdot \mathbf{E}_{q^*_{\sigma_i}}(\sigma_i^{-2})\mathbf{E}_{q^*_{-\tau_{ig}}}\Big(\beta_i^{gT}\mathbf{D}_{\lambda_i}^{-1}\beta_i^g\Big), \qquad g = 1, \cdots, G.
\end{aligned}$$

$$\begin{aligned}
q^*_{\sigma_i}(\sigma_i^{-2}) &\propto \exp\Big\{\mathbf{E}_{q^*_{-\sigma_i}}\Big[\ln p(\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}, Y_i)\Big]\Big\} \\
&\propto \exp\Big\{\mathbf{E}_{q^*_{-\sigma_i}}\Big[\ln p(Y_i|\beta_i, \sigma_i^{-2})\Big]\Big\} \\
&\qquad\qquad \cdot \exp\Big\{\mathbf{E}_{q^*_{-\sigma_i}}\Big[\ln p(\beta_i|\lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2})\Big]\Big\} \cdot p(\sigma_i^{-2}) \\
&\propto \Big(\sigma_i^{-2}\Big)^{\frac{n}{2}} \cdot (\sigma_i^{-2})^{\frac{s_i}{2}} \exp\Big\{ - \frac{\sigma_i^{-2}}{2}\mathbf{E}_{q^*_{\beta_i}}(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)\Big\} \\
&\qquad\qquad \cdot \exp\Big\{ - \frac{\sigma_i^{-2}}{2}\mathbf{E}_{q^*_{-\sigma_i}}\Big(\beta_i^T\mathbf{D}_{\tau_i\lambda_i}^{-1}\beta_i\Big)\Big\} \cdot (\sigma_i^{-2})^{c-1}\exp\Big\{ - d(\sigma_i^{-2})\Big\}
\end{aligned}$$

$$
\begin{aligned}
&\propto (\sigma_i^{-2})^{c+\frac{n}{2}+\frac{s_i}{2}-1} \\
&\cdot \exp\Big\{ -\Big[d + \frac{1}{2}\mathbf{E}_{q^*_{-\sigma_i}}\Big(\beta_i^T \mathbf{D}^{-1}_{\tau_i\lambda_i}\beta_i\Big) + \frac{1}{2}\mathbf{E}_{q^*_{\beta_i}}(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)\Big](\sigma_i^{-2})\Big\}
\end{aligned}
$$

Hence, $\sigma_i^{-2}|Y_i \sim \Gamma(c_i^*, d_i^*)$ where

$$
\begin{aligned}
c_i^* &= c + \frac{n}{2} + \frac{s_i}{2}, \\
d_i^* &= d + 0.5 \cdot \mathbf{E}_{q^*_{-\sigma_i}}\Big(\beta_i^T \mathbf{D}^{-1}_{\tau_i\lambda_i}\beta_i\Big) + 0.5 \cdot \mathbf{E}_{q^*_{\beta_i}}(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\beta_i|\, Y_i &\sim \mathrm{N}\Big(\beta_i^*, \Sigma_i^*\Big), & \\
\lambda_{i,t}|\, Y_i &\sim \Lambda_{\lambda_{it}}, & t = 1, \cdots, s_i, \\
\tau_{i,g}^{-2}|\, Y_i &\sim \Gamma(a_{i,g}^*, b_{i,g}^*), & g = 1, \cdots, G, \\
\sigma_i^{-2}|\, Y_i &\sim \Gamma\Big(c_i^*, d_i^*\Big), &
\end{aligned}
\tag{3.13}
$$

1.2. Variational lower bound.

Let's denote by $\mathcal{L}_i$ the variational lower bound on the log-marginal likelihood. It is

$$
\begin{aligned}
\mathcal{L}_i &= \mathbf{E}_{q^*} \log \frac{p(Y_i, \theta_i)}{q(\theta_i)} \\
&= \mathbf{E}_{q^*} \log p(Y_i|\, \theta_i) + \mathbf{E}_{q^*} \log p(\theta_i) - \mathbf{E}_{q^*} \log q(\theta_i) \\
&= \mathbf{E}_{q^*} \log p(Y_i|\, \beta_i, \sigma_i^{-2}) + \mathbf{E}_{q^*} \log p(\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}) \\
&\qquad\qquad - \mathbf{E}_{q^*} \log q(\beta_i, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}) \\
&= \mathbf{E}_{q^*} \log p(Y_i|\, \beta_i, \sigma_i^{-2}) + \mathbf{E}_{q^*} \log p(\beta_i|\, \lambda_{i,1}, \cdots, \lambda_{i,s_i}, \tau_{i,1}^{-2}, \cdots, \tau_{i,G}^{-2}, \sigma_i^{-2}) \\
&\quad + \sum_{t=1}^{s_i} \mathbf{E}_{q^*} \log p(\lambda_{i,t}) + \sum_{g=1}^{G} \mathbf{E}_{q^*} \log p(\tau_{i,g}^{-2}) + \mathbf{E}_{q^*} \log p(\sigma_i^{-2}) \\
&\quad - \mathbf{E}_{q^*} \log q(\beta_i) - \sum_{t=1}^{s_i} \mathbf{E}_{q^*} \log q(\lambda_{i,t}) - \sum_{g=1}^{G} \mathbf{E}_{q^*} \log q(\tau_{i,g}^{-2}) - \mathbf{E}_{q^*} \log q(\sigma_i^{-2}).
\end{aligned}
$$

The sum elements can be found to satisfy:

$$
\begin{aligned}
\mathbf{E}_{q^*} \log p(Y_i|\, \beta_i, \sigma_i^{-2}) &= -\frac{n}{2}\log(2\pi) + \frac{n}{2}\mathbf{E}_{q^*}\Big[\log(\sigma_i^{-2})\Big] \\
&\qquad - \frac{1}{2}\mathbf{E}_{q^*}(\sigma_i^{-2})\mathbf{E}_{q^*}\Big[(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)\Big],
\end{aligned}
$$

$$\begin{aligned}
&\mathbf{E}_{q^*}\log p(\beta_i|\,\lambda_{i,1},\cdots,\lambda_{i,s_i},\tau_{i,1}^{-2},\cdots,\tau_{i,G}^{-2},\sigma_i^{-2}) = \\
&-\frac{s_i}{2}\log(2\pi)+\frac{s_i}{2}\mathbf{E}_{q^*}\Big[\log(\sigma_i^{-2})\Big]+\sum_{g=1}^{G}\frac{s_i^g}{2}\mathbf{E}_{q^*}\Big[\log(\tau_{i,g}^{-2})\Big]+\sum_{t=1}^{s_i}\mathbf{E}_{q^*}\Big[\log(\lambda_{i,t}^{-1})\Big] \\
&\qquad -\frac{1}{2}\mathbf{E}_{q^*}(\sigma_i^{-2})\mathbf{E}_{q^*}\Big(\beta_i^T\mathbf{D}_{\tau_i\lambda_i}^{-1}\beta_i\Big),
\end{aligned}$$

$$\mathbf{E}_{q^*}\log q(\beta_i) = -\frac{s_i}{2}\log(2\pi)-\frac{1}{2}\log|\Sigma_i^*|-\frac{s_i}{2},$$

$$\begin{aligned}
\mathbf{E}_{q^*}\log q(\lambda_{i,t}) = \log\left[\frac{\pi}{\exp(l_{it})E_1(l_{it})}\right]+\mathbf{E}_{q^*}\Big[\log(\lambda_{i,t}^{-1})\Big]+\mathbf{E}_{q^*}\log p(\lambda_{i,t}) \\
- l_{it}\mathbf{E}_{q^*}(\lambda_{i,t}^{-2}),
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}_{q^*}\log q(\tau_{i,g}^{-2}) = \log\left[\frac{{b_{i,g}^*}^{a_{i,g}^*}}{\Gamma(a_{i,g}^*)}\cdot\frac{\Gamma(a_g)}{{b_g}^{a_g}}\right]+\frac{s_i^g}{2}\mathbf{E}_{q^*}\Big[\log(\tau_{i,g}^{-2})\Big]+\mathbf{E}_{q^*}\log p(\tau_{i,g}^{-2}) \\
-\frac{1}{2}\mathbf{E}_{q^*}(\sigma_i^{-2})\mathbf{E}_{q^*}\Big({\beta_i^g}^T\mathbf{D}_{\lambda_i}^{-1}\beta_i^g\Big)\cdot\mathbf{E}_{q^*}(\tau_{i,g}^{-2}),
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}_{q^*}\log q(\sigma_i^{-2}) = \log\left[\frac{{d_i^*}^{c_i^*}}{\Gamma(c_i^*)}\cdot\frac{\Gamma(c)}{d^c}\right]+\Big(\frac{n}{2}+\frac{s_i}{2}\Big)\mathbf{E}_{q^*}\Big[\log(\sigma_i^{-2})\Big]+\mathbf{E}_{q^*}\log p(\sigma_i^{-2}) \\
-\frac{1}{2}\mathbf{E}_{q^*}\Big(\beta_i^T\mathbf{D}_{\tau_i\lambda_i}^{-1}\beta_i\Big)\mathbf{E}_{q^*}(\sigma_i^{-2})-\frac{1}{2}\mathbf{E}_{q^*}\Big[(Y_i-X_i\beta_i)^T(Y_i-X_i\beta_i)\Big]\mathbf{E}_{q^*}(\sigma_i^{-2}).
\end{aligned}$$

Replacing the sum elements by their respective expression the variational lower bound simplifies to

$$
\begin{aligned}
\mathcal{L}_i = & -\frac{n}{2}\log(2\pi) - s_i\log(\pi) + \frac{1}{2}\log|\Sigma_i^*| + \frac{1}{2}s_i \\
& + \sum_{g=1}^{G}(a_g\log b_g - \log\Gamma(a_g) - a_{i,g}^*\log b_{i,g}^* + \log\Gamma(a_{i,g}^*)) \\
& + c\log d - \log\Gamma(c) - c_i^*\log d_i^* + \log\Gamma(c_i^*) \qquad (3.14)\\
& + \sum_{g=1}^{G}\Big(\frac{1}{2}\mathbf{E}_{q_\sigma^*}(\sigma_i^{-2})\mathbf{E}_{q_\tau^*}(\tau_{i,g}^{-2})\mathbf{E}_{q^*}(\beta_i^{gT}\mathbf{D}_{\lambda_i}^{-1}\beta_i^g)\Big) \\
& + \sum_{t=1}^{s_i}\Big(\log E_1(l_{it}) + \frac{1}{\exp(l_{it})E_1(l_{it})}\Big),
\end{aligned}
$$

where we used the result $\mathbf{E}_{q^*}(\lambda_{i,t}^{-2}) = \frac{1}{l_{it}\cdot\exp(l_{it})\cdot E_1(l_{it})} - 1$ from *Lemma* 1 of the main manuscript.

2. Global empirical Bayes estimation for prior parameters.

We consider the criterion

$$
\begin{aligned}
\alpha = (a_1, b_1, \cdots, a_G, b_G) \mapsto & \sum_{i=1}^{p}\mathbf{E}_q\log\frac{p_\alpha(Y_i,\theta_i)}{q(\theta_i)} \qquad (3.15)\\
= & \sum_{i=1}^{p}\mathbf{E}_q\log\frac{p(Y_i|\,\theta_i)}{q(\theta_i)} + \sum_{i=1}^{p}\mathbf{E}_q\log p_\alpha(\theta_i). \qquad (3.16)
\end{aligned}
$$

For fixed $q = q^*$ the far right side of the preceding display depends on α only through its second term, which is

$$\sum_{i=1}^{p}\mathbf{E}_{q^*}\Big[\log p_\alpha(\tau_{i,1}^{-2}) + \cdots + \log p_\alpha(\tau_{i,G}^{-2})\Big].$$

Since all prior densities are Gamma densities, we find that (a_g, b_g) maximizes, for $g = 1, \cdots, G$,

$$
\begin{aligned}
(a_g, b_g) \mapsto & \sum_{i=1}^{p}\mathbf{E}_{q^*}\Big[(a_g-1)\log\tau_{i,g}^2 - b_g\tau_{i,g}^2 + a_g\log b_g - \log\Gamma(a_g)\Big] \\
= & \sum_{i=1}^{p}\Big[(a_g-1)\Big(\Psi(a_{i,g}^*) - \log b_{i,g}^*\Big) - b_g\frac{a_{i,g}^*}{b_{i,g}^*} + a_g\log b_g - \log\Gamma(a_g)\Big] \\
= & \sum_{i=1}^{p}\Big[(a_g-1)\Big(\Psi(a_{i,g}^*) - \log b_{i,g}^*\Big) - b_g\frac{a_{i,g}^*}{b_{i,g}^*}\Big] + p\Big(a_g\log b_g - \log\Gamma(a_g)\Big)
\end{aligned}
$$

$$= L_g(a_g, b_g).$$

where $\Psi = \Gamma'/\Gamma$ denotes the digamma function and recall $\tau^2_{i,g}$ possesses a $\Gamma(a^*_{i,g}, b^*_{i,g})$-distribution under q^* for $g = 1, \cdots, G$.

