Bayesian Analysis (2014) **9**, Number 1, pp. 221–258

Incorporating Grouping Information in Bayesian Variable Selection with Applications in Genomics

Veronika Rockova [*] and Emmanuel Lesaffre [†]

Abstract. In many applications it is of interest to determine a limited number of important explanatory factors (representing groups of potentially overlapping predictors) rather than original predictor variables. The often imposed requirement that the clustered predictors should enter the model simultaneously may be limiting as not all the variables within a group need to be associated with the outcome. Within-group sparsity is often desirable as well. Here we propose a Bayesian variable selection method, which uses the grouping information as a means of introducing more equal competition to enter the model within the groups rather than as a source of strict regularization constraints. This is achieved within the context of Bayesian LASSO (least absolute shrinkage and selection operator) by allowing each regression coefficient to be penalized differentially and by considering an additional regression layer to relate individual penalty parameters to a group identification matrix. The proposed hierarchical model therefore enables inference simultaneously on two levels: (1) the regression layer for the continuous outcome in relation to the predictors and (2) the regression layer for the penalty parameters in relation to the grouping information. Both situations with overlapping and non-overlapping groups are applicable. The method does not assume within-group homogeneity across the regression coefficients, which is implicit in many structured penalized likelihood approaches. The smoothness here is enforced at the penalty level rather than within the regression coefficients. To enhance the potential of the proposed method we develop two rapid computational procedures based on the expectation maximization (EM) algorithm, which offer substantial time savings in applications where the high-dimensionality renders Markov chain Monte Carlo (MCMC) approaches less practical. We demonstrate the usefulness of our method in predicting time to death in glioblastoma patients using pathways of genes.

Keywords: Bayesian shrinkage estimation, EM algorithm, Bayesian LASSO, Minorization-maximization

1 Introduction

Rapid advances in the development of biomedical technologies have facilitated the availability of complex genomic data, which have continued posing significant challenges for statistical practitioners particularly because of their high dimensionality. Simultaneous

[*] Department of Biostatistics, Erasmus MC, Erasmus University Rotterdam v.rockova@erasmusmc.nl

[†] Department of Biostatistics, Erasmus MC, Erasmus University Rotterdam and L-BioStat KU Leuven e.lesaffre@erasmusmc.nl

DOI:10.1214/13-BA846

selection of genomic features associated with a clinical outcome as well as development of an interpretable prediction rule are commonplace in routine analysis of genomic data. Current statistical toolkits rely heavily on methodological developments in variable selection, among which the regularization approaches (Tibshirani 1994; Zou and Hastie 2005; Fan and Li 2001) have enjoyed particular attention. Despite the practical value of these approaches, one of their limitations is the inability to effectively utilize existing structural information about the predictors.

Modern genomic applications often deal with complicated covariate structures such as gene network topologies or partitions into groups, which may overlap. In cancer genomics, for example, DNA mutations are detected along the DNA sequence, where the location in the chromosome provides a linear ordering of the observations. It is reasonable to assume that adjacent measurements measure the same genetic effect and therefore should be grouped (Li and Zhang 2010). Gene expression data yield another example of a highly structured covariate space. Biologically related genes are known to form groups called pathways. Functional interactions between genes within/between pathways give rise to a gene interaction network, another type of structural information which has proven beneficial to incorporate in variable selection (Li and Li 2008).

Nowadays, many databases are available which store biological information from experimental research. These databases are continuously being updated with newly emerging information, providing a compendium of existing knowledge on how genes and gene products interact with each other. These interactions can be represented either as a network, where vertices represent genes/gene products and edges indicate a regulatory relationship, or as a list of pathway memberships. Existing databases of gene networks include among others the KEGG gene regulatory network (Kanehisa et al. 2002).

It is recognized that incorporation of the supplementary covariate information in the analysis of genomic data is key to more accurate prediction and improved interpretability of the results (Stingo et al. 2011; Pan et al. 2010). Several methods have been proposed that account for the gene network topology structures. Li and Li (2008) and Pan et al. (2010) proposed network-based penalties in linear regression, which induce both sparsity as well as smoothness of estimated effects within the pathways. These penalties have a Bayesian interpretation in that the prior on regression coefficients corresponds to the Gaussian conditional autoregressive model (Gelfand and Vounatsou 2003). Structural information among the predictors has been considered in the context of Bayesian variable selection by multiple authors including Li and Zhang (2010), Stingo and Vannucci (2011) and Stingo et al. (2011), who consider a Markov random field (MRF) prior on variable selection indicators with a neighboring structure defined by the network.

The limitation of MRF prior specification is that the effects of individual pathways cannot be separated from each other. The MRF network consists of multiple overlaying pathways, where the overlap makes it difficult to quantify the respective pathway contributions. It is often of interest to evaluate importance of pathways and simultaneously perform within-pathway gene selection. Recently, Stingo et al. (2011) proposed a partial least squares approach for pathway and gene selection using variable selection priors and Markov chain Monte Carlo (MCMC) for computation. In this paper we consider an alternative approach, which utilizes pathway membership information as a source of group-driven shrinkage. This is achieved within the context of Bayesian LASSO (least absolute shrinkage selection operator) (Park and Casella 2008), where individual penalty parameters are considered for each regression coefficient. An additional regression layer is then specified to relate these penalties to the grouping information. The motivation there is that penalties for coefficients within a group should share a common hyper-regression parameter, which puts the within-group coefficients on more equal footing in terms of penalization. These hyper-regression coefficients can be interpreted as “pathway effects”, which explain how the overall amount of penalization is distributed across the groups. The model extends the normal-exponential-gamma (NEG) prior of Griffin and Brown (2012) by embedding the grouping information in the prior distribution on the penalties to induce structured shrinkage. As opposed to the overlapping group LASSO approaches (Jacob et al. 2009), where either a whole group of predictors enters the model or is left out, here we rather introduce a more equal competition for genes within the same pathways to enter the model. As such, we let the likelihood of a variable to be selected be dependent on the pathway effects rather than its neighbors in the undirected graph. The estimated pathway effects then quantify respective pathway importance, adding to the biological interpretability. Group sparsity can be enforced through priors on the pathway weights, where the posterior serves a prerequisite for performing variable selection in a hierarchical manner by first selecting pathways and then selecting genes within the pathways.

An important point of contrast between our method and the penalized regression approaches for structured variable selection such as group LASSO (Yuan and Lin 2006) or Markov random field models on regression coefficients (Pan et al. 2010; Li and Li 2008) is that the latter two enforce smoothness in the regression coefficients rather than in the penalty parameters. This discrepancy may have important practical implications in situations where there is no reason to assume homogeneity in regression coefficients within groups or between neighbors in the graph.

We also investigate asymptotic implications of rescaling the NEG shrinkage prior by a factor dependent on the sample size and consider an alternative formulation of the model, which guarantees a non-vanishing penalization effect. We show that the maximum a posteriori (MAP) estimator in the rescaled model possesses the oracle property (Fan and Li 2001) and demonstrate its satisfactory finite sample size performance on simulated examples.

The implementations of Bayesian methods for shrinkage estimation have relied heavily on the developments of MCMC strategies, which may suffer from high computational time requirements when the cardinality of the predictor space is large. In this paper we consider an alternative computational strategy, the maximum a posteriori estimation based on the expectation maximization (EM) algorithm. We build on work done previously by Griffin and Brown (2012) and we extend their algorithm by including structural grouping information. Similar to Armagan et al. (2012) we present two versions of the algorithm, the first one based on iteratively solving ridge regression, while the other one is based on LASSO (Tibshirani 1994). The two algorithms are seen to converge rapidly even in situations where the number of predictors p greatly exceeds the number of observations n.

The outline of the paper is as follows. Section 2 describes the background notation and motivation for our method. Sections 3 and 4 deal with model formulation and computation. Section 5 is devoted to a brief discussion on the properties of the NEG prior and selection of tuning parameters. In Section 6, a simple implementation is demonstrated on simulated data. Application to the data is presented in Section 7. Finally, Section 8 concludes the paper with a brief discussion.

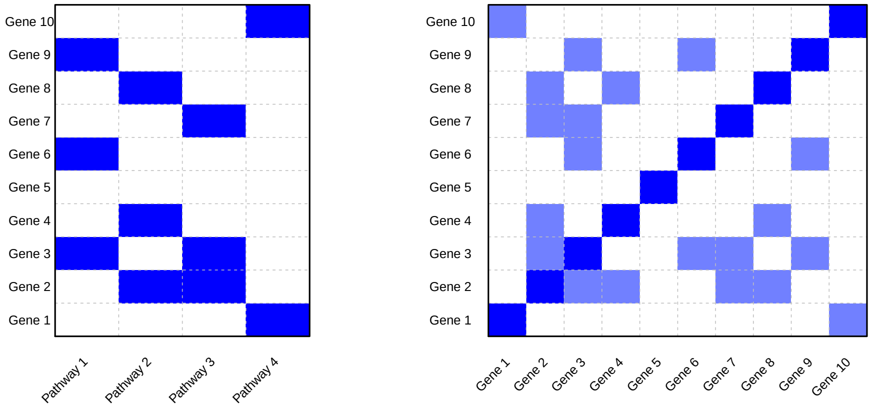
2 The Method

Consider the canonical multiple linear regression setting, where the $(n \times 1)$ vector of centered responses $\boldsymbol{Y}$ is linked to the $(n \times p)$ matrix $\boldsymbol{X}$ of standardized regressors (mean zero and variance one) through the relation $\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{X\beta}, \sigma^2 \mathrm{I}_n)$, where $\boldsymbol{\beta}$ denotes the $(p \times 1)$ vector of regression coefficients and σ^2 is an unknown scalar. We focus on the "large p small n" situation arising often in genomic and proteomic studies, where the number of predictors greatly exceeds the number of observations. The regression vector $\boldsymbol{\beta}$ is believed to be sparse in that only a small subset of predictors contributes to explaining the variability of the response. Apart from the sparsity requirement, we wish to impose additional regularization constraints as dictated by available prior knowledge about the structure among predictors.

The key ingredient in the model formulation is the introduction of a design/loading matrix $\boldsymbol{Z}$ $(p \times q)$ consisting of q columns of dummy variables coding for group membership. Given that the predictors form a network structure attributable to the existence of few shared underlying factors, the involvement of genes within each of the q factors/pathways can be encoded through a pattern of zeroes in the loading matrix $\boldsymbol{Z}$. Here we assume that the number of the latent factors as well as patterns of zeros in the loading matrix can be retrieved from external scientific knowledge.

An illustrative example for 10 genes and 4 pathways is depicted in Figure 1(a), where colored fields indicate functional gene-pathway relationships. Assuming that two genes are related if and only if they share at least one underlying pathway, we obtain an undirected graphical structure characterized as a set of edges $\mathcal{E} = \{(i,j) : 1 \leq i < j \leq p\}$, where $(i,j) \in \mathcal{E}$ whenever X_i is a neighbor of X_j. Such a structure can be represented by a symmetric $p \times p$ matrix $M = (m_{ij})_{i,j=1}^{p,p}$, where $m_{ij} \neq 0$ whenever $(i,j) \in \mathcal{E}$. The zero patterns in matrix M are depicted for our simple example in Figure 1(b). It is easy to see that the off-diagonal elements in M copy the pattern of zeroes in the matrix $\boldsymbol{ZZ}'$. This undirected graph however assumes that all the genes within the pathway are connected. In reality, the genes within pathways are far more sparsely connected and it is the union of these small network components that gives rise to a gene association network.

Assume that the k-th pathway is assigned a weight coefficient b_k, which summarizes its activity. In order to induce simultaneous shrinkage of coefficients sharing the same underlying pathways we let the likelihood of a gene to be selected depend on a combination



(a) Pathway loading matrix

(b) Undirected graph

Figure 1: Loading matrix and undirected graph representations of gene interactions

of the active pathway effects. In our simple example, for instance, Gene 2 is involved in the activity of Pathway 2 and Pathway 3 and therefore the degree of shrinkage of β_2 towards zero is affected by the combination of the pathway weights b_2 and b_3. In case there are singletons, which do not belong to any pathway, such as Gene 5 in our example, we consider an additional shared parameter b_0, which controls the overall sparsity for all genes. In the following paragraph we put down a mathematical formulation for this mechanism.

3 Model Formulation

We consider the problem of Bayesian shrinkage estimation in structured high-dimensional covariate spaces. Our proposal extends the Normal-Exponential-Gamma (NEG) prior of Griffin and Brown (2012) by embedding the structural covariate information (encoded in $\boldsymbol{Z}$) within the sparsity inducing regularization. The model formulation is as follows:

$$\begin{aligned}
\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2 &\sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\mathrm{I}_n),\\
\beta_j|\sigma^2,\tau_j &\overset{\text{ind}}{\sim} \mathrm{N}(0,\sigma^2\tau_j^2),\\
\tau_1^2,\dots,\tau_p^2|\lambda_1^2,\dots,\lambda_p^2 &\sim \prod_{j=1}^{p}\lambda_j^2\exp(-\lambda_j^2\tau_j^2)\mathrm{I}(\tau_j>0),\\
\lambda_j^2|\boldsymbol{b} &\overset{\text{ind}}{\sim} \Gamma\left[a,h(\boldsymbol{Z}_j'\boldsymbol{b})\right],\\
b_l &\overset{\text{ind}}{\sim} \pi(\boldsymbol{\theta}),\ l=0,\dots,q,\\
\sigma^2 &\sim \mathrm{IGamma}(c,d),
\end{aligned}$$

where $\boldsymbol{Z}_j$ denotes the j-th row of the $p \times (q+1)$ matrix $(\mathbf{1}_p, \boldsymbol{Z})$ and $\Gamma(a, b)$ (resp. $\mathtt{IGamma}(a, b)$) denotes the gamma (resp. inverse gamma) distribution with shape a and scale b. The regression coefficients arise from the conditional Laplace distribution (expressed as a scale mixture of normals), given the variance σ^2 and a vector of penalty parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$. An important ingredient in this formulation is the conjugacy, where including the variance σ^2 in the prior for regression coefficients yields nice analytical simplifications in the derivation of the EM algorithm. Furthermore, it guarantees the unimodality of the joint conditional posterior distribution $\pi(\boldsymbol{\beta}, \sigma^2 | \boldsymbol{y}, \boldsymbol{\lambda})$ (as shown by Park and Casella (2008)), which may better mitigate the local mode problems associated with the EM algorithm. As opposed to the Bayesian LASSO model (Park and Casella 2008), where only one common penalty is used to regularize all the coefficients, we allow unique parameters for each individual coefficient by analogy with the adaptive LASSO (Zou 2006). Griffin and Brown (2012) further suggest imposing a gamma hyper-prior distribution on the coefficient-specific penalties with fixed shape and scale. Here we go a step further and assume that the scale parameter is random and varies from coefficient to coefficient.

