Bayesian Variable Selection

in High-dimensional Applications

Veronika Ročková

2013

Bayesian Variable Selection in High-dimensional Applications

ISBN: 978-90-9027731-8

The work presented in this thesis was performed at the Department of Biostatistics and Department of Hematology at the Erasmus Medical Center in Rotterdam and was partially financially supported by the Center for Translational Molecular Medicine (CTMM).

Layout: Veronika Ročková
Printing: Silueta s.r.o. Pardubice

Bayesian Variable Selection in High-dimensional Applications

Bayesiaanse variabele selectie in hoog-dimensionale toepassingen

Proefschrift

ter verkrijging van de graad van doctor aan de

Erasmus Universiteit Rotterdam

op gezag van de rector magnificus

Prof.dr. H.A.P. Pols

ingevolge het besluit van het College voor Promoties

in het openbaar te verdedigen op

dinsdag 12 november 2013 om 11:30 uur

door

Veronika Ročková

geboren op 1 augustus 1985
te Pardubice (Tsjechië)

Promotiecomissie

Promotoren:

Prof.dr. E.M.E.H. Lesaffre
Prof.dr. B. Löwenberg

Overige leden:

Prof.dr. R. Paap
Prof.dr. C.J.F. ter Braak
Prof.dr. Z. Shkedy

"In the time of your life, live – so that in that good time there shall be no ugliness for yourself or for any life your life touches. Seek goodness everywhere, and when it is found, bring it out of its hiding place and let it be free and unashamed.

In the time of your life, live – so that in that wondrous time you shall not add to the misery and sorrow of the world, but shall smile to the infinite delight and mystery of it."

William Saroyan

Dedicated to my beloved parents
Ilona and Václav

CONTENTS

CHAPTER 1

GENERAL INTRODUCTION

1.1 Perspectives on High-Dimensional Variable Selection

Advances in research technologies over the past few decades have encouraged the proliferation of massive datasets, revolutionizing statistical perspectives on high-dimensionality. High-throughput technologies have become pervasive in diverse scientific disciplines and continued to generate data of increasingly complex phenomena, altering the course of statistical developments both in methodology and theory. A major focus of the intensive methodological research has centered around *variable selection*, which has become fundamental to knowledge extraction from such challenging data.

The problem of variable selection refers to the statistical endeavor of selecting a subset of observed characteristics, which collectively provide a good description of an observed phenomenon. Of particular interest are settings where such a subset is parsimonious. We take the perspective of regarding variable selection as a special form of model selection within a given regression framework, where models differ in their configuration of the contributing variables. Diverse model optimality criteria can be specified to tailor variable selection to a specific problem at hand. However, there are two main tasks in high-dimensional statistical analysis, where variable selection has become essential to knowledge discovery: (a) construction of an effective method to predict future observations, (b) accurate estimation of model parameters in order to gain insights on the contributions of individual variables to the response. Achieving both of these goals simultaneously is typically not possible, since prediction accuracy is often compromised by conciseness and interpretability of the data analyzer. Optimal prediction will be rarely achieved by a single regression model without some form of model averaging. In this respect, variable selection cannot and should not be regarded a general-purpose technique but rather as a means to find useful middle ground or solutions for a specific purpose. Conceptually related methodological developments encompassing such versatile variable selection solutions have occurred in the context of penalized likelihood estimation and Bayesian variable selection, the latter being the focus of this thesis.

High-dimensional data are plentiful in contemporary research disciplines traversing fields as diverse as computational biology or financial risk management. The statistical characterization of high-dimensionality describes the property of growing data dimension along with sample size, where the number of measured attributes typically greatly exceeds the number of observations. Such settings have necessitated reconsideration of traditional asymptotics as well as systematic investigation of finite-sample operational characteristics. Performance properties of statistical procedures can be characterized by aspects as important as accuracy of statistical inference or computational complexity. Whereas in the familiar situation, when n (number of observations) exceeds p (number of variables), neither of these two properties needs to be sacrificed for the benefit of the other. The reverse scenario (p much larger than n) has required careful design of statistical procedures and in depth understanding of their strengths and limitations. The challenge high-dimensional data pose result from multiple, in-

tricate factors. The data size demands (a) examination of the excess limits in dimensionality, where such methods are no longer meaningful to consider, (b) characterization of relevant optimality attributes of variable selection procedures, and (c) implementation of reliable inferential tools that scale efficiently with the dimensionality. In this thesis we focus on this third objective and address how deterministic methods can be used to provide flexible computationally efficient analogues to stochastic methods for Bayesian learning in high-dimensional variable selection.

In the instance of a limited amount of data, there is concern that the dimensionality of parametrization will still yield an adequate representation of the phenomenon. One of the crucial assumptions, which facilitates statistical inference in such situation, is the one of sparsity, where the regression function can be parametrized using only a few coefficients, which correspond to the essential covariates that should not be disregarded by the model. The notion of sparsity is central to the implementation of variable selection, which can provide insights into the properties of the observed phenomenon and effectively recover a sparse underlying structure, whenever it exists. The assumption of sparsity is not unreasonable in many practical contexts including genomic applications, where it is generally believed that only a fraction of measured genomic features actually impact the observed response. The analysis of genomic data is a recurrent theme in this thesis, in which we provide numerous demonstrations of practical instances where variable selection generates meaningful interpretation of the data and where sparsity is in concordance with biological intuition.

In the remainder of this chapter, we lay down the groundwork for the forthcoming chapters. We begin by describing the principle of penalization and its connection to Bayesian regularization and variable selection, ideas which will reappear in the next chapters. We confine our explanation to the basic principles, leaving more detailed outlines to the introductory paragraphs of each chapter. The connecting thread throughout this thesis is the proposal of deterministic computational methods for rapid posterior calculations in Bayesian shrinkage estimation and variable selection, crossing the borders from the linear regression framework to multivariate statistical techniques for retrieving sparse genomic networks.

1.2 The Statistical Concept of Penalization

We will be dealing mostly with the case of multiple linear regression, where it is of interest to find a linear model representation for a $(n \times 1)$ response vector Y in terms of a subset of potential predictors $X = [X_1, \ldots, X_p]$. Penalized likelihood methods for inferring the active variables set rely on the full model specification $Y \sim \mathrm{N}_n(X\beta, \sigma^2 \mathrm{I}_n)$, where redundant variables are eliminated by determining which regression coefficient estimates are zero. Such regularized sparse solutions are obtained by constraining the set of admissible coefficient vectors, where the the boundary optima possess the variable selection property. In situations when $p > n$, some restrictions need to be imposed on the model solutions in order to guarantee problem determinacy. Apart from sparsity, various requirements can be induced to reflect

personalized preferences on the solutions, such as limited model size, limited length of the regression vector, smoothness among coefficients or linear constraints for cost-restricted variable selection. The constrained optimization is typically solved by the method of Lagrange multipliers, where the Lagrangian corresponds to the penalized log-likelihood function. One form of the penalized log-likelihood problem in linear regression arises as solving the following optimization

$$\max_{\beta\in\mathbb{R}^p}\left\{-\frac{1}{2}||Y-X\beta||^2-\sum_{j=1}^{p}\text{pen}_\lambda(|\beta_j|)\right\}, \tag{1.2.1}$$

where $||\cdot||^2$ denotes the l^2 norm and $\text{pen}_\lambda(\cdot)$ designates the penalty function indexed by the regularization parameter $\lambda\geq 0$. The last few decades of intensive statistical research have witnessed an explosion of penalized likelihood approaches, as is evidenced by the emergence of increasingly intricate penalties motivated by arguments from asymptotic theory. A lingering issue is to characterize non-asymptotic justifications for these procedures.

By finding the optimum of the penalized likelihood function (1.2.1), we hope to simultaneously perform variable selection (by determining the nonzero coefficient estimates) as well as estimate the associated regression coefficients with as little bias as possible. Of interest are penalties that posses the variable selection property such as the l^0 penalty $l^0_\lambda(|\beta_j|)=\lambda\,\text{I}(|\beta_j|\neq 0)$, which arises naturally in many classical model selection methods such as AIC or BIC. However, computation of the l^0 optimization problem has NP complexity due to the many combinatorial possibilities when evaluating all 2^p model configurations. Pragmatic alternatives have emerged in the form of continuous approximations to the l_0 penalty, such as the bridge penalty $l^q_\lambda(|\beta_j|)=\lambda\,|\beta_j|^q$ for $0<q\leq 2$ (Frank and Friedman, 1993), which bridges the best subset l^0 regularization and the l^2 ridge regression. The particular case for $q=1$ corresponds to the LASSO variable selection (Tibshirani, 1994), which has become one of the benchmark feature extraction methods.

With the availability of so many penalty function it has become easy to be misled. To address this issue, Fan and Li (2001) advocated penalty functions fulfilling the following three properties: (a) *sparsity*, where the estimator automatically sets unimportant coefficients to zero and thereby accomplishes variable selection; (b) *unbiasedness*, where the estimator avoids overshrinkage of large effects and thereby avoids unnecessary modeling bias; (c) *continuity* in data, which makes the estimator robust against small perturbations in the data which may cause instability in prediction. Antoniadis and Fan (2001) characterized formally the three conditions in terms of behavior of the penalty function. Namely, (a) sparsity is guaranteed whenever $\min_{t\geq 0}\{t+\text{pen}'_\lambda(t)\}>0$, (b) the near unbiasedness occurs whenever $\text{pen}'_\lambda(t)\to 0$ with $t\to\infty$, (c) continuity applies if and only if $\arg\min_{t\geq 0}\{t+\text{pen}'_\lambda(t)\}=0$, where $\text{pen}_\lambda(\cdot)$ is nondecreasing and continuously differentiable on $\mathbb{R}^+$. According to these criteria, a good penalty function should have a singularity at the origin (needed to generate sparse solutions) and should be concave to reduce estimation bias. It is well recognized that bridge penalties fall short in fulfilling all these three conditions simultaneously (Fan and Lv, 2010). The l^q penalty does not satisfy: the sparsity condition for $q>1$, the unbiasedness

condition for $q = 1$, the continuity condition for $0 \leq q < 1$. This observation has motivated proposals for more elaborate penalties which do possess all three desirable characteristics (such as the smoothly slipped absolute deviation penalty (SCAD) of Fan and Li (2001) or the minimax concave penalty (MCP) of Zhang (2010). Under certain conditions on the concave penalty functions and with the regularization parameter behaving appropriately, the penalized likelihood estimators guarantee the variable selection consistency property and asymptotic normality, which serves as a prerequisite for uncertainty assessment, both in the finite parameter case (Fan and Li, 2001) and with a diverging number of parameters (Peng and Fan, 2004).

Similar criteria can be applied for the assessment of Bayesian procedures and will be examined in **Chapter 4** in the context of Bayesian shrinkage priors.

1.3 Bayesian Analogues to Penalization

A wide spectrum of sparsity inducing penalty functions emerge naturally in hierarchical Bayesian models for shrinkage estimation. This follows from the well known observation that penalized likelihood estimation can be cast as the problem of maximum a posteriori (MAP) estimation within the Bayesian framework, the advantage there being that one can leverage the extensive methodology developed for this field. Determining the shrinkage properties then transfers to the study of properties of prior distributions on the regression coefficients $\pi(|\beta_j|)$, where it is the part of $\log \pi(|\beta_j|)$ depending on $|\beta_j|$ that gives rise to the frequentist penalties.

Preferable have become heavy-tailed prior densities arising as scale mixtures of normals, which bypass implementation difficulties by allowing for computationally tractable Bayesian mechanisms, both stochastic and deterministic. In conjugate Bayesian linear models $Y \sim \mathrm{N}_n(X\beta, \sigma^2 \mathrm{I}_n)$ with a hierarchical prior on the regression coefficients $\beta \sim \mathrm{N}_p(0_p, \sigma^2\lambda^2)$ with $\lambda^2 \sim \pi(\lambda^2)$ distributed accordingly, the posterior distribution $\pi(\beta, \sigma \,|\, y)$ corresponds to recognized penalized likelihood methods (1.2.1). A broader overview of such prior specifications is postponed until **Chapter 2**. As a modification to the conjugate formulation, many authors recommend replacing σ^2 in the prior variance for regression coefficients by a so-called global shrinkage parameter τ^2, which is assigned an independent distribution inducing substantial mass near zero (Polson and Scott, 2010). Such "global-local" specifications are often better suited for negotiating underlying sparsity, where the small global parameter pulls unimportant coefficients towards zero and the heavy tails of the local parameter simultaneously provide enough support for large coefficients to escape the gravitation and avoid over-shrinkage. Recent proposals include among others (Carvalho and Polson, 2010; Armagan et al., 2012) variants of normal-exponential distributions (Griffin and Brown, 2012). An extension of the Normal-Exponential-Gamma (NEG) prior to account for grouping information is proposed in **Chapter 4**.

1.4 Bayesian Variable Selection Priors

The absolutely continuous shrinkage priors outlined in the previous section offer computational advantages allowing for the implementation of standard deterministic or stochastic search inferential techniques. However, although the MAP parameter estimates associated with shrinkage priors are sparse, stochastic search samples from the posterior distribution are not. This may create practical difficulties when it is of interest to determine which parameters are exactly zero. Better characterization of variable selection uncertainty is obtained by priors which induce positive probability on sparse solutions. Such prior specifications allow uncertainty assessment not only around coefficient estimates, but also around the models themselves. A natural choice of such a prior is the “spike and slab" two component mixture prior, where the first component drives the coefficients to zero and the second component allows for nonzero entries. The mixing proportion between the two components can be regarded as an analogue to the global shrinkage parameter, quantifying probabilistically the degree of overall sparsity and determining the shrinkage properties of the prior. The accurate characterization of sparsity within the spike and slab framework demands the spike distribution to be concentrated at zero, whereas the slab constitutes is a uniform proper prior or a heavy tailed distribution. Although labeled as a methodological ideal (Carvalho and Polson, 2010), the point mass spike and slab formulation poses significant computational difficulties. Practically useful relaxations have been proposed which replace the Dirac delta spike at zero by a continuous prior with a small variance (the Stochastic Search Variable Selection (SSVS) prior of George and McCulloch (1993)). A more detailed overview of spike and slab formulations is presented in **Chapter 2**. In **Chapter 3** we adopt the Bayesian variable selection perspective using the SSVS continuous relaxation of the point mass prior and formulate an adaptation of the EM algorithm for rapid, dynamic posterior inference.

1.5 Computational Aspects

1.5.1 Non-convex Optimization for Bayesian MAP Estimation

Computation in penalized likelihood problems, or equivalently Bayesian posterior mode finding, is challenging unless the penalty function is convex, in which case the problem translates as a convex optimization task that is easily addressed by existing powerful algorithms (Efron et al., 2004). As one possible approach to solving (1.2.1) for the case of concave penalties, Zou and Li (2008) propose to proceed iteratively by solving a series of reweighted convex penalization problems using either local linear or local quadratic approximations to the penalty function.

The latter entails approximating the penalty by a quadratic function at the current parameter vector $\beta^{(k)}$, where for each $\beta_j \approx \beta_j^{(k)}$ we have

$$\text{pen}_\lambda(|\beta_j|) \approx \text{pen}_\lambda(|\beta_j^{(k)}|) + \frac{\text{pen}'_\lambda(|\beta_j^{(k)}|)}{2|\beta_j^{(k)}|}\left(\beta_j^2 - \beta_j^{(k)2}\right). \tag{1.5.2}$$

Substituting the term which depends on β_j on the right hand side of (1.5.2) into the penalized likelihood (1.2.1), we obtain generalized ridge regression with coefficient-specific penalties that depend on the derivative of the penalty function, which admits a closed form solution. Hunter and Li (2005) note that the local quadratic approximation (LQA) is an instance of the EM algorithm (Dempster et al., 1977) and more generally a minorization/maximization algorithm, by pointing out that the approximated penalized likelihood is a convex minorizing function. For various penalty functions arising from hierarchical Bayesian shrinkage priors, Griffin and Brown (2005) formulate an EM algorithm, which corresponds to LQA. An extension of this algorithm to account for grouping among predictors is proposed in **Chapter 4**. Whereas LQA has been considered exclusively for absolutely continuous shrinkage priors, in **Chapter 3** we note that LQA approximation can be exploited also for continuous spike and slab mixture priors. There we propose a novel dynamic model exploratory mechanism based on the EM algorithm for simultaneous MAP estimation and posterior model mode detection. We refer to our proposed method as EMVS, the EM approach to variable selection. The core practical ingredient in the EMVS procedure is expeditious updating of ridge regression solutions, where exact answers may be too costly if both n and p are considerably large. In **Chapter 3** we describe a variant of the EM algorithm, which involves rapid approximative ridge solutions obtained with the assistance of a conjugate stochastic dual coordinate ascent algorithm (Shalev-Shwartz and Zhang, 2013).

An alternative to LQA can be obtained by approximating the penalty locally by a linear function, where for each $\beta_j \approx \beta_j^{(k)}$ we have

$$\text{pen}_\lambda(|\beta_j|) \approx \text{pen}_\lambda(|\beta_j^{(k)}|) + \text{pen}'_\lambda(|\beta_j^{(k)}|)\left(|\beta_j| - |\beta_j^{(k)}|\right). \tag{1.5.3}$$

This approximation replaces the ridge optimization at every iteration by the adaptive LASSO optimization, for which efficient computational methods exist (Efron et al., 2004). In continuous sparsity priors, the linear approximation is of great practical advantage since it possesses variable selection property by generating solutions with zeroes at every iteration. Similarly as LQA, the local linear approximation (LLA) also corresponds to the EM algorithm (Hunter and Li, 2005), where it provides the minimum (tightest) convex minorant of the concave objective function. A particular implementation of LLA is proposed in **Chapter 4** in the context of Bayesian group shrinkage estimation.

1.5.2 Sampling Schemes

Although the analytical intractability of the posterior distributions in Bayesian variable selection and shrinkage estimation precludes exact inference, approximate answers can be obtained by sampling from the posterior with the assistance of MCMC methods (Robert and Casella, 1999), which have become pervasive in Bayesian inference. Whereas Bayesian shrinkage models often admit efficient implementation of the Gibbs sampler (Geman and Geman, 1984), posterior calculations in spike and slab models are typically more involved, since they entail simultaneous exploration of parameter and model space and face difficulties in traversing dimensions. George and McCulloch (1993) were the first to introduce the Gibbs sampler in the context of spike and slab variable selection, where they laid down the foundations for stochastic model search. In order to avoid expensive updating of the regression coefficient vector in high-dimensions, Smith and Kohn (1996) and George and McCulloch (1997) suggested integrating over the regression parameters to sweep only through the model space. One-site Gibbs samplers and Metropolis Hastings routines (George and McCulloch, 1997; Madigan et al., 1994) have been successively applied to rapidly evaluate posterior model selection uncertainty in problems of a manageable size. Spatial dependence between variable inclusion probabilities was introduced in the context of the Ising prior by Smith et al. (2003), Smith and Fahrmeir (2007) and more generally by Li and Zhang (2010), who considered predictors forming an undirected graph and characterized one-site Gibbs sampling algorithm. Goldsmith et al. (2013) considered a variant of their sampler, which does not require costly computation of matrix determinants. A more detailed overview of recent methodological contributions in stochastic search is given in **Chapter 2**.

Despite these elaborations of stochastic search posterior model inference techniques, the practicality of their implementation has remained a challenge in high-dimensional applications. In **Chapter 3** we present a deterministic model exploration tool, which is based on the EM algorithm (EMVS procedure) and which identifies high-probability models at a fraction of time required for MCMC computation.

1.6 Structured Variable Selection

Variable selection procedures which treat each explanatory variable individually fail to account for knowledge on existing structural organization among the predictors such as hierarchy, network topology, competitive predictors or grouping. In the context of penalized likelihood estimators, custom-made penalties have been proposed, which induce (a) similarity in coefficients that are neighbors on a lattice or an undirected graph (Li and Li, 2008; Pan et al., 2010), (b) homogeneity in within-group coefficients (Meier et al., 2008) or (c) hierarchical constraints using structured analogues of the LASSO penalty (Choi et al., 2010). The implicit assumption in many of these methods being that related regression coefficients are similar or at least sign-consistent, which may not be realistic in many practical situations (Li and Zhang,

2010). Another known problem associated with LASSO-based approaches for grouped variable selection (Meier et al., 2008; Jacob et al., 2009) is their inability to produce estimates that are sparse within groups. Recent proposals have corrected for this (Friedman et al., 2010a) by incorporating an additional within group penalty. Another approach is presented in **Chapter 4**, which embeds grouping (possibly with overlap) into sparsity inducing regularization using a particular Bayesian shrinkage prior.

Whereas incorporation of prior knowledge on the sparsity patterns is less intuitive in the penalized likelihood framework, Bayesian variable selection constitutes a coherent framework for transmitting such information by re-distributing probabilities over the model space. Before proceeding, it is useful to distinguish between two types of prior structural knowledge.

The first type arises in the form of strict structural constraints as dictated by commonly used principles of a model building process such as heredity when dealing with interactions (Choi et al., 2010; Chipman, 1996). For instance, the strong inheritance principle requires the presence of both main terms in order that the interaction be allowed in the model (Yuan et al., 2009). Another situation, admitting only a subset of model configurations, arises in determining the order of auto-regression in transition longitudinal models, where configurations that are not triangular (monotone with no intermittent patterns of zeroes) are typically not allowed. Limiting the size of the model as well as dealing with competing (mutually exclusive) predictors also imposes zero probabilities on certain models. Such strict constraints may hamper the required irreducibility of some model states and thereby complicate posterior calculation using sampling techniques.

Instead of considering models as either permitted or allowed, the structural information can be used to smoothly re-distribute the prior distribution on the model space, so that certain combinations of predictors are more likely to occur together. The underlying assumption there being that predictors are more probable to be true positives if they cluster within groups or are connected on a graph. Structured model prior distributions have been proposed in the Bayesian variable selection literature, where spatial or within-group smoothing is induced on variable selection probabilities (Stingo and Vannucci, 2011; Li and Zhang, 2010). Introducing smoothness at the penalty level rather than within the regression coefficients is a distinguishing feature of Bayesian analogues to structured penalized likelihood methods.

In **Chapter 3** we discuss the Ising prior on the model space for covariates that lie on an undirected graph, as well as the independent logistic regression prior to incorporate grouping in the context of the EMVS procedure. We demonstrate that our proposed algorithm gears the model search towards models that are more homogeneous with respect to the underlying structural architecture. In **Chapter 5** we propose a structured prior on the model space, where competitiveness is introduced in the variable inclusion probabilities. This prior is considered in the context of clustering multiple responses in the sparse factor modeling framework.

1.7 Variable Selection in Biomedical Sciences

The practical relevance of variable selection in biomedical sciences is enormous, encompassing tasks as important as biomarker discovery, prognostic assessment, design of risk-stratification rules, or modeling association networks. Conclusions drawn from variable selection may have serious practical implications, particularly in disciplines where statistical analysis is followed by a series of costly validation experiments. In cancer research, variable selection generates valuable targets for therapeutic decisions, where careful considerations need to be exercised to draw valid conclusions from data that may not be optimally powered. The mainstream statistical practice typically requires implementation of standard tools, where novel approaches have only begun to permeate. The proliferation of the Bayesian methods, particularly in genomic applications, is increasing and will be accelerated with the emergence of rapid computational techniques. With the availability of so many inferential methods, it has become difficult to establish a single preferential tool and even simplistic methods can very often lead to valid conclusions. Towards the end of the thesis we turn to the clinical and bioinformatics applications of variable selection and we present one standard analysis in **Chapter 6**, which has had profound clinical implications.

This thesis presents numerous analyses using data on patients diagnosed with acute myeloid leukemia (AML). The past decades of intensive biomedical research at the department of hematology at Erasmus Medical Centre in Rotterdam gave rise to a unique collection of high-throughput and clinical data on what is undoubtedly one of the biggest cohorts of patients with this very rare disease. We had the excellent opportunity to combine expertise in both statistics and biology to enhance existing knowledge about AML.

1.7.1 Variable selection in Acute Myeloid Leukemia

Acute myeloid leukemia describes a heterogeneous group of hematopoietic disorders, which are collectively characterized by a proliferation of immature myeloid blood cells. The past years of intensive research have accumulated a large body of evidence for multifactorial pathogenesis of AML. The multiple contributing factors engage molecular mechanisms as diverse as epigenetic alterations, cytogenetic abnormalities and other genetic aberrations leading to impaired expression of oncogenic genes. The characterization of molecular processes underlying AML is far from being completed, as is evidenced by the continuous emergence of novel prognostic markers. The intricate biological mechanisms involved in AML pathogenesis have engaged biomedical researchers for decades and despite a lot has been learned there is still far more to be discovered. A role of a statistician in these efforts has been instrumental in (a) providing supporting evidence about validity of biological hypotheses and (b) generating targets for biological validation. The first of the two objectives is exemplified by an integrative analysis of established molecular and genomic markers in **Chapter 6**. With the assistance of variable selection, we discern a few relevant markers that are capable of

stratifying otherwise very heterogeneous patients into two populations with similar survival outcomes. The result of this analysis has had impacts as important as including one particular marker in a daily diagnostic practice. The analysis is revised in **Chapter 2**, where consistent findings were generated using Bayesian methods. The second objective is exemplified by an analysis in **Chapter 5**, where we set out to discover associations between two sets of genomic features. Recent studies have begun associating microRNAs with specific AML regulatory mechanisms. MicroRNAs are negative regulators of gene expression, decreasing the stability of target RNAs or limiting their translation (Fabian et al., 2010). The AML dataset provides a set of snapshot gene and microRNA expression measurements. This rich collection of expression data has motivated our work presented in **Chapter 5**, where by proposing a novel integrative model in conjunction with the fast EM algorithm for variable selection, we introduce a new framework for Bayesian learning about the likely “dynamics" of the microRNA mediated gene regulation.

CHAPTER 2

HIERARCHICAL BAYESIAN FORMULATIONS FOR SELECTING VARIABLES IN REGRESSION MODELS

Adapted version of a research article:
Rockova, V., Lesaffre, E., Luime, J. and Löwenberg, B. 2011. **Hierarchical Bayesian formulations for selecting variables in regression models**. *Statistics in Medicine* 31:213-232

Abstract

The objective of finding a parsimonious representation of the observed data by a statistical model that is also capable of accurate prediction is commonplace in all domains of statistical applications. The parsimony of the solutions obtained by variable selection is usually counterbalanced by a limited prediction capacity. On the other hand, methodologies that assure high prediction accuracy usually lead to models that are neither simple nor easily interpretable. Regularization methodologies have proven to be useful in addressing both prediction and variable selection problems. The Bayesian approach to regularization constitutes a particularly attractive alternative as it is suitable for high-dimensional modeling, offers valid standard errors and enables simultaneous estimation of regression coefficients and complexity parameters via computationally efficient MCMC techniques. Bayesian regularization falls within the versatile framework of Bayesian hierarchical models, which encompasses a variety of other approaches suited for variable selection such as spike and slab models and the MC^3 approach. In this chapter, we review these Bayesian developments and evaluate their variable selection performance in a simulation study for the classical small p large n setting. The majority of the existing Bayesian methodology for variable selection deals only with classical linear regression. Here we present two applications in the contexts of binary and survival regression, where the Bayesian approach was applied to select markers prognostically relevant for the development of rheumatoid arthritis and for overall survival in acute myeloid leukemia patients.

2.1 Introduction

The simultaneous assessment of the associations between multiple disease factors and a health outcome is an important topic in epidemiological research. The two fundamental objectives implicit in these investigations are: (a) determining which predictors are prognostically or diagnostically important, (b) selecting a combination of factors capable of accurate prediction of the disease outcome. The two goals are somewhat at odds with each other. Models that possess high prediction accuracy are usually not easily interpretable, might even contain insignificant variables and their estimated effects may be biased (Copas, 1983). When the focus shifts from prediction to explanation, usually parsimonious models are preferred consisting of only variables that are truly influential for the outcome. Finding such a model in the regression framework can be recast as a problem of variable selection.

The customary variable selection strategies involving the sequential search (forward selection, backward elimination or stepwise selection) or all-subset regression using different optimization criteria have several well-acknowledged deficiencies. They become increasingly ineffective and impractical in higher dimensions and they exhibit high sensitivity towards small changes in the data (Breiman, 1996; Fan and Li, 2001). The stepwise selection procedures are also prone to getting trapped in locally optimal models (Hocking, 1976) and face

problems in designs with complex patterns of multicollinearity (Hans et al., 2007). Despite the drawbacks, they are still the immediate choice in routine data analysis.

Recently, a great deal of attention has been devoted to the development of different regularization methods for simultaneous variable selection and coefficient estimation (Tibshirani, 1994; Yuan and Lin, 2006; Zou and Hastie, 2005). The statistical concept of regularization can be vaguely characterized as imposing additional requirements on the regression solutions in that the more "useful" solutions are preferred over other ones. What is meant by "useful" depends on the purpose. If variable selection is the ultimate goal, sparse solutions (i.e. solutions with the redundant coefficients effectively zeroed out) are more desirable. The preference requirements can take the form of restrictions on the space of the solutions (which is equivalent to imposing the frequentist penalty term to the log-likelihood being maximized) or, in a Bayesian way, putting a suitable prior on the regression coefficients.

The two regularization concepts are closely related to each other. The general principle behind the frequentist regularization is to maximize $\text{logLik}(\theta|y) - \text{pen}(\theta)$ with respect to the vector of unknown parameters $\theta = (\theta_1, \ldots, \theta_q)'$, where $\text{logLik}(\cdot)$ denotes the logarithm of the likelihood and $\text{pen}(\cdot)$ is a regularization term, which controls the complexity of the solution. The most popular penalty terms are the l_p penalties, $l_p(\theta) = \sum_{i=1}^{q} |\theta_i|^p$, with $p = 1$ (the LASSO penalty (Tibshirani, 1994)) and $p = 2$ (the ridge penalty (Hoerl and Kennard, 1970)). The solution to the penalized maximum likelihood estimation using the l_p penalties possesses a Bayesian interpretation (Tibshirani, 1994; Park and Casella, 2008). It coincides with the mode of the joint posterior distribution of regression coefficients arising from independent individual priors of the form $p(\theta_j|\eta_j) \approx \exp(-\tau_j|\eta_j|^p)$, better known as exponential power priors (Frank and Friedman, 1993; Fu, 1998). However, the fully Bayesian approach to regularization entails evaluation of the whole posterior distribution, rather than finding just its mode. Such exploration is most often achieved by Markov Chain Monte Carlo (MCMC) methodology.

The Bayesian regularization constitutes only a fraction of Bayesian methodology currently available for variable selection. In the Bayesian paradigm, the task of variable selection is recast as parameter estimation in hierarchical models. In fact, the classical variable selection methods based on penalization of likelihood with a fixed multiple of model dimension (e.g. using *AIC*, C_p and *BIC* criteria) can be regarded as special cases of hierarchical Bayesian model selection under a particular class of priors with fixed choices of hyper-parameters (George and Foster, 1997). Alternatively, George and Foster (1997) proposed to estimate the hyper-parameters from the data to obtain adaptive penalty criteria. The versatility of the hierarchical formulations together with the availability of numerous sophisticated MCMC techniques have lead to the development of a variety of Bayesian variable selection strategies (Carlin and Chib, 1995; Ishwaran and Rao, 2003; Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997). The appeal of the Bayesian approach resides in several features: (a) the inference is purely probabilistic, as opposed to the frequentist hypotheses testing, (b) it provides a natural framework for the assessment of model uncertainty and thereby creates a basis for eventual model averaging, (c) it enables the incorporation of past external informa-

tion through priors, (d) it extends naturally to settings with multivariate responses and (e) it is applicable for high-dimensional variable selection ("small n large p" setting).

In this chapter we provide an overview of several Bayesian variable selection methods in the unified framework of Bayesian hierarchical models and we highlight discrepancies and connections between them. The empirical performance (with regard to variable selection accuracy) of the presented Bayesian methods was evaluated and compared to the classical strategies in a simulation study. The results demonstrate that Bayesian variable selection offers improved performance in detecting the true underlying model. The majority of Bayesian developments for variable selection occurred in the context of the classical linear model. The concept can be applied in other regression settings as well. To illustrate the application of Bayesian variable selection in binary and survival regression, we present an application from rheumatoid arthritis and from acute myeloid leukemia.

2.2 Bayesian Hierarchical Formulations for Variable Selection

Consider an outcome random variable Y that we want to relate to the set of explanatory variables $X_1,\ldots,X_p$ by means of a regression model. The regression framework encompasses a variety of modeling platforms for different types of responses (Gaussian, time-to-event, binary), where the distribution of the response is related to the linear combination of covariates in a way which is specific for the type of outcome. Most often, only a subset of the available predictors play an important role in explaining the variability of the response and the goal of the analysis is to identify these variables.

Each regression model is uniquely characterized by a vector of binary inclusion variables $\gamma=(\gamma_1,\ldots,\gamma_p)'$ indicating whether or not the variable enters the model. Each model γ is then characterized by a specific linear combination of covariates of the form $\beta_0+X'_\gamma\beta_\gamma$, where X_γ and β_γ denote subvectors of covariates and model parameters corresponding to the configuration γ and β_0 is the intercept.

In the Bayesian framework, variables are selected based on posterior information obtained from hierarchical mixture models. Given the set of all plausible models $\{\gamma_s : s\in S\}$, the hierarchical setup starts by assigning a prior probability $p(\gamma_s)$ to each of the individual models, proceeds with choosing a prior distribution $p(\beta_\gamma|\gamma=\gamma_s)$ over coefficients within each model and is completed by the specification of the likelihood $p(Y|\beta_\gamma,\gamma)$. The various Bayesian variable selection strategies emerge by considering different prior specifications and by choosing the actual posterior processing strategy.

2.2.1 The "Model Space" Approach

A natural way to compare models is by inspecting the individual posterior model probabilities

$$p(\gamma_s|Y)=\frac{p(Y|\gamma_s)p(\gamma_s)}{\sum_{k\in S}p(Y|\gamma_k)p(\gamma_k)},$$

where

$$p(Y|\gamma_k) = \int p(Y|\gamma_k, \beta_0, \beta_\gamma) p(\beta_0, \beta_\gamma|\gamma) \mathrm{d}(\beta_0, \beta_\gamma) \tag{2.2.1}$$

denotes the marginal likelihood. The posterior model probabilities quantify the posterior evidence for selecting each particular model and as such immediately suggest models with the highest values as suitable candidates. With an increasing number of predictors, the exhaustive evaluation of the whole model space to find these models becomes impractical. As an alternative to the deterministic solutions based on stepwise search (Madigan et al., 1994) stochastic alternatives have been suggested that exploit MCMC techniques to simulate a chain of models to find interesting regions of the model space with an accumulation of posterior mass. The most popular and intuitively appealing MCMC strategy adapted for this setting is MC^3 (Markov Chain Monte Carlo Model Composition) originally proposed in the context of graphical models (Madigan et al., 1995). The procedure results in a sequence $\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(T)}$ of visited models, which are generated according to the Metropolis-Hastings (MH) routine (Metropolis et al., 1953; Hastings, 1970). The MH proposal distribution is concentrated at close proximity of the current state γ, thereby restricting to models differing by an inclusion or exclusion of just one variable. The candidate model γ^* sampled from the proposal distribution is then accepted with probability $\alpha = \min\left[1, \frac{p(\gamma^*|Y)}{p(\gamma|Y)}\right]$. The posterior model ratio is obtainable in closed form in conjugate regression designs. Otherwise, suitable approximations to the marginal likelihood in (2.2.1), e.g. *BIC* approximation, (Schwarz, 1978) can be used.

The prior distribution over models $p(\gamma)$ is an important ingredient in MC^3 and other Bayesian variable selection procedures. The common choice of this prior distribution assumes independence amongst the binary inclusion indicators $\gamma_1, \ldots, \gamma_p$ and follows a product of individual Bernoulli distributions, i.e.$p(\gamma) = \prod_{j=1}^{p} w_j^{\gamma_j}(1-w_j)^{1-\gamma_j}$, where w_j is the prior probability that the j-th variable is in the model. In some hierarchical setups the probability of inclusion w_j is assigned another prior layer. Keeping the parameters w_j fixed and equal to $1/2$, we obtain a uniform prior on the model space.

The actual variable selection can proceed in several ways. Two strategies most often applied in practice are: (a) to pick a model with the highest estimated posterior probability $\widehat{p(\gamma|y)} = \sum_{t=1}^{T} \mathrm{I}(\gamma^{(t)} = \gamma)/T$ (the highest posterior density (HPD) model), (b) to pick variables with estimated posterior marginal inclusion probabilities $\widehat{p(\gamma_k|y)} = \sum_{t=1}^{T} \mathrm{I}(\gamma_k^{(t)} = 1)/T$ higher than 0.5 (the median probability model (MPM) (Barbieri and Berger, 2004)). The appropriateness of HPD model selection was studied by Barbieri and Berger (2004). The authors have shown that in orthogonal linear regression settings the optimal model from a Bayesian predictive viewpoint was the MPM rather than the HPD model.

The model space approach can also be implemented using some other MCMC samplers such as reversible jump (RJ) MCMC (Green, 1995; Gramacy and Pantaleo, 2010; Lunn et al., 2009) that explores simultaneously model and parameter space and dynamically adjusts for the differences in dimensionality of the sampled vectors β_γ. Another strategy, based on the Gibbs sampler, was proposed by Carlin and Chib (1995).

2.2.2 Elaborations on the Model Space Approach

One of the limitations of the stochastic search for variable selection is their inability to escape from steep local posterior peaks, or to discover relevant but isolated regions of the model space. The difficulties with multimodal posterior landscapes can be mitigated with the assistance of population-based MCMC algorithms (Jasra et al., 2007). The main idea is to run a population of chains in parallel, each chain typically associated with a particular "heated/tempered version" of the target distribution. In the model selection context, the natural target distribution is the posterior distribution over the model space. A useful approximation to the posterior model probability $p(\gamma|Y)$ can be obtained using the BIC_γ criterion of a model γ, where $p(\gamma|Y) \approx C \exp(-\frac{1}{2} BIC_\gamma)$. The heated approximate target distribution for a given temperature t is then naturally defined as $p_t(\gamma|Y) \propto \exp\left(-\frac{1}{2t} BIC_\gamma\right)$. The tempering has the effect of flattening the peaks of the true target distribution. The higher the temperature t, the easier it is for the chain to escape from abrupt peaks. Furthermore, the parallel chains interact and learn from each other, making the exploration of the model space more efficient. The interaction is achieved by altering/swapping model configurations between/within the chains with different temperature at each MCMC iteration. Liang and Wong (2000) suggested a hybrid procedure that combines the idea of parallel tempering together with genetic algorithms in the method called Evolutionary MCMC (EMC). See also Bottolo and Richardson (2010) for the application of EMC in Bayesian model selection.

Parallel tempering is closely related to simulated annealing (Kirkpatrick et al., 1983), where only a single chain is used to sample from a joint distribution of the temperatures and the target distribution and where only values with a "zero" temperature are recorded.

Alternative to parallel tempering techniques, Hans et al. (2007) suggested Shotgun Stochastic Search (SSS) that is capable of sampling from vast discrete spaces of regression models. SSS can be regarded as a hybrid procedure that combines Occam's razor principle with Metropolis-Hastings ingredients. Similar to the Occam's window (Madigan et al., 1994), SSS neither focuses on finding a point estimate nor it aims at closely approximating the posterior model distribution. The goal is rather to determine a bigger set of best models. As opposed to Occam's window the search is not entirely deterministic, since the explored models are subject to a randomized proposal mechanism (similarly as in the MH routines). In comparison to one-site Metropolis-Hasting routines (Madigan et al., 1994), in SSS all models from the neighborhood of the current state are evaluated and multiple models are stored at each iteration. In direct parallel to Occam's window, at each iteration the set of best models is deterministically updated by better models found in the neighborhood.

2.2.3 Spike and Slab Models

In many practical situations it is desirable to estimate the values of selected coefficients after the model configuration has been chosen. In MC^3 and related strategies the focus rests purely on variable selection, leaving the inference about model parameters aside. The parameter

estimates then can be obtained by "post-model selection estimation" (using e.g. posterior means or least square estimates). However, such strategy leads to biased estimates as it ignores the uncertainty in the model selection (George and Foster, 1997). Alternatively, one could consider estimates that are not conditional on one selected model but rather averaged over all or highly probable models. Model averaging on the other hand does not provide sparse representation as it yields nonzero estimates of all coefficients regardless of how many of them are actually zero. A convenient solution would be to combine the model averaging and variable selection in one estimation process. This can be achieved in the Bayesian context using the so called variable selection priors.

Variable selection priors, better known as spike and slab priors (George and McCulloch, 1993, 1997; Ishwaran and Rao, 2003), induce a positive prior probability on the hypotheses $H_0 : \beta_k = 0$. In the original formulation (Leamer, 1978; Mitchell and Beauchamp, 1988), the spike and slab distribution is defined as a mixture of a Dirac measure concentrated at zero and a uniform diffuse component. Similarly as in Ishwaran and Rao (2005), we will slightly deviate from the original definition here. By a spike and slab prior we understand any prior that is a mixture of two continuous distributions, implying high prior probability close to zero. These peak-shaped mixtures can be regarded as approximations to the point mass priors, which are computationally feasible for more conventional MCMC samplers. Such priors can be represented as conditionally Gaussian, i.e. normal scale mixtures specified through the prior on hyper-variances. The different variants of the spike and slab formulations emerge by considering different priors for the hyper-variance (a two-point or continuous distribution). We now elaborate in more detail on the two most popular spike and slab priors: Stochastic Search Variable Selection prior of George and McCulloch (1993) and Normal Mixture of Inverse Gamma of Ishwaran and Rao (2003).

2.2.3.1 □ Stochastic Search Variable Selection (SSVS)

Stochastic Search Variable Selection (SSVS) was proposed by George and McCulloch George and McCulloch (1993) for variable selection in the context of linear regression. In SSVS, the model coefficients β_k are assumed to have a mixture prior of "spike" and "slab" Gaussian components. The mathematical formulation of the SSVS hierarchical prior setup is the following:

$$\begin{aligned}
\beta_k|\lambda_k &\sim \mathrm{N}(0,\lambda_k),\\
\lambda_k|c_k,\tau_k^2,\gamma_k &\sim (1-\gamma_k)\delta_{\tau_k^2}(\cdot)+\gamma_k\delta_{c_k^2\tau_k^2}(\cdot),\\
\gamma_k|w_k &\sim \mathrm{Bernoulli}(w_k),\\
w_k &\sim \mathrm{Uniform}[0,1],
\end{aligned}$$

where $\delta_x(\cdot)$ denotes the Kronecker delta concentrated at point x. The "spike" element concentrates closely around zero, reflecting the actual absence of the variable in the model (γ_k equals zero). The "slab" component has a sufficiently large variance to allow the "nonzero" coefficients to spread over larger values.

The degree of separation between the two components is regulated by two tuning parameters τ_k and c_k, where $\tau_k^2 > 0$ is the variance in the spike component and $c_k^2\tau_k^2 > 0$ the variance in the slab component. In order to guide the choice of τ_k and c_k, it helps to note that the two Gaussian densities intersect at the points $\pm\delta_k = \tau_k\varepsilon_k$, where $\varepsilon_k = \sqrt{2(\log c_k)c_k^2/(c_k^2-1)}$. The point δ_k can be regarded as a threshold for declaring practical significance in that all coefficients falling into the interval $[-\delta_k, \delta_k]$ can be interpreted as "practically zero". Given the parameter c_k, the variance τ_k^2 can be selected such that the intersection point reflects our perception of practical significance.

Due to the non-conjugacy, the analytical simplification of posterior distributions $p(\beta_k|y)$ and $p(\gamma_k|y)$ is not tractable. George and McCulloch (George and McCulloch, 1993) suggested a MCMC approximation to the posteriors using the Gibbs sampler, which yields a chain of regression coefficients and visited models $(\beta^{(1)}, \gamma^{(1)}), \ldots, (\beta^{(T)}, \gamma^{(T)})$. Variable selection is then achieved through posterior model probabilities, posterior inclusion probabilities or the posterior distribution of the individual regression coefficients. Processing the MCMC information in $p(\beta_k|y)$ is complicated by the fact that the distribution can be multimodal, which makes the interpretation of posterior summary statistics less meaningful. Nevertheless, in case of strong evidence against the inclusion of the variable, the spike will dominate the posterior which will effectively shrink the posterior mean towards zero. The decision on whether or not a variable enters the model can be done by hard shrinkage (hard thresholding/selection shrinkage) (Fan and Li, 2001; Johnstone and Silverman, 2004), where variables are included whenever the absolute value of the estimated coefficient (e.g. posterior mean) exceeds some threshold value.

One particular variant of SSVS called Gibbs variable selection (GVS) was considered by Dellaportas (Dellaportas et al., 2002), who suggested introducing the binary inclusion indicators also in the likelihood so that only the variables that are literally present in the model contribute to the linear predictor, which now equals $\beta_0 + \sum_{j=1}^{q} \gamma_j\beta_j X_j$. Apart from that, the prior setup for regression coefficients is analogous to SSVS. Examples of an application of SSVS priors in other than linear regression settings can be found in George et al. (1996) and (Ntzoufras et al., 2000).

2.2.3.2 □ Normal Mixture of Inverse Gamma (NMIG)

In the SSVS prior formulation, the variances λ_k have a discrete distribution with a support $\{\tau_k^2, c_k\tau_k^2\}$, which implies a two-point Gaussian mixture prior for the regression coefficient. In the context of linear regression, Ishwaran and Rao (2003) suggested to move the spike and slab element down in the hierarchy and place it on the variances rather than on the regression coefficients. They argued that considering a continuous bimodal distribution for the variance introduces more uncertainty, which might potentially diminish the sensitivity towards the tuning of hyper-parameters. In the original formulation (Ishwaran and Rao, 2003, 2005), the variance was parametrized as a product of two random variables, one having a two point distribution and the second one having an inverse gamma (IG) distribution. Similarly as

Fahrmeir et al. (Fahrmeir et al., 2010) we adopt a different parametrization using a two-point mixture of inverse gammas. This yields the following hierarchical model

$$\begin{aligned}\beta_k|\lambda_k &\sim \mathrm{N}(0,\lambda_k),\\ \lambda_k|v_0,v_1,\gamma_k,a,b &\sim (1-\gamma_k)\mathrm{IG}\left(a,\frac{v_0}{b}\right)+\gamma_k\mathrm{IG}\left(a,\frac{v_1}{b}\right),\\ \gamma_k|w_k &\sim \mathrm{Bernoulli}(w_k),\\ w_k &\sim \mathrm{Uniform}[0,1].\end{aligned}$$

The role of τ_k^2 and c_k in SSVS is now taken by the parameters v_0 and v_1. Ishwaran and Rao (2003) suggested to use $v_1 = 1$ by default for standardized covariates and rescaled responses in the linear model. Similarly as in SSVS the "practical significance" argument can be applied to specify the other hyper-parameters. Note that the marginal prior for the regression coefficients obtained by integrating out the variance is a two-point mixture of scaled t-distributions (with $2a$ degrees of freedom and respective scales $s_1 = \sqrt{\frac{bv_0}{a}}$ and $s_2 = \sqrt{\frac{bv_1}{a}}$). The two densities intersect at the points $\delta = \pm\sqrt{\frac{2a(1-r)}{\frac{r}{s_2^2}-\frac{1}{s_1^2}}}$, where $r = \left(\frac{s_2}{s_1}\right)^{\frac{2}{2a+1}}$. Similarly as in SSVS, the preferred threshold for practical significance can be achieved by a suitable constellation of the hyper-parameters a, b, v_0 and v_1. The extensions to non-Gaussian and hazard rate models were considered by Konrath et al. (2008, unpublished manuscript). For an application in the context of additive regression models see Fahrmeir et al. (2010).

2.2.4 Bayesian Regularization

In spike and slab hierarchies, all possible models are embodied within one hierarchical formulation and the inference for variable selection can be done model-wise or from selection shrinkage. Whereas in the spike and slab formulations the peaked shape of the prior is achieved somewhat artificially by assuming a mixture distribution, it is possible to approximate the spike and slab shape with just one continuous prior component, e.g. using the exponential power priors (Park and Casella, 2008; Box and Tiao, 1973) of the form $p(\beta_j|\eta_j) \approx \exp(-\eta_j|\beta_j|^p)$, where $p > 0$ and η_j is some variance-related parameter. The most popular powered exponential priors are the Laplace prior (Tibshirani, 1994; Park and Casella, 2008) with $p = 1$ and the ridge prior with $p = 2$. If $0 < p \leq 1$, the prior has a singularity at origin, which promotes an intensive shrinkage towards the zero prior mean. For $0 < p \leq 2$, these distributions can be represented as scale mixtures of normals (Andrews and Mallows, 1974). The class of normal scale mixtures has been recognized to generate many popular procedures for regularized regression, most notably the LASSO (Tibshirani, 1994; Park and Casella, 2008), which is equivalent to the MAP estimation under normal/exponential (Laplace) prior. More recent normal scale mixture priors proposed for the shrinkage estimation in linear regression are the normal/gamma (Griffin and Brown, 2010), the normal/Jeffreys

(Bae and Mallick, 2004; Figueiredo, 2002) or the horseshoe prior (Carvalho and Polson, 2010), where the mixing density belongs to the class of inverted beta distributions.