Taking the derivative of L_g with respect to b_g yields

$$\frac{\partial L_g}{\partial b_g} = p\frac{a_g}{b_g} - \sum_{i=1}^{p} \frac{a^*_{i,g}}{b^*_{i,g}}$$

and we get by setting this to zero

$$b^*_g = a^*_g \Big(\frac{1}{p}\sum_{i=1}^{p} \frac{a^*_{i,g}}{b^*_{i,g}}\Big)^{-1} = a^*_g M$$

Where $M = p / \sum_{i=1}^{p} \frac{a^*_{i,g}}{b^*_{i,g}}$. Now we get by substituting b_g by b^*_g in L_g

$$\begin{aligned} L_g(a_g, Ma_g) =& \sum_{i=1}^{p} \Big[(a_g - 1)\Big(\Psi(a^*_{i,g}) - \log b^*_{i,g}\Big) - Ma_g \frac{a^*_{i,g}}{b^*_{i,g}}\Big] \\ &+ p\Big(a_g \log(Ma_g) - \log\Gamma(a_g)\Big) \end{aligned}$$

which by differentiating with respect to a_g yields

$$\begin{aligned} \frac{\partial L_g}{\partial a_g} &= p\Big(1 + \log(Ma_g) - \Psi(a_g)\Big) + \sum_{i=1}^{p}\Big[\Big(\Psi(a^*_{i,g}) - \log b^*_{i,g}\Big) - M\frac{a^*_{i,g}}{b^*_{i,g}}\Big] \\ &= p\Big(1 + \log(a_g) + \log M - \Psi(a_g) - 1\Big) + \sum_{i=1}^{p}\Big(\Psi(a^*_{i,g}) - \log b^*_{i,g}\Big) \end{aligned}$$

Setting the derivative to zero, we obtain

$$\log(a^*_g) - \Psi(a^*_g) = \frac{1}{p}\sum_{i=1}^{p}\Big(\log b^*_{i,g} - \Psi(a^*_{i,g})\Big) - \log M.$$

Using the approximation $\log(a^*_g) - \Psi(a^*_g) \approx \frac{1}{2a^*_g}$, we finally find

$$a^*_g \approx \tfrac{1}{2}\Big(\log(a^*_g) - \Psi(a^*_g)\Big)^{-1} = \tfrac{1}{2}\Big(\frac{1}{p}\sum_{i=1}^{p}\Big(\log b^*_{i,g} - \Psi(a^*_{i,g})\Big) - \log M\Big)^{-1}.$$

Chapter 4

Borrow network information between observational and time-course studies: explorations

The transcriptional dynamics of the cell are modelled by a first-order vector autoregression model. It is explored how this model (or aspects thereof) could be learned from observational data, while having time-series data from the same system observed in a possibly different environment available. With existing machinery it is investigated whether incorporation of topological aspects of the dynamical system, as inferred from the time-series data, aid in their reconstruction from the observational one. Subsequently proposed strategies, making assumptions on the relationship between the model parameters of the two enviroments, aim to learn the dynamic parameters from the observational data. Throughout this chapter, cell line time-series and human observational gene expression data from cervical cancer studies are used to illustrate what and how much can be borrowed from the cell lines to enhance knowledge on the transcriptional dynamics in the human environment.

The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

4.1 Introduction

The behaviour and interactions over time of the molecules in a cell are conceived as a dynamical system and modelled as such. A simple but general, stochastic description of a dynamical system is offered by a first-order Vector Auto Regressive model, in short: VAR(1) model. It is used to explain the changes in expression levels of $j = 1, \ldots, p$ genes at time $t+1$, represented by the p-dimensional random vector $\mathbf{Y}_{t+1}$, in terms of a linear combination of the expression of the previous time point $\mathbf{Y}_t$ and an error term:

$$\mathbf{Y}_{t+1} = \mathbf{A}\mathbf{Y}_t + \boldsymbol{\varepsilon}_{t+1}, \tag{4.1}$$

where $\mathbf{A}$ is the matrix with lag-one auto-regression coefficients and $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_\varepsilon)$ for all t. The right-hand side of the VAR(1) model thus comprises an exogeneous part, $\boldsymbol{\varepsilon}_{t+1}$, representing an incoming signal, while an endogeneous part, $\mathbf{A}\mathbf{Y}_t$, tells how this signal is processed by the system. Under stationarity assumption, the covariance matrix of $\mathbf{Y}_t$, denoted $\mathbf{\Sigma}_y$, satisfies the Lyapunov equation: $\mathbf{\Sigma}_y = \mathbf{A}\mathbf{\Sigma}_y\mathbf{A}^\top + \mathbf{\Sigma}_\varepsilon$ and it is readily verified – by substitution – that $\mathbf{\Sigma}_y = \sum_{t=0}^{\infty} \mathbf{A}^t\mathbf{\Sigma}_\varepsilon(\mathbf{A}^t)^\top$.

Two archetypical experiments provide information on the dynamical system of the cell:

- An *observational* study in which multiple independent individuals are profiled at a single, randomly sampled time point. For example, cancer patients generally undergo surgery once at which point the tumor material is recovered and subsequently molecularly interrogated.
- A *time-course* study in which one (or more) individual is tracked over time and profiled at various instances. For instance, a cell line is transfected with an oncogenic agent (e.g., the human papilloma virus) and cultured in a petri dish from which cells are sampled at various time points and molecularly interrogated.

The experiments thus differ by *i)* the environment in which the cells are studied: *in vivo* vs. *in vitro*, and *ii)* the sampling scheme. These differences have implications for what can be learned from each experiment. With data from humans being harder to acquire and more revelant for medical practice, the remainder of this work concentrates on methods that learn the dynamics from the human observational data using the information from the cell line time-course data.

Both study types aim to shed light on the same cellular system and a VAR(1) model is thus assumed for both. Parameters, however, need not be identical due to the different environments of both study types and are – for the moment – to be without restrictions. The data from the observational and time-course studies, $\{\mathbf{Y}_{t,i_h}^{(h)}\}_{t=1,i_h=1}^{T_h,n_h}$ and $\{\mathbf{Y}_{t,i_c}^{(c)}\}_{t=1,i_c=1}^{T_c,n_c}$ with indices h and c referring to the $h$uman and $c$ell line environments, are then described by the following VAR(1) models:

$$\begin{array}{rcll} \mathbf{Y}_{t+1,i_h}^{(h)} & = & \mathbf{A}^{(h)}\mathbf{Y}_{t,i_h}^{(h)} + \boldsymbol{\varepsilon}_{t+1,i_h}^{(h)} & \text{and} \\ \mathbf{Y}_{t+1,i_c}^{(c)} & = & \mathbf{A}^{(c)}\,\mathbf{Y}_{t,i_c}^{(c)} + \boldsymbol{\varepsilon}_{t+1,i_c}^{(c)}, & \end{array}$$

respectively. The $\mathbf{A}^{(h)}$ and $\mathbf{A}^{(c)}$ are the study type specific autoregression parameters. Moreover, the errors are assumed to follow zero-mean Gaussians with possibly different covariances: $\boldsymbol{\varepsilon}_{t+1,i_h}^{(h)} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{(h)})$ and $\boldsymbol{\varepsilon}_{t+1,i_c}^{(c)} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_{\boldsymbol{\varepsilon}}^{(c)})$.

Here we explore what and how can be learned on (aspects of) the ‘human’ VAR(1) model from the observational data. In these explorations cell line time-course data of the same system are assumed available, thus enabling the borrowing of information between the two environments. The remainder of the manuscript is organized as follows. In section 4.2 it is investigated using existing machinary whether information on the conditional independence graph of the system, as inferred from the cell line data, benefits the reconstruction of that graph from human data. An illustration on cervical data compares different methods that serve this end. Section 4.3 proposes strategies to recover information on the dynamic parameter from the human observational data. These strategies make parametric assumptions on the relation among the model parameters of the *in vitro* and *in vivo* environments. Section 4.3.4 investigates the empirical validity of one crucial assumption common to these strategies. Finally, these strategies are applied to the same cervical cancer studies of both environments as before. Analysis with the proposed strategies indicate that *in vitro* cell line studies harbour useful information on the dynamical system of the human cell *in vivo*.

4.1.1 Related work

The statistical literature appears to devote little to no attention to the development of methodology for combining observational and time-course studies. We have found only one work that considers both studies types jointly [146]. This work presents an algorithm referred to as `cMIKANA` (*combined* `MIKANA`) for gene regulatory networks inference which combines steady-state and time-series gene expression data. Their proposed algorithm `cMIKANA` combines two versions of their previous algorithm (`MIKANA`) [61, 126], which reconstructs gene regulatory networks from either steady-state datasets or temporal datasets. Motivated from ordinary differential equations, with either a steady state assumption or a discretized differential operator, the `cMIKANA` algorithm fits a system of nonlinear equations to the both data types simultaneously. However, in their model the exogeneous part (pertubation) is set to zero, thus, not allowing for randomness.

4.2 Shared precision information

First efforts concentrate on the recovery of the conditional independence graph (CIG) of the variates of the system. Such a graph constitutes of nodes, representing the molecules of the system, and edges, corresponding to interactions between these molecules. A graph $\mathcal{G}$ is specified by the pair $(\mathcal{V}, \mathcal{E})$ with node set $\mathcal{V} = \{1, \ldots, p\}$ and edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. A graph is a CIG if and only if the absence/presence of an edge implies conditional in/dependence between the random variables by the nodes un/connected by the edge, given all other random variables. When the vector of variates follows a multivariate Gaussian law, the Inverse Variance lemma [152] implies that two variates are conditionally independent – given all other variates – if the corresponding element in the inverse covariance matrix is zero. The inverse of the marginal variance $\mathbf{\Sigma}_y$ of the VAR(1) process can then be thought of as harbouring the 'global' conditional independencies, that may be determined by the verification of the Wermuth condition from the the temporal and contemporaneous conditional independencies depicted by the time-series chain graph [29].

The information on the CIG of the system is contained in the data from both study types. The normality of the VAR(1) model (4.1) implies:

$$\mathbf{Y}_{t,i_h}^{(h)} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_y^{(h)}) \qquad \text{and} \qquad \mathbf{Y}_{t,i_c}^{(c)} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_y^{(c)}).$$

The VAR(1) model also yields structured – in terms of $\mathbf{A}$ and $\mathbf{\Sigma}_\varepsilon$ – covariance matrices, that harbour the dependencies represented by the time series chain graphs, but this fact is explored in later sections. The inverse covariance matrices, also known as precision matrices and denoted by $\mathbf{\Omega}_y^{(h)}$ and $\mathbf{\Omega}_y^{(c)}$, may be estimated from the data through maximization of the log-likelihood and the pseudo-likelihood (obtained by multiplying the marginal densities of the $\mathbf{Y}_t$):

$$\log(|\mathbf{\Omega}_y^{(h)}|) - \operatorname{tr}(\mathbf{S}_y^{(h)}\mathbf{\Omega}_y^{(h)}) \text{ and } \log(|\mathbf{\Omega}_y^{(c)}|) - \operatorname{tr}(\mathbf{S}_y^{(c)}\mathbf{\Omega}_y^{(c)}),$$

where $\mathbf{S}_y^{(h)}$ and $\mathbf{S}_y^{(c)}$ are the sample covariance matrices from the observational and time-course studies defined by:

$$\mathbf{S}_y^{(h)} = \tfrac{1}{n_h} \sum_{i_h=1}^{n_h} \mathbf{Y}_{t,i_h}^{(h)} [\mathbf{Y}_{t,i_h}^{(h)}]^\top \quad \text{and} \quad \mathbf{S}_y^{(c)} = \tfrac{1}{n_c T} \sum_{i_c=1}^{n_c} \sum_{t=1}^{T} \mathbf{Y}_{t,i_c}^{(c)} [\mathbf{Y}_{t,i_c}^{(c)}]^\top,$$

respectively. The maximization is hampered when the data are high-dimensional. This is usually resolved by log-likelihood augmentation with a penalty. The commonly employed lasso and ridge penalties amount to adding the ℓ_1- and ℓ_2-norm, respectively, of the parameter of interest to the likelihood [43, 139]. Both penalties shrink the elements of the precision matrix towards zero. The lasso penalty may shrink these to exactly zero, thus performing variable selection. The ridge penalization requires a post-hoc step to sparsify its precision matrix estimate, for which amongst others the local FDR procedure of [36] provides. Alternative procedures to reconstruct the conditional independence graph are Bayesian [50, 68], limited information [73, 97], or a combination of the two [33, 80].

4.2.1 Joint precisions estimation

Estimation of multiple precision matrices in a fused penalized manner [10, 30] facilitates the borrowing of information among the various groups/datasets from which these parameters are to be learned. Fusion aims to retain common features of the parameters when the data provide evidence for it, but renounces sharing when the data does not. In contrast to separate *in vitro*- and *in vivo* precision estimation, a fused approach acknowledges possible similarities between environments while not ignoring possible differences and may improve the recovery of the (human) precision matrix by making use of that from the cell line study.

Fused penalized estimation amounts to jointly maximizing of the log- and pseudo-likelihood:

$$\mathcal{L}(\{\mathbf{\Omega}_y^{(h)}, \mathbf{\Omega}_y^{(c)}\}) \quad = \quad n_h[\log(|\mathbf{\Omega}_y^{(h)}|) - \mathrm{tr}(\mathbf{S}_y^{(h)}\mathbf{\Omega}_y^{(h)})] + n_c T[\log(|\mathbf{\Omega}_y^{(c)}|) - \mathrm{tr}(\mathbf{S}_y^{(c)}\mathbf{\Omega}_y^{(c)})]$$

augmented by, e.g. the fused ridge penalty

$$\lambda_2(\|\mathbf{\Omega}_y^{(h)}\|_F^2 + \|\mathbf{\Omega}_y^{(c)}\|_F^2) + \lambda_{2,f}\|\mathbf{\Omega}_y^{(h)} - \mathbf{\Omega}_y^{(c)}\|_F^2.$$

where λ_2, $\lambda_{2,f}$ are nonnegative tuning/penalty parameters. The first tuning parameter λ_2 shrinks the precision matrices towards the null matrix $\mathbf{0}_{pp}$, while large values of $\lambda_{2,f}$ force the estimates $\hat{\mathbf{\Omega}}_y^{(h)}$ and $\hat{\mathbf{\Omega}}_y^{(c)}$ towards each other. The tuning parameters are determined in a data-driven manner, e.g. through cross-validation.

Neither for lasso nor for ridge fused penalization does an explicit expression of the precision estimator exist. However, it is a (strict) convex optimization problem and one may – in line with [10, 30] – use an iterative procedure alternating between the estimation of $\mathbf{\Omega}_y^{(h)}$ and $\mathbf{\Omega}_y^{(c)}$ (given the other) through penalized likelihood maximization. Then, for the ridge case at each step an explicit expression of the estimator exists [10], from which it can be seen how exactly the information is borrowed between the two studies.

4.2.2 Borrowing information from the time-course study

With primary interest in the human CIG, the balanced interest implicit in the fused case may overweigh the cell line information for the reconstruction of the human CIG. Knowledge of the dynamics learned from the *in vitro* study may serve as a prior and enhance the reconstruction of the CIG from the human data. Such an approach is facilitated by the method of [72], also discussed in Chapter 2, and outlined next.