More specifically, we assume an additional regression layer in the hierarchy to relate the penalty parameters to the matrix $\boldsymbol{Z}$. Each λ_j is independently assigned a gamma distribution with expected value $\mathsf{E}\lambda_j = a\, h(\boldsymbol{Z}_j' \boldsymbol{b})$, where coefficients $\boldsymbol{b} = (b_0, \ldots, b_q)'$ are unknown and subject to estimation. The intercept b_0 can be regarded as a global shrinkage hyper-parameter determining the baseline level of shrinkage. The individual regression coefficients are then locally influenced by the remaining coefficients in the linear predictor $\boldsymbol{Z}_j' \boldsymbol{b}$.

Assume for a moment that $\boldsymbol{Z}$ encodes for q non-overlapping groups, i.e. $\{1, \ldots, p\} = \bigcup_{k=1}^{q} \mathcal{Q}_k$, where $\mathcal{Q}_k \cap \mathcal{Q}_l = \emptyset$ for $k \neq l$. Then, $\forall j \in \mathcal{Q}_k$ we have $\mathsf{E}\lambda_j = ah(b_0 + b_k)$. The parameter b_k hence quantifies the additional amount of shrinkage attributable to the k-th group and puts the within-group coefficients on more equal footing in terms of penalization. For overlapping groups, the shape parameter is an additive summary of the weights for all active pathways, i.e. $\mathsf{E}\lambda_j = a\, h(b_0 + \sum_{k=1}^{q} \mathrm{I}[j \in \mathcal{Q}_k] b_k)$.

Various link functions $h(\cdot)$ can be considered in the hierarchical formulation. However, in order to interpret the higher values b_k as more evidence for pathway importance, we need to consider a link function decreasing in $\boldsymbol{b}$, such as an inverse or an inverse exponential link function. The choice of the link function has implications for the selection of appropriate prior distributions $\pi(\theta)$. We are not necessarily restricted to the conjugate class of priors, which would be a natural candidate for posterior sampling in the generalized linear model (GLM) setting (Chen and Ibrahim 2003). The (inverse) exponential link functions slow down the convergence of the EM algorithm, therefore we consider only inverse and identity links with pathway weights restricted to be positive. Since for a fixed shape parameter a, the gamma distribution is conjugate for the rate parameter $1/s$ in $\Gamma(a, s)$, we opt for independent gamma priors $\Gamma(\alpha, 1/\gamma)$ on the elements of $\boldsymbol{b}$ in the inverse link and for inverse gamma priors $\mathtt{IGamma}(\alpha, 1/\gamma)$ in the identity link. The hyper-parameters α and γ can be tuned according to the expected degree of group "sparsity". In the inverse link, we might want to assure sufficient spread over a wider range of values in situations when many groups are assumed predictive. Other choices for α and γ would be more appropriate if the solution is expected to be group "sparse", in which case the prior $\Gamma(\alpha, 1/\gamma)$ should accumulate more density on pathway weights close to zero.

Finally, the weights b_k summarize the relevance of the respective groups/pathways, when related to clinical outcomes. In gene networks, predictive disease co-regulation patterns can be found by locating high-evidence pathways, as determined by the magnitude of these pathway weights.

A similar prior construction was considered by Stingo et al. (2010), who proposed a hierarchical Bayesian graphical model for microRNA targets, where the prior probability of variable inclusion is related to a linear combination of external association scores through a logistic regression formulation.

4 EM Algorithm for NEG Prior with External Covariate Information

The practicality of implementation is one of the most important aspects when analyzing high-dimensional data. In this regard, MCMC algorithms for Bayesian shrinkage estimation have become increasingly computationally cumbersome as the number of covariates has escalated. Several authors have considered alternative strategies based on the EM algorithm (Figueiredo 2003; Kiiveri 2003; Griffin and Brown 2012). To adapt these to our situation, we have the additional difficulty of estimating the pathway weights $\boldsymbol{b}$, which requires extensions of existing approaches.

In the EM algorithm, modified for Bayesian modal estimation, the logarithm of the incomplete data likelihood is augmented by the logarithm of the prior density (McLachlan and Krishnan 1996, p. 26). The incompleteness here refers to unobserved latent variables rather than missing observations. The MAP estimates are then values that maximize the so called log-incomplete data posterior density, here $\log p(\boldsymbol{\beta}, \boldsymbol{b}, \sigma^2 \mid \boldsymbol{y})$. These values are obtained by iteratively maximizing the conditional expectation of the log complete posterior $\log p(\boldsymbol{\beta}, \boldsymbol{b}, \sigma^2, \boldsymbol{w} \mid \boldsymbol{y})$ with respect to the conditional distribution of the latent variables $\boldsymbol{w}$ given the current parameter estimates and the observed data.

Since the parameters $\boldsymbol{\beta}, \boldsymbol{b}$ and σ are of interest, the candidates for the latent unobserved data are either $\boldsymbol{\tau}^2$ and $\boldsymbol{\lambda}^2$. Instead of assuming that both $\boldsymbol{\tau}^2$ and $\boldsymbol{\lambda}^2$ are missing, we integrate out either one of the two sets of parameters from the model. This leads to nice simplifications, as will become clearer later on. Similarly as in Armagan et al. (2012), we consider two variants. First, we integrate over the penalty parameters $\boldsymbol{\lambda}^2$ and treat the latent variances $\boldsymbol{\tau}^2$ as missing. This formulation exploits the normal-scale mixture representation of the NEG prior. In the second version, we integrate over $\boldsymbol{\tau}^2$ and treat the penalty parameters $\boldsymbol{\lambda}$ as missing, which corresponds to the Laplace prior formulation. We will see that the latter possesses many convenient properties, such as a naturally sparse representation and flexibility in the selection of the link function $h(\cdot)$.

4.1 EM Algorithm Using the Normal Mixture Representation

The E-step of the algorithm entails the computation of the conditional expectation of the log complete posterior distribution given the observed data and current values $\boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}$ at the k-th iteration. This objective function, which is to be maximized in the subsequent M-step, takes the following form:

$$\begin{aligned} Q\left(\boldsymbol{\beta}, \boldsymbol{b}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) &= \mathsf{E}_{\boldsymbol{\tau}^2}\left[\log p(\boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{\tau}^2 \mid \boldsymbol{y}) \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}, \boldsymbol{y}\right] \\ &= C + Q_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) + Q_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right), \end{aligned}$$

where

$$\begin{aligned} Q_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) &= -\frac{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{j=1}^{p} \beta_j^2 \, \mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot}\left(\frac{1}{\tau_j^2}\right) \\ &\quad - \frac{n + p + 2c + 2}{2}\log(\sigma^2) - \frac{d}{\sigma^2}, \\ Q_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) &= \sum_{j=1}^{p}\left\{\log[a\, h(\boldsymbol{Z}_j'\boldsymbol{b})] - (a+1)\,\mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot}\log[1 + \tau_j^2 h(\boldsymbol{Z}_j'\boldsymbol{b})]\right\} \\ &\quad + \sum_{l=0}^{q}\log \pi(b_l) \end{aligned}$$

and $\mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot}(\cdot)$ denotes the conditional expectation $\mathsf{E}_{\boldsymbol{\tau}^2}\left(\cdot \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}, \boldsymbol{y}\right)$.

As a result of our hierarchical prior formulation, where coefficients $\boldsymbol{\beta}$ depend on the coefficients $\boldsymbol{b}$ only through the penalty parameters $\boldsymbol{\lambda}$, the objective function $Q(\cdot)$ is separable with respect to $\boldsymbol{b}$ and $(\boldsymbol{\beta}, \sigma)'$. This implies that the M-step can be performed by separately maximizing each of the functions $Q_1(\cdot)$ and $Q_2(\cdot)$. It is worth noting that $Q_1(\cdot)$ corresponds to the log-posterior distribution resulting from a Bayesian ridge regression, whose maximum can be expressed analytically. The maximization of $Q_2(\cdot)$ with respect $\boldsymbol{b}$ is complicated by the unavailability of the conditional expectation $\mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot}\log[1 + \tau_j^2 h(\boldsymbol{Z}_j'\boldsymbol{b})]$ in closed form. This problem could be circumvented by approximating the integral either analytically or using MCMC methods. However, this would impose an additional computational burden and we do not elaborate on such alternatives further. In the following paragraph we show how to maximize this function without approximations, assuming the identity link function $h(\boldsymbol{Z}'\boldsymbol{b}) = \boldsymbol{Z}'\boldsymbol{b}$. Recall that for the identity link we use independent inverse gamma priors on the elements of $\boldsymbol{b}$, i.e. $\log \pi(b) = -(\alpha + 1)\log b - \gamma/b + \texttt{const}$.

In the spirit of a generalized EM algorithm (Dempster et al. 1977), instead of finding the value that globally maximizes the function $Q_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right)$ we choose $\boldsymbol{b}^{(k+1)}$ such that

$$Q_2\left(\boldsymbol{b}^{(k+1)} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) \geq Q_2\left(\boldsymbol{b}^{(k)} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right). \tag{1}$$

Such a condition on $\boldsymbol{b}^{(k+1)}$ is sufficient to guarantee the monotonicity property, i.e. the incomplete data log posterior distribution is not decreased after the k-th iteration. The update $\boldsymbol{b}^{(k+1)}$ that satisfies property (1) can be found by maximizing a surrogate minorizing function, the definition of which is given below.

Definition 4.1. *Let $\boldsymbol{b}^{(k)} \in \mathrm{D} \subset \mathbb{R}^{q+1}$ represent a fixed value of the parameter vector $\boldsymbol{b}$ and let $f(\boldsymbol{b}; \boldsymbol{b}^{(k)})$ denote a real-valued function. Then $f(\boldsymbol{b}; \boldsymbol{b}^{(k)})$ is said to be minorizing a real valued function $g(\boldsymbol{b})$ at $\boldsymbol{b}^{(k)}$ in domain D if and only if*

$$\begin{aligned} f(\boldsymbol{b}; \boldsymbol{b}^{(k)}) &\leq g(\boldsymbol{b}), \quad \forall \boldsymbol{b} \in D, \\ f(\boldsymbol{b}^{(k)}; \boldsymbol{b}^{(k)}) &= g(\boldsymbol{b}^{(k)}). \end{aligned}$$

From the definition of the minorizing function, it easily follows that $g(\boldsymbol{b}^{(k+1)}) \geq g(\boldsymbol{b}^{(k)})$, where $\boldsymbol{b}^{(k+1)}$ maximizes the surrogate function $f(\boldsymbol{b}; \boldsymbol{b}^{(k)})$. The question remains how to construct a suitable minorizing function for $Q_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right)$. The answer is given by the following theorem.

Theorem 4.1. *Let $\boldsymbol{b}^{(k)} \in \mathbb{R}_+^{q+1}$ represent a fixed value of the parameter vector $\boldsymbol{b}$. Denote*

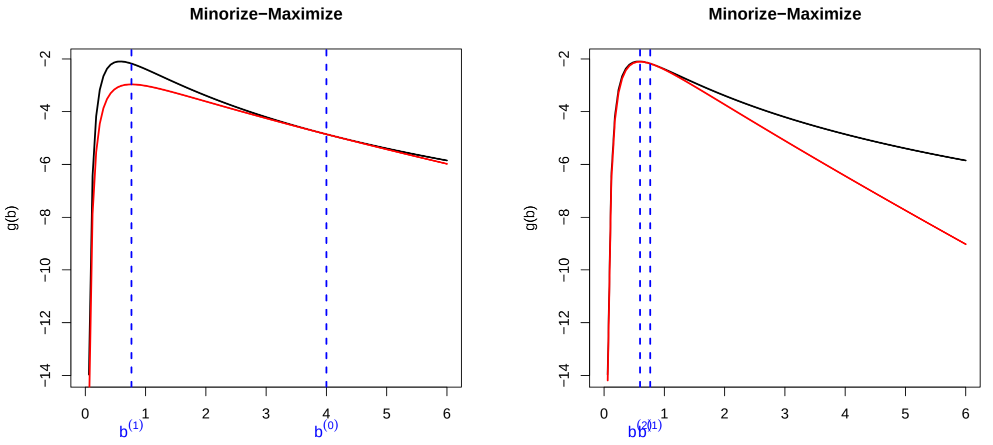
$$\begin{aligned} M_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) = &\sum_{j=1}^{p} \left[\log(a\, \boldsymbol{Z}_j' \boldsymbol{b}) - (a+1) \mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot} \left(\frac{\tau_j^2}{1 + \tau_j^2 \boldsymbol{Z}_j' \boldsymbol{b}^{(k)}}\right) \boldsymbol{Z}_j'(\boldsymbol{b} - \boldsymbol{b}^{(k)})\right] \\ &- \sum_{j=1}^{p} (a+1)\, \mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot} \log[1 + \tau_j^2 \boldsymbol{Z}_j' \boldsymbol{b}^{(k)}] + \sum_{l=0}^{q} [-(\alpha+1)\log b_l - \gamma/b_l]. \end{aligned}$$

Then the function $M_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right)$ minorizes $Q_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right)$ at $\boldsymbol{b}^{(k)}$ in $\mathbb{R}_+^{q+1}$.

Proof. For $j \in \{1, \ldots, p\}$ denote $g_j(\boldsymbol{b}) = -(a+1)\log(1 + \tau_j^2 \boldsymbol{Z}_j' \boldsymbol{b})$. Each of the functions $g_j(\boldsymbol{b})$ is convex in $\mathbb{R}_+^{q+1}$ (i.e. the function $g_j^*(t) = g_j(\boldsymbol{b} + t\boldsymbol{c})$ is convex $\forall \boldsymbol{b}, \boldsymbol{c} \in \mathbb{R}_+^{q+1}$ and $\forall t \in \mathbb{R}$ such that $\boldsymbol{b} + t\boldsymbol{c}$ is in the domain of $g_j(\cdot)$). The convexity implies that the first order Taylor approximation at the point $\boldsymbol{b}^{(k)}$ is a global underestimator of the function $g_j(\cdot)$. The fact that $\mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\, X \geq \mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\, Y$, whenever $X \geq Y$ *a.s.* completes the proof.