Unlike in the model space or spike and slab approaches, the sparsity approach avoids the specification of priors over models or individual hypotheses $H_{0k} : \beta_k = 0$. The variable selection rests purely on the inspection of the posterior behavior of the model coefficients. The posterior summary measures (mean or median) are never zero with a positive probability and zeroing the redundant variables out then needs to be done through hard shrinkage. Several authors augmented the shrinkage priors to include a point mass at zero (Hans, 2010). Conceptually, these approaches belong to the spike and slab framework discussed in the previous section.

2.2.4.1 □ Bayesian LASSO: The Laplace Prior

The Laplace (LASSO) prior arises as a scale normal mixture assuming exponentially distributed variances (Andrews and Mallows, 1974). A conjugate variant

$$\beta_k|\lambda_k \sim \mathrm{N}(0, \sigma^2\lambda_k),$$
$$\lambda_k|\tau_k^2 \sim \frac{\tau_k^2}{2}\mathrm{e}^{-\lambda_k\tau_k^2/2}\mathrm{I}(\lambda_k > 0),$$

which corresponds to a conditional Laplace prior $p(\beta_k|\sigma^2, \tau_k) = \frac{\tau_k}{2\sigma}\mathrm{e}^{-\tau_k|\beta_k|/\sigma}$, was considered in the context of linear regression by multiple authors including Carlin and Polson (1991), Park and Casella (2008) or Hans (2009). Instead of considering separate shrinkage parameters it is customary to assume that $\tau_1^2 = \cdots = \tau_p^2 = \tau^2$. The parameter τ^2 then takes the role of the complexity parameters in the frequentist LASSO (Tibshirani, 1994). Whereas the frequentist perception of regularization assumes the shrinkage parameter fixed, the Bayesian LASSO allows to learn about the amount of shrinkage from the data by treating the parameter τ^2 as a random variable with its own prior distribution. Hans (2010) complemented the LASSO prior with the point mass at zero and provided Gibbs sampling schemes alternative to the approach of Park and Casella (2008) (see also (Hans, 2009)).

Keeping the variances λ_k equal and fixed, the MAP estimation corresponds to the frequentist ridge regression. However, such prior is not flexible enough to accommodate different shrinkage patterns for the individual coefficients. Assuming priors for the idiosyncratic variances assures more adaptivity. The fully Bayesian setup for ridge regression assumes the conjugate inverse gamma prior distribution for the variances, which implies a marginal scaled Student prior distribution for the individual regression coefficients.

Recent efforts in generalizing the penalization methodology to more complex data structures crystalized in several innovations of LASSO, which can be in turn transformed into MAP estimation in Bayesian hierarchical models. In linear regression setting, Tibshirani et al. (2005) proposed fused LASSO for predictors that have a natural ordering, where the penalty is a linear combination on l_1 penalty on coefficients themselves and l_1 penalty on their first order differences. Such penalty induces similarity between neighboring coefficients. In case

grouping among regression coefficients is suspected, but unknown, Zou and Hastie (2005) suggested elastic net, which combines LASSO and ridge into one penalty and as such tends to keep the related variables in the model as a group. When the groups among predictors are known (e.g. group of dummy variables or spline coefficients), Yuan and Lin (2006) proposed a grouped LASSO, which penalizes elliptical norms of the coefficients for each group. The Bayesian counterparts of these LASSO alternatives emerge by considering adequate alternations of powered exponential priors, that can be again represented as scale mixtures of normals (Kyung et al., 2010).

2.2.4.2 □ The Elastic Net Prior

Bayesian elastic net, proposed by (Zou and Hastie, 2005; Li and Lin, 2010) in the context of linear regression, constitutes a compromise between the LASSO and ridge enjoying the advantages of the two. The elastic net prior inherits the sparsity property from the LASSO, since it is also not differentiable at zero, and at the same time encourages grouping as typical for the ridge prior. By a grouping effect we refer to the ability to retain a group of highly correlated variables in a model and keeping their estimated coefficients nearly equal (up to a change of sign for negatively correlated ones). This behavior is appreciated in modeling, for instance, gene expression data where related genes should enter the model as a group. The frequentist penalty term l_{net} for the "naive" elastic net (Zou and Hastie, 2005) is the linear combination of l_1 and l_2 penalties, i.e. $l_{net}(\beta) = a_1 \sum_{k=1}^{q} |\beta_k| + a_2 \sum_{k=1}^{q} \beta_k^2$, which corresponds to the marginal MAP estimation implied by the following prior hierarchy:

$$\beta_k|\tau_k \sim \mathrm{N}\left(0, \left[\frac{a_2}{\sigma^2}\frac{\tau_k}{\tau_k - 1}\right]^{-1}\right),$$

$$\tau_k \sim \mathrm{Gamma}\left[0.5, \frac{8a_2\sigma^2}{a_1^2}, (1,\infty)\right],$$

where $\mathrm{Gamma}[a,b,(c,d)]$ refers to the truncated gamma distribution with shape a, scale b and with a support restricted to the interval (c,d). Diffuse hyperpriors for the two penalization parameters a_1 and a_2 can be added in the formulation to circumvent the uncertainty in their selection. Similarly as for the LASSO prior, Hans (2008, unpublished manuscript) augmented the elastic net prior to include the point mass at zero and suggested a Gibbs sampling algorithm in Gaussian regression models.

■ 2.2.5 Extensions to Other than Linear Regression Settings

The variable shrinkage/selection priors outlined in the previous sections can be applied in other regression modeling settings where the response is not Gaussian. In probit regression, data augmentation strategies (Albert and Chib, 1993) assuming a linear model on latent continuous data greatly facilitate the implementation of efficient MCMC schemes. A similar approach can be adapted for handling ordered categorical data (Albert and Chib, 1993). Data

augmented Bayesian logistic regression was enabled by the introduction of Kolmogorov-Smirnov random variables (Holmes and Held, 2006). Holmes and Held (2006) further describe variable selection approach in logistic regression using reversible jump MCMC. The regularized logistic regression was dealt by Gramacy and Polson (2012). Variable selection in binary regression models was considered by many authors including (Sha et al., 2004; Yang and Song, 2010; Zhou et al., 2004; Bae and Mallick, 2004). In survival regression context, Sha et al. (2006) applied the variable selection priors in accelerated failure time model.

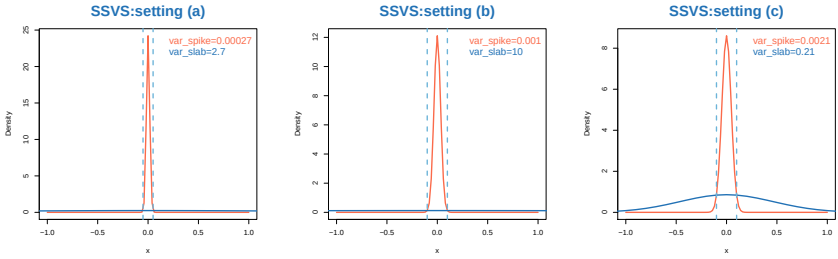
2.3 Simulation Study

The empirical variable selection performance of the outlined Bayesian methodology (*SSVS*, *NMIG*, *GVS*, MC^3, Bayesian LASSO, ridge and elastic net) was evaluated in a simulation study carried out to: (a) compare the classical versus Bayesian variable selection, (b) assess the sensitivity of spike and slab priors to the choice of tuning parameters, (c) to compare the different approaches to processing of the posterior information. The simulation study was performed on data with binary responses, generated according to the latent variable probit regression scheme

$$\begin{aligned} Z_i &\overset{\text{ind}}{\sim} \mathrm{N}(x_i'\beta, \sigma^2), \\ Y_i &= \mathrm{I}(Z_i > 0), \quad (i = 1, \ldots, n), \end{aligned}$$

which is equivalent to assuming $Y_i \sim \texttt{Bernoulli}\left[\Phi(x_i'\beta/\sigma)\right]$. We assume four linear regression models for the latent continuous data Z, which reflect settings with different degree of sparsity, magnitude of the main effects and the pattern of collinearity among the predictors. The first three designs were adopted from the original LASSO paper of Tibshirani Tibshirani (1994). The predictors were drawn independently from $\mathrm{N}_8(0,\Sigma)$ with $\Sigma = (\sigma_{ij})_{i,j}$ and $\sigma_{ij} = \rho^{|i-j|}$. The first and third model (Design 1 and Design 3) mimic rather sparse situations with relatively large values of nonzero coefficients. The parameters were chosen as follows: $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ with $\rho = 0.5$ and $\sigma = 3$ in Design 1 and $\beta = (5, 0, 0, 0, 0, 0, 0, 0)'$ with $\rho = 0.5$ and $\sigma = 2$ in Design 3. In Design 2, all the 8 predictors are weakly informative, i.e. $\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)', \rho = 0.5$ and $\sigma = 3$. In the last design, the predictors were generated as follows: $x_{ij} = m_{ij} + z_1, j = 1, \ldots, 5$, $x_{ij} = m_{ij} + z_2, j = 6, \ldots, 10$, and $x_{ij} = m_{ij} + z_3, j = 11, \ldots, 15$, where m_i were drawn independently from $\mathrm{N}_{15}(0, \mathrm{I}_{15})$ with I_{15} the identity matrix and $z_i\,(i = 1, 2, 3)$ are standard normal. Such specification induces correlations of about 0.5 within the three blocks of predictors. The vector of coefficients was chosen equal to $\beta = (3, 3, 3, 3, 3, 0, 0, 0, 0, 0, -3, -3, -3, -3, -3)'$.

For modeling the relationship between the binary response and the predictors, we used probit regression (data augmentation formulation (Albert and Chib, 1993), which assumes the latent normal linear model).



(a) $\delta = 0.05, c = 100$ (b) $\delta = 0.1, c = 100$ (c) $\delta = 0.1, c = 10$

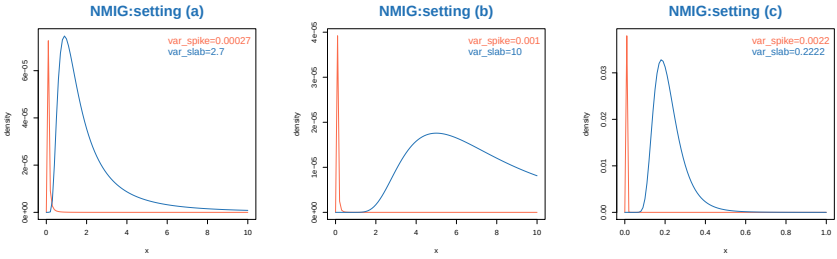
Figure 2.1: Tuning parameters for the SSVS mixture priors

For each of the four models 50 datasets were simulated, each consisting of $n = 100$ observations. To evaluate the variable selection properties, we keep track of the following quantities: (1) FDN (number of false discoveries), which is the number of coefficients falsely identified as nonzero, (2) FNN (number of false nondiscoveries), which stands for the number of unrevealed nonzero coefficients and (3) DIM (dimension of the model), which is the number of nonzero coefficients.

2.3.1 Settings

In order to assess the sensitivity of the spike and slab priors to the choice of tuning parameters we considered three sets of hyperparameters. For SSVS, these were selected considering different values of the intersection point δ of the two normal mixture components and different ratio c^2 of the slab versus spike variance. We have the following settings: (a) $\delta = 0.05$ and $c = 100$ (spike variance $\mathsf{Var}_{sp} = 0.00027$ and slab variance $\mathsf{Var}_{sl} = 2.7$), (b) $\delta = 0.1$ and $c = 100$ ($\mathsf{Var}_{sp} = 0.001$, $\mathsf{Var}_{sl} = 10$) and (c) $\delta = 0.1$ and $c = 10$ ($\mathsf{Var}_{sp} = 0.0021$ and $\mathsf{Var}_{sl} = 0.21$). The three mixture densities are depicted in Figures 2.1(a), 2.1(b) and 2.1(c). Similar settings were used in NMIG, where the parameters were chosen so that the intersection point of the scaled t-distributions and the ratio of the variances match to each of the three previous SSVS settings. The NMIG mixture variance priors are depicted in Figures 2.2(a), 2.2(b) and 2.2(c). The mixture prior for GVS was selected as in SSVS (b). In the Bayesian LASSO and Bayesian elastic net, the regularization parameters τ, a_1 and a_2 are assigned prior $\mathsf{Exp}(0.01)$, where $\mathsf{Exp}(\mu)$ denotes the exponential distribution with the expectation $1/\mu$. A noninformative prior $\mathrm{N}(0, 1000)$ is used for the intercept term. Whenever applicable, we used the uniform prior on the model space.

For spike and slab models as well as for GVS, the highest posterior model (HPD) selection, the median probability model (MPM) selection and the hard shrinkage rule (HS) were investigated. Bayesian regularization (LASSO, ridge regression and elastic net) enables only HS selection, whereas only MPM and HPD are applicable in MC^3. The interval decision criterion for HS was based on one standard deviation interval around the posterior mean. Only



(a) $a = 2, b = 2.7, v_0 = 10^{-5}$ (b) $a = 3, b = 20, v_0 = 10^{-5}$ (c) $a = 10, b = 2, v_0 = 0.01$

Figure 2.2: Tuning parameters for the NMIG mixture priors

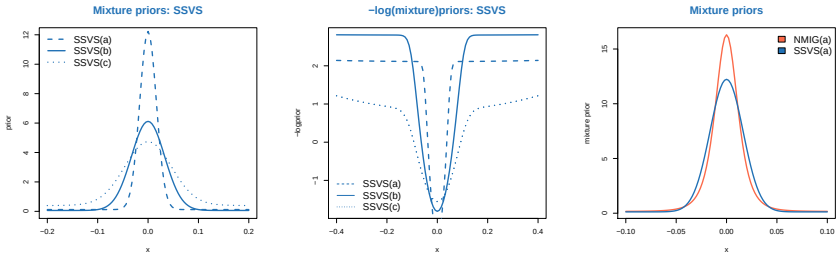
coefficients whose decision interval covers zero were excluded from the final model. Finally, the Bayesian methodology was contrasted with F-to-out backward selection with $p = 0.05$ (STEP1) and $p = 0.1$ (STEP2) and exhaustive evaluation using AIC. The MC^3 variable selection was based on the run of 1 000 MCMC iterations. The remaining Bayesian hierarchical models were estimated using 10 000 iterations with 1 000 burn-in period and 10 fold thinning.

2.3.2 Software

The majority of the available software for Bayesian variable selection deals only with linear regression models. Shrinkage estimation using sparsity priors (ridge, Laplace, normal/gamma, horseshoe) coupled with the reversible jump variable selection is obtainable through the R package `monomvn` of Gramacy and Pantaleo (2010). Spike and slab variable selection with NMIG priors can be found in the package `spikeSlabGAM` of Scheipl (2011, manuscript under revision). Bayesian model averaging as well as MC^3 for linear regression models has been implemented in the package `BMA` Madigan et al. (1994). Bayesian regularized logistic regression applicable for high-dimensional data has been implemented in the package `regloglogit` of Gramacy and Polson (2012). The frequentist regularization for generalized linear models can be found in package `glmnet` Friedman et al. (2010b). A hybrid spike and slab variable selection procedure for linear (high-dimensional) regression has been made available in the package `spikeslab` of Ishwaran and Rao (2010, unpublished manuscript). To implement the spike and slab models, Bayesian regularization and GVS in the probit (Weibull) regression context, we used WinBUGS. The code for SSVS has been adapted from the BUGS code of Ntzoufras (Ntzoufras, 2002). The MC^3 for probit (Weibull) regression has been implemented in R.

2.3.3 Results

The results for median probability model selection are summarized in Figure 2.4 and for hard shrinkage in Figure 2.5. For each of the methods, the figures present a triplet of bars. The



(a) SSVS (b) -log(SSVS) (c) SSVS & NMIG

Figure 2.3: Left panel: minus logarithm of the three SSVS mixture priors; middle panel: minus logarithm of SSVS(a) prior and NMIG(a) prior; right panel: SSVS(a) and NMIG(a) priors

left one represents the number of times a correct model has been identified (out of the 50 simulated datasets). These numbers relate to the vertical axis on the left. The middle and right bars correspond to the average FDN and FNN, respectively. These values relate to the vertical axis on the right from each graph. The average model dimension estimated by each of the methods is attached at the top of the three bars. Results for the highest posterior model selection greatly overlap with the median probability model selection in first three designs (results not presented). The difference, however, emerged in Design 4, where the HPD model selection for all spike and slab models as well as MC^3 and GVS did not correctly identify the right model in any of the 50 repetitions.

Looking at the two figures, several observations can be made:

(1) The difference between HS and MPM model selection is less apparent in the first three designs. Discrepancies again occur in Design 4, where MPM in “properly calibrated" spike and slab models outperforms the regularization priors. In Design 4, as expected, the elastic net performed the best among the regularization priors in including the groups of correlated regressors in the model.

(2) The choice of tuning parameters is influential on spike and slab variable selection, which is particularly evident in Design 2 and Design 4, where the performance increases with a decreasing variance in the slab component. This is a little at odds with the intuition that high hypervariance represents the prior belief that the coefficient can attain “arbitrarily" large values. To explain this behavior, it suffices to note that in spike and slab models the high slab hypervariance induces a stronger penalization on weak nonzero effects and hence expresses the prior opinion that many of the coefficients will be zero. To support this statement, we plotted the three mixture (SSVS) densities (Figure 2.3(a)) as well as their minus logarithms, which are proportional to the frequentist penalty functions (Figure 2.3(b)). Among the mixture priors, the setting with the lowest slab variance (SSVS (c)) places more prior emphasis on smaller effects. This forces the penalty to elevate more gradually with an increasing distance from origin

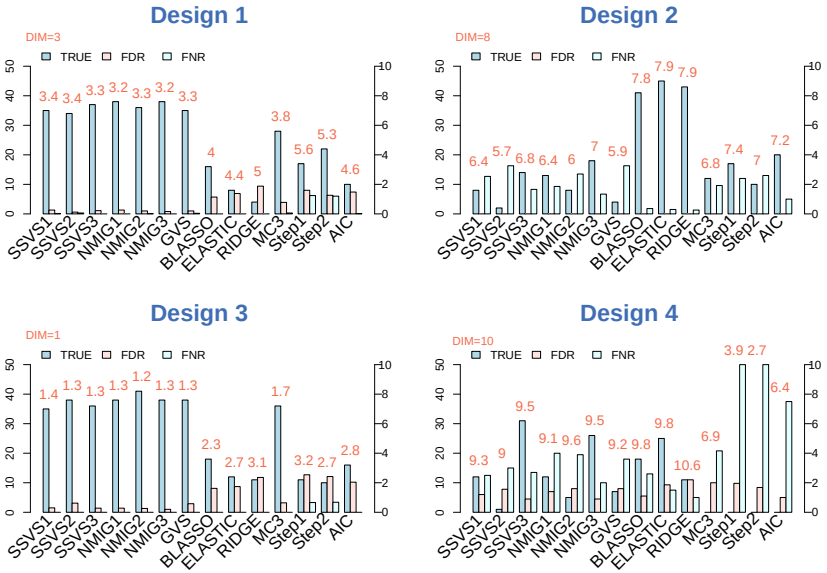


Figure 2.4: Simulation results: MPM model selection

and indeed causes the small nonzero effects to be penalized to a lesser extend. On the other hand, the shape of the penalty function arising from the "narrow spike wide slab" prior (setting (a)) provides the closest approximation to the l_0 type of penalty, which penalizes nonzero effects equally regardless their magnitude.

(3) The distributional assumption underlying the constitution of spike and slab can influence the variable selection. Comparing the Gaussian and Student mixtures with matched variances and intersection points, the t-mixture implies weaker penalization of larger effects due to heavier slab tails (Figure 2.3(c)). The impact of this behavior is particularly evident in the non-sparse designs (Design 2 and Design 4). On the other hand, the two spike and slab models exhibit similar mixing properties (evaluated by the number of visited models) in all the four designs and their computation time in our implementation was comparable.

(4) The classical frequentist variable selection was outperformed by spike and slab variable selection (regardless the posterior inference) in the two sparse regression designs. The Bayesian regularization, on the other hand, emerged as more accurate (compared to the backward selection and *AIC* full subset selection) in finding the true underlying model, as long as there were many nonzero effects.

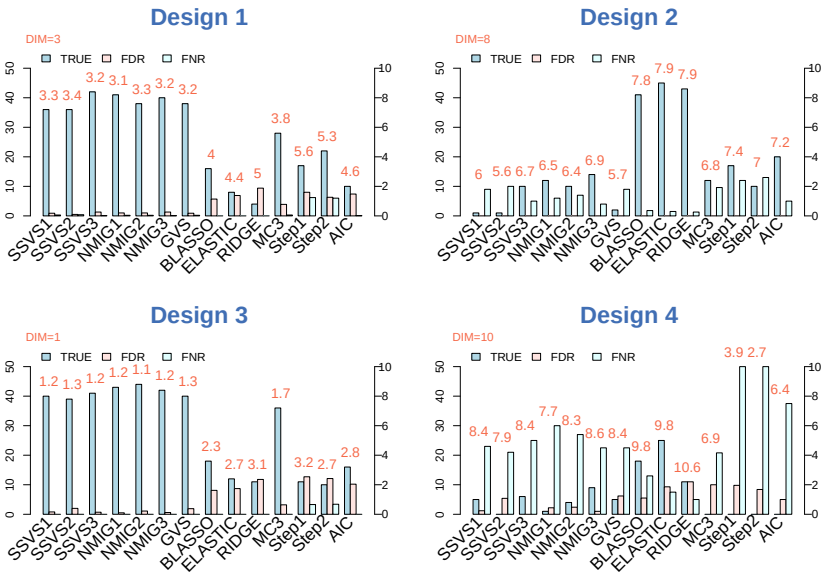


Figure 2.5: Simulation results: HS model selection

2.4 The Data

The Bayesian methodology for variable selection is applied here on two datasets. The goal of the first analysis is to identify markers predictive for development of rheumatoid arthritis, whereas the second analysis deals with joint assessment of prognostic capability of preselected mutation and gene expression markers for overall survival in patients with acute myeloid leukemia.

2.4.1 REACH Data

Rheumatoid arthritis (RA) is an autoimmune disease characterized by chronic synovial inflammation and destruction of cartilage and bone in the joints. The Rotterdam Early CoHort study (REACH) was initiated in 2004 to investigate the development of RA in patients with early manifestations of joint impairment. Information regarding basic patient characteristics, serological measurements and patterns of disease involvement at baseline has been gathered in 681 recruited patients. It is of interest to know which of the following 12 factors are potentially associated with the development of rheumatoid arthritis considered as a binary (yes/no)

outcome: ACCP (cyclic citrullinated peptide antibody), age, ESR (erythrocyte sedimentation rate), DC (duration of complaints in days), stiffness (duration of morning stiffness in minutes), RF (rheumatoid factor), gender, Sym (symmetrical pattern of joint inflammation yes/no), SJC (swollen joint count), TJC (tender joint count), BCPH (bilateral compression pain in hands yes/no) and BCPF (bilateral compression pain in feet yes/no).

The standard approach to analyze these data would be to use logistic/probit regression combined with some off-the-shelf variable selection method. The F-to-out backward selection with $p = 0.05$ yields a model with the following variables: ACCP, ESR, DC, Sym, SJC, BCPH. The model with the most favorable value of the *AIC* criterion selected after an exhaustive model evaluation contains two extra variables: RF and stiffness. Which of these models provides the best approximation to the true underlying relationships is, if at all possible, difficult to assess. In the Bayesian approach, however, these individual models can be effectively compared using one particular measure, the posterior model probability, which quantifies the amount of confidence in each of the given models. The Bayesian analysis of the REACH data is presented in Section 4.

2.4.2 AML Data

Acute myeloid leukemia (AML) describes a group of hematopoietic disorders characterized by the expansion of immature myeloid blood cells. Risk stratification and therapy decision making is nowadays based mainly on karyotype information. However, about 45 percent of patients lack any cytogenetical aberration. These patients exhibit various responses to therapy and therefore more targeted treatment protocols are required to improve their survival outcome. Identification of prognostic markers associated with survival in these “intermediate risk" patients would contribute to improved risk stratification. Recently, various markers have been individually identified as prognostically relevant. These include various mutation markers (FLT3ITD, FLT3TKD, NPM1, NRAS, IDH1, IDH2 and CEBPA single (SM) and double (DM) mutation) as well as gene expression markers (ABCB1, BCL2, BAALC, ERG, EVI1, CD34, MN1, FLT3, INDO and WT1). These markers were assessed and/or measured in a series of 318 AML patients with normal karyotype or a karyotype of no recognized prognostic value. Here we focus on the joint assessment of the prognostic importance and the selection of a combination of the markers to be used for prediction/stratification.

For modeling the relationship between the markers and survival, we used a parametric Weibull model. Backward selection ($p = 0.05$) identified variables CD34, ERG, BCL2 and CEBPA DM as relevant, whereas the *AIC* selection selects in addition also NPM1, FLT3ITD and IDH2.

In the Bayesian approach, the research question can be formulated and answered in a variable specific way rather than model-wise. The conclusion about which variables are important for the survival outcome then again follows from posterior probabilities rather than from p-values. The Bayesian analysis of this data is presented in Section 4.

	ACCP	Age	ESR	DC	Stiff.	RF	Sex	Sym.	SJC	TJC	BCPH	BCPF
Highest posterior model selection												
SSVS1	■	□	■	□	□	□	□	■	□	□	■	□
SSVS2	■	□	■	□	□	□	□	■	□	□	■	□
SSVS3	■	□	■	□	□	■	□	■	■	□	■	□
NMIG1	■	□	■	□	□	■	□	■	■	□	■	□
NMIG2	■	□	■	□	□	□	□	■	□	□	■	□
NMIG3	■	□	■	□	□	■	□	■	■	□	■	□
MC3	■	□	■	■	□	□	□	■	■	□	■	□
GVS	■	□	■	□	□	□	□	■	□	□	■	□
Hard shrinkage												
SSVS1	■	□	■	□	□	□	□	■	□	□	■	□
SSVS2	■	□	■	□	□	□	□	■	□	□	■	□
SSVS3	■	□	■	■	□	■	□	■	■	□	■	□
NMIG1	■	□	■	□	□	■	□	■	■	□	■	□
NMIG2	■	□	■	□	□	□	□	■	□	□	■	□
NMIG3	■	□	■	■	□	■	□	■	■	□	■	□
BLASSO	■	□	■	■	■	■	□	■	■	□	■	■
ELASTIC	■	□	■	■	■	■	□	■	■	□	■	■
RIDGE	■	□	■	■	■	■	□	■	■	□	■	■
Frequentist												
STEP1	■	□	■	■	□	□	□	■	■	□	■	□
STEP2	■	□	■	■	■	■	□	■	■	□	■	□
AIC	■	□	■	■	■	■	□	■	■	□	■	□

Table 2.1: The table of models selected by the different BVS methods

2.5 Data Analysis

In the previous section, we presented the frequentist analysis of the two datasets. Whereas in the classical approach, the emphasis is often put on finding a single representation of the data by one model, the Bayesian approach enables to assess uncertainty surrounding such decision and prepares grounds for the eventual model averaging. In the analysis of the REACH data, we apply the Bayesian model selection and uncertainty assessment via posterior model probabilities, as well as the shrinkage and MPM variable selection. In the AML data we elaborate further on the shrinkage approaches and the "model averaged" Bayesian variable selection. In both the analyses, all continuous regressors were standardized. The estimation was based on Markov chains of the length 10 000 for MC^3 and 15 000 with a burn-in 5 000 thinned by 10 for all the other Bayesian models.

■ 2.5.1 Bayesian Analysis of REACH Data

In the simulation study, we have seen that the variable selection (implied by hard shrinkage, posterior inclusion probabilities, or posterior model probabilities) is influenced by the prior specification, both in terms of the choice of the prior (mixture) distribution and hyperparameter calibration. To amplify this point, we applied the same prior settings on the REACH data. Variables selected by the highest posterior model selection and hard shrinkage, if applicable, for each of the Bayesian variable selection method are presented in Table 2.5. The median probability model selection is not presented separately as it yields the same models as the HPD selection for all the methods.

Again, we see how sensitive the Bayesian variable selection can be towards the prior settings. In HPD variable selection, this "sensitivity" connects to the mixing properties of the chain sampling individual models. There the actual model selection depends on the chain's ability to find interesting regions of the posterior model space. One particular stochastic

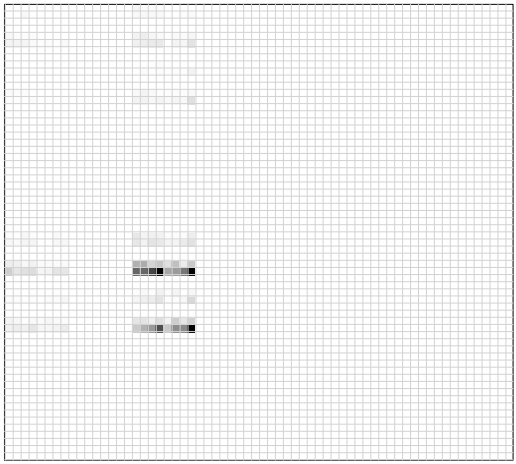
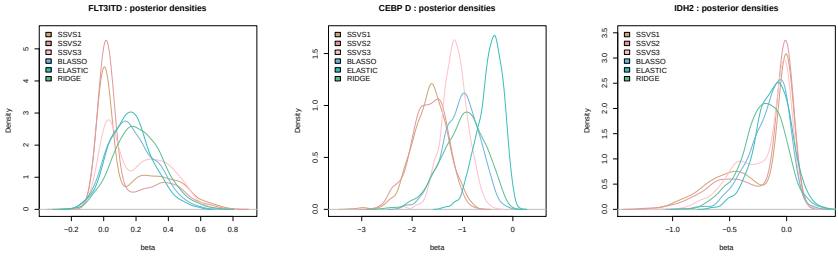


Figure 2.6: MC^3 stochastic approximation to the posterior distribution over all models for the REACH data; figure plots a heatmap of posterior probabilities, where all possible models have been sorted in a matrix and similar models are located close to each other; each entry in the matrix corresponds to a single model, where darker shades of grey correspond to a higher accumulation of posterior density

approximation to the posterior model distribution for the REACH data, which was obtained by MC^3, is depicted on Figure 2.6. The individual models were sorted in a matrix, where the simplest one (no covariates at all) is located in the lower right corner and the full model in the upper left corner. The logic of the sorting is so that related models create blocks in the matrix. The darker the grey color of each squared spot (model), the more often the model was visited by the MC^3 Markov chain. The darkest spot then corresponds to the estimated HPD model (with covariates ACCP, ESR, DC, Sym, SJC and BCPH), which the chain occupied 576 times during the run of 10 000 iterations. There are clearly more candidate models with a high number of visits. In fact, the second most frequent model was visited 571 times. Therefore, the posterior evidence contained within the estimated model probabilities from MC^3 does not vote unequivocally for the selection of just one model. Out of the 4 096 possible models, only 229 different models were encountered by the MC^3 Markov chain. The number of visited models in GVS was only 33, whereas in the three versions of SSVS (resp. NMIG) there were 80, 56 and 175 (resp. 126, 74 and 176) different models among the 1 000 recorded MCMC iterations. The fast mixing ability of the last SSVS and NMIG specification is to be expected, since the variance of the slab is sufficiently small allowing the chain to escape from the spike more easily. The sharpness of the prior spike together with the magnitude of the prior slab variance hence influence how many models will be visited and implicitly determine the shape of the posterior distribution of individual coefficients. If models with a particular variable included were not often encountered in the sequence of sampled models, the spike will dominate the



(a) (b) (c)

Figure 2.7: Approximation to the posterior distribution of three selected coefficients

posterior shape of the corresponding coefficient, which will result in shrinkage of the posterior mean towards zero.

The regularization priors such as ridge, elastic net or LASSO induce rather “soft" shrinkage, leading to many nonzero selected coefficients, whereas the shrinkage from spike and slab priors is more aggressive, especially when the slab versus spike variance ratio is sufficiently large. If the preference is to select a model with all the included variables strongly associated with the outcome, we might opt for spike and slab variable selection with a “sharp spike and flat slab" shape. In this case we would end up with a model with only 4 covariates ACCP, ESR, Sym and BCPH. Relaxing the requirements for the model parsimony and giving preference to a model suitable rather for prediction, we might choose the model indicated by the Bayesian regularization, which contains one extra variable compared to the *AIC* model.

A similar interplay between the model complexity and practical significance of included factors can be achieved by selecting different significance thresholds in stepwise selection. However, the Bayesian variable selection (HS and MPM) in spike and slab models accounts for the uncertainty introduced by the model selection process, since the posterior distribution, on which the decision is based on, is averaged over more candidate models.

2.5.2 Bayesian Analysis of AML Data

Unlike the analysis of REACH data, where we applied all three types of posterior inference for variable selection (i.e. HPD, MPM and HS), in the AML data we assess the variable selection only via posterior inclusion probabilities (MPM) and hard shrinkage (HS). We suspect that the approximation to the posterior model distribution provided by the spike and slab models may not be sufficiently accurate to find the highest posterior model in the setting with this many variables. Furthermore, the posterior inclusion probabilities and posterior distribution of coefficients provide model-averaged decision criterion that is potentially more reliable than just comparing individual model probabilities.

The results of BVS applied on the AML data are summarized in Figure 2.8. The upper panel depicts estimated marginal inclusion probabilities. According to the median probability

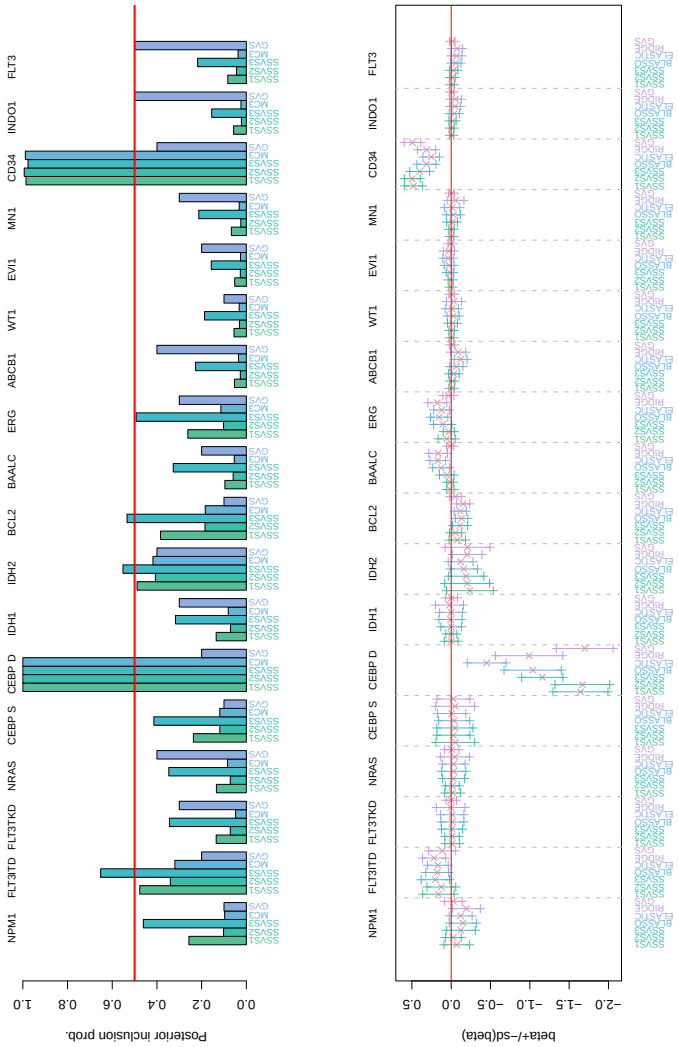


Figure 2.8: AML data: in the upper panel there are estimated inclusion probabilities for each of the markers, the lower panel depicts the estimated coefficients together with ± sd interval

model selection rule, a variable is included in a model whenever the inclusion probability exceeds 0.5 (indicated by the horizontal line). The lower panel displays point estimates (posterior means) for each regression coefficient, accompanied with $\pm sd$ inclusion interval. The point estimates are weighted averages of estimated posterior means arising from visited models with underlying slab (resp. spike) prior on the present (resp. absent) coefficients. The weights are determined from the frequencies of visits of each model. Intervals which exclude zero (again marked by the horizontal line) imply the inclusion of the variable in the model by the hard shrinkage rule. The NMIG appeared to show poorer mixing (358, 23 and 118 visited models for each of the settings compared to 763, 127 and 626 models for SSVS). That is why we present the results only from SSVS spike and slab models, as we believe they are more reliable.

To compare the different shrinkage behavior of the spike and slab and regularization priors, we depicted the approximations to the posterior distribution of three selected coefficients (for variables FLT3ITD, CEBPA DM and IDH2) on Figure 2.7. For the first coefficient (FLT3ITD), the evidence for the inclusion is not strongly convincing and therefore the spike and slab posteriors are bimodal. We observe the sharpest posterior spike (i.e. the strongest penalization of larger values) for the SSVS with the biggest prior slab variance (setting (b)). Less stringent penalization of the SSVS setting (c) is evident from the pronounced bimodal shape of the posterior. In case of CEBPA DM, the Bayesian regularization places heavier penalties on larger effects (compared to the inflated-slab-variance priors), which inevitably introduces estimation bias.

Conclusively, the Bayesian approach gives strong evidence for the two markers CEBPA DM and CD34. The variables FLT3ITD, ERG or BCL2 were included by some of the methods, but their estimated effects are rather small. In the frequentist approach, we ended up with models that are quite complex, whereas the Bayesian approach points at more parsimonious models, enables to quantify the importance of each individual marker by means of a posterior inclusion probability or posterior distribution of the coefficients. These summaries are averaged over posterior model uncertainty and therefore provide more objective quantitative assessment than p-values.

2.6 Discussion

The purpose of this chapter was to survey the evolution of Bayesian variable selection and highlight some of its recent developments. The list of the discussed methodology is surely not exhaustive as the methodology is continuously evolving and its potential has only begun to be realized. We have restricted our attention to the general discussion on the principles rather than technical details on implementation using sophisticated MCMC techniques. We have omitted discussion on the nonparametric relaxations of considered hierarchical models using the Dirichlet process priors (Nott, 2008a,b; Kim et al., 2009) as well as application of the prior

hierarchies on factor analytic models Carvalho and Polson (2010), additive regression models (Fahrmeir et al., 2010) etc.

The practical utility of the Bayesian methodology (regularization and spike and slab models) would be particularly appreciated in the analysis of high-dimensional data (genomics, proteomics), where the estimation in Bayesian hierarchical models constitutes a coherent alternative to approaches based on corrections of multiple testing. Here, we have confined our application to the classical regression settings and in the simulation study demonstrated that non-negligible practical gain can be obtained also in these, yet less involved, modeling tasks. Among the outlined methodology, the spike and slab models constitute an approach that is particularly conceptually appealing. They are closely connected to the Bayesian regularization in the sense that they provide a Bayesian framework that gives rise to the similar type of penalties as the l_0 frequentist complexity penalty. The Bayesian formalism for these penalties has been pursued by Abramovich et al. (2007) in the context of high-dimensional normal means models. The spike and slab models, however, provide a different perspective on the l_0 frequentist penalization. The connection between penalized l_0 estimation and Bayesian spike and slab models follows quite analogously as between Bayesian MAP estimates from the Laplace priors and the frequentist LASSO. The frequentist implementation of the optimization problem in l_0 penalized models is hampered by the non-singularity and discontinuity of this penalty at origin. Continuous approximations to this frequentist penalty have been suggested that facilitate the computation (Liu and Wu, 2007). On the other hand, the penalty induced by the Bayesian mixture priors (which is proportional to the logarithm of the mixture prior) can be regarded as another type of continuous approximation to the l_0 type of penalty. The advantage of the Bayesian formulation is that the MCMC machinery can be used to obtain the approximation to the whole posterior distribution, which is typically feasible when $p < n$.

In this chapter we have restricted our attention to $p < n$ setting. Nevertheless, the modern applications of Bayesian variable selection deal mostly with high-dimensional data. The complexity of such problems renders several presented Bayesian variable selection methods less appealing from the computational time and storage efficiency standpoints. Adaptations of SSVS algorithm suitable for high-dimensional data that avoid sampling the individual regression coefficients have been considered by Kwon et.al Kwon et al. (2011) and Yang and Song (2010). Despite the advances in high-dimensional stochastic model search (Hans et al., 2007; Bottolo and Richardson, 2010), the shrinkage approaches (eventually accompanied with the reversible jump sampling) might be preferred in such situations (Gramacy and Pantaleo, 2010). Alternatively, the involved MCMC computation in hierarchical shrinkage models can be avoided using EM algorithm (Kiiveri, 2003). A proposal of an EM approach for variable selection is presented in the next chapter.

We exemplified the Bayesian hierarchical models for variable selection in probit regression and Weibull regression. Previously, the BVS methods have been discussed in the context of probit regression models by e.g. Sha et al. (2004), Kwon et al. (2007), Yang and Song (2010), Zhou et al. (2004), Bae and Mallick (2004) and in survival models by Sha et al.

(2006). Despite our WinBUGS programs offer a working solution to fitting the hierarchical models with sparsity/variable selection priors in low-dimensional settings, customized algorithms/implementations are needed in higher dimensions. For instance, the Bayesian regularized logistic regression has been implemented in package `reglogit`. Nevertheless, the majority of the discussed hierarchical constructions are still awaited to be transferred to/implemented in other than linear regression settings.

In the simulation study we demonstrated that Bayesian variable selection leads to improved performance in identifying the true underlying model, when compared with the frequentist methods. We used several Bayesian variable selection approaches, none of which could be postulated as the methodological ideal for all the considered simulation settings and neither it should be. The choice of the particular Bayesian approach should be context dependent as some of the discussed methodologies are customized for particular data structures (groups of correlated predictors) and inferential goals (prediction rather than variable selection). Information regarding the correlation structure and the expected dimension of the solution can be beneficial when finding the "true" pattern of sparsity.

In the theoretical discussion we focused mainly on absolutely continuous priors, also within the spike and slab context. The point mass spike and slab priors (Gramacy and Pantaleo, 2010; Mitchell and Beauchamp, 1988) on the other hand offer a correct characterization of the model uncertainty and avoid making subjective choices on tuning hyperparameters. These facts have contributed to the fact that the point mass priors have begun to be realized as benchmark for Bayesian variable selection. Recently, point mass shrinkage priors have been made available through standard software (Gramacy and Pantaleo, 2010) for linear regression.

Despite the conceptual appeal of Bayesian variable selection, the wide acceptance of BVS as the preferred variable selection strategy has been hampered by the unavailability of implementation in standard software. Catalyzed by advances in the MCMC computation, the methodology has become no longer problematic to implement in Bayesian software such as WinBUGS for the classical regression settings. However, the computational challenges increase with the dimensionality of the data, where developments in numerical approximations and/or MCMC techniques will hopefully make the methodology more approachable for more practically oriented users.

CHAPTER 3

EMVS: THE EM APPROACH TO BAYESIAN VARIABLE SELECTION

Adapted version of a research article:
Rockova, V., George, E. 2012. **EMVS: The EM Approach to Bayesian Variable Selection**. Tentatively accepted by the *Journal of the American Statistical Association (Theory and Methods)*

Abstract

Despite rapid developments in stochastic search algorithms, the practicality of Bayesian variable selection methods has continued to pose challenges. High-dimensional data are now routinely analyzed, typically with many more covariates than observations. To broaden the applicability of Bayesian variable selection for such high-dimensional linear regression contexts, we propose EMVS, a deterministic alternative to stochastic search based on an EM algorithm which exploits a conjugate mixture prior formulation to quickly find posterior modes. Combining a spike-and-slab regularization diagram for the discovery of active predictor sets with subsequent rigorous evaluation of posterior model probabilities, EMVS rapidly identifies promising sparse high posterior probability submodels. External structural information such as likely covariate groupings or network topologies is easily incorporated into the EMVS framework. Deterministic annealing variants are seen to improve the effectiveness of our algorithms by mitigating the posterior multi-modality associated with variable selection priors. The usefulness the EMVS approach is demonstrated on real high-dimensional data, where computational complexity renders stochastic search to be less practical.

3.1 Introduction

Bayesian variable selection for the normal linear model typically requires two main ingredients, a prior to induce a posterior distribution over subsets of potential predictors, and an approach to extract information from this posterior in order to identify promising subset models. When the number of potential predictors is large and/or the posterior is simply intractable, this latter step is often carried out by some form of Markov chain Monte Carlo (MCMC) stochastic search that is used to discover high probability models. See, for example, Bottolo and Richardson (2010), Hans et al. (2007), Li and Zhang (2010) and Stingo and Vannucci (2011) from the large literature about such methods.

The main thrust of this chapter is to propose an approach called EMVS (EM Variable Selection), a deterministic alternative to MCMC stochastic search based on the EM algorithm, that can be used to rapidly identify promising high posterior models. Ideally suited for high-dimensional "$p > n$" settings with many potential predictors, EMVS succeeds in finding interesting candidate models at a fraction of the time required for stochastic search. Furthermore, EMVS can be deployed to effectively identify the sparse high probability models, which are of increasing interest in high-dimensional settings.

EMVS is based on one of the earliest Bayesian variable selection prior formulations, the continuous conjugate version of the "spike-and-slab" normal mixture formulation underlying the SSVS (Stochastic Search Variable Selection) approach of George and McCulloch (1993, 1997). The continuity of the spike distribution is essential in the derivation of rapidly computable closed form expressions for the EM algorithm. Furthermore, increasing the variance of the spike distribution serves to absorb negligible coefficients, thereby reducing posterior

multimodality and exposing sparse high probability subsets. The speed of the algorithm makes it feasible to carry out dynamic posterior exploration for the identification of posterior modes over a sequence of mixture priors with increasing spike variances. For the visualization of the progressively sparser sequence of associated high probability submodels, we propose new spike-and-slab regularization diagrams. To further determine which of the discovered submodels is best supported by the data, we return to a point mass spike distribution for model evaluation.

Although EMVS is anchored by the original SSVS prior, extension to more modern elaborations of the prior are straightforward. Heavy tailed slab distributions such as the Cauchy or double exponential are obtained with little computational cost by extending the algorithm to average the slab distribution variance over an additional prior. Structured priors on variable inclusion probabilities at the top level of the hierarchical model such as the logistic regression product prior of Stingo et al. (2010) or the Markov random field prior of Li and Zhang (2010) are also easily incorporated. Finally, the performance of EMVS can be further enhanced by a deterministic annealing variant, which improves upon the potential problem of entrapment in local modes.

3.2 Conjugate Spike-and-Slab Formulations for EMVS

The data for the setup under consideration consists of y, an $n \times 1$ response vector, and $X = [x_1, \ldots, x_p]$, an $n \times p$ matrix of p potential predictors. We assume throughout that y is related to X by a Gaussian linear model

$$f(y \mid \alpha, \beta, \sigma) = \mathrm{N}_n(1_n\alpha + X\beta, \sigma^2 \mathrm{I}_n), \tag{3.2.1}$$

where 1_n is an $n \times 1$ vector of 1's, α is an unknown scalar intercept, β is a $p \times 1$ vector of unknown regression coefficients, and σ is an unknown positive scalar. It will often be sensible to standardize the predictors to have mean zero and variance one before proceeding.

As with many Bayesian variable selection approaches for this problem, EMVS is facilitated by the introduction of a vector of binary latent variables $\gamma = (\gamma_1, \ldots, \gamma_p)'$, $\gamma_i \in \{0, 1\}$, where $\gamma_i = 1$ indicates that x_i is to be included in the model. Combined with suitable prior distributions over α, β, σ and γ, the induced posterior distribution $\pi(\gamma|y)$ then summarizes all post-data variable selection uncertainty.

The EMVS approach is anchored by prior formulations stemming from the conjugate version of the hierarchical SSVS prior of George and McCulloch (1997), (hereafter GM97). The cornerstone of this formulation is the "spike-and-slab" Gaussian mixture prior on β,

$$\pi(\beta \mid \sigma, \gamma, v_0, v_1) = \mathrm{N}_p(0, D_{\sigma,\gamma}), \tag{3.2.2}$$

where $D_{\sigma,\gamma} = \sigma^2 \mathrm{diag}(a_1, \ldots, a_p)$ with $a_i = (1 - \gamma_i)v_0 + \gamma_i v_1$ for $0 \leq v_0 < v_1$. GM97 recommended setting the hyper-parameters v_0 and v_1 to be small and large fixed values, respectively,

to distinguish those β_i values which warrant exclusion of x_i from those that warrant inclusion of x_i.

Although the variance parameter v_0 of the spike distribution is commonly set equal to zero in practice, GM97 proposed consideration of small but positive $v_0 > 0$ to encourage the exclusion of unimportant nonzero effects. We make use of both v_0 specifications, first using a sequence of $v_0 > 0$ values to identify promising subsets, and then using $v_0 = 0$ to evaluate the submodels corresponding to those subsets. As will be seen, positive v_0 values not only tend to expose the sparser subsets by increasing their posterior probability, but also allow for the construction of a closed form EM algorithm that can rapidly identify those subsets.

For the variance parameter v_1 of the slab distribution, we consider two possibilities: (i) fixing it at a large enough value to accommodate all plausible β values, or (ii) treating it as random with respect to a prior $\pi(v_1)$ to induce heavy tailed slab alternatives such as the double exponential or Cauchy distributions. As will be seen, such a $\pi(v_1)$ can be incorporated by folding it iteratively into our EM algorithm, which is at each step based on fixed values of v_0 and v_1.