In the work of [72] prior information on a CIG is operationalized simply as an adjacency matrix. Such a matrix is obtained from the *in vitro* CIG, which is inferred by means of the graphical lasso estimator of the inverse covariance matrix $\mathbf{\Omega}_y^{(c)}$ from the cell line data through the maximization of the penalized pseudo-likelihood:

$$\log(|\mathbf{\Omega}_y^{(c)}|) - \operatorname{tr}(\mathbf{S}_y^{(c)}\mathbf{\Omega}_y^{(c)}) - \lambda_1\|\mathbf{\Omega}_y^{(c)}\|_1. \tag{4.2}$$

For each value of the penalty parameter λ_1, this (4.2) yields an estimate $\widehat{\mathbf{\Omega}}(\mathbf{E})$ corresponding to a graph with a given edge set $\mathbf{E}$. To choose between these estimates (i.e. select the parameter λ_1) we use the BIC and (eBIC), ($e$xtended) Bayesian Information Criterion, for reasons of consistency and computational efficiency [24, 42, 46]. For a sample of n i.i.d observations, the eBIC criterion takes the form:

$$\mathrm{BIC}_\gamma(\mathbf{E}) = -2\mathcal{L}_n[\widehat{\mathbf{\Omega}}(\mathbf{E})] + |\mathbf{E}|\log n + 4|\mathbf{E}|\gamma\log p,$$

where $\gamma \in [0, 1]$, $\mathbf{E}$ is the edge set of a candidate graph, $|\mathbf{E}|$ denotes its cardinality, and $\mathcal{L}_n[\widehat{\mathbf{\Omega}}(\mathbf{E})]$ represents the maximized log-likelihood of the associated model. The special case $\gamma = 0$ yields the classical BIC. In this work, we chose the model with the smallest BIC (respectively, eBIC using $\gamma = 0.5$ [42]).

With an adjacency matrix obtained from the cell line data at hand, we invoke the method of [72] to infer the human CIG. The method exploits the equivalence between the Gaussian graphical model and a formulation of this model as system of regression equations (see also Chapter 2). It fits in a Bayesian manner the latter while incorporating prior information on the absent and present edges – that correspond one-to-one with the regression coefficients in the model – obtained from an independent information source. The method has been shown to increase the reconstruction performance when the prior graph is relevant but does not harm if not. This thus invites the reconstruction of the CIG from the time-course data and use that *in vitro* information to learn the graph from the *in vivo* data.

4.2.3 Illustration

Here we investigate, using the aforementioned methods, whether the cell line data provides useful information for the reconstruction of the human CIG of the apoptosis pathway.

The investigation uses data from an observational human and time-course cell line study, [38] and [153] respectively, into cervical cancer. The observational data comprise 43 cervical cancer samples and are publicly available from the Gene Expression Omnibus (GEO) repository under accession number GSE39001. The cell line data, on the other hand, consists of $n = 4$ cell lines interrogated at $T = 8$ time points and are also available via GEO (accession number GSE78279). Both data sets are limited to those genes that *i)* map to the apoptosis pathway according to the KEGG repository and *ii)* and are present in both data sets. This leaves $p = 78$ genes.

The human CIG is now reconstructed from the cervical cancer data described above in the following ways:

- The human and the cell line precision matrices are jointly estimated (as described in Section 4.2.1) by the fused ridge procedure using the R package `rags2ridges` [109]. The associated ridge and fused penalty parameters λ_2 and fused $\lambda_{2,f}$ are chosen through leave-one-out cross-validation (LOOCV). The resulting estimates are denoted by $\widehat{\mathbf{\Omega}}^{(h)}_{joint}$ and $\widehat{\mathbf{\Omega}}^{(c)}_{joint}$ for the *in vivo* and *in vitro* environments, respectively.
- The human CIG is inferred as outlined in Section 4.2.2: by means of the method of [72], using prior information from the cell line data. Hereto first the cell line precision matrix is estimated from the time-course sample covariance matrix $\mathbf{S}^{(c)}_y$ by means of the graphical lasso (4.2), [43] with the penalty parameter chosen on

the basis of the (e)BIC. The resulting cell line precision estimates are denoted by $\widehat{\mathbf{\Omega}}^{(c)}_{bic}$ and $\widehat{\mathbf{\Omega}}^{(c)}_{ebic}$. Because the latter pair of estimates will have edge sets of different sizes, we reduced these edge sets for the clarity of comparison, to the same number 50 or 100, of edges. For this reduction we used the single ranking of all possible edges in the full graph, determined by the full regularization path of the graphical lasso (i.e. when decreasing λ_1 in (4.2) to zero the edges will enter one by one into the model, thus giving their ordering). These resulting estimates form the adjacency matrix that is to be used as prior information in the reconstruction of the human CIG with the method of [72]. The estimated human precision matrices are denoted by $\widehat{\mathbf{\Omega}}^{(h)}_{bic}$ and $\widehat{\mathbf{\Omega}}^{(h)}_{ebic}$.

- For reference the previous ([72]) approach is also used without prior information from the cell line data. This has actually been proposed by [80] (which was expanded by the work of [72]). The resulting human precision estimate of it is denoted by $\widehat{\mathbf{\Omega}}^{(h)}_{bsem}$, with the subscript referring to the name of the method of [80].

Performance of the reconstructed human precision matrices is assessed through reproducibility of edges. Precisely, we split the human data set into two equal and independent subsets. Then both above-mentioned methods are fit to these two subsets separately and the overlapping edge set from the two subsets for each method is reported. The procedure is repeated 100 times. Table 4.1 report the average number of overlapping edges between the two subsets for each method when the total number of edges selected by each method in each subset is fixed to either 50, 100 or 200.

# edges	$\widehat{\mathbf{\Omega}}^{(h)}_{joint}$	$\widehat{\mathbf{\Omega}}^{(h)}_{bsem}$	$\widehat{\mathbf{\Omega}}^{(h)}_{ebic}$	$\widehat{\mathbf{\Omega}}^{(h)}_{bic}$	$\widehat{\mathbf{\Omega}}^{(h)}_{ebic}$	$\widehat{\mathbf{\Omega}}^{(h)}_{bic}$
50	3.8	11.5	19.0	26.7	20.7	32.8
100	9.0	24.9	32.3	37.8	34.1	48.7
200	27.3	55.9	64.6	67.4	66.4	79.5

Table 4.1: *Reproducibility of the apoptosis pathway. The first column contains the number of edges selected in each split, subsequent columns show the (average) number of overlapping edges between two equally-sized splits of the observational human data for various methods (indicated by the subscript of the precision parameter in the top row). The last two pair of columns both use the method of [72] but with fifty and hundred, respectively, nonzero edges adopted from the inferred cell line CIG.*

Table 4.1 shows joint learning of the precision matrices yields the poorest reproducibility, even compared to method of [80] that uses no cell line data at all. When contrasting the reproducibility results of the latter to that of [72], the method of [72] shows clear improvement, which can only be due to the incorporation of the cell line data. This suggests the relevance of the cell line data for the reconstruction of the human CIG.

4.3 Shared parameter information

Rather than studying a derivative of the parameters, i.e. the CIG, efforts may concentrate on the transfer of information on the parameter themselves between the two study types. In this section this amounts to the estimation of the VAR(1) model parameters from the cell line study and using these to learn those of the human observational study. That is, given estimates of $\hat{\mathbf{A}}^{(c)}$ and $\hat{\mathbf{\Sigma}}_{\varepsilon}^{(c)}$, can we – under some assumptions – obtain $\hat{\mathbf{A}}^{(h)}$ and $\hat{\mathbf{\Sigma}}_{\varepsilon}^{(h)}$? The former is briefly outlined after which the latter is elaborated.

The estimation of the *in vitro* VAR(1) model parameters proceeds by maximization of the log-likelihood:

$$\begin{aligned}\mathcal{L}(\{\mathbf{Y}_{t,i_c}^{(c)}\}_{t=1,i_c=1}^{T,n_c}; \mathbf{A}^{(c)}, \mathbf{\Omega}_{\varepsilon}^{(c)}) \quad &\propto \quad n_c(T-1)\log(|\mathbf{\Omega}_{\varepsilon}^{(c)}|) \\ &\quad - \sum_{i_c=1}^{n_c}\sum_{t=2}^{\mathcal{T}} \big(\mathbf{Y}_{*,t,i_c}^{(c)} - \mathbf{A}^{(c)}\mathbf{Y}_{*,t-1,i_c}^{(c)}\big)^{\top}\mathbf{\Omega}_{\varepsilon}^{(c)}\big(\mathbf{Y}_{*,t,i_c}^{(c)} - \mathbf{A}^{(c)}\mathbf{Y}_{*,t-1,i_c}^{(c)}\big).\end{aligned} \tag{4.3}$$

Explicit expression for the estimators can be derived [92, 99]. However, the data may be high-dimensional and the log-likelihood needs to be regularized to ensure well-defined estimators of $\mathbf{A}^{(c)}$ and $\mathbf{\Omega}_{\varepsilon}^{(c)}$. Abegaz and Wit, (2013) [1] and Miok et al (2017) [99] present lasso and ridge penalized maximum likelihood estimation, respectively, of the VAR(1) model. In the remainder we resort to the latter. Miok et al [99] reported that its performance is (slightly) better in terms of loss and on a par with respect to selection than its lasso counterpart [1]. But practically, the implementation of the ridge approach can handle higher dimensional data sets where the lasso implementation fails to converge (cf. the `ragt2ridges` and `sparseTSCGM`-packages, respectively). The ridge penalty that augments the log-likelihood (4.3) is:

$$P(\mathbf{A}^{(c)}, \mathbf{\Omega}_{\varepsilon}^{(c)}, \lambda_a, \lambda_\omega) = -\frac{1}{2}n_c(T-1)\lambda_a \mathrm{tr}[\mathbf{A}^{(c)\top}\mathbf{A}^{(c)}] - \frac{1}{2}n_c(T-1)\lambda_\omega \mathrm{tr}[\mathbf{\Omega}_{\varepsilon}^{(c)\top}\mathbf{\Omega}_{\varepsilon}^{(c)}]. \tag{4.4}$$

Analytic expressions for the ridge estimator of the parameters can be found in [99].

4.3.1 Symmetric or triangular $\mathbf{A}^{(h)}$ and known $\mathbf{\Sigma}_{\varepsilon}^{(h)}$

A first approach to learn the human VAR(1) parameters assumes, perhaps rather boldly, $\mathbf{A}^{(h)}$ to be symmetric but unknown while $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ is assumed known, possibly obtained from the estimate of $\mathbf{\Sigma}_{\varepsilon}^{(c)}$. With $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ known and $\mathbf{A}^{(h)}$ symmetric the $\frac{1}{2}p(p+1)$ parameters of $\mathbf{A}^{(h)}$ need to be estimated from the Lyapunov equation. When $\mathbf{\Sigma}_{y}^{(h)}$ has been estimated from the data of the observational study, this is feasible under the assumption of stability, as the Lyapunov equation then contains exactly $\frac{1}{2}p(p+1)$ degrees of freedom.

If both $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ and $\mathbf{\Sigma}_{y}^{(h)}$ (or a – penalized – estimate thereof) are known and positive definite, there exists a unique solution $\mathbf{A}^{(h)}$ to the Lyapunov equation. The Lyapunov equation then reduces to: $\mathbf{\Sigma}_{y}^{(h)} = \mathbf{A}^{(h)}\mathbf{\Sigma}_{y}^{(h)}\mathbf{A}^{(h)} + \mathbf{\Sigma}_{\varepsilon}^{(h)}$. Post-multiplication of this equation by $\mathbf{\Sigma}_{y}^{(h)}$ yields: $(\mathbf{\Sigma}_{y}^{(h)} - \mathbf{\Sigma}_{\varepsilon}^{(h)})\mathbf{\Sigma}_{y}^{(h)} = (\mathbf{A}^{(h)}\mathbf{\Sigma}_{y}^{(h)})^2$. Solving for $\mathbf{A}^{(h)}$ gives:

$$\mathbf{A}^{(h)} \;=\; [(\mathbf{\Sigma}_{y}^{(h)} - \mathbf{\Sigma}_{\varepsilon}^{(h)})\mathbf{\Sigma}_{y}^{(h)}]^{1/2}(\mathbf{\Sigma}_{y}^{(h)})^{-1} \;=\; (\mathbf{\Sigma}_{y}^{(h)})^{-1}[\mathbf{\Sigma}_{y}^{(h)}(\mathbf{\Sigma}_{y}^{(h)} - \mathbf{\Sigma}_{\varepsilon}^{(h)})]^{1/2}.$$

Of course, $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ is generally unknown and the symmetry of $\mathbf{A}^{(h)}$ is not very plausible.

The symmetry assumption on $\mathbf{A}^{(h)}$ is effectively one to guarantee identifiability. Identifiability may be acheived in other ways. For instance, a lower triangular $\mathbf{A}^{(h)}$ also has $\frac{1}{2}p(p+1)$ parameters. Consider the Cholesky decompositions of $\mathbf{\Sigma}_{y}^{(h)} = \mathbf{L}_{\sigma_y}\mathbf{L}_{\sigma_y}^{\top}$ and $\mathbf{\Sigma}_{y}^{(h)} - \mathbf{\Sigma}_{\varepsilon}^{(h)} = \mathbf{L}_{\sigma_{y,\varepsilon}}\mathbf{L}_{\sigma_{y,\varepsilon}}^{\top}$. Substitute this in the Lyapunov equation and obtain:

$$\mathbf{A} = \mathbf{L}_{\sigma_{y,\varepsilon}}\mathbf{L}_{\sigma_y}^{-1},$$

where we have used that *i)* the product of two lower triangular matrix is itself a lower triangular matrix and *ii)* the inverse of a lower triangular matrix is one too. Biologically, a lower triangular $\mathbf{A}^{(h)}$ may be plausible as it could represent a signalling pathway. It would, however, in addition to knowledge of $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ require knowledge of the ordering of the variates to form this particular support.

4.3.2 $\mathbf{A}^{(h)} = \delta\mathbf{A}^{(c)}$ and a diagonal $\mathbf{\Sigma}_{\varepsilon}^{(h)}$

Slightly more realistically one may assume that the endogeneous part of the system is largely preserved across environments, but that the exogeneous part differs considerably. This is operationalized as

- *i)* $\mathbf{A}^{(h)} = \delta\mathbf{A}^{(c)}$, i.e. the endogeneous part differs only by a scalar between environments, and

- *ii)* an *in vivo* diagonal error covariance $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ being unrelated to its *in vitro* counterpart $\mathbf{\Sigma}_{\varepsilon}^{(c)}$.