Several observations can be made based on the result of Theorem 4.1. First, the minorizing function $M_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right)$ no longer entails the evaluation of an integral which depends on the unknown parameter values $\boldsymbol{b}$. All the integrals in the minorizing functions depend only on the current parameter values $\boldsymbol{b}^{(k)}$. Furthermore, the cumbersome expectation $\mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot} \log(1 + \tau_j^2 \boldsymbol{Z}_j' \boldsymbol{b}^{(k)})$ does not need to be computed, as the summand involving this term does not depend on $\boldsymbol{b}$. Second, the values maximizing the minorizing function can be regarded as MAP estimates in a Bayesian regression with exponentially distributed responses $(a+1)/a\, \mathsf{E}_{\boldsymbol{\tau}^2 \mid \cdot} \left(\frac{\tau_j^2}{1+\tau_j^2 \boldsymbol{Z}_j' \boldsymbol{b}^{(k)}}\right)$, which are related to the regression matrix $a\boldsymbol{Z}$ via an inverse link function, and where the regression coefficients $\boldsymbol{b}$ are assumed to be independently gamma distributed. Third and most importantly, the expectations $\mathsf{E}_{\boldsymbol{\tau}^2\,|\,\cdot}\left(\frac{\tau_j^2}{1+\tau_j^2 \boldsymbol{Z}_j' \boldsymbol{b}^{(k)}}\right)$ can be expressed analytically using hypergeometric cofluent functions (Gradshteyn and Ryzhik 2000, p. 278).

The graphical representation of the "minorization-maximization" (MM) algorithm is given in Figure 2. The black curve corresponds to the function $g(b) = -\log(b) - 2\log(1+b) - 1/b$, which depicts the behavior of the function $Q_2(\cdot)$ for $p = q = a = \alpha = \gamma = 1$ assuming that $\tau_1 = 1$ almost surely and $\boldsymbol{Z}_1 = 1$. We have the initial estimate $b^{(0)} = 4$, at which we construct the minorizing function $f(b; b^{(0)})$ according to Theorem 4.1 (depicted by the red curve). This function has its maximum at the value $b^{(1)} = 0.76$. Repeating the minorization-maximization at the new point $b^{(1)}$ (Figure 2(b)), we obtain a surrogate function $f(b; b^{(1)})$, whose maximum $b^{(2)} = 0.59$ lies in the close vicinity of the true global maximum $\widehat{b} = 0.57$ of the function $g(b)$.



(a) Minorization-maximization iteration 1

(b) Minorization-maximization iteration 2

Figure 2: Graphical representation of the minorization-maximization algorithm

Unfortunately, this convenient computation can be only applied for the identity link function. Considering either inverse, exponential, or inverse exponential link functions, we lose the convexity property of the function $\log[1 + \tau_j^2 h(\boldsymbol{Z}'\boldsymbol{b})]$, which is necessary to assure the monotonicity of the update based on the Taylor expanded surrogate function.

Summary of the EM Algorithm Using the Normal Mixture Representation

The parameters are initialized with starting values $\boldsymbol{\beta}^{(0)}, \boldsymbol{b}^{(0)}, \sigma^{(0)}$. The steps described below are then repeated until a convergence criterion is satisfied (e.g. $|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}|_{l_1} + |\boldsymbol{b}^{(k+1)} - \boldsymbol{b}^{(k)}|_{l_1} < \varepsilon$).

E-step

In the E-step, we first evaluate the conditional expectations $\mathsf{E}_{\boldsymbol{\tau}^2\,|\,\cdot}\left(\frac{1}{\tau_j^2}\right)$. Following Griffin and Brown (2012), we obtain (proof in Appendix A)

$$\mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\left(\frac{1}{\tau_j^2}\right) = \frac{2(a+0.5)\sigma^{(k)}\sqrt{\boldsymbol{Z}_j'\boldsymbol{b}^{(k)}}}{|\beta_j|^{(k)}} \frac{D_{-2(a+1)}\left(\frac{|\beta_j^{(k)}|\sqrt{\boldsymbol{Z}'_j\boldsymbol{b}^{(k)}}}{\sigma^{(k)}}\right)}{D_{-2(a+0.5)}\left(\frac{|\beta_j^{(k)}|\sqrt{\boldsymbol{Z}'_j\boldsymbol{b}^{(k)}}}{\sigma^{(k)}}\right)}, \tag{2}$$

where $D_\eta(x)$ denotes the parabolic cylinder function (Gradshteyn and Ryzhik 2000, p. 256). We then denote $\boldsymbol{\Omega}^{(k)} = \operatorname{diag}\left[\mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\left(1/\tau_j^2\right)\right]_{j=1}^p$ the diagonal matrix with the entries (2) on the diagonal. Next, we compute $\mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\left(\frac{\tau_j^2}{1+\tau_j^2\boldsymbol{Z}'_j\boldsymbol{b}^{(k)}}\right)$ for $j = 1,\ldots,p$ and we stack the values in a $p \times 1$ vector $\boldsymbol{\Lambda}^{(k)}$. We obtain (proof in Appendix B)

$$\mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\left[\frac{\tau_j^2}{1+\tau_j^2\boldsymbol{Z}'_j\boldsymbol{b}^{(k)}}\right] = \frac{a\,\Gamma\left(a+0.5\right)}{\sigma^{(k)}\sqrt{2\pi\boldsymbol{Z}'_j\boldsymbol{b}^{(k)}}}\boldsymbol{\Psi}\left(a+0.5, -\frac{1}{2}; \frac{\beta_j^{(k)2}\boldsymbol{Z}'_j\boldsymbol{b}^{(k)}}{2\sigma^{(k)2}}\right)\frac{1}{p(\beta_j^{(k)}|\boldsymbol{b}^{(k)},\sigma^{(k)})}, \tag{3}$$

where $\boldsymbol{\Psi}(a, b; x)$ denotes the hypergeometric cofluent function (Gradshteyn and Ryzhik 2000, p. 543).

M-step

The value $\boldsymbol{\beta}^{(k+1)}$ which globally maximizes $Q_1\left(\boldsymbol{\beta},\sigma \mid \boldsymbol{\beta}^{(k)},\boldsymbol{b}^{(k)},\sigma^{(k)}\right)$ can be regarded as a solution to the ridge regression problem

$$\boldsymbol{\beta}^{(k+1)} = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\{|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}|_{l_2}+|\boldsymbol{\Omega}^{(k)1/2}\boldsymbol{\beta}|_{l_2}\}, \tag{4}$$

where $\boldsymbol{\Omega}^{(k)1/2}$ denotes the square root of the matrix $\boldsymbol{\Omega}^{(k)}$. The computation of the closed form solution $(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{\Omega}^{(k)})^{-1}\boldsymbol{X}'\boldsymbol{Y}$ can be, for $p > n$, facilitated by utilizing the Sherman-Morrison-Woodbury formula (Golub and van Loan 1996), which requires the inversion of only an $n \times n$ matrix. The variance is updated as $\sigma^{2(k+1)} = (|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k+1)}|_{l_2}+|\boldsymbol{\Omega}^{(k)1/2}\boldsymbol{\beta}^{(k+1)}|_{l_2}+2d)/(n+p+2c+2)$. Finally, the pathway weights are updated according to Theorem (4.1) as values maximizing the function $M_2(\boldsymbol{b}|\boldsymbol{\beta}^{(k)},\boldsymbol{b}^{(k)},\sigma^{(k)})$. Keeping only the summands in $M_2(\cdot)$, which depend on $\boldsymbol{b}$, we obtain $\boldsymbol{b}^{(k+1)}$ as

$$\boldsymbol{b}^{(k+1)} = \operatorname{argmax}_{\boldsymbol{b}\in\mathbb{R}_+^{q+1}}\sum_{j=1}^p\left[\log(a\,\boldsymbol{Z}'_j\boldsymbol{b}) - (a+1)\Lambda_j^{(k)}\boldsymbol{Z}'_j\boldsymbol{b}\right] + \sum_{l=0}^q[-(\alpha+1)\log b_l - \gamma/b_l], \tag{5}$$

which is a box-constrained optimization problem solvable using optimization routines implemented in standard packages (`optimize` function in R).

This EM algorithm corresponds to the algorithm of Zou and Li (2008) for the computation of penalized likelihood estimates with nonconvex penalties, using the local quadratic approximation to the penalty function.

4.2 EM Algorithm Using the Laplace Representation

The ease of the computation of the normal-mixture-based algorithm applies only for the identity link function. The difficulty in using the identity link is the interpretability of the pathway weights $\boldsymbol{b}$, where small values indicate more evidence for the importance of the pathway. Another limitation is the inability to estimate the coefficients directly at zero, due to the ridge regression updates. Fan and Li (2001) suggested that if $\beta_j^{(k)}$ is very close to zero, say $|\beta_j^{(k)}| < \varepsilon$, then the MAP estimate is set $\widehat{\beta}_j = 0$ and the j-th component is removed from the next iteration. The drawback of this approach is that once deleted, the covariate is ultimately excluded from the model. Moreover, the selection threshold ε, which determines the sparsity of the solution, can be regarded as an additional parameter, which requires tuning. Similarly to Armagan et al. (2012), we consider an alternative version of the EM algorithm, which benefits from the LASSO rather than ridge regression solutions and therefore produces a naturally sparse solution without unnecessary thresholding. Furthermore, it allows for richer choices of the link functions.

In the previous version of the EM algorithm, we integrated over the penalty parameters $\boldsymbol{\lambda}^2$ and treated the latent variances $\boldsymbol{\tau}^2$ as missing data. Now we do exactly the opposite, we integrate over $\boldsymbol{\tau}^2$ and treat the penalties $\boldsymbol{\lambda}^2$ as missing.

The objective function, i.e. the conditional expectation of the complete log posterior distribution given the observed data and current values $\boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}$ and $\sigma^{(k)}$ at the k-th iteration now corresponds to:

$$
\begin{aligned}
\widetilde{Q}\left(\boldsymbol{\beta}, \boldsymbol{b}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) &= \mathsf{E}_{\boldsymbol{\lambda}^2}\left[\log p(\boldsymbol{\beta}, \boldsymbol{b}, \sigma, \boldsymbol{\lambda}^2 \mid \boldsymbol{y}) \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}, \boldsymbol{y}\right] \\
&= C + \widetilde{Q}_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) + \widetilde{Q}_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\widetilde{Q}_1\left(\boldsymbol{\beta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) &= -\frac{(\boldsymbol{Y} - \boldsymbol{X\beta})'(\boldsymbol{Y} - \boldsymbol{X\beta})}{2\sigma^2} - \frac{\sqrt{2}}{\sigma}\sum_{j=1}^{p} |\beta_j| \mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_j \\
&\quad - \frac{n+p+2c+2}{2}\log(\sigma^2) - \frac{d}{\sigma^2}, \\
\widetilde{Q}_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right) &= \sum_{j=1}^{p}\left[-a \log h(\boldsymbol{Z}_j'\boldsymbol{b}) - \frac{\mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_j^2}{h(\boldsymbol{Z}_j'\boldsymbol{b})}\right] + \sum_{l=0}^{q}[(\alpha-1)\log b_l - \gamma b_l]
\end{aligned}
$$

and $\mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}(\cdot)$ denotes the conditional expectation $\mathsf{E}_{\boldsymbol{\lambda}^2}\left(\cdot \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}, \boldsymbol{y}\right)$.

The expected log complete posterior distribution is again separable with respect to $\boldsymbol{b}$ and $(\boldsymbol{\beta}, \sigma)'$. In contrast to the previous version of the EM algorithm, the coefficients

$\boldsymbol{\beta}^{(k+1)}$ at the k-th iteration solve the "adaptive" LASSO problem, where differential penalties are considered for each regression coefficient. This algorithm relates to the algorithm of Zou and Li (2008) for the computation of nonconcave penalized likelihood problems using the local linear approximation to the penalty function.

Summary of the EM Algorithm Using the Laplace Representation

The parameters are initialized with starting values $\boldsymbol{\beta}^{(0)}, \boldsymbol{b}^{(0)}, \sigma^{(0)}$. The steps described below are then repeated until a convergence criterion is satisfied (e.g. $|\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}|_{l_1} + |\boldsymbol{b}^{(k+1)} - \boldsymbol{b}^{(k)}|_{l_1} < \varepsilon$).

E-step

The E-step entails the calculation of $\mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_j$ and $\mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_j^2$, which can be evaluated using known functions (proof in Appendix C). For $s = 1, 2$, we have

$$\mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_j^s = \frac{[h(\boldsymbol{Z}_j'\boldsymbol{b}^{(k)})]^{(s+1)/2}}{\sigma^{(k)}\Gamma(a)2^{a+s/2}} \exp\left(\frac{\beta_j^{(k)2} h(\boldsymbol{Z}_j'\boldsymbol{b}^{(k)})}{4\sigma^{(k)2}}\right) D_{-(2a+1+s)}\left(\frac{|\beta_j|\sqrt{h(\boldsymbol{Z}_j'\boldsymbol{b}^{(k)})}}{\sigma^{(k)}}\right). \tag{6}$$

M-step

In the M-step, we begin with the update $\boldsymbol{\beta}^{(k+1)}$, a solution to the problem

$$\boldsymbol{\beta}^{(k+1)} = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\{|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}|_{l_2} + 2\sqrt{2}\sigma^{(k)}|\boldsymbol{D}^{(k)}\boldsymbol{\beta}|_{l_1}\},$$

where $\boldsymbol{D}^{(k)} = \operatorname{diag}\left[\mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_1, \ldots, \mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_p\right]$. The solution can be obtained easily after reweighing the regression matrix and applying standard LASSO computation (Zou 2006). The M-step continues by updating $\sigma^{(k+1)}$ according to

$$\sigma^{(k+1)} = \frac{\sqrt{2}|\boldsymbol{D}^{(k)}\boldsymbol{\beta}^{(k+1)}|_{l_1} + \sqrt{2(|\boldsymbol{D}^{(k)}\boldsymbol{\beta}^{(k+1)}|_{l_1})^2 + 4(|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^{(k+1)}|_{l_2} + 2d)(n+p+2c+2)}}{n+p+2c+2}.$$

Finally, the updates $\boldsymbol{b}^{(k+1)} = \operatorname{argmax}_{\boldsymbol{b}\in\mathbb{R}^{q+1}}\widetilde{\mathcal{Q}}_2\left(\boldsymbol{b} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{b}^{(k)}, \sigma^{(k)}\right)$ can be computed using box-constrained optimization routines. Assuming $a = 1$, this function corresponds to the log posterior for Bayesian regression with exponentially distributed variables $\mathsf{E}_{\boldsymbol{\lambda}|\cdot}\lambda_j$, which are related to the regression matrix $\boldsymbol{Z}$ through the $h(\cdot)$ link function, assuming independent gamma distributed priors on the regression coefficients $\boldsymbol{b}$.

5 Hierarchical Variable Selection

A natural strategy for variable selection based on the posterior output $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{b}}, \widehat{\sigma})$ is by screening out variables with a zero estimated (or negligible) regression coefficient $\widehat{\beta}$. As an alternative practical guidance for selecting variables, we suggest proceeding hierarchically from the top of the hierarchical model to the bottom. In the first step, we select relevant pathways. This is achieved by disregarding groups with pathway weights $\widehat{b}$ that are estimated at the zero boundary of the parameter space (or are negligibly small). Given that the weights correlate with the proportion of relevant genes within each pathway (simulated study in Appendix D) it will often be sensible to ignore all the genes within the non-predictive pathways. The second step then proceeds by selecting only from variables that are located in the predictive groups. This selection can be anchored by either thresholding or identification of zeroes in the vector of posterior estimates $\widehat{\boldsymbol{\beta}}$, depending on which version of the EM algorithm has been used. This recommended strategy in our simulated examples leads to a dramatic reduction of false discoveries.