For the prior on α we adopt a uniform improper prior over α. This prior is formally justified here because α is a location parameter that appears in every submodel (when $v_0 = 0$), and the improper uniform prior is the right-Haar prior for the location invariance group. See Berger et al. (1998) for details. To facilitate our development, we will from here on assume that α has been margined out with respect to this prior, and proceed with the induced marginal likelihood $f(y \mid \beta, \sigma)$. This is equivalent to centering Y at 0 and treating it as a constrained multivariate Gaussian realization with mean $X\beta$.

For the prior on σ^2, we follow GM97 and use an inverse gamma prior

$$\pi(\sigma^2 \mid \gamma) = \mathrm{IG}(\nu/2, \nu\lambda/2) \tag{3.2.3}$$

with $\nu = 1$ and $\lambda = 1$ to make it relatively noninfluential. Further choices of ν and λ as recommended by GM97 may also be of interest.

The remaining component of the hierarchical prior specification is completed with a prior distribution $\pi(\gamma)$ over the 2^p possible values of γ. For this purpose, we shall be interested in hierarchical specifications of the form

$$\pi(\gamma) = \mathsf{E}_{\pi(\theta)}\pi(\gamma \mid \theta) \tag{3.2.4}$$

where θ is a (possibly vector) hyperparameter. In the absence of structural information about the predictors, i.e., when their inclusion is apriori exchangeable, a useful default choice for $\pi(\gamma \mid \theta)$ is the i.i.d. Bernoulli prior form

$$\pi(\gamma \mid \theta) = \theta^{|\gamma|}(1-\theta)^{p-|\gamma|}, \tag{3.2.5}$$

where $\theta \in [0,1]$ and $|\gamma| = \sum_i \gamma_i$. With this form, any marginal $\pi(\gamma)$ in (3.2.4) will be exchangeable on the components of γ. Of particular interest to us will be the exchangeable

priors obtained with a beta prior $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, $a,b>0$, (3.2.5) which yields beta-binomial priors $\pi(\gamma)$ that favor parsimony, see Scott and Berger (2010). As will be seen, EMVS can be applied to locate promising candidate subsets under these priors by exploiting the conditional independence of the intermediate Bernoulli form.

Beyond (3.2.5), when structural information about the predictors is available, more flexible priors can be used to transmit this information. In particular two recent useful forms of $\pi(\gamma \mid \theta)$ for this purpose are the logistic regression product prior considered by Stingo et al. (2010) and the Markov random field prior considered by Li and Zhang (2010) and Stingo and Vannucci (2011), both of which were used to incorporate external biological information in a genetic context. We will consider these forms further in Section 3.7 and show how they can be folded into EMVS.

3.3 A Closed Form EM Algorithm

EMVS is based on an EM algorithm alternative to the commonly used MCMC stochastic search approaches to extracting information from the posterior distribution induced by the prior formulations described in Section 2. Geared towards finding posterior modes of the parameter posterior $\pi(\beta,\theta,\sigma|y)$ rather than simulating from the entire model posterior $\pi(\gamma|y)$, the EM algorithm derived here offers potentially enormous computational savings over stochastic search alternatives, especially in problems with a large number p of potential predictors. In Section 4, we show how EMVS threshholds the modal estimates of (β,θ,σ) to identify the associated high posterior loci of $\pi(\gamma|y)$ when $v_0 = 0$.

Our implementation of the EM algorithm maximizes $\pi(\beta,\theta,\sigma|y)$ indirectly, proceeding iteratively in terms of the "complete-data" log posterior, $\log \pi(\beta,\theta,\sigma,\gamma \mid y)$, where the latent inclusion indicators γ are treated as "missing data". As this function is unobservable, it is at every iteration replaced by its conditional expectation given the observed data and current parameter estimates, the so called E-step. This is followed by an M-step that entails the maximization of the expected complete-data log posterior with respect to (β,θ,σ). Iterating between these two steps, the EM algorithm generates a sequence of parameter estimates, which under regularity conditions converge monotonically towards a local maximum of $\pi(\beta,\theta,\sigma \mid y)$.

More precisely, our EM algorithm indirectly maximizes $\pi(\beta,\theta,\sigma|y)$ by iteratively maximizing the objective function

$$Q\left(\beta,\theta,\sigma \mid \beta^{(k)},\theta^{(k)},\sigma^{(k)}\right) = \mathsf{E}_{\gamma|\cdot}\left[\log \pi(\beta,\theta,\sigma,\gamma|y) \mid \beta^{(k)},\theta^{(k)},\sigma^{(k)},y\right] \tag{3.3.6}$$

where $\mathsf{E}_{\gamma|\cdot}(\cdot)$ denotes the conditional expectation $\mathsf{E}_{\gamma|\beta^{(k)},\theta^{(k)},\sigma^{(k)},y}(\cdot)$. At the kth iteration, given $(\beta^{(k)},\theta^{(k)},\sigma^{(k)})$, an E-step is first applied, which computes the expectation of the right side of (3.3.6) to obtain Q. This is followed by an M-step, which maximizes Q over (β,θ,σ) to yield the values of $(\beta^{(k+1)},\theta^{(k+1)},\sigma^{(k+1)})$.

For the conjugate spike-and-slab hierarchical prior formulations described in Section 2, the objective function Q in (3.3.6) is of the form

$$Q\left(\beta,\theta,\sigma\,|\,\beta^{(k)},\theta^{(k)},\sigma^{(k)}\right) = C + Q_1\left(\beta,\sigma\,|\,\beta^{(k)},\theta^{(k)},\sigma^{(k)}\right) \tag{3.3.7}$$

$$+ Q_2\left(\theta\,|\,\beta^{(k)},\theta^{(k)},\sigma^{(k)}\right), \tag{3.3.8}$$

where

$$\begin{aligned} Q_1\left(\beta,\sigma|\beta^{(k)},\theta^{(k)},\sigma^{(k)}\right) &= -\frac{(y-X\beta)'(y-X\beta)}{2\sigma^2} - \frac{n-1+p+\nu}{2}\log(\sigma^2) \\ &\quad - \frac{\nu\lambda}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{i=1}^{p}\beta_i^2\,\mathsf{E}_{\gamma|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right], \\ Q_2\left(\theta|\beta^{(k)},\theta^{(k)},\sigma^{(k)}\right) &= \sum_{i=1}^{p}\log\left(\frac{\theta}{1-\theta}\right)\mathsf{E}_{\gamma|\cdot}\,\gamma_i \\ &\quad + (a-1)\log(\theta) + (b+p-1)\log(1-\theta). \end{aligned}$$

Note that Q_2 above corresponds to the beta-binomial prior on γ. Different expressions for Q_2 will be described in Section 3.7 where we consider alternative forms for $\pi(\gamma\,|\,\theta)$.

Two features of this objective function lead to substantial simplifications which facilitate the E-step and M-step calculations described below. First, for the E-step calculation of the expectation in (3.3.6), the hierarchical posterior distribution of γ given $(\beta^{(k)},\theta^{(k)},\sigma^{(k)},y)$ depends on y only through the current estimates $(\beta^{(k)},\theta^{(k)},\sigma^{(k)})$, so that

$$\mathsf{E}_{\gamma|\cdot}(\cdot) = \mathsf{E}_{\gamma|\beta^{(k)},\theta^{(k)},\sigma^{(k)},y}(\cdot) = \mathsf{E}_{\gamma|\beta^{(k)},\theta^{(k)},\sigma^{(k)}}(\cdot). \tag{3.3.9}$$

Second, the separability of (3.3.7) into a pair of distinct functions, Q_1 of (β,σ) and Q_2 of θ, yields an M-step that is obtained by maximizing each of these functions separately.

3.3.1 The E-step

The E-step proceeds by computing the conditional expectations $\mathsf{E}_{\gamma|\cdot}\,\gamma_i$ and $\mathsf{E}_{\gamma|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right]$ and for Q_2 and Q_1, respectively. Considering the latter first, it follows from (3.3.9) that

$$\mathsf{E}_{\gamma|\cdot}\,\gamma_i = \mathsf{P}(\gamma_i = 1\,|\,\beta^{(k)},\theta^{(k)},\sigma^{(k)}) = p_i^{\star}, \tag{3.3.10}$$

where

$$p_i^{\star} = \frac{\pi(\beta_i^{(k)}\,|\,\sigma^{(k)},\gamma_i=1)\mathsf{P}(\gamma_i=1\,|\,\theta^{(k)})}{\pi(\beta_i^{(k)}\,|\,\sigma^{(k)},\gamma_i=1)\mathsf{P}(\gamma_i=1\,|\,\theta^{(k)}) + \pi(\beta_i^{(k)}\,|\,\sigma^{(k)},\gamma_i=0)\mathsf{P}(\gamma_i=0\,|\,\theta^{(k)})}, \tag{3.3.11}$$

Under (3.2.5), the conditional independence of the γ_i's $(i = 1,\ldots,p)$ leads to $\mathsf{P}(\gamma_i = 1 \mid \theta^{(k)}) = \theta^{(k)}$, greatly facilitating the computation of $p_i^\star$. Note that (3.3.11) is equivalent to the posterior update of mixing proportions for fitting a two-point Gaussian mixture to $\beta^{(k)}$ with the conventional EM algorithm.

The other conditional expectation is computed simply as a weighted average of the two precision parameters with weights determined by the posterior distribution $\pi(\gamma_i|\beta^{(k)},\sigma^{(k)},\theta)$, i.e.

$$\mathsf{E}_{\gamma|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right] = \frac{\mathsf{E}_{\gamma|\cdot}(1-\gamma_i)}{v_0} + \frac{\mathsf{E}_{\gamma|\cdot}\gamma_i}{v_1} = \frac{1-p_i^\star}{v_0} + \frac{p_i^\star}{v_1} = d_i^\star. \tag{3.3.12}$$

3.3.2 The M-step

Maximization with respect to (β,σ) is facilitated by the separability of the objective function, as noted above, and by the conjugacy of the prior formulation which led to the tractable closed form expressions. Beginning with the maximization of Q_1, the $\beta^{(k+1)}$ value that globally maximizes Q_1, regardless of $\sigma^{(k+1)}$, is obtained quickly by the well-known solution to the ridge regression problem

$$\beta^{(k+1)} = \text{argmin}_{\beta\in\mathbb{R}^p}\{||y - X\beta||^2 + ||D^{\star 1/2}\beta||^2\}, \tag{3.3.13}$$

where $||\cdot||^2$ is the l_2 norm and $D^{\star 1/2}$ denotes the square root of the $p\times p$ diagonal matrix $D^\star = \text{diag}\{d_i^\star\}_{i=1}^p$ with diagonal entries $d_i^\star > 0$ from (3.3.12). The solution

$$\beta^{(k+1)} = (X'X + D^\star)^{-1}X'y \tag{3.3.14}$$

is a generalized ridge estimator (GRR) with ridge matrix $D^\star$ which allows a unique penalty parameter $d_i^\star$ for each individual coefficient β_i. This induces a "selective shrinkage" property which shrinks the smaller coefficient estimates much more sharply towards zero compared to the larger coefficients, a consequence of the spike-and-slab prior, see Ishwaran and Rao (2005). An important property of the estimator (3.3.14) is that it is well defined even when $X'X$ is not invertible.

In problems where $p >> n$, the calculation cost of (3.3.14) can be substantially reduced by using the Sherman-Morrison-Woodbury formula to obtain

$$\beta^{(k+1)} = \left[D^{\star -1} - D^{\star -1}X'\left(\mathrm{I}_{n\times n} + XD^{\star -1}X'\right)^{-1}XD^{\star -1}\right]X'y, \tag{3.3.15}$$

an expression which requires an $n\times n$ matrix inversion rather than a $p\times p$ matrix inversion. Alternatively, as described in Section 3.8, the solution of (3.3.13) can be obtained even faster with the stochastic dual coordinate ascent algorithm of Shalev-Shwartz and Zhang (2013).

The maximization of $Q_1\left(\beta,\sigma|\beta^{(k)},\theta^{(k)},\sigma^{(k)}\right)$ with respect to (β,σ) is then completed with the simple update

$$\sigma^{(k+1)}=\sqrt{\frac{||y-X\beta^{(k+1)}||^2+||D^{\star 1/2}\beta^{(k+1)}||^2+\nu\lambda}{n+p+\nu}}. \tag{3.3.16}$$

Note that the convenient M-step forms for $\beta^{(k+1)}$ and $\sigma^{(k+1)}$ resulted from proceeding conditionally on σ throughout the EM algorithm. Had we initially margined out σ over its prior, the resulting posterior under spike-and-slab mixtures of t distributions would have been prohibitively expensive to maximize.

Turning to Q_2, its maximization is obtained by the closed form solution of

$$\theta^{(k+1)}=\text{argmax}_{\theta\in\mathbb{R}}\Big\{\sum_{i=1}^{p}p_i^{\star}\log\left(\frac{\theta}{1-\theta}\right)+(a-1)\log(\theta) \\ +(p+b-1)\log(1-\theta)\Big\},$$

namely

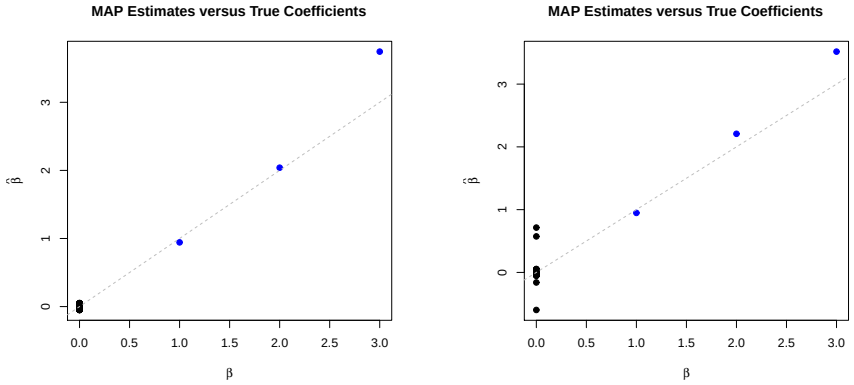
$$\theta^{(k+1)}=\frac{\sum_{i=1}^{p}p_i^{*}+a-1}{a+b+p-2}.$$

The EM algorithm has been previously considered in the context of Bayesian shrinkage estimation under sparsity priors (Figueiredo (2003)), Kiiveri (2003), Griffin and Brown (2012, 2005). Literature on similar computational procedures for spike and slab models is far more sparse. EM-like algorithms using point mass variable selection priors were considered by Hayashi and Iwata (2010) and Bar et al. (2010), but were limited by the unavailability of the closed form E-step.

3.4 The EMVS Approach

In this section we outline the EMVS approach for variable selection. This entails dynamic posterior exploration over a sequence of nested spike-and-slab priors as $v_0>0$ is gradually increased. For each value of v_0, the EM algorithm is deployed to identify a posterior mode $(\widehat{\beta},\widehat{\theta},\widehat{\sigma})$ which is then thresholded to obtain a closely associated submodel. The detailed description of the thresholding rule to obtain the lower dimensional submodels is given in Section 3.4.1. Section 3.4.2 then describes the "spike-and-slab regularization diagram", which captures the evolution of the modal estimates as well as the model configurations and their posterior probabilities over the sequence of different $v_0>0$.

For clarity of exposition, we illustrate the various steps of this approach with a simple simulated dataset consisting of $n=100$ observations and $p=1\,000$ predictors. Predictor values for each observation were simulated from $\text{N}_p(0,\Sigma)$ where $\Sigma=(\rho_{ij})_{i,j=1}^{p}$ with $\rho_{ij}=$



(a) $v_0 = 0.5, v_1 = 1000, \theta \sim U(0,1)$ (b) $v_0 = 0.5, v_1 = 1000, \theta = 0.5$

Figure 3.1: Modal estimates of the regression coefficients; (a) beta binomial prior, (b) Bernoulli prior with fixed $\theta = 0.5$.

$0.6^{|i-j|}$. Response values were then generated according to the linear model $y = X\beta + \varepsilon$ where $\beta = (1,2,3,0,0,\ldots,0)'$ and $\varepsilon \sim \mathrm{N}_n(0, \sigma^2 \mathrm{I}_n)$ with $\sigma^2 = 3$.

Beginning with an illustration of the EM algorithm from Section 3.3, we apply it to the simulated data using the spike-and-slab prior (3.2.2) with a single value $v_0 = 0.5$, $v_1 = 1000$, and the beta-binomial variable inclusion prior with $\theta \sim U(0,1)$. The starting values for the EM algorithm were set to $\beta^{(0)} = 1_p$ and $\sigma^{(0)} = 1$. After merely 4 iterations, the algorithm obtained the modal coefficient estimates $\widehat{\beta}$ depicted in Figure 3.1(a). Note that although they are all nonzero because of $v_0 > 0$, many of them are small in magnitude, a consequence of the ridge regression shrinkage induced by the spike-and-slab prior. The associated modal estimates of $\widehat{\theta}$ and $\widehat{\sigma}$ were 0.003 and 0.037, respectively.

For comparison, we applied the same formulation except with the Bernoulli prior (3.2.5) under fixed $\theta = 0.5$ (Figure 3.1(b)). Note the inferiority of the estimates near zero due to the lack of adaptivity of the Bernoulli prior in determining the degree of underlying sparsity.

3.4.1 Thresholding the EM Output for Variable Selection

Looking at Figure 3.1(a), it seems intuitively reasonable that the submodel most closely associated with the EM estimate $\widehat{\beta}$ is the one that includes only the variables corresponding to the three large estimates. This intuition is supported by defining the submodel $\widehat{\gamma}$ associated with $(\widehat{\beta}, \widehat{\theta}, \widehat{\sigma})$ to be the most probable γ given $(\beta, \theta, \sigma) = (\widehat{\beta}, \widehat{\theta}, \widehat{\sigma})$, namely

$$\widehat{\gamma} = \arg\max_{\gamma} \mathrm{P}(\gamma \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}). \tag{3.4.17}$$

To obtain $\widehat{\gamma}$, note that

$$\mathsf{P}(\gamma \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}) = \prod_{i=1}^{p} \mathsf{P}(\gamma_i \,|\, \widehat{\beta}_i, \widehat{\theta}, \widehat{\sigma}), \tag{3.4.18}$$

where the component conditional inclusion probabilities are given by

$$\mathsf{P}(\gamma_i \,|\, \widehat{\beta}_i, \widehat{\theta}, \widehat{\sigma}) = \frac{\pi(\widehat{\beta}_i \,|\, \widehat{\sigma}, \gamma_i)\mathsf{P}(\gamma_i \,|\, \widehat{\theta})}{\pi(\widehat{\beta}_i \,|\, \widehat{\sigma}, \gamma_i = 1)\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\theta}) + \pi(\widehat{\beta}_i \,|\, \widehat{\sigma}, \gamma_i = 0)\mathsf{P}(\gamma_i = 0 \,|\, \widehat{\theta})}. \tag{3.4.19}$$

Thus, (3.4.17) is obtained by maximizing each component probability, namely

$$\widehat{\gamma}_i = 1 \iff \mathsf{P}(\gamma_i = 1 \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}) \geq 0.5. \tag{3.4.20}$$

It may be of interest to note that $\widehat{\gamma}$ is a local version of the median probability model of Barbieri and Berger (2004).

Selection of $\widehat{\gamma}$ via (3.4.20) is equivalent to thresholding the $\widehat{\beta}_i$ values because $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\beta}_i, \widehat{\theta}, \widehat{\sigma})$ is a monotone increasing function of $|\widehat{\beta}_i|$. This thresholding can be seen to occur at the intersection points $\pm\beta_i^*$ of the $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\theta})$ weighted mixture of the spike-and-slab priors, namely

$$\pm\beta_i^*(v_0, v_1, \widehat{\theta}, \widehat{\sigma}) = \pm\widehat{\sigma}\sqrt{2v_0 \log(\omega_i c)\, c^2/(c^2 - 1)}, \tag{3.4.21}$$

where $c^2 = v_1/v_0$ and $\omega_i = [1 - \mathsf{P}(\gamma_i = 1 \,|\, \widehat{\theta})]/\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\theta})$. Thus, (3.4.20) is equivalent to

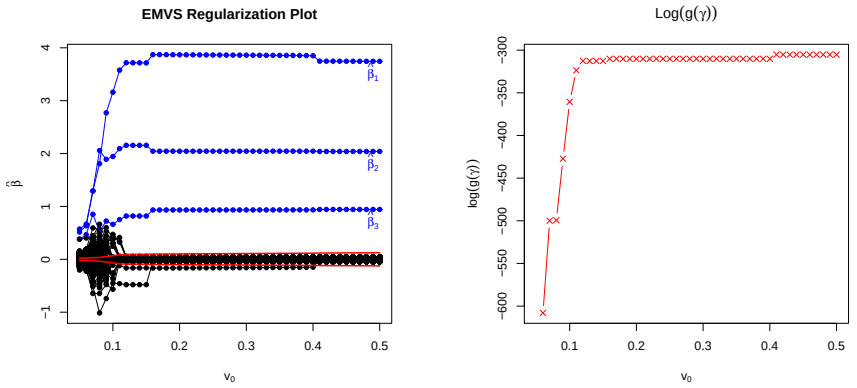
$$\widehat{\gamma}_i = 1 \iff |\widehat{\beta}_i| \geq \beta_i^*(v_0, v_1, \widehat{\theta}, \widehat{\sigma}). \tag{3.4.22}$$

Applying this thresholding rule to the estimates in Figure 3.1(a) yields the correct three predictor submodel, in contrast to Figure 3.1(b) where some of the small coefficients are not thresholded out. The increased weighting on parsimonious models induced by the beta-binomial formulation has proved to be beneficial here.

We should point out that although (3.4.21) may vary across variables with different inclusion probabilities, it will not vary under the beta-binomial prior, where $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\theta}) \equiv \widehat{\theta}$, the overall conditional probability of inclusion. Because the values of $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma})$ accumulate around zero and one for $\widehat{\beta}_i$'s far from either of $\pm\beta_i^*$, it is likely that such selection will not be too sensitive to the threshold of 0.5 in (3.4.20). Nonetheless, it may be useful to also consider larger threshold values to obtain sparser models.

3.4.2 Variable Selection with a Spike-and-Slab Regularization Plot

Rather than restricting attention to selection based on a single value for v_0, the speed of the EM algorithm makes it feasible to consider a sequence of selected submodels as v_0 is varied over a set V of values, a strategy we recommend for EMVS. The effect of increasing v_0



(a) EMVS Regularization Plot

(b) Submodel Evaluations

Figure 3.2: (a) plot of estimated regression coefficients for varying choices of v_0, red lines correspond to the varying benchmark threshold; (b) logarithm of $g(\gamma)$ for models with selected variables outside the threshold.

serves to absorb more of the negligible coefficients into the spike distribution, thereby reducing posterior multimodality and exposing sparse high probability subsets for thresholding identification.

To illustrate how this works with our simulated data, we consider the grid of v_0 values $V = \{0.01 + k \times 0.01 : k = 0, \ldots, 50\}$ again with $v_1 = 1000$ fixed and the same beta-binomial inclusion prior. Figure 3.2(a) shows the modal estimates of the regression coefficients obtained for each $v_0 \in V$. As v_0 increases, more variables fall within the $\pm\beta_i^*$ threshold limits depicted by the two red lines, and the estimates of the large effects stabilize. It it worth noting the difficulty of subset identification when v_0 is small and no clear model emerges.

By analogy with LASSO regularization plots that display the effect of an increasing penalty parameter (Tibshirani, 1994), we refer to plots such as Figure 3.2(a) as (spike-and-slab) regularization plots since they provide a visualization of the effect of an increasing v_0. Indeed, both the LASSO penalty and v_0 serve to pull coefficient estimates towards zero although they do so in very different ways. Increasing the LASSO penalty parameter corresponds to decreasing the variance of single unimodal prior thereby shrinking all coefficients towards zero. In contrast, an increasing v_0 corresponds to increasing the variance of the spike component of the spike-and-slab mixture. This has the effect of shrinking the smaller coefficients with the spike distribution without very much affecting the larger coefficients which are supported more by the slab distribution.

For each $v_0 \in V$, the thresholded EM output determines an active set of variables $\mathscr{S}_{v_0} = \{x_i : |\widehat{\beta}_i| > \beta_i^*(v_0, v_1, \widehat{\theta}, \widehat{\sigma}\}$. Letting $\widehat{\gamma}_{v_0}$ denote the submodel identified by $\mathscr{S}_{v_0}$, the full procedure thus effectively generates a solution path $\{\widehat{\gamma}_{v_0} : v_0 \in V\}$ through model space. To select

the "best" γ from this solution path, a natural criterion is the marginal probability of γ under the prior with $v_0 = 0$, a marginal we denote by $\pi_0(\gamma|y)$. The appeal of $\pi_0(\gamma|y)$ is that it evaluates $\gamma = (\gamma_1, \ldots, \gamma_p)'$ according to the submodel containing only those variables for which $\gamma_i = 1$. This would not be the case for the marginal probability under $v_0 > 0$, which would always evaluate γ on the basis of a full model where coefficient estimates corresponding to $\gamma_i = 0$ were shrunk only to be small. In effect, we are contemplating that the statistician would have preferred a full comparison of all models using $\pi_0(\gamma|y)$, but to avoid the difficulties associated with the implementation of such an analysis, has used the thresholded EM procedure as a device to identify promising submodels.

As shown by GM97, except for an unknown normalizing constant C, a rapidly computable closed form

$$g_0(\gamma) = C\,\pi_0(\gamma|y) \tag{3.4.23}$$

is available. This $g_0(\gamma)$ serves our purposes perfectly since it suffices for identifying the $\gamma \in \{\widehat{\gamma}_{v_0} : v_0 \in V\}$ for which $g_0(\gamma)$ is largest. To illustrate how this would work on our simulated dats, Figure 3.2(b) plots $\log g_0(\gamma)$ values for all models visited along the solution path. We observe a clearly escalating trend, where the largest posterior probability is obtained by the correct model, namely the model which includes only x_1,x_2 and x_3.

3.4.3 A Speed Comparison with Stochastic Search

It may be of interest to consider how stochastic search Bayesian variable selection would fare on the same simulated data used throughout this section. For this purpose, we considered the same conjugate spike-and-slab prior with $v_0 = 0$, $v_1 = 1\,000$ and beta-binomial model prior with $\theta \sim U(0,1)$, and implemented a Metropolis-Hastings (MH) sampler with a one-step random scan proposal to simulate from the marginal posterior on γ. To put EMVS and the MH sampler on equal footing in terms of initialization, we started the sampler at $\gamma^{(0)} = 0_{1\,000}$, which is the local median probability model obtained by thresholding the EMVS initialization $\beta^{(0)} = 1_{1\,000}$, $\sigma^{(0)} = 1$, $\theta^{(0)} = 0.5$ when $v_0 = 0.5$ and $v_1 = 1\,000$.

We ran the MH algorithm for the same amount of time it took EMVS to generate the entire regularization path (consisting of 51 v_0 values) in Figure 2. In this time, the MH algorithm generated 50 000 iterations with an acceptance rate 0.0001 for $v_1 = 1\,000$. The model including only the predictors $\{2,3\}$, rather than $\{1,2,3\}$, was obtained as both the maximum $g_0(\gamma)$ model and the median probability model. Repeating the stochastic search with $v_1 = 1, 10, 100$ yielded higher acceptance rates, but still always identified $\{2,3\}$ as the model and median model. Repeating the stochastic search initialized at the full model $\gamma^{(0)} = 1_p$, (the local median probability model for the EMVS initialization with $v_0 = 0.1$), was disappointing. Performing merely 10 iterations with a zero acceptance rate due to the complexity of evaluating $g_0(\gamma)$ for rich models, the MH sampler never identified a model even close to $\{1,2,3\}$. In a setting where EMVS rapidly identified the correct model, the MH sampler failed to do so in a comparable amount of time, even when initialized in the close vicinity of the true mode. It

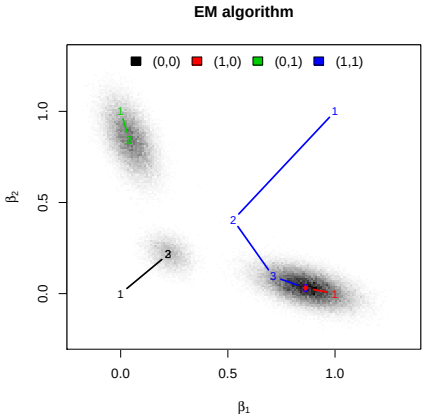


Figure 3.3: Posterior distribution $p(\beta \mid Y)$ arising from a conjugate SSVS model with $v_0 = 0.005, v_1 = 1\,000$ together with EM iterative steps for 4 different initializations.

should also be noted that, in contrast to the MH sample, the deterministic nature of the EMVS computation would always yield reproducible results.

3.5 Mitigating Multimodality with Deterministic Annealing

A potential drawback of the EM algorithm occurs in multimodal posterior landscapes where it can be prone to entrapment in local maximum modes. To illustrate this undesirable phenomenon, we investigate the performance of the EM algorithm on a simple simulated example. We construct $n = 100$ observations on $p = 2$ predictors according to $\mathrm{N}_p(0, \Sigma)$ with $\Sigma = (\rho_{ij})_{i,j=1}^{p}$ and $\rho_{ij} = 0.9^{|i-j|}$. We consider the following regression vector $\beta = (1,0)'$ and generate responses according to $\mathrm{N}_n(X\beta, \sigma^2 \mathrm{I}_n)$ with $\sigma^2 = 3$. The resulting maximum likelihood estimates are $\widehat{\beta}_{MLE} = (0.52, 0.4)'$ and $\widehat{\sigma}_{MLE} = 1.8$. For this problem, we apply the EM algorithm for posterior modal estimation under the conjugate SSVS model. For small enough values v_0 we expect the posterior distribution $p(\beta \mid Y)$ to be multimodal, partially due to the high correlation between the predictors. Setting $v_1 = 1\,000$ and $v_0 = 0.005$, we proceed to explore the posterior distribution $p(\beta \mid Y)$ using the Gibbs sampler as described in GM97.

Figure 3.3 depicts the MCMC approximation to the posterior distribution $p(\beta \mid Y)$ obtained after 100 000 iterations. The accumulation of the posterior probability is displayed in various degrees of grey, where darker areas are associated with higher posterior modes. The global mode (marked with a red dot) is located at the point $\widehat{\beta}_{MAP} = (0.86, 0.03)'$. As opposed to the maximum likelihood estimate, the posterior mode better recovers the underlying

regression structure, a consequence of the selective shrinkage property of the spike and slab prior. Figure 3.3 also plots the iterative process of the EM algorithm using 4 different starting points ($\beta^{(0)} = (0,0)', \beta^{(0)} = (0,1)', \beta^{(0)} = (1,0)'$ and $\beta^{(0)} = (1,1)'$). Expectedly, we observe that for initial values located at the close vicinity of local modes, the EM algorithm fails to converge to the global maximum.

To mitigate this issue, a general recommendation (McLachlan and Basford, 2004) is to run the algorithm for a wide choice of starting values. To further improve the chances of finding a global mode, one might also consider the deterministic annealing variant of the EM algorithm (DAEM) proposed by Ueda and Nakano (1998).

Using the principle of maximum entropy and an analogy with statistical mechanics, the DAEM algorithm aims at finding a minimum of a tempered version of the objective function, often called the free energy function. In our context, this is equivalent to finding the maximum of the negative free energy function

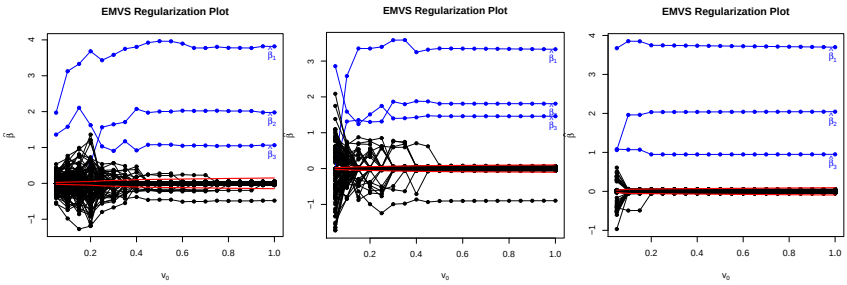
$$H_t(\beta, \theta, \sigma) = \frac{1}{t} \log \sum_{\gamma} \pi(\beta, \theta, \sigma, \gamma \,|\, y)^t \quad \text{with} \quad 0 < t \leq 1, \tag{3.5.24}$$

which embeds the actual log incomplete posterior as a special case when $t = 1$. In (3.5.24), $1/t$ corresponds to a temperature parameter and determines the degree of separation between the multiple modes of H_t. Starting with large enough temperatures which smooth away the local modes of H_t, as the temperature is decreased, multiple modes begin to appear and H_t gradually resembles the actual incomplete posterior. Thus, the influence of poorly chosen starting values can be weakened by keeping the temperature high at the early stage of computation and gradually decreasing it during the iteration process. Alternatively, (3.5.24) can be optimized for a decreasing sequence of temperature levels $1/t_1 > 1/t_2 > \cdots > 1/t_k$, where the solution at $1/t_i$ serves as the starting point for the computation at $1/t_{i+1}$. Provided that the new global maximum is close to the previous one, this strategy can increase the chances of finding the true global maximum.

To extend our EM algorithm to incorporate DAEM iterations, the M-step remains unchanged. However, the E-step requires the computation of the expected complete log posterior density with respect to a modified posterior distribution. This distribution, derived using the maximum entropy principle, is proportional to a current estimate of the conditional complete posterior given the observed data raised to the power t. Particularly easy to derive for mixtures (Ueda and Nakano, 1998), in our context this distribution is simply obtained by replacing $p_i^\star$ in (3.3.11) with

$$p_{i,t}^\star = \frac{\pi(\beta_i^{(k)} \,|\, \sigma^{(k)}, \gamma_i = 1)^t \mathsf{P}(\gamma_i = 1 \,|\, b^{(k)})^t}{\pi(\beta_i^{(k)} \,|\, \sigma^{(k)}, \gamma_i = 1)^t \mathsf{P}(\gamma_i = 1 \,|\, b^{(k)})^t + \pi(\beta_i^{(k)} \,|\, \sigma^{(k)}, \gamma_i = 0)^t \mathsf{P}(\gamma_i = 0 \,|\, b^{(k)})^t}. \tag{3.5.25}$$

The deterministic annealing version of EMVS, which we shall refer to as DAEMVS, is obtained by making this substitution in (3.3.11). At high temperatures (t close to zero)



(a) EMVS (b) DAEMVS $t = 0.2$ (c) DAEMVS $t = 0.1$

Figure 3.4: Regularization plots for the simulated example from Section 3.4 using EMVS and the deterministic annealing EMVS (DAEMVS) considering randomly generated starting vector $\beta^{(0)} \sim \mathrm{N}_{1000}(0, \mathrm{I})$

the probabilities (3.5.25) become nearly uniform, as can be seen from the limiting behavior $\lim_{t\to 0} p^{\star}_{i,t} \equiv 0.5$. Thus tempering induces more equal penalties on all the coefficients through (3.3.13) regardless of their magnitude.

Finally, under (3.5.25) as $t \to 0$, the unique posterior mode $\widehat{\beta}$ turns out to be a very promising general initialization value for EMVS. This mode is easily obtained as the M-step ridge regression solution (3.3.14) with equal penalties $\frac{v_0+v_1}{2v_0v_1}$, namely

$$\widehat{\beta}_{t=0} = \left[X'X + \frac{v_0+v_1}{2v_0v_1} I_p\right]^{-1} X'y. \tag{3.5.26}$$

3.5.1 Revised Analysis of the Simulated Example

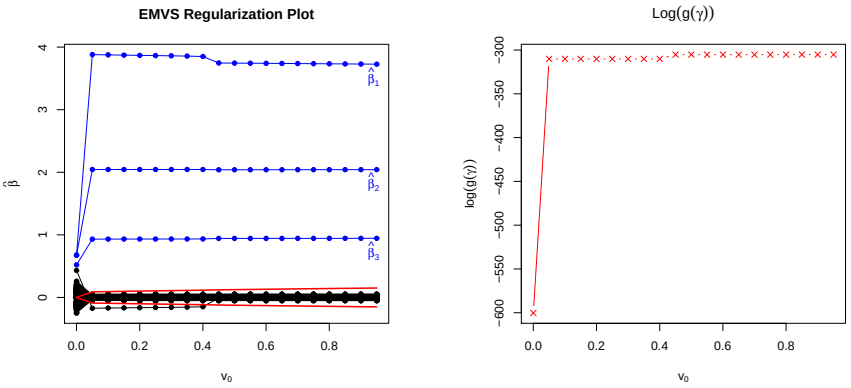
In Section 3.4 we illustrated the EMVS procedure on a simple simulated example with a single set of starting values $\widehat{\beta}^{(0)} = 1_p$. Here we apply EMVS and its tempered version DAEMVS on the same data using a randomly generated starting vector $\beta^{(0)} \sim \mathrm{N}_{1000}(0, \mathrm{I})$ in order to demonstrate the sensitivity of EMVS to initialization and the potential of deterministic annealing. In the process, it is also seen how posterior multimodality is diminished as v_0 is increased, making it easier to find global modes. For all these illustrations, the slab parameter was set to $v_1 = 1\,000$.

The resulting regularization diagrams in Figure 3.4 for (DA)EMVS at temperatures $1/t =$ 5 and 10 show that EMVS is postponing the detection of sparse models until larger values of v_0. In contrast, deterministic annealing lessens multimodality for smaller values v_0, exposing the correct model more quickly. Note the increasing success of all three algorithms as v_0 gets larger.

To further illustrate the impact of initial values scattered farther away from the true coefficient vector, we considered two other randomly generated starting vectors $\beta^{(0)} \sim \mathrm{N}_{1000}(0, 3\times$

	$\beta^{(0)} = 1_{1000}$			$\beta^{(0)} \sim \mathrm{N}_{1000}(0, \mathrm{I})$			$\beta^{(0)} \sim \mathrm{N}_{1000}(0, 3 \times \mathrm{I})$			$\beta^{(0)} \sim \mathrm{N}_{1000}(0, 5 \times \mathrm{I})$		
	#Iter	#Var	$\log g_0(\gamma)$	#Iter	#Var	$\log g_0(\gamma)$	#Iter	#Var	$\log g_0(\gamma)$	#Iter	#Var	$\log g_0(\gamma)$
$v_0 = 0.2$												
EMVS	4*	4*	-310.16*	13*	73*	-529.06*	18	148	-565.32	8	202	-587.36
DAEMVS ($t = 0.2$)	29*	10*	-313.07*	43*	12*	-329.85*	16	71	-597.32	7	95	-515.42
DAEMVS ($t = 0.1$)	**6**	**3**	-305.24	**7**	**3**	-305.24	9	12	-342.18	27*	33*	-428.34*
$v_0 = 0.6$												
EMVS	**5**	**3**	-305.24	5*	9*	-335.17*	10	81	-579.11	13	102	-523.59
DAEMVS ($t = 0.2$)	**5**	**3**	-305.24	**5**	**3**	-305.24	9*	6*	-316.35*	24*	9*	-329.32*
DAEMVS ($t = 0.1$)	**5**	**3**	-305.24	**5**	**3**	-305.24	**6**	**3**	-305.24	**6**	**3**	-305.24
$v_0 = 1$												
EMVS	**4**	**3**	-305.24	5*	4*	-308.54*	9	19	-369.24	9	77	-606.51
DAEMVS ($t = 0.2$)	**5**	**3**	-305.24	**5**	**3**	-305.24	**8**	**3**	-305.24	8*	4*	-310.41*
DAEMVS ($t = 0.1$)	**6**	**3**	-305.24	**6**	**3**	-305.24	**6**	**3**	-305.24	**5**	**3**	-305.24

Table 3.1: Performance evaluation of the EMVS and deterministic annealing DAEMVS procedures considering different temperature parameters and starting value sets on a simulated data example from Section 3.4. Numbers of iterations until convergence are tabulated, as well as numbers of selected variables and $\log g_0$ evaluated at each selected model. Correctly identified model is indicated with bold font. All the other models which include the three true predictors are designated with a star.



(a) EMVS Regularization Plot

(b) Submodel Evaluations

Figure 3.5: (a) plot of estimated regression coefficients for varying choices of v_0, red lines correspond to the varying benchmark threshold; (b) $\log g_0(\gamma)$ for models with selected variables outside the threshold.

I) and $\beta^{(0)} \sim \mathrm{N}_{1000}(0, 5 \times \mathrm{I})$. For three different values of v_0 (0.2, 0.6 and 1), we applied EMVS and DAEMVS at temperatures $1/t = 5$ and 10, keeping track of the number of iterations to convergence, the number of selected active predictors, and $\log g_0$ evaluated over the solution path of models. These quantities are tabulated in Table 3.5.1.

We observe that depending on the choice of starting vector, the EMVS algorithm converged to a different solution for each v_0. In contrast, at higher temperatures and larger values of v_0, DAEMVS converged to the correct model even from distant starting values. Evidently, tempering together with larger v_0 act in conjunction to reduce posterior multimodality and gravitate smaller coefficient estimates towards zero.

Finally, we note that $\widehat{\beta}_{t=0}$ in (3.5.26) fared superbly as a starting value on this data. Indeed, EMVS without any tempering very quickly detected the correct model as is evidenced by regularization plot Figure 3.5(a). We recommend this starting value as a general choice for consideration in practice.

3.6 A Heavy-Tailed Slab Distribution

Under the spike-and-slab prior, we would ideally like the slab distribution to leave large coefficient estimates relatively unaffected. To this end, the prior needs to limit shrinkage of the large effects, while providing enough support to keep them away from being shrunk by the spike distribution. This can be achieved under our formulation by adding a prior $\pi(v_1)$ to induce a heavy tailed slab distribution.

To gain insight into the shrinkage properties of our spike-and-slab prior formulation, consider that the induced MAP estimates are regularized estimates arising as solutions to the penalized least squares problem

$$\widehat{\beta} = \text{argmin}_{\beta} \left\{ \frac{1}{2\sigma^2} ||y - X\beta||^2 + \sum_{i=1}^{p} pen_{v_0,v_1,\gamma_i}(|\beta_i|) \right\}, \tag{3.6.27}$$

where $pen_{v_0,v_1,\gamma_i}(|\beta_i|)$ contains the term in $-\log \pi(\beta_i|v_0,v_1,\gamma_i)$ which depends on β_i. When the columns of X are orthonormal, the problem (3.6.27) can be solved component-wise (Fan and Li, 2001):

$$\widehat{\beta}_i = \text{argmin}_{\beta_i} \left\{ \frac{1}{2}(\widetilde{\beta}_i - \beta_i)^2 + \sigma^2 pen_{v_0,v_1,\gamma_i}(|\beta_i|) \right\}, \tag{3.6.28}$$

where $\widetilde{\beta} = X'y$. Taking the first derivative of (3.6.28) with respect to β_i, it can be seen that the term $pen'_{v_0,v_1,\gamma_i}(|\beta_i|) = \frac{\partial pen_{v_0,v_1,\gamma_i}(|\beta_i|)}{\partial |\beta_i|}$ biases estimates towards zero. Fan and Li (2001) characterize bias-reducing penalty functions as those for which $pen'_{v_0,v_1,\gamma_i}(|\beta_i|)$ approaches zero at a fast rate as $|\beta_i| \to \infty$.

Because the Gaussian tails of the spike prior go to zero so quickly, the tail behavior of the spike-and-slab prior is for large enough $|\beta_i|$ dominated by the tails of the slab component, and so it suffices to focus on the slab distribution. A Gaussian slab prior is less appealing as the derivative of the penalty is an increasing function of $|\beta_i|$. A Laplace prior on the other hand implies constant bias irrespective of the magnitude of $|\beta_i|$. Griffin and Brown (2005) propose alternative shrinkage distributions arising from normal scale mixtures by considering various mixing distributions for the variance parameter. Similar distributions were considered by other authors including Strawderman (1971), Carvalho and Polson (2010).

To induce a heavy tailed slab prior for EMVS, we consider adding the prior proposed in the g-prior context by Maruyama and George (2011),

$$\pi(v_1) = \frac{v_1^b (1+v_1)^{-a-b-2}}{\text{B}(a+1,b+1)} \text{I}_{(0,\infty)}(v_1), \tag{3.6.29}$$

a Pearson Type VI or beta-prime distribution under which $1/(1+v_1)$ has a Beta distribution $\text{Be}(a+1,b+1)$. See also Cui and George (2008) and Liang et al. (2008) who proposed the special case of (3.6.29) with $b = 0$.

The marginal spike-and-slab prior on β_i obtained after integrating out the parameter v_1 with respect to the prior distribution (3.6.29) can be for $a > -1.5$ and $b > -0.5$ (Gradshteyn and Ryzhik, 2000, p. 362) written as

$$\pi(\beta_i|v_0,\sigma,\gamma) = (1-\gamma_i)\text{N}(0,\sigma^2 v_0) + \gamma_i \widetilde{\pi}_{a,b,\sigma}(\beta_i). \tag{3.6.30}$$

Here $\widetilde{\pi}_{a,b,\sigma}(\beta_i)$ denotes a density function

$$\widetilde{\pi}_{a,b,\sigma}(\beta_i) = \frac{\Gamma\left(a+\frac{3}{2}\right)}{B(a+1,b+1)} \frac{\exp\left(\frac{\beta_i^2}{4\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \left(\frac{\beta_i^2}{2\sigma^2}\right)^{\frac{b}{2}-\frac{1}{4}} \times W_{-a-\frac{b}{2}-\frac{5}{4},-\frac{b}{2}-\frac{1}{4}}\left(\frac{\beta_i^2}{2\sigma^2}\right), \tag{3.6.31}$$

where $W_{\eta,\psi}$ is the Whittaker function (Abramowitz and Stegun, 1972, p. 505). When $b = 0$, (3.6.31) is equivalent to the normal-exponential-gamma (NEG) prior (Griffin and Brown, 2012) obtained by imposing the hierarchical distribution $p(v_1|\lambda) = \lambda \exp(-\lambda v_1)$ with $p(\lambda) = \frac{1}{\Gamma(a+1)}\lambda^a \exp(-\lambda)$. The density of the NEG distribution

$$\widetilde{\pi}_{a,0,\sigma}(\beta_i) = \frac{(a+1)2^{a+\frac{3}{2}}}{\sqrt{2\pi\sigma^2}} \Gamma\left(a+\frac{3}{2}\right) \exp\left(\frac{\beta_i^2}{4\sigma^2}\right) D_{-2\left(a+\frac{3}{2}\right)}\left(|\beta_i|/\sigma\right), \tag{3.6.32}$$

follows from the identity $D_\eta(z) = 2^{\frac{1}{4}+\frac{\eta}{2}} W_{\frac{1}{4}+\frac{\eta}{2},-\frac{1}{4}}\left(\frac{z^2}{2}\right) z^{-\frac{1}{2}}$ (Gradshteyn and Ryzhik, 2000, p. 1018), where D_η denotes the parabolic cylinder function (Abramowitz and Stegun, 1972, p. 685).

It is illuminating to study the limiting behavior of the implicit bias term $\widetilde{pen}'_{v_0,v_1,\gamma_i}(|\beta_i|)$ as $|\beta_i| \to \infty$. It is desirable that the bias term diminishes rapidly as coefficients get farther away from zero. The asymptotic properties of the bias term are summarized in the following theorem.

Theorem 3.6.1. *Let $\widetilde{\pi}_{a,b,\sigma}(\beta_i)$ be the distribution given in* (3.6.31) *with $a > -\frac{3}{2}$ and $b > \frac{1}{2}$. Denote $\widetilde{pen}'_{a,b,\sigma}(|\beta_i|) = \frac{\partial \log \widetilde{\pi}_{a,b,\sigma}(\beta_i)}{\partial |\beta_i|}$. Then*

$$\widetilde{pen}'_{a,b,\sigma}(|\beta_i|) = \frac{\sqrt{2}\left(a+\frac{3}{2}\right)}{\sigma} \frac{W_{-a-\frac{b}{2}-\frac{7}{4},-\frac{b}{2}-\frac{1}{4}}\left(\frac{\beta_i^2}{2\sigma^2}\right)}{W_{-a-\frac{b}{2}-\frac{5}{4},-\frac{b}{2}+\frac{1}{4}}\left(\frac{\beta_i^2}{2\sigma^2}\right)} \tag{3.6.33}$$

and $\widetilde{pen}'_{a,b,\sigma}(|\beta_i|) = \mathscr{O}\left(\frac{1}{|\beta_i|}\right)$ as $|\beta_i| \to \infty$.