The parameters of the *in vivo* VAR(1) model, $\mathbf{A}^{(h)}$ and $\mathbf{\Sigma}_{\varepsilon}^{(h)}$, are learned from the Lyapunov equation, in which the aforementioned assumptions on the parameters are used together with a – penalized – estimate of $\mathbf{\Sigma}_{y}^{(h)}$ obtained from the observational data (as outlined before in Section 4.2). Under stability the parameters $\mathbf{A}^{(h)}$ and $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ should satisfy the Lyapunov equation:

$$\widehat{\mathbf{\Sigma}}_{y}^{(h)} = \mathbf{A}^{(h)} \widehat{\mathbf{\Sigma}}_{y}^{(h)} (\mathbf{A}^{(h)})^{\top} + \mathbf{\Sigma}_{\varepsilon}^{(h)}, \tag{4.5}$$

where the estimate of $\mathbf{\Sigma}_{y}^{(h)}$ has been substituted. It now seems reasonable to choose $\mathbf{A}^{(h)}$ and $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ such that this equation is fulfilled best. This invites the definition of the following loss criterion for the estimation of the parameter (temporarily refraining from the substitution of the assumption $\mathbf{A}^{(h)} = \delta \mathbf{A}^{(c)}$ as the right-hand side is revisited in the next section with a different assumption on $\mathbf{A}^{(h)}$):

$$(\delta, \mathbf{\Sigma}_{\varepsilon}^{(h)}) \mapsto \| \widehat{\mathbf{\Sigma}}_{y}^{(h)} - \mathbf{A}^{(h)} \widehat{\mathbf{\Sigma}}_{y}^{(h)} (\mathbf{A}^{(h)})^{\top} - \mathbf{\Sigma}_{\varepsilon}^{(h)} \|_{F}^{2},$$

where $\|.\|_F$ denotes the Frobenius norm. Using the diagonality of $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ and the element-wise formulation of the Lyapunov equation

$$[\widehat{\mathbf{\Sigma}}_{y}^{(h)}]_{j_1,j_2} = (\mathbf{A}^{(h)})_{j_1,*} \widehat{\mathbf{\Sigma}}_{y}^{(h)} [(\mathbf{A}^{(h)})]_{*,j_2}^{\top} + [\mathbf{\Sigma}_{\varepsilon}^{(h)}]_{j_1,j_2},$$

the loss criterion can be written as:

$$\begin{aligned} & \sum_{j=1}^{p} \Big[(\widehat{\mathbf{\Sigma}}_{y}^{(h)})_{jj} - (\mathbf{A}^{(h)})_{j,*} \widehat{\mathbf{\Sigma}}_{y}^{(h)} [(\mathbf{A}^{(h)})^{\top}]_{*,j} - (\mathbf{\Sigma}_{\varepsilon}^{(h)})_{jj} \Big]^2 \\ & + 2 \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \Big[(\widehat{\mathbf{\Sigma}}_{y}^{(h)})_{j_1,j_2} - (\mathbf{A}^{(h)})_{j_1,*} \widehat{\mathbf{\Sigma}}_{y}^{(h)} [(\mathbf{A}^{(h)})^{\top}]_{*,j_2} \Big]^2. \end{aligned} \tag{4.6}$$

Arrive at the estimating equations through substitution of the assumptions $\mathbf{A}^{(h)} = \delta \mathbf{A}^{(c)}$ and a diagonal $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ and equation of the derivatives w.r.t δ^2 and the diagonal elements of $\mathbf{\Sigma}_{\varepsilon}^{(h)}$, respectively, to zero:

$$0 \;=\; 4 \sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} (\mathbf{A}^{(c)})_{j_1,*} \widehat{\mathbf{\Sigma}}_{y}^{(h)} [(\mathbf{A}^{(c)})^{\top}]_{*,j_2} \Big[(\widehat{\mathbf{\Sigma}}_{y}^{(h)})_{j_1,j_2} - \delta^2 (\mathbf{A}^{(c)})_{j_1,*} \widehat{\mathbf{\Sigma}}_{y}^{(h)} [(\mathbf{A}^{(c)})^{\top}]_{*,j_2} \Big],$$

$$\begin{array}{rcl} & & + \, 2 \displaystyle\sum_{j=1}^{p} (\mathbf{A}^{(c)})_{j,*} \widehat{\boldsymbol{\Sigma}}_y^{(h)} [(\mathbf{A}^{(c)})^{\top}]_{*,j} \Big[(\widehat{\boldsymbol{\Sigma}}_y^{(h)})_{jj} - \delta^2 (\mathbf{A}^{(c)})_{j,*} \widehat{\boldsymbol{\Sigma}}_y^{(h)} [(\mathbf{A}^{(c)})^{\top}]_{*,j} - [\boldsymbol{\Sigma}_{\varepsilon}^{(h)}]_{jj} \Big] \\ 0 & = & (\widehat{\boldsymbol{\Sigma}}_y^{(h)})_{jj} - \delta^2 (\mathbf{A}^{(c)})_{j,*} \widehat{\boldsymbol{\Sigma}}_y^{(h)} [(\mathbf{A}^{(c)})^{\top}]_{*,j} - (\boldsymbol{\Sigma}_{\varepsilon}^{(h)})_{jj} \qquad \text{for } j = 1, \cdots, p. \end{array}$$

Solve these equations for δ^2 and the diagonal elements of $\boldsymbol{\Sigma}_{\varepsilon}^{(h)}$ and obtain their estimates. Hereto isolate $\boldsymbol{\Sigma}_{\varepsilon}^{(h)}$ from the second equation of the preceeding display and substitute it in the first, from which δ^2 is then easily solveable and can be used to obtain an estimator of $\boldsymbol{\Sigma}_{\varepsilon}^{(h)}$. This gives:

$$\begin{array}{rcl} \hat{\delta} & = & \pm \sqrt{\dfrac{\sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} (\mathbf{A}^{(c)})_{j_1,*} \widehat{\boldsymbol{\Sigma}}_y^{(h)} [(\mathbf{A}^{(c)})^{\top}]_{*,j_2} (\widehat{\boldsymbol{\Sigma}}_y^{(h)})_{j_1,j_2}}{\sum_{j_1=1}^{p-1} \sum_{j_2=j_1+1}^{p} \Big[(\mathbf{A}^{(c)})_{j_1,*} \widehat{\boldsymbol{\Sigma}}_y^{(h)} [(\mathbf{A}^{(c)})^{\top}]_{*,j_2} \Big]^2}}, \\ (\widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{(h)})_{jj} & = & (\widehat{\boldsymbol{\Sigma}}_y^{(h)})_{jj} - \hat{\delta}^2 (\mathbf{A}^{(c)})_{j,*} \widehat{\boldsymbol{\Sigma}}_y^{(h)} ([\mathbf{A}^{(c)}]^{\top})_{*,j}, \end{array}$$

There are two estimators of δ. In practice, we suggest to consider the positive one. Heuristically, that is the only sensible one, as the negative one refers to the case where all *in vivo* relationships have a sign opposite of their *in vitro* counterparts. Furthermore, the numerator inside the square root of the estimator of δ may not be positive. A real solution then does not exists. We then suggest to set $\hat{\delta} = 0$. This indicates that $\mathbf{A}^{(h)}$ cannot be learned from the data. Or, in other words, that $\mathbf{A}^{(c)}$ does not provide (proportional) information on $\mathbf{A}^{(h)}$.

The estimator of δ^2 can loosely be interpreted as a regression-type estimator. The analogy starts with the loss criterion which can be rewritten as the following sum-of-squares criterion:

$$\sum_{j_1,j_2=1}^{p} \Big\{ (\widehat{\boldsymbol{\Sigma}}_y^{(h)} - \widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{(h)}))_{j_1,j_2} - \delta^2 [\mathbf{A}^{(c)} \widehat{\boldsymbol{\Sigma}}_y^{(h)} (\mathbf{A}^{(c)})^{\top}]_{j_1,j_2} \Big\}^2.$$

It resembles the loss criterion of the linear regression model, in the sense that the elements of the estimates of the marginal variance corrected for the error, $\widehat{\boldsymbol{\Sigma}}_y^{(h)} - \widehat{\boldsymbol{\Sigma}}_{\varepsilon}^{(h)}$, are regressed on the $\mathbf{A}^{(c)} \widehat{\boldsymbol{\Sigma}}_y^{(h)} (\mathbf{A}^{(c)})^{\top}$ – under the proportionality assumption. From this perspective the estimator of δ^2 can then indeed be seen as a ratio of a 'covariance' and 'variance'.

4.3.3 Sparse $\mathbf{A}^{(h)}$, shared support and a diagonal $\boldsymbol{\Sigma}_{\varepsilon}^{(h)}$

Communality of the human and cell line regulatory systems may also be assumed at the level of their topology. For instance, gene A regulates gene B in the cell line *in vitro* system if and only if it does so in the human *in vivo* system. Information on the regulatory relationships (as captured by the autoregression parameter $\mathbf{A}$) can be learned from the cell line data and not from the human data. Consequently, the remainder of this section takes the topology inferred from the former as a template for the latter. This is then used when learning $\mathbf{A}^{(h)}$ from the human data.

A common regulatory topology implies a shared support between $\mathbf{A}^{(c)}$ and $\mathbf{A}^{(h)}$. The support of $\mathbf{A}^{(c)}$ is assumed known. Moreover, its support is assumed to be sparse, a requirement to ensure identifiability (see end of this subsection) of the nonzero elements of $\mathbf{A}^{(h)}$. The knowledge of $\mathbf{A}^{(c)}$'s support carries over to $\mathbf{A}^{(h)}$ and used as parameter constraint in its estimation via the minimization of loss criterion (4.6). The problem of the estimation of $\mathbf{A}^{(h)}$ and a diagonal $\mathbf{\Sigma}_{\varepsilon}^{(h)}$ under the above formulated assumption is thus operationalized as:

$$\min_{\mathbf{A}^{(h)}, \mathbf{\Sigma}_{\varepsilon}{}^{(h)}} \|\widehat{\mathbf{\Sigma}}_{y}^{(h)} - \mathbf{A}^{(h)}\widehat{\mathbf{\Sigma}}_{y}^{(h)}(\mathbf{A}^{(h)})^{\top} - \mathbf{\Sigma}_{\varepsilon}^{(h)}\|_F^2$$

$$\text{s.t. } \operatorname{supp}(\mathbf{A}^{(h)}) = \operatorname{supp}(\mathbf{A}^{(c)}), \ \mathbf{\Sigma}_{\varepsilon}^{(h)} \text{ diagonal.}$$

Define $\mathbf{U} := \widehat{\mathbf{\Sigma}}_{y}^{(h)} - \mathbf{A}^{(h)}\widehat{\mathbf{\Sigma}}_{y}^{(h)}(\mathbf{A}^{(h)})^{\top} - \mathbf{\Sigma}_{\varepsilon}^{(h)}$ and write $g(\mathbf{U}) = \|\mathbf{U}\|_F^2$. To avoid notational clutter in the derivation of the estimator the (h)-superscript is dropped in the remainder of this subsection. We thus write, e.g., $\mathbf{U} := \widehat{\mathbf{\Sigma}}_y - \mathbf{A}\widehat{\mathbf{\Sigma}}_y\mathbf{A}^{\top} - \mathbf{\Sigma}_{\varepsilon}$.

The constrained optimization problem above can be solved by optimization of the objective with respect to only the nonzeros elements of $(\mathbf{A}, \mathbf{\Sigma}_{\varepsilon})$. In particular, using the local convexity of the loss function the nonzero parameters of $\mathbf{A}$ are optimized one-at-the-time, while keeping the other (temporarily) fixed. In fact, for any fixed, nonzero element we have a (at least local) convex optimization problem. The minimizers will, thus, be all updated at each iteration and this till convergence.

At each step the derivative of $g(\mathbf{U})$ with respect to a nonzero parameter of $\mathbf{A}$ or $\mathbf{\Sigma}_{\varepsilon}$ is equated to zero. The root of the resulting equation is the updated parameter estimate. The derivative of $g(\mathbf{U})$ with respect to the (i,j)-th element of $\mathbf{A}$ can be found to be:

$$\frac{\partial g(\mathbf{U})}{\partial a_{ij}} = \operatorname{tr}\Big[\Big(\frac{\partial g(\mathbf{U})}{\partial \mathbf{U}}\Big)^{\top} \frac{\partial \mathbf{U}}{\partial a_{ij}}\Big] \quad = \quad -4\mathbf{U}_{i,*}(\mathbf{A}\widehat{\mathbf{\Sigma}}_y)_{*,j}$$

where $\mathbf{X}_{i,*}$ and $\mathbf{X}_{*,j}$ represents the i-th row (resp. the j-the column) of the matrix $\mathbf{X}$, respectively. Substitute $\mathbf{U}$ by its expression in the derivative above and, after a little rewritting, obtain:

$$\frac{\partial g(\mathbf{U})}{\partial a_{ij}} = 4[(\mathbf{A}\widehat{\mathbf{\Sigma}}_y\mathbf{A}^{\top})_{i,*}(\mathbf{A}\widehat{\mathbf{\Sigma}}_y)_{*,j} - (\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_{\varepsilon})_{i,*}(\mathbf{A}\widehat{\mathbf{\Sigma}})_{*,j}].$$

This derivative simplifies (cf. Appendix I) to:

$$\frac{\partial g(\mathbf{U})}{\partial a_{ij}} = \mathbf{K}_3 a_{ij}^3 + \mathbf{K}_2 a_{ij}^2 + \mathbf{K}_1 a_{ij} + \mathbf{K}_0, \tag{4.7}$$

where the $\mathbf{K}_0$, $\mathbf{K}_1$, $\mathbf{K}_2$, $\mathbf{K}_3$ are known expression in terms of the other (temporarily fixed) parameters of $\mathbf{A}$ and $\mathbf{\Sigma}_\varepsilon$ (see Appendix I). Since we are interested in real roots, we can have one or three real roots by setting (4.7) to zero which respectively corresponds to one global minimum of g or two local minima of g. In the latter case we suggest to evaluate the objective function g on the three real roots in order to find the overall minimum. Similarly, the derivative of $g(\mathbf{U})$ with respect to the elements of $\mathbf{\Sigma}_\varepsilon := \text{diag}(\sigma_{11}^2, \cdots, \sigma_{pp}^2)$ can be found to be:

$$\frac{\partial g(\mathbf{U})}{\partial \sigma_{ii}^2} = \text{tr}\Big[\Big(\frac{\partial g(\mathbf{U})}{\partial \mathbf{U}}\Big)^\top \frac{\partial \mathbf{U}}{\partial \sigma_{ii}^2}\Big] = -2\mathbf{U}_{ii} = -2(\widehat{\mathbf{\Sigma}}_y - \mathbf{\Sigma}_\varepsilon - \mathbf{A}\widehat{\mathbf{\Sigma}}_y\mathbf{A}^\top)_{ii}.$$

The above is combined into an iterative procedure that runs over the to-be-estimated parameter and updates one-at-the-time until convergence. The parameter values at convergence are the estimates $\hat{\mathbf{A}}$ and $\widehat{\mathbf{\Sigma}}_\varepsilon$.