6 Some Properties of the NEG Prior

The hierarchical prior construction introduced in Section 2.1 differs from the original formulation of the NEG prior (Griffin and Brown 2012) in the assumption that the scale parameter (further denoted as s) in the gamma prior density $\Gamma(a, s)$ is unknown and subject to estimation. In this section, we discuss some of the properties of the NEG prior in relation to the choice of the shape and scale hyper-parameters. Recall that the NEG distribution has the following density function (Griffin and Brown (2012)):

$$p_{a,s,\sigma}(\beta) = \frac{a\,2^a\sqrt{s}}{\sqrt{\pi\sigma^2}}\Gamma\left(a+0.5\right)\exp\left(\frac{\beta_j^2 s}{4\sigma^2}\right)D_{-2(a+0.5)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right). \tag{7}$$

The shape parameter a controls the heaviness of the tails, where the prior density becomes more peaked and lighter tailed with increased a, which may cause unwanted bias in estimation of large effects. Decreasing the scale parameter, the density (7) becomes flatter, losing the ability to shrink noise signals due to a less pronounced peak at zero. With both a and $1/s$ approaching zero, we obtain the Normal-Jeffreys limiting case (Griffin and Brown 2012). With both a and $1/s$ approaching infinity at the same rate, the density converges to the Laplace prior. This property is formally summarized by the following theorem.

Theorem 6.1. *Let $p_{a,s,\sigma}(\beta)$ denote the density function in (7). Then for $0 < s/a = \lambda' < \infty$ we have $\lim_{a\to\infty} p_{a,s,\sigma}(\beta) = \frac{\sqrt{\lambda'}}{2\sigma}\exp(-\sqrt{\lambda'}|\beta|/\sigma)$.*

Proof. Let us consider the characteristic function of the $\Gamma(a, s)$ distribution $\psi(t) = (1 - \mathrm{i}ts)^{-a}$, where $\mathrm{i}^2 = -1$. Since $s = \frac{\lambda'}{a}$, we have $\forall t \in \mathbb{R}$

$$\lim_{a\to\infty}\left[1 - \frac{\mathrm{i}t\lambda'}{a}\right]^{-a} = \lim_{a\to\infty}\exp\left[a\log\left(1 + \frac{\mathrm{i}t\lambda'}{a - \mathrm{i}t\lambda'}\right)\right] = \exp(\mathrm{i}t\lambda'),$$

which follows from l'Hospital rule. The limit is the characteristic function of a Dirac distribution concentrated at λ'. Denote by $p_{a,s}(\lambda^2)$ the gamma density function with shape a and scale s. Then $\lim_{a\to\infty} p_{a,\lambda'/a}(\lambda^2) = \delta_{\lambda'}(\lambda^2)$. This altogether gives

$$\lim_{a\to\infty}\int_{\lambda^2}\int_{\tau^2} p(\beta \mid \sigma,\tau^2)p(\tau^2 \mid \lambda^2)p_{a,\lambda'/a}(\lambda^2)\mathrm{d}\,\tau^2\mathrm{d}\,\lambda^2 = \int_{\tau^2} p(\beta \mid \sigma,\tau^2)p(\tau^2 \mid \lambda')\mathrm{d}\,\tau^2$$
$$= \frac{\sqrt{\lambda'}}{2\sigma}\exp(-\sqrt{\lambda'}|\beta_j|/\sigma),$$

which is the density of the Laplace distribution. Switching the limit and integral signs is justified by the bounded convergence theorem and noting that $p_{a,\lambda'/a}(\lambda^2) < \lambda'$ for all $a > 1$.

Remark 6.1. *A similar bridging property between the Laplace and Normal-Jeffreys priors has been observed for the Generalized Double Pareto distribution (Armagan et al. 2012).*

To gain more insights about the properties of the NEG prior, we consider for a moment a simple normal mean situation, i.e. $\boldsymbol{Y}|\boldsymbol{\beta},\sigma^2 \sim \mathrm{N}(\boldsymbol{\beta}, œ^2)$ and $\beta_j|\tau_j^2,\sigma \sim \mathrm{N}(0,\sigma^2\tau_j^2), j = 1,\ldots,n$. According to Fan and Li (2001), a sufficient condition for the unbiasedness of the MAP estimator is that $\pi_{a,s,\sigma}(|\beta_j|) = 0$ for large $|\beta_j|$, where $\pi_{a,s,\sigma}(|\beta_j|) = \frac{\partial \log p_{a,s,\sigma}(|\beta_j|)}{\partial|\beta_j|}$ and $p_{a,s,\sigma}(\cdot)$ denotes the marginal prior distribution (7). As given in Griffin and Brown (2012),

$$\pi_{a,s,\sigma}(|\beta_j|) = \frac{(2a+1)\sqrt{s}}{\sigma}\frac{D_{-2(a+1)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)}{D_{-2(a+0.5)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)}. \tag{8}$$

It is desirable that $\pi_{a,s,\sigma}(|\beta_k|)$ approaches zero rapidly as $|\beta_k|\to\infty$ to avoid unnecessary modeling bias. The asymptotic properties of the bias term are summarized by the following theorem.

Theorem 6.2. *For $\pi_{a,s,\sigma}(|\beta|)$ in (8), $\pi_{a,s,\sigma}(|\beta|) = \mathcal{O}\left(\frac{1}{|\beta|}\right)$ as $|\beta|\to\infty$.*

Proof. The limiting behavior of the term $\pi'_{a,s,\sigma}(|\beta|)$ can be better understood using the Poincare expansion of the Parabolic cylinder function for large $|\beta|$ (Gradshteyn and Ryzhik 2000, p. 1016), namely

$$D_\eta(x) \sim \exp(-x^2/4)x^\eta\left(1 - \frac{\eta(\eta-1)}{2x^2} + \frac{\eta(\eta-1)(\eta-2)(\eta-3)}{2.4x^4} - \ldots\right) \tag{9}$$

where the symbol $\sim$ indicates that the Parabolic cylinder function is equal to the series in the limit as $|x|\to\infty$. As a consequence, we have

$$\lim_{|x|\to\infty}\frac{D_\eta(x)}{\exp\left(-\frac{x^2}{4}\right)x^\eta} = 1.$$

This altogether enables us to rewrite $\lim_{|\beta|\to\infty} \pi'_{a,s,\sigma}(|\beta|)$ as

$$\lim_{|\beta|\to\infty} \frac{(2a+1)\sqrt{s}}{\sigma} \frac{\exp\left(-\frac{\beta_i^2 s}{4\sigma^2}\right)\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)^{-2(a+1)}}{\exp\left(-\frac{\beta_i^2 s}{4\sigma^2}\right)\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)^{-2(a+0.5)}} = \lim_{|\beta|\to\infty} \frac{2a+1}{|\beta|},$$

which was to be demonstrated.

Remark 6.2. *The bias hence decreases less rapidly for higher values of the shape parameter a, which is expected since a determines the heaviness of the tails.*

In order to better understand how the choice of a and s affects the shrinkage properties of the NEG prior, we investigated the behavior of the "shrinkage factor" $\kappa_j = \frac{1}{1+\tau_j^2}$. In the conjugate normal means model, this random coefficient determines how much shrinkage towards zero is put on the regression coefficient β_j once we have observed the data (Carvalho and Polson 2010). The interpretation follows from the identity $\mathsf{E}(\beta_j|y_j,\tau_j^2) = (1-\kappa_j)y_j$, which marginally becomes $\mathsf{E}(\beta_j \mid y_j, \sigma^2) = [1-\mathsf{E}(\kappa_j \mid y_j, \sigma^2)]y_j$. The shape of the prior distribution $p(\kappa_j)$ indicates how much shrinkage is to be expected a priori. Inspecting the prior density of the NEG shrinkage factor

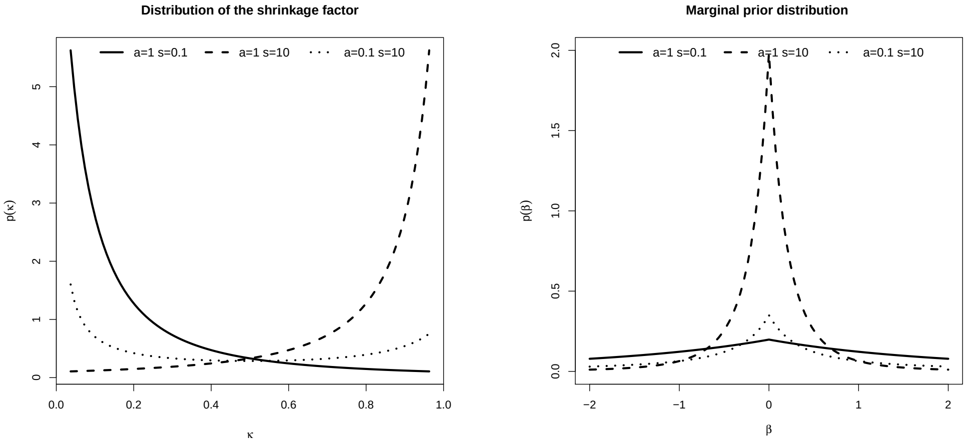
$$p_{a,s}(\kappa_j) = \frac{as}{\kappa_j^2}\left[1 + s\left(\frac{1-\kappa_j}{\kappa_j}\right)\right]^{-a-1}$$

for various choices of shape and scale parameters (Figure 3(a)) gives us an idea how the two parameters affect the ability of the NEG prior to distinguish between signal and noise. Increasing the shape parameter a for fixed s, the distribution $p(\kappa_j)$ concentrates more densely around one, implying that the NEG prior is more aggressive in shrinking small noise-like signals towards zero. A similar effect can be achieved by increasing the scale parameter s for fixed a. Decreasing the shape parameter a, more probability mass is accumulated near zero, which in turn induces heavier tails of the NEG prior. It is possible to select a configuration of the two parameters, which induces a "horseshoe-like" shape, where both tail robustness and ability to shrink noise are retained simultaneously (Carvalho and Polson 2010). The corresponding prior densities for the regression coefficients assuming $\sigma^2 = 1$ are depicted on Figure 3(b).

The delicate interplay between the hyper-parameters a and s in determining the shrinkage characteristics of the NEG prior is further complicated by the presence of the unknown global variance parameter σ^2. This parameter affects the posterior distribution of the shrinkage factor

$$p(\kappa_j \mid y_j, \sigma^2) = \frac{\sqrt{\kappa_j}}{\sigma} \exp\left(-\frac{y_j^2 \kappa_j}{2\sigma^2}\right) p_{a,s}(\kappa_j),$$

where small values σ^2 distribute more posterior mass on κ_j's near zero. The consequence is that small σ^2 may cause under-shrinkage of noise.



(a) Prior distribution of shrinkage factor

(b) Marginal prior distributions $p_{a,s,1}(\beta_k)$

Figure 3: Prior distribution of shrinkage factor and regression coefficients

In the context of multiple linear regression, the small fixed values of σ^2 may increase the number of false positives. In our EM algorithm, small values $\sigma^{(k)}$ at the k-th iteration imply smaller penalties on the regression coefficients (as seen from equation (2)) and thereby increased likelihood of false discoveries. This may be problematic in high-dimensional settings ($p > n$), where variance estimates at each iteration are typically very small. Possible remedies for this problem are: (a) to consider higher values of the shape parameter a, (b) to specify an informative prior on the variance, such as a flat prior within an interval bounded away from zero, (c) to add a fixed multiplying factor g to the prior variance $\mathsf{Var}\,(\beta_j|\sigma^2, \tau_j) = g\,\tau_j^2\sigma^2$. The parameter g resembles the hyperparameter in the g-prior (Liang et al. 2008), but its role is fundamentally different. Zellner (1986) and other authors have recommended treating g as a function of sample size to prevent the g-prior from asymptotically dominating the likelihood. Whereas in the g-prior context it is desirable for g to grow with n, we will see that the NEG prior benefits from letting g decrease with n in order to achieve a non-vanishing penalization effect.

Multiplying the prior variance of the regression coefficient by the factor g is equivalent to imposing the NEG prior with shape a and scale s/g. In the following theorem we show that considering $g = 1/n^2$ guarantees, for suitably chosen scale parameters s, variable selection consistency and asymptotic normality of the MAP estimator under mild regularity conditions for multiple regression with fixed p. For simplicity we will assume that σ is fixed to one and let the scale parameter s vary according to the sample size.

Theorem 6.3. *Assume the regularity conditions (A)-(C) in Fan and Li (2001) and denote by $\widehat{\boldsymbol{\beta}}_n$ the MAP estimator arising from the hierarchical model under the NEG*$(a, n^2 s_n)$

prior. Let $\mathcal{A}_n = \{j : \widehat{\beta}_j \neq 0\}$ *and* $\mathcal{A} = \{j : \beta_j \neq 0\}$*, where* $\boldsymbol{\beta}$ *is the true coefficient vector. Then for* $s_n \to 0$ *and* $\sqrt{n}s_n \to \infty$ *as* $n \to \infty$ *the MAP estimator* $\widehat{\boldsymbol{\beta}}_n$ *satisfies:*

(a) Consistency in variable selection: $\lim_{n\to\infty} \mathsf{P}(\mathcal{A}_n = \mathcal{A}) = 1$*,*

(b) Asymptotic normality: $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}_n} - \boldsymbol{\beta}_{\mathcal{A}}) \to \mathrm{N}(\mathbf{0}, \boldsymbol{I}_{\mathcal{A}^{-1}})$*, where* $\boldsymbol{\beta}_{\mathcal{A}}$ *denotes the nonzero elements in* $\boldsymbol{\beta}$ *and* $\boldsymbol{I}_{\mathcal{A}}$ *is the Fisher information knowing* $\beta_j = 0$ *for* $j \notin \mathcal{A}$*.*

Proof. The MAP estimate $\widehat{\boldsymbol{\beta}}_n$ under the $\mathrm{NEG}(a, n^2 s_n)$ prior can be regarded as the coefficient vector minimizing the penalized least squares

$$\frac{1}{2}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + n\sum_{j=1}^{p} pen_{a,s_n}(|\beta_j|),$$

where the penalty term consists of the summands in negative $\mathrm{NEG}(a, n^2 s_n)$ density, which depend on $|\beta|$, divided by n. According to (7), the penalty term for $\sigma^2 = 1$ takes the following form:

$$pen_{a,s_n}(|\beta|) = -\frac{|\beta|^2 n s_n}{4} - \frac{1}{n}\log D_{-2(a+0.5)}\left(|\beta| n\sqrt{s_n}\right). \qquad (10)$$

Denote by $pen'_{a,s_n}(|\beta|)$ and $pen''_{a,s_n}(|\beta|)$ the first and second derivatives of (10) with respect to $|\beta|$. In order to demonstrate asymptotical normality and consistency, it suffices to show that the penalty function satisfies the following three conditions (Fan and Li 2001):

(a) $\lim_{n\to\infty} pen'_{a,s_n}(|\beta|) = 0$ for all $\beta \neq 0$,

(b) $\lim_{n\to\infty} pen''_{a,s_n}(|\beta|) = 0$ for all $\beta \neq 0$,

(c) $\liminf_{n\to\infty} \liminf_{\beta\to 0+} pen'_{a,s_n}(|\beta|)/s_n > 0$.

