Prior information and variational Bayes in high dimensional statistical network inference

Kpogbezan, G.B.

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 $g = 1, \ldots, G$: 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 $i = 1, \ldots, p$ 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):

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

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 $a_1, b_1, \ldots, a_G, b_G, c, d$ for all $i = 1, \ldots, p$. These hyperparameters will be estimated from the combined data $(Y_1, X_1), \ldots, (Y_p, X_p)$ 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 ($i = 1, \ldots, s_i$). 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 $g = 1, \ldots, G$, 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 $\tau_{i,g}^2$. 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 $Y^1, \ldots, Y^n$ of p-vectors satisfying

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

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 $1, 2, \ldots, p$ 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 $Y^j = (Y_1^j, \ldots, Y_p^j)^T$, and next collecting the observations Y_i^j per gene i, giving the n-vector $Y_i = (Y_i^1, \ldots, Y_i^n)^T$, for $i = 1, \ldots, p$. 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: $P_{i,t} = 1$ if variable i is connected with variable t and $P_{i,t} = 0$ 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

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

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)

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

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)

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

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

Variational algorithm with sparse local-global shrinkage priors

1: **Initialize**
$a_g^{(0)} = b_g^{(0)} = 10^{-3}$, $g \in \{1, \cdots, G\}$ and $\forall i \in \mathcal{I}$, $b_{i,g}^* = d_i^* = 10^{-3}$, $\epsilon = 10^{-3}$, $\mathrm{M} = 10^3$ and $k = 1$

2: **while** $\max|\mathcal{L}_i^{(k)} - \mathcal{L}_i^{(k-1)}| \geq \epsilon$ **and** $2 \leq k \leq \mathrm{M}$ **do**

E-step: Update variational parameters

3: **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 $\mathcal{L}_i^{(k)}$; $\quad \forall g$ and $\forall t$ in that order
end for

M-step: Update hyperparameters

4: $a_g^{(k)}$, $b_g^{(k)}$; $\quad \forall g$

5: $k \leftarrow k + 1$

6: **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 $Y^1, \ldots, Y^n$ 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