Proof. The proof of the expression (3.6.33) is facilitated by noting that

$$\widetilde{pen}'_{a,b,\sigma}(|\beta_i|) = \frac{\partial \widetilde{\pi}_{a,b,\sigma}(\beta_i)/\partial|\beta_i|}{\widetilde{\pi}_{a,b,\sigma}(\beta_i)}.$$

The denominator can be for $b > -\frac{1}{2}$ and $a > -\frac{3}{2}$ rewritten using the expression for marginal prior distribution in (3.6.31). The numerator can be expressed as

$$\frac{|\beta_i|}{B(a+1,b+1)\sqrt{2\pi\sigma^2}\sigma^2} \int_0^\infty v_1^{b-\frac{3}{2}} \exp\left(-\frac{\beta_i^2}{2\sigma^2 v_1}\right)(1+v_1)^{-a-b-2} \mathrm{d}v_1.$$

This identity follows from the Leibnitz integral rule, which is justified since the integrand is a positive integrable function on $(0,\infty)$ for $b > \frac{1}{2}$ and $a > -\frac{5}{2}$. According to (Gradshteyn and Ryzhik, 2000, p. 362), we can then for $b > \frac{1}{2}$ and $a > -\frac{5}{2}$ write $\frac{\partial \widetilde{\pi}_{a,b,\sigma}(\beta_i)}{\partial |\beta_i|}$ as

$$\frac{|\beta_i|}{B(a+1,b+1)\sqrt{2\pi\sigma^2}\sigma^2}\left(\frac{\beta_i^2}{2\sigma^2}\right)^{\frac{b}{2}-\frac{3}{4}}\Gamma\left(a+\frac{5}{2}\right)\exp\left(\frac{\beta_i^2}{4\sigma^2}\right)\times W_{-a-\frac{b}{2}-\frac{7}{4},-\frac{b}{2}+\frac{1}{4}}. \tag{3.6.34}$$

The identity (3.6.34) together with the expression for the marginal distribution $\widetilde{\pi}_{a,b,\sigma}(\beta_i)$ then completes the proof of the equation (3.6.33).

The limiting behavior of the term $\widetilde{pen}'_{a,b,\sigma}(|\beta_i|)$ can be better understood using the Poicare expansion of Whittaker function for large $|z|$ (Gradshteyn and Ryzhik, 2000, p. 1016), namely

$$W_{\eta,\psi}(z) \sim \exp\left(-\frac{z}{2}\right) z^{\eta}\left(1+\sum_{k=1}^{\infty}\frac{[\psi^2-(\eta-\frac{1}{2})^2]\ldots[\psi^2-(\eta-k+\frac{1}{2})^2]}{k!z^k}\right), \tag{3.6.35}$$

where $\sim$ sign indicates that the Whittaker function is equal to the series in the limit as $|z| \to \infty$. As a consequence, we have

$$\lim_{|z|\to\infty}\frac{W_{\eta,\psi}(z)}{\exp\left(-\frac{z}{2}\right)z^{\eta}} = 1.$$

This altogether enables us to rewrite the $\lim_{|\beta_i|\to\infty}\widetilde{pen}'_{a,b,\sigma}(|\beta_i|)$ as

$$\lim_{|\beta_i|\to\infty}\frac{\sqrt{2}(a+\frac{3}{2})}{\sigma}\frac{\exp\left(-\frac{\beta_i^2}{4\sigma^2}\right)\left(\frac{\beta_i^2}{2\sigma^2}\right)^{-a-\frac{b}{2}-\frac{7}{4}}}{\exp\left(-\frac{\beta_i^2}{4\sigma^2}\right)\left(\frac{\beta_i^2}{2\sigma^2}\right)^{-a-\frac{b}{2}-\frac{5}{4}}} = \lim_{|\beta_i|\to\infty}\frac{2a+3}{|\beta_i|},$$

which was to be demonstrated.

Remark 3.6.1. The asymptotic expansion of the Whittaker function is useful in determining the asymptotic tail behavior of the prior distribution $\widetilde{\pi}_{a,b,\sigma}(\beta_i)$. From (3.6.31) and (4.6.12) it follows that $\widetilde{\pi}_{a,b,\sigma}(\beta_i) = \mathscr{O}\left[\left(\frac{\beta_i^2}{2\sigma^2}\right)^{-a-\frac{5}{4}}\right]$. The tail behavior is therefore unaffected by b, finding noted previously by Maruyama and George (2011) in the g-prior context. As a controls the heaviness of the tails, with lighter tails for large values a, it is intuitive that the bias term in Theorem 4.6.2 diminishes faster for smaller a.

Remark 3.6.2. Similar expression for the bias term implied by the NEG prior was shown previously by Griffin and Brown (2012). In that case $\widetilde{pen}'_{a,0,\sigma}(|\beta_i|) = \frac{(2a+3)}{\sigma}\frac{D_{-2(a+2)}\left(\frac{|\beta_i|}{\sigma}\right)}{D_{-2\left(a+\frac{3}{2}\right)}\left(\frac{|\beta_i|}{\sigma}\right)}$,

which follows from the relationship between Whittaker and parabolic cylinder function and the fact that $W_{\eta,\psi} = W_{\eta,-\psi}$. Since the asymptotic behavior in Theorem 4.6.2 is independent of b, it applies to the NEG prior as a special case.

Margining out parameter v_1 complicates the maximization with respect to β as the logarithm of the prior distribution (3.6.31) does not yield a tractable closed form. Instead, we proceed by estimating the parameter v_1 together with the remaining parameters. The E-step remains unchanged, just with the value v_1 implicit in the computation of (3.3.11) replaced by the current estimate at the k-th iteration $v_1^{(k)}$. The M-step involves one additional computation for finding the value $v^{(k+1)}$. Given the estimates $\beta^{(k+1)}, \sigma^{(k+1)}$, we can find $v_1^{(k+1)}$ as

$$\text{argmax}_{v_1} \Big\{ -\frac{||P^{\star 1/2}\beta||^2}{2\sigma^{(k+1)}} \frac{1}{v_1} + (b - 1/2 \sum_{i=1}^{p} p_i^{\star}) \log(v_1) - (a+b+2)\log(1+v_1) \Big\},$$

where $P^{\star} = \text{diag}\{p_1^{\star}, \ldots, p_p^{\star}\}$. This can be numerically maximized by fast routine methods.

3.7 Structured Prior Information Forms for $\pi(\gamma \mid \theta)$

The beta-binomial prior based on the Bernoulli form (3.2.5) for $\pi(\gamma \mid \theta)$ is suitable for modeling exchangeable variable inclusion probabilities. However, sometimes a priori structural information indicates that certain combinations of variables are more likely to be included together. For example, in the context of genomics, scientific studies have indicated that certain groups of functionally related genes form network topology structures called pathways. In such cases, prior forms more structured than than the Bernoulli can be used to transmit such information. In this section, we consider two such forms which have been recently proposed for stochastic search Bayesian variable selection methodology. As will be seen, these forms are incorporated naturally into the EMVS approach.

3.7.1 The Independent Logistic Regression Prior

The first structured prior form we consider for $\pi(\gamma \mid \theta)$ is the independent logistic regression prior,

$$\pi(\gamma \mid \theta) = \prod_{i=1}^{p} \left(\frac{\exp(Z_i'\theta)}{1+\exp(Z_i'\theta)} \right)^{\gamma_i} \left(\frac{1}{1+\exp(Z_i'\theta)} \right)^{1-\gamma_i}, \tag{3.7.36}$$

a product of independent logistic regression function. A special case of this prior was proposed by Stingo et al. (2010) to incorporate external biological information in a genetic context. In (3.7.36), Z_i is a $q \times 1$ vector of covariates which may influence the model inclusion probability of x_i, and θ is a $q \times 1$ vector of regression coefficients. Letting $Z = [Z^1, \ldots, Z^q]$ be the $p \times q$ matrix whose i^{th} row is equal to Z_i', θ_j is the weight assigned to Z^j, the jth column

of Z. As will be illustrated in Section 3.7.3 below, the columns of Z can conveniently be used to represent potential variable inclusion groupings by using dummy variables to represent potential inclusion. With the addition of a prior $\pi(\theta)$, posterior estimates of θ can yield additional information about the relative influence of the Z_j^{th} grouping.

The choice of a prior for θ is motivated by observing that when the Z_i identify nonoverlapping groupings, (3.7.36) can be reparameterized to be an equally weighted mixture of Bernoulli forms. Indeed, when all the predictors are designated to belong to a single group, i.e. $Z_i \equiv 1$, the prior (3.7.36) simplifies to the exchangeable Bernoulli form (3.2.5) with the success probability $\theta^* = \exp(\theta)/[1+\exp(\theta)]$. Thus the natural choice of the beta distribution for θ^* in the Bernoulli case, translates to

$$\pi(\theta) = \frac{1}{B(a,b)}\left[\frac{\exp(\theta)}{1+\exp(\theta)}\right]^a\left[\frac{1}{1+\exp(\theta)}\right]^b, \tag{3.7.37}$$

which we will refer to as the "logistic-beta prior" on θ. For the general case where θ is $q\times 1$, we generalize this to the multivariate conjugate form

$$\pi(\theta) = \frac{1}{B(a,b)}\left[\frac{\exp(1'\theta)}{1+\exp(1'\theta)}\right]^a\left[\frac{1}{1+\exp(1'\theta)}\right]^b. \tag{3.7.38}$$

The EMVS algorithm under the form (3.7.36) with the prior (3.7.38), is then obtained by replacing Q_2 in (3.3.7) with

$$Q_2^{LR}\left(\theta|\beta^{(k)},\theta^{(k)},\sigma^{(k)}\right) = \sum_{i=1}^{p}\left\{Z_i'\theta\,\mathsf{E}_{\gamma|\cdot}\gamma_i - \log[1+\exp(Z_i'\theta)]\right\} + \{a1'\theta - (a+b)\log[1+\exp(1'\theta)]\}$$

Using the fact that

$$E_{\gamma|\cdot}\gamma_i = \mathsf{P}(\gamma_i = 1\,|\,\theta^{(k)}) = \frac{\exp(Z_i'\theta^{(k)})}{1+\exp(Z_i'\theta^{(k)})}, \tag{3.7.39}$$

maximization of Q_2^{LR} by routine methods can be used to update $\theta^{(k+1)}$.

3.7.2 The Markov Random Field Prior

The second structured prior form we consider for $\pi(\gamma|\,\theta)$ is the Markov random field (MRF) prior proposed by Li and Zhang (2010) to model apriori genetic network information. Representing such information by an undirected graph where predictors x_i and x_j are allowed to interact if and only if i and j are connected by an edge within the edge set $\mathscr{E} = \{(i,j) : 1 \le i \ne j \le p\}$, they proposed the MRF prior

$$\pi(\gamma|\,\theta) = \exp\left[\theta_1'\gamma + \gamma'\theta_2\gamma - \psi(\theta_1,\theta_2)\right], \tag{3.7.40}$$

where $\theta_1 = (\theta_1, \ldots, \theta_p)'$ is a vector of sparsity parameters, $\theta_2 = \left(\theta_{ij}\right)_{i,j=1}^{p}$ is a symmetric matrix of real numbers with $\theta_{ij} = 0 \Leftrightarrow (i,j) \neq \mathscr{E}$, and $\theta = (\theta_1, \theta_2)$. The matrix θ_2 regulates the smoothness of the distribution (3.7.40) by controlling the inclusion probability of a variable based on the selection status of its neighbors. If all the genes are disconnected, so that $\theta_2 = 0$, then the prior (3.7.40) reduces to an independent product of Bernoulli distributions with parameters $p_i = \exp(\theta_i)/[1+\exp(\theta_i)]$. The normalizing constant $\psi(\theta_1, \theta_2)$, known as the partition function, is typically intractable due to the many combinatorial possibilities when summing over all 2^p model configurations.

Letting $\gamma_{\setminus i} = \{\gamma_j : j \neq i\}$ denote the subvector containing all but the i^{th} inclusion indicator, the distribution (3.7.40) implies a simple form for the conditional distributions

$$\pi(\gamma_i \,|\, \gamma_{\setminus i}) = \frac{\exp(\theta_i + \sum_{j \neq i} \theta_{ij}\gamma_j)}{1+\exp(\theta_i + \sum_{j \neq i} \theta_{ij}\gamma_j)}, \tag{3.7.41}$$

which enables Gibbs sampling algorithms (Li and Zhang, 2010) for stochastic search. As a fast practical alternative to such stochastic search, we show how this MRF prior can be incorporated into EMVS to handle challenging high-dimensional problems.

To implement the EMVS algorithm under the MRF prior (3.7.40), the key calculation for the E-step is the evaluation of $\mathsf{E}_{\gamma|\cdot}\gamma_i = \mathsf{P}(\gamma_i = 1 \,|\, \beta^{(k)}, \theta^{(k)}, \sigma^{(k)}) = p_i^*$ in (3.3.11). Because this evaluation is complicated by the dependence among the components in γ under the MRF prior, we approximate it as follows. To begin with, note that the $\mathsf{P}(\gamma_i = 1 \,|\, \beta^{(k)}, \theta^{(k)}, \sigma^{(k)})$ values here arise as marginal means under the joint conditional distribution

$$\begin{aligned}\pi(\gamma \,|\, \beta^{(k)}, \theta^{(k)}, \sigma^{(k)}) \propto \exp\Big[\Big(\frac{1}{2}\log(v_0/v_1)\,1' - \frac{v_0 - v_1}{2\sigma^{(k)2} v_1 v_0}\beta^{(k)\prime}\,\mathrm{diag}\{\beta_i^{(k)}\}_{i=1}^{p} \\ + \theta_1^{(k')}\Big)\gamma + \gamma'\theta_2^{(k)}\gamma\Big].\end{aligned} \tag{3.7.42}$$

The first two terms in the exponent follow directly from the prior distribution $\pi(\beta \,|\, \sigma, \gamma) = \mathrm{N}_p(0, D_{\sigma,\gamma})$, rewriting the determinant of the matrix

$$D_{\sigma,\gamma}^{-1/2} = \frac{\mathrm{diag}\left\{\gamma_i/\sqrt{v_1} + (1-\gamma_i)/\sqrt{v_0}\right\}_{i=1}^{p}}{\sigma}$$

as

$$\begin{aligned}|D_{\sigma,\gamma}|^{-1/2} &= \exp\left[-p\log\sigma - \frac{1}{2}\sum_{i=1}^{p}\left(\gamma_i \log v_1 + (1-\gamma_i)\log v_0\right)\right] \\ &= \exp\left(-p\log\sigma + \frac{1}{2}\log(v_0/v_1)1'\gamma - \frac{p}{2}\log v_0\right).\end{aligned}$$

The conditional distribution in (3.7.42) can be regarded as an MRF distribution with adjusted parameters $\theta_1^\star = (\theta_1^\star, \ldots, \theta_p^\star)'$ and $\theta_2^\star = \theta_2$, where $\theta_i^\star = \frac{1}{2}\log(v_0/v_1) - \frac{v_0 - v_1}{2\sigma^{(k)2} v_1 v_0}\beta_i^{(k)2} + \theta_i$.

Because the partition function $\psi(\theta_1,\theta_2)$ is a normalizing factor in the exponential family, it follows that $\mathsf{E}(\gamma|\beta^{(k)},\theta^{(k)},\sigma^{(k)}) = \frac{\partial\psi(\theta_1,\theta_2)}{\partial\theta_1}|_{\theta_1=\theta_1^\star}$. Although this vector is not analytically tractable, a useful approximation can be obtained using mean field methods (Wainwright and Jordan, 2008).

Recall that mean field approximation refers to a class of variational methods that approximate a distribution on a graph, here $\pi(\gamma|\beta^{(k)},\theta^{(k)},\sigma^{(k)})$, with a simpler distribution for which it is feasible to do exact inference. Here we make use of the naive mean field method, which restricts to a class of tractable approximating distributions assuming completely disconnected graphs. In other words, we assume approximating distributions of the form $q(\gamma|\mu) = \prod_i \mu_i^{\gamma_i}(1-\mu_i)^{1-\gamma_i}$, where $\mu = (\mu_1,\ldots,\mu_p)' \in [0,1]^p$ denotes the vector of mean parameters.

It can be shown (Wainwright and Jordan, 2008) that the parameter vector $\widehat{\mu}$, for which $q(\gamma|\widehat{\mu})$ best approximates $\pi(\gamma|\beta^{(k)},\theta^{(k)},\sigma^{(k)})$ within the class of tractable functions, where the quality of the approximation is measured by the KL divergence, satisfies the set of equations

$$\widehat{\mu}_i = \frac{\exp(\theta_i^\star + \sum_{j\neq i}\theta_{ij}\widehat{\mu}_j)}{1+\exp(\theta_i^\star + \sum_{j\neq i}\theta_{ij}\widehat{\mu}_j)}, \quad 1\leq i\leq p. \tag{3.7.43}$$

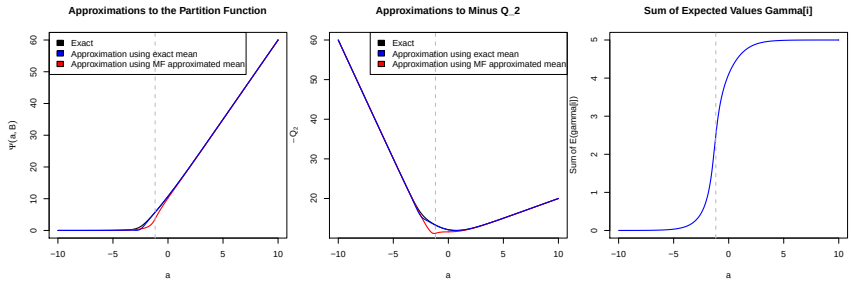
Each of the equations (3.7.43) can be regarded as an averaged version of the expression in (3.7.41). The solution can be found by iteratively updating (3.7.43), which can be seen as a type of coordinate ascent algorithm. Each value $\widehat{\mu}_i$ then provides the mean field approximation to p_i^* in (3.3.11).

The hyperparameters of the MRF distribution have until now been assumed to be fixed. In order to enhance the adaptability of the procedure we may consider the sparsity parameters θ_1 to be unknown (arising from a prior distribution $\pi(\theta_1)$). In what follows, we restrict attention to vectors of type $\theta_1 = \theta(1,\ldots,1)'$. A natural candidate prior distribution $\pi(\theta)$, which corresponds to the beta-binomial prior in case $\theta_2 = 0$, is the logistic-beta distribution (3.7.37). The M-step of the algorithm then requires the additional step of updating the parameter θ by finding the maximum of the function

$$Q_2^{MRF}(\theta\,|\,\beta^{(k)},\theta^{(k)},\sigma^{(k)}) = \theta\left(\sum_{i=1}^p p_i^\star + a\right) + \psi(\theta,\theta_2) - (a+b)\log[1+\exp(\theta)].$$

Maximizing Q_2^{MRF} w.r.t. θ is complicated by the unavailability of the partition function in a closed form. The mean field theory can be again used to obtain an approximate solution. According to Wainwright and Jordan (2008), the mean field approximation to the partition function for the MRF model can be expressed as

$$\psi(\theta,\theta_2) \approx \theta\sum_{i=1}^p \mu_i + \mu'\theta_2\mu - \psi^\star(\mu), \tag{3.7.44}$$



(a) $\psi(\theta, \theta_2)$ (b) $Q_2^{MRF}(\cdot)$ (c) $\sum_{i=1}^{p} \mathsf{E}_{\theta,\theta_2} \gamma_i$

Figure 3.6: Plots of phrase transition function $f(\theta, \theta_2)$ as well as approximated and true partition functions $\psi(\theta, \theta_2)$, $Q_2^{MRF}(\theta, \theta_2)$ in relation to parameter θ

where $\mu_i = \mathsf{E}_{\theta,\theta_2} \gamma_i$ and $\psi^\star$ denotes the conjugate dual function to ψ, which has an explicit form for the approximating product distributions, i.e.

$$\psi^\star(\mu) = \sum_{i=1}^{p} [\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i)].$$

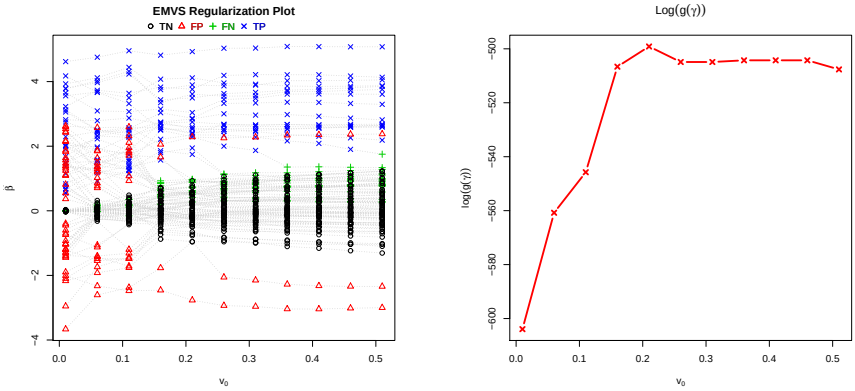
The mean values $\mu_i = \mathsf{E}_{\theta,\theta_2} \gamma_i$ for each specific value θ can be obtained from (3.7.43).

It is widely known that the MRF prior is susceptible to phase transitions, where small increments in θ may lead to massive increments in the size of the selected model. Stingo and Vannucci (2011) suggest putting prior mass on θ values in a neighborhood of the transition point to improve mixing of the MCMC sampler. In our EM context, the transition point θ_{trans} can be regarded as the value at which $f(\theta, \theta_2) = \sum_{i=1}^{p} \mathsf{E}_{\theta,\theta_2} \gamma_i$ exhibits rapid growth or even a jump. There may be multiple transition points in situations when the matrix θ_2 has complicated structural patterns.

In order to visually assess the quality of the approximation to the partition function, Figure 3.6(a) plots the approximated and true function $\psi(\theta, \theta_2)$ for varying θ with θ_2 a 5×5 symmetric zero diagonal matrix with 5 randomly placed nonzero entries in the upper triangle. We consider two approximations, where either true means or mean field approximated means are plugged in the equation (3.7.44). The values of $p_i^\star$ are set to one and $a = b = 1$.

We observe that the approximation (3.7.44) with imputed approximated mean values loses the convexity property (Figure 3.6(a)). Moreover, the approximation is impaired in the closed neighborhood of the transition point $\theta_{trans} = -1.17$, which was detected from the plot of the function $f(\theta, \theta_2)$ in Figure 3.6(c). The plots of the true and approximated $Q_2^{MRF}(\cdot)$ function in Figure 3.6(b) suggest that the update $\theta^{(k+1)}$ is likely to be estimated at the transition point, if we use the mean field approximation of the partition function.

As for the deterministic annealing versions of EMVS under structured priors, the tempered E-step for the logistic regression prior remains the same as for the beta-binomial case.



(a) EMVS Regularization Plot (b) Submodel Evaluations

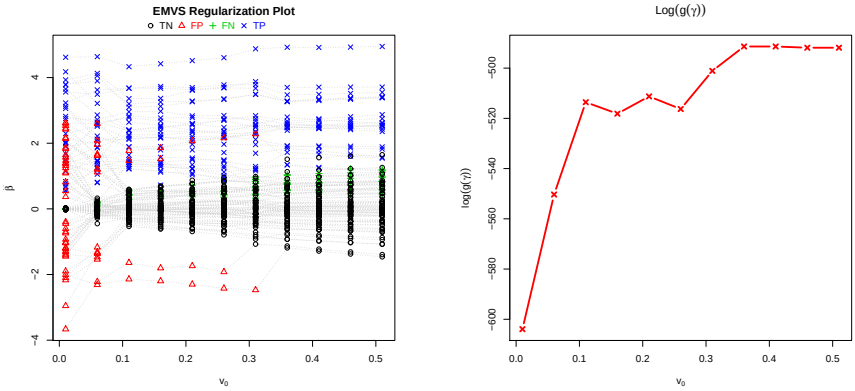
Figure 3.7: (a) plot of estimated regression coefficients for varying choices of v_0, TN/FP/FN/TP stand for true negatives/false positives/false negatives/true positives; (b) $\log g_0(\gamma)$ for models selected with $\mathrm{P}(\gamma_i = 1 \mid \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}) > 0.5$.

Under the MRF prior, the E-step is performed with parameters $\theta_1^\star$ and $\theta_2^\star$ multiplied by an inverse temperature parameter.

3.7.3 Simulated Example for Structured Priors

To illustrate the potential of the structured variable inclusion priors from Sections 3.7.1 and 3.7.2, we compare them with the benchmark beta-binomial prior on simulated data with substantial grouped structure. For this purpose, we simulated $Y \sim \mathrm{N}_n(X\beta, \sigma^2 \mathrm{I}_n)$ with $n = 100$ and $\sigma^2 = 5$. The $n \times p$ predictor matrix X consisted of $p = 99$ normally distributed predictors generated as three equicorrelated groups $\{x_1, \ldots, x_{33}\}$, $\{x_{34}, \ldots, x_{66}\}$ and $\{x_{67}, \ldots, x_{99}\}$ with pairwise correlations of 0.8 within each group and zero correlations between the groups. For the $p \times 1$ regression vector we set the components $\beta_i = 2 \times \mathbb{I}_{[1;33]}(i)$ so that only the first group of predictors is actually explaining the variability of the response Y.

For this setting, we considered the following three forms for the pair $\pi(\gamma \mid \theta)$ and $\pi(\theta)$ to reflect varies degrees of prior knowledge: (a) The Bernoulli form (3.2.5) coupled with the uniform prior on θ which yields the beta-binomial prior with $a = b = 1$ on γ. This unstructured exchangeable choice ignores the potential grouping information; (b) The independent logistic regressions form (3.7.36) with the three grouping vector choices Z^1, Z^2, Z^3 where the ith component of Z^j is given by $z_{ij} = \mathbb{I}_{[33\,(j-1)+1;\,33\,j]}(i)$ for $1 \leq i \leq p$. To this we add the logistic-beta prior (3.7.38) with $a = b = 1$ on $\theta = (\theta_1, \theta_2, \theta_3)'$. This prior conveys the information that there are three possible groupings, of which only the first is correct for the simulated data here; (c) The MRF prior (3.7.40) with sparsity parameter $\theta_1 = \theta 1$, where θ



(a) EMVS Regularization Plot

(b) Submodel Evaluations

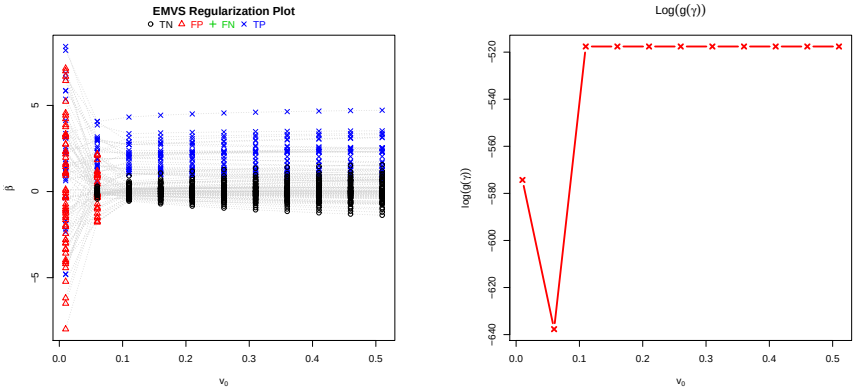
Figure 3.8: (a) plot of estimated regression coefficients for varying choices of v_0, TN/FP/FN/TP stand for true negatives/false positives/false negatives/true positives; (b) $\log g_0(\gamma)$ for models selected with $\mathsf{P}(\gamma_i = 1 \mid \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}) > 0.5$.

is assigned prior (3.7.37) with $a = b = 1$, and with fixed $\theta_2 = (1_{33\times 33} - \mathrm{I}_{33}) \otimes \mathrm{I}_3$, where $1_{33\times 33}$ is a 33×33 matrix of ones and $\otimes$ denotes the Kronecker matrix product. This prior also conveys the information that there are three possible groupings, where all within-group predictors are neighbors on an undirected graph. Note that (b) and (c) would be equivalent to (a) when $Z^1 = 1_p, Z^2 = 0, Z^3 = 0$ and $\theta_2 = 0$.

To carry out the EMVS search and regularization algorithm with each of these three prior choices, we considered the grid of v_0 values $V = \{0.01 + k \times 0.05 : k = 0, \ldots, 10\}$. Rather than setting v_1 to a large fixed value, we applied the prior $\pi(v_1)$ in (3.6.29) with $a_{v_1} = 0.5$ and $b_{v_1} = 250$ (under which the prior mode $\widehat{v_1} = \frac{b_{v_1}}{2+a_{v_1}}$ equals 100). The starting values $\beta^{(0)}$ were selected according to (3.5.26) with $v_0 = 1$ and $v_1 = 1\,000$, $\sigma^{(0)} = 1$, $\theta^{(0)} = 1_3$ for the logistic prior and $\theta^{(0)} = \theta_{trans}$ for the MRF prior. To evaluate the solution path of models $\{\widehat{\gamma}_{v_0} : v_0 \in V\}$ generated under each prior, we used the same g_0 function from (3.4.23) corresponding to the posterior under $v_0 = 0$ and $v_1 = 1\,000$ obtained with the uniform beta-binomial model prior in order to allow for a fair comparison of every model.

For EMVS under the beta-binomial prior (a) with no structural information, we obtain the regularization plot in Figure 3.7. The best visited model (corresponding to $v_0 = 0.21$) identified 21 true predictors together with 12 false negatives and 2 false positives.

For EMVS under the independent logistic regression prior (b) which conveyed structural information in an additive matter, we obtain the regularization plot in Figure 3.8. Performing better than the beta binomial prior, the best visited model here (corresponding to $v_0 = 0.36$) contains 22 correctly identified predictors together with 11 false negatives and zero false pos-



(a) EMVS Regularization Plot (b) Submodel Evaluations

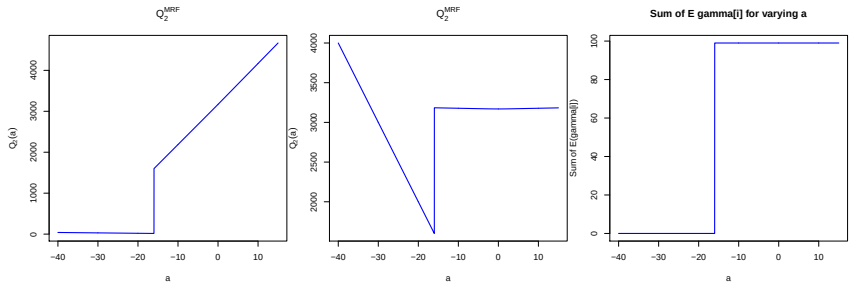
Figure 3.9: (a) plot of estimated regression coefficients for varying choices of v_0, TN/FP/FN/TP stand for true negatives/false positives/false negatives/true positives; (b) $\log g_0(\gamma)$ for models selected with $\mathsf{P}(\gamma_i = 1 \mid \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}) > 0.5$.

itives. The posterior estimates $\widehat{\theta} = (0.93, -3.35, -3.44)'$, further indicate that the posterior adaptively increased the inclusion probabilities for predictors within the first group, the single correct grouping for our data.

Finally, for EMVS under the MRF prior (c) with $\theta = \theta_{trans}$, which assumes that all predictive covariates are interconnected on an undirected graph, we obtain the regularization plot in Figure 3.9. The best found model correctly identifies all the 33 predictors with zero false discoveries and zero false non-discoveries.

In order to understand the phase shift behavior, we plot the $f(\theta, \theta_2)$ function for varying values of θ (Figure 3.10(c)). We observe a jump at the transition point at $\theta_{trans} = -16.03$. Next, we plot the approximated Q_2^{MRF} function considering $p_i^\star = 0, (i = 1, \ldots, p)$ (Figure 3.10(b)) and $p_i^\star = 1, (i = 1, \ldots, p)$ (Figure 3.10(a)) for $a = b = 1$. We observe that the minimum is attained in both cases at the value of the transition point. This behavior is seen irrespective of the choices of a and b. The sparsity parameter θ is therefore likely to be estimated directly at the transition point, which rather resembles applying the procedure for θ fixed to this value.

It is worth noting that in densely connected networks that are sparse for predictive variables, we have observed a tendency for the MRF prior to increase the number of false positives. In such scenarios, the logistic prior can better negotiate the within group sparsity and improve variable selection over the independent beta-binomial prior.



(a) Q_2^{MRF} (b) Q_2^{MRF} (c) $f(\theta, \theta_2)$

Figure 3.10: Approximated Q_2 function together with the phase transition function for the simulated data example

3.8

Stochastic Dual Coordinate Ascent for EMVS

The efficiency of the EMVS implementation relies on expeditious updating of the ridge regression solutions, which constitutes the most expensive operation in the EM algorithm. The matrix inversion needed to obtain the closed form solution may be very prohibitive when both n and p are large. Approximative solutions to ridge and other regularized loss minimization problems can be obtained using conjugate gradient descent methods or dual coordinate ascent algorithms at only a fraction of the runtime. Here we describe the stochastic version of the dual coordinate ascent algorithm (SDCA) of Shalev-Shwartz and Zhang (2013), which has been shown to possess strong theoretical guarantees. In conjunction with the fast E-step, the SDCA greatly enhances the rapidity of the EMVS procedure.

We begin by describing the generic optimization problem associated with regularized linear predictor minimizers of a general loss function. Denote the original data by y, a $(n \times 1)$ response vector, and $X^* = [x_1^*, \ldots, x_p^*]$, a $(n \times p)$ regression matrix, where $||x_j^*|| \equiv 1$. The constrained loss function to be minimized is then

$$P^*(\beta^*) = \left[\sum_{i=1}^{n} \phi_i(x_i^{*'}\beta^*) + ||D^{1/2}\beta^*||^2 \right], \tag{3.8.45}$$

where D is a diagonal matrix of individual penalty parameters for each regression coefficient. Throughout the section we refer to the objective function in (3.8.45) as the generalized ridge regularized loss, where each coordinate in the regression vector is penalized differentially. In case of ridge regularized linear regression, which interests us in the context of EMVS, the loss function takes the form $\phi(a) = (a - y_i)^2$. The optimizer of the generalized ridge regression problem can be obtained from a solution to a classical ridge regression after reweighing the

columns of the regression matrix. Let $\beta = D^{1/2}\beta^*$ and $X = X^*D^{-1/2}$. Then the minimizer $\widehat{\beta^*}$ of $P^*(\beta^*)$ corresponds to the minimizer $\widehat{\beta}$ of the ridge regularized loss function with unit penalty

$$P(\beta) = \left[\sum_{i=1}^{n} \phi_i(x_i'\beta) + ||\beta||^2\right]. \tag{3.8.46}$$

We now describe the dual formulation of the optimization problem associated with (3.8.46). We begin with rewriting the objective function in terms of $\eta_i = \phi_i(x_i'\beta)$ and introducing the Lagrange multipliers α_i for every one of the corresponding constraints $\eta_i - \phi_i(x_i'\beta) = 0$. Augmenting the objective function (3.8.46) by the weighted sum of the constraint functions we obtain following Lagrangian

$$L(\beta, \eta, \alpha) = \left[\sum_{i=1}^{n} \eta_i^2 + ||\beta||^2 + \sum_{i=1}^{n} \alpha_i[\phi(x_i'\beta) - \eta_i]\right] \tag{3.8.47}$$

and the associated dual Lagrange function

$$D(\alpha) = \inf_{\beta,\eta} L(\beta, \eta, \alpha).$$

Differentiating the Lagrangian (3.8.47) in β and η, we obtain conditions

$$\beta(\alpha) = \frac{1}{2}\sum_{i=1}^{n} \alpha_i x_i \quad \text{and} \quad \eta_i(\alpha) = \frac{\alpha_i}{2},$$

which after substitution in (3.8.47) give the dual Lagrangian

$$D(\alpha) = \left[\sum_{i=1}^{n} -\phi_i^*(-\alpha_i) + \left\|\frac{1}{2}\sum_{i=1}^{n} \alpha_i x_i\right\|^2\right] \tag{3.8.48}$$

where $\phi_i^*(u) = \max_z(zu - \phi_i(z))$ is the convex conjugate of $\phi_i(\cdot)$. Let $\widehat{\alpha}$ denote a maximizer of $D(\alpha)$. Then it is known that $\beta(\widehat{\alpha}) = \widehat{\beta}$ and $P(\widehat{\beta}) = D(\widehat{\alpha})$. It also holds that $P(\beta) \geq D(\alpha)$ for all β and α, which implies that the duality gap $P[\beta(\alpha)] - D(\alpha)$ constitutes an upper bound of the sub-optimality $P[\beta(\alpha)] - P(\alpha)$.

In the following, we restrict the attention to the squared loss in the linear regression setting, where the dual function takes the form

$$D(\alpha) = \left[\sum_{i=1}^{n} y_i\alpha_i - \frac{1}{4}\sum_{i=1}^{n} \alpha_i^2 + \left\|\frac{1}{2}\sum_{i=1}^{n} \alpha_i x_i\right\|^2\right]. \tag{3.8.49}$$

A nearly optimal value $\widehat{\alpha}$, and hence nearly optimal $\widehat{\beta}$, can be found by a applying a coordinate descent algorithm (CDA) on the dual Lagrangian function. We describe the stochastic version

SDCA Procedure

(1) Initialize $\beta^{(0)} = \beta(0)$

(2) Iterate for $t = 1, 2, \ldots, T$

  (a) Select randomly i from $\{1, \ldots, n\}$

  (b) Set $\Delta\alpha_i = \frac{2(y_i - x_i'\beta^{(t-1)}) - \alpha_i^{t-1}}{1+||x_i||^2}$

  (c) $\alpha^{(t)} \leftarrow \alpha^{(t-1)} + \Delta\alpha_i e_i$

  (d) $\beta^{(t)} \leftarrow \beta^{(t-1)} + \frac{1}{2}\Delta\alpha_i x_i$

(3) Output $\bar{\alpha} = \frac{1}{T-T0}\sum_{t=T0}^{T}\alpha^{(t)}$

(4) Output $\bar{\beta} = \frac{1}{T-T0}\sum_{t=T0}^{T}\beta^{(t)}$

Table 3.2: Steps of the SDCA algorithm.

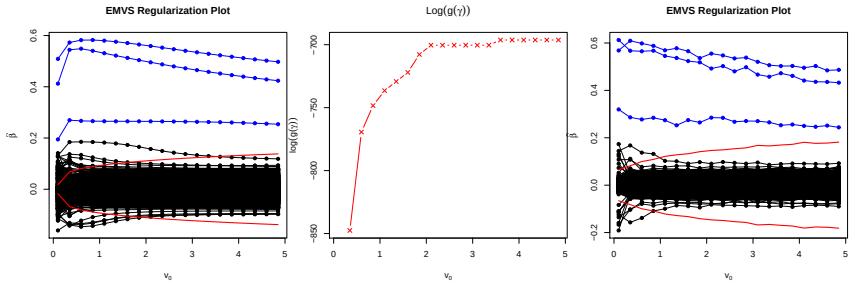
(SDCA), where at each iteration the coordinate to be updated is chosen at random. The steps of the algorithm are summarized in the Table 3.8.

For the γ-smooth loss functions (differentiable and with derivative γ-Lipschitz), Shalev-Shwartz and Zhang (2013) show that SDCA requires at least $T = 2(n + n\gamma/2)\log(1/\varepsilon)$ iterations in order the to have an expected duality gap $\mathsf{E}[P(\bar{\beta}) - D(\bar{\alpha})] \leq \varepsilon$ for $\bar{\beta}$ and $\bar{\alpha}$ averaged over last $T0 = T/2$ iterations. Since the squared loss is 2-smooth, it suffices to perform at least $T = 4n\log(1/\varepsilon)$ iterations.

3.8.1 Timing Comparisons

We consider simulated datasets on $p = 1000$ explanatory variables, where only the first three are predictive with a corresponding regression vector $\beta = (2, 3, 4, 0, \ldots, 0)'$. We generated three datasets with $n = 100, 500, 2000$ and compared the computational time required to obtain a generalized ridge regression solution using (a) classical closed form expression inverting $p \times p$ matrix, (b) Woodbury-Sherman matrix formula inverting $n \times n$ matrix, (c) SDCA implementation in R, (d) SDCA implementation in C. The regression matrices are generated with rows drawn independently from $\mathrm{N}_p(0, \Sigma)$, where $\Sigma = \left(0.6^{|i-j|}\right)_{i,j=1}^{p}$. The predictor matrices were further rescaled so that $||x_i||^2 \leq 1$, which is one of the requirements for the theoretical guarantees to hold. The response vector was for each of the three sample sizes created according to the generating model $\mathrm{N}_n(X\beta, \mathrm{I}_n)$ and further normalized so that $||y||^2 = 1$.

The vector of penalty coefficients was generated through random sampling from Gamma distribution with shape 1 and scale 0.5. Table 1 reports on the computation runtime in seconds obtained on a $3GHz$ server as well as distances between the exact and approximate solutions. A stopping rule $T = 4n\log(1/\varepsilon)$ was implemented to obtain at most $\varepsilon = 0.1$ expected duality gap.



(a) (b) (c)

Figure 3.11: (a) Exact EMVS regularization plot, (b) model exploration based on the exact regularization plot, (c) approximated EMVS regularization plot

$p = 1000$	**CR**	**WS**	**SDCA (R)**	**SDCA (C)**	$\|\|\tilde{\beta} - \beta_{ridge}\|\|^2$
$n = 100$	2.44	0.38	0.17	0.02	0.004
$n = 500$	4.41	3.69	1.27	0.16	0.005
$n = 2000$	9.93	10.24	5.28	0.66	0.002

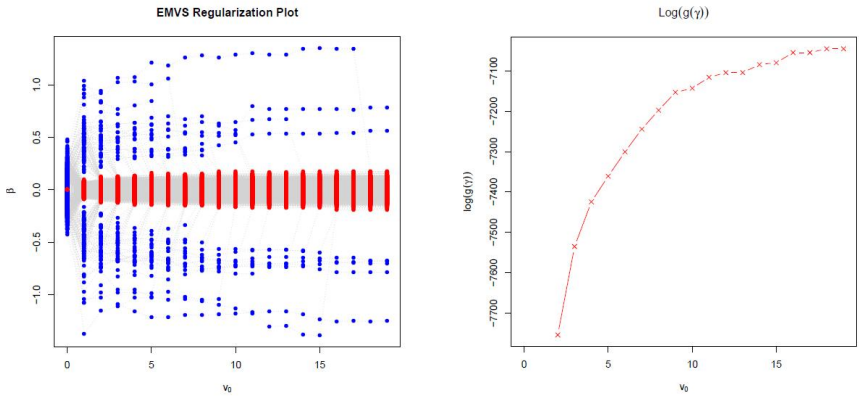
Table 3.3: Computational time in seconds of the generalized ridge regression solutions, CR: classical ridge, WS: Woodbury-Shermann

We take the second dataset ($n = 500$) and apply the EMVS procedure assuming $v_1 = 10$ and $v_0 \in \{0.1 + k \times 0.25; 0 \leq k \leq 20\}$. We obtain the EMVS regularization plot displaying the evolution of the posterior modal estimates as the spike variance increases (Figure 3.11(a)). The log-posterior model probabilities of subsets obtained after screening out coefficients that are small in magnitude (outside the threshold boundary depicted in red) are plotted in Figure 5.8. The approximate regularization plot obtained using the SDCA procedure ($T = 4n\log(10)$) in the M-step is depicted in Figure 3.11(c).

Under the convergence criterion $\max|\beta^{(k)} - \beta^{(k-1)}| < 0.05$, the exact evaluation of the whole regularization plot requires 80 iterations taking 295 seconds using the Woodbury-Sherman updates in the R-implementation. The approximation requires 94 ridge regression updates along the regularization path taking altogether 119 seconds in R-implementation and merely 15 seconds in the C-implementation.

3.9 Finding DNA Regulatory Motifs Using EMVS

In this section, we apply the EMVS procedure to detect DNA nucleotide sequences that act as binding sites for transcription factors and thereby coordinate expression of genes in whose regulatory region they appear. Transcription factors are proteins which are known to either inhibit



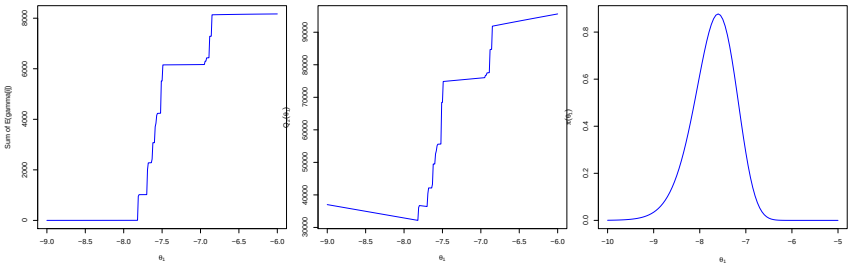
(a) EMVS Regularization Plot (b) Submodel Evaluations

Figure 3.12: (a) plot of estimated regression coefficients for varying choices of v_0, estimates for variables with conditional posterior inclusion probability $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma})$ above (below) 0.5 depicted in blue (red); (b) logarithm of $g(\gamma)$ for models with selected variables with $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}) > 0.5$.

or enhance transcription of genes by binding to their promoter region sequences. Spellman et al. (1998) conducted a series of yeast experiments to identify transcription factor binding sites whose occurrence in the genome drives the periodic expression pattern associated with the cell cycle. This data set has been analyzed in literature by multiple authors including Li and Zhang (2010), Bussemaker et al. (2001) or Tadesse et al. (2004).

The data consists of gene expression measurements collected longitudinally at 18 time points spanning over two cell cycles. Following the approach of Li and Zhang (2010), we use first principal component scores to compress the gene expression over time for each of the 1 568 genes. The response vector Y then consists of $n = 1\,568$ continuous measurements of the summarized expression levels. About half of the genes were previously recognized as associated with the cell cycle, whereas the other half does not exhibit any differential expression across time and is included as a reference. Upstream regulatory regions of each gene have been screened for the presence of short regulatory motifs. A motif is considered to be a word of length 7 consisting of letters $\{A, G, T, C\}$, where each word and its reversed complementary sequence represent the same biological motif. The predictor matrix X then consists of numbers of occurrences of each of the $p = 4^7/2 = 8\,192$ motifs in the promoter region of each gene.

The predictors are assumed to cluster based on the similarity in their sequence as can be determined by the Hamming distance (Li and Zhang, 2010). Motifs with a similar content are likely to attract the same transcription factors and thereby influence the gene expression in a similar manner. This phenomenon has been incorporated in the linear model for motif



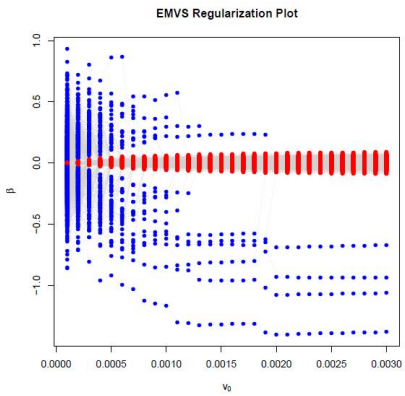
(a) $\sum_{i=1}^{p} \mathsf{E}_{\theta,\theta_2} \gamma_i$ (b) $Q_2^{MRF}(\theta)$ (c) $\pi(\theta)$

Figure 3.13: Approximated $Q_2(\cdot)$ function together with the phase transition function and prior distribution $\pi(\theta)$ for the Spellman data

detection through the MRF prior by Li and Zhang (2010). We similarly regard two related motifs (differing by at most one letter regardless the location of the mismatch) to be two vertices connected by an edge in an undirected graph. The $8\,192 \times 8\,192$ smoothing matrix θ_2 then consists of 144 896 nonzero entries.

We apply the EMVS procedure assuming both exchangeable and structured variable selection indicators under the beta-binomial and MRF priors. In both analyses, we treat the slab variance parameter v_1 as unknown and we consider the prior distribution (3.6.29) with a_{v_1} and b_{v_1} selected so that the mode $\widehat{v}_1 = b_{v_1}/(2+a_{v_1})$ of the prior distribution is 100. We examined the sensitivity of the results to the choice of a_{v_1} and b_{v_1} and found them to be quite robust. The two parameters are seen to influence the number of iterations rather than selected model configurations. We considered $a_{v_1} = 0.5$ and $b_{v_1} = 250$, for which the number of iterations was moderate. (We also considered some deterministic annealing variants with these settings, not reported here, which essentially yielded similar findings). For submodel evaluations, we used $g_0(\gamma)$ with $v_1 = 1\,000$ and a uniform beta distribution on the success probability. In both analyses, we set the vector of starting values for the regression coefficients equal to the ridge regression solution corresponding to the limiting case of deterministic annealing, as given in (3.5.26), with $v_0 = 1$ and $v_1 = 1\,000$.

For the exchangeable variable selection indicators, we consider a grid of values $v_0 \in \{0.001 + k \times 1 : k = 0, \ldots, 20\}$. The regularization diagram together with model evaluation is depicted in Figure 3.12. As v_0 increases, $g_0(\gamma)$ continues to escalate and sparser models are revealed, leaving us with only 7 motifs at $v_0 = 20.001$ (`ACGCGTT, CGCGTTT, GACGCGT, GGACGAT, TTCGCGT, TTTATCG, TTTCGCG`). Other interesting candidates are found, corresponding to more moderate v_0 values. For $v_0 = 9.001$, 18 motifs were screened out (Table 3.4), among which 3 are connected on the graph and several have been previously identified (Li and Zhang, 2010) or experimentally validated (according to Sacharomyces Cerevisae Promoter Database (SCPD) of Zhu and Zhang (1999) available at



(a) EMVS Regularization Plot (b) Submodel Evaluations

Figure 3.14: (a) plot of estimated regression coefficients for varying choices of v_0, estimates for variables with conditional posterior inclusion probability $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma})$ above (below) 0.5 depicted in blue (red); (b) logarithm of $g(\gamma)$ for models with selected variables with $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\beta}, \widehat{\theta}, \widehat{\sigma}) > 0.5$.