The property (a) follows from the asymptotic expansion of the Parabolic cylinder function, which gives that $\forall\beta \neq 0$ and for $n\sqrt{s_n} \to \infty$ as $n \to \infty$ (which follows from the assumption $\sqrt{n}s_n \to \infty$)

$$\lim_{n\to\infty} pen'_{a,s_n}(|\beta|) = \lim_{n\to\infty} (2a+1)\sqrt{s_n}\frac{D_{-2(a+1)}\left(|\beta| n\sqrt{s_n}\right)}{D_{-2(a+0.5)}\left(|\beta| n\sqrt{s_n}\right)} = \lim_{n\to\infty} \frac{2a+1}{n|\beta|} = 0.$$

In order to show the validity of condition (b) it is helpful to reexpress the derivatives of the Parabolic cylinder function using the recursion formulas (Abramowitz and Stegun 1972, p.688). After some algebra we obtain the following expression for the second derivative of the penalty function:

$$\begin{aligned} pen''_{a,s_n}(|\beta|) =& n^2 s_n\sqrt{s_n}(2a+1)|\beta|\frac{D_{-2(a+1)}\left(|\beta| n\sqrt{s_n}\right)}{D_{-2(a+0.5)}\left(|\beta| n\sqrt{s_n}\right)} \\ & - n s_n(2a+1) + n s_n(2a+1)^2\left(\frac{D_{-2(a+1)}\left(|\beta| n\sqrt{s_n}\right)}{D_{-2(a+0.5)}\left(|\beta| n\sqrt{s_n}\right)}\right)^2. \end{aligned}$$

Applying again the Poincare asymptotic expansion we conclude that as $n \to \infty$: (a) the third summand in $pen''_{a,s_n}(|\beta|)$ is asymptotically $o(n)$, and (b) the first summand is asymptotically equivalent to $ns_n(2a+1)$. This altogether implies that the limit of $pen''_{a,s_n}(|\beta|)$ is zero as n grows to infinity.

In order to verify the last condition it is helpful to note that $D_{-\eta-1/2}(0) = \sqrt{\pi}\frac{2^{-\eta/2-1/4}}{\Gamma(3/4+\eta/2)}$ (Abramowitz and Stegun 1972, p.687). Then for $s_n \to 0$ as $n \to \infty$ we have

$$\liminf_{n\to\infty} \liminf_{\beta\to 0+} pen'_{a,s_n}(|\beta|)/s_n = \liminf_{n\to\infty} \frac{(2a+1)\Gamma(a+1)\sqrt{2}}{\Gamma(a+1.5)\sqrt{s_n}} > 0.$$

Remark 6.3. *In the penalized likelihood setting with a diverging number of parameters the "oracle" properties of the NEG penalty (without scaling) were shown in Griffin and Brown (2012). Here we considered a modified penalized likelihood function, which corresponds to an actual posterior distribution in the hierarchical Bayesian context.*

Remark 6.4. *Instead of tuning the prior as a function of sample size, Ishwaran and Rao (2005) suggest an alternative way to avoid vanishing effect of the prior in spike and slab models by rescaling the responses by a factor $\sqrt{n}$ and adding a variance inflation factor.*

Remark 6.5. *Fan and Li (2001) suggest a sandwich standard error formula for the non-zero penalized likelihood estimates, which can be applied also for the MAP coefficients arising from the rescaled NEG prior in Theorem 6.3.*

7 Simulated Examples

The purpose of this section is to illustrate the application of the proposed method on two simulated examples and to demonstrate its potential as a variable selection tool. In the first example, the predictors are assumed to cluster within known non-overlapping groups, whereas the second example deals with the overlapping case. Throughout the section we assume that the number of predictors p is much larger than the number of observations n, whereas the number of informative predictors is smaller than n. The estimation is in both examples conducted using the Laplace version of the EM algorithm with an inverse link function. The threshold for convergence ε is set to 10^{-5}.

7.1 Non-overlapping Groups

In the first example, we assume $p = 1\,000$ and $n = 100$. The matrix of predictors $\boldsymbol{X}$ has been generated with rows drawn independently from $\mathrm{N}_p(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ and $\sigma_{ij} = \rho^{|i-j|}$ with $\rho = 0.5$. We assume throughout that the regression vector consists of two blocks of informative coefficients with all remaining values set to zero. Namely, we consider the following set of regression coefficients $\boldsymbol{\beta} = (1, 2, 3, 4, 5, \mathbf{0}'_{15}, 1, 2, 3, 4, 5, \mathbf{0}'_{975})'$, where $\mathbf{0}_m$ is an $m \times 1$ vector of zeroes, and we construct the responses according to the generating linear model $\mathrm{N}_n(\boldsymbol{X\beta}, 3 \times \mathrm{I}_n)$.

	Grouping 1							**Grouping 2**								**NEG**	
Size				5	15	5	975					10	20	30	940		1000
Sparsity				1	0	1	0					1/2	1/4	0	0		1/100
	FD	FDH	$\widehat{b}_0$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	FD1	FD2	FDH	$\widehat{b}_0$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	FD	$\widehat{b}_0$
No scaling $g = 1$																	
a=0.5	51	0	0.001	2.399	0	2.388	0	40	8	8	0.001	1.235	0.016	0	0	52	0.001
a=1	42	0	0.001	4.828	0	4.780	0	40	17	17	0.001	3.133	1.349	0	0	45	0.001
a=3	34	0	0.002	13.138	0	12.805	0	31	17	19	0.002	11.590	9.332	0.002	0	33	0.002
Rescaled prior $g = 1/n^2$																	
a=0.5	49	0	0.001	2.412	0.002	2.412	0	45	4	4	0.001	1.012	0.050	0	0	49	0.001
a=1	45	0	0.054	4.850	0.004	4.850	0	43	2	4	0.053	2.333	1.408	0.007	0	45	0.06
a=3	35	0	2.122	12.613	0	12.613	0	34	1	3	2.111	10.389	8.976	0.035	0	35	2.609

Table 1: Analysis summary of the simulated data, FD/FD1/FD2/FDH refer to number of false positives overall/in non-predictive groups/in predictive groups/overall after hierarchical selection. The size and sparsity relate to the number of predictors within each group and proportion of predictive explanators.

Two non-overlapping grouping patterns were considered, where either the whole groups of predictors should enter the model (Grouping 1) or only a subset of variables within each predictive group is relevant (Grouping 2). Our first grouping scenario perfectly separates informative from uninformative predictors by clustering them into four groups identified by the following sets of indices: $\mathcal{Q}_1^{(1)} = \{1, \ldots, 5\}$, $\mathcal{Q}_2^{(1)} = \{6, \ldots, 20\}$, $\mathcal{Q}_3^{(1)} = \{21, \ldots, 25\}$ and $\mathcal{Q}_4^{(1)} = \{26, \ldots, 1\,000\}$. The second clustering mechanism is characterized by the following four sets of indices $\mathcal{Q}_1^{(2)} = \{1, \ldots, 10\}$, $\mathcal{Q}_2^{(2)} = \{11, \ldots, 30\}$, $\mathcal{Q}_3^{(2)} = \{31, \ldots, 60\}$ and $\mathcal{Q}_4^{(2)} = \{61, \ldots, 1\,000\}$, which differ not only in size but also in the proportion of relevant predictors within each group $(1/2, 1/4, 0$ and $0)$. Lastly, we conduct the analysis assuming no grouping is available, i.e. all p predictors belong to only one group. This model corresponds to an extended NEG prior with an estimable scale parameter. We compare our method to LASSO (R package `lars`) and group LASSO (R package `grpreg`).

We consider the following values for the hyper-parameters $c = d = \alpha = \gamma = 1$ and three choices of the shape parameter $a = 0.5,\ 1,\ 3$. The EM algorithm is initiated with the following starting values: $\boldsymbol{\beta}^{(0)} = \mathbf{1}_p$, $\boldsymbol{b}^{(0)} = \mathbf{1}_5$ and $\sigma^{(0)} = 1$.

In all considered settings, the 10 relevant predictors were correctly identified. Table 1 summarizes the number of false discoveries (FD), which are in the second grouping scenario divided into within non-predictive group false discoveries (FD1) and within predictive group false discoveries (FD2).

Focusing on the estimates of the pathway weights, several observations can be made based on the reported estimates in Table 1. First, the estimates corresponding to the non-relevant groups are typically at the zero boundary of the parameter space $(0 \approx 10^{-10})$, which illustrates the method's ability to correctly identify the predictive groups. Second, we observe that the magnitude of the estimated weights $\widehat{b}_1$ and $\widehat{b}_2$ in the second grouping scenario reflects the proportion of important within group variables, which is a desirable property. Third, the estimated nonzero group weights increase with the increased shape parameter a. This is expected since higher weights together with the inverse link function compensate for the large amount of penalization induced by the larger shape parameter.

It is interesting to note in Table 1 how the shape parameter a affects the within-group and overall sparsity. Assuming that all predictors within an important group are relevant (Grouping 1), increasing a gradually decreases the number of false discoveries (FD). In the presence of within-group sparsity (Grouping 2) there are noticeable differences before and after rescaling the prior. In the first case, increasing the shape parameter forces all grouped predictors to enter the model simultaneously (FD2 increases), while the number of false discoveries in non-predictive groups goes down (FD1 decreases). This suggests that larger a would be advisable in situations where we have a strong belief that the predictive groups are not sparse. For small a, we obtain sparsity within groups but might include unnecessarily many irrelevant coefficients. This is not the case after rescaling the prior distribution by the factor $g = 1/n^2$, where the within group sparsity is well preserved.

It is instructive to see how the performance can be improved by performing the hierarchical variable selection (as explained in Section 5). In the first step, we screen out pathways with a zero/small estimated weight. In the second step we select variables with nonzero estimated regression coefficients within the selected groups. This strategy in our simulated example leads to a dramatic reduction of false discoveries as compared to the plain NEG prior (FDH values in Table 1). By not performing the hierarchical selection, the NEG prior may gain in reduction of false discoveries but lose the interpretability of the group predictive pattern of the covariates.

Leave-one-out cross-validation for LASSO variable selection leads to a model with 77 false positives. The group LASSO after cross-validation selected a null model (Grouping 2) and a model with 11 false positives (Grouping 1). The group LASSO in the latter case may have benefitted from the sign consistency of the nonzero within group coefficients.

In the current case of non-overlapping groups with a "complete partition" (each variable is in one and only one group), we might not need the intercept shrinkage parameter. However, in our experience deleting this coefficient does not substantially influence the variable selection performance. The main difference is that the non-informative group weights are typically not at the boundary of the parameter space, although they are very small. Truncating these small estimates would then serve the purpose of selecting groups in the hierarchical selection scenario.

Turning to the perfect grouping scenario (Grouping 1), the majority of false discoveries has occurred in the last group consisting of 975 variables. Due to the zero estimated pathway weight, all regression coefficients in this group are penalized by the intercept weight. An estimate of this parameter is in our simulated example and is very similar to the overall shrinkage parameter in the NEG prior without the grouping, yielding a comparable number of false discoveries in this very large group. More marked differences in terms of false discoveries and non-discoveries between the plain and group versions of the NEG prior can be observed in less sparse situations, such as the ones presented in Appendix E.

As a consequence of an asymptotically vanishing effect of the prior on the posterior in

			Grouping 1											**NEG**	
Size				10	10	22	21	15	15	14	13	12	12		
Sparsity				1	1	0.091	0	0.33	0	0.143	0	0	0		
	FD	FDH	$\widehat{b}_0$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	$\widehat{b}_5$	$\widehat{b}_6$	$\widehat{b}_7$	$\widehat{b}_8$	$\widehat{b}_9$	$\widehat{b}_{10}$	FD	$\widehat{b}_0$
No scaling $g = 1$															
a=0.5	45	8	0.001	4.696	4.776	0.001	0	0.008	0	0.001	0	0	0	45	0.001
a=1	36	6	0.001	9.361	9.497	0.001	0	0.008	0	0.002	0.001	0	0.004	38	0.001
a=3	43	19	0.002	25.348	25.526	0.011	0	2.017	0	0	0	0	0.417	31	0.002
Rescaled prior $g = 1/n^2$															
a=0.5	42	7	0.001	4.905	4.933	0.001	0	0.002	0	0.001	0	0	0	42	0.001
a=1	40	6	0.001	9.624	9.754	0	0	0.008	0	0.003	0	0.001	0.004	40	0.001
a=3	23	2	1.290	28.297	28.530	0.017	0	0.495	0	0	0	0	0	23	3.935

Table 2: Analysis summary of the simulated data, FD/FDH refer to number of false positives without hierarchical selection/after hierarchical selection, size and sparsity relate to the group size and proportion of predictive explanators.

the unscaled model, the pathway coefficients in the inverse link decrease with growing sample size, where the whole linear predictor asymptotically approaches a value bounded away from zero. In order to preserve the shrinkage effect in the limit, we have considered a rescaled NEG prior, where the scale parameter is multiplied by a factor n^2. According to Theorem 6.3, the scale parameter (inverted linear predictor) in the modified model should ideally approach zero and its root-n multiple grow to infinity as $n \to \infty$. Evidence for this behavior was observed in a simulated experiment described in Appendix D. It is interesting to note the relationship of the pathway weights to the group size, where the estimated coefficients represent the proportion of predictive coefficients within each pathway. Larger pathways have typically smaller estimated coefficients as compared to smaller pathways with the same (number of) predictive variables. This behavior was also evident in the results of the simulation study in Appendix D. It is worth mentioning that the regression on the scale parameter is less influential in the rescaled version of the model ($g = 1/n^2$). The overall performance in terms of false discoveries and non-discoveries there is very similar for the grouped NEG and the plain NEG priors. We speculate that rescaling the prior, the regression on the scale parameter has little influence on the model search and rather helps to effectively discriminate between the predictive and the non-predictive groups. The pathway weights are seen to correctly represent the grouping structure and serve as a useful prerequisite for group selection that isolates discoveries in non-predictive groups.