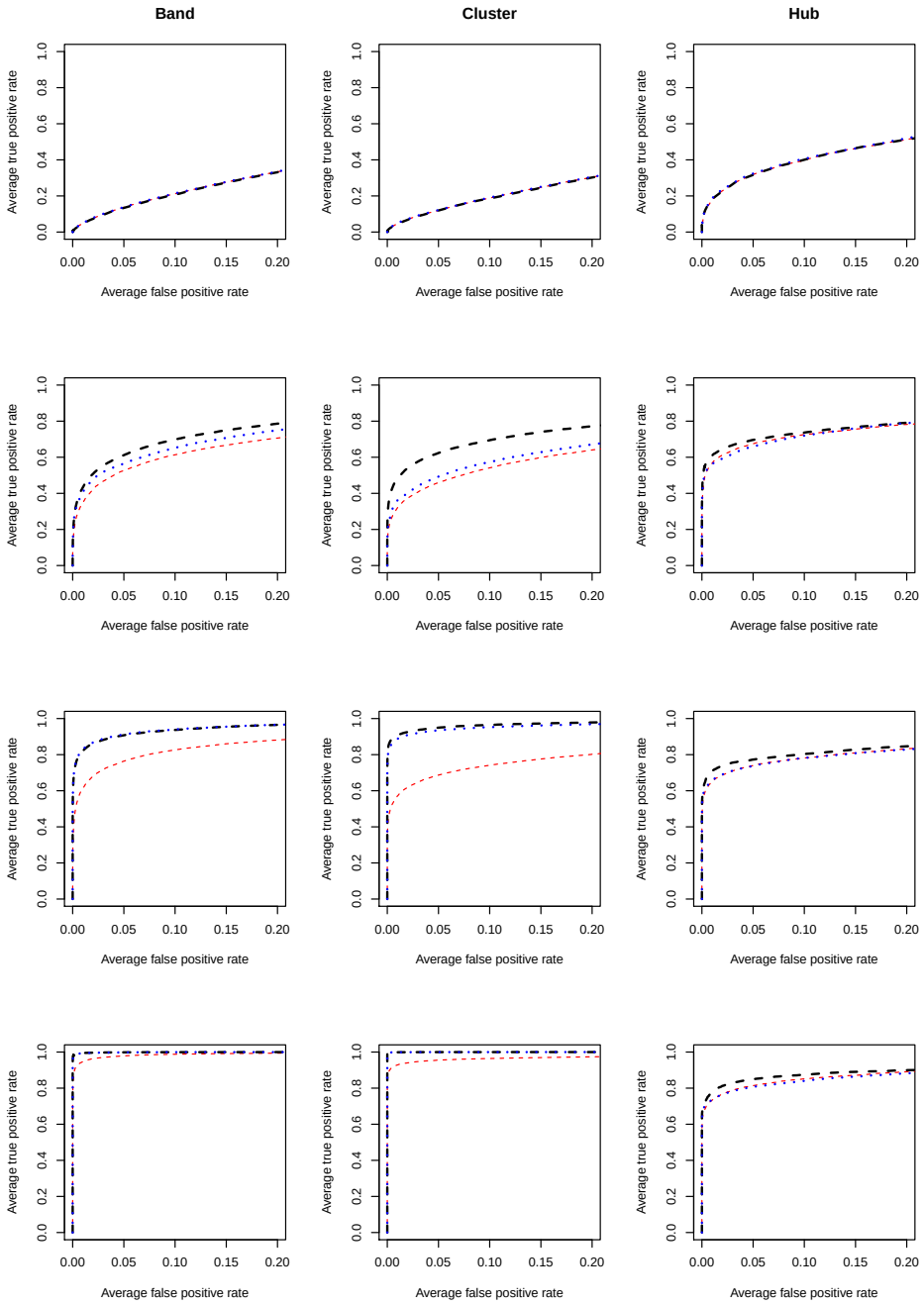

Figure 3.1: *Average partial-ROC curves comparing performance of ShrinkNet (dashed red), pInc2 (dashed black) and pInc (dashed blue) where* $n \in \{10, 100, 200, 500\}$ *and* $p = 100$*. First, second, third and fourth rows correspond respectively to the performances of* $n = 10$*,* $n = 100$*,* $n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

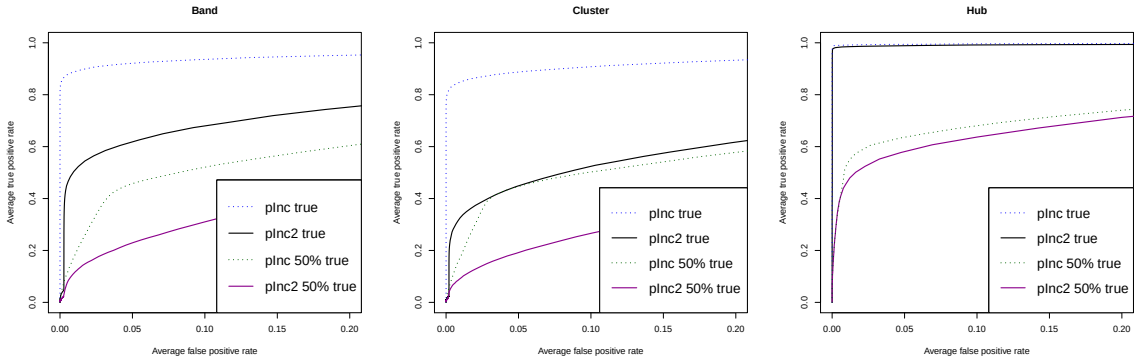

Figure 3.2: *Average partial-ROC curves comparing performance of pInc using perfect prior information (dashed blue), pInc2 using perfect prior information (black), pInc using* 50% *true edge information (dashed dark green) and pInc2 using* 50% *true edge information (darkmagenta). Sample size and network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