`http://cb1.utdallas.edu/SCPD/`).

We proceed to apply the EMVS procedure under the MRF prior. Following Li and Zhang (2010), we set the nonzero elements in θ_2 equal to 0.83. We assume $\theta_1 = \theta(1, \ldots, 1)'$ and specify an appropriate distribution $\pi(\theta)$ according to (3.7.37), which locates the majority of its mass in the phase transition region. The plot of the transition function in Figure 3.13(a) indicates multiple transition points, a consequence of the structure θ_2 which allows for overlapping components. Selecting the hyperparameter values $a_\theta = 5$ and $b_\theta = 10000$ guarantees accumulation of the prior distribution within boundaries $[-9, -6]$ (Figure 3.13(c)). The corresponding $Q_2(\cdot)$ function for $p_i^\star \equiv 1$ is plotted in Figure 3.13(b). In order to better observe the gradual sparsification of the explored models, we consider a more refined grid of smaller values $v_0 \in \{10^{-5} + k \times 10^{-5} : k = 0, \ldots, 30\}$.

The corresponding regularization plot together with the evolution of $g_0(\gamma)$ is displayed in Figure 3.14. Among the explored models, the highest value of $g_0(\gamma)$ was obtained for a model with 4 motifs (`ACGCGTT, CGCGTTT, GACGCGT, TTTCGCG`). Table 3.4 summarizes two other motif sets of dimensions 18 and 7, which have been identified along the regularization path. In comparison with the models of the same size found by the beta-binomial model, we observe that the MRF EMVS biases the search towards models with more interconnected predictors.

The execution time to obtain the modal estimates for a single mixture prior varied depending on the magnitude of v_0. Generally, more iterations were needed for smaller v_0 values, where multimodality appeared to hamper convergence towards a single local mode. The

median number of iterations (for the considered set of v_0 values) needed to achieve convergence under the criterion $\max_{1\leq i\leq p}\{|\beta_i^{(k+1)} - \beta_i^{(k)}|\} < 10^{-4}$ was 12 for the beta-binomial version and 5 for the MRF version of the EMVS procedure. Fewer iterations were needed for EMVS with a fixed v_1 (the median number of iterations was 7 for the beta-binomial model with $v_1 = 1000$). One iteration of the beta-binomial (resp. MRF) model took 107 (resp. 210) seconds using an R implementation on a $3GHz$ linux server. The execution time for the largest v_0 values considered was 21.4 minutes for the beta-binomial model and 17.5 minutes for the MRF model. In sharp contrast the stochastic search MCMC approach of Li and Zhang (2010) took more than 12 hours to obtain marginal inclusion estimates for a single mixture prior with $v_0 = 0$ [personal communication].

18 **Selected Motifs**		7 **Selected Motifs**		
BB	**MRF**	**BB**	**MRF**	**Known**
$GACGCGT^1$	$GACGCGT^1$	$GACGCGT^1$	$GACGCGT^1$	×
$TACGCGT^1$	$TACGCGT^1$		$TACGCGT^1$	×
$TTCGCGT^1$	$TTCGCGT^1$	$TTCGCGT^1$	$TTCGCGT^1$	×
	$TTACGCG^2$			
$TTTCGCG^2$	$TTTCGCG^2$	$TTTCGCG^2$	$TTTCGCG^2$	×
	$TGACGCG^2$			
$TTAGCAG$				
$ACGCGTT$	$ACGCGTT$	$ACGCGTT$	$ACGCGTT$	
$CCGCTTG$	$CCGCTTG$			
$CCGTCCT$	$CCGTCCT$			
$CGCGTTT$	$CGCGTTT$	$CGCGTTT$	$CGCGTTT$	
$CGTCCCT$	$CGTCCCT$			
$CTGATGG$	$CTGATGG$			
$GAATTAT$	$GAATTAT$			
$GACAGGT$				
$GCCATTT$	$GCCATTT$			
	$GCGTTTT$			
$GGACGAT$	$GGACGAT$	$GGACGAT$		×
$GTCCTCT$				
$TACACAG$	$TACACAG$			×
$TTTATCG$	$TTTATCG$	$TTTATCG$	$TTTATCG$	

Table 3.4: Selected motifs by betabinomial (BB) and MRF versions of EMVS for selected v_0 values that along the regularization path lead to selection of 18 and 7 predictors; known or previously identified motifs (Li and Zhang (2010), Zhu and Zhang (1999)) are marked with a cross; motifs that form a subnetwork of connected components are marked with a superscript ([1] Group of known MCB cell cycle regulatory motifs, [2] Group of known SCB cell cycle regulatory motifs)

3.10 Discussion

The main thrust of this work has been to propose EMVS, a practical deterministic approach for posterior model mode discovery under spike-and-slab formulations for Bayesian variable selection in high dimensional regression settings. Through dynamic posterior exploration with a fast EM algorithm, EMVS can be used to find sparse high probability models in complicated settings with structured prior information and a large number of potential predictors, settings where alternative methods such as MCMC stochastic search would, at best, be much slower.

The core ingredients of EMVS are the continuous conjugate spike-and-slab formulation, the regularization scheme and an EM algorithm tailored for non-convex Bayesian maximum a posteriori optimization. As opposed to point mass variable selection priors, a continuous spike distribution serves to absorb smaller unimportant coefficients and to reveal sparser candidate subsets. The gradual sparsification of the explored models for increasing spike variance is captured by the regularization diagram, where each of the discovered subsets is subsequently evaluated by its posterior model probability. For posterior computation, our EM algorithm converges quickly, effectively identifying sets of high-posterior models and regression coefficient estimates. On both real and simulated examples, we have demonstrated that EMVS is capable of identifying promising models, while still providing computational tractability, a crucial feature for high-dimensional model spaces. We have also illustrated the generality of EMVS, how it can accommodate a variety of hierarchical model prior constructions, from exchangeable priors that are uniform over model size to flexible structured priors driven by existing external knowledge.

Extensions of EMVS to frameworks beyond linear regression provide rich new directions for methodological developments. For example, a straightforward probit extension for classification of binary responses can be derived using data augmentation with an additional E-step to obtain expected values of the latent continuous data. Other generalized linear models such as logistic regression and Poisson regression become feasible with the dual coordinate ascent algorithm (Shalev-Shwartz and Zhang, 2013) for approximating the M-step. Further interesting directions will be to consider EMVS for Gaussian graphical model determination or for factor analytic augmentation of multivariate regression models.

Another important avenue for future research will be the development of uncertainty reports to accompany EMVS model selection. Although full posterior inference has been sacrificed for computational feasibility, posterior variability assessments will still be available.

To begin with, conditionally on the posterior, EMVS selection uncertainty could be addressed by considering multiple starting values for the EM algorithm. This might be done locally by reinitializing EMVS over a set of perturbed modal estimates, or more globally over a set of spread out values obtained from a preselected grid or by random sampling. The speed of our EM algorithm would allow for as many starting values as tens to hundreds. In multimodal posterior landscapes without a dominating posterior mode, EMVS model selection

will be more sensitive to such reinitializations, leading to a variety of different modal models. The relative posterior probabilities obtained by $g_0(\gamma)$ in (3.4.23) for such selected models would provide an informative model uncertainty report, and could be used as a basis for model averaging or for the approximation of a median probability model.

For any given EMVS selected mode $\widehat{\gamma}$ one could carry out local MCMC posterior simulations in a neighborhood of $\widehat{\gamma}$ in order to gauge the relative accumulation of posterior probability. The closed form posterior expression $g_0(\gamma)$ would be useful for this simulation. Note that such posterior accumulations would provide a further basis for the comparison of multiple modes obtained through the reinitialization described above.

Finally, the ability of EMVS to quickly find posterior modes in high dimensional settings makes it a potentially powerful complement for other methods. For example, general MCMC simulation in multimodal settings may be substantially enhanced with EMVS selected posterior modes as starting values.

Our software implementation of EMVS was written in R with a prototype version in C as a shared library loadable from R. Both are available from the authors upon request.

CHAPTER 4

INCORPORATING GROUPING IN BAYESIAN VARIABLE SELECTION WITH APPLICATIONS IN GENOMICS

Rockova, V., Lesaffre, E. 2013. **Incorporating Grouping Information in Bayesian Variable Selection with Applications in Genomics**. To appear in *Bayesian Analysis*

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 Bayesian LASSO context 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 EM algorithm, which offer substantial time savings in applications where the high-dimensionality renders the MCMC approaches less practical. We demonstrate the usefulness of our method in predicting time to death in glioblastoma patients using pathways of genes.

4.1 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 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 can be beneficial for 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 MCMC for computation. In this chapter we consider an alternative approach, which utilizes pathway membership information as a source of group-driven shrinkage. This is achieved within the Bayesian LASSO context (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 being 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 to 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.

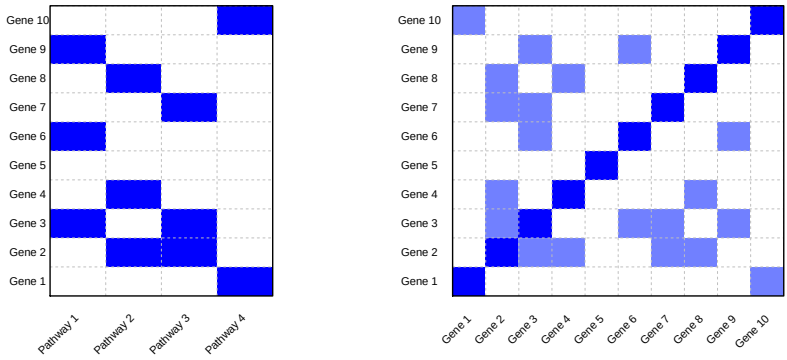
An important point of contrast between our method and the penalized regression approaches for structured variable selection such as group LASSO (Yuan and Lin, 2006) or Markov random field models on regression coefficients (Pan et al., 2010; Li and Li, 2008) is that the latter two enforce smoothness in 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 chapter we consider an alternative computational strategy, the maximum a posteriori estimation based on the EM algorithm. We build on work done previously by Griffin and Brown (2012) and we extend their algorithm by including structural grouping information. Similar as 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.

4.2 The Method

Consider the canonical multiple linear regression setting, where the $(n \times 1)$ vector of centered responses Y is linked to the $(n \times p)$ matrix X of standardized regressors (mean zero and variance one) through the relation $Y \sim \mathrm{N}_n(X\beta, \sigma^2 \mathrm{I}_n)$, where β 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 β is believed to be sparse in that only a small subset of predictors contributes to explaining the variability of the response.



(a) Pathway loading matrix

(b) Undirected graph

Figure 4.1: Loading matrix and undirected graph representations of gene interactions

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 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 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 4.1(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 $\mathscr{E} = \{(i,j) : 1 \leq i < j \leq p\}$, where $(i,j) \in \mathscr{E}$ 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 $(i,j) \in \mathscr{E}$. The zero patterns in matrix M are depicted for our simple example in Figure 4.1(b). It is easy to see that the off-diagonal elements in M copy the pattern of zeroes in the matrix 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 to depend on a combination of the ac-

tive 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.

4.3 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 Z) within the sparsity inducing regularization. The model formulation is as follows:

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

where Z_j denotes the j-th row of the $p\times(q+1)$ matrix $[1_p,Z]$ and $\Gamma(a,b)$ (resp. $\mathrm{IGamma}(a,b)$) 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 $\lambda=(\lambda_1,\ldots,\lambda_p)'$. An important ingredient in this formulation is the conjugacy, whereby 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 $\pi(\beta,\sigma^2|y,\lambda)$ (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 Z. Each λ_j is independently assigned a gamma distribution with expected value $\mathsf{E}\lambda_j = a\,h(Z_j'b)$, where coefficients $b = (b_0, \ldots, b_q)'$ 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 $Z_j'b$.

Assume for a moment that Z encodes for q non-overlapping groups, i.e. $\{1, \ldots, p\} = \bigcup_{k=1}^{q} \mathscr{Q}_k$, where $\mathscr{Q}_k \cap \mathscr{Q}_l = \emptyset$ for $k \neq l$. Then, $\forall j \in \mathscr{Q}_k$ we have $\mathsf{E}\lambda_j = a\,h(b_0 + b_k)$. 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. $\mathsf{E}\lambda_j = a\,h(b_0 + \sum_{k=1}^{q} \mathrm{I}[j \in \mathscr{Q}_k] b_k)$.

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 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 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 b in the inverse link and for inverse gamma priors $\texttt{IGamma}(\alpha, 1/\gamma)$ 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 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 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.

4.4 EM Algorithm for the Extended NEG Prior

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 adopt these to our situation, we have the additional difficulty of estimating the pathway weights b, which requires extensions of existing approaches.

In the EM algorithm, modified for Bayesian modal estimation (McLachlan and Krishnan, 1996, p. 26), the logarithm of the incomplete data likelihood is augmented by the logarithm of the prior density. 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 $\log p(\beta, b, \sigma^2 \mid y)$. These values are obtained by iteratively maximizing the conditional expectation of the log complete posterior $\log p(\beta, b, \sigma^2, w \mid y)$ with respect to the conditional distribution of the latent variables w given the current parameter estimates and the observed data.

Since the parameters β, b and σ are of interest, the candidates for the latent unobserved data are either τ^2 and λ^2. Instead of assuming that both τ^2 and λ^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 λ^2 and treat the latent variances τ^2 as missing. This formulation exploits the normal-scale mixture representation of the NEG prior. In the second version, we integrate over τ^2 and treat the penalty parameters λ as missing, which corresponds to the Laplace prior formulation.

4.4.1 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 $\beta^{(k)}, b^{(k)}, \sigma^{(k)}$ at the k-th iteration. This objective function, which is to be maximized in the subsequent M-step, takes the following form:

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

where

$$\begin{aligned} Q_1\left(\beta, \sigma \mid \beta^{(k)}, b^{(k)}, \sigma^{(k)}\right) =& -\frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2} - \frac{1}{2\sigma^2}\sum_{j=1}^{p} \beta_j^2 \, \mathsf{E}_{\tau^2 \mid \cdot}\left(\frac{1}{\tau_j^2}\right) \\ &- \frac{n + p + 2c + 2}{2}\log(\sigma^2) - \frac{d}{\sigma^2}, \end{aligned}$$

$$Q_2\left(b\,|\,\beta^{(k)},b^{(k)},\sigma^{(k)}\right)=\sum_{j=1}^{p}\left\{\log[a\,h(Z_j'b)]-(a+1)\,\mathsf{E}_{\tau^2\,|\,\cdot}\log[1+\tau_j^2 h(Z_j'b)]\right\}+\sum_{l=0}^{q}\log\pi(b_l)$$

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

As a result of our hierarchical prior formulation, where coefficients β depend on the coefficients b only through the penalty parameters λ, the objective function $Q(\cdot)$ is separable with respect to b and $(\beta,\sigma)'$. 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 b is complicated by the unavailability of the conditional expectation $\mathsf{E}_{\tau^2\,|\,\cdot}\log[1+\tau_j^2 h(Z_j'b)]$ 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 $h(Z'b)=Z'b$. Recall that for the identity link we use independent inverse gamma priors on the elements of 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 $Q_2\left(b\,|\,\beta^{(k)},b^{(k)},\sigma^{(k)}\right)$ we choose $b^{(k+1)}$ such that

$$Q_2\left(b^{(k+1)}\,|\,\beta^{(k)},b^{(k)},\sigma^{(k)}\right)\geq Q_2\left(b^{(k)}\,|\,\beta^{(k)},b^{(k)},\sigma^{(k)}\right). \tag{4.4.1}$$

Such a condition on $b^{(k+1)}$ 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 $b^{(k+1)}$ that satisfies property (4.4.1) can be found by maximizing a surrogate minorizing function, the definition of which is given below.

Definition 4.4.1. Let $b^{(k)}\in \mathrm{D}\subset\mathbb{R}^{q+1}$ represent a fixed value of the parameter vector b and let $f(b;b^{(k)})$ denote a real-valued function. Then $f(b;b^{(k)})$ is said to be minorizing a real valued function $g(b)$ at $b^{(k)}$ in domain D if and only if

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

From the definition of the minorizing function, it easily follows (McLachlan and Krishnan, 1996, p. 278) that $g(b^{(k+1)})\geq g(b^{(k)})$, where $b^{(k+1)}$ maximizes the surrogate function $f(b;b^{(k)})$. The question remains how to construct a suitable minorizing function for $Q_2\left(b\,|\,\beta^{(k)},b^{(k)},\sigma^{(k)}\right)$. The answer is given in the following theorem.

Theorem 4.4.1. *Let $b^{(k)} \in \mathbb{R}_+^{q+1}$ represent a fixed value of the parameter vector b. Denote*

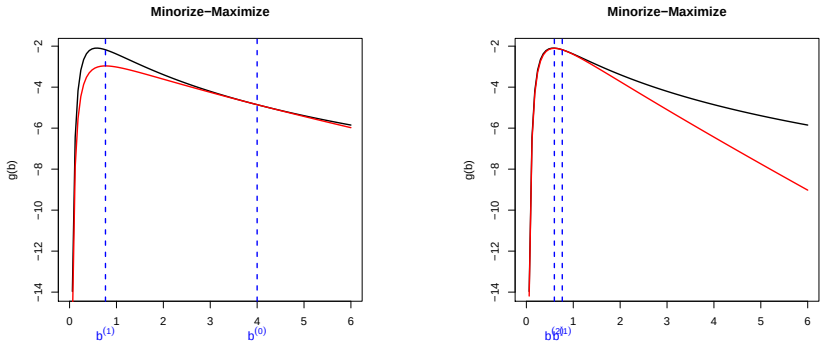
$$M_2\left(b \,|\, \beta^{(k)}, b^{(k)}, \sigma^{(k)}\right) = \sum_{j=1}^{p}\left[\log(aZ_j'b) - (a+1)\mathsf{E}_{\tau^2|\cdot}\left(\frac{\tau_j^2}{1+\tau_j^2 Z_j' b^{(k)}}\right) Z_j'(b-b^{(k)})\right] - \sum_{j=1}^{p}(a+1)\,\mathsf{E}_{\tau^2|\cdot}\log[1+\tau_j^2 Z_j' b^{(k)}] + \sum_{l=0}^{q}[-(\alpha+1)\log b_l - \gamma/b_l].$$

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

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

Several observations can be made based on the result of Theorem 4.4.1. First, the minorizing function $M_2\left(b \,|\, \beta^{(k)}, b^{(k)}, \sigma^{(k)}\right)$ no longer entails the evaluation of an integral which depends on the unknown parameter values b. All the integrals in the minorizing functions depend only on the current parameter values $b^{(k)}$. Furthermore, the cumbersome expectation $\mathsf{E}_{\tau^2|\cdot}\log(1+\tau_j^2 Z_j' b^{(k)})$ does not need to be computed, as the summand involving this term does not depend on b. Second, the values maximizing the minorizing function can be regarded as MAP estimates in a Bayesian regression with exponentially distributed responses $(a+1)/a\,\mathsf{E}_{\tau^2|\cdot}\left(\frac{\tau_j^2}{1+\tau_j^2 Z_j' b^{(k)}}\right)$, which are related to the regression matrix aZ via an inverse link function, and where the regression coefficients b are assumed to be independently gamma distributed. Third and most importantly, the expectations $\mathsf{E}_{\tau^2|\cdot}\left(\frac{\tau_j^2}{1+\tau_j^2 Z_j' b^{(k)}}\right)$ 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 4.2. The solid 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 $Z_1 = 1$. We have the initial estimate $b^{(0)} = 4$, at which we construct the minorizing function $f(b; b^{(0)})$ according to Theorem 4.4.1 (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 4.2(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 4.2: 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 $\log[1+\tau_j^2 h(Z'b)]$, 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 $\beta^{(0)}, b^{(0)}, \sigma^{(0)}$. The below described steps are then repeated until a convergence criterion is satisfied (e.g. $|\beta^{(k+1)}-\beta^{(k)}|_{l_1}+|b^{(k+1)}-b^{(k)}|_{l_1}<\varepsilon$).

☐ E-step

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

$$\mathsf{E}_{\tau^2|\cdot}\left(\frac{1}{\tau_j^2}\right)=\frac{2(a+0.5)\sigma^{(k)}\sqrt{Z_j'b^{(k)}}}{|\beta_j|^{(k)}}\frac{D_{-2(a+1)}\left(\frac{|\beta_j^{(k)}|\sqrt{Z_j'b^{(k)}}}{\sigma^{(k)}}\right)}{D_{-2(a+0.5)}\left(\frac{|\beta_j^{(k)}|\sqrt{Z_j'b^{(k)}}}{\sigma^{(k)}}\right)}, \tag{4.4.2}$$

where $D_\eta(x)$ denotes the parabolic cylinder function (Gradshteyn and Ryzhik, 2000, p. 256). We then denote $\Omega^{(k)}=\mathrm{diag}\left[\mathsf{E}_{\tau^2|\cdot}\left(1/\tau_j^2\right)\right]_{j=1}^{p}$ the diagonal matrix with the entries (4.4.2) on

the diagonal. Next, we compute $\mathsf{E}_{\tau^2|\cdot}\left(\frac{\tau_j^2}{1+\tau_j^2 Z_j' b^{(k)}}\right)$ for $j=1,\ldots,p$ and we stack the values in a $p\times 1$ vector $\Lambda^{(k)}$. We obtain (proof in Appendix B)

$$\mathsf{E}_{\tau^2|\cdot}\left[\frac{\tau_j^2}{1+\tau_j^2 Z_j' b^{(k)}}\right] = \frac{a\Gamma(a+0.5)}{\sigma^{(k)}\sqrt{2\pi Z_j' b^{(k)}}}\frac{1}{p(\beta_j^{(k)}|b^{(k)},\sigma^{(k)})}\times \tag{4.4.3}$$

$$\Psi\left(a+0.5,-\frac{1}{2};\frac{\beta_j^{(k)2} Z_j' b^{(k)}}{2\sigma^{(k)2}}\right), \tag{4.4.4}$$

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

□ M-step

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

$$\beta^{(k+1)} = \text{argmin}_{\beta\in\mathbb{R}^p}\{|Y-X\beta|_{l_2}+|\Omega^{(k)1/2}\beta|_{l_2}\}, \tag{4.4.5}$$

where $\Omega^{(k)1/2}$ denotes the square root of the matrix $\Omega^{(k)}$. The computation of the closed form solution $(X'X+\Omega^{(k)})^{-1}X'Y$ 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

$$\sigma^{2(k+1)} = (|Y-X\beta^{(k+1)}|_{l_2}+|\Omega^{(k)1/2}\beta^{(k+1)}|_{l_2}+2d)/(n+p+2c+2).$$

Finally, the pathway weights are updated according to Theorem (4.4.1) as values maximizing the function $M_2(b\,|\,\beta^{(k)},b^{(k)},\sigma^{(k)})$. Keeping only the summands in $M_2(\cdot)$, which depend on b, we obtain $b^{(k+1)}$ as

$$b^{(k+1)} = \text{argmax}_{b\in\mathbb{R}_+^{q+1}}\Big\{\sum_{j=1}^{p}\Big[\log(aZ_j'b)-(a+1)\Lambda_j^{(k)}Z_j'b\Big] \tag{4.4.6}$$

$$+\sum_{l=0}^{q}[-(\alpha+1)\log b_l-\gamma/b_l]\Big\}, \tag{4.4.7}$$

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.

4.4.2 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 b, where small values indicate more evidence for the importance of the pathway. Another limitation is the inability to estimate the coefficients directly as 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 λ^2 and treated the latent variances τ^2 as missing data. Now we do exactly the opposite, we integrate over τ^2 and treat the penalties λ^2 as missing.

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

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

where

$$\begin{aligned}\widetilde{Q}_1\left(\beta,\sigma \,|\, \beta^{(k)},b^{(k)},\sigma^{(k)}\right) =& -\frac{(Y-X\beta)'(Y-X\beta)}{2\sigma^2} - \frac{\sqrt{2}}{\sigma}\sum_{j=1}^{p}|\beta_j|\mathsf{E}_{\lambda^2|\cdot}\lambda_j \\ &- \frac{n+p+2c+2}{2}\log(\sigma^2) - \frac{d}{\sigma^2}\end{aligned}$$

and

$$\begin{aligned}\widetilde{Q}_2\left(b \,|\, \beta^{(k)},b^{(k)},\sigma^{(k)}\right) =& \sum_{j=1}^{p}\left[-a\log h(Z_j'b) - \frac{\mathsf{E}_{\lambda^2|\cdot}\lambda_j^2}{h(Z_j'b)}\right] \\ &+ \sum_{l=0}^{q}\left[(\alpha-1)\log b_l - \gamma b_l\right]\end{aligned}$$

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

The expected log complete posterior distribution is again separable with respect to b and $(\beta, \sigma)'$. In contrast to the previous version of the EM algorithm, the coefficients $\beta^{(k+1)}$ 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 $\beta^{(0)}, b^{(0)}, \sigma^{(0)}$. The below described steps are then repeated until a convergence criterion is satisfied (e.g. $|\beta^{(k+1)} - \beta^{(k)}|_{l_1} + |b^{(k+1)} - b^{(k)}|_{l_1} < \varepsilon$).

□ E-step

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

$$\mathsf{E}_{\lambda^2|\cdot}\lambda_j^s = \frac{[h(Z_j' b^{(k)})]^{(s+1)/2}}{\sigma^{(k)}\Gamma(a)2^{a+s/2}} \exp\left(\frac{\beta_j^{(k)2} h(Z_j' b^{(k)})}{4\sigma^{(k)2}}\right) \times \tag{4.4.8}$$

$$D_{-(2a+1+s)}\left(\frac{|\beta_j|\sqrt{h(Z_j' b^{(k)})}}{\sigma^{(k)}}\right). \tag{4.4.9}$$

□ M-step

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

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

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

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

Finally, the updates $b^{(k+1)} = \text{argmax}_{b \in \mathbb{R}^{q+1}} \widetilde{\mathscr{Q}}_2\left(b \,|\, \beta^{(k)}, b^{(k)}, \sigma^{(k)}\right)$ 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 $\mathsf{E}_{\lambda|\cdot}\lambda_j$, which are related to the regression matrix Z through the $h(\cdot)$ link function, assuming independent gamma distributed priors on the regression coefficients b.

4.5 Hierarchical Variable Selection

A natural strategy for variable selection based on the posterior output $(\widehat{\beta},\widehat{b},\widehat{\sigma})$ 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 $\widehat{\beta}$, 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.

4.6 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)):

$$p_{a,s,\sigma}(\beta) = \frac{a\,2^a\sqrt{s}}{\sqrt{\pi\sigma^2}}\Gamma(a+0.5)\exp\left(\frac{\beta_j^2 s}{4\sigma^2}\right)D_{-2(a+0.5)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right). \tag{4.6.10}$$

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 (4.6.10) becomes flatter, loosing the ability to shrink noise signals due to a less pronounced peak at zero. With both for 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 in the following theorem.

Theorem 4.6.1. *Let $p_{a,s,\sigma}(\beta)$ denote the density function in* (4.6.10)*. 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 $\psi(t) = (1-\mathrm{i}ts)^{-a}$. 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 the l'Hospital rule. The limit is a characteristic function of a Dirac distribution concentrated at λ'. Denote $p_{a,s}(\lambda^2)$ 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

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

which is a 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 4.6.1. 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. $Y|\beta,\sigma^2 \sim \mathrm{N}(\beta,\sigma^2)$ 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 (4.6.10). As given in Griffin and Brown (2012),

$$\pi_{a,s,\sigma}(|\beta_j|) = \frac{(2a+1)\sqrt{s}}{\sigma}\frac{D_{-2(a+1)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)}{D_{-2(a+0.5)}\left(\frac{|\beta_j|\sqrt{s}}{\sigma}\right)}. \tag{4.6.11}$$

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

Theorem 4.6.2. *Let $\pi_{a,s,\sigma}(|\beta|)$ denote the term in* (4.6.11)*, then $\pi_{a,s,\sigma}(|\beta|) = \mathscr{O}\left(\frac{1}{|\beta|}\right)$ as $|\beta| \to \infty$.*

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

$$D_\eta(x) \sim \exp(-x^2/4)x^\eta \left(1 - \frac{\eta(\eta-1)}{2x^2} + \frac{\eta(\eta-1)(\eta-2)(\eta-3)}{2.4x^4} - \dots\right) \tag{4.6.12}$$

where $\sim$ sign indicates that the Parabolic cylidner 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 the $\lim_{|\beta|\to\infty} \pi'_{a,s,\sigma}(|\beta|)$ as

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

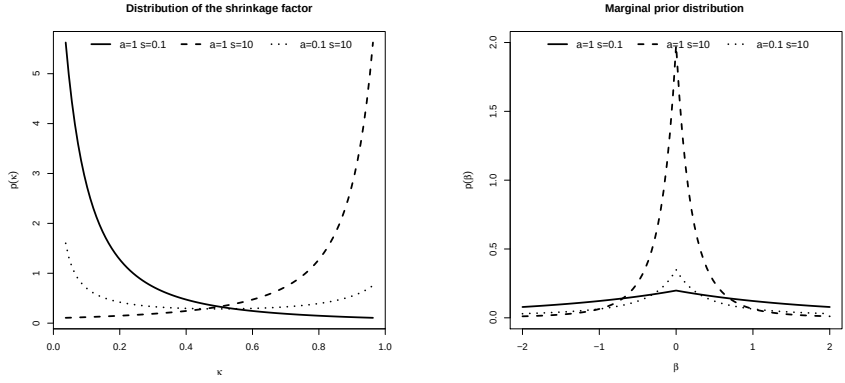
which was to be demonstrated.

Remark 4.6.2. 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 $\mathsf{E}(\beta_j\,|\,y_j,\sigma^2) = [1-\mathsf{E}(\kappa_j\,|\,y_j,\sigma^2)]y_j$. 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 4.3(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,



(a) Prior distribution of shrinkage factor

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

Figure 4.3: Prior distribution of shrinkage factor and regression coefficients

which in turn induces heavier tails of the NEG prior. It is possible to select a configuration of the two parameters, which implies 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 4.3(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

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

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

In the context of multiple linear regression, the small fixed values σ^2 may increase the number of false positives. In our EM algorithm, small values $\sigma^{(k)}$ at k-th iteration imply smaller penalties on the regression coefficients (as seen from equation (4.4.2)) and thereby increased likelihood of false discoveries. This may be problematic in high-dimensional settings ($p > n$), where the variance estimates at each iteration are typically very small. Possible remedies to this problem are: (a) considering higher values of the shape parameter a, (b) specifying an informative prior on the variance, such as flat prior within an interval bounded away from zero, (c) adding 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 hyper-parameter 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 that g grows 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 on 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 asymptotical normality of the MAP estimator under mild regularity conditions in the multiple regression considering fixed p. For simplicity we will assume that σ is fixed to one and we let the scale parameter s vary according to the sample size.

Theorem 4.6.3. *Assume the regularity conditions (A)-(C) in Fan and Li (2001) and denote $\widehat{\beta}_n$ the MAP estimator arising from the hierarchical model under NEG$(a, n^2 s_n)$ prior. Denote $\mathscr{A}_n = \{j : \widehat{\beta}_j \neq 0\}$ and $\mathscr{A} = \{j : \beta_j \neq 0\}$, where β 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 $\widehat{\beta}_n$ satisfies:*

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

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

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

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

where the penalty term consists of the summands in negative NEG$(a, n^2 s_n)$ density, which depend on $|\beta|$, divided by n. According to (4.6.10), the penalty term for $\sigma^2 = 1$ takes the following form:

$$pen_{a,s_n}(|\beta|) = -\frac{|\beta|^2 n s_n}{4} - \frac{1}{n}\log D_{-2(a+0.5)}\left(|\beta| n\sqrt{s_n}\right). \tag{4.6.13}$$

Denote $pen'_{a,s_n}(|\beta|)$ and $pen''_{a,s_n}(|\beta|)$ the first and second derivatives of (4.6.13) with respect to $|\beta|$. In order to demonstrate the 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 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:

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

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 $o(n)$, (b) the first summand is asymptotically equivalent to $ns_n(2a+1)$. This altogether implies that the limit $pen''_{a,s_n}(|\beta|)$ 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 4.6.3. The "oracle" properties of the NEG penalty (without scaling) were in the penalized likelihood setting with a diverging number of parameters 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 4.6.4. 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 4.6.5. 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 4.6.3.

4.7 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}.

4.7.1 Non-overlapping Groups

In the first example, we assume $p = 1\,000$ and $n = 100$. The matrix of predictors X has been generated with rows drawn independently from $\mathrm{N}_p(0,\Sigma)$, 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 $\beta = (1,2,3,4,5,0'_{15},1,2,3,4,5,0'_{975})'$, where 0_m is a $m \times 1$ vector of zeroes, and we construct the responses according to the generating linear model $\mathrm{N}_n(X\beta, 3 \times \mathrm{I}_n)$.

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 $\beta^{(0)} = 1_p$, $b^{(0)} = 1_5$ and $\sigma^{(0)} = 1$.

In all considered settings, the 10 relevant predictors were correctly identified. Table 4.1 and 4.2 summarize 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).

	Grouping 1							NEG	
Size				5	15	5	975		1000
Sparsity				1	0	1	0		1/100
	FD	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	52	0.001
a=1	42	0	0.001	4.828	0	4.780	0	45	0.001
a=3	34	0	0.002	13.138	0	12.805	0	33	0.002
Rescaled prior $g = 1/n^2$									
a=0.5	49	0	0.001	2.412	0.002	2.412	0	49	0.001
a=1	45	0	0.054	4.850	0.004	4.850	0	45	0.06
a=3	35	0	2.122	12.613	0	12.613	0	35	2.609

Table 4.1: Analysis summary of the simulated data, FD/FD1/FD2/FDH refer to number of false positives overal/in non-predictive groups/in predictive groups/overally after hierarchical selection. The size and sparsity relate to the number of predictors within each group and proportion of predictive explanators.

Focusing on the estimates of the pathway weights, several observations can be made based on the reported estimates in Table 4.1 and 4.2. 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 4.1 and 4.2 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 4.1 and 4.2). 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 being 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 parameters is in our simulated example 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 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 4.6.3, 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 contemplate that rescaling the prior, the regression on the scale parameter has a 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 a useful prerequisite for group selection that isolates discoveries in non-predictive groups.

	Grouping 2							
Size					10	20	30	940
Sparsity					1/2	1/4	0	0
	FD1	FD2	FDH	$\widehat{b}_0$	$\widehat{b}_1$	$\widehat{b}_2$	$\widehat{b}_3$	$\widehat{b}_4$
No scaling $g = 1$								
a=0.5	40	8	10	0.001	1.235	0.016	0	0
a=1	40	17	19	0.001	3.133	1.349	0	0
a=3	31	17	19	0.002	11.590	9.332	0.002	0
Rescaled prior $g = 1/n^2$								
a=0.5	45	4	4	0.001	1.012	0.050	0	0
a=1	43	2	4	0.053	2.333	1.408	0.007	0
a=3	34	1	3	2.111	10.389	8.976	0.035	0

Table 4.2: Analysis summary of the simulated data, FD/FD1/FD2/FDH refer to number of false positives overal/in non-predictive groups/in predictive groups/overally after hierarchical selection. The size and sparsity relate to the number of predictors within each group and proportion of predictive explanators.

4.7.2 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 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 select 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 will consider two possible scenarios: (1) all genes within the predictive pathways are assumed to contribute in explaining of the variability of the response (Grouping 1), (2) predictive pathways are sparse (Grouping 2). We shall 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 Z (associated with Grouping 2), we first generate the covariance matrix $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})_{i,j=1}^{p}$, where $\widetilde{\Sigma} = Z\mathrm{diag}\{\rho_1,\ldots,\rho_q\}Z' + \mathrm{I}_p$, 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 X is then generated according to $\mathrm{N}_n(0,\Sigma)$. The observations on the response variable are created according to the relation $\mathrm{N}_n(X\beta,\sigma^2\mathrm{I}_n)$. We keep $\sigma^2 = 1$, $n = 100$ and we assume (a) relatively high signal to noise ratio, (b) medium correlation within non-predictive pathways, (c) high correlation within predictive pathways. Namely, the nonzero entries in the regression vector β equal 2. In order to obtain average correlation of 0.8 and 0.3 within the predictive and non-predictive pathways,

			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

				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	9	0.001	0.175	0.246	0.001	0	0.007	0	0.001	0	0	0
a=1	36	4	8	0.001	4.430	5.440	0	0	0.003	0	0.001	0	0	0
a=3	23	0	2	3.198	18.049	19.272	0	0	0.729	0	0	0	0	0

Table 4.3: Analysis summary of the simulated data, FD/FD1/FD2/FDH refer to number of false positives overally/in non-predictive groups/in predictive groups/overally after hierarchical selection. The size and sparsity relate to the number of predictors within each group and proportion of predictive explanators.

we assume $\rho_j = 0.1 \times \mathrm{I}\left(j \notin \bigcup_{k=1}^{3} \mathscr{Q}_k\right) + 2 \times \mathrm{I}\left(j \in \bigcup_{k=1}^{3} \mathscr{Q}_k\right)$.

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

The summary of the analysis for the non-sparse clusters (Grouping 1) is in Table 4.3. 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 4.3). Again, no false-nondiscoveries were observed.

In the second grouping scenario (Table 4.3) 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 Table 4.3 again show the superiority of the rescaled model, both in the accuracy of determining important pathways, as well as 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 BIC criterion, 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.

4.8 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

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 4.4: LASSO selected genes together with 10 top represented pathways.

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 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).

On 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 4.4 together with information on their pathway involvement (top 10 represented pathways with at least 3 genes).

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 start-

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

ing vector $\beta^{(0)} = 1_p$ and 10 initial values randomly sampled from from $\mathrm{N}_p(0, \mathrm{I})$. The starting values for the pathway weights and variance parameter are $b^{(0)} = 1$ and $\sigma^{(0)} = 1$.

The highest located log-posterior mode (10595.72 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 4.5). 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 4.5 reports a subset of 10 pathways with the highest numbers of identified genes together with the estimated weights $\widehat{b}$, which represent the proportion of within group predictive genes. The complete list of gene-pathway interactions for all the 21 pathways is in Appendix 3.

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 CX3CL1, controlling neuronal survival and neuron transmission (Sciumč 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 specific type of glioma (McDonald et al., 2005), FRAP1 is a member of 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 4.9 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 molecular basis for malignant gliomas (Paulus and Tonn, 1995). Expression of `cell adhesion molecules` (binding proteins) has been shown consistently altered in glioblastoma as compared the normal brain tissue (Gingras et al., 1995). The full list of the 21 identified pathways is deferred to the Appendix E.

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) lead to a lower log posterior mode (9428.231 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 6780 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 BIC criterion. The list of these pathways is in Appendix E. 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 β and b. The update β 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 between $20-40$ 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 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. To be contrasted with MCMC implementation of the Bayesian LASSO (R-package `monomvn`), drawing 100 samples from the posterior took around 20 seconds.

4.9 Discussion

In this chapter 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 (West, 1987), 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 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 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 $\exp(-n\,\mathrm{pen}_\lambda(|\boldsymbol{\beta}|))$. The penalty function $\mathrm{pen}_\lambda(|\boldsymbol{\beta}|)$ 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 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 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.

4.10

Appendix

4.10.1 A: Proof of Equation (4.4.2)

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 $\mathsf{E}_{\tau^2|\cdot}\left(\frac{1}{\tau_j^2}\right)$ 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:

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

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

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

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

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

4.10.2 B: Proof of Equation (4.4.3)

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

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

4.10.3 C: Proof of Equation (4.4.8)

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

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

4.10.4 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

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

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

			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 4.6: 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.

rows drawn independently from $\mathrm{N}_p(0, \mathrm{I}_p)$. Ten response vectors were generated according to $Y \sim \mathrm{N}(X\beta, 3 \times \mathrm{I}_n)$ for each sample size. The average estimated group weights after in Table 4.6 below.

We observe a decreasing trend in the estimated weights with the growing sample size in the unscaled model. After rescaling, the weights are seen to increase, where the scale parameter (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.

	Perfect Grouping			**Imperfect Grouping**			**NEG**	
Sparsity	FD	FN	FDH	FD	FN	FDH	FD	FN
β_1	44.2	10.6	0	52.8	20.9	9	53.5	34.6
β_2	49.9	4.4	0	60.6	17.8	11.4	57.1	21.2
β_3	52.6	0.2	0	60.9	2.9	19.8	53.1	3.7

Table 4.7: Simulation study with different degrees of sparsity. FD/FN/FDH stand for false discoveries/false non-discoveries/false discoveries after the hierarchical variable selection

■ 4.10.5 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:

$$\beta_1 = (\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{470},\underbrace{1,\ldots,1}_{30},\underbrace{0,\ldots,0}_{470})',$$
$$\beta_2 = (\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{480},\underbrace{1,\ldots,1}_{20},\underbrace{0,\ldots,0}_{480})',$$
$$\beta_3 = (\underbrace{1,\ldots,1}_{10},\underbrace{0,\ldots,0}_{490},\underbrace{1,\ldots,1}_{10},\underbrace{0,\ldots,0}_{490})'.$$

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 $\mathcal{Q}_1 = \{1,\ldots,40\}$, $\mathcal{Q}_2 = \{41,\ldots,500\}$, $\mathcal{Q}_3 = \{501,\ldots,540\}$, $\mathcal{Q}_4 = \{541,\ldots,1000\}$. We consider a covariance matrix $\Sigma = \left\{0.5^{|i-j|}\right\}_{i,j=1}^{p}$ to generate predictors from $\mathrm{N}_p(0,\Sigma)$.

Results are summarized in Table 4.7, 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 exemption was the least sparse model associated with β_1 in Table 4.7. 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 β_3 in Table 4.7 and also simulated examples in our manuscript).

4.10.6 F: Complete description of gene/pathway information

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

Table 4.8: 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 4.9: Results obtained from NEG grouping model. Table reports the involvement of 21 identified genes within 21 identified pathways

CHAPTER 5

FAST DYNAMIC POSTERIOR EXPLORATION FOR FACTOR AUGMENTED MULTIVARIATE REGRESSION

Rockova, V., Lesaffre, E. 2013. **Fast Dynamic Posterior Exploration for Factor Augmented Multivariate Regression**. Manuscript in preparation

Abstract

Advancements in high-throughput experimental techniques have facilitated the availability of diverse genomic data, which provide complementary information regarding the function and organization of gene regulatory mechanisms. The massive accumulation of data has increased demands for more elaborate modeling approaches that combine the multiple data platforms. We consider a sparse factor regression model, which augments the multivariate regression approach by adding a latent factor structure, thereby allowing for dependent patterns of marginal covariance between the responses. In order to enable the identification of parsimonious structure, we impose spike and slab priors on the individual entries in the factor loading and regression matrices. The continuous relaxation of the point mass spike and slab enables the implementation of a rapid EM inferential procedure for dynamic posterior model exploration. This is accomplished by considering a nested sequence of spike and slab priors and various factor space cardinalities. Identified candidate models are evaluated by a conditional posterior model probability criterion, permitting trans-dimensional comparisons. Patterned sparsity manifestations such as an orthogonal allocation of zeros in factor loadings are facilitated by structured priors on the binary inclusion matrix. The model is applied to a problem of integrating two genomic datasets, where expression of microRNA's is related to the expression of genes with an underlying connectivity pathway network.

5.1 Introduction

The systematic analysis of the cancer genome and transcriptome has over the past decades identified profound modifications of expression homeostasis involving both coding and non-coding genes. Modulation of gene expression can be mediated by many intricate biological processes, study of which has become instrumental in characterizing disease pathogeneses. The recent explosion of data of diverse genomic phenomena has provided complementary information about the function and organization of gene regulatory mechanisms. This massive accumulation of information has resulted in forces towards building more comprehensive models, that integrate multiple data platforms. The increasing capacity to quantify the expression dynamics of the transcriptome will soon begin to open new opportunities to model the regulatory mechanisms using dynamic statistical models. Nowadays, the majority of routinely analyzed data are limited by design or technology, forcing the design of simplistic statistical models that rely strongly on unrealistic biological assumptions. This work presents one alternative approach to generating insights about microRNA mediated modulation of gene expression from two sets of static expression data. Turning to the pragmatic challenges that these data pose, it is the high-dimensionality which most complicates the computational tractability and precludes the use of standard methods, for which an enormous effort would have to be exercised to discern the relevant from the noise. We implement an expeditious inferential procedure that profoundly facilitates computation in high volume data.

Our work is anchored by the development of a factor regression framework for elucidating associations between two (high-dimensional) sets of variables, where directionality exists that designates one set as predictors and the other set as responses. The multivariate regression model is augmented by adding latent factors to decipher the pattern of marginal covariance after adjustment for the predictors. In addition, we address uncertainty about the cardinality of the factor space. The factor model is better suited for the interpretation rather than to the accurate prediction of the marginal covariance matrix. We envision that a practitioner would like to recover an interpretable pattern of sparsity, when it exists. The key instrument to detecting sparsity is the continuous relaxation of the point mass spike and slab prior, which admits implementation of rapid inferential schemes (Rockova and George, 2013). Inducing a mixture prior on every single regression coefficient and factor loading, a binary inclusion matrix is used to encode the factor model configuration. Patterned variable selection is then facilitated by structured priors on the model matrix. In order to yield a complementary allocation of zeroes and to mitigate factor splitting in over-parametrized models, we implement a row-wise multinomial-Dirichlet prior on the factor loading matrix.

Our principal contribution is the development of a rapid inferential algorithm for factor model exploration based on the EM algorithm, which leverages existing tools developed for probabilistic principal components (Tipping and Bishop, 1999) and variable selection in linear regression (Rockova and George, 2013). The proposed exploratory procedure is a multivariate factor extension of the EMVS procedure of Rockova and George (2013) and admits computation using closed form expressions. For greater flexibility in detecting high posterior models, we proceed sequentially with a series of nested of spike and slab priors and considering various factor space cardinalities. The posterior identification of candidate models, based on a local median probability model rule (Barbieri and Berger, 2004), is followed by model evaluation using a conditional posterior model probability criterion assuming a point mass at zero. We contemplate that the EM algorithm is likely to generate more interesting candidate models when the assumed factor cardinality is close to the effective dimension. Inference about factor dimensionality can be guided by the sparsity pattern in over-parametrized models and grounded by the evaluation criterion, which admits trans-dimensional comparisons.

The usefulness of the model will be demonstrated on the problem of describing microRNA regulatory networks in acute myeloid leukemia. MicroRNAs are short non-coding RNA's that down-regulate expression of their gene targets through complementary base pairing. Apart from automated algorithms, which predict putative targets by just their genomic content, there has been an emergence of statistical prediction models that also take experimental data into account (Stingo et al., 2010; Huang and Morris, 2007; Zacher et al., 2012). Many of these approaches rest on the simplifying assumption of conditional independence between genes, given the microRNAs. Recent biological evidence suggests that related genes with similar genomic recognition elements (not necessarily coding for a protein) can attract similar microRNAs. This competition for the limited pool of microRNAs then induces a distorted balance in concentration of the competing genes. Such between gene communication is difficult to capture from the snapshot expression measurements and statistical evidence for

this mechanism is still missing. Despite this complexity of microRNA regulation, we use the microRNA and gene expression datasets to illustrate our developed methodology.

5.2 Factor Regression Model Structure

The data setup under consideration consists of a $n \times G$ matrix $Y = [y_1, \ldots, y_n]'$ containing n independent observations on G related responses and a $n \times p$ predictor matrix $X = [x_1, \ldots, x_n]'$. Before proceeding, it will be beneficial to center the columns in X and Y around zero and standardize X to have unit column-wise variances. We assume throughout that y_i's arise as independent realizations from a latent factor regression model, where the responses are mapped linearly on a space spanned by both observed explanatory variables and unobserved (latent) factors. Given ω_i, a $(d \times 1)$ vector of latent variables for the case i, we assume

$$f(y_i \mid \omega_i, A, B, \Sigma) = \mathrm{N}_G(Ax_i + B\omega_i, \Sigma), \quad 1 \leq i \leq n, \tag{5.2.1}$$

where the $G \times p$ matrix A consists of unknown regression coefficients, $\Sigma = \mathsf{diag}\{\sigma_j^2\}_{j=1}^G$ is a diagonal matrix of unknown positive scalars and the $G \times d$ matrix B contains factor loadings weighting the contributions of individual factors. Following the standard assumption, the latent vectors are considered to arise through random sampling from a Gaussian distribution $\mathrm{N}_d(0, \sigma_\omega \mathrm{I}_d)$. The variance parameter σ_ω is typically set to unity as a supplement to the identifiability constraints, a convention that we adopt here. The equation (5.2.1) induces a corresponding Gaussian distribution on the observations $f(y_i \mid A, B, \Sigma) = \mathrm{N}_G(Ax_i, BB' + \Sigma)$, $1 \leq i \leq n$. This permits dependent patterns of covariance among y_i, to be attributed to the common latent factors.