7.2 Overlapping Groups

In our second simulated example we assume that the predictors correspond to known genes and cluster within known pathways. The list of gene/pathway interactions was generated from the KEGG database using the R Bioconductor library `hgu133plus2`. A subset of size $p = 1\,000$ was randomly selected from a set of genes analyzed in the next section. Focusing only on known pathways consisting of at least 10 genes, we selected at random $q = 10$ pathways for the construction of the grouping structure.

Two of these pathways were randomly selected to be predictive. Similarly as in the previous example we consider two possible scenarios: (1) all genes within the predic-

				Grouping 2										
Size					27	25	22	21	15	15	14	13	12	12
Sparsity					0.37	0.4	0.091	0	0.33	0	0.143	0	0	0
	FD1	FD2	FDH	$\widehat{b}_0$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$	$\widehat{b}_5$	$\widehat{b}_6$	$\widehat{b}_7$	$\widehat{b}_8$	$\widehat{b}_9$	$\widehat{b}_{10}$
							No scaling $g = 1$							
a=0.5	31	25	33	0.001	0.781	1.229	0.001	0	0.011	0.001	0.002	0.001	0.001	0
a=1	30	25	33	0.001	3.782	4.971	0.007	0	0.188	0.003	0.008	0.001	0	0.003
a=3	30	30	48	0.002	19.169	21.811	0.183	0.483	2.521	0	0.698	0	0	0
							Rescaled prior $g = 1/n^2$							
a=0.5	38	4	4	0.001	0.175	0.246	0.001	0	0.007	0	0.001	0	0	0
a=1	36	4	4	0.001	4.430	5.440	0	0	0.003	0	0.001	0	0	0
a=3	23	0	0	3.198	18.049	19.272	0	0	0.729	0	0	0	0	0

Table 3: Analysis summary of the simulated data, FD/FD1/FD2/FDH refer to number of false positives overall/in non-predictive groups/in predictive groups/overall after hierarchical selection. The size and sparsity relate to the number of predictors within each group and proportion of predictive explanators.

tive pathways are assumed to contribute in explaining the variability of the response (Grouping 1), (2) predictive pathways are sparse (Grouping 2). We assume that in each of the two predictive pathways (sized 27 and 25), there are only 10 relevant predictors. The second grouping pattern corresponds to the pathway loading matrix generated from the KEGG database. Limiting the size of the predictive pathways to 10, we obtain a modified grouping pattern that we associate with the first grouping scenario.

Given the binary pathway loading matrix $\boldsymbol{Z}$ (associated with Grouping 2), we first generate the covariance matrix $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})_{i,j=1}^{p}$, where $\widetilde{\Sigma} = \boldsymbol{Z}\text{diag}\{\rho_1, \ldots, \rho_q\}\boldsymbol{Z}' + \mathrm{I}_p$, which is positive definite and symmetric. Note that genes that do not share any underlying pathway have zero pairwise correlations. The values $\rho_i > 0$ (not bounded to lie within an interval $[0, 1]$) regulate the magnitude of the within-pathway correlations. The correlation matrix $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ is obtained by setting $\sigma_{ij} = \widetilde{\sigma}_{ij}/\sqrt{\widetilde{\sigma}_{ii}\widetilde{\sigma}_{jj}}$. The predictor matrix $\boldsymbol{X}$ is then generated according to $\mathrm{N}_n(\mathbf{0}, \Sigma)$. The observations on the response variable are created according to the relation $\mathrm{N}_n(\boldsymbol{X\beta}, \sigma^2\mathrm{I}_n)$. We keep $\sigma^2 = 1$, $n = 100$ and we assume (a) a relatively high signal to noise ratio, (b) medium correlation within non-predictive pathways, and (c) high correlation within predictive pathways. Namely, the nonzero entries in the regression vector $\boldsymbol{\beta}$ equal 2. In order to obtain an average correlation of 0.8 and 0.3 within the predictive and non-predictive pathways, we assume $\rho_j = 0.1 \times \mathrm{I}\left(j \notin \bigcup_{k=1}^{3} \mathcal{Q}_k\right) + 2 \times \mathrm{I}\left(j \in \bigcup_{k=1}^{3} \mathcal{Q}_k\right)$.

The values of hyper-parameters were considered to be the same as in the previous example. The starting values for the algorithm are again $\boldsymbol{\beta}^{(0)} = \mathbf{1}_p, \boldsymbol{b}^{(0)} = \mathbf{1}_q$ and $\sigma^{(0)} = 1$.

The summary of the analysis for the non-sparse clusters (Grouping 1) is in Table 2. Due to the overlap between the groups, some of the "non-predictive" pathways contain important coefficients as well. The magnitude of the estimated pathway weights again reflects the degree of predictiveness of each group, typically leaving the unimportant pathways with a zero weight. The numbers of false discoveries (without applying the hierarchical selection) are comparable to the plain NEG prior. Under the hierarchical selection after removing pathways with a zero estimated weight, the respective numbers of false discoveries without rescaling are $8, 6$ and 19 for $a = 0.5, 1, 3$ (FDH values in Table 2). Again, no false-nondiscoveries were observed.

In the second grouping scenario (Table 3) we again observe higher within group false positives for larger values a, a consequence of strongly enforced smoothness in within-group penalties. The hierarchical selection reduces the false positives in this simulated example to $33, 33$ and 48 for $a = 0.5, 1, 3$.

It is interesting to compare the results before and after rescaling the model with the factor $g = 1/n^2$. The results in Tables 2 and 3 again show the superiority of the rescaled model, both in the accuracy of determining important pathways, and in controlling within group false discoveries. The hierarchical selection performs superbly in identifying the underlying sparsity. In contrast, applying leave-one-out cross-validated LASSO variable selection, we obtain 75 false positives. We also implemented the overlapping group LASSO of Jacob et al. (2009) by duplicating the columns in the regression matrix, which appear in more than one group, and applying the standard group LASSO computation (R-package `grpreg`). Selecting the optimal penalty parameter using the Bayesian information criterion (BIC), we obtain a model with 28 false positives for Grouping 1 and 32 false positives for Grouping 2, which is more than for the rescaled grouped NEG model with an appropriately chosen shape parameter a.

8 Application

We demonstrate the practical usefulness of the proposed method on a microarray gene expression data set with glioblastoma patients (Horvath et al. (2006)). Glioblastoma is a primary malignant brain tumor, which classifies as one of the most lethal tumors in adults. Diagnosed patients have a median survival of 15 months despite various treatments. The data consists of two sets of measurements coming from two independent studies. Similarly as in Pan et al. (2010) and Li and Li (2008), we shall use only the first set, which appears to carry more information related to time to death from glioblastoma. We select a subset of 50 patients (out of 55) with the observed clinical outcome. The logarithm of time to death (in days) is treated as the response. Gene expression profiles were obtained using the Affymetrix platform and further normalized using the RMA methods (Irizarry et al. 2003). Li and Li (2008) focused on a subset of 1 533 genes, which were involved in gene pathways. Using the R Bioconductor library `hgu133plus2` we retrieved the functional gene/pathway interactions from the KEGG database. For each gene, a list of active pathways was generated and translated into a pattern of zeros in the $p \times q$ matrix $\boldsymbol{Z}$, where rows correspond to $p = 1\,533$ genes and columns to $q = 103$ pathways (only pathways consisting of at least 20 genes were considered for the analysis).

In order to determine genes predictive of time to death we first run the LASSO method (R library `lars`), selecting the optimal penalty parameter as the value which minimizes the leave-one-out cross-validated prediction mean-squared error. As a result, we obtain 21 genes reported in Table 4 together with information on their pathway involvement (the top 10 represented pathways with at least 3 genes).

IKBKG	CASP5	CAMK2D	TRH	CX3CL1	CCL21	PRKCG	PPP3R1	PIK3C2G	ZAK	PDPK1	CLDN11	KLKB1	ITGB7	IRF3	FOXO1A	EPHB4	CTNNB1	CTLA4	COMP	ISGF3G	
·	·	·	·	·	·	×	·	·	·	×	·	·	×	·	·	·	×	·	×	·	Focal adhesion
×	·	·	·	·	·	×	×	·	×	·	·	·	·	·	·	·	·	·	·	·	MAPK signaling pathway
·	·	×	·	·	·	×	×	·	·	·	·	·	·	·	·	·	×	·	·	·	Wnt signaling pathway
·	·	·	·	·	·	×	·	·	×	·	×	·	·	·	·	·	×	·	·	·	Tight junction
×	·	·	·	·	·	·	·	·	·	·	×	·	·	×	·	·	·	·	·	×	Hepatitis C
×	·	·	·	·	·	×	·	·	·	·	·	·	·	·	×	·	×	·	·	·	Pathways in cancer
×	·	·	·	·	·	·	·	·	·	×	·	·	·	·	×	·	×	·	·	·	Prostate cancer
·	·	×	·	·	·	×	×	·	·	·	·	·	·	·	·	·	·	·	·	·	Calcium signaling pathway
×	·	·	·	×	×	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	Chemokine signaling pathway
·	·	·	·	·	·	·	·	·	·	·	×	·	×	·	·	·	·	×	·	·	Cell adhesion molecules

Table 4: LASSO selected genes together with 10 top represented pathways.

We then repeat the analysis using the Laplace version of the EM algorithm with an inverse link function to incorporate the gene-pathway membership information. Based on the experience from the simulated examples we choose $a = 5$ and apply the rescaled version of the model with the scaling factor $g = 1/n^2$. In order to mitigate the problem of finding a locally suboptimal solution, we run the algorithm for multiple choices of starting values and select the solution which corresponds to the highest log posterior mode (which can be evaluated up to an additive constant). Considering the following values of hyper-parameters $c = d = \alpha = \gamma = 1$ and setting the convergence threshold ε to 10^{-5}, we consider a unit starting vector $\boldsymbol{\beta}^{(0)} = \mathbf{1}_p$ and 10 initial values randomly sampled from from $\mathrm{N}_p(\mathbf{0}, \mathrm{I})$. The starting values for the pathway weights and variance parameter are $b^{(0)} = 1$ and $\sigma^{(0)} = 1$.

The highest located log-posterior mode (10 595.72 plus a common additive constant) is associated with a model consisting of 21 predictors, of which 13 overlap with the LASSO analysis (marked with blue in Table 5). We identified 21 predictive pathways with a nonzero estimated weight, where each of the selected genes is involved in at least one of these pathways. Table 5 reports a subset of 10 pathways with the highest numbers of identified genes together with the estimated weights $\widehat{\boldsymbol{b}}$, which represent the proportion of within group predictive genes. The complete list of gene-pathway interactions for all the 21 pathways is in Appendix F.

Both LASSO and our method identified genes previously associated with malignant brain tumors such as `FOXO1A`, which is a transcription factor linked to glioblastoma (Choe et al. 2003), or `PRKCG` and `CAMK2D`, which are members of the glioma pathway. Other genes were found to be related to various brain molecular processes such as

TNFRSF7	**CASP5**	**CAMK2D**	SLIT1	**CX3CL1**	**PRKCG**	WNT4	**ZAK**	**CLDN11**	**KLKB1**	**ITGB7**	ITGAL	**IRF3**	INPP5D	LAMA1	FRAP1	**FOXO1A**	DFFB	**CTNNB1**	**COMP**	**ISGF3G**		$\widehat{\beta}$
·	·	·	·	·	×	·	×	×	·	·	·	·	·	·	·	·	·	×	·	·	Tight junction	0.153
·	·	·	·	·	×	·	·	×	·	·	×	·	·	·	·	·	·	×	·	·	Leukocyte transendothelial migration	0.188
·	·	×	·	·	×	×	·	·	·	·	·	·	·	·	·	·	·	×	·	·	Melanogenesis	0.341
·	·	·	·	·	·	·	·	·	·	×	·	·	·	×	·	·	·	·	×	·	ECM-receptor interaction	0.360
·	·	·	·	·	·	·	·	×	·	×	×	·	·	·	·	·	·	·	·	·	Cell adhesion molecules (CAMs)	0.019
·	·	·	·	·	·	·	·	·	·	·	·	·	×	·	×	×	·	·	·	·	Insulin signaling pathway	0.029
·	·	·	·	·	·	·	·	×	·	·	·	×	·	·	·	·	·	·	·	×	Hepatitis C	0.089
·	·	×	·	·	×	·	·	·	·	·	·	·	·	·	×	·	·	·	·	·	Glioma	0.102
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	×	×	·	×	·	·	Prostate cancer	0.174
×	·	·	·	×	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	Cytokine-cytokine receptor interaction	0.008

Table 5: Analysis using the NEG prior with grouping, selected genes together with top 10 identified pathways.

CX3CL1, controlling neuronal survival and neuron transmission (Scium et al. 2010), and CTNNB1, found to be differentially expressed in brain tumors (Nikuseva-Martic et al. 2010).

Focusing on the genes that were missed by LASSO: DFFB is an apoptosis regulator, identified as a contributing factor in development of a specific type of glioma (McDonald et al. 2005), FRAP1 is a member of the glioma pathway, SLIT1 is an axon guidance gene whose epigenetic changes were associated with glioma (Dickinson et al. 2004).

Several of the 10 pathways reported in Table 9 in Appendix F were recognized to be linked with brain molecular processes underlying malignant tumors. Tight junctions, which mediate blood-brain barriers and whose impairment may cause brain edema, have been reported defective in glioblastoma (Schneider et al. 2004). The ECM (extracellular matrix) pathway has a confirmed role in cellular processes associated with neuronal survival, axon guidance and synapse formation. Impaired activity of the ECM receptors may create a molecular basis for malignant gliomas (Paulus and Tonn 1995). Expression of cell adhesion molecules (binding proteins) has been shown consistently to be altered in glioblastoma as compared to the normal brain tissue (Gingras et al. 1995). The full list of the 21 identified pathways is deferred to Appendix F.

Whereas the post hoc pathway analysis for the LASSO selected genes revealed MAPK signaling pathway, which is an important glioblastoma related pathway (Nakada et al. 2011), it did not appear in the 21 pathways selected by our method. Since the estimated pathway weights corresponds to the proportion of predictive genes, perhaps smaller pathways involving a similar set of genes may have had a selective advantage.

The plain rescaled NEG prior (without grouping structure) led to a lower log posterior mode (9 428.231 plus the common additive constant) associated with 22 genes, of which 12 overlap with the model including the grouping. We implemented the overlapping group LASSO by augmenting the regression matrix with duplicates of columns, which occur in more than one group. This leads to a new regression matrix with 6 780 columns. Applying the group LASSO computation (`R`-package `grpreg`) we identified 17 pathways consisting of 608 different genes after selecting the optimal penalty parameter based on the BIC. The list of these pathways is in Appendix F. Since group LASSO does not assume within group sparsity, many of the identified genes are likely to be false positives.