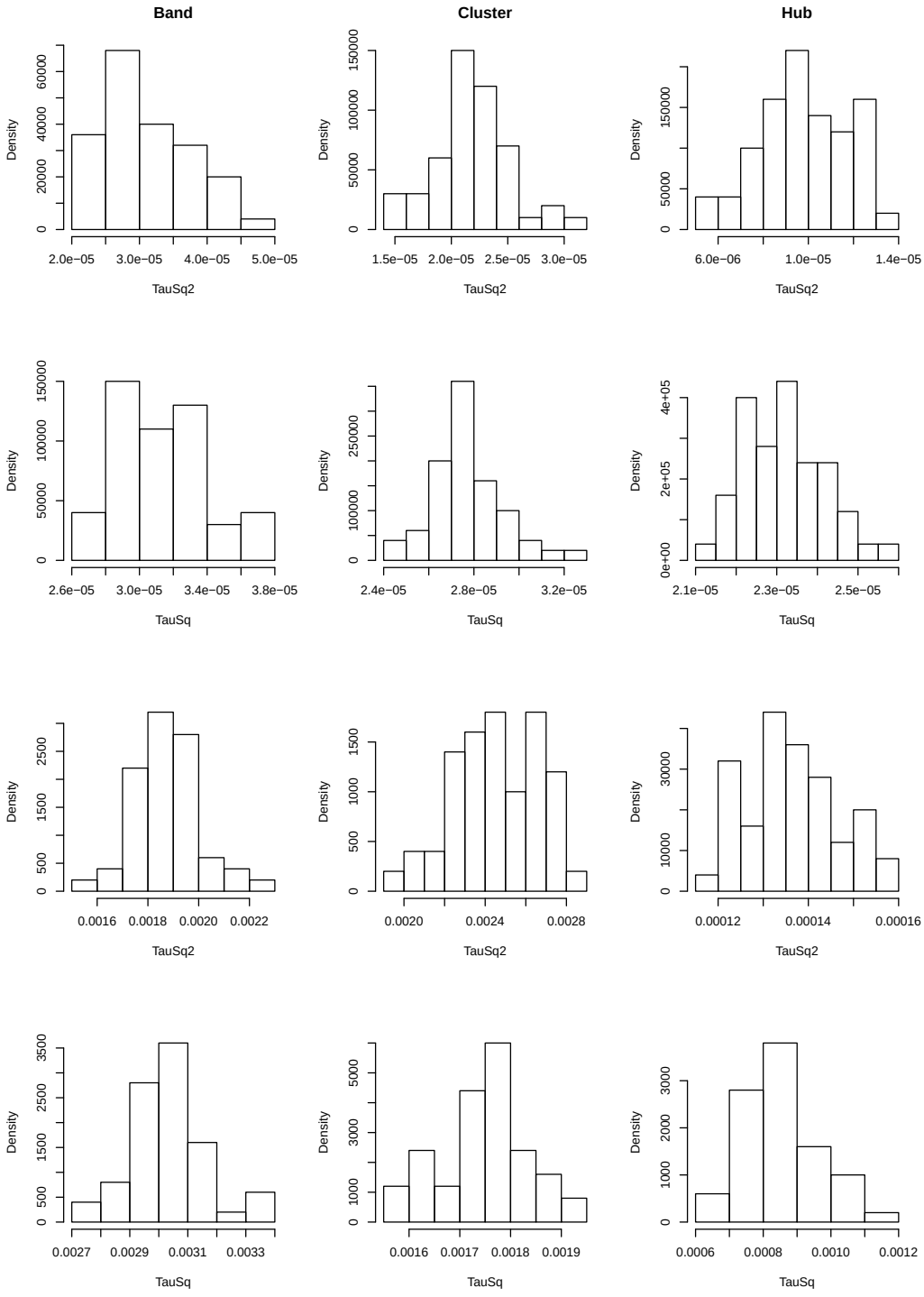

Figure 3.3: *Histograms of the global variance parameter τ^2 estimates by EB by pInc (TauSq) and by pInc2 (TauSq2) across* 50 *simulation replicates. First, second and third columns correspond respectively to Band, Cluster and Hub structures for the precision matrix. First row ($n = 10$) and third row ($n = 200$) 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 $\leftrightarrow$ FASLG, MYD88 $\leftrightarrow$ CSF2RB, PIK3R2 $\leftrightarrow$ CHUK, TNFRSF10B $\leftrightarrow$ CHP1, PRKAR1B $\leftrightarrow$ AKT2, PIK3R2 $\leftrightarrow$ NGF, TRAF2 $\leftrightarrow$ BAX, TNF $\leftrightarrow$ IL1B, PRKAR2B $\leftrightarrow$ AKT3, and TRAF2 $\leftrightarrow$ 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-