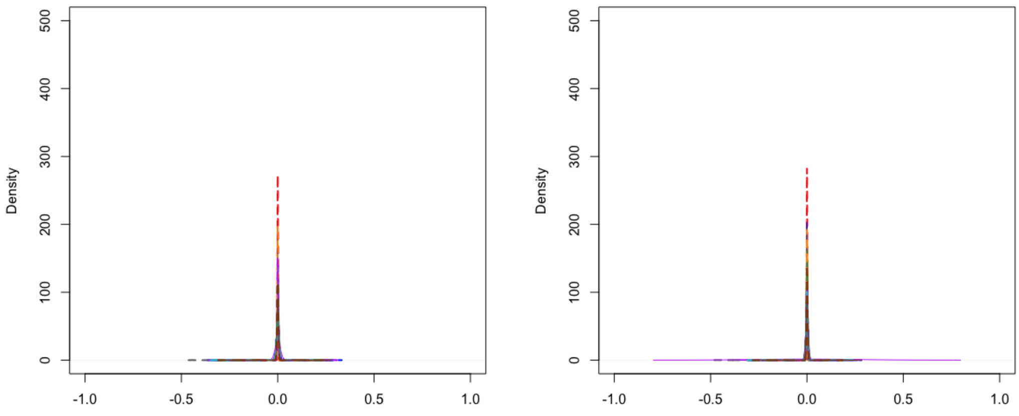
The sparsity assumption is again born out of the need for identifiability of the problem. Too few zeros in the autoregression parameter require the estimation of too many parameters while there are only $\frac{1}{2}p(p+1)$ degrees of freedom (inherited from the estimate of $\mathbf{\Sigma}_y$). In particular, as p degrees of freedom are to be reserved for the estimation of diagonal elements of $\mathbf{\Sigma}_\varepsilon$, the cardinality of the support of $\mathbf{A}^{(c)}$ (resp. $\mathbf{A}^{(h)}$) needs to satisfy $|\text{supp}(\mathbf{A}^{(h)})| \leq \frac{1}{2}p(p-1)$. This, however, is a necessary and not necessarily a sufficient condition. The latter is provided by the Implicit Function Theorem [117]. This theorem states (translated to the current context) that the map from the parameter space to that of a 'parameter × observations' space is one-to-one if its Jacobian is non-singular. This aligns with identifiability: each choice of the parameters yields a unique value in its range. The translation of this non-singularity condition on the Jacobian into tangible and general constraints is not deemed to provide extra insight at this point, as it depends much on the numerics at hand. But intuition suggests, though with a lot of handwaving, that a sparse $\mathbf{A}$ is more likely to induce a non-singular Jacobian (as it is less likely that two elements from the lower dimensional parameter domain may then map to the same point in the outcome space). Pragmatically, the non-singularity could be considered as being violated when the iterative estimation procedure does not converge.

4.3.4 Evaluation of the diagonal $\mathbf{\Sigma}_\varepsilon$ assumption

The strategies presented here hinge upon the assumption of a diagonal error covariance matrix of the vector autoregressive model. It is investigated whether there is any empirical ground for this assumption. This comprises the estimation of (the error covariance matrix of) the VAR(1) model from data of twenty signalling pathways (as those tend to be best defined) from two *in vitro* time-course studies. The twenty pathways and their dimensions have been defined on the basis of the KEGG repository and the presence of their genes in the data sets. The data sets consist of the previously described cervical cancer study and a novel colon cancer data. The latter consists of $n = 7$ cell lines interrogated at $T = 3$ time points and available from GEO, accession number GSE13059.

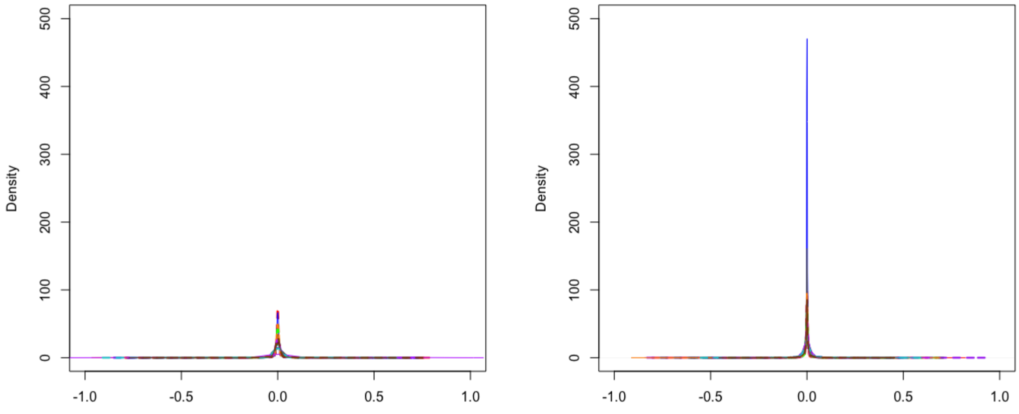
The VAR(1) model is estimated from each pathway's time-course data in penalized fashion as the dimension of most pathway exceeds that of the number of cell lines and time points involved. For reasons explained in Section 4.3, we use ridge penalized maximum likelihood to fit the VAR(1) model (see [99]). The ridge penalty parameters are chosen by (LOOCV) to optimize the cross-validated log-likelihood. The thus estimated error covariance matrices are inverted and standardized to obtain the partial correlations. For comparison, the VAR(2) model is fitted in similar fashion to the twenty signalling pathways' data from both time-course studies. Ridge-type estimates are non-sparse (although the partial correlations are all close to zero (Figures 4.1 & 4.2)), thus a sparsification step is needed. This is done by the local false discovery rate (lfdr) procedure [36, 129].

Figures 4.1 and 4.2 show the densities of the estimated partial correlations of the twenty pathways, for both data sets, and for the VAR(1) and VAR(2) models. All pathways have a partial correlation density with most mass tightly concentrated around zero. Only a few densities have some mass a little away from zero. This suggests that there are only relatively few off-diagonal elements in the estimated $\mathbf{\Sigma}_\varepsilon^{-1}$ that are possibly nonzero. Inclusion of a second order autoregressive term does not change the overall conclusion. However, for some pathways it may reduce a few nonzero partial correlations. In all, the plots suggest that a diagonal $\mathbf{\Sigma}_\varepsilon$ could be considered a reasonable working assumption when assumming a VAR(1) model.



(a) VAR(1) (b) VAR(2)

Figure 4.1: *Cervical cancer cell line data: densities of the partial correlations derived from the ridge penalized estimate of* $\mathbf{\Sigma}_{\varepsilon}$ *of the VAR(1) and VAR(2) model (left and right panel, respectively). Each panel contains twenty densities, one per signalling pathway and represented by different colors and line styles.*



(a) VAR(1) (b) VAR(2)

Figure 4.2: *Colon cancer cell line data: densities of the partial correlations derived from the ridge penalized estimate of* $\mathbf{\Sigma}_{\varepsilon}$ *of the VAR(1) and VAR(2) model (left and right panel, respectively). Each panel contains twenty densities, one per signalling pathway and represented by different colors and line styles.*

4.3.5 Illustration

Here the approaches outlined in Sections 4.3.2 and 4.3.3 are illustrated on the data from the observational and time-course cervical cancer study, described and used in the first illustration in Section 4.2.3. We first apply the approach of Section 4.3.2. It assumes the *in vivo* autogression parameter to be proportional to its *in vitro* counterpart, while considering the human error covariance matrix to be diagonal. Knowledge, e.g. an estimate, of $\mathbf{A}^{(c)}$ is required for this approach. This is obtained from the time-course cell line data by fitting the VAR(1) model using the ridge penalized maximimum likelihood approach of [99] and implemented in the `ragt2ridges`-package, thereby accommodating the high-dimensionality of the data. The resulting $\hat{\mathbf{A}}^{(c)}$ is then used to estimate the *in vivo* autoregression parameter under the proportionality assumption $\mathbf{A}^{(h)} = \delta \mathbf{A}^{(c)}$. This yields $\hat{\delta} = 0.223$. This is different from zero, which suggests that the cell lines harbour information on the *in vivo* dynamics. The strength of this information, as reflected in the exact value of this estimate, is hard to interpret. Finally, it should be kept in mind that we have selected the positive root of the estimate of δ^2, and the negative one ($\hat{\delta} = -0.223$) cannot be ruled out on the basis of the approach of Section 4.3.2.

The approach of Section 4.3.3 requires knowledge of support of the *in vitro* autoregression parameter. This is obtained from the ridge penalized maximum likelihood estimate of $\hat{\mathbf{A}}^{(c)}$, from which the largest (in an absolute sense) nonzero elements form the support. This support is now used to recover $\mathbf{A}^{(h)}$ with the same support using the approach Section 4.3.3. We only consider the top ten (resp. top fifteen) nonzero elements, as from around twenty nonzero elements onwards the proposed iterative algorithm did not converge. The actual values of thus obtained $\hat{\mathbf{A}}^{(h)}$ are not of interest here. But they may be compared to the corresponding values of its *in vitro* counterpart $\hat{\mathbf{A}}^{(c)}$, which is re-estimated obeying the same support. This aims to yield numeric values of the nonzero elements of $\hat{\mathbf{A}}^{(c)}$ and $\hat{\mathbf{A}}^{(h)}$ that are on a comparable scale (which was hampered by the penalized estimation of the orginal estimate of $\hat{\mathbf{A}}^{(c)}$). Figure 4.3 shows the scatter plots of the nonzero elements of both estimates. Both panels of this figure show a (weak) positive correlation between the two estimates. As only the support information is borrowed between the environments, the positive correlation indicates that information on the transcriptional dynamics is preserved between the *in vitro* and *in vivo* environments. Moreover, it corroborates with the analysis using approach of Section 4.3.2, which also pointed in the same direction of the dynamics of the *in vitro* environment being informative for the other. Finally, the positive correlation depicted in Figure 4.3 may be used to settle the sign issue of the analysis result from the approach of Section 4.3.2 discussed above: the positive sign is indeed more likely.

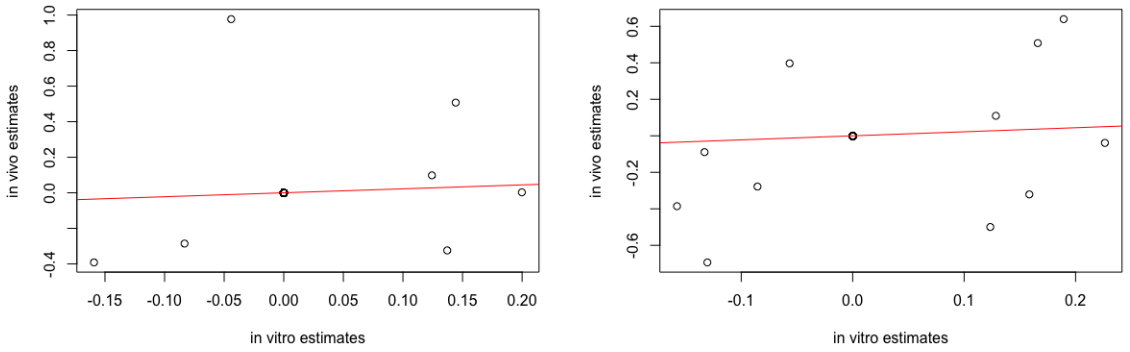


Figure 4.3: *Plots of* $\mathbf{A}^{(c)}$ *(in vitro) estimate values against* $\mathbf{A}^{(h)}$ *(in vivo) estimates using the method in Section 4.3.3. The number of nonzero elements was fixed to 10 (left subfigure) and 15 (right subfigure). The red line has slope* $\hat{\delta}$ *(Section 4.3.2).*

4.4 Recovering the precision matrix from the human VAR(1) model parameters

In this section we explore how the CIG can be efficiently computed given the *human* VAR(1) model parameters estimated from previous sections. Indeed, would the support of the VAR(1) model parameters $\mathbf{A}$ and $\mathbf{\Sigma}_\epsilon$ be known, then that of the precision matrix $\mathbf{\Omega}_y$ is implied. This is illustrated by a few examples.

Ex. 1) Let $\mathbf{A}$ be symmetric (cf. [148], for examples of this assumption) and $\mathbf{\Sigma}_\varepsilon = \mathbf{I}_p$. Then:

$$(4.8)\quad \mathbf{\Omega}_y^{-1} = \sum_{t=0}^{\infty} \mathbf{A}^t(\mathbf{A}^t)^\top = \sum_{t=0}^{\infty} \mathbf{A}^{2t} = (\mathbf{I}_{pp} - \mathbf{A}^2)^{-1} \quad \text{or} \quad \mathbf{\Omega}_y = \mathbf{I}_{pp} - \mathbf{A}^2.$$

This identity reveals that $\mathbf{A}^2$ and $\mathbf{\Omega}_y$ share the same off-diagonal elements. In particular, they have the same support.