The factor model (5.2.1) is not identifiable without further constraints. One requirement is for B to be full rank in order to avoid identification problems arising through translational invariance of the factor model (Geweke and Singleton, 1980). Additional restrictions need to be imposed to guarantee that the number of free parameters does not exceed the number of parameters in the unrestricted covariance matrix $\mathrm{Var}(Y)$. Lastly, the parametrization needs to be invariant under invertible linear transformation of the factor vectors. The common convention has become to assume that B is zero upper-triangular with positive or even unit elements on the diagonal (Lopes and West, 2004). Despite the convenient interpretation, inference under this assumption depends on the ordering of the responses, an undesirable phenomenon. In order to mitigate the influence of the ordering in determining the leading variable for each factor, Frühwirth-Schnatter and Lopes (2009) considered alternative identifiability conditions, allowing for more relaxed patterns of zeroes. Nevertheless, these constraints complicate inference using deterministic computational tools, which typically require knowing beforehand over which parameters the optimization takes place. Unlike the use of factor analytic models to merely produce an accurate estimates of the covariance matrix, the interpretation of factor loadings is of the uttermost interest here. We aim to find a sparse interpretable approximation

to the factor loading matrix corresponding to the (global) posterior mode within a class of constrained lower-triangular matrices.

The factor regression model (5.2.1) may be contrasted with the orthogonal factor formulation of Yoshida and West (2010), where the loadings are required to be orthonormal. Whereas their formulation has many convenient properties such as shared pattern of zeroes in marginal concentration and covariance matrices, our formulation benefits from its resemblance to the multivariate linear regression, which will be exploited in designing the deterministic tool for model exploration. Instead of assuming orthonormal in factor loadings, in Section 4 we will induce an orthogonal allocation of zeroes by means of a structured prior on the loading matrix.

5.3 Sparsity Modeling with Spike and Slab Priors

The Bayesian approach to defining sparse latent and regression structures uses priors on the individual elements in $B = \{b_{jl}\}_{j,l=1}^{G,d}$ and $A = \{a_{jl}\}_{j,l=1}^{G,p}$ that induce either zeroes (Carvalho et al., 2008; Frühwirth-Schnatter and Lopes, 2009) or values close to zero with high-probability. We take the latter approach, exploiting the continuous relaxation of the point-mass mixture prior (George and McCulloch, 1993). The continuous spike and slab formulation is essential for derivation of efficient deterministic inferential tools (Stegle et al., 2000; Rockova and George, 2013). Brown et al. (1998) introduced matrix-variate spike and slab priors for multivariate variable selection using the notation of Dawid (1981), assuming that all the rows have the same correlation structure. Instead, we permit each row to be treated differentially, which creates more flexibility in characterizing the sparsity pattern in the regression and loading matrices. Denote $[A,B] = [\beta_1,\ldots,\beta_G]'$, where

$$\beta_j = (a_j', b_j')' = (\underbrace{a_{j1},\ldots\ldots\ldots,a_{jp}}_{\text{regression coefficients}},\underbrace{b_{j1},\ldots,b_{jd}}_{\text{factor loadings}})', \quad 1 \le j \le G.$$

Then each β_j is assigned a conjugate Gaussian mixture prior

$$\pi(\beta_j) \sim \mathscr{N}_{p+d}(0, \sigma_j^2 D_j) \tag{5.3.2}$$

where $D_j = \mathsf{diag}\{(1-\gamma_{jl})v_0 + \gamma_{jl}v_1\}_{l=1}^{p+d}$ and variance parameters v_0 and v_1 are set to small and large to distinguish the β_{jl} values which warrant a functional relationship between j-th response and l-th predictor (factor). Typically, we would like to make a distinction between the loadings and regression coefficients and allow for different values of v_0 and v_1. However, standardizing the predictors and assuming that latent factors arise from the standard normal distribution, it will often be sensible to use the same spike and slab prior. Here

$$\gamma_j = (\gamma_{ja}', \gamma_{jb}')' = (\underbrace{\gamma_{j1},\ldots\ldots\ldots,\gamma_{jp}}_{\text{predictor indicators}},\underbrace{\gamma_{jp+1},\ldots,\gamma_{jp+d}}_{\text{factor indicators}})', \quad 1 \le j \le G,$$

denotes the vector of inclusion indicators, which characterizes the binary selection status of each predictor and factor in relation to the j-th response. Stacking the inclusion vectors for each response in one matrix $\Gamma = [\gamma_1, \ldots, \gamma_G]'$, we obtain model configurations characterized by an active set of binary indicators. For σ_j^2, the diagonal elements in Σ, we assume independent inverse Gamma priors $\text{IG}(\eta/2, \eta\hat{\lambda}/2)$. The prior specification is completed by characterizing priors on the $G \times (p+d)$ model matrix $\Gamma = [\Gamma_a, \Gamma_b]$, consisting of two blocks for regression coefficients and factor loadings. One of the main thrusts of our modeling approach rests in inducing structure in the prior distributions in matrices Γ_a and Γ_b to encourage the manifestation of patterned sparsity.

5.4 Priors on the Binary Inclusion Matrix

The evidence for selecting variables and factors is aggregated in the posterior inclusion matrix $\Gamma = [\Gamma_a, \Gamma_b]$, given the observed data. If we were to make the strong assumption that all the elements γ_{ij} are independent, we would treat each indicator individually by assigning independent Bernoulli priors with a global inclusion probability θ. Such exchangeability has simplifying implications for inference since we can simply ignore the location of the indicator within the matrix. Sometimes, the exchangeability is too strong of an assumption and in the presence of prior knowledge about the patterned or collective behavior among the explanators, one might want to induce structure in the matrix Γ by making the indicators a priori dependent. Such relaxations include the partially exchangeable situation, where auxiliary partitions exist and define exchangeable sets of indicators. Partitions that arise naturally in our context occur across the columns in the matrix Γ.

To define column-wise exchangeable prior distributions (Frühwirth-Schnatter and Lopes, 2009) on $\Gamma = (\gamma_1, \ldots, \gamma_G)'$, we use a hierarchical prior which allows different occurrence probabilities $\theta = (\theta_1, \ldots, \theta_{p+d})'$ of non-zero elements in the different columns of $[A, B]$, e.g.

$$\pi(\Gamma \mid \theta) = \prod_{l=1}^{p+d} \theta_l^{\sum_{j=1}^G \gamma_{jl}} (1-\theta_l)^{G - \sum_{j=1}^G \gamma_{jl}} \tag{5.4.3}$$

and θ_l's are assigned independent Beta distributions $\mathscr{B}(a,b)$. Going further, there are many possible variations and extensions of exchangeable and partially exchangeable models. We would ideally like to have two separate mechanisms for the patterned sparsity (a) in the supervised learning about the regression coefficients and (b) in the unsupervised learning about the factor loadings.

5.4.1 Structured Multivariate Regression

The rectangular matrix of regression inclusion indicators Γ_a can be regarded as an adjacency matrix in a bipartite graph with two finite sets of nodes and directed arrows connecting them.

In the instance of structured predictors, which operate in groups or networks, preferable are priors that interconnect the subsets of related indicators within the rows of Γ_a. The assumption of row-wise exchangeability may not be warranted in cases where we have some additional information about the properties of the responses, such as the fact that they were produced in a particular temporal sequence, or reflect a known pattern of covariance. Natural extensions consider proliferation of inclusion probabilities not only within rows, but also within columns. Following the notation of Rockova and George (2013), we denote the partially exchangeable logistic regression product prior on the model matrix as

$$\pi(\Gamma_a \mid \theta) = \prod_{j=1}^{g}\prod_{k=1}^{p}\left(\frac{\exp(Z'_{G\times(j-1)+k}\theta)}{1+\exp(Z'_{G\times(j-1)+k}\theta)}\right)^{\gamma_{jk}}\left(\frac{1}{1+\exp(Z'_{G\times(j-1)+k}\theta)}\right)^{1-\gamma_{jk}}, \quad (5.4.4)$$

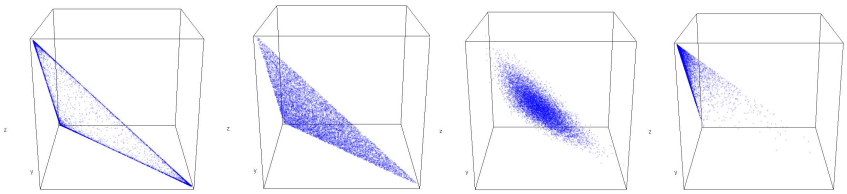
where $Z=[Z^1,\ldots,Z^q]$ is a $(Gp)\times q$ group identification matrix, and $z_{G\times(j-1)+k,l}=1$ if and only if γ_{jk} corresponds to the l-th group. The parameters $\theta=(\theta_1,\ldots,\theta_q)'$ then quantify the respective contributions of each of the q groups. The prior distribution on θ that corresponds to the beta-binomial prior in case of non-overlapping groups is the independent inverse logistic beta distribution (Rockova and George, 2013). A special case of this formulation was proposed by Stingo et al. (2010) for incorporating biological knowledge in multivariate regression. Given that the entries within the matrix Γ_a lie on a network connecting related entries by undirected edges, we can define the matrix-variate MRF prior, which is uniquely determined (Besag, 1974) by the conditional distributions of each indicator γ_{jk}, given its set of neighbors $\gamma_{jk\cdot}$,

$$\pi(\gamma_{jk} \mid \theta, \gamma_{jk\cdot}) = \left(\frac{\exp(\theta_1+\theta_2\sum_k\gamma_{jk\cdot})}{1+\exp(\theta_1+\theta_2\sum_k\gamma_{jk\cdot})}\right)^{\gamma_{jk}}\left(\frac{1}{1+\exp(\theta_1+\theta_2\sum_k\gamma_{jk\cdot})}\right)^{1-\gamma_{jk}}.$$

Here the parameter θ_1 regulates the sparsity and θ_2 the smoothness of the probability proliferations. Whereas the partially exchangeable prior admits closed form calculations in our inferential procedure, the MRF requires approximations (as discussed in Rockova and George (2013)).

5.4.2 Orthogonal Sparsity in Factor Loadings

Focusing on the explanation of the factor model rather than on an accurate prediction of the marginal covariance matrix, we would ideally like to generate interpretable patterns of sparsity in the factor loading matrix. The patterned allocation of zeroes can be encouraged by structured priors on the matrix of factor indicators Γ_b. Factor analytic models, even under strict identifiability conditions, often exhibit a phenomenon of factor splitting, where the activity of a single latent variable is smoothly redistributed across multiple factors. This is undesirable since we would prefer to obtain the minimal sparse representation leading to the same pattern of zeroes in the marginal covariance matrix. In many practical situations, it would even be



(a) $\alpha = (0.1, 0.1, 0.1)'$ (b) $\alpha = (1, 1, 1)'$ (c) $\alpha = (10, 10, 10)'$ (d) $\alpha = (0.1, 1, 10)'$

Figure 5.1: Scatterplot of simulated data from a 3-dimensional Dirichlet distribution with various concentration vectors

desirable to yield a complementary allocation of zeros in the columns of the loading matrix. Such representations would be useful when one is interested in finding the likely grouping of responses based on their affinity for a single latent factor.

The smooth structured prior parametrizations (two examples outlined in the previous section) are less convenient for modeling mutually exclusive or competitive indicators. The dependence can be introduced indirectly by allowing each indicator γ_{jk} to have an individual inclusion probability θ_{jk} and by subjecting the matrix $\Theta = (\theta_{jk})_{j,k}^{G,d}$ to a set of constraints. A useful parametrization can be obtained by allowing every response to interact with one and only one factor by assuming a multinomial distribution $\mathscr{M}ult(\gamma_{j1}, \ldots, \gamma_{jd}; 1; \theta_{j1}, \ldots, \theta_{jd})$ for every row $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jd})'$ in the matrix Γ_b. To put this down formally, denote $\gamma_{j\cdot} = \sum_{l=1}^{d} \gamma_{jl}$ and define

$$\pi(\gamma_j; \theta_j) = \begin{cases} \prod_{l=1}^{d} \theta_{jl}^{\gamma_{jl}} & \text{if} \quad \sum_{l=1}^{d} \gamma_{jl} = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5.4.5}$$

The allocation proportions in this multinomial distribution are constrained to sum up to one and typically equipped with the Dirichlet prior distribution $\mathscr{D}(\alpha_1, \ldots, \alpha_d)$. Adopting this convention, we specify row-wise independent Dirichlet priors in the matrix Θ with a common vector of positive concentration parameters $\alpha = (\alpha_1, \ldots, \alpha_d)'$, i.e.

$$\pi(\theta_j) \sim \mathscr{D}(\alpha) = \begin{cases} \frac{\Gamma(\sum_{k=1}^{d} \alpha_k)}{\prod_{k=1}^{d} \Gamma(\alpha_k)} \prod_{k=1}^{d} \theta_{jk}^{\alpha_k - 1} & \text{if} \quad \theta_{jk} > 0 \,\text{and}\, \sum_{k=1}^{d} \theta_{jk} = 1, \\ 0 & \text{otherwise.} \end{cases}$$

The multinomial-Dirichlet prior induces competitiveness among the factors which helps to better mitigate the issue of factor splitting that occurs in saturated factor models (Geweke and Singleton, 1980). It is worth noting that such dynamics would not be possible to capture by considering only one set of allocation proportions θ shared among the rows in Γ_b. It is also instructive to see how the choice of the concentration parameters affects the probability mass distribution of the Dirichlet prior. The marginal means are proportional to the entries

in α where higher values, generating the same mean vector, increase the concentration of the probability mass. The mode of the Dirichlet distribution is not well defined unless $\alpha_i > 1$. In symmetric distributions ($\alpha_1 = \cdots = \alpha_d$) the values (a) below one (Figure 5.1(a)) smooth away the distribution from the center towards the borders of the simplex, (b) equal to one (Figure 5.1(b)) generate a uniform distribution, (c) above one (Figure 5.1(c)) gravitate the distribution towards the center of the simplex. In the asymmetrical distributions (unequal concentration parameters), the mass is pulled towards a side of the simplex in the direction of the smallest value α_i (Figure 5.1(d)). This suggests that a more dramatic reduction in factor splitting would be encouraged by assuming $\alpha_1 > \cdots > \alpha_d$, inducing also a decreasing trend in the prior means of γ_j.

5.5 EM Algorithm for Sparse Bayesian Factor Regression

Bayesian learning in factor analytic models has relied heavily on the developments in stochastic search methods for posterior exploration (Carvalho et al., 2008; Frühwirth-Schnatter and Lopes, 2009). The computational complexity there is hugely challenged as the number of factors and the dimensionality of the response vector increases. As an alternative to stochastic search, we propose a deterministic approach to finding high posterior probability models associated with modes of the posterior distribution $\pi(A, B, \Sigma, \theta \mid Y)$ using an EM algorithm. The observed data is augmented by the latent factors $\Omega = [\omega_1, \ldots, \omega_n]'$ as well as the latent binary model matrix Γ. The principal ingredient to obtaining a closed form EM algorithm is positioning the variable selection indicators at the level of regression parameters, yielding a convenient hierarchical separability of the prior. Our factor model formulation is reminiscent of the probabilistic principal components approach (Tipping and Bishop, 1999), for which a closed form EM algorithm exists. Recently, Rockova and George (2013) proposed an expeditious EM-based procedure for Bayesian variable selection using continuous spike and slab priors. Augmenting the factor model with continuous mixture priors, we can combine the ingredients of the two algorithms to obtain a closed form EM inferential procedure for factor model exploration. The mathematical formalism of the EM procedure is described below and the model identification based on the maximum a posteriori output is postponed until the next section.

The EM algorithm locates posterior modes by iteratively maximizing the objective function:

$$Q(A, B, \theta, \Sigma) = \mathsf{E}_{\Gamma, \Omega \mid \cdot} \left[\log \pi \left(\underbrace{A, B, \Sigma, \theta}_{\text{unknown parameters}} , \overbrace{\Gamma, \Omega}^{\text{missing data}} \mid \underbrace{Y}_{\text{observed data}} \right) \right],$$

where $\mathsf{E}_{\Gamma, \Omega \mid \cdot}(\cdot)$ denotes the conditional expectation of the latent data, given the observed data and current parameter estimates at the k-th iteration. It is worth noting that due to the

separability of the prior, where the binary factor model matrix depends on the latent factors only through factor loadings, Ω and Γ are conditionally independent.

Denote D_{ja} and D_{jb} the diagonal blocks of the matrix D_j in (5.3.2) that correspond to the regression coefficients and factor loadings, respectively. We can write

$$\begin{aligned} Q(A,B,\ \Sigma,\theta) = \frac{1}{2}\,\mathsf{E}_{\gamma,\Omega|\cdot}\Big[&- \sum_{i=1}^{n}(y_i - Ax_i - B\omega_i)'\Sigma^{-1}(y_i - Ax_i - B\omega_i) \\ &- \sum_{j=1}^{G}||D_{ja}^{-1/2}a_j|| - \sum_{j=1}^{G}||D_{jb}^{-1/2}b_j|| - (n+p+d+\eta)\sum_{j=1}^{G}\log\sigma_j^2 \\ &- \eta\lambda\sum_{j=1}^{G}\frac{1}{\sigma_j^2} + \log\pi(\Gamma\,|\,\theta) + \log\pi(\theta)\Big]. \end{aligned} \tag{5.5.6}$$

For convenience of notation, let $\langle X\rangle$ be the conditional expectation $\mathsf{E}_{\Gamma,\Omega|\cdot}(X)$. The latent variables entering (5.5.6) linearly can be replaced by their conditional expectations. The expectation of the emerging quadratic terms of the latent data is evaluated as $\langle\omega_i\omega_i'\rangle = \langle\omega_i\rangle\langle\omega_i\rangle' + M^{(k)}$, where $M^{(k)} = \mathsf{Var}_{\Gamma,\Omega|\cdot}(\omega_i)$ denotes the conditional covariance matrix of the factor vector. Then we can write

$$\begin{aligned} Q(A,B,\ \Sigma,\theta) = -\frac{1}{2}\Big[&\sum_{i=1}^{n}(y_i - Ax_i - B\langle\omega_i\rangle)'\Sigma^{-1}(y_i - Ax_i - B\langle\omega_i\rangle) \\ &+ \sum_{j=1}^{G}||\langle D_{ja}^{-1/2}\rangle a_j|| + \sum_{j=1}^{G}||\langle D_{jb}^{-1/2}\rangle b_j|| + n\,\mathsf{tr}(B'\Sigma^{-1}BM^{(k)}) \\ &+ (n+p+d+\eta)\sum_{j=1}^{G}\log\sigma_j^2 + \eta\lambda\sum_{j=1}^{G}\frac{1}{\sigma_j^2} + \langle\log\pi(\Gamma\,|\,\theta)\rangle + \log\pi(\theta)\Big], \end{aligned} \tag{5.5.7}$$

where $\mathrm{tr}(\cdot)$ designates the trace of a matrix. The first two rows in (5.5.7) correspond to a penalized likelihood in multivariate regression with a predictor matrix of observed explanators and imputed latent factors and with ridge penalty matrices. To see this, it suffices to replace the row summations by column summations in (5.5.7) to obtain

$$\begin{aligned} Q(A,B,\ \Sigma,\theta) = -\frac{1}{2}\Big[&\sum_{j=1}^{G}\frac{||y^j - Xa_j - \langle\Omega\rangle b_j||}{\sigma_j^2} + \sum_{j=1}^{G}\frac{||\langle D_{ja}\rangle^{-1/2}a_j||}{\sigma_j^2} \\ &+ \sum_{j=1}^{G}\frac{||(\langle D_{jb}\rangle^{-1/2} + \sqrt{n}M^{(k)1/2})b_j||}{\sigma_j^2} \\ &+ (n+p+d+\eta)\sum_{j=1}^{G}\log\sigma_j^2 + \eta\lambda\sum_{j=1}^{G}\frac{1}{\sigma_j^2} + \langle\log\pi(\Gamma\,|\,\theta)\rangle + \log\pi(\theta)\Big], \end{aligned} \tag{5.5.8}$$

where y^j denotes the j-th column in the matrix Y and $\langle\Omega\rangle = [\langle\omega_1\rangle,\ldots,\langle\omega_n\rangle]'$.

5.5.1 Closed Form E-step

The E-step entails computation of the expectations involved in (5.5.8), namely (a) the conditional mean and covariance of the latent factors, (b) the conditional expectation of the diagonal penalty matrices $\langle D_j \rangle = \{(1-\langle \gamma_{jl} \rangle)v_0 + \langle \gamma_{jl} \rangle v_1\}_{l=1}^{p+d}$ and (c) terms involved in $\langle \log \Gamma \mid \theta \rangle$. The expected logarithm of the partially exchangeable matrix priors (5.4.3) and (5.4.4) requires only the computation of marginal inclusion probabilities $\langle \gamma_{jl} \rangle$, which can be obtained in closed form. Assuming (5.4.3), these calculations simplify to

$$\langle \omega_i \rangle = M^{(k)} B^{(k)\prime} \Sigma^{(k)-1} \left(y_i - A^{(k)} x_i \right), \tag{5.5.9}$$

$$M^{(k)} = \left(B^{(k)\prime} \Sigma^{(k)-1} B^{(k)} + \mathrm{I}_d \right)^{-1}, \tag{5.5.10}$$

$$\langle \gamma_{jl} \rangle = \frac{\phi(\beta_{jl}^{(k)}; 0, \sigma_j^{(k)2} v_1)\theta_l^{(k)}}{\phi(\beta_{jl}^{(k)}; 0, \sigma_j^{(k)2} v_1)\theta_l^{(k)} + \phi(\beta_{jl}^{(k)}; 0, \sigma_j^{(k)2} v_0)(1-\theta_l^{(k)})}, \tag{5.5.11}$$

where $\phi(x; 0, \sigma^2)$ denotes the zero mean Gaussian density with variance σ^2 evaluated at x. Under the logistic partially exchangeable prior (5.4.4), the update in (5.5.11) for $1 \leq l \leq p$ replaces $\theta_l^{(k)}$ by the inverse logistic transformation of the current estimate of the linear predictor at the k-th iteration. Under the multinomial-dirichlet prior, the conditional expectation of the factor indicators for $p+1 \leq l \leq p+d$ uses updates (5.5.11) with $\theta_l^{(k)}$ replaced by $\theta_{jl}^{(k)}$.

5.5.2 Closed Form M-step

The M-step requires computation of ridge regression solutions with penalties induced by the posterior averaged spike and slab precisions and the covariance matrix of the latent factors. For the simplicity of exposition, denote $D_j^\star = \begin{pmatrix} 0 & 0 \\ 0 & M^{(k)} \end{pmatrix} + \langle D_j \rangle$. Assuming (5.4.3) with the beta-binomial prior on the inclusion probabilities, the M-step for $j > d$ consists of updates:

$$\beta_j^{(k+1)} = \left([X, \langle \Omega \rangle]' [X, \langle \Omega \rangle] + D_j^\star \right)^{-1} [X, \langle \Omega \rangle]' y^j, \tag{5.5.12}$$

$$\sigma_j^{(k+1)} = \sqrt{\frac{||y^j - [X, \langle \Omega \rangle]\beta_j^{(k+1)}|| + ||D_j^{\star 1/2} \beta_j^{(k+1)}|| + v\lambda}{n+p+d+v}}, \tag{5.5.13}$$

$$\theta_l^{(k+1)} = \frac{\sum_{j=1}^G \langle \gamma_{jl} \rangle + a - 1}{a+b+G-2}. \tag{5.5.14}$$

For $j \leq d$, each vector $\beta_j^{(k+1)}$ is confined by the lower triangular structure in the factor loadings. The updates again require a ridge regression solution, only with a subset predictor matrix and a modified response after regressing out the factor with the unit diagonal element in B.

The updates $\sigma_j^{(k+1)}$ then change correspondingly. In the instance that $p+d > n$, the ridge regression solutions (5.5.12) can be obtained more efficiently by the Woodburry-Sherman matrix inversion trick. If both n and $p+d$ are formidably large, fast approximate solutions can be obtained with the assistance of dual coordinate ascent methods (George et al. (2013), Tong et al. (2012)). For the logistic partially exchangeable prior (5.4.4) with the inverse logistic beta prior on θ, the step (5.5.14) is replaced by the maximization of

$$\sum_{j=1}^{g}\sum_{k=1}^{p}\Big\{\langle\gamma_{jk}\rangle\theta' Z_{G\times(j-1)+k} - \log\Big[1+\exp(\theta' Z_{G\times(j-1)+k})\Big]\Big\} + a1'\theta - (a+b)\log[1+\exp(1+1'\theta)],$$

which admits closed form solutions unless the groups overlap, in which case routine optimization methods can be used. The Dirichlet-multinomial prior on the factor loadings leads for $p+1 \le l \le p+d$ to closed form updates of elements in $\Theta^{(k)}$, namely

$$\theta_{jl}^{(k)} = \frac{\langle\gamma_{jl}\rangle + \alpha_l - 1}{\sum_{l=1}^{d}(\gamma_{jl}+\alpha_l) - d}. \tag{5.5.15}$$

5.6 Factor Model Exploration and Evaluation

In our perspective, model building of factor analytic models entails two related modeling decisions: (a) determining the effective factor dimensionality, (b) allocating zeroes in the matrix of factor loadings. The first decision is typically resolved by fitting the factor model for different d and performing comparisons by some model selection criterion. Alternatively, inference about factor cardinality can be facilitated by trans-dimensional inferential algorithms (Lopes and West, 2004; Bhattacharya and Dunson, 2011). Similarly as Frühwirth-Schnatter and Lopes (2009), inference on the effective dimension will be anchored by the pattern of sparsity in the factor loading matrix, induced by structured priors under identifiability constraints. In way too generous factor models, the overly sparse loading columns serve as an indicator for factor reduction.

5.6.1 Recovering Sparsity

If we were to consider the factor regression model (5.2.1) with a presumed factor dimensionality d, we could exploit the EM algorithm to detect suitable candidate models. Different choices of tuning parameters would likely lead to different candidates. Turning to the issue of calibration, it is sensible to keep the slab variance v_1 fixed to a large value or to treat it as unknown and impose a prior. The choice of the spike variances v_0 is far more consequential. Varying the spike variance v_0 changes dramatically the character of the posterior landscape (Rockova and George, 2013). Small values v_0 encourage concentration of the posterior around

sharp isolated peaks of local maxima. Large values v_0 induce a wide posterior spike around zero, reducing multimodality by swallowing local modes associated with small regression coefficients. In order to capture the dynamics of such posterior mass reallocation, we consider a grid of values $v_0 \in V$ and run the EM algorithm for each v_0 to generate a set of candidate models $\{\Gamma^d_{v_0} = [\Gamma^d_{v_0 a}, \Gamma^d_{v_0 b}] : v_0 \in V\}$. Gradually increasing v_0 leads inevitably to model sparsification and provides an accumulation of evidence for the truly important associations.

Each model configuration $\Gamma^d_{v_0}$ can be retrieved from the maximum-a-posteriori (MAP) matrices $\widehat{A}^d_{v_0}$ and $\widehat{B}^d_{v_0}$ by thresholding the individual entries that are small in magnitude. By virtue of selective shrinkage, induced by the spike and slab prior, unimportant coefficients are pulled towards zero, deactivating the binary inclusion indicators. In contrast where the slab has dominated the posterior, the MAP coefficients will be large and instead activate the binary indicators. Although not actually sparse, the MAP estimates serve a useful prerequisite for identifying the associated high-probability posterior model. As a by-product of the EM algorithm, we obtain the matrix $\langle\Gamma\rangle^d_{v_0} = [\langle\Gamma^d_{v_0 a}\rangle, \langle\Gamma^d_{v_0 b}\rangle]$ of the conditional inclusion probabilities $\mathsf{P}(\gamma_{jk} = 1 \,|\, \widehat{A}^d_{v_0}, \widehat{B}^d_{v_0}, \widehat{\Sigma}^d_{v_0}, Y)$ given the MAP estimates and the observed data. A natural way to locate the high probability model is by screening out the entries $\langle\Gamma\rangle^d_{v_0}$ that are below a selection threshold 0.5. This corresponds to a local variant of the median posterior model rule (Barbieri and Berger, 2004). As explained in Rockova and George (2013), this rule is equivalent to thresholding the MAP estimates $\widehat{A}^d_{v_0}$ and $\widehat{B}^d_{v_0}$ based on the intersection point between the two densities in the posterior weighted spike and slab mixture.

The median probability model rule does not guarantee the orthogonal factor model matrix $\Gamma^d_{v_0 b}$ even under the multinomial-Dirichlet prior, because the rows in $\langle\Gamma^d_{v_0 b}\rangle$ do not sum to one. Using the Dirichlet prior, the rows in $\Theta^{(k)}$ do sum to one and allow only one entry above 0.5. This motivates an alternative practical guidance to identify orthogonal factor model by thresholding the matrix $\Theta^{(k)}$.

We contemplate that the practitioner would prefer to run the EM procedure for a set of factor dimensionalities. The EM algorithm will presumably generate more interesting models, when the assumed factor cardinality is close to the effective dimensionality. Having obtained a series of candidates, the lingering issue remains how to effectively distinguish between models identified along the regularization path $v_0 \in V$ and how to perform inference on the number of necessary factors.

5.6.2 Trans-dimensional Model Comparisons

In our Bayesian approach to factor augmented multivariate regression, it is the EM exploratory algorithm that enables us to rapidly elicit suitable candidate models. Under the order-inducing identifiability constraints, every lower-dimensional factor model can be embedded within a richer factor model by augmenting the factor loading matrix with additional zero columns to fill in the dimensionality gap. Thereby models with different factor dimensionalities can be set on an equal footing for model evaluation, for instance by the posterior model probability

$\mathsf{P}(\Gamma \,|\, Y)$. Whereas the continuous spike distribution ($v_0 > 0$) is key to implementing the model discovery, the "objective" model evaluation (irrespective of the choice $v_0 > 0$) will be based on the assumption that $v_0 = 0$. It is the point mass spike that in conjugate linear regression models leads to a closed form expression for the posterior model probability, up to a norming constant (George and McCulloch, 1997).

In the multivariate regression with uncorrelated responses, the posterior model probability simplifies to an independent product of individual model probabilities for each regression. In the latent factor model, the posterior model evaluation is far more challenging where tractable closed forms are no longer available. A useful approximation can be obtained through Monte Carlo integration, based on the integral representation

$$\mathsf{P}(\Gamma \,|\, Y) \propto \mathsf{P}(\Gamma) \int \mathsf{P}(Y \,|\, \Omega, \Gamma) \pi(\Omega \,|\, \Gamma) \mathrm{d}\Omega \tag{5.6.16}$$

$$\propto \mathsf{P}(\Gamma) \int \mathsf{P}(Y \,|\, \Omega, \Gamma) \pi(\Omega) \mathrm{d}\Omega. \tag{5.6.17}$$

The identity (5.6.17) follows from (5.6.16) by the argument of hierarchical separability, where the factor selection indicators depend on the latent data only through the factor loadings. Once we knew the latent factors, we could employ the standard computation to evaluate conditional posterior probabilities. This motivates our investigation of approximations to (5.6.17) by means of an empirical average, given a finite set of draws from the prior distribution $\pi(\Omega)$. We define the approximated criterion for model comparison as

$$\overline{G}(\Gamma) = \frac{\mathsf{P}(\Gamma)}{M} \sum_{m=1}^{M} \mathsf{P}(Y \,|\, \Omega_m, \Gamma) \pi(\Omega_m). \tag{5.6.18}$$

This strategy, however, increases the computational complexity, which may be formidable if the cardinality of the factor space is large. Alternative schemes can be obtained based on the integral representation

$$\mathsf{P}(\Gamma \,|\, Y) = \int \mathsf{P}(\Gamma \,|\, \Omega, Y) \mathsf{P}(\Omega \,|\, Y) \mathrm{d}\Omega. \tag{5.6.19}$$

Replacing the intractable distribution $\mathsf{P}(\Omega \,|\, Y)$ by the conditional distribution given the MAP estimates $\mathsf{P}(\Omega \,|\, \widehat{A}, \widehat{B}, \widehat{\Sigma}, Y)$, which is tractable and easy to sample from, we could employ the Monte Carlo integration similarly to what was done previously. Ultimately, we propose an alternative surrogate criterion, which follows from (5.6.19) by replacing $\mathsf{P}(\Omega \,|\, Y)$ with a dirac measure concentrated at $\langle \Omega \rangle$, given the MAP estimates and the observed data. This leads to a rapidly computable criterion

$$\widetilde{G}(\Gamma) = \mathsf{P}(\Gamma) \mathsf{P}(Y \,|\, \langle \Omega \rangle, \Gamma) \propto \mathsf{P}(\Gamma \,|\, \langle \Omega \rangle, Y). \tag{5.6.20}$$

As will be demonstrated on simulated data, this conditional criterion can be used to perform (trans-dimensional) model comparisons effectively and also efficiently.

5.7 The EM Strategy for Factor Model Selection

In this section we exemplify the EM strategy for factor model selection and explore empirically its performance on a simple simulated data set. The performance is evaluated by three metrics: the ability (a) to detect sets of predictors relevant for multiple responses, (b) to determine effective factor dimensionality, (c) to recover underlying latent structure.

Here we analyze a synthetic data set consisting of $n = 100$ observations generated from the factor regression model (5.2.1), assuming $G = 200$ responses, $p = 20$ predictors and $d = 5$ generating factors. The rows of the regression matrix X were drawn independently from $\mathcal{N}_{20}(0, \mathrm{I}_{20})$ and the latent factors were obtained through random sampling from $\mathcal{N}_5(0, \mathrm{I}_5)$. We assume homoscedastic residual variances $\sigma_1^2 = \cdots = \sigma_G^2 = 4$. The nonzero entries in the regression matrix $A = \mathrm{I}_5 \otimes 1_{40\times 4}$ are placed in blocks along the main diagonal, where $\otimes$ denotes the Kronecker matrix product and $1_{40\times 4}$ is a (40×4) matrix of ones. We begin by assuming that each gene is driven by only one underlying factor and generate a block-wise matrix $B = \mathrm{I}_5 \otimes 1_{40}$, where 1_{40} is a (40×1) vector of ones. For the identifiability, we then set the diagonal elements $\{b_{kk}\}_{k=1}^d$ to one. Throughout the course of this section, the EM procedure will be initialized with unit elements in $B^{(0)}$ and $\sigma^{(0)}$. The starting values $A^{(0)}$ for the matrix of regression coefficients are for given v_0 and v_1 generated as row-wise ridge regression solutions corresponding to the limiting case of deterministic annealing (Rockova and George, 2013)

$$A^{(0)} = Y'X\left(X'X + \frac{v_0 + v_1}{2v_0v_1}\right)^{-1}.$$

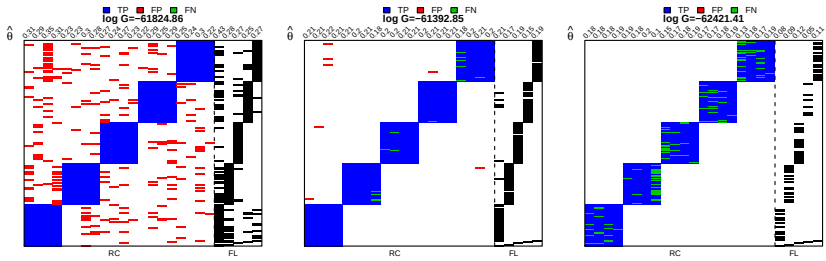
The data is pre-processed by centering columns in X and Y around zero and scaling X to be within a unit variance range.

■ 5.7.1 Model Exploration

First, we implement the EM algorithm to generate candidate models considering various $v_0 > 0$. We begin with the vanilla column-wise exchangeable prior on the model matrix. Later, we demonstrate the benefits of the Dirichlet-multinomial formulation. The model exploration is performed under three scenarios, where the assumed factor dimensionality is (a) correct ($d = 5$), (b) under-determined ($d = 3$), or (c) over-determined ($d = 7$). The model elicitation is followed by model evaluation by the surrogate posterior model probability criterion (5.5.8) based on $v_0 = 0$.

5.7.1.1 □ Column-wise Partially Exchangeable Prior

We begin by considering the partially exchangeable model prior (5.4.3), where the indicators in Γ are Bernoulli trials with inclusion probabilities $\theta = (\theta_1, \ldots, \theta_{p+d})'$ that are individual for each column. Each θ_l is assigned a Beta prior $\mathcal{B}(1,1)$. We proceed by setting $v_1 = 100$



(a) $v_0 = 0.01$ (b) $v_0 = 0.03$ (c) $v_0 = 0.05$

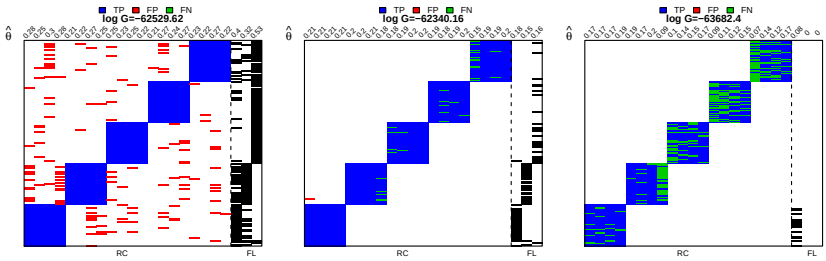
Figure 5.2: Estimated patterns of zeroes in the model matrix Γ for $d = 5$

and considering spike variances $v_0 \in V = \{0.01, 0.03, 0.05\}$. For every $v_0 \in V$ and $d \in D = \{3, 5, 7\}$ we run the EM exploratory procedure to elicit a candidate model $\Gamma^d_{v_0}$ by the local median probability model rule.

We first assume that the true dimensionality of the latent factors is known in an advance (i.e. $d = 5$). We run the EM algorithm individually for each $v_0 \in \{0.01, 0.03, 0.05\}$ and obtain model configurations portrayed in Figure 5.2. The blank entries in the matrix correspond to zeroes, whereas nonzero values are filled with color. In the panel for regression coefficients, blue denotes the true positives whereas green and red denote false negatives and false positives, respectively. Increasing the spike variance, the EM algorithm generates sparser models with fewer false positive findings. This reduction in model complexity comes at the expense of impaired ability to detect signal, particularly in the latent structure. The estimated coefficients $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_{p+d})'$, superimposed on top of each column, report close to the true proportion of responses affected by each individual variable/factor.

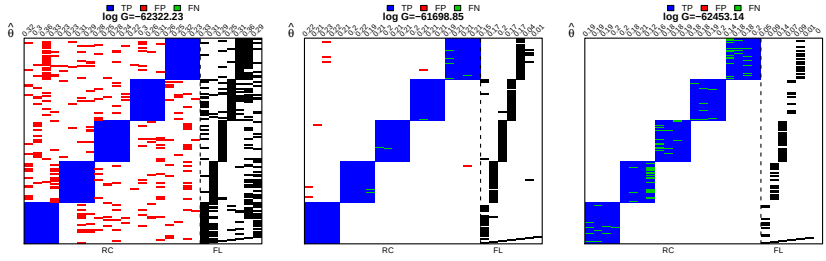
It is instructive to investigate, how the searching algorithm performs when the factor dimensionality is under-determined. We set $d = 3$ and repeat the calculations with the same settings as before. The results depicted in Figure 5.3 confirm that by considering a larger spike, the EM is unable to capture the latent structure. Smaller values, on the other hand, lead to factor merging.

Finally, we perform the EM exploration when there are too many assumed factors by setting $d = 7$. The triplet of models, displayed in Figure 5.4, demonstrates that in an over-determined model with a small spike variance, the activity of what was supposed to be a single latent variable is distributed among multiple columns, a phenomenon called factor splitting. Larger spikes again fail to detect the latent signal. The close to zero estimates $\widehat{\theta}$ associated with the last two latent factors in $\Gamma^7_{0.03}$ and $\Gamma^7_{0.05}$ suggest factor reduction. In fact, the model $\Gamma^7_{0.05}$ has only one nonzero entry in the last two loading columns, which implies that the effective dimensionality is essentially 5.



(a) $v_0 = 0.01$ (b) $v_0 = 0.03$ (c) $v_0 = 0.05$

Figure 5.3: Estimated patterns of zeroes in the model matrix Γ for $d = 3$



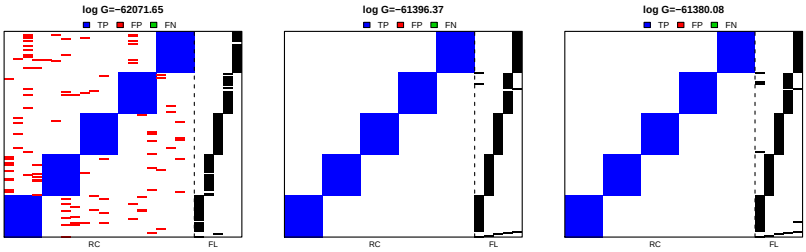
(a) $v_0 = 0.01$ (b) $v_0 = 0.03$ (c) $v_0 = 0.05$

Figure 5.4: Estimated patterns of zeroes in the model matrix Γ for $d = 7$

5.7.1.2 □ Dirichlet-Multinomial Prior

The two undesirable phenomena emerging in the previous analysis are (a) factor splitting in an over-determined model and (b) the inability to capture the latent signal for larger values of the parameter v_0. These observations support the argument that the loadings and regression coefficients should be treated individually. In the revised analysis, we will consider the row-wise multinomial-Dirichlet prior (5.4.5) on the loading inclusion matrix. The allocation of zeroes in the factor model matrix is based on the 0.5 thresholded maximum-a-posteriori matrix $\widehat{\Theta}$.

Regarding the prior on the regression model matrix, we exploit the fact that the predictors operate in clusters to predict groups of related responses. In our example, there are 5 groups of predictive coefficients placed in blocks along the main diagonal. The remaining coefficients



(a) $v_0 = 0.01$ (b) $v_0 = 0.03$ (c) $v_0 = 0.05$

Figure 5.5: Estimated patterns of zeroes in the model matrix Γ for $\widetilde{d} = 5$ and structured model prior

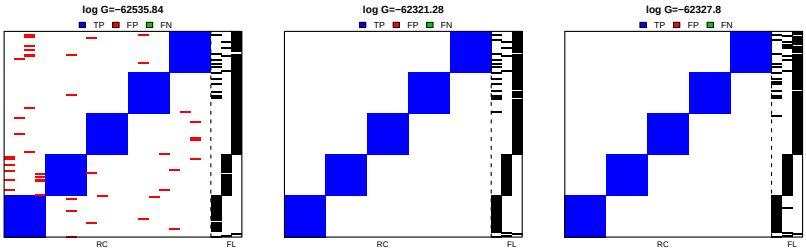
are deemed to belong to a single non-predictive group. Assuming that all the within-group indicators share the same inclusion probability, we consider the partially exchangeable prior (5.4.4) driven by this grouping. We expect that inducing the dependence within the matrix of binary indicators, the EM will generate candidates that better correspond to the true generating model.

Once again we perform the exploration under the three scenarios $d \in \{3,5,7\}$. It is worth noting that for $d = 5$, the structured prior with suitably chosen v_0 leads to the identification of all nonzero regression elements without any false positives (Figure 5.5). The identified latent structure there is nearly identical to the generating model. In the under-determined model (Figure 5.6), the factor merging is even more encouraged, leading to models with a richer latent structure as compared to the previous analysis (Figure 5.3). In the over-determined case (Figure 5.7), the factor splitting is mitigated by performing the factor selection based on the matrix $\widehat{\Theta}$. The last two factors hardly contribute to the explanation of responses, which suggests that it might be preferable to ignore them.

The goal of these exploratory analyses has been to demonstrate that the EM algorithm finds better candidate models when the factor dimensionality is close to the truth, and that the sparsity pattern in factor loadings provides evidence complimentary to the $\widetilde{G}$-function in determining the convenient dimensionality of the factor model. Rigorous model evaluation is presented in the next section.

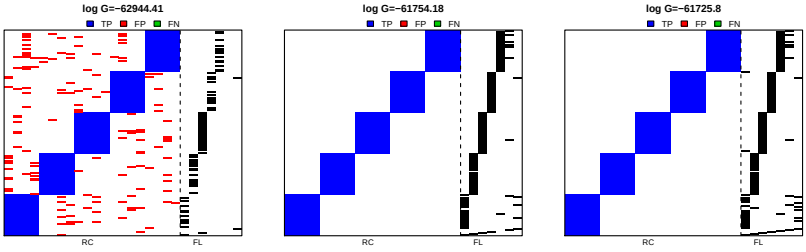
5.7.2 Model Evaluation

The recommended practical guidance for comparing identified candidate models is based on the surrogate posterior model probability criterion (5.6.20). The values $\widetilde{G}(\cdot)$ for all detected models from the previous section are tabulated in Table 5.7.2 together with information on the computational time and iteration history. The EM searching machanism is seen to locate better candidate models when $d = 5$. These models exhibit a good compromise between



(a) $v_0 = 0.01$ (b) $v_0 = 0.03$ (c) $v_0 = 0.05$

Figure 5.6: Estimated patterns of zeroes in the model matrix Γ for $d = 3$ and structured model prior



(a) $v_0 = 0.01$ (b) $v_0 = 0.03$ (c) $v_0 = 0.05$

Figure 5.7: Estimated patterns of zeroes in the model matrix Γ for $d = 7$ and structured model prior

false negatives and positives, which is also supported by higher values of $\widetilde{G}$. The search can be improved when guided by the multinomial-Dirichlet prior and correct grouping, where the best model is the one with zero false positives and negatives in the regression structure and nearly accurate latent structure. This model also has the most favorable value of the $\widetilde{G}$-function among all considered models, which supports the evidence that this criterion can be used to effectively and efficiently perform comparisons between factor models.

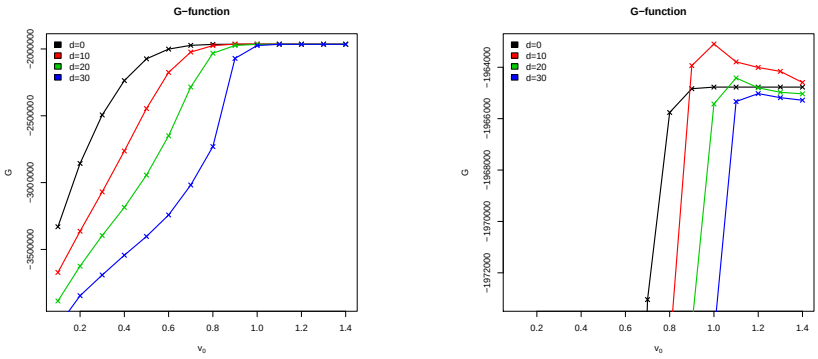
	$d=5$		$d=3$		$d=7$	
	(a)	**(b)**	**(a)**	**(b)**	**(a)**	**(b)**
			$v_0 = 0.01$			
$\widetilde{G}$**-function**	-61824.86	-62071.65	-62529.62	-62535.84	-62322.23	-62944.41
#iterations	27	27	21	21	31	31
CPU(sec)	3.15	8.29	2.7	7.42	4.31	9.86
			$v_0 = 0.03$			
$\widetilde{G}$**-function**	-61392.85	-61396.37	-62340.16	-62321.28	-61698.85	-61754.18
#iterations	27	27	21	21	31	31
CPU(sec)	3.14	8.18	2.7	6.99	4.31	10.31
			$v_0 = 0.05$			
$\widetilde{G}$**-function**	-62421.41	**-61380.08**	-63682.4	-62327.8	-62453.14	-61725.8
#iterations	27	27	21	21	31	31
CPU(sec)	3.1	9.62	2.69	7.36	4.28	10.4

Table 5.1: (a) EM search with column-wise partial exchangeability, (b) EM search with multinomial-dirichlet prior and partially exchangeable prior with correct grouping; the Table reports values of the $\widetilde{G}$-function as well as the number of iterations until convergence and the computational time in seconds on a 3.0 GHz processor desktop computer using a *R* implementation; the largest value of the $\widetilde{G}$-function marked in bold font

5.8 AML MicroRNA Regulatory Network

Acute myeloid leukemia (AML) describes a heterogeneous group of hematopoietic disorders, characterized by the proliferation of immature progenitors that have lost their ability to differentiate into functional myeloid cells. Past decades of an intensive biomedical research have accumulated a large body of evidence for the multifactorial pathogenesis of AML. The multiple contributing factors engage molecular mechanisms as diverse as epigenetic alterations, cytogenetic abnormalities and other genetic aberrations leading to impaired expression of oncogenic genes. Recent studies have also begun associating microRNAs with specific AML regulatory mechanisms (Jongen-Lavrencic et al., 2008; Sun et al., 2013a). MicroRNAs are negative regulators of gene expression, decreasing the stability of target RNAs or limiting their translation (Fabian et al., 2010). Assuming that microRNAs disrupt the gene expression homeostasis associated with the normal hematopoiesis, we set out to detect an “active" set of microRNA’s whose elevated levels imply a modulation of gene expression. Our expression dataset provides snapshot measurements of the gene and microRNA levels in a heterogeneous group of 212 patients diagnosed with AML.

The 212 AML samples, collected at the department of hematology at Erasmus Medical Center in Rotterdam, were analyzed for (1) the expression of $M = 256$ microRNA’s using real-time quantitative PCR, as described previously (Jongen-Lavrencic et al., 2008), (2) gene expression using high-density Affymetrix arrays, as described previously (Valk et al., 2004). The gene expression dataset (online access `http://www.ncbi.nlm.nih.gov/geo/` with the accession GSE6891) was normalized using the RMA methods (Irizarry et al., 2003).