Regarding the computational time, the most expensive operations are the updates of coefficients $\boldsymbol{\beta}$ and $\boldsymbol{b}$. The update of $\boldsymbol{\beta}$ is in the Laplace EM algorithm based on solving the LASSO problem, which using the `lars` package took 0.41 seconds in the glioblastoma dataset on a 2.533GHz server. For the multiple selected starting values, the EM algorithm converged in $20 - 40$ iterations with an average of 26. This time would compare to performing $20 - 40$ fold cross-validation in the LASSO analysis. The time needed to update $\boldsymbol{b}$ will barely matter for a small number of pathways (< 10). In the glioblastoma data with 104 pathways, one update took on average 5 seconds per iteration using routine `R` optimization techniques. In contrast with the MCMC implementation of the Bayesian LASSO (`R`-package `monomvn`), drawing 100 samples from the posterior took around 20 seconds.

9 Discussion

In this paper we proposed a method for Bayesian shrinkage estimation in linear regression, which incorporates grouping information within the sparsity inducing regularization. We demonstrated on two simulated examples that the method is capable of retrieving groups of informative predictors through the identification of nonzero group weights. However, we expect that the performance will be influenced by the level of agreement between the external structural information and actual “group predictive behavior”. In case no such information is available, the pathway loading matrix could be obtained from e.g. a sparse factor analytic model (Carvalho et al. 2008), where nonzero entries in the loading matrix indicate functional interaction with latent factor/ pathway activity.

We have opted for the EM algorithm as our computational strategy, which offers substantial time savings. Moreover, the Laplace version of the algorithm provides a naturally sparse solution, which identifies sets of active predictors that correspond to a particular model. As such, this EM algorithm can be regarded as a deterministic model search machine, which during the iterative process drives the search towards more interesting models. However, due to the multimodality of the posterior finding the global mode is not guaranteed. The choice of an initial value is likely to influence the results and the speed of the convergence. Running the procedure for multiple choices of starting values and selecting the mode associated with the highest posterior value (which can be computed up to a constant) may increase our chances of finding the global mode. An alternative solution based on deterministic annealing was suggested by Ueda and Nakano (1998) in the context of normal mixtures. The authors suggest performing the E-step with respect to a perturbed version of the posterior distribution, which is proportional to the log-complete data posterior raised to the power of an inverse temperature. Such an E-step can be still obtained in a closed form.

By using the EM algorithm we are trading the benefits of the Bayesian inference based on the full posterior (in particular confidence assessment) for computational efficiency. Similarly as in sparse penalized likelihood techniques, our method outputs merely a sparse point estimate of the coefficient vector. One possibility to perform (frequentist) uncertainty assessment for our method is through asymptotics borrowed from the established theory on penalized likelihood estimators. Fan and Li (2001) and Peng and Fan (2004) developed asymptotic theory showing model selection consistency and asymptotic normality of specific sparse penalized likelihood estimators, both for fixed p as well as for a diverging number of parameters. These results can be transferred to the Bayesian MAP estimation framework directly in instances where the marginal prior on the regression coefficients takes the form $\exp(-n\,\mathrm{pen}_\lambda(|\beta|))$. The penalty function $\mathrm{pen}_\lambda(|\beta|)$ then needs to fulfill certain conditions in order for the oracle property of the MAP estimator to be guaranteed. Although the plain $\mathrm{NEG}(a, s)$ prior does not meet these requirements, multiplying the scale parameter s by a factor depending on the sample size warrants the desired properties. We showed that the penalty function implied by the rescaled prior $\mathrm{NEG}(a, n^2 s)$ satisfies the conditions in Fan and Li (2001) for root-n consistency and asymptotic normality of the Bayesian MAP estimator, which creates a basis of sandwich-like standard errors. One disadvantage of this approach is that it disregards the uncertainty around the zero estimates by setting their standard errors to zero. Moreover, the finite sample distributions for some penalized likelihood estimators have been shown to be severely deviated from the approximating normal distribution (Leeb and Potscher 2005). An alternative way to compute the standard errors, not only for the regression coefficients but also for the pathway weights $\boldsymbol{b}$, is through bootstrapping. However, this can lead to inconsistent standard errors if the true regression coefficient values are zero, as shown in the LASSO context by Kyung et al. (2010).

Our proposed model selection procedure outputs a sparse point estimate of the regression vector, which forms the basis for a potential prediction rule. In practical implementations, the sparse model-selectors/predictors such as LASSO are typically tuned to achieve optimal prediction accuracy. Whereas tuning parameters in some hierarchical models can be directly related to Akaike information criterion (AIC) and BIC penalties (George and Foster 1997), the tradeoff between prediction and model selection accuracy is more difficult to control in our model. The scale penalty parameter is adaptively determined from the data, where appropriate limiting behavior guarantees identification of the true model with probability converging to one. We believe that the main practical value of our method rests in improved interpretation of the collective behavior of the predictors in the effort of finding a sparse representation of the data rather than in accurate prediction.

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions*. Dover Publications, 1 edition. 238, 239

Armagan, A., Dunson, D., and Lee, J. (2012). "Generalized Double Pareto Shrinkage." Technical report, Duke University. 223, 227, 232, 235

Carvalho, C. and Polson, N. (2010). "The Horseshoe Estimator for Sparse Signals." *Biometrika*, 97(476): 465–480. 236

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., and West, M. (2008). "High-Dimensional Sparse Factor Modelling: Applications in Gene Expression Genomics." *Journal of the American Statistical Association*, 103(484): 1438–1456. 247

Chen, M.-H. and Ibrahim, J. G. (2003). "Conjugate priors for generalized linear models." *Statistica Sinica*, 13(2): 461–476. 226

Choe, G., Horvath, S., Cloughesy, T., Crosby, K., Seligson, D., Palotie, A., Inge, L., Smith, B., Sawyers, C., and Mischel, P. (2003). "Analysis of the phosphatidylinositol 3'-kinase signaling pathway in glioblastoma patients in vivo." *Cancer Research*, 63(2): 2742–2746. 245

Dempster, A., Laird, N., and Rubin, D. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society, Series B*, 39(1): 1–38. 228

Dickinson, R., Dallol, A., Bieche, I., Krex, D., Morton, D., Maher, E., and Latif, F. (2004). "Epigenetic inactivation of SLIT3 and SLIT1 genes in human cancers." *British Journal of Cancer*, 13: 2071–2078. 246

Fan, J. and Li, R. (2001). "Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties." *Journal of the American Statistical Association*, 96: 1348–1360. 222, 223, 232, 235, 237, 238, 248

Figueiredo, M. A. (2003). "Adaptive Sparseness for Supervised Learning." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25: 1150–1159. 227

Gelfand, A. and Vounatsou, P. (2003). "Proper Multivariate Conditional Autoregressive Models for Spatial Data Analysis." *Biostatistics*, 4: 11–15. 222

George, E. and Foster, D. (1997). "Calibration and Empirical Bayes Variable Selection." *Biometrika*, 87: 731–747. 248

Gingras, M., Roussel, E., Bruner, J., Branch, C., and Moser, R. (1995). "Comparison of cell adhesion molecule expression between glioblastoma multiforme and autologous normal brain tissue." *Journal of Neuroimmunology*, 57: 143–153. 246

Golub, G. and van Loan, C. (1996). *Matrix Computations*. The John Hopkins University Press, 1 edition. 231

Gradshteyn, I. and Ryzhik, E. (2000). *Table of Integrals Series and Products*. Academic Press, 6 edition. 230, 231, 235

Griffin, J. E. and Brown, P. J. (2012). "Bayesian Hyper-LASSOS with Non-convex Penalization." *Australian & New Zealand Journal of Statistics*, 53: 423–442. 223, 225, 226, 227, 231, 234, 235

Horvath, S., Zhang, B., Carlson, M., Lu, K., Zhu, S., Felciano, R., Laurance, M., Zhao, W., Qi, S., Chen, Z., Lee, Y., Scheck, A., Liau, L., Wu, H., Geschwind, D., Febbo, P., Kornblum, H., Cloughesy, T., Nelson, S., and Mischel, P. (2006). "Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Molecular Target." *Proceedings of the National Academy of Sciences of the United States of America*, 103: 17402–17407. 244

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003). "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics*, 4: 249–264. 244

Ishwaran, H. and Rao, S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *The Annals of Statistiscs*, 33: 730–773. 239

Jacob, L., Obozinski, G., and Vert, J. (2009). "Group LASSO with Overlap and Graph LASSO." *Proceedings of the 26th International Conference on Machine Learning*, 55: 1–8. 223, 244

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). "The KEGG Databases at GenomeNet." *Nucleic Acids Research*, 30: 42–46. 222

Kiiveri, H. (2003). "A Bayesian Approach to Variable Selection When the Number of Variables is Very Large." *Institute of Mathematical Statistics Lecture Notes-Monograph Series*, 40: 127–143. 227

Kyung, M., Gilly, J., Ghosh, M., and Casella, G. (2010). "Penalized Regression, Standard Errors, and Bayesian Lassos." *Bayesian Analysis*, 5: 369–412. 248

Leeb, H. and Potscher, B. M. (2005). "Model Selection and Inference: Facts and Fiction." *Econometric Theory*, 21: 21–59. 248

Li, C. and Li, H. (2008). "Network-constrained Regularization and Variable Selection for Analysis of Genomic Data." *Biometrics*, 24(9): 1175–1182. 222, 223, 244

Li, F. and Zhang, N. R. (2010). "Bayesian Variable Selection in Structured High-dimensional Covariate Spaces with Applications in Genomics." *Journal of the American Statistical Association*, 105(3): 1978–2002. 222

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). "Mixtures of g-priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 410–423. 237

McDonald, J., Dunmire, V., Taylor, R., E. Sawaya, Bruner, J., Fuller, G., Aldape, K., and Zhang, W. (2005). "Attenuated Expression of DFFB is a Hallmark of Oligodendrogliomas with 1p-Allelic Loss." *Molecular Cancer*, 4: 1476–1498. 246

McLachlan, G. J. and Krishnan, T. (1996). *The EM Algorithm and Extensions*. Wiley-Interscience, 2 edition. 227

Nakada, M., Kita, D., Watanabe, T., Hayashi, Y., Teng, L., Pyko, I., and Hamada, J. (2011). "Aberrant Signaling Pathways in Glioma." *Cancers*, 3: 3242–3278. 246

Nikuseva-Martic, T., Beros, V., Pecina-Slaus, N., Pecina, H. I., and Bulic-Jakus, F. (2010). "Genetic changes of CDH1, APC, and CTNNB1 found in human brain tumors." *Pathology - Research and Practice*, 203(11): 779–787. 246

Pan, W., Benhuai, X., and Xiaotong, S. (2010). "Incorporating Predictor Network in Penalized Regression with Application to Microarray Data." *Biometrics*, 66(2): 474–484. 222, 223, 244

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 223, 226

Paulus, W. and Tonn, J. (1995). "Interactions of glioma cells and extracellular matrix." *Journal of Neuro-Oncology*, 24: 87–91. 246

Peng, H. and Fan, J. (2004). "Nonconcave penalized likelihood with a diverging number of parameters." *The Annals of Statistics*, 32(3): 928–961. 248

Schneider, S., Ludwig, T., Tatenhorst, L., Braune, S., Oberleithner, H., Senner, V., and Paulus, W. (2004). "Glioblastoma cells release factors that disrupt blood-brain barrier features." *Acta Neuropathologica*, 107: 272–276. 246

Scium, G., Soriani, A., Piccoli, M., Frati, L., Santoni, A., and Bernardini, G. (2010). "CX3CL1 axis negatively controls glioma cell invasion and is modulated by transforming growth factor-beta1." *Neuro-Oncology*, 111(2): 3626–3634. 246

Stingo, F., Chen, Y., Tadesse, M., and Vannucci, M. (2011). "Incorporating Biological Information into Linear Models: A Bayesian Approach to the Selection of Pathways and Genes." *The Annals of Applied Statistics*, 5: 1202–1214. 222

Stingo, F., Chen, Y., Vannucci, M., Barrier, M., and Mirkes, P. (2010). "A Bayesian Graphical Modeling Approach to MicroRNA Regulatory Network Inference." *Annals of Applied Statistics*, 4: 2024–2048. 227

Stingo, F. and Vannucci, M. (2011). "Variable Selection for Discriminant Analysis with Markov Random Field Priors for the Analysis of Microarray Data." *Bioinformatics*, 27(4): 495–501. 222

Tibshirani, R. (1994). "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society, Series B*, 58: 267–288. 222, 223

Ueda, N. and Nakano, R. (1998). "Deterministic annealing EM algorithm." *Neural Networks*, 11: 271–282. 248

Yuan, M. and Lin, Y. (2006). "Model selection and estimation in regression with grouped variables." *Journal of the Royal Statistical Society, Series B*, 68: 49–67. 223

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *In Bayesian Inference and Decision Techniques*. 237

Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association*, 101(476): 1418–1429. 226, 233

Zou, H. and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B*, 67: 301–320. 222

Zou, H. and Li, R. (2008). "One-step Sparse Estimates in Nonconcave Penalized Likelihood Models." *The Annals of Statistics*, 36(4): 1509–1533. 232, 233

Acknowledgments

The authors would like to thank Hongzhe Li for kindly providing the gene expression data.