Ex. 2) Let $\mathbf{A}$ be a block diagonal square matrix with blocks $\mathbf{A}_1, \ldots, \mathbf{A}_k$ and $\mathbf{\Sigma}_\varepsilon = \mathbf{I}_p$. Then, $\mathbf{A}^t$, $\mathbf{A}^\top$, $\mathbf{A}^{-1}$ are also block diagonal matrices. Hence, $\mathbf{\Omega}_y^{-1} = \sum_{t=0}^{\infty} \mathbf{A}^t(\mathbf{A}^t)^\top$ and $\mathbf{\Omega}_y$ are block diagonal with blocks of the same size as those of $\mathbf{A}$. In particular, $\mathbf{A}$ and $\mathbf{\Omega}_y$ share their support.

For a more general treatment on the relation between the VAR(1) model parameters and conditional (in)dependence (as implied by the support of the process precision matrix) the reader is referred to [28, 29].

4.5 Conclusion

This chapter explores several approaches to reconstruct gene regulatory networks from combinations of observational and time-course cell line gene expression data. The dynamics of the human cell are assumed to obey a first-order vector autoregression model and it is investigated how the underlying model parameters can be efficiently learned using the two types of data. Both existing and novel proposed strategies have been used to this end. The proposed strategies here hinge upon the assumption of a diagonal error covariance matrix. This assumption has been investigated in a large-scale simulation, results of which supported the diagonality assumption.

We observed in an application to real data that reconstruction of the conditional independence graph by borrowing information from the cell line data improves significantly. Moreover, our newly proposed strategies appear to be consistent in our data-driven analysis and indicate preserved transcriptional dynamics between the *in vitro* and *in vivo* environments.

4.6 Appendix I

Here the details of the derivation of the derivative of $\partial g(\mathbf{U}) = \|\mathbf{U}\|_2^2$ with $\mathbf{U} = \widehat{\boldsymbol{\Sigma}}^{(h)} - \mathbf{A}^{(h)}\widehat{\boldsymbol{\Sigma}}^{(h)}(\mathbf{A}^{(h)})^\top - \boldsymbol{\Sigma}_\varepsilon^{(h)}$ are provided. Note that below, for notational clarity, when subsetting a matrix the row and columns indices are separated by a comma, e.g. $\mathbf{A}_{i,j}$ instead of $\mathbf{A}_{ij}$. Substitution of $\mathbf{U}$ in the derivative of $g(\mathbf{U})$ yields:

$$\begin{aligned}
\frac{\partial g(\mathbf{U})}{\partial a_{ij}} &= -4(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon - \mathbf{A}\widehat{\boldsymbol{\Sigma}}\mathbf{A}^\top)_{i,*}(\mathbf{A}\widehat{\boldsymbol{\Sigma}})_{*,j} \\
&= -4[(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,*}(\mathbf{A}\widehat{\boldsymbol{\Sigma}})_{*,j} - (\mathbf{A}\widehat{\boldsymbol{\Sigma}}\mathbf{A}^\top)_{i,*}(\mathbf{A}\widehat{\boldsymbol{\Sigma}})_{*,j}] \\
&= 4[(\mathbf{A}\widehat{\boldsymbol{\Sigma}}\mathbf{A}^\top)_{i,*}(\mathbf{A}\widehat{\boldsymbol{\Sigma}})_{*,j} - (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,*}(\mathbf{A}\widehat{\boldsymbol{\Sigma}})_{*,j}] \\
&:= 4(\mathbf{I} - \mathbf{II})
\end{aligned}$$

Next, this expression (notably the terms $\mathbf{I}$ and $\mathbf{II}$) are simplified to separate terms involving the to-be-updated element of $\mathbf{A}$ from the others. In the following, $\mathbf{X}_{i,-k}$ and $\mathbf{X}_{-k,j}$ represents the i-th row and the j-th column, respectively, of the matrix $\mathbf{X}$ but with the k-element excluded.

$$\begin{aligned}
\mathbf{II} &= (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,*}(\mathbf{A}\widehat{\boldsymbol{\Sigma}})_{*,j} \\
&= (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}\mathbf{A}_{i,*}\widehat{\boldsymbol{\Sigma}}_{*,j} + (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,-i}\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j} \\
&= (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}(\mathbf{A}_{i,j}\widehat{\boldsymbol{\Sigma}}_{j,j} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}) + (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,-i}\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j} \\
&= (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}\widehat{\boldsymbol{\Sigma}}_{j,j}a_{ij} + (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j} + (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,-i}\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}
\end{aligned}$$

$$\begin{aligned}
\mathbf{I} &= \sum_{k\neq i}(\mathbf{A}_{i,*}\widehat{\boldsymbol{\Sigma}}(\mathbf{A}^\top)_{*,k})(\mathbf{A}_{k,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) + (\mathbf{A}_{i,*}\widehat{\boldsymbol{\Sigma}}(\mathbf{A}^\top)_{*,i})(\mathbf{A}_{i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
&= \mathbf{A} + \mathbf{B} \\
\mathbf{A} &= \sum_{k\neq i}(\mathbf{A}_{i,*}\widehat{\boldsymbol{\Sigma}}(\mathbf{A}^\top)_{*,k})(\mathbf{A}_{k,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
&= \sum_{k\neq i}\Big[(\mathbf{A}_{i,j}\widehat{\boldsymbol{\Sigma}}_{j,*}(\mathbf{A}^\top)_{*,k})(\mathbf{A}_{k,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) + (\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,*}(\mathbf{A}^\top)_{*,k})(\mathbf{A}_{k,*}\widehat{\boldsymbol{\Sigma}}_{*,j})\Big] \\
&= a_{ij}\Big(\widehat{\boldsymbol{\Sigma}}_{j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) + \Big(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
\mathbf{B} &= (\mathbf{A}_{i,*}\widehat{\boldsymbol{\Sigma}}(\mathbf{A}^\top)_{*,i})(\mathbf{A}_{i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
&= \Big[\mathbf{A}_{i,j}(\widehat{\boldsymbol{\Sigma}}_{j,*}(\mathbf{A}^\top)_{*,i}) + \mathbf{A}_{i,-j}(\widehat{\boldsymbol{\Sigma}}_{-j,*}(\mathbf{A}^\top)_{*,i})\Big]\Big[\mathbf{A}_{i,j}\widehat{\boldsymbol{\Sigma}}_{j,j} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\Big] \\
&= \Big[\mathbf{A}_{i,j}\Big(\widehat{\boldsymbol{\Sigma}}_{j,j}(\mathbf{A}^\top)_{j,i} + \widehat{\boldsymbol{\Sigma}}_{j,-j}(\mathbf{A}^\top)_{-j,i}\Big) + \mathbf{A}_{i,-j}\Big(\widehat{\boldsymbol{\Sigma}}_{-j,j}(\mathbf{A}^\top)_{j,i} + \widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\Big)\Big] \\
&\quad \Big[a_{ij}\widehat{\boldsymbol{\Sigma}}_{j,j} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\Big] \\
&= \Big[\mathbf{A}_{i,j}\widehat{\boldsymbol{\Sigma}}_{j,j}(\mathbf{A}^\top)_{j,i} + \mathbf{A}_{i,j}\widehat{\boldsymbol{\Sigma}}_{j,-j}(\mathbf{A}^\top)_{-j,i} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}(\mathbf{A}^\top)_{j,i} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\Big]
\end{aligned}$$

$$
\begin{aligned}
& \Big[a_{ij}\widehat{\boldsymbol{\Sigma}}_{j,j} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\Big] \\
= \; & \Big[\widehat{\boldsymbol{\Sigma}}_{j,j}a_{ij}^2 + 2\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}a_{ij} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\Big]\Big[a_{ij}\widehat{\boldsymbol{\Sigma}}_{j,j} + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\Big] \\
= \; & \Big(\widehat{\boldsymbol{\Sigma}}_{j,j}^2 a_{ij}^3 + 2\widehat{\boldsymbol{\Sigma}}_{j,j}\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}a_{ij}^2 + \mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\widehat{\boldsymbol{\Sigma}}_{j,j}a_{ij}\Big) + \\
& \Big(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\widehat{\boldsymbol{\Sigma}}_{j,j}a_{ij}^2 + 2(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j})^2 a_{ij} + (\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i})(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j})\Big) \\
= \; & \widehat{\boldsymbol{\Sigma}}_{j,j}^2 a_{ij}^3 + 3\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\widehat{\boldsymbol{\Sigma}}_{j,j}a_{ij}^2 + \Big(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\widehat{\boldsymbol{\Sigma}}_{j,j} + 2(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j})^2\Big)a_{ij} \\
& + (\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i})(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j})
\end{aligned}
$$

Aggregate the above to arrive at:

$$
\begin{aligned}
\mathbf{I} \; = \; & \widehat{\boldsymbol{\Sigma}}_{j,j}^2 a_{ij}^3 \\
& + 3\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\widehat{\boldsymbol{\Sigma}}_{j,j}a_{ij}^2 \\
& + \Big[\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\widehat{\boldsymbol{\Sigma}}_{j,j} + 2(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j})^2 + \Big(\widehat{\boldsymbol{\Sigma}}_{j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j})\Big]a_{ij} \\
& + (\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i})(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}) + \Big(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j})
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\frac{\partial g(\mathbf{U})}{\partial a_{ij}} \; = \; & 4\Big\{\widehat{\boldsymbol{\Sigma}}_{j,j}^2 a_{ij}^3 \\
& + 3\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\widehat{\boldsymbol{\Sigma}}_{j,j}a_{ij}^2 \\
& + \Big[\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\widehat{\boldsymbol{\Sigma}}_{j,j} + 2(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j})^2 + \Big(\widehat{\boldsymbol{\Sigma}}_{j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
& - (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}\widehat{\boldsymbol{\Sigma}}_{j,j}\Big]a_{ij} \\
& + (\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i})(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}) + \Big(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
& - (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j} - (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,-i}\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}\Big\} \\
= \; & \mathbf{K}_3 a_{ij}^3 + \mathbf{K}_2 a_{ij}^2 + \mathbf{K}_1 a_{ij} + \mathbf{K}_0,
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{K}_0 \; = \; & 4\Big\{(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i})(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}) + \Big(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
& - (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j} - (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,-i}\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}\Big\}, \\
\mathbf{K}_1 \; = \; & 4\Big[\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,-j}(\mathbf{A}^\top)_{-j,i}\widehat{\boldsymbol{\Sigma}}_{j,j} + 2(\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j})^2 + \Big(\widehat{\boldsymbol{\Sigma}}_{j,*}(\mathbf{A}^\top)_{*,-i}\Big)(\mathbf{A}_{-i,*}\widehat{\boldsymbol{\Sigma}}_{*,j}) \\
& - (\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}_\varepsilon)_{i,i}\widehat{\boldsymbol{\Sigma}}_{j,j}\Big], \\
\mathbf{K}_2 \; = \; & 12\mathbf{A}_{i,-j}\widehat{\boldsymbol{\Sigma}}_{-j,j}\widehat{\boldsymbol{\Sigma}}_{j,j}, \\
\mathbf{K}_3 \; = \; & 4\widehat{\boldsymbol{\Sigma}}_{j,j}^2.
\end{aligned}
$$

Bibliography

[1] Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, 14(3):586–599.

[2] Allen, G. and Liu, Z. (2013a). A local poisson graphical model for inferring networks from sequencing data. *NanoBioscience, IEEE Transactions on*, 12(3):189–198.

[3] Allen, G. I. and Liu, Z. (2013b). A local poisson graphical model for inferring networks from sequencing data. *NanoBioscience, IEEE Transactions on*, 12:189–198.

[4] Andersson, S., Madigan, D., and Perlman, M. (2001). Alternative markov properties for chain graphs. *Scandinavian Journal of Statistics*, 28(1):33 – 85.

[5] Archambeau, C. and Bach, F. (2008). Sparse probabilistic projections. In *Advances in Neural Information Processing Systems 21*, pages 73–80. Curran Associates, Inc.

[6] Armagan, A. (2009). Variational bridge regression. *Journal of Machine Learning Research, Workshop and Conference Proceedings*, 5:17–24.

[7] Attias, H. (1999). Inferring parameters and structure of latent variable models by variational bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, pages 21–30. Morgan Kaufmann Publishers Inc.

[8] Berenson, L. and et al. (2006). Selective requirement of p38alpha mapk in cytokine-dependent, but not antigen receptor-dependent, th1 responses. *J. Immunol.*, 176:4616–4621.

[9] Bhattacharya, A., Chakraborty, A., and Mallick, B. (2016). Fast sampling with gaussian scale-mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

[10] Bilgrau, A. and et al. (2015). Targeted fused ridge estimation of inverse covariance matrices from multiple high-dimensional data classes. *arXiv preprint arXiv:1509.07982*.

[11] Blei, D. and Jordan, M. (2011). Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1:121 – 143.

[12] Blei, D., Kucukelbir, A., and McAuliffe, J. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

[13] Blei, D. M. and Jordan, M. I. (2006). Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–144.

[14] Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *J. Amer. Statist. Assoc.*, 107(500):1610–1624.

[15] Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.

[16] Cai, X., Huang, A., and Xu, S. (2011). Fast empirical bayesian lasso for multiple quantitative trait locus mapping. *BMC Bioinformatics*, 12(1):211.

[17] Camby, I., Le Mercier, M., Lefranc, F., and Kiss, R. (2006). Galectin-1: a small protein with major functions. *Glycobiology*, 16(11):137R–157R.

[18] Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

[19] Carvalho, C., Polson, N., and Scott, J. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research, W&CP*, 5:73–80.

[20] Carvalho, C., Polson, N., and Scott, J. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480.

[21] Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., and Schultz, N. (2012). The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404.

[22] Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., Schultz, N., Bader, G. D., and Sander, C. (2011). Pathway commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(suppl 1):D685–D690.

[23] Chen, J. and Chen, Z. (2008a). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771.

[24] Chen, J. and Chen, Z. (2008b). Extended bayesian information criterion for model selection with large model space. *Biometrika*, 95:759–771.

[25] Chen, S., Witten, D. M., and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika*.

[26] Cheung, V., Spielman, R., Ewens, K., Weber, T., Morley, M., and Burdick, J. (2005). Mapping determinants of human gene expression by regional and genome-wide association. *Nature*, 437:1365 – 1369.

[27] Cordes, C., Bartling, B., Simm, A., Afar, D., Lautenschläger, C., Hansen, G., Silber, R.-E., Burdach, S., and Hofmann, H.-S. (2009). Simultaneous expression of cathepsins b and k in pulmonary adenocarcinomas and squamous cell carcinomas predicts poor recurrence-free and overall survival. *Lung Cancer*, 64(1):79 – 85.