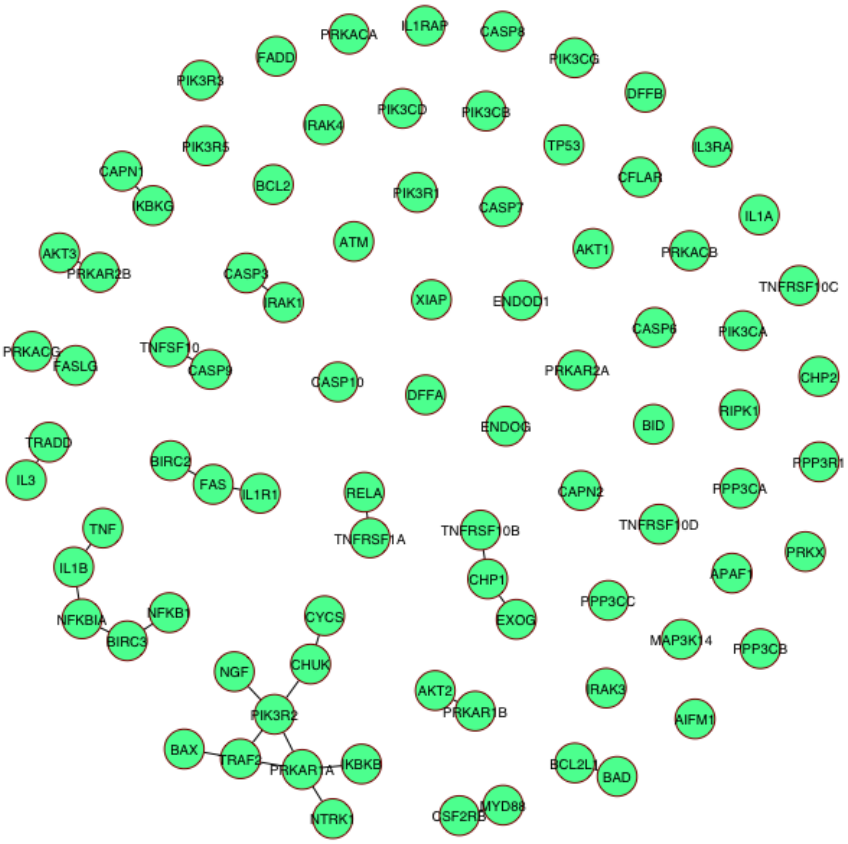

Figure 3.4: *Apoptosis network reconstructed for the 84 genes by pInc.*

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

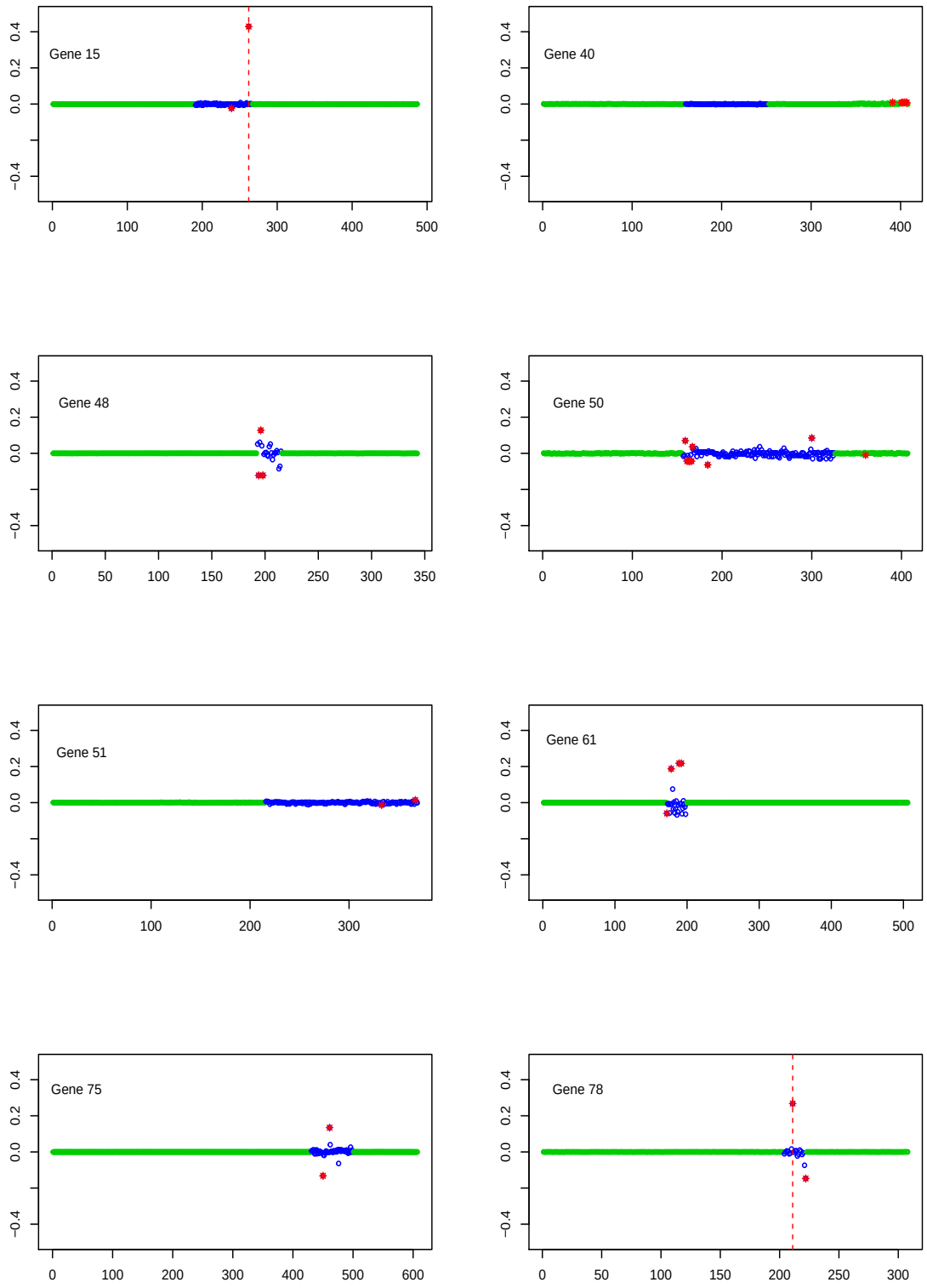

Figure 3.5: *Estimates of SNP effects on genes 15, 40, 48, 50, 51, 61, 75 and 78 using* pInc. *Green dots indicate effects estimates for SNPs outside the gene range and blue dots for SNPs inside the gene range. Red 'stars' indicate selected SNP effects. Dashed vertical lines indicate the 6 largest effects.*