(a) $\log(\widetilde{G}) \sim v_0$ (b) Zoomed $\log(\widetilde{G}) \sim v_0$

Figure 5.8: Evolution plot of the $\log(\widetilde{G})$-function for the different factor cardinalities.

Only $G = 4\,245$ genes with a known biological function, that is genes involved in at least one pathway recognized by the KEGG database, were analyzed. The columns in Y were further standardized to have a mean zero and a variance one. The raw microRNA expression values were obtained using the relative quantification method ($2^{-\Delta C_t}$ values with respect to an endogenous control), further log-transformed and again standardized. We select a subset of $p = 177$ microRNA's which have at least 50 observations above a lower detection limit.

In the EM exploration algorithm, we implement the homogeneous Dirichlet-multinomial prior with a concentration parameter $\alpha = 1.01$. The prior on the regression model matrix assumes the column-wise exchangeability. This is because we want to quantify the relevance of each individual microRNA in jointly predicting the multiple responses. This assessment is facilitated by the MAP estimates $\widehat{\theta}$, which correlate with the proportion of genes affected by each microRNA. Instead of the column-wise exchangeable prior, we could consider more elaborate structures that relate inclusion probabilities of microRNA's with a similar set of binding elements (e.g. using the Hamming distance as in Li and Zhang (2010)). Alternatively, Stingo et al. (2010) incorporated information from multiple microRNA target prediction algorithms, which induce a selection advantage for the predicted associations. Their prior formulation is potentially better equipped to guide the search for putative microRNA targets. In our analysis, however, we are not searching for the direct microRNA targets. Such task is hugely challenged by the heterogeneity of the AML regulatory mechanisms and the unavailability of temporal data in order to capture the dynamics of the microRNA-mediated regulation. We merely explore important AML-related microRNAs and infer their likely biological function.

The EM exploration is initialized with the same choice of starting values as in our simulated examples. We proceed by considering a sequence of spike variance parameters $v_0 \in \{0.1 + k \times 0.1; k = 0, \ldots, 13\}$ and we set $v_1 = 100$ for the model exploration as well as eval-

	$d=0$			$d=10$			$d=20$			$d=30$		
v_0	$\log(\widetilde{G})$	**#iter**	**#nonzero**	$\log(\widetilde{G})$	**#iter**	**#nonzero**	$\log(\widetilde{G})$	**#iter**	**#nonzero**	$\log(\widetilde{G})$	**#iter**	**#nonzero**
0.1	-3330298	4	314299	-3672100	6	462762	-3885701	7	531544	-4074521	5	587690
0.2	-2856687	5	179277	-3363892	6	338616	-3625152	7	432136	-3846295	5	507374
0.3	-2493883	6	97756	-3069493	7	244369	-3396171	7	348275	-3691427	7	453435
0.4	-2236896	7	47007	-2763675	9	164463	-3185536	9	280864	-3543189	8	402877
0.5	-2073895	8	17816	-2446249	11	93287	-2944016	11	213175	-3402932	9	357133
0.6	-2001167	7	5773	-2175522	17	39043	-2649743	14	140400	-3242056	14	307975
0.7	-1973042	6	1320	-2023462	17	11366	-2285081	16	61490	-3018197	15	244364
0.8	-1965760	18	181	-1974979	18	3054	-2032146	19	12625	-2730665	19	170183
0.9	-1964834	16	24	-1963936	20	1121	-1974093	23	2556	-2070711	37	19395
1	-1964772	19	1	-1963096	20	581	-1965429	26	498	-1974473	33	2263
1.1	-1964772	10	0	-1963792	18	266	-1964415	15	229	-1965334	39	413
1.2	-1964772	7	0	-1964008	11	186	-1964800	11	100	-1965025	25	109
1.3	-1964772	7	0	-1964161	11	129	-1964973	14	54	-1965183	17	69
1.4	-1964772	8	0	-1964592	11	66	-1965035	11	43	-1965283	15	52

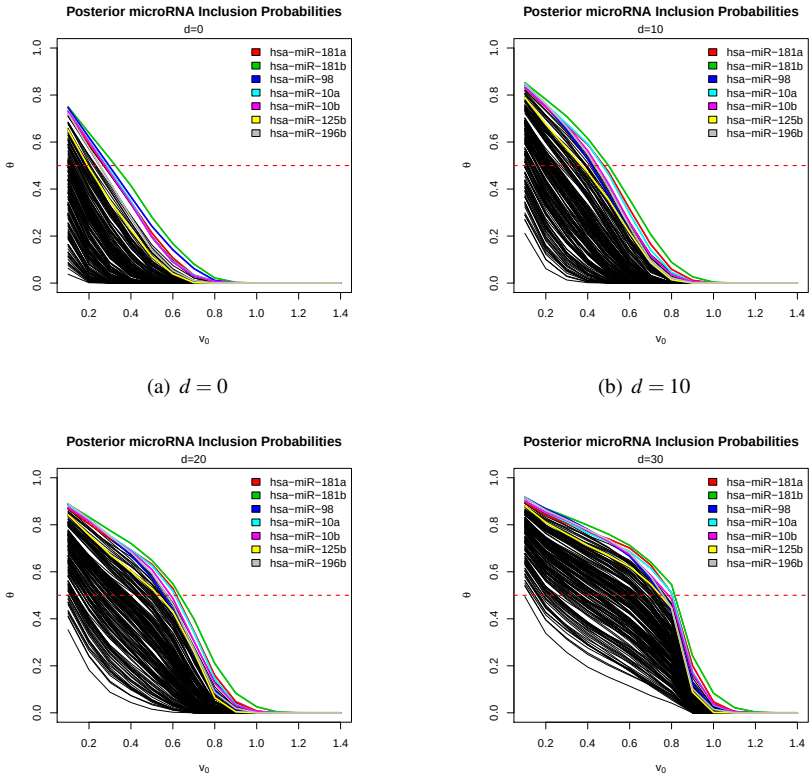
Table 5.2: Models selected along the regularization path considering $v_0 \in \{0.1 + k \times 0.1; k = 0, \ldots, 13\}$. The table reports the logarithm of the $\widetilde{G}$-function, the number of iterations and the number of the nonzero entries in Γ. Computational time for one iteration in seconds on a 3.0 GHz processor desktop computer using a *R* implementation is: 62.01 ($d=0$), 68.33 ($d=10$), 74.71 ($d=20$), 82.25 (d=30).

uation . We begin by considering the plain multivariate regression model with uncorrelated responses ($d = 0$) and proceed by fitting the factor regression model with $d \in \{10, 20, 30\}$. Model comparisons are performed by the (conditional) posterior probability criterion $\widetilde{G}(\cdot)$. The nonzero elements are selected by the local median probability model rule in the regression matrix and by thresholding the matrix $\widehat{\Theta}$ in the loading matrix.

The results are summarized in Table 5.2, which reports the log-values of the $\widetilde{G}$-function together with the iteration history, the computational time and the number of nonzero elements in each model matrix Γ. The evolution plot of the $\log(\widetilde{G})$ for an increasing spike variance is for all the considered factor dimensionalities depicted in Figure 5.8. One observation to be made is that up until $v_0 = 0.7$ the EM algorithm assuming $d = 0$ generates models with the highest $\log(\widetilde{G})$ values. Towards the end of the regularization path, the $\log(\widetilde{G})$ values are dominated by the factor models associated with $d = 20$.

Turning to the assessment of the relevance of individual microRNAs in modulating gene expression, the supporting evidence is aggregated in the MAP posterior estimates $\widehat{\theta}$. The evolution plots of these estimates in relation to the spike variance are depicted on Figure 5.9, where the upper curves correspond to the more relevant microRNAs. Among the top rated candidates, we consistently identify members of the `miR-181`, `miR-10` or `miR-125` families. The `miR-181` family has been shown to be associated with a favorable outcome in cytogenetically normal AML patients (Li et al., 2012b), where the up-regulation has been hypothesized to correlate with an acquisition of the *CEBPA* mutation (Marcucci et al., 2011), which associates with a milder course of AML. There has also been an experimental evidence suggesting that `miR-181b` promotes apoptosis and inhibits proliferation of leukemic cells. Among the other identified relevant microRNAs we found the `miR-10` family, which is implicated in malignant transformations across a wide range of tissues. Recently, `miR-10a` has been associated with a nucleophosmin mutation *NPM*1 (Bryant et al., 2012), which is a positive prognostic factor in AML Another top ranked candidate recurrent in our analysis is `miR-125b`, which has been shown to be implicated in specific chromosomal translocations leading to AML (Sun et al., 2013b; Bousquet et al., 2008). Among the most influential microRNAs we identified as well `miR-196b`, whose overexpression has been associated with aggressive leukemia in mice and poor prognosis in acute myeloid leukaemia (AML) patients (Li et al., 2012a). Although the evolution curve of `miR-98` stands out as very influential, the role of this microRNA in the context of AML has largely remained unknown.

Turning to the interpretation of the association network, we select one of the candidates which is associated with a high value of the $\log(\widetilde{G})$-function in Table 5.2 and which is rich enough to perform a sensible knowledge extraction. We select a model $\Gamma^{10}_{0.9}$ (depicted in Figure 5.10), where we identified 227 microRNA-gene associations involving 23 microRNA (6 previously established ones) and 204 genes. The list of genes associated with 4 top represented microRNAs is listed in Table 5.3. Focusing on the `miR-181` family, we identified several genes associated with `miR-181b` that are involved in cancer pathways, cell apoptosis, hematopoietic cell lineage, or both acute and myeloid leukemias (such as `CDK6, ZBTB16, IL3RA, PAK2`). Several genes associated with `miR-181a` in our model have been found



(a) $d = 0$ (b) $d = 10$

(c) $d = 20$ (d) $d = 30$

Figure 5.9: Evolution plots of $\widehat{\theta}$ for the different factor dimensionalities

to relate to the function of lymphoid blood cells and the cell cycle (such as `EDC4, NCF4, PPP3CA, MAD2L1, MAD1L1`). Apart from the gene-microRNA interactions, we extracted 10 gene clusters that share an underlying latent factor and thereby potentially relate to a similar biological function. Genes that cluster according to the three most frequent factors (`Factor 2`, `Factor 4` and `Factor 8`) are listed in Table 5.4 in the Appendix. Many of the genes involved with `Factor 2` were found to be associated with RNA degradation and Huntington disease. The two dominant pathways in `Factor 4` were phagosome and osteoclast differentiation. `Factor 8` clustered many genes involved in the MAPK signaling pathway. The MAPK's are a family of proteins that play an essential role in connecting cell-surface receptors

mir-181b								
ABI2	ADCY4	ALDH1A2	ALDH2	B3GALTL	BACE2	C1QC	CACNG4[12]	CD27
CDK6[1,4]	CKB	CMPK2	COQ3	DGAT1	DKK2	DMD	DNAJC1	F11R
F13A1	FECH	FGF9[4,12]	GLRB	HIST1H2BC	HIST3H2A	HLA-DQB1[5]	HYAL2	IL17RB
IL1RAP[3]	IL3RA[3,6]	IRGM	ITPR3	MAP3K13[12]	NKX3-1[4]	NOX3	NUP93	OPRL1
PAK2[4,7,12]	PIGG	PLXNB2	PSTPIP1	SGSH	SLA	SLC27A1	SNRPC	SORD
TFDP1	THBS3	TNFSF11	TUBAL3	TXNRD2	VDAC1	YARS	ZBTB16[2,4]	HSPA5
BID[3,4,9]	DCP1B[7]	HYAL3	RBL2	ROCK2	SIAH1	TAF4	TAF5	CFD
UBL5	MDN1	NOD2	TLR8	ABAT	ACOX3	EPS15	GALNT7	PIK3C2A
AP2B1	HARS	NUDT9	PRDX6	SRP19	SNCA	C1QA	C1QB	IL10RA
CTSO								
mir-181a								
ADA[13]	ALAS2	CBR1	CCDC12	EDC4[7]	ENTPD4	FBXW7	FDFT1	IQSEC1
MDH1	MLYCD	MPO	MRE11A	NCF4[10]	PIGH	PPP3CA[3,8,9,12]	SQLE	TPO[6]
DCI	POP4	PSMA2	MAD2L1[11]	NANP	FCGR2A	PKM2	HSD17B7	PAPOLG
COQ5	MUTYH	WDR61[7]	KCNMB4	MAD1L1[11]	UTP6	SLC6A19	UBA3	
mir-10a								
EFNA1	HIST1H2BB	PRKG2	STT3A	NCF2[10]	VCAN	SNRNP40	EPX	ITGA7
ITGA9	P2RX7	PAPSS1	STK36[4]	STS	TRAC	XCL1	IL17RA	
mir-10b								
GNAI1	SLC38A2	SOCS5	TJP2	TRAF5[5]	SUV420H1	EPX	ITGA7	ITGA9
P2RX7	PAPSS1	STK36[4]	STS	TRAC	XCL1	IL17RA	CTSO	

Table 5.3: (1) chronic myeloid leukemia, (2) acute myeloid leukemia, (3) apoptosis, (4) pathways in cancer, (5) allograft rejection, (6) hematopoietic cell lineage, (7) RNA degradation, (8) T cell receptor signaling pathway, (9) natural killer cell mediated cytotoxicity, (10) leukocyte transendothelial migration, (11) cell cycle, (12) MAPK signaling pathway, (13) primary immunodeficiency

to changes in transcriptional programs. Over the last decade, extensive work has established that these proteins play critical roles in the regulation of a wide variety of cellular processes including cell growth, migration, proliferation, differentiation, and survival. Correct regulation of MAPK signaling is hypothesized to be essential in the regulation of multiple processes involved in hematopoiesis, where aberrant MAPK activation can lead to pathogenesis of various myeloid malignancies (Geest and Coffer, 2009).

The heterogeneity of the biological processes underlying AML in the diverse group of analyzed patients precludes the interpretation of the microRNA-gene interactions in terms of direct targets. Such conclusions would be better obtained with data generated by carefully controlled knock-out time course experiments. However, the disadvantage there is typically a limited amount of measured samples which would preclude advanced statistical modeling. In our approach, we believe that by borrowing the strength among the multiple patients, we were able to identify microRNA's that are influential in common processes underlying the course of AML. We have been able to learn about the likely target gene groups that might provide some useful biological insights about the biological function of these microRNAs.

5.9 Discussion

High-dimensional Bayesian factor modeling has often been challenged by the practicality of MCMC implementations as well as by the inference on the unknown number of latent factors. In the presented work, we propose a rapid deterministic factor model exploratory tool that leverages the existing EM inferential procedure of Rockova and George (2013) for vari-

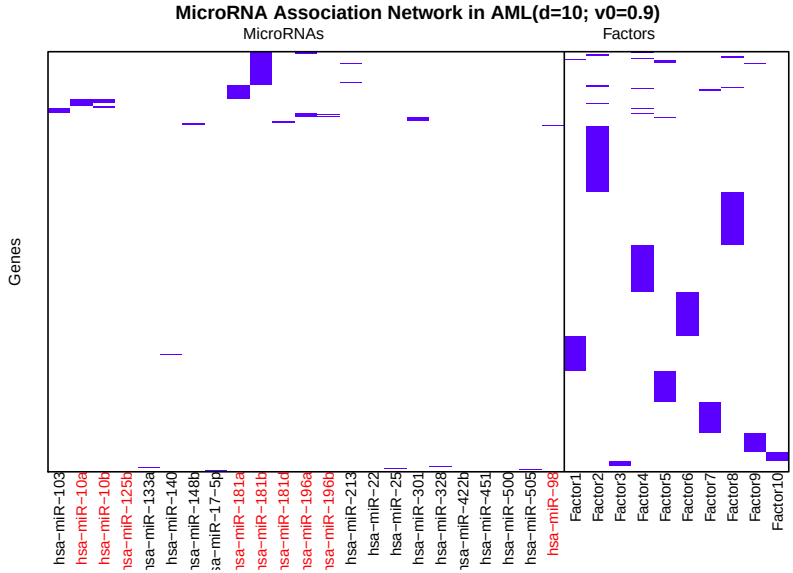


Figure 5.10: Micro-RNA-gene factor interaction network. Model assumes 10 factors and was located assuming $v_0 = 0.9$.

able selection in linear regression. By considering a nested sequence of continuous spike and slab priors and various factor space cardinalities, we dynamically explore the posterior model space and generate a series of candidate models. These are subsequently evaluated by a conditional posterior model probability criterion, which performs comparisons also across various factor dimensionalities. Patterned variable and factor selection is encouraged by structured priors on the model matrix, which also enable an orthogonal allocation of zeroes in factor loadings. We demonstrated the usefulness of our approach on a simulated data set, where we effectively recovered the sets of predictive explanators as well as the generating latent factor architecture. In a real data example, we implemented our EM factor procedure in the context of estimating a microRNA-gene interaction network in acute myeloid leukemia. We identified a set of active microRNAs, which were previously identified as influential in the pathogenesis (or related to subtypes) of AML. The practicality of the implementation renders our EM tool as an efficient approach to perform effective model selection in factor analytic models in combination with multivariate regression.

5.10

Appendix

Factor 2								
AAAS	ACAD8	ACADM	ACAP1	ADSL	AGK	AGL	AGPAT1	AHCY
ALDH5A1	ALG6	APEX1	APRT	AQP3	ATP5A1[16]	ATP5B[16]	ATP5D[16]	ATP5E[16]
ATP5G3[16]	ATP5J[16]	BCAP31	BCAS2	BPGM	BUB3[11]	CD3E[6]	CDC16[11]	CDC26[11]
CMPK1	CNOT10[7]	COX15	COX7C[16]	CPSF3	CPT2	CREB3L4[16]	CSNK1G3	CSNK2B
CUL4B	DAD1	DARS	DCTD	DLAT	DVL2[4]	EGFR[4,12]	EIF3E	EIF3H
ENO2[7]	EXOSC5[7]	EXOSC6[7]	EXOSC9[7]	FARSA	FAU	FBL	GBE1	GCDH
GMPR2	GNPAT	GNPNAT1	GOSR2	GTF2H4	HDAC1[4,11,16]	HEMK1	HIBCH	HSP90AA1[4]
HSPA4	HSPA8[12]	HTRA2	IDH3B	IMPDH2	IRF3	ITPA	KARS	LCK
LCMT1	LIAS	LSM2[7]	LSM4[7]	MAP2K2[2,4,12]	MTMR1	NAE1	NCBP2	NDUFA7[16]
NDUFB6[16]	NDUFV1[16]	NFYB	NHP2L1	NME1	NOB1	NOP56	NUDT21	NUP133
NUP37	NUP43	ORC3L[11]	PAF1	PARN[7]	PDHA1	PDHB	PFN1	PIGP
PLRG1	PNPO	POLR1D	POLR2C[16]	POLR2D[16]	POLR2H[16]	POLR2I[16]	POLR2L[16]	POLR3D
POLR3GL	POMGNT1	PPID[16]	PREB	PRIM2	PRKCI	PRKCQ	PRMT5	PRPF19
PSENEN	PSMB6	QARS	RAD50	RBM22	RBM8A	RBMX	RIOK2	RPA1
RPA2	RPL10A	RPL22L1	RPL30	RPL34	RPL36	RPL36AL	RPP40	RPS10
RPS5	RPS9	RUVBL1	RXRB[4]	SARS	SDHC[16]	SF3A2	SLC25A5[16]	SLC25A6[16]
SNRPF	SOD1[16]	SRPRB	SSB	SSR3	STAT4	SUMO2	SVIP	TAF9
TARS2	TRAM1	TTC37[7]	UQCRC1	UQCRFS[16]1	UTP18	VTA1	XAB2	XRCC1
ZMAT2	ZNRD1	SNRNP40	COQ5	MUTYH	WDR61[7]	AP2B1	HARS	NUDT9
PRDX6	SRP19							

Factor 4								
AADAT	AGPAT2	ALDH9A1	AP1M1	AP1S2	AQP9	ARF6	ARHGDIA	ARHGDIB
ARPC1B[3]	ARPC3[3]	ARPC4[3]	ATP6V0D1[15]	ATP6V1F[15]	C12orf5	C5AR1	CASP1	CCR1
CD14[6,12,15]	CD1D[6]	CD36[6,15]	CD4[6]	CD68	CD86[5]	CDA	CFL1	CLEC7A[15]
CORO1A[15]	CR1[6]	CTSH	CTSS[15]	CYBA[10,14,15]	CYBB[10,14,15]	CYP51A1	DPYD	EHD4
ENTPD1	ERCC1	FBP1	FBXO6	FCER1G	FCGR2C[14,15]	FGR	FLOT1	FLOT2
FPR1	FPR2	GLT25D1	GM2A	GNB2	GNS	GSR	HCLS1[3]	HCST
HERPUD1	HMOX1	IFI30	IFNGR2[14]	IL10RB	IMPDH1	IQGAP1	LILRA1[14]	LILRA2[14]
LILRA6[14]	LILRB1[14]	LILRB4[14]	LRP1	LRRK2	LYN	LYPLA1	MAP2K1[1,2,4,12,14]	MAPK3[1,2,4,12,14,2]
MARCO[15]	MYD88	MYL12B[10]	NADK	NPC2	OAS1	OSCAR[14]	PAPSS2	PDXK
PGAM1	PGK1	PPT1	PREX1	PRKACA[4,10,12]	PSAP	PSMC2	PTAFR	RAP1A[4,10,12]
RASGRP4[12]	RHOA[4,10]	RXRA[4]	SEC61G[15]	SEMA4A	SERPINA1	SGMS2	SIRPA[14]	SLC11A1
SLC7A7	SPI1[2,4,14]	SRA1	STX10	STX11	STX1[15]2	SUMO1	TBXAS1	TK2
TLR4[15]	TLR5	TNFRSF1B	TREX1	TYMP[4]	TYROBP[14]	UBE2D1	VASP[10]	YWHAB
YWHAZ	LILRB2[14]	IFNGR1[14]	SIRPB[14]1	NCF2[10,14,15]	VCAN	IL17RA	FCGR2A[14,15]	PKM2
NOD2[3]	TLR8	IL10RA						

Factor 8								
ABLIM2	ALG10B	ANAPC7[17]	APC[4]	AQR	ATF2[12]	ATP6V1C1	CARD6	CASP3[4,12]
CCNG2	CD79B	CDC23[17]	CHERP	CISH	CLDN5	CNOT6	COL18A1	CPSF2
CREB1	CRNKL1	CSNK1D	CSTF1	CSTF3	CUL3[17]	CUL5[17]	DCLRE1C	DDX58
DHX36	DTX3L	DUSP5[12]	DYNC1LI2	EDEM3	EHHADH	ENPP7	ERCC4	EXOC1
EXOC7	FPGT	FRAT2	GART	GLCE	GLS	GNAQ	GOSR1	GTF2E1
GXYLT1	HEATR1	HERC4[17]	HMGCR	IKBKB[4,12]	IMPA1	KLHL9[17]	LARS2	LILRA4
MAP3K1[12,17]	MAP4K3[12]	MAPK8[4,12]	MAT2A	MBTPS1	ME3	MFSD8	MTR	MYLK3
NAT1	NBN	NCK1	NLK	NMNAT1	NUMB	OR2L13	PAFAH1B1	PANK3
PHAX	PIAS1[4,17]	PIGB	PIGM	PIK3CA[4]	PIK3CB[4]	POLR1B	PPM1B[12]	PPP1R12A
PPP2R2A	PRKAA1	PSMD12	RASA1[12]	RCHY1[17]	REXO1	RIPK1	RPS6KA5[12]	SAP130
SC4MOL	SCNN1A	SEC24B	SEC24D	SEPSECS	SETD2	SGMS1	SIKE1	SLC33A1
SMAD4[4]	SMAD5	SMG1	SMURF2[17]	SNAP23	SOCS4	SOS1[4,12]	SSTR5	STAM
STAM2	SYNJ1	TAF2	TBK1	TOMM40	TPR[4]	TRIM21	TRNT1	UBE2W[17]
UBQLN2	UBR5[17]	UPF2	USP8	VPS37A	VPS4B	WDR36	WDR75	WWP1[17]
XIAP[4,17]	XRCC2	XRCC3	XRN1	SUV420H1	HSD17B7	PAPOLG	UBA3[17]	ABAT
ACOX3	EPS15	GALNT7	PIK3C2A					

Table 5.4: Genes associated with Factor 2: (1) chronic myeloid leukemia, (2) acute myeloid leukemia, (3) apoptosis, (4) pathways in cancer, (5) allograft rejection, (6) hematopoietic cell lineage, (7) RNA degradation, (8) T cell receptor signaling pathway, (9) natural killer cell mediated cytotoxicity, (10) leukocyte transendothelial migration, (11) cell cycle, (12) MAPK signaling pathway, (13) primary immunodeficiency, (14) osteoclast differentiation,(15) phagosome, (16)huntington disease, (17) ubiquitin mediated proteolysis

CHAPTER 6

RISK-STRATIFICATION OF INTERMEDIATE-RISK ACUTE MYELOID LEUKEMIA

Rockova, V., Abbas, S., Wouters, B., Erpelinck, C., Beverloo, B., Delwel, R., Putten, W., Lowenberg, B., Valk, P. 2011. **Risk-stratification of intermediate-risk acute myeloid leukemia: integrative analysis of a multitude of gene mutation and gene expression markers**. *Blood* 118:1069-1076

Abstract

Numerous molecular markers have been recently discovered as potential prognostic factors in acute myeloid leukemia (AML). It has become of critical importance to thoroughly evaluate their interrelationships and relative prognostic importance. We set out to investigate a comprehensive set of biomarkers with an emphasis on the statistical assessment of their collective utility in the stratification of intermediate risk AML using model selection in the frameworks of survival tree and regression methodologies. Gene expression profiling was conducted in a well-characterized cohort of 439 patients under age 60 with newly diagnosed AML to determine expression levels of $EVI1$, $WT1$, $BCL2$, $ABCB1$, $BAALC$, $FLT3$, $CD34$, $INDO$, ERG and $MN1$. A variety of AML-specific mutations were evaluated, i.e. $FLT3$, $NPM1$, $NRAS$, $KRAS$, $IDH1$, $IDH2$ and $CEBPADM/SM$ (double/single). Univariable survival analysis shows that (I) patients with $FLT3$ITD mutations have inferior overall survival (OS) and event free survival (EFS), whereas $CEBPADM$ and $NPM1$ mutations indicate favourable OS and EFS in intermediate risk AML, (II) high transcript levels of $BAALC$, $CD34$, $MN1$, $EVI1$ and ERG predict inferior OS and EFS. In multivariable survival analysis, $CD34$, ERG and $CEBPADM$ remain significant. Using survival tree methodology, we show that a reduced combination of $CEBPADM$, $CD34$ and $IDH2$ is capable of separating the intermediate group into two AML subgroups with highly distinctive survival characteristics (OS at 60 months: 51.9% versus 14.9%). The integrated statistical approach demonstrates that from the multitude of biomarkers a greatly condensed subset can be selected for improved stratification of intermediate risk AML.

6.1 Introduction

It is widely accepted that certain cytogenetic abnormalities consistently associate with particular subsets of AML that carry distinct responses to therapy (Marcucci et al., 2011). Approximately 40% of all AML patients are currently classified into distinct groups with variable prognosis based on the presence or absence of specific recurrent cytogenetic abnormalities. AML without favorable and particular unfavorable cytogenetic aberrations is classified as intermediate prognosis. The intermediate risk cytogenetic subclass of AML includes cytogenetically normal and AML with other cytogenetic abnormalities and accounts for approximately 60% of all AML patients, and according recent gene-mutation and gene-expression studies represents a mixture of leukemias with favorable and unfavorable prognosis.

In recent years a variety of novel molecular markers have refined the risk-stratification of intermediate risk AML. For instance, mutations in $FLT3$ (Levis and Small, 2003), $NPM1$ (Döhner et al., 2005), CEBPA (Wouters et al., 2009) all carry variable prognostic value. Recently, $IDH1$ and $IDH2$ mutations were identified but for the time being the prognostic value of these mutations appears to be controversial (Abbas et al., 2010).

Besides acquired mutations, a number of individual genes have been proposed as important prognostic expression markers, i.e., specific gene expression levels were shown to be

associated with treatment outcome in AML. For instance, expression of *EVI*1 (Lugthart et al., 2008), *BAALC* (Baldus et al., 2003), *ERG* (Marcucci et al., 2007) and *MN*1 were proposed as indicators for treatment outcome in AML (Langer et al., 2009). Some expression markers such as *WT*1 (Bergmann et al., 1997), *BCL*2 (Karakas et al., 1998), *INDO* (Chamuleau et al., 2008), *CD*34 (Kanda et al., 2000), *ABCB*1 (van den Heuvel-Elbrink et al., 1997) and *FLT*3 (Ozeki et al., 2004) have been put forward as clinical markers, but their applicability has been less well-established or has been controversial.

Previous studies have often assessed the prognostic value of various biomarkers on an individual basis or in a limited collective context. For the purpose of risk stratification and understanding of the relative prognostic importance it has become crucial to integrate them in a joint analysis. In the present study we investigate the role of gene expression markers *EVI*1, *WT*1, *BAALC*, *ERG*, *BCL*2, *ABCB*1, *INDO*, *CD*34, *BCL*2 and *MN*1 (evaluated using standardized micro-array analysis) as well as somatic gene mutations in *FLT*3, *NRAS*, *CEBPASM*, *CEBPADM*, *NPM*1, *IDH*1 and *IDH*2 in survival prognosis in cytogenetically defined intermediate risk AML. In addition to univariable and multivariable analysis we adopted a statistical approach that is capable of deriving a simplified prognostic index that can be used for the risk stratification of the intermediate risk group.

6.2

Methods

6.2.1 Patients, Cell Samples and Molecular Analyses

We investigated a cohort of 439 patients (age below 60 years) with a diagnosis of primary AML or RAEB(-t) (n=17) (Figure 6.2 in Appendix). All patients were treated according to the HOVON (Dutch-Belgian Hematology-Oncology Cooperative group) protocols between 1987 and 2006 (http://www.hovon.nl) (Löwenberg et al., 1997).

All AML cases in this study were also included in (Wouters et al., 2009; Lugthart et al., 2008) and subsets of cases have also been investigated in Valk et al. (2004) (online supplementary materials[1], Table 1). The earlier studies had different study objectives, i.e., dealing with individual markers or selected subsets of leukemia (for instance cytogenetically normal AML).

AML was cytogenetically classified into the following prognostic categories: (I) favorable: t(8;21) and inv(16); (II) very unfavorable: monosomal karyotypes (MK) as defined in Breems et al. (2008); (III) intermediate risk I: cytogenetically normal (CN) and (IV) intermediate risk II: the remaining AML cases (CA).

After informed consent, bone marrow aspirates or peripheral blood samples were taken at diagnosis. Blasts and mononuclear cells were purified by Ficoll-Hypaque (Nygaard, Oslo, Norway) centrifugation and cryopreserved. The AML samples contained $80-100$ percent

[1] http://bloodjournal.hematologylibrary.org/content/118/4/1069/suppl/DC1

blast cells after thawing, regardless of the blast count at diagnosis. Mutational analyses were all performed as described previously (Abbas et al., 2010).

6.2.2 Gene Profiling and Quality Control for Assessment of Gene Expression Variations

439 AML samples were analyzed using Affymetrix U133Plus2.0 GeneChips (Affymetrix, Santa Clara, CA) that contains 54675 probe sets, representing 20650 unique genes. The methods have been reported in detail elsewhere (Valk et al., 2004). The differences between the scaling/normalization factors of the GeneChips in complete cohort was less than 3-fold (0.62 ($\pm$ 0.20)). All additional measures of quality-percent genes present (39.8 ($\pm$ 3,5)), GAPDH 3' to 5' ratio (1.08 ($\pm$ 0.15)) and actin 3' to 5' ratio (1.30 ($\pm$ 0.26))-indicated high overall sample and assay quality in the complete AML cohort.

Informative probe sets detecting expression of various genes were selected. Only those probe sets with accurate annotation and genomic localization using the ENSEMBL genome browser (http://www.ensembl.org/) were included, i.e., *ABCB*1: 209993_at, 209994_s_at; *WT*1: 206067_s_at, 216953_s_at; *BCL*2: 203684_s_at, 203685_at; *BAALC*: 218899_s_at, 222780_s_at; *ERG*: 213541_s_at, 241926_s_at; *EVI*1: 221884_at, 226420_at; *FLT*3:206674 _at; *CD*34: 209543_s_at; *MN*1: 205330_at, *INDO*: 210029_at.

6.2.3 Data Preparation

Each of the mutation markers is coded as a binary variable, i.e. mutation present (+) or absent. The gene expression of each selected gene was determined from either a single probe or a combination of multiple probes linked to that gene. Probe sets for each expression marker were selected from the Affymetrix U133Plus2.0 GeneChip, based on an accurate annotation and localization using the ENSEMBL genome browser. If one probe per a gene was available (*MN*1, *CD*34, *FLT*3 and *INDO*1), the probe expression intensity was log2 transformed and scaled so that the minimal value equals 0 and the maximal value equals 1. In case multiple probe sets were annotated for a single gene (*BAALC*, *BCL*2, *ABCB*1, *EVI*1, *WT*1 and *ERG*), we reduced the number of variables by performing a factor analysis per gene using the log2 transformed expression data of all 439 AML patient samples. This resulted in a factor score, comprising of the expression values from all the representative probe sets, for each individual expression marker. The factor scores were also rescaled so that the minimal value of the score for each gene is 0 and the maximal value is 1.

6.2.4 Statistical Analysis

Statistical analyses were performed with R (version 2.9.2). Both overall survival (with failure defined as death due to any cause) and event-free survival (with failure defined as no complete remission (set at day 1), relapse, or death in first complete remission) were considered as endpoints for survival analyses.

To determine the prognostic value of the markers Cox proportional hazard (PH) regression model was used in univariable and multivariable analyses. To further inspect the prognostic importance and/or redundancy of the markers we applied a variable selection in Cox proportional hazards model, namely the AIC-based stepwise variable selection and the LASSO (Tibshirani, 1994), where the optimal penalty parameter was chosen so that it maximizes the cross-validated partial log-likelihood (20-fold cross-validation). To further evaluate the hierarchy of the prognostic importance we utilized tree-structured survival modeling (unbiased recursive partitioning approach of Hothorn et al. (2006)). Estimated probabilities of OS and EFS were calculated using the Kaplan-Meier method. Partial likelihood ratio test was used to evaluate differences between survival distributions.

The bimodal shape of the *EVI*1 expression distribution (online supplementary material, Figure 3) suggests that there are two populations of patients with high and low *EVI*1 expression. A mixture model fit with normally distributed components supports the evidence for this observation (online supplementary material, Figure 3). The intersection point of the two superimposed densities naturally suggests a threshold (c=1.15) to decide whether the *EVI*1 was over-expressed or not. *EVI*1 expression based on RQ-PCR was treated as a categorical variable in previous reports (Lugthart et al., 2008). Here we also use a categorical *EVI*1 in survival analyses using the reference value 1.15. A penalized spline fit in Cox PH regression suggests a nonlinear behavior of *EVI*1 (p-value of a test for linearity 0.04, best degrees of freedom 2.1 determined by AIC criterion). Despite a piecewise constant transformation might not be the best approximation of the true relationship it manages to separate the distinctive survival characteristic of the small group of patients (8.8%) with high *EVI*1, which would be masked if we treated *EVI*1 as linear. The remaining markers are treated on a continuous scale in accordance with their actual distribution pattern. Unless otherwise stated, with “high expression" we refer to high values extreme with respect to the distribution of each marker.

Pair-wise associations between binary markers were assessed by means of Chi-square test (or Fisher (Halton-Freeman) exact tests when the expected count number in at least one of the cells dropped below 5). The direction of the observed associations was measured by a ϕ coefficient. Spearman correlation coefficient was used to asses the pair-wise correlations between gene expression markers. Differential gene expression across patient sub-categories was tested by means of Wilcoxon sum rank test (two categories only) and Kruskal-Wallis test (more than two categories). The level 0.05 has been utilized as a threshold to declare the statistical significance.

6.3 Results

6.3.1 Distribution Across Cytogenetically Defined AML Subsets

Details on the molecular and clinical characteristics of the investigated cohort of 439 patients are summarized in Figure 6.2 in Appendix.

The distribution of the recurrent mutations among the cytogenetically-defined AML subsets is summarized in Table 2 in online supplementary materials. We note increased frequencies of $FLT3$ITD and $NPM1$ mutations in CN AML as well as the common occurrence of $FLT3$ITD and $FLT3$TKD in AML with t(15;17). The prevalence of $FLT3$TKD and $NRAS$ mutations is higher in AML with inv(16). $IDH1$ and both $CEBPASM$ and $CEBPADM$ were observed exclusively in intermediate risk cytogenetic categories (CN and CA). $KRAS$ mutations are relatively rare in AML and were not considered in further analyses.

The majority of expression markers genes show a differential expression in the cytogenetically defined AML subsets (online supplementary materials, Figure 1). Expression marker genes $INDO1$ and $FLT3$ do not have distinctive expression patterns in relationship to the cytogenetically defined subgroups (p-values of the overall Kruskal-Wallis test 0.301 and 0.204 respectively). Compared to the normal karyotype group, significantly higher expression of $WT1$ is associated with t(15;17) ($p < 0.001$), relatively low $BCL2$ expression is observed in AML with t(8;21) ($p < 0.001$) and high expression of both $BAALC$ and $CD34$ was detected in t(8;21) and inv(16) groups ($p < 0.001$ for both comparisons). We further noticed elevated $MN1$ expression in AML with inv(16) compared to the cytogenetically normal group ($p < 0.001$).

6.3.2 Associations Between Mutation and Expression Markers

The summary of pair-wise associations between the binary mutation markers is given in Table 3 (online supplementary materials). $FLT3$ITD, $FLT3$TKD and $IDH1$ mutations appear significantly over-represented in $NPM1$ mutant group (ϕ coefficients 0.36, 0.15 and 0.28, respectively). On the other hand, $FLT3$ITD are more prevalent in AML without $FLT3$TKD (ϕ=-0.11), $NRAS$ (ϕ=-0.2) or $IDH2$ mutations (ϕ=-0.11).

Spearman correlation analysis between the gene expression markers in Figure 6.1 in Appendix revealed the following associations: (i) the expression of the marker genes $BAALC$, $CD34$, $MN1$, ERG and $ABCB1$ are relatively strongly associated, (ii) $BAALC$ exhibits the strongest positive correlation with $CD34$ expression (correlation coefficient Φ equals 0.78) and $MN1$ (Φ=0.76), (iii) moderate associations are also observed between ERG and $WT1$ (Φ=0.43), ERG and $BCL2$ (Φ=0.4) and $BCL2$ and $WT1$ (Φ=0.36), (iv) $INDO1$ appears to be inversely associated with $EVI1$ (Φ=-0.25), $WT1$ (Φ=-0.28) and ERG (Φ=-0.14).

The summary of the association analysis between the mutation and gene expression markers is given in Table 4 (online supplementary materials). $NPM1$ mutant patient group is significantly associated with higher $WT1$ expression. In contrast, the expression of $BCL2$, $BAALC$, ERG, $ABCB1$, $CD34$ and $MN1$ is elevated in $NPM1$ wild-type AML. Other associations that we observe are e.g. decreased $BAALC$, $CD34$ and $MN1$ expression as well as increased $FLT3$ and $WT1$ expression in $FLT3$ITD AML. Increased $ABCB1$ expression associates with $CEBPADM$ AML. Likewise, $BAALC$ and $CD34$ expression are higher in $IDH1$ wild type AML and $BCL2$ expression is higher in $IDH2$ mutant AML.

6.3.3 Survival Analyses in Intermediate-risk AML

Univariable survival analysis (Table 6.3 in Appendix) indicated inferior OS in intermediate risk AML patients with *FLT*3ITD mutations (hazard ratio (HR) 1.41; p=0.017), whereas *CEBPADM* (OS: HR=0.38, p=0.004; EFS: HR=0.45, p=0.007) and *NPM*1 mutations (OS: HR=0.73, p=0.03; EFS: HR=0.69; p=0.006) were found indicative of favorable OS and EFS. The positive prognostic impact of *NPM*1 mutations becomes even more pronounced in *FLT*3ITD negative AML (OS: HR=0.63, p=0.022; EFS: HR=0.64, p=0.018). Univariable analysis of the gene expression markers demonstrates that increased expressions of *BAALC*, *CD*34, *MN*1, *EVI*1 and *ERG* are significant negative indicators for OS and EFS (all hazard ratios >1.5 and p<0.01). Univariable survival analysis for CN AML is given in Table 5 (online supplementary materials). The negative predictive effect of *FLT*3ITD, *BAALC*, *CD*34, *EVI*1 and *ERG* is retained in CN AML.

The multivariable Cox regression analysis (Table 6.4 in Appendix) shows that *CD*34, *ERG* and *CEBPADM* remain significant predictors for OS and EFS after the correction for the remaining markers (respective p-values for OS p=0.004, p=0.036, p<0.001 and EFS p=0.005, p=0.032, p<0.001), whereas neither increased *BAALC*, increased *MN*1 or *EVI*1 expression or the presence of *FLT*3ITD are no longer indicative of adverse OS and EFS in intermediate risk AML. The multivariable survival analysis for CN AML is summarized in Table 6 (online supplementary materials). When we control for the remaining prognostic markers, only *CEBPADM* and *CD*34 remain significant in CN AML.

To investigate which minimal subset/combination of markers is sufficient for assessing prognosis, variable selection in Cox proportional hazards model was performed. The LASSO variable selection with an optimal value 8.7 of the penalization parameter identified the following markers: *CD*34, *CEBPADM*, *IDH*2, *BCL*2, *ERG*, *NPM*1, *EVI*1, *FLT*3ITD and *INDO*1. Estimated regression coefficients of the penalized Cox proportional hazards model for different values of the penalization parameter for OS are depicted in Figure 2 (online supplementary materials). The plot indicates that amongst the considered series of markers *CD*34 and *CEBPADM* play a predominant role in survival prognosis. The AIC based stepwise selection identified a similar set of markers, i.e. *CD*34, *CEBPADM*, *IDH*2, *BCL*2 and *ERG*. The variables recognized as important by the recursive binary partitioning in the survival tree methodology were *CD*34, *CEBPADM* and *IDH*2 (Figure 6.6 in Appendix). Similar results were obtained for EFS. The tree model is in accordance with the penalized Cox regression approach in that *CD*34 and *CEBPADM* were again identified as the most important predictors.

The survival tree model in Figure 6.6 naturally suggests stratification of the intermediate risk AML into subgroups with more homogeneous survival characteristics. According to the model, the intermediate risk group could be divided into four categories: (I) “low *CD*34" (defined as *CD*34<0.398), (II) “high *CD*34" (defined as *CD*34>0.398), *IDH*2 wild type and *CEBPADM* absent, (III) “high *CD*34" (>0.398), *IDH*2 mutated and no *CEBPADM* and (IV) “high *CD*34" (>0.398) and *CEBPADM*. Out of the 4 categories, the groups (I), (III) and (IV) have statistically indistinguishable survival characteristics (p-value of a 2df partial likelihood

ratio test 0.298 (OS) and 0.333 (EFS)). The three groups (I), (III) and (IV) together could be aggregated as the favorable intermediate risk group (estimated OS and EFS at 60 months 51.9% and 41.5%, respectively). In contrast, the estimated OS and EFS at 60 months in the group (II) is 14.9% and 8.3%, respectively, which indicates unfavorable prognosis. The latter group has been designated: poor intermediate risk group. The survival characteristics in the proposed strata when compared with the survival profile of the established cytogenetical prognostic stratification (as described in Methods) is given in Figure 6.6. The difference in survival between favorable and intermediate favorable prognostic groups is not statistically significant (p-value of 1df likelihood ratio test 0.153 (OS), 0.44 (EFS)). The survival characteristics between poor and poor intermediate group are significantly different (p-value of 1df likelihood ratio test 0.012 for both OS and EFS).

6.4 Discussion

AML is a group of neoplasms characterized by a variety of genetic and epigenetic aberrations and variable responses to therapy (Marcucci et al., 2011). The pretreatment karyotype of leukemic blasts is currently a key determinant for therapy decision-making in AML. Usually, the largest cytogenetic subclass of AML, i.e., those patients with a normal karyotype and patients with prognostically non-informative cytogenetic aberrations, are categorized as intermediate risk. In recent years a number of novel markers have been identified as putative classifiers for these AML patients. These markers include a wide-variety of acquired mutations as well as expression changes in specific genes.

In previous studies prognostic risk assessments were put forward based on various expression markers *BAALC*, *ERG*, *MN*1 and *EVI*1. These studies have postulated risk algorithms mainly for CN AML and included only few out of the wide-variety of mutations and expression markers. Studies addressing the relative importance of the various postulated mutations and expression markers are limited54.

In this study we investigated the role of a wide series of genomic biomarkers that included mutations in *FLT*3, CEBPA, *NPM*1 and *WT*1 genes as well as high-expression of *EVI*1, *WT*1, *BCL*2, *ABCB*1, *BAALC*, *FLT*3, *CD*34, *INDO*, *ERG* and *MN*1 in the risk-stratification of intermediate risk AML. The results reveal particular associations between some of these markers that may strongly affect the collective use of these markers in risk assessment. For instance, we demonstrate an inverse association between *NPM*1 mutations and Affymetrix HGU133 Plus2.0-derived *CD*34 expression, as was shown by others (Verhaak et al., 2005). Importantly, relatively strong associations exist between expression levels of *CD*34, *BAALC*, *MN*1, *ERG* and *ABCB*1. Consequently, expression values of all these markers inversely correlate with the presence of mutant *NPM*1. These interactions indicate that these markers will have similar value in risk stratification of AML and should therefore be taken into account when prognostic scores based on selected markers are constructed.

By univariable analyses we confirmed the prognostic ability of previously established markers in intermediate risk AML, i.e., *CEBPADM* and *NPM*1 mutations as indicators for favorable OS and EFS and *FLT*3ITD mutations as markers for poor response to therapy. High expression of *BAALC*, *CD*34, *MN*1 and *ERG* all express unfavorable prognostic value with regard to OS and EFS, which is in line with earlier publications (Marcucci et al., 2011). Importantly, expression of *CD*34 mRNA strongly associates with poor OS and EFS.

In multivariable analyses, *CEBPADM* independently predicts favorable outcome, whereas *CD*34 and *ERG* are independent predictors for inferior OS and EFS. *ERG* expression has emerged as a strong negative predictor in multivariable analyses previously, however, in this model *CD*34 expression is the strongest expression marker for poor outcome. By conducting a model selection in both Cox proportional hazards regression models and survival trees, it becomes evident that *CEBPADM* and *CD*34 expression stand out as the most prominent predictors for treatment outcome. Although the value of *CD*34 protein expression has been controversial (Kanda et al., 2000), *CD*34 mRNA appears to be notably valuable in AML risk stratification.

Although stratification based on expression levels is challenging, the usage of standardized protocols and Affymetrix GeneChips may facilitate the implementation of gene expression level analyses. In fact, since many laboratories currently use Affymetrix GeneChips the results of these types of analyses may be relatively easily implemented.

We developed a simplified stratification rule of intermediate risk AML, which identifies two distinctive groups of patients with survival characteristics being similar to the generally established favorable and poor risk cytogenetic subgroups, respectively. We acknowledge that the proposed stratification needs further validation in future studies and will be probably improved with new emerging knowledge. Nevertheless, the model presented here discloses several particularly interesting associations with respect to the hierarchy of the prognostic importance of a scale of molecular biomarkers and adds to the understanding of the heterogeneity of intermediate risk AML.

6.5 Appendix

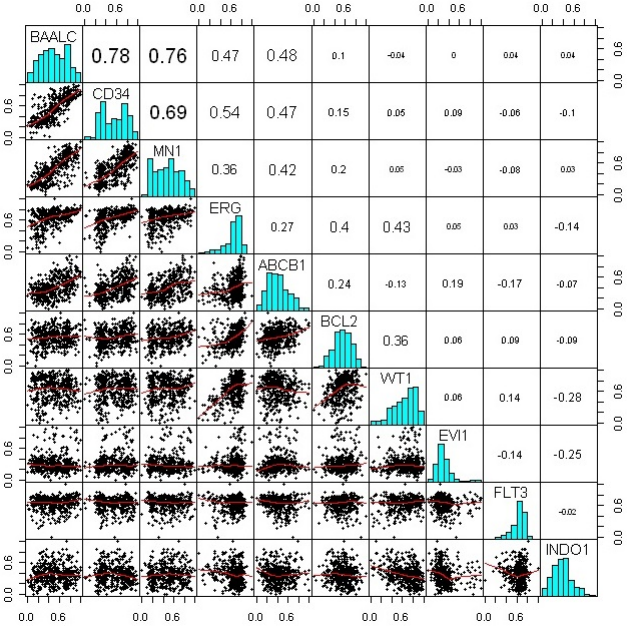


Figure 6.1: Associations between the gene expression markers. Lower triangle is a scatter-plot matrix of the markers, where the red lines are the lowess smoothing curves. Upper triangle encapsulates pair-wise Spearman correlation coefficients. On the diagonal there are histograms of each of the markers.