[28] Dahlhaus, R. (2000). Graphical interaction models for multivariate time series. *Metrika*, 51(2):157–172.

[29] Dahlhaus, R. and Eichler, M. (2003). *Causality and graphical models in time series analysis.* Oxford Statistical Science Series.

[30] Danaher, P., Wang, P., and Witten, D. (1999). The joint graphical lasso for inverse covariance estimation across multiple classes. *J R Stat Soc Series B Stat Methodol.*, 76(2):373–397.

[31] Datta, J. and Ghosh, J. (2013). Asymptotic properties of bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132.

[32] de Angelis, L. and et al. (2005). Regulation of vertebrate myotome development by the p38 map kinase-mef2 signaling pathway. *Dev. Biol.*, 283:171–179.

[33] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90:196 – 212.

[34] Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.*, 106(496):1418–1433.

[35] Dodd, L. E. and Pepe, M. S. (2003). Partial AUC estimation and regression. *Biometrics*, 59(3):614–623.

[36] Efron, B. (2004). Large-scale simultaneous hypothesis testing. *Journal of the American Statistical Association*, 99:96–104.

[37] Efron, B. (2010). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction.* Institute of Mathematical Statistics Monographs 1. Cambridge: Cambridge University Press. xii, 263 p., Cambridge.

[38] Espinosa, A. and et al. (2013). Mitosis is a source of potential markers for screening and survival and therapeutic targets in cervical cancer. *PLOS ONE*, 8(2):1–21.

[39] Feng, Y. and et al. (2018). Novel genetic variants in the p38mapk pathway gene zak and susceptibility to lung cancer. *Mol. Carcinog.*, 57(2):216–224.

[40] Fortin, S., Le Mercier, M., Camby, I., Spiegl-Kreinecker, S., Berger, W., Lefranc, F., and Kiss, R. (2010). Galectin-1 is implicated in the protein kinase c epsilon/vimentin-controlled trafficking of integrin-beta1 in glioblastoma cells. *Brain Pathology*, 20(1):39–49.

[41] Foygel, R. and Drton, M. (2010a). Extended Bayesian information criteria for Gaussian graphical models. In Lafferty, J., Williams, C. K. I., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 604–612.

[42] Foygel, R. and Drton, M. (2010b). Extended bayesian information criteria for gaussian graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1*, NIPS'10, pages 604–612, USA. Curran Associates Inc.

[43] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441.

[44] Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1–22.

[45] Galassi, M. (2009). *GNU Scientific Library : reference manual for GSL version 1.12*. Network Theory, Bristol, UK, 3rd edition.

[46] Gao, X., Pu, D., Wu, Y., and Xu, H. (2012). Tuning parameter selection for penalized likelihood estimation of gaussian graphical model. *Statistica Sinica*, 22:1123 – 1146.

[47] Geiger, D. and Heckerman, D. (2002). Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *Ann. Statist.*, 30(5):1412–1440.

[48] Giordano, R., Broderick, T., and Jordan, M. (2017). Covariances, robustness, and variational bayes. *arXiv preprint arXiv:1709.02536*.

[49] Giraud, C. (2008). Estimation of Gaussian graphs by model selection. *Electron. J. Stat.*, 2:542–563.

[50] Giudici, P. and Green, P. J. (1999). Decomposable graphical gaussian model determination. *Biometrika*, 86:785 – 801.

[51] Gole, B., Huszthy, P. C., Popović, M., Jeruc, J., Ardebili, Y. S., Bjerkvig, R., and Lah, T. T. (2012). The regulation of cysteine cathepsins and cystatins in human gliomas. *International Journal of Cancer*, 131(8):1779–1789.

[52] Gradshteyn, I. and Ryzhik, I. (1994). *Tables of Integrals, Series, and Products*. Academic Press, San Diego, California, 5th edition.

[53] Hahn, P. R. and Carvalho, C. M. (2015). Decoupling shrinkage and selection in bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, 110(509):435–448.

[54] Hamid, J., Hu, P., Roslin, N., Ling, V., Greenwood, C., and Beyene, J. (2009:869093). Data integration in genetics and genomics: Methods and challenges. *Human Genomics and Proteomics: HGP*.

[55] Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell*, 100:57–70.

[56] Hankin, R. (2007). Special functions in r: introducing the **gsl** package. *R News*, 6.

[57] Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC.

[58] Hawkins, R., Hon, G., and Ren, B. (2010). Next-generation genomics: an integrative approach. *Nat. Rev. Genet.*, 11(7):476–86.

[59] Heit, B., Tavener, S., Raharjo, E., and Kubes, P. (2002). An intracellular signaling hierarchy determines direction of migration in opposing chemotactic gradients. *J. Cell Biol.*, 159:91–102.

[60] Heuertz, R., Tricomi, S., Ezekiel, U., and Webster, R. (1999). C-reactive protein inhibits chemotactic peptide-induced p38 mitogen-activated protein kinase activity and human neutrophil movement. *J. Biol. Chem.*, 274:17968–17974.

[61] Hurley, D. and et al. (2012). Gene network inference and visualization tools for biologists: application to new human transcriptome datasets. *Nucleic Acids Research*, 40(6):2377–2398.

[62] Hwang, D., Jang, B., Yu, G., and Boudreau, M. (1997). Expression of mitogen- inducible cyclooxygenase induced by lipopolysaccharide: mediation through both mitogen-activated protein kinase and nf-kappab signaling pathways in macrophages. *Biochem. Pharmacol.*, 54:87–96.

[63] Igney, F. and Krammer, P. (2002). Death and anti-death: tumour resistance to apoptosis. *Nat Rev Cancer.*, 2:277–88.

[64] Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S., and Miyano, S. (2003). Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. In *Computational Systems Bioinformatics. CSB2003. Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, pages 104–113.

[65] Isci, S., Dogan, H., Ozturk, C., and Otu, H. (2013). Bayesian network prior: Network analysis of biological data using external knowledge. *Bioinformatics*, pages 860–867.

[66] Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *Ann. Statist.*, 33(2):730–773.

[67] Jacobsen, A. (2013). *cgdsr: R-Based API for accessing the MSKCC Cancer Genomics Data Server (CGDS).* R package version 1.1.30.

[68] Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high dimensional graphical models. *Statist. Sci.*, 20:388 – 400.

[69] Kallunki, T., Olsen, O., and Jäättelä, M. (2013). Cancer-associated lysosomal changes: friends or foes? *Oncogene*, 32(16):1995–2004.

[70] Kass, R. E. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90:928–934.

[71] Kim, S. and Xing, E. (2012). Tree-guided group lasso for multi-response regression with structured sparsity, with an application to eqtl mapping. *Ann. Appl. Stat.*, 6(3):1095 – 1117.

[72] Kpogbezan, G., van der Vaart, A., van Wieringen, W., Leday, G., and van de Wiel, M. (2017). An empirical bayes approach to network recovery using external knowledge. *Biom. Journal*, 59(5):932–947.

[73] Krämer, N. and et al. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10:384.

[74] Krämer, N., Schafer, J., and Boulesteix, A.-L. (2009). Regularized estimation of large-scale gene association networks using graphical gaussian models. *BMC Bioinformatics*, 10(1):384.

[75] Krammer, P. H., Galle, P. R., Moller, P., and Debatin, K. M. (1998). Cd95(apo-1/fas)-mediated apoptosis in normal and malignant liver, colon, and hematopoietic cells. *Adv. Cancer Res.*, 75:251–273.

[76] Landi, M. T., Dracheva, T., Rotunno, M., Figueroa, J., H. Liu, A. D., Mann, F., Fukuoka, J., Hames, M., Bergen, A., Murphy, S., Yang, P., Pesatori, A., Consonni, D., Bertazzi, P., Wacholder, S., Shih, J., Caporaso, N., and Jen, J. J. (2008). Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS ONE*, 3:e1651. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10072.

[77] Lappalainen, T. and et al. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511.

[78] Lauritzen, S. (1996). *Graphical models.* The Clarendon Press, Oxford University Press, New York.

[79] Lauritzen, S., Dawid, A., Larsen, B., and Leimer, H.-G. (1990). Independence properties of directed markov fields. *Networks*, 20:491 – 505.

[80] Leday, G., de Gunst, M., Kpogbezan, G., van der Vaart, A., van Wieringen, W., and van de Wiel, M. (2017). Gene network reconstruction using global-local shrinkage priors. *Ann. Appl. Stat.*, 11(1):41 – 68.

[81] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411.

[82] Lee, J. and et al. (1994). A protein kinase involved in the regulation of inflammatory cytokine biosynthesis. *Nature*, 372:739–746.

[83] Lee, S., Pe'er, D., Dudley, A., Church, G., and Koller, D. (2006). Identifying regulatory mechanisms using individual variation reveals key role for chromatin modification. *Proc. Natl. Acad. Sci. USA*, 103:14062 – 14067.

[84] Lehne, B., Lewis, C., and Schlitt, T. (2011). From snps to genes: Disease association at the gene level. *PLos ONE*, 6(6):e20133.

[85] Lewis, C., Brault, C., Peck, B., Bensaad, K., Griffiths, B., Mitter, R., Chakravarty, P., East, P., Dankworth, B., Alibhai, D., et al. (2015). Srebp maintains lipid biosynthesis and viability of cancer cells under lipid-and oxygen-deprived conditions and defines a gene signature associated with poor survival in glioblastoma multiforme. *Oncogene.*

[86] Li, G., Shabalin, A., Rusyn, I., Wright, F., and Nobel, A. (2018). An empirical bayes approach for multiple tissue eqtl analysis. *Biostatistics*, 19(3):391–406.

[87] Li, S., Wu, L., and Zhang, Z. (2006). Constructing biological networks through combined literature mining and microarray analysis: a LMMA approach. *Bioinformatics*, 22:2143 – 2150.

[88] Li, Y., Jiang, B., Ensign, W., Vogt, P., and Han, J. (2000). Myogenic differentiation requires signalling through both phosphatidylinositol 3-kinase and p38 map kinase. *Cell. Signal.*, 12:751–757.

[89] Lian, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *J. Statist. Plann. Inference*, 141(8):2839–2848.

[90] Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g-priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

[91] Lim, K., Lim, K., Price, A., Orr, B., Eberhart, C., and Bar, E. (2013). Inhibition of monocarboxylate transporter-4 depletes stem-like glioblastoma cells and inhibits hif transcriptional response in a lactate-independent manner. *Oncogene.*

[92] Luetkepohl, H. (2005). *The New Introduction to Multiple Time Series Analysis.* Springer, Berlin.

[93] Luo, S., Song, R., and Witten, D. (2014). Sure Screening for Gaussian Graphical Models. *arXiv:1407.7819.*

[94] MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, New York, NY, USA.

[95] Madhankumar, A., Slagle-Webb, B., Mintz, A., Sheehan, J. M., and Connor, J. R. (2006). Interleukin-13 receptor–targeted nanovesicles are a potential therapy for glioblastoma multiforme. *Molecular Cancer Therapeutics*, 5(12):3162–3169.

[96] McGrory, C. and Titterington, D. (2007). Variational approximations in bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*, 51:5352 – 5367.

[97] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34:1436–1462.

[98] Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 72(4):417–473.

[99] Miok, V., Wilting, S., and van Wieringen, W. (2017). Ridge estimation of the var(1) model and its time series chain graph from multivariate time-course omics data. *Biometrical Journal*, 59(1):172–191.

[100] Mohammadi, A. and Wit, E. C. (2015a). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Anal.*, 10(1):109–138.

[101] Mohammadi, A. and Wit, E. C. (2015b). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Anal.*, 10:109 – 138.

[102] Mukherjee, S. and Speed, T. (2008). Network inference using informative priors. *Proceedings of the National Academy of Sciences*, 105:14313–14318.

[103] Oates, C. J. and Mukherjee, S. (2012). Network inference and biological dynamics. *Ann. Appl. Stat.*, 6(3):1209–1235.

[104] Ormerod, J. and Wand, M. (2010a). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

[105] Ormerod, J. and Wand, M. (2010b). Explaining variational approximations. *The American Statistician*, 64:140–153.

[106] Ormerod, J. T. and Wand, M. P. (2010c). Explaining variational approximations. *Amer. Statist.*, 64(2):140–153.

[107] Park, T. and Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.*, 103(482):681–686.

[108] Pearl, J. and Paz, A. (1986). Graphoids: A graph-based logic for reasoning about relevancy relations. In *Proceedings of the European Conference on Artificial Intelligence.*

[109] Peeters, C. and van Wieringen, W. (2014). *rags2ridges: Ridge Estimation of Precision Matrices from High-Dimensional Data.* R package version 1.4.

[110] Peng, J., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.*, 104(486):735–746.

[111] Polson, N., Scott, J., and J., W. (2014). The bayesian bridge. *Journal of the Royal Statistical Society: Series B*, 76:713 – 733.

[112] Porstmann, T., Santos, C. R., Griffiths, B., Cully, M., Wu, M., Leevers, S., Griffiths, J. R., Chung, Y.-L., and Schulze, A. (2008). Srebp activity is regulated by mtorc1 and contributes to akt-dependent cell growth. *Cell metabolism*, 8(3):224–236.

[113] Pujana, M., Han, J., Starita, L., Stevens, K., Tewari, M., Ahn, J., Rennert, G., Moreno, V., Kirchhoff, T., and Gold, B. e. a. (2007). Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics*, 39:1338 – 1349.

[114] Rajagopalan, M. and Broemeling, L. (1983). Bayesian inference for the variance components in general mixed linear models. *Comm. Statist. A—Theory Methods*, 12(6):701–723.

[115] Ravikumar, P., Wainwright, M. J., and Lafferty, J. D. (2010). High-dimensional ising model selection using l1-regularized logistic regression. *Ann. Statist.*, 38:1287–1319.

[116] Reif, D., White, B., and Moore, J. (2004). Integrated analysis of genetic, genomic and proteomic data. *Expert Rev. Proteomics*, 1:67–75.

[117] Rudin, W. (1991). *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill Inc., New York, second edition.

[118] Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):319–392.

[119] Schaefer, J., Opgen-Rhein, R., and Strimmer, K. (2006). Reverse engineering genetic networks using the GeneNet package. *R News*, 6/5:50–53.