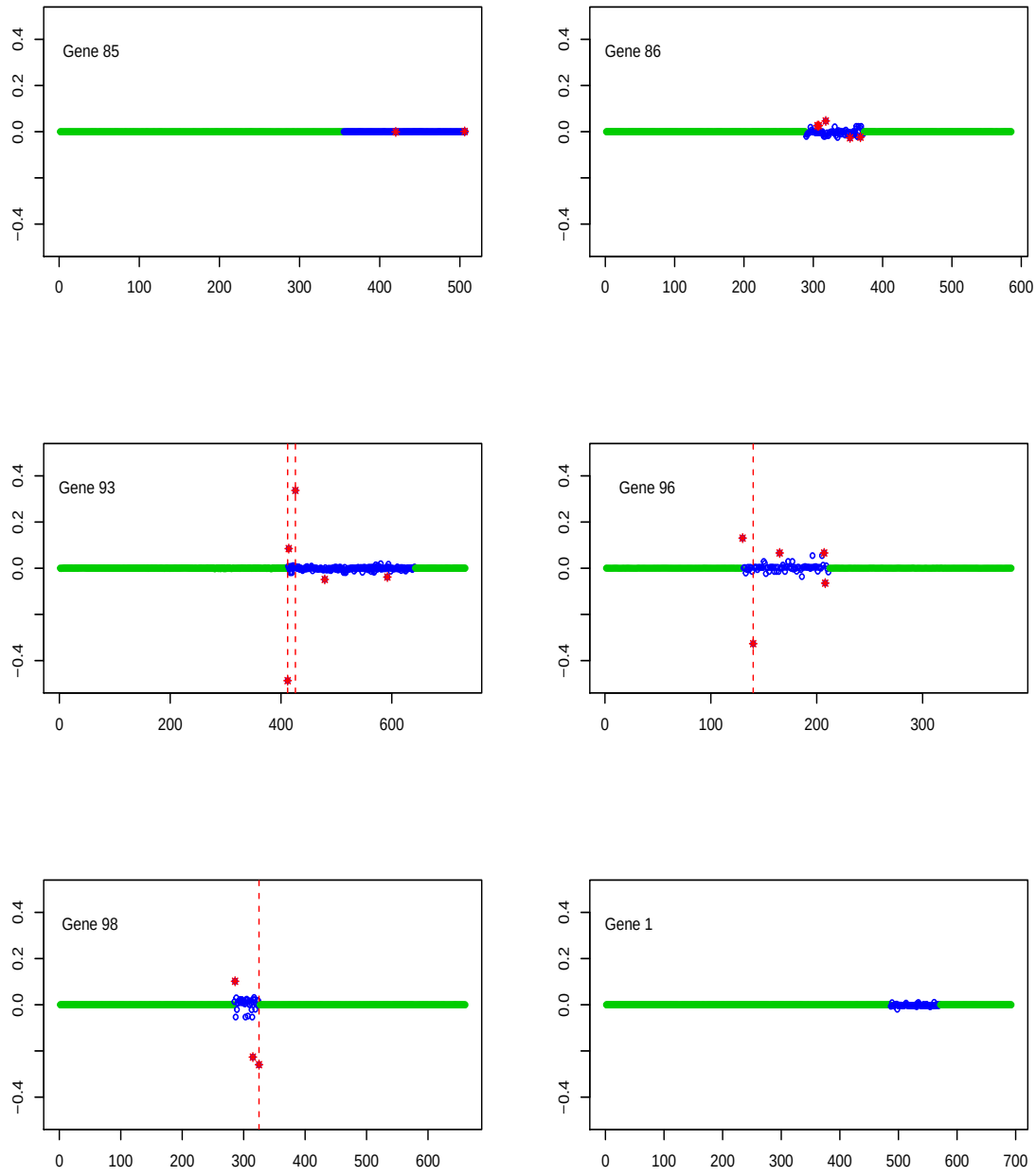

Figure 3.6: *Estimates of SNP effects on genes 75, 78, 85, 86, 93, 96, 98 and 1 using* pInc. *Green dots indicate effects estimates for SNPs outside the gene range and blue dots for SNPs inside the gene range. Red 'stars' indicate selected SNP effects. Dashed vertical lines indicate the 6 largest effects.*

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

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

where MSE_0 is the MSE of the empty model and $\text{MSE}(m_i)$ 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