Appendix A: Proof of Equation (2)

Denote by a and s the shape and scale of the NEG distribution. As shown in Griffin and Brown (2012), in order to evaluate the conditional expectation $\mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\left(\frac{1}{\tau_j^2}\right)$ it suffices to note the connection to the derivative of the logarithm of the NEG prior density. We have

$$\begin{aligned}-\frac{\partial \log p_{a,s,\sigma}(|\beta_j|)}{\partial|\beta_j|} &= -\left(\frac{\partial p_{a,s,\sigma}(|\beta_j|)}{\partial|\beta_j|}\right)\frac{1}{p_{a,s,\sigma}(|\beta_j|)} = \int_0^\infty \frac{|\beta_j|}{\sigma^2\tau_j^2}\frac{p(\beta_j \mid \tau_j^2,\sigma^2)p(\tau_j^2|a,s)}{p_{a,s,\sigma}(|\beta_j|)}\mathrm{d}\,\tau_j^2\\ &= \int_0^\infty \frac{|\beta_j|}{\sigma^2\tau_j^2}p(\tau_j^2|\beta_j,a,s,\sigma)\mathrm{d}\,\tau_j^2 = \frac{|\beta_j|}{\sigma^2}\mathsf{E}\left(\frac{1}{\tau_j^2}|\beta_j,a,s,\sigma\right).\end{aligned}$$

The marginal prior distribution $p_{a,s,\sigma}(|\beta_j|)$ can be obtained in a closed form (using Gradshteyn and Ryznik (2000), page 334, equation 7) as follows:

$$\begin{aligned}p_{a,s,\sigma}(|\beta_j|) &= \int_0^\infty \frac{s\,x^{a-1/2}}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{x\beta_j^2}{2\sigma^2}\right)\frac{a}{(x+s)^{a+1}}\mathrm{d}\,x\\ &= \frac{a\,2^a\sqrt{s}}{\sqrt{\pi\sigma^2}}\Gamma\left(a+0.5\right)\exp\left(\frac{\beta_j^2 s}{4\sigma^2}\right)D_{-2(a+0.5)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right).\end{aligned}$$

The derivative of the marginal distribution can be again obtained analytically (Gradshteyn and Ryznik (2000), page 334, equation 6) according to

$$\begin{aligned}-\frac{\partial p_{a,s,\sigma}(|\beta_j|)}{\partial|\beta_j|} &= \int_0^\infty \frac{s}{\sqrt{2\pi\sigma^2}}\frac{|\beta_j|}{\sigma^2}x^{a+1/2}\exp\left(-\frac{\beta_j^2}{2\sigma^2}x\right)\frac{a}{(x+s)^{a+1}}\mathrm{d}\,x\\ &= \frac{a\,2^{a+1}s}{\sqrt{\pi\sigma^2}}\Gamma\left(a+1.5\right)\exp\left(\frac{\beta_j^2 s}{4\sigma^2}\right)D_{-2(a+1)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right).\end{aligned}$$

Combining these expressions for the NEG prior and its derivative, we obtain

$$\begin{aligned}\mathsf{E}\left(\frac{1}{\tau_j^2}|\beta_j,a,s,\sigma\right) &= -\left(\frac{\partial p_{a,s,\sigma}(|\beta_j|)}{\partial|\beta_j|}\right)\frac{1}{p_{a,s,\sigma}(|\beta_j|)}\frac{\sigma^2}{|\beta_j|}\\ &= \frac{2(a+0.5)\sigma\sqrt{s}}{|\beta_j|}\frac{D_{-2(a+1)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)}{D_{-2(a+0.5)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)}.\end{aligned}$$

Appendix B: Proof of Equation (3)

Denote by a and s the shape and scale of the NEG distribution. The computation of the conditional expectation follows from Gradshteyn and Ryznik (2000), page 334, equation

5. More precisely, it holds that

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\tau}^2|\cdot}\left(\frac{\tau_j^2}{1+\tau_j^2 s}\right) &= \int_0^\infty \frac{a\,s^{-1/2}}{\sqrt{2\pi\sigma^2}}\frac{z^{a-0.5}}{(1+z)^{a+2}}\exp\left(-\frac{\beta_j^2 s}{2\sigma^2}z\right)\mathrm{d}\,z\\ &= \frac{a\,\Gamma\left(a+0.5\right)}{\sigma\sqrt{2\pi s}}\boldsymbol{\Psi}\left(a+0.5,-\frac{1}{2},\frac{\beta_j^2 s}{2\sigma^2}\right)\frac{1}{p(\beta_j|\boldsymbol{b},\sigma)}.\end{aligned}$$

Appendix C: Proof of Equation (5)

Denote by a and s the shape and scale of the NEG distribution and $p_{a,s,\sigma}(\beta)$ the marginal NEG distribution. According to Gradshteyn and Ryznik (2000), page 360, equation 1 we have

$$\begin{aligned}\mathsf{E}_{\boldsymbol{\lambda}^2|\cdot}\lambda_j^r &= \frac{1}{p_{a,s,\sigma}(\beta_j)}\int_0^\infty \frac{\lambda_j^{r+1}}{\sqrt{2\sigma^2}}\exp\left(-\frac{\sqrt{2}\lambda_j|\beta_j|}{\sigma}\right)\frac{\lambda_j^{2(a-1)}}{\Gamma(a)s^a}\exp\left(-\frac{\lambda_j^2}{s}\right)\mathrm{d}\,\lambda_j^2\\ &= \frac{2}{\Gamma(a)s^a\sqrt{2\sigma^2}p_{a,s,\sigma}(\beta_j)}\int_0^\infty z^{2a+r}\exp\left(-\frac{\sqrt{2}|\beta_j|z}{\sigma}-\frac{z^2}{s}\right)\mathrm{d}\,z\\ &= \frac{2}{\Gamma(a)s^a\sqrt{2\sigma^2}p_{a,s,\sigma}(\beta_j)}\left(\frac{2}{s}\right)^{-\frac{2a+r+1}{2}}\Gamma\left(2a+r+1\right)\exp\left(\frac{\beta_j^2 s}{4\sigma^2}\right)\times\\ &\qquad D_{-(2a+r+1)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right).\end{aligned}$$

Appendix D: Effects of sample size and pathway size on estimated pathway weights

To illustrate the effects of increasing sample size as well as pathway size for the fixed number of predictors we designed a small simulated experiment. We consider $p = 100$ predictors which cluster within groups that differ not only in the number of elements but also in the proportion of predictive variables. The assumed true coefficient vector is

$$\boldsymbol{\beta} = (1,2,3,4,5,\underbrace{0,\ldots\ldots,0}_{15},1,2,3,4,5,\underbrace{0,\ldots\ldots,0}_{15},1,2,3,4,5,0,\ldots\ldots,0)'.$$

The grouping structure divides the 100 predictors into 6 non-overlapping groups consisting of $5, 10, 15, 20, 25$ and 25 predictors with predictive proportions $1, 0, 1/3, 1/4, 0$ and 0. For each of the three considered sample sizes $n = 50, 500, 1\,000$, we generate the regression matrix with rows drawn independently from $\mathrm{N}_p(\mathbf{0}, \mathrm{I}_p)$. Ten response vectors were generated according to $\boldsymbol{Y} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}, 3 \times \mathrm{I}_n)$ for each sample size. The average estimated group weights are in Table 6 below.

We observe a decreasing trend in the estimated weights as the sample size grows in the unscaled model. After rescaling, the weights are seen to increase as the scale parameter

	Grouping					
Size	5	10	15	20	25	25
Sparsity	1	0	1/3	1/4	0	0
	$\widehat{b_1}$	$\widehat{b_2}$	$\widehat{b_3}$	$\widehat{b_4}$	$\widehat{b_5}$	$\widehat{b_6}$
No scaling $g = 1$						
$n = 50$	3.968	0.018	1.813	1.266	0.026	0.004
	(0.136)	(0.032)	(1.059)	(0.718)	(0.052)	(0.005)
$n = 500$	1.406	0.001	0.072	0.039	0.001	0.001
	(0.039)	(<0.001)	(0.008)	(0.007)	(<0.001)	(<0.001)
$n = 1\,000$	1.346	0.001	0.068	0.036	0.001	0.001
	(0.017)	(<0.001)	(0.005)	(0.005)	(<0.001)	(<0.001)
Rescaled model $g = 1/n^2$						
$n = 50$	4.231	0.007	0.397	0.303	0.016	0.018
	(0.746)	(0.009)	(0.275)	(0.303)	(0.023)	(0.024)
$n = 500$	4.958	0.032	1.154	0.707	0.079	0.078
	(0.004)	(0.020)	(0.159)	(0.192)	(0.055)	(0.053)
$n = 1\,000$	4.952	0.079	1.472	1.072	0.194	0.197
	(0.003)	(0.034)	(0.163)	(0.203)	(0.084)	(0.090)

Table 6: Results from a simulation study to evaluate effects of sample size and group size. Table reports average estimated pathway weights with standard deviations in brackets.

(inverted linear predictor) goes down, which is according to Theorem 5 a desirable property. In both models, the size of pathway weights reflects the proportion of important coefficients.

Appendix E: Simulated examples with different degrees of sparsity

In order to investigate the practical gains in more realistic scenarios, we considered a set of simulation experiments with three different degrees of sparsity and a lower signal to noise ratio. We assume $a = 1, p = 1\,000, \sigma^2 = 5$ and three sparsity settings for the unscaled version of the model:

$$\boldsymbol{\beta}_1 = (\underbrace{1,\ldots\ldots,1}_{30},\underbrace{0,\ldots\ldots,0}_{470},\underbrace{1,\ldots\ldots,1}_{30},\underbrace{0,\ldots\ldots,0}_{470})',$$

$$\boldsymbol{\beta}_2 = (\underbrace{1,\ldots\ldots,1}_{20},\underbrace{0,\ldots\ldots,0}_{480},\underbrace{1,\ldots\ldots,1}_{20},\underbrace{0,\ldots\ldots,0}_{480})',$$

$$\boldsymbol{\beta}_3 = (\underbrace{1,\ldots\ldots,1}_{10},\underbrace{0,\ldots\ldots,0}_{490},\underbrace{1,\ldots\ldots,1}_{10},\underbrace{0,\ldots\ldots,0}_{490})'.$$

For each scenario we consider (a) the NEG prior without the grouping, (b) the correct grouping (the perfect separation of predictive blocks), and (c) the imperfect grouping according to $\mathcal{Q}_1 = \{1,\ldots,40\}$, $\mathcal{Q}_2 = \{41,\ldots,500\}$, $\mathcal{Q}_3 = \{501,\ldots,540\}$, $\mathcal{Q}_4 = \{541,\ldots,1\,000\}$. We consider a covariance matrix $\Sigma = \left\{\sigma_{ij} = 0.5^{|i-j|}\right\}_{i,j=1}^{p}$ to generate predictors from $\mathrm{N}_p(\mathbf{0},\Sigma)$.

	Perfect Grouping			**Imperfect Grouping**			**NEG**	
Sparsity	FD	FN	FDH	FD	FN	FDH	FD	FN
$\boldsymbol{\beta}_1$	44.2	10.6	0	52.8	20.9	9	53.5	34.6
$\boldsymbol{\beta}_2$	49.9	4.4	0	60.6	17.8	11.4	57.1	21.2
$\boldsymbol{\beta}_3$	52.6	0.2	0	60.9	2.9	19.8	53.1	3.7

Table 7: Simulation study with different degrees of sparsity. FD/FN/FDH stand for false discoveries/false non-discoveries/false discoveries after the hierarchical variable selection

Results are summarized in Table 7, where the average numbers of false discoveries, false non-discoveries and false discoveries after applying the hierarchical selection are reported from 10 simulated repetitions. The number of false non-discoveries remains the same after the hierarchical selection.

We clearly see the benefit of including the grouping in the reduction of false non-discoveries. The lowest number is seen for the correct grouping, followed by the imperfect grouping and then by the plain NEG prior. The model with the correct grouping has consistently the lowest number of false discoveries, which even drop down to zero after the hierarchical selection. Regarding the false discoveries, the NEG prior benefits from the incorrect grouping only after the hierarchical selection. The exception was the least sparse model associated with $\boldsymbol{\beta}_1$ in Table 7. As explained in the manuscript, the model without the scaling tends to increase the number of within group false discoveries in the sparse groups. It is worth noting that the NEG prior without the grouping performs well in very sparse situations (viz. the sparsity pattern associated with $\boldsymbol{\beta}_3$ in Table 7 and also simulated examples in our manuscript).

Appendix F: Complete description of gene/pathway information

Glycerophospholipid metabolism	Phosphatidylinositol signaling system
Protein processing in endoplasmic reticulum	mTOR signaling pathway
ECM-receptor interaction	Adherens junction
Complement and coagulation cascades	RIG-I-like receptor signaling pathway
Intestinal immune network for IgA production	Insulin signaling pathway
Aldosterone-regulated sodium reabsorption	Salivary secretion
Gastric acid secretion	Prion diseases
Prostate cancer	Systemic lupus erythematosus
Hypertrophic cardiomyopathy (HCM)	

Table 8: Pathways identified by the overlapping group LASSO

TNFRSF7	**CASP5**	**CAMK2D**	SLIT1	**CX3CL1**	**PRKCG**	WNT4	**ZAK**	**CLDN11**	**KLKB1**	**ITGB7**	ITGAL	**IRF3**	INPP5D	LAMA1	FRAP1	**FOXO1A**	DFFB	**CTNNB1**	**COMP**	**ISGF3G**		$\widehat{\beta}$
·	·	·	·	·	·	·	·	·	·	×	·	·	·	×	·	·	·	·	×	·	ECM-receptor interaction	0.36
·	·	×	·	·	×	×	·	·	·	·	·	·	·	·	·	·	·	×	·	·	Melanogenesis	0.341
·	·	·	·	·	·	·	·	·	·	·	×	·	·	·	·	·	·	·	×	·	Malaria	0.297
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	×	·	·	·	·	·	Type II diabetes mellitus	0.226
·	·	·	·	·	×	·	·	×	·	·	×	·	·	·	·	·	·	×	·	·	Leukocyte transendothelial migration	0.188
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	×	×	·	×	·	·	Prostate cancer	0.174
·	·	·	·	·	×	·	×	×	·	·	·	·	·	·	·	·	·	×	·	·	Tight junction	0.153
·	×	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	NOD-like receptor signaling pathway	0.123
·	·	×	·	·	×	·	·	·	·	·	·	·	·	·	×	·	·	·	·	·	Glioma	0.102
·	·	·	·	·	·	·	·	×	·	·	·	×	·	·	·	·	·	·	·	×	Hepatitis C	0.089
·	·	·	·	·	×	·	·	·	·	·	·	·	×	·	·	·	·	·	·	·	Fc gamma R-mediated phagocytosis	0.07
·	·	·	·	·	·	·	·	·	×	·	·	·	·	·	·	·	·	·	·	·	Complement and coagulation cascades	0.062
·	·	·	·	·	·	·	·	·	·	·	·	×	·	·	·	·	·	·	·	·	Cytosolic DNA-sensing pathway	0.048
·	·	·	·	·	×	·	·	·	·	·	·	·	×	·	·	·	·	·	·	·	Phosphatidylinositol signaling system	0.036
·	·	·	·	·	·	·	·	·	·	·	·	·	×	·	×	×	·	·	·	·	Insulin signaling pathway	0.029
·	·	·	·	·	·	·	·	×	·	×	×	·	·	·	·	·	·	·	·	·	Cell adhesion molecules (CAMs)	0.019
·	·	·	·	·	·	×	·	·	·	·	·	·	·	·	·	·	·	×	·	·	Basal cell carcinoma	0.016
·	·	·	×	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	Axon guidance	0.011
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	×	·	·	·	Apoptosis	0.011
×	·	·	·	×	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	Cytokine-cytokine receptor interaction	0.008
·	·	·	·	·	·	·	·	·	·	×	·	·	·	·	·	·	·	·	·	·	Intestinal immune network for IgA production	0.003

Table 9: Results obtained from NEG grouping model. Table reports the involvement of 21 identified genes within 21 identified pathways.