[120] Schäfer, J., Opgen-Rhein, R., and Strimmer, K. (2006). Reverse engineering genetic networks using the GeneNet package. *R News*, 6:50–53.

[121] Schäfer, J. and Strimmer, K. (2005a). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

[122] Schäfer, J. and Strimmer, K. (2005b). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 32, 28 pp. (electronic).

[123] Schröder, N. and Schumann, R. (2005). Single nucleotide polymorphisms of toll-like receptors and susceptibility to infectious disease. *Lancet Infect Dis.*, 5:156 – 64.

[124] Scutari, M. (2013). On the prior and posterior distributions used in graphical modelling. *Bayesian Analysis*, 8(1):1–28.

[125] Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, 34:166 – 178.

[126] Srividhya, J., Crampin, E., McSharry, P., and Schnell, S. (2007). Reconstructing biochemical pathways from time course data. *Proteomics*, 7(6):828–38.

[127] Steele, E., Tucker, A., 't Hoen, P., and Schuemie, M. (2009). Literature-based priors for gene regulatory networks. *Bioinformatics*, 25(14):1768–1774.

[128] Stranger, B., Forrest, M., Clark, A., Minichiello, M., Deutsch, S., and Lyle, R. e. a. (2005). Genome-wide associations of gene expression variation in humans. *PLoS Genetics*, 1:695 – 704.

[129] Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9(1):303.

[130] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

[131] Turner, R. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, T., and Chiappa, S., editors, *Bayesian Time series models*. Cambridge University Press.

[132] Valpola, H. and Honkela, A. (2006). Hyperparameter adaptation in variational bayes for the gamma distribution. Technical report, Helsinki University of Technology.

[133] van de Wiel, M., Leday, G., Pardo, L., Rue, H., van der Vaart, A., and van Wieringen, W. (2013a). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14(1):113–128.

[134] van de Wiel, M., Leday, G., Pardo, L., Rue, H., van der Vaart, A., and van Wieringen, W. (2013b). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*, 14:113 – 128.

[135] van de Wiel, M., Te Beest, D., and Münch, M. (2018). Learning from a lot: Empirical bayes for high-dimensional model-based prediction. *Scandinavian Journal of Statistics*, pages 1–24.

[136] van der Pas, S., Kleijn, B., and van der Vaart, A. (2014). The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

[137] van der Pas, S., Szabó, B., and van der Vaart, A. (2017a). Adaptive posterior contraction rates for the horseshoe. *Electron. J. Stat.*, 11(2):3196–3225.

[138] van der Pas, S., Szabó, B., and van der Vaart, A. (2017b). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Anal.*, 12(4):1221–1274. With a rejoinder by the authors.

[139] van Wieringen, W. and Peeters, C. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics & Data Analysis*, 103:284 – 303.

[140] van Wieringen, W. N. and Peeters, C. F. W. (2014). Ridge estimation of inverse covariance matrices from high-dimensional data. *ArXiv e-prints*.

[141] Wainwright, M. and Jordan, M. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.

[142] Wand, M., Ormerod, J., Padoan, S., and Frühwirth, R. (2011). Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900.

[143] Wang, B. and Titterington, M. (2004). Inadequacy of interval estimates corresponding to variational bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*.

[144] Wang, H. and Li, S. Z. (2012). Efficient Gaussian graphical model determination under G-Wishart prior distributions. *Electron. J. Stat.*, 6:168–198.

[145] Wang, Y. and Blei, D. (2017). Frequentist consistency of variational bayes. *arXiv preprint arXiv:1705.03439*.

[146] Wang, Y. and et al. (2013). Integration of steady-state and temporal gene expression data for the inference of gene regulatory networks. *PLOS ONE*, 8(8):1–11.

[147] Warton, D. I. (2008). Penalized normal likelihood and ridge regularization of correlation and covariance matrices. *J. Amer. Statist. Assoc.*, 103(481):340–349.

[148] Weiner, I., Schmitt, N., and Highhouse, S. (2012). *Handbook of Psychology, Industrial and Organizational Psychology*, volume 12. John Wiley and Sons.

[149] Werhli, A. and Husmeier, D. (2007). Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge. *Stat Appl Genet Mol Biol*, 6.

[150] West, M. (2003). Bayesian factor regression models in the “large p, small n” paradigm. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 733–742. Oxford Univ. Press, New York.

[151] Westling, T. and McCormick, T. (2015). Establishing consistency and improving uncertainty estimates of variational inference through m-estimation. *arXiv preprint arXiv:1510.08151.*

[152] Whittaker, J. (1990). *Graphical models in applied multivariate statistics*. Wiley, Chichester.

[153] Wilting, S. and et al. (2016). Aberrant methylation-mediated silencing of micrornas contributes to hpv-induced anchorage independence. *Oncotarget*, 7(28):43805.

[154] Wu, Z. and et al. (2000). p38 and extracellular signal-regulated kinases regulate the myogenic program at multiple steps. *Mol. Cell. Biol.*, 20:3951–3964.

[155] Yajima, M., Telesca, D., Ji, Y., and Muller, P. (2012). Differential patterns of interaction and gaussian graphical models. *COBRA Preprint Series*, (91).

[156] Yang, E., Allen, G., Liu, Z., and Ravikumar, P. K. (2012). Graphical models via generalized linear models. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 1358–1366. Curran Associates, Inc.

[157] Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, 11:2261–2286.

[158] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35.

[159] Yuan, Y., Curtis, C., Caldas, C., and Markowetz, F. (2012). A sparse regulatory network of copy-number driven gene expression reveals putative breast cancer oncogenes. *IEEE/ACM Trans Comput Biol Bioinform*, 9(4):947–954.

[160] Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian statistics: Proceedings of the first international meeting held in Valencia (Spain)*, volume 1. Valencia: University Press.

[161] Zetser, A., Gredinger, E., and Bengal, E. (1999). p38 mitogen-activated protein kinase pathway promotes skeletal muscle differentiation. participation of the mef2c transcription factor. *J. Biol. Chem.*, 274:5193–5200.

[162] Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 17, 45 pp.

[163] Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in r. *J. Mach. Learn. Res.*, 13:1059 – 1062.

[164] Zhou, S., Rütimann, P., Xu, M., and Bühlmann, P. (2011). High-dimensional covariance estimation based on Gaussian graphical models. *J. Mach. Learn. Res.*, 12:2975–3026.

[165] Zhu, J., Zhang, B., Smith, E., Drees, B., Brem, R., Kruglyak, L., Bumgarner, R., and Schadt, E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. *Nature Genetics*, 40:854–861.

Samenvatting

In dit proefschrift hebben we statistische methoden voor de analyse van hoog-dimensionele data ontwikkeld. In het bijzonder besteden we aandacht aan reconstructie van hoog-dimensionele netwerken. In de genetica is de identificatie van gen-regulerende netwerken cruciaal voor het begrijpen van genfunctie, en derhalve belangrijk zowel in behandeling als voorspelling van ziekten. Hoog-dimensionele netwerkreconstructie is een zeer uitdagende taak aangezien het aantal mogelijke grafen exponentieel groeit met het aantal variabelen (e.g. genen). Echter, een aantal van de verbanden tussen deze variabelen zijn mogelijk bekend vanuit de literatuur. Zo zijn de huidige overtuigingen op het gebied van interacties tussen genen bijvoorbeeld geconcentreerd in kennisbanken als KEGG en Reactome. We introduceren een raamwerk waarin het opnemen van dergelijke *a-priori* informatie in de reconstructie mogelijk wordt op een zachte manier: het informeert de analyse wanneer correct, maar kan gecompenseerd worden indien volledig incompatibel met de data. We behandelen ook de onderwerpen genetische associatie studies (*eQTL mapping*) en data integratie.

In hoofdstuk 1 introduceren we een nieuwe *global-local shrinkage prior* van het type *ridge* voor niet-gerichte netwerk reconstructie gebaseerd op *SEMs* met *a-posteriori* selectie van zijden. De voorgestelde aanpak is computationeel snel en doet het beter dan bekende concurrenten zoals de *graphical lasso*.

In hoofdstuk 2 breiden we hoofdstuk 1 uit door prior informatie toe te voegen in de reconstructie van niet-gerichte netwerken. Er wordt rekening gehouden met deze prior kennis op een zachte manier die de beschikbare data toelaat om de prior informatie te compenseren indien deze niet relevant is. Bovendien is de voorgestelde methode in staat om de overeenstemming tussen de beschikbare data en de prior informatie expliciet te schatten, hetgeen een noviteit is bij het toevoegen van prior informatie in netwerk inferentie.

In hoofdstuk 3 introduceren we een raamwerk voor het simultaan analyseren van meerdere gerelateerde hoog-dimensionele en complexe datasets. Zulke analyses omvatten onder andere gen-regulerende netwerk reconstructies, genetische associatie studies (e.g. eQTL mapping) en data integratie in genomica. Om de analyse van kleine n relatief tot grote p mogelijk te maken introduceren we de *hoefijzer* prior die *sparsity* toelaat; een gewenste eigenschap voor de analyse van zulke data. We illustreren de procedure met twee toepassingen, namelijk: de reconstructie van gen-regulerende netwerken en eQTL mapping.

In hoofdstuk 4 verkennen we diverse methoden die gen-regulerende netwerken reconstrueren door combinatie van observationele (*in vivo*) en temporele cellijn gen-expressie (*in vitro*) data. De dynamiek van de humane cel wordt aangenomen een eerste-orde vector autoregressie VAR(1) model te volgen en er wordt onderzocht hoe de onderliggende model parameters efficient geleerd kunnen worden door gebruik van de twee soorten datasets. We zien in een toepassing op echte data dat de reconstructie van de conditionele onafhankelijkheidsgraaf significant verbetert door informatie van de cellijn te lenen. Bovendien lijken onze voorgestelde strategieën om de VAR(1) model parameters te leren consistent in onze data-gedreven analyse en tonen behoud van transcriptionele dynamiek tussen de *in vivo* en *in vitro* omgevingen.

Summary

In this thesis we developed statistical methods for the analysis of high dimensional data. We particularly focussed on high dimensional networks reconstruction. In genomics, the identification of gene regulatory networks is crucial for understanding gene function, and hence important for both treatment and prediction of diseases. High dimensional networks reconstruction is a very challenging task since the number of possible graphs grows exponentially with the number of variables (e.g. genes). However, some of the relationships between these variables may be known from the literature. For instance, the current beliefs on interactions among genes is condensed in repositories like KEGG and Reactome. We introduce a framework which allows the incorporation of such prior information in the reconstruction in a soft manner such that it informs the analysis if correct, but can be overruled if completely incompatible with the data. We also treat the subjects of genetic association studies (eQTL mapping) and data integration.

In chapter 1 we introduce a new global-local shrinkage ridge-type prior for undirected networks reconstruction based on SEMs with posterior edge selection. The proposed approach is computationally fast and outperforms known competitors such as the *graphical lasso*.

In chapter 2 we extend chapter 1 to include prior information in reconstructing undirected networks. The incorporation of the prior knowledge is done in a soft manner allowing the data at hand to overrule the prior information if not relevant. Furthermore, the proposed method is able to explicitly estimate the agreement of the prior knowledge with the data at hand which is a novelty in incorporating prior information in network inference.

In chapter 3 we introduce a framework for simultaneously analysing multiple related high dimensional and complex datasets. Such analyses include gene regulatory network reconstruction, genetic association studies (e.g. eQTL mapping) and data integration in genomics, to name but a few. To enable the analysis for small n relative to large p, we introduce the *horseshoe* prior which allows for sparsity; a desired property for the analysis of such data. We illustrate the approach by two applications, namely: to the reconstruction of gene regulatory networks and to eQTL mapping.

In chapter 4 we explore several approaches to reconstruct gene regulatory networks from combining observational (*in vivo*) and time-course cell line (*in vitro*) gene expression data. The dynamics of the human cell are assumed to obey a first-order vector autoregression VAR(1) model and it is investigated how the underlying model parameters can be efficiently learned using the two types of datasets. We see in an application to real data that reconstruction of the conditional independence graph by borrowing information from the cell line data improves significantly. Moreover, our newly proposed strategies to learn the VAR(1) model parameters are able to indicate preserved transcriptional dynamics between the *in vitro* and *in vivo* environments.

Acknowledgements

I'm sincerely grateful to many people for their help in the completion of this thesis. First, I wish to express my gratitude to my supervisors Aad van der Vaart, Mark van de Wiel and Wessel van Wieringen for their patience in guiding and working with me during my PhD time. I particularly thank Aad for giving me all the time I need and also for his power to speed up things everytime we want to. This thesis may not exist without him.

Je voudrais aussi remercier Gwen avec qui j'ai eu l'opportunité de collaborer durant la première année de ma thèse et même après son départ pour Cambridge. Je ne pouvais pas espérer un meilleur début pour ma thèse. Nos échanges souvent en français ont été très utiles pour moi puisqu'en ce moment mes connaissances en anglais étaient faibles.

I would like to thank all the staff at both the Mathematical Institute (MI) at Leiden University and the Vrije Universiteit medical center (VUmc). I really enjoyed switching between both institutes. I particularly thank Maarten, Stéphanie, Kevin, Carel and Armin, to name but a few.

I am grateful to the members of the reading committee for their time spent in reading the manuscript and for their feedback.

Mon dernier mot de remerciements va à ma famille pour leur soutien. Je remercie particulièrement Tata Vêdoko que j'appelle affectueusement 'Tanto' pour avoir été une personne ressource pour moi pendant toute cette période de mon doctorat. Je remercie également Julien Allognikou pour avoir été d'un soutien constant.

Gino B. Kpogbezan
Leiden, November 2018

Curriculum vitae

Gino Bertrand Kpogbezan was born on September 5, 1981 in Cotonou, Benin. He attended the Berlin Institute of Technology (TU Berlin) where he completed a Master degree in Mathematics in 2012. In 2013, he worked as research assistant in the field of Mathematical Statistics at the Institut für Mathematik of Potsdam University. He started to work as PhD student in 2014 at the Mathematical Institute of Leiden University under the supervision of Prof.dr. Aad van der Vaart, spending part time at the Vrije Universiteit medical center under the supervision of Prof.dr. Mark van de Wiel and dr. Wessel van Wieringen.