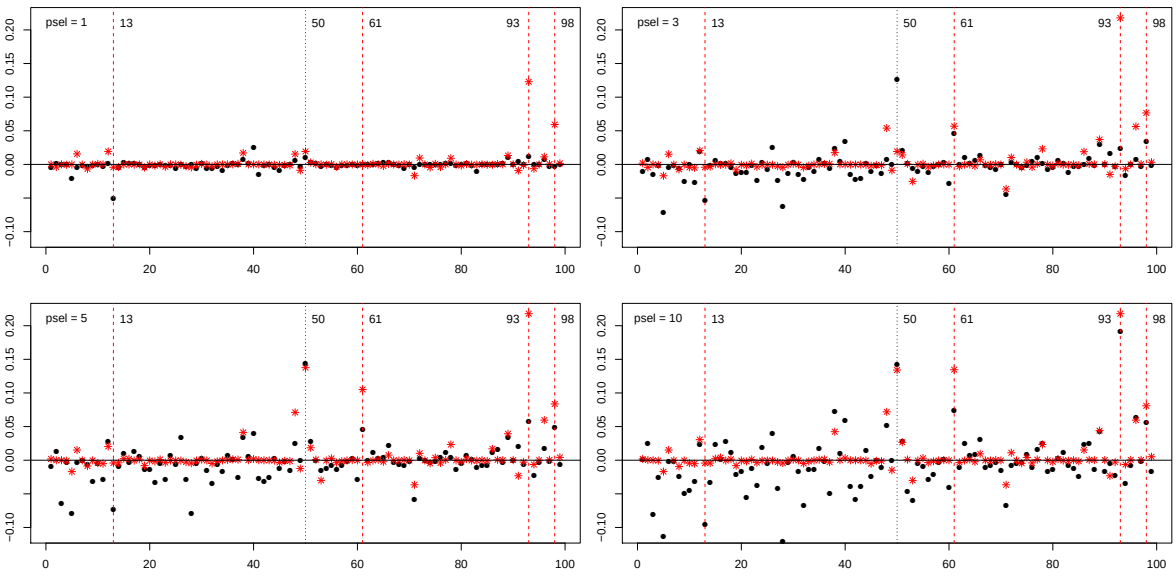

Figure 3.7: *Relative reduction of MSE (y-axis) for the lasso (black dots) and* pInc-DSS *(red stars) for all genes* $i = 1, \ldots, 99$ *(x-axis) when maximal number of variables is fixed to 1, 3, 5, or 10 (top-left, top-right, bottom-left, bottom-right). The genes with the large differences are highlighted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

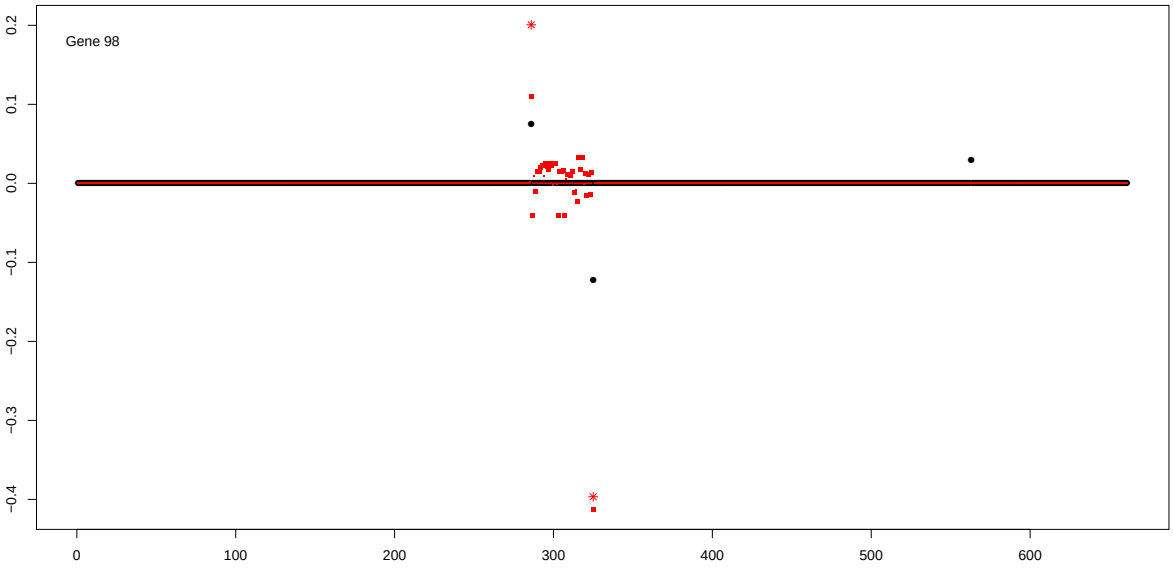

Figure 3.8: *Estimates of SNP effects on gene 98 using* pInc *(red squares), and* pInc-DSS *(red stars) and the lasso (black dots) with 3 predictor variables for the latter two. X-axis denotes SNP index.*

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:

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

Hence,

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

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):

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

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:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Therefore,

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

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:

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

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

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

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

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

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

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

simplifies to

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

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

$$
\alpha = (a_1, b_1, \cdots, a_G, b_G) \mapsto \sum_{i=1}^{p}\mathbf{E}_q\log\frac{p_\alpha(Y_i,\theta_i)}{q(\theta_i)} \tag{3.15}
$$

$$
= \sum_{i=1}^{p}\mathbf{E}_q\log\frac{p(Y_i|\,\theta_i)}{q(\theta_i)} + \sum_{i=1}^{p}\mathbf{E}_q\log p_\alpha(\theta_i). \tag{3.16}
$$

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 $\tau_{i,g}^2$ 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

$$a_g^* \approx \tfrac{1}{2}\Big(\log(a_g^*) - \Psi(a_g^*)\Big)^{-1} = \tfrac{1}{2}\Big(\frac{1}{p}\sum_{i=1}^{p}\Big(\log b_{i,g}^* - \Psi(a_{i,g}^*)\Big) - \log M\Big)^{-1}.$$