Clinical variables		Range	Mean/Median	Number of Patients
White blood cell count	(10^9/l)	0.3-278	52.04/29.8	
Bone marrow blast count	(percentage)	0-98	62.08/66	
Platelet count	(10^9/l)	3-998	78.92/52	
Patient characteristics		Range	Mean/Median	Number of Patients
Age		15-60	42.11/43	
Gender	(female)			219
FAB classification			%	Number of Patients
M0			3.6	16
M1			19.1	84
M2			23.2	102
M3			5	22
M4			18.5	81
M4Eo			6.2	27
M5			23.7	104
M6			1.1	5
RAEB			0.9	4
Not determined			4.8	21
Cytogenetics			%	Number of Patients
t(8;21)			8	35
inv(16)			8.2	36
t(15;17)			5.7	25
Cytogenetically normal (CN)			43.47	192
Cytogenetically abnormal (CA)			28.7	126
Monosomal karyotype (MK)*			5.7	25
Mutations			%	Number of Patients
NPM1+			29.6	130
FLT3^{ITD}+			26.9	118
FLT3^{TKD}+			10.7	47
N-RAS+			987	43
K-RAS+			0.9	4
CEBPA^{SM}+			1.6	7
CEBPA^{DM}+			5.2	23
IDH1+			7.2	32
IDH2+			8.2	36
NPM1+ FLT3^{ITD}+			15.3	67
NPM1+ FLT3^{ITD}-			14.4	63
NPM1- FLT3^{ITD}+			11.6	51
NPM1- FLT3^{ITD}-			58.8	258

Figure 6.2: Mutation present (resp. absent) groups denoted with (+) (resp. (-)). Abbreviations: RAEB: refractory anemia with excess blasts, FAB: French-American-British, CN: normal cytogenetics or -X or -Y as the sole abnormality, M4Eo: M4 category with inv(16); * MK category contains 16 AML patients classified as complex karyotype, 17 other cases with complex karyotypes are in the CA (n=13), inv(16) (n=2), t(8;21) (n=1) and t(15;17) (n=1) categories.

		Overall Survival				Event Free Survival			
Variable		**Hazard Ratio**	**Lower**	**Upper**	**p-value**	**Hazard Ratio**	**Lower**	**Upper**	**p-value**
NPM1	+	0.73	0.55	0.97	**0.03**	0.69	0.53	0.9	**0.006**
FLT3^{ITD}	+	1.41	1.06	1.86	**0.017**	1.3	0.99	1.7	0.059
FLT3^{TKD}	+	0.82	0.51	1.32	0.418	0.74	0.47	1.16	0.192
N-RAS	+	0.94	0.57	1.54	0.798	1.23	0.77	1.94	0.386
CEBPA^{SM}	+	1.01	0.38	2.72	0.984	0.8	0.3	2.16	0.662
CEBPA^{DM}	+	0.38	0.19	0.74	**0.004**	0.45	0.25	0.81	**0.007**
IDH1	+	0.83	0.52	1.31	0.414	0.97	0.64	1.47	0.877
IDH2	+	0.74	0.47	1.17	0.199	0.79	0.51	1.21	0.273
FLT3$^{ITD}\times$ **NPM1**	+ -	1.67	1.13	2.46	**0.01**	1.76	1.21	2.58	**0.003**
	- +	0.63	0.42	0.94	**0.022**	0.64	0.45	0.93	**0.018**
	+ +	1.03	0.72	1.47	0.875	0.9	0.64	1.27	0.549
EVI1*	+	1.78	1.17	2.7	**0.007**	2.01	1.34	3.02	**<0.001**
BAALC		3.16	1.74	5.72	**<0.001**	2.9	1.63	5.16	**<0.001**
CD34		3.81	2.17	6.67	**<0.001**	3.57	2.11	6.05	**<0.001**
MN1		2.41	1.37	4.23	**0.002**	2.51	1.46	4.32	**<0.001**
ERG		3.69	1.65	8.26	**0.001**	3.48	1.63	7.43	**0.001**
ABCB1		0.99	0.51	1.93	0.983	0.92	0.49	1.73	0.798
BCL2		1.07	0.5	2.3	0.861	1.19	0.57	2.46	0.644
INDO1		0.65	0.32	1.36	0.254	0.69	0.35	1.38	0.3

Figure 6.3: Univariable survival analysis in the intermediate risk group.

		Overall survival				Event free survival			
Variable		**Hazard Ratio**	**Lower**	**Upper**	**p-value**	**Hazard Ratio**	**Lower**	**Upper**	**p-value**
FLT3$^{ITD}\times$ **NPM1**	+ -	1.23	0.79	1.92	0.37	1.54	1	2.37	0.05
	- +	0.73	0.43	1.25	0.25	0.7	0.42	1.17	0.17
	+ +	1.09	0.66	1.79	0.74	0.94	0.58	1.52	0.81
FLT3^{TKD}	+	1.13	0.67	1.93	0.64	0.99	0.6	1.63	0.95
N-RAS	+	0.99	0.59	1.68	0.99	1.28	0.79	2.08	0.32
CEBPA^{SM}	+	0.99	0.35	2.81	0.99	0.84	0.3	2.36	0.75
CEBPA^{DM}	+	0.21	0.1	0.44	**<0.001**	0.26	0.14	0.51	**<0.001**
IDH1	+	1.09	0.67	1.79	0.73	1.31	0.83	2.06	0.25
IDH2	+	0.66	0.4	1.1	0.11	0.79	0.49	1.27	0.32
EVI1	+	1.15	0.72	1.83	0.56	1.3	0.83	2.04	0.26
BAALC		1.43	0.43	4.72	0.56	1.07	0.35	3.28	0.91
CD34		5.09	1.73	15.01	**<0.001**	4.47	1.62	12.32	**<0.001**
MN1		0.59	0.21	1.67	0.32	0.7	0.25	1.95	0.49
ERG		4.13	1.01	16.86	**0.05**	3.93	1.05	14.62	**0.04**
ABCB1		0.84	0.32	2.22	0.72	0.67	0.26	1.68	0.39
BCL2		0.34	0.12	0.96	**0.04**	0.42	0.15	1.15	0.09
WT1		0.68	0.28	1.64	0.39	0.66	0.29	1.53	0.34
FLT3		0.75	0.28	2	0.56	0.66	0.25	1.74	0.41
INDO1		0.7	0.3	1.61	0.4	0.69	0.32	1.51	0.36

Figure 6.4: Multivariable survival analysis of the intermediate risk group Mutation present (resp. absent) groups denoted with (+) (resp. (-)). The reference category for binary mutation markers is mutation absent (-). The reference category for the combined aberration in *FLT*3ITD and *NPM*1 is both mutations absent. $*$ *EVI*1 expression categorized with the reference category "*EVI*1<1.15".

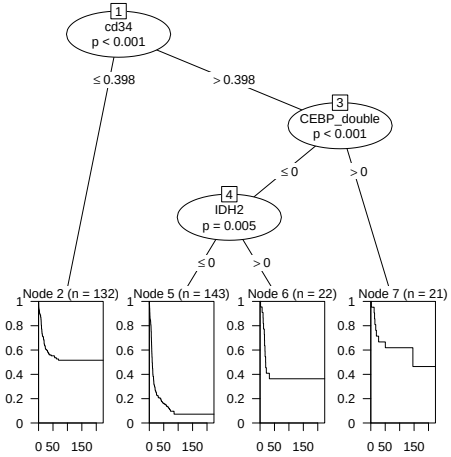


Figure 6.5: Graphical representation of the survival tree model (OS) The tree depicts the partitioning of the 318 intermediate risk AML into four groups with more similar survival characteristics. Kaplan-Meier estimates of the survival curves for each of the groups attached at the bottom of the tree. The group I (n=132) consists of patients with *CD*34 expression ? 0.398, group II (n=143) are patients with *CD*34 expression > 0.398, *IDH*2 and *CEBPADM* wild types, group III is characterized by *CD*34 expression > 0.398, *IDH*2 mutation present and no *CEBPADM*, group IV includes patients with *CD*34 expression > 0.398 and *CEBPADM*.

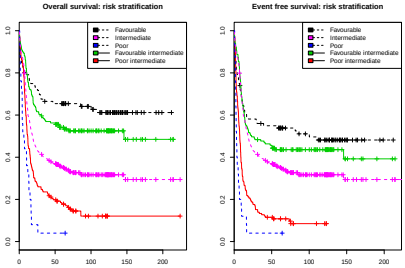


Figure 6.6: Risk stratification of intermediate risk AML The left (right) panel presents Kaplan-Meier survival curve estimates for the overall survival (event free survival) in five AML subsets. Black lines indicate survival curves for favorable (solid line), intermediate (dashed line) and unfavorable (dotted line) cytogenetic risk subgroups of AML as defined in Methods. The red curve corresponds to the poor intermediate group defined as *CD*34 expression > 0.398, *IDH*2 mutation and no *CEBPADM*. The green line refers to the favorable intermediate group defined as either (a) *CD*34 expression < 0.398, (b) *CD*34 expression > 0.398 and *CEBPADM*, or (c) *CD*34 expression > 0.398, no *CEBPADM* and *IDH*2 mutant.

CHAPTER 7

GENERAL DISCUSSION

7.1

Concluding Remarks

In this last chapter we conclude with a summary of principal research contributions presented in the preceding chapters and with a discussion on possible future extensions. The *variable selection* has been the leitmotif throughout the thesis, where taking the Bayesian perspective we presented fast inferential algorithms for manageable computation in high-volume data. The thesis accommodates numerous adaptations of the EM algorithm for variable selection and introduces a rich deterministic framework for posterior model mode detection and MAP estimation. The recurrent motives of Bayesian variable selection and EM algorithm are to reappear one more time in the overview of methodological contributions presented in the following section.

7.1.1 Summary of the Methodology

The growing cross-disciplinary abundance of high-dimensional data has necessitated developments of reliable data exploratory techniques to perform sensible knowledge extraction in order to generate valid conclusions. Due to the complexity of the data it is often challenging to fully explore the model uncertainty and to determine how many and which elements best explain the observed phenomenon.

In **Chapter 2** we examine how uncertainty surrounding such selections can be quantified using coherent probabilistic mechanisms under the Bayesian paradigm. The richness and versatility of the Bayesian formalism for model selection have led to an explosion of increasingly elaborate approaches, a brief overview of which is given in **Chapter 2**. There we further examine the performance of the selected Bayesian and classical variable selection methods on synthetic data and demonstrate practical gains of casting variable selection in terms of posterior probabilities, while highlighting the methodological benefits of spike and slab models. We also shed light on practical challenges associated with posterior computation, which can be formidable particularly in situations when the set of possible models is large. This issue is addressed in **Chapter 3**.

Posterior model exploration typically requires careful implementation of stochastic search algorithms, that scale efficiently with the data and guarantee sufficient coverage over the set of important models. In vast model spaces, interesting high-probability models can be scattered among a few isolated peaks of accumulated posterior density. Stochastic model exploration is unlikely to be effective without some form of parallelization or distributed computing, imposing a heavy computational burden. In **Chapter 3** we demonstrate that by taking the conjugate continuous spike and slab prior, it is feasible to implement a deterministic model exploratory tool, which locates high-posterior models and at the same time leads to dramatic timing improvements over the previously proposed stochastic search methods. The rapid model exploration device is obtained with the assistance of the EM algorithm, where the observed data is augmented with the “missing" variable selection indicators. As opposed to the previous

approaches suggested for the point mass priors, the continuous relaxation is essential to obtaining a hierarchical separation of the prior to yield a closed form E-step and M-step. A simple probabilistic rule enables locally identifying high posterior models by thresholding local modes of the full regression vector according to the conditional median probability model rule. The EM machinery thereby enables a rapid exploration of the model space through locating high-probability models associated with the shrunken full regression vector. Sharply spiked prior distributions typically produce highly multimodal posterior landscapes, whereof smoothing can be accomplished by increasing the spike variance. This observation has motivated our consideration of a whole sequence of spike and slab priors with an increasingly wider spike, where small unimportant coefficients are absorbed within the spike and only a few nonzero coefficients are exposed. This dynamic posterior model exploration enables us to capture the evolution of selected subsets in a novel spike and slab regularization diagram. The regularization plot displays a series of increasingly sparser local candidate subsets, which can be evaluated based on their posterior probability, an analogue to the cross-validation criterion for optimal penalty selection in frequentist regularization methods. We refer to the dynamic procedure for model exploration and evaluation as EMVS, the EM approach to variable selection as a deterministic counterpart to the stochastic search variable selection (SSVS) of George and McCulloch (1993). The multimodal nature of the posterior further motivated our examination of deterministic annealing, which generates smooth objective functions by tempering the posterior and thereby increases chances of finding the global mode. We have presented numerous analyses of synthetic data, where the EMVS procedure leads to enormous computational savings while simultaneously providing a dynamic perspective on the variable selection status of each variable. In **Chapter 3** we also embark on prior modeling of the variable selection indicators by employing the Markov random field prior to encourage patterned sparsity in order to make inferences about possible predictive network covariate structures.

Returning to the point of patterned variable selection, many methods have been proposed to encourage concurrent selection of variables that cluster within groups. In **Chapter 4** we propose a Bayesian variable selection variant, which embeds the grouping within the shrinkage estimation. We relax the following assumptions that are typical for many existing procedures for grouped variable selection: (1) smoothness of regression coefficients within a group, (2) predictiveness of all variables within a group, (3) orthogonality of the group identification matrix. We analyze properties of two proposed computational methods based on the EM algorithm, where variable selection is accomplished through the identification of MAP estimates that are zero or at its close proximity. We allow each group to be characterized by an unknown shrinkage parameter, which drives the within-group regression coefficients towards zero. By jointly modeling the regression of responses on the predictors and the regression of the penalty parameters on the group identification matrix, we aim to simultaneously identify predictive groups as well as relevant variables within the groups. Applications on real and simulated data demonstrate that substantial practical gains can be obtained over existing approaches.

Finally, in **Chapter 5** we investigate the situation where multiple related responses are available and we would like to select subsets of predictive explanatory variables, while si-

multaneously decoding the likely grouping structure responsible for the residual covariance pattern. We extend the multivariate regression approach by augmenting the model with latent factors to capture the residual collinearity among the multiple responses. Following the main theme in the thesis, we induce sparsity by imposing variable selection priors on individual entries in the matrix of regression coefficients and factor loadings. We demonstrate that the EM algorithm developed for probabilistic principal components combines conveniently with the EMVS procedure to enable rapid exploration of candidate factor regression models. Similar existing proposals have so far relied on heavy computation using MCMC.

Chapter 6 presents an analysis of biomarkers in acute myeloid leukemia, that demonstrates the immense practical utility of variable selection in biomedical applications.

7.2 Future Research Directions

The work presented in the preceding sections provides foundations for interesting future research avenues and extensions. In the following, we highlight several important aspects which were not discussed in the body of the thesis and which provide potentially interesting directions for future investigation.

7.2.1 Variational Bayesian Methods

Throughout the course of the thesis, we have seen numerous implementations of the EM algorithm. The EM bypasses difficulties with the global posterior exploration by generating merely a point estimate of the regression vector. Conditionally on this point estimate, our EM algorithm for variable selection outputs a locally optimal model, the “local" median probability model. Alternatively, the EM framework could be extended to allow for a full posterior inference using variational approximations. Variational methods yield an estimate of the full posterior distribution by finding a closely fitting approximation within a class of distributions for which it is easier to do inference. The quality of the approximation is judged by the Kullback-Leibler divergence. The problem translates as an optimization task taking place over a set of parameters that are iteratively updated. Variational schemes were developed for sparse factor analytic models using continuous spike and slab priors by Stegle et al. (2000). Similar schemes could be considered also in the classical variable selection framework, where they could potentially yield a useful approximation to the global median probability model.

7.2.2 Predictors Forming a Directed Acyclic Graph

In **Chapter 3** we discussed patterned variable selection where spatial or network dependencies are induced among the variable selection indicators. In undirected networks, the structured model priors can be accommodated conveniently within an exponential family framework, where one can leverage an extensive theory developed for approximate inference. The directed graphs pose some additional difficulties which are discussed below.

Genomic data often give rise to graph-structured covariates together with supplementary information on the causal relationships among the predictors. The formalism of directed acyclic graphs provides a suitable framework for representing prior distributions on the model space, where the binary inclusion status can be transmitted from a parent to a child node in the graph. Following the notation introduced in **Chapter 3**, we define a graph $G = (V, E)$ to be a collection of vertices $V = \{\gamma_1, \ldots, \gamma_p\}$ and a set of edges $\mathscr{E} = \{(i, j) : 1 \leq i \neq j \leq p\}$, which represents a subset of interacting pairs. The edge set is characterized by an adjacency matrix $\theta_2 = (\theta_{ij})_{i,j=1}^p$, where $\theta_{ij} \neq 0$ if and only if $(i, j) \in \mathscr{E}$. Whereas in the undirected case there is no distinction between edges (i, j) and (j, i), directed graphs are characterized by an asymmetrical adjacency matrix, where each entry θ_{ij} is nonzero if and only if γ_j is a parent of the child node γ_i. For every directed graph that is also acyclic we can find a partial ordering of the vertices by the notion of ancestry, where each edge is directed from a higher-ordered node towards a lower-ordered node. The joint distribution $\pi(\gamma \mid \theta)$ can be then factorized as a product of local conditional distributions for each vertex, given its parents on the graph. For binary networks, the logistic distribution is convenient to parametrize these conditional probabilities. As opposed to the independent logistic regression product prior described in **Chapter 3**, here the parent selection indicators themselves are predictors in the link function, i.e.

$$\mathsf{P}[\gamma_i = 1 \mid \theta, \mathscr{P}(\gamma_i)] = \frac{\exp(\theta_i + \sum_{j \sim i}^p \theta_{ij}\gamma_j)}{1 + \exp(\theta_i + \sum_{j \sim i}^p \theta_{ij}\gamma_j)}, \tag{7.2.1}$$

where $j \sim i$ designates the parent-child relation between jth and ith node. Here the parameters θ_i regulate the sparsity and entries in the adjacency matrix regulate smoothness of the ancestral transitions. The joint distribution can be then written in the product form

$$\pi(\gamma \mid \theta) = \prod_{i=1}^p \left(\frac{\exp(\theta_i + \sum_{j=1}^p \theta_{ij}\gamma_j)}{1 + \exp(\theta_i + \sum_{j=1}^p \theta_{ij}\gamma_j)} \right)^{\gamma_i} \left(\frac{1}{1 + \exp(\theta_i + \sum_{j=1}^p \theta_{ij}\gamma_j))} \right)^{1-\gamma_i}$$

or more compactly

$$\pi(\gamma \mid \theta) = \exp[\theta_1'\gamma + \gamma'\theta_2\gamma - \widetilde{\psi}(\theta_1, \theta_2, \gamma)], \tag{7.2.2}$$

where $\widetilde{\psi}(\theta_1, \theta_2, \gamma)) = \sum_{i=1}^p \log[1 + \exp(\theta_i + \sum_{j=1}^p \theta_{ij}\gamma_j)]$. As opposed to the MRF prior discussed in **Chapter 3**, here the adjacency matrix θ_2 is asymmetrical. Also note that the distribution (7.2.2) is no longer exponential family, since the "partition function" $\widetilde{\psi}(\theta_1, \theta_2, \gamma)$ also depends on γ. Large multi-layer networks preclude exact inference as the marginal distributions become intractable due to many combinatorial possibilities when summing over the model configurations. The product form of the distribution implies readily available conditionals $\mathsf{P}[\gamma_i \mid \theta, \gamma_j, 1 \leq j \neq i \leq p] = \mathsf{P}[\gamma_i = 1 \mid \theta, \mathscr{N}(\gamma_i)]$, which facilitate application of Gibbs sampling schemes to generate samples from the posterior model distribution (George and McCulloch, 1993; Li and Zhang, 2010). A deterministic inference to obtain approximations to the marginal quantities has been studied within the variational framework (Wainwright and Jordan, 2008) and large deviation bounds (Kearns and Saul, 1998). These approaches allow

obtaining approximate marginal expectations as solutions to a series of non-linear regressions. These inferential tools provide a promising first steps towards extensions of the EMVS procedure for directed graphs. The derivation of the E-step is further complicated by the fact that the prior distribution (7.2.2) does not combine nicely with the data to produce a conditional posterior distribution of the same probability class. Extensions of EMVS to directed acyclic graphs is a challenging future avenue.

7.2.3 Beyond Linear Regression

Throughout the thesis we focused mainly on variable selection in the linear regression case. Relaxing the distributional assumption, the method can be extended naturally to the generalized linear model framework. Whereas the data augmentation approach to probit regression permits a rather straightforward extension of the EMVS procedure for binary responses, generalizations to other GLMs are less obvious. Due to the separability of the EM algorithm, which places the variable selection indicators at the level of regression coefficients, the E-step does not necessitate further adjustments. The part which becomes problematic is the M-step, where the maximization with respect to the regression and dispersion coefficients is typically not obtainable in a closed form. A promising strategy to perform an approximate M-step is with the assistance of the conjugate dual coordinate ascent method introduced in **Chapter 3**, which allows for general loss functions. A closer investigation of this and other possibilities is of huge practical relevance, since the majority of research developments for variable selection merely permit normally distributed responses.

7.2.4 Sparse Precision Matrix Estimation

Straightforward extensions of the EMVS procedure can be implemented for estimating patterns of association between a set of random variables. One possible strategy to elucidate such relationships is by identifying nonzero elements in the inverse covariance matrix (precision/concentration matrix). The off-diagonal elements are the conditional covariances, given the remaining set of variables, where zero entries under Gaussianity imply conditional independence. The pattern of nonzero entries in the precision matrix can be interpreted as an undirected graph, where vertices (random variables) are connected if and only if the conditional covariance is nonzero.

There is a significant literature on model selection and parameter estimation in Gaussian concentration graphical models, starting with the seminal paper of Dempster et al. (1977). The traditional methods (Whittaker, 1990) are based on two steps: (1) identifying the pattern of sparsity, (2) estimating the parameters. One standard approach to estimating the pattern of association is the greedy stepwise selection achieved by sequential hypotheses testing (Edwards, 2000). Such a strategy is not applicable in the context $p >> n$ and is highly instable, as recognized by many authors including Breiman (1996). The model selection and estimation can be accomplished simultaneously with the assistance of penalized likelihood methods (Yuan and Lin, 2006; Rothman et al., 2008; Friedman et al., 2007), where penalties are induced on the

individual entries in the precision matrix and where the resulting estimate is guaranteed to be positive definite. Friedman et al. (2007) proposed a method called GLASSO and implemented an efficient coordinate descent procedure, exploiting dual representation and resemblance to the LASSO regression.

The maximum likelihood approach for sparse precision matrix estimation that extends the GLASSO method (Friedman et al., 2007) by considering non-concave penalties was proposed by Fan et al. (2009). The computation rests on iteratively estimating weighted penalized regressions in a way that is very much similar to the local approximation methods for non-concave variable selection. This is analogous to the posterior computation using an EM algorithm with sparsity priors. Considering a Laplace spike and slab mixture on the individual precision matrix entries, we can derive an EM algorithm which solves the weighted GLASSO at every iteration of the M-step. Exploiting the efficient GLASSO solutions in combination with the closed form E-step provides a promising rapid strategy to perform Bayesian model selection in Gaussian graphical models.

Apart from the maximum likelihood methods, there has been an emergence of alternative approaches which also benefit from the resemblance to variable selection, but do so in a different way. Meinshausen and Bühlmann (2006) proposed the neighborhood selection strategy, where every variable is regressed on the other variables in a separate linear regression model. Because each regression coefficient is proportional to the associated entry in the precision matrix, the identification of zeroes can be accomplished by applying variable selection to each regression. This strategy however neglects the symmetry of the problem and can lead to sign inconsistency in the estimated partial correlation matrix. As a remedy, Peng et al. (2009) suggested a joint sparse regression model, that estimates only diagonal and upper triangle of the precision matrix. There is also a class of methods that benefit from the natural ordering of the variables (longitudinal data, spatial data) (Wu and Pourahmadi, 2003; Huang et al., 2007). These methods typically exploit the modified Cholesky decomposition of the precision matrix, which leads to fitting a sequence of lagged regressions, such as in the spike and slab approach of Smith and Kohn (2002).

Due to the versatility of formulations for precision matrix estimation, there is a huge potential for the implementation of EMVS variants for model selection in Gaussian graphical models.

CHAPTER 8

NEDERLANDSE SAMENVATTING, CV AND ACKNOWLEDGMENTS

Nederlandse Samenvatting

Hoog dimensionele data komen in de laatste decennia veelvuldig voor. Echter, om op een valide wijze relevante informatie te extraheren uit zulke, vaak enorme, databestanden en zo te komen tot goed onderbouwde conclusies, zijn specifieke exploratieve statistische technieken nodig. Het is echter een uitdaging om het meest geschikte model te vinden en te bepalen welke en hoeveel factoren de respons(en) het best bepalen.

In **Hoofdstuk 2** onderzoeken we hoe de onzekerheid die gepaard gaat met het selecteren van het meest geschikte model kan uitgedrukt worden als een coherente kansmaat onder het Bayesiaanse paradigma. We tonen aan dat het veelzijdige karakter van het Bayesiaanse formalisme tot een explosie van modelselectietechnieken geleid heeft. Hiervan wordt een beknopte samenvatting gegeven in Hoofdstuk 2. In dat hoofdstuk bekijken we ook het gedrag van een aantal Bayesiaanse variabelenselectietechnieken en vergelijken deze met klassieke technieken. Deze oefening werd uitgevoerd op gesimuleerde data. We tonen hierbij de praktische voordelen aan van het gebruik van a posteriori modelkansen als middel om variabelen te selecteren. Verder tonen we ook aan dat de 'spike en slab' techniek methodologische voordelen biedt. Tot slot illustreren we in dit hoofdstuk de praktische problemen bij het bepalen van de a posteriori kansen, met name hun rekenintensief karakter dat belemmerend kan worden in het geval van een groot aantal covariaten. Dit probleem wordt aangepakt in **Hoofdstuk 3**.

Het opsporen van interessante modellen noodzaakt de implementatie van efficiënte zoekstrategieën die rekening houden met de aard van de gegevens en met de grootte van het databestand maar ook met grote zekerheid de belangrijkste modellen kunnen selecteren. Echter in het geval van hoog-dimensionele modelruimten, liggen de meest interessante modellen verspreid in de ruimte rondom een beperkt aantal geïsoleerde pieken van relatief hoge a posteriori kans. Daardoor is het bijna onvermijdelijk om computationele trucs te hanteren, zoals bijvoorbeeld "parallel computing" of "distributed computing". Deze technieken hebben echter een zware rekenintensieve impact. In **Hoofdstuk 3** tonen we aan dat met geconjugeerde continue "spike and slab" priorverdelingen, men een deterministisch mechanisme kan hanteren dat de a posteriori meest waarschijnlijke modellen kan opsporen met een enorme tijdswinst in vergelijking met de bestaande stochastische methoden. Deze deterministische techniek is gebaseerd op het populaire Expectance-Maximisation (EM) algoritme. Hierbij worden aan de geobserveerde data "missende" indicatoren voor variable inclusie toegevoegd. Het gebruik van continue "spike and slab" priorverdelingen is essentieel voor het bekomen van analytische uitdrukkingen van de E- en M-step van het algoritme. Een eenvoudige kansregel leidt dan tot het identificeren van lokaal zeer waarschijnlijke a posteriori modellen door het vergelijken van de lokale modes met grenswaarden aangegeven door de conditionele mediaan kansmodel regel. Het EM mechanisme laat een snelle exploratie toe van de meest waarschijnlijke modellen en produceert gekrompen geschatte regressieparameters. De keuze van de prior verdelingen is bepalend voor de posterior verdeling van de modellen. Namelijk, scherp gepiekte priorverdelingen induceren typisch meerdere modes in de posteriorverdeling, daarentegen worden meer gladde posteriorverdelingen bekomen met priorverdelingen die een

minder scherpe piek vertonen. Dit laatste wordt bekomen door de piekvariantie te vergroten. Dit motiveerde ons om een reeks van 'spike and slab' priorverdelingen te beschouwen met een vergrotende piekvariantie, waarbij onbelangrijke regressiecoëfficiënten geabsorbeerd worden in de piekverdeling en zo slechts een klein aantal regressiecoëfficiënten overblijven. Dit dynamisch exploreren van modellen laat ons toe om de reeks van geselecteerde modellen weer te geven in nieuw 'spike and slab' regularizatie diagram. Deze figuur toont dan een sequentie van in toenemende mate lokaal schaarse modellen die kunnen geëvalueerd worden via hun a posteriori kans. Dit levert een techniek op analoog aan het cross-validatie criterium voor het selecteren van de optimale 'penalty' frequentistische regularizatie methoden. De techniek van het dynamisch exploreren en evalueren van modellen hebben we EMVS genoemd. We beschouwen deze EM methode als de determinische tegenhanger van de stochastische variable selectie zoekmethode (SSVS) van George en McCulloch (1993). Het multimodale karakter van de posteriorverdeling motiveerde ons om ook deterministiche tempering te bestuderen. Deze techniek genereert gladde doelfuncties door het temperen van de posteriorverdeling waardoor de kans om een globale mode te vinden verhoogd wordt. Om de enorme rekenwinst die EMVS met zich meebrengt en de elegantie van het dynamisch selecteren van variabelen te illustreren, hebben we onze techniek toegepast op tal van gesimuleerde gegevens. In **Hoofdstuk 3** hebben we ook Markov Random Velden ingeschakeld voor het modelleren van de priorverdelingen van de variabelenselectieindicatoren. Dit laat toe om een netwerk structuur op te sporen in de covariaten.

Verschillende technieken werden voorgesteld om aan gestructureerde variabele selectie te doen. In dit verband stellen een groot aantal methodes voor variabelen op gelijktijdige en gegroepeerde manier te selecteren. In **Hoofdstuk 4** stellen we een Bayesiaanse variabele selectie variant voor, waarbij het gegroepeerd selecteren ingebed is in een krimpingsschattingstechniek. Echter onze veronderstellingen zijn minder restrictief dan wat de meeste gegroepeerde variabelenselectie technieken onderstellen, namelijk: (1) gladheid van de regressiecoëfficiënten binnen een groep, (2) voorspellingkracht van alle variabelen binnen een groep, (3) orthogonaliteit van de groepidentificatiematrix. We analyseren de eigenschappen van twee EM algoritmes, waarbij variabelenselectie verkregen wordt via de identificatie van de meest a posteriori waarschijnlijk schatters die nul of bijna nul zijn. Elke groep van variabelen is gekarakteriseerd door een te bepalen krimpingsfactor, die alle regressiecoëfficiënten binnen de groep globaal naar nul drijft. Door het gezamenlijk regresseren van de responsen op de predictoren en de regressie van de krimpingsfactoren op de groepidentificatiematrix, proberen we zowel predictieve groepen als predictieve variabelen binnenin groepen te ontdekken. Toepassingen op reële en gesimuleerde data tonen het voordeel van onze techniek aan tegenover bestaande technieken.

Tot slot onderzoeken we in **Hoofdstuk 5** de situatie waarbij meerdere gerelateerde responsen beschikbaar zijn waarvoor we groepen van predictieve covariaten willen selecteren en tegelijkertijd de groepstructuur van de residuele covariantiematrix willen ontdekken. Hiervoor breiden we de klassieke multivariate regressie techniek uit door latente factoren in te sluiten die de residuele collineariteit tussen de meerdere responsen kunnen beschrijven. In

de geest van deze thesis, verkrijgen we schaarse modellen door het introduceren van variabelenselectie priorverdelingen op de regressiecoëfficiënten en de factorladingen. We tonen aan dat het EM algoritme dat ontwikkeld werd voor probabilistische hoofdcomponentenanalyse elegant combineert met de EMVS procedure. Dit maakt het mogelijk om op een snelle wijze de belangrijkste factorregressiemodellen te exploreren, dat in tegenstelling tot de bestaande technieken die op MCMC berekeningen gebaseerd zijn.

Hoofdstuk 6 beschrijft een integrale analyse van bekende biomarkers in acute myeloïde leukemie die het enorme praktische nut laat zien van variabelenselectie in biomedische toepassingen.

Curriculum Vitae

Personal Information

First Name	Veronika
Family Name	Ročková
Date of Birth	August 1st, 1985
Place of Birth	Pardubice, Czech Republic
Nationality	Czech

Professional Activities

10/2009-10/2013	**Ph.D. Student in Training**
	Erasmus MC, Erasmus University Rotterdam
	The Netherlands
Doctoral Thesis	*Bayesian Variable Selection in High-dimensional Applications*
Supervisors	Prof. Emmanuel Lesaffre and Prof. Bob Löwenberg

Education

10/2007-9/2010	**M.Sc. in Mathematical Statistics**
	Faculty of Mathematics and Physics, Charles University in Prague
	Czech Republic
Master Thesis	*Non-negative Time Series*
Supervisor	Prof. Jiri Andel
10/2008-9/2009	**M.Sc. in Biostatistics**
	Faculty of Sciences, Universiteit Hasselt
	Belgium
Master Thesis	*Prediction of Rheumatoid Arthritis on the Basis of Laboratory and Clinical Parameters*
Supervisor	Prof. Emmanuel Lesaffre
10/2007-9/2010	**B.Sc. in General Mathematics**
	Faculty of Mathematics and Physics, Charles University in Prague
	Czech Republic
Bachelor Thesis	*Testing for Periodicity in Time Series*
Supervisor	Prof. Jiri Andel

Invited Talks and Seminars

6/2013 **Research Seminar**
Centre de Recherche en Economie et Statistique, Paris, France

6/2013 **Workshop on Bayesian Nonparametrics**
University of Leiden (Lorentz Center), The Netherlands

5/2013 **BAYES 2013**
Erasmus University Rotterdam, The Netherlands

12/2012 **Workshop on Variable Selection**
Universiteit Hasselt, Belgium

10/2012 **Statistics Seminar**
University of Pennsylvania (The Wharton School), Philadelphia, USA

5/2012 **Statistics Seminar**
University of Pennsylvania (The Wharton School), Philadelphia, USA

3/2012 **Spring Symposium in Biostatistics**
Erasmus University Rotterdam, The Netherlands

Poster Presentations at International Meetings

6/2013 **High-Dimensional Inference with Applications**
University of Kent, Great Britain

6/2013 **9th Conference on Bayesian Nonparametrics**
Amsterdam, The Netherlands

Contributed Talks at International Meetings

8/2013 **Joint Statistical Meeting (JSM)**
Montreal, Canada

7/2013 **International Workshop in Statistical Modeling (IWSM)**
Palermo, Italy

8/2012 **Conference of International Society of Clinical Biostatisics (ISCB)**
Bergen, Norway

6/2012 **International Workshop in Statistical Modeling (IWSM)**
Prague, Czech Republic

2/2012 **Dutch Hematology Congress (DHC)**
Arnhem, The Netherlands

9/2010 **Conference of International Society of Clinical Biostatistics (ISCB)**
Montpellier, France

Participation at International Meetings

3/2013 **Recent Advances in Statistical Inference: Theory and Case Studies**
Padova, Italy

7/2012 **Joint Statistical Meeting**
San Diego, USA

4/2011 **Symposium on Joint Modelling**
Rotterdam, The Netherlands

1/2011 **Workshop on Bayesian Inference for Latent Gaussian Models**
Zurich, Switzerland

4/2010 **Symposium on Bayesian Methods**
Rotterdam, The Netherlands

3/2010 **Workshop on Mixed Modelling and Convergence**
Dublin, Ireland

Manuscripts and Peer Reviewed Publications

Faster Spike-and-Slab Variable Selection with Dual Coordinate Ascent EM
George E., Rockova V. and Lesaffre E. (2013)
Proceedings of the 28th International Workshop in Statistical Modeling, ISBN: 978-88-96251-47-8, pages 165-171

Sparse Bayesian Factor Regression Approach to Genomic Data Integration
Rockova V. and Lesaffre E. (2013)
Proceedings of the 28th International Workshop in Statistical Modeling ISBN: 978-88-96251-47-8, pages 337–343

Incorporating Grouping Information in Bayesian Variable Selection with Applications in Genomics
Rockova V. and Lesaffre E. (2012)
To appear in
Bayesian Analysis

EMVS: The EM Approach to Bayesian Variable Selection
Rockova V. and George E. (2012)
Tentatively accepted
Journal of American Statistical Association (Theory and Methods)

Incorporating Prior Biological Knowledge in Bayesian Modeling of Sparse Networks
Rockova V. and Lesaffre E. (2012)
Proceedings of the 27th International Workshop in Statistical Modeling (2012), ISBN: 978-80-263-0250-6, pages 291– 296

Bayesian Hierarchical Formulations for Selecting Variables in Regression Models
Rockova V., Lesaffre E., Luime J. and Löwenberg B.
Statistics in Medicine (2012), **31**(11), 1221–1237

Risk-stratification of Intermediate-risk Acute Myeloid Leukemia: Integrative Analysis of a multitude of gene mutation and expression markers
Rockova V., Abbas S., Wouters B.J., Erpelinck C., Beverloo B., Delwel R., van Putten W., Löwenberg B. and Valk P.
Blood (2011), **118**(4), 1069-76

The Prognostic Relevance of miR-212 Expression with Survival in Cytogenetically and Molecularly Heterogeneous AML
Sun S., Rockova V., Bullinger L., Dijkstra M., Döhner H., Löwenberg B., Jongen-Lavrencic M.
Leukemia (2013), **27**(1), 100-6

Mutant DNMT3A: a Marker of Poor Prognosis in Acute Myeloid Leukemia
Ribeiro A., Pratcorona M., Erpelinck C., Rockova V., Sanders M., Abbas S., Figueroa M., Zeilemaker Z., Melnick A., Löwenberg B., Valk P. and Delwel R.
Blood (2012), **119**(24), 5824-31

Retroviral Integration Mutagenesis in Mice and Comparative Analysis in Human AML Identify Reduced PTP4A3 Expression as a Prognostic Indicator
Beekman E., Valkhof M., Erkeland S., Taskesen E., Rockova V., Peeters J., Valk P., Löwenberg B. and Touw I.
PLoS ONE (2011), **6**(10), e26537

Deregulated Expression of EVI1 Defines a Poor Prognostic Subset of MLL-Rearranged Acute Myeloid Leukemias: A Study of the German-Austrian Acute Myeloid Leukemia Study Group and the Dutch-Belgian-Swiss HOVON/SAKK Cooperative Group
Groschel S., Schlenk R., Engelmann J., Rockova V., Teleanu V., Kühn M., Eiwen K., Erpelinck C., Havermans M., Lubbert M., Germing U., Schmidt-Wolf I., Beverloo B., Schuurhuis G., Ossenkoppele G., Schlegelberger B., Verdonck L., Vellenga E., Verhoef G., Vandenberghe P., Pabst T., Bargetzi M., Krauter J., Ganser A., Valk P., Löwenberg B., Döhner K., Döhner H., Delwel R.
Journal of Clinical Oncology (2013), **31**(1), 95-103

Acknowledgments

This thesis is a testimony to four gratifying years of research at Erasmus University Rotterdam. It is where I embarked on the journey of a junior researcher, where I best learned about the rewards of perseverance, multidisciplinary research and independent scientific pursuits. I look proudly upon those years, marked by the ripening of ideas and sustained efforts that have culminated into this document.

Although the years at Erasmus University were the most creative and formative, it was much earlier during my graduate studies at Charles University in Prague, where I first identified my fondness towards mathematics and statistics. I am very grateful to all my tutors at the Faculty of Mathematics and Physics, especially my graduate thesis advisor prof. **Jiří Anděl**, for fully engaging my interest in these disciplines in my early student days. Nevertheless, it was not until my expedition to Belgium, University Hasselt, where I discovered the immense potential of statistics in biomedical applications. My graduate thesis advisor, prof. **Emmanuel Lesaffre**, played a pivotal role in introducing me into the fascinating world of biostatistics by offering me the wonderful opportunity to pursue a doctorate under his supervision at the Department of Biostatistics at Erasmus University Rotterdam. Emmanuel's enthusiasm, incredible breadth of expertise, approachable and friendly attitude to his students, and genuine care for solving relevant practical problems have been a great source of inspiration for me. It was Emmanuel's excellent instruction of the Bayesian course at University Hasselt and his proposed graduate thesis topic on Bayesian CART, which were instrumental for me to begin studying Bayesian methods for model selection, a direction which turned out to be so gratifying. During many of our illuminating discussions, Emmanuel recognized my inclination towards more technical aspects of statistical modeling and directed me towards interesting topics on high-dimensional genomic data. This was the beginning of what turned out to be a unique bi-departmental collaboration between the Department of Biostatistics and the Department of Hematology at Erasmus MC. This challenging pursuit opened a whole new world of learning opportunities for me. A part of this unique experience was the honorable opportunity to collaborate with prof. **Bob Löwenberg**, an authority whose kindness as my advisor has served me as an example in many different ways. My interaction with the Hematology Department provided me with an invaluable experience. It made me realize the huge practical consequences of statistics and deepen my respect towards practically oriented biostatistical research. The plentiful scientific departmental discussions and research meetings, especially with **Bob**, **Peter**, **Ruud**, **Mojca**, were very illuminating in my understanding of the challenging background of top level cancer research. I would also like to thank **Ivo** and **Stefan** for all their genuine and insightful comments during my work discussion presentations. I am very grateful for this enriching experience, which served to broaden my field of knowledge. I am happy to have contributed to the variety of expertise at the Hematology Department.

The course of my doctorate would have been a lot different if it had not been for the thoughtful interventions of my advisors. I am particularly indebted to Emmanuel for his continuous efforts to establish bi-departmental and bi-continental collaborations. These generated

multiple opportunities for me to visit many exciting places such as Cornell University and the University of Pennsylvania. The latter turned out to be more than influential in shaping the research contained within this thesis. I have been privileged to have the honorable opportunity to work with prof. **Ed George**. During my many visits to the Statistics Department at the Wharton School, we set the ground for a solid and fruitful collaboration, whose results form a good portion of work presented herein. I am deeply indebted to Ed for giving me the invaluable opportunity to discuss my thoughts about Bayesian variable selection. His insightful comments, astonishing enthusiasm, generosity, never-ending encouragement and support were essential for developing my doctoral research.

On more personal note, the years at Erasmus University have blessed me with excellent colleagues, many of them now turned into very good friends. Especially, my long-time officemates **Susan** and **Elrozy** provided me with such a warm and friendly working environment. Thank you so much for being so supportive when times were not all that great. I feel privileged to have the two of you as my “paranimfen". The list of good colleagues/friends goes even further. **Sten**, you are one of the people at the department who has known me the longest. Thank you for the many interesting illuminating research discussions. I am thrilled to have you as my friend and I am very happy that we have wrapped our Erasmus experience by being office mates for the last few months. Among other biostats people, I would like to thank all my fellow students: **Marek**, **Kazem**, **Nahid**, **Nicole**, **Li**, **Johan**, **Siti**, **Magdalena** and **Karolina** for creating a lovely atmosphere at the department. To continue, I would like to thank **Eline** for her genuine care, **Dimitris**, **Joost** and **Paul** for their valuable scientific contributions.

One of the great things about splitting my time in between the Hematology and Biostatistics Departments was the opportunity to meet so many kind, wonderful people. Among my peers, there have been plenty that I care about as my dear friends and with whom the interactions outside office were so much fun. Thank you: **Marshall**, for hosting the best Halloween party, **Noemi** for the kind personal advice on Italian and other matters, **Kasia** for being an incredibly kind and fair friend, **Roberto** for the patient hours of discussion, support and advice. I will never forget the nights out with **Onno**, **Joyce**, **Jana**, **Amiet**, **Roel**, **Helen**, **Erik**, **Rasty**, **Suming**, the great ski trip and the multitude of other events we have done together. I would also like to thank all the guys I shared my hematology office with over the years: **Mark**, for the numerous microRNA discussions over e-mail and **Elwin**, **Stefan**, **Erik**, **Amiet** for the plentiful jokes in the office. Lastly, my thanks go also to **Mr. Green**, **Mr. Lion** and **Alvin** and all the people that made my Rotterdam experience memorable.

Finally and most importantly, my deepest gratitude goes to my parents **Ilona** and **Václav**, who supported me throughout all of my study adventures far away from my home. Thank you for being there for me no matter what, for your multiple visits to Benelux, for raising me in the warmest loving environment, and for tolerating my obsession with studies. Děkuji za vaši neustálou podporu při mých studijích stovky kilometrů od domova, za vaši trpelivost s mým pracovním nasazením, za vaši starostlivost a péči, které si nadevše cením. Bez vaší pomoci a lásky by žadný z mých smělých počinů nebyl vůbec možný. My deepest gratitude also goes to my soul-mate sister, **Ilona**, whose personality has always been very influential

and inspirational to me and who I admire for her incredible strength. Ilonko, děkuji moc za tvoji lásku.

BIBLIOGRAPHY

Abbas, S., Lugthart, S., Kavelaars, F., and et al. (2010). Acquired mutations in the genes encoding IDH1 and IDH2 both are recurrent aberrations in acute myeloid leukemia: prevalence and prognostic value. *Blood*, 116:2122–2126.

Abramovich, F., Angelini, C., and De Canditiis, D. (2007). On optimality of Bayesian testimation in the normal means problem. *Annals of Statistics*, 35:2261–2286.

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

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Andrews, D. R. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B*, 36:99–102.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96:939–967.

Armagan, A., Dunson, D., and Lee, J. (2012). Generalized double Pareto shrinkage. Technical report, Duke University.

Bae, K. and Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, 20:3423–3430.

Baldus, C., Tanner, S., Ruppert, A., and et al. (2003). BAALC expression predicts clinical outcome of de novo acute myeloid leukemia patients with normal cytogenetics. *Blood*, 102:1613–1618.

Bar, H., Booth, J., and Wells, M. (2010). An empirical Bayes approach to variable selection and QTL analysis. *In Proceedings of the 25th International Workshop on Statistical Modelling*, pages 63–68.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32:870–897.

Berger, J., Pericchi, L., and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *The Indian Journal of Statistics, Series A*, 60:307–321.

Bergmann, L., Miething, C., Maurer, U., and et al. (1997). High levels of Wilm's tumor gene (WT1) mRNA in acute myeloid leukemias are associated with a worse long-term outcome. *Blood*, 90:1217–1225.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B*, 36:192–236.

Bhattacharya, A. and Dunson, D. (2011). Sparse Bayesian infinite factor models. *Biometrika*, 98:291–306.

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5:583–618.

Bousquet, M., Quelen, C., Rosati, R., and et al., M.-D. M. (2008). Myeloid cell differentiation arrest by miR-125b-1 in myelodysplastic syndrome and acute myeloid leukemia with the t(2;11)(p21;q23) translocation. *Journal of Experimental Medicine*, 11:2499–506.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Boston.

Breems, D., van Putten, W., de Greef, G., and et al. (2008). Monosomal karyotype in acute myeloid leukemia: a better indicator of poor prognosis than a complex karyotype. *Journal of Clinical Oncology*, 26:4791–4797.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383.

Brown, P., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of Royal Statistical Society. Series B*, 60:627–641.

Bryant, A., Palma, C., Jayaswal, V., Yang, Y., Lutherborrow, M., and Ma, D. (2012). MiR-10a is aberrantly overexpressed in nucleophosmin1 mutated acute myeloid leukaemia and its suppression induces cell death. *Molecular Cancer*, 11:1–11.

Bussemaker, H., Li, H., and Siggia, E. (2001). Regulatory elements detection using correlation with expression. *Nature Genetics*, 27:167–171.

Carlin, B. P. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B*, 57:473–484.

Carlin, B. P. and Polson, N. G. (1991). Inference for nonconjugate Bayesian models using the Gibbs sampler. *The Canadian Journal of Statistics*, 19:399–405.

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

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*, 103:1438–1456.

Chamuleau, M., van de Loosdrecht, A., Hess, C., and et al. (2008). High INDO (indoleamin 2,3-dioxygenase) mRNA level in blasts of acute myeloid leukemic patients predicts poor clinical outcome. *Haematologica*, 93:1894–1989.

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

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24:17–36.

Choe, G., Horvath, S., Cloughesy, T., Crosby, K., and et al. (2003). Analysis of the phosphatidylinositol 3'-kinase signaling pathway in glioblastoma patients in vivo. *Cancer Research*, 63:2742–2746.

Choi, N., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105:354–364.

Copas, J. B. (1983). Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B*, 45:311–354.

Cui, W. and George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138:888–900.

Dawid, P. (1981). Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68:265–274.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12:27–36.

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*, 39:1–38.

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.

Döhner, K., Schlenk, R., Habdank, M., and et al. (2005). Mutant nucleophosmin (NPM1) predicts favorable prognosis in younger adults with acute myeloid leukemia and normal cytogenetics: interaction with other gene mutations. *Blood*, 106:3740–3746.

Edwards, D. (2000). *Introduction to Graphical Modelling*. Springer-Verlag.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, 32:407–499.

Fabian, M., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA translation and stability by microRNAs. *Annual Review of Biochemistry*, 79:351–379.

Fahrmeir, L., Kneib, T., and Konrath, S. (2010). Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing*, 20:203–219.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *Annals of Applied Statistics*, 3:521–541.

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.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148.

Figueiredo, M. (2002). Adaptive sparseness using Jeffreys prior. *Advances in Neural Information Processing Systems*, 14:697–704.

Figueiredo, M. A. (2003). Adaptive sparseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1150–1159.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35:109–135.

Friedman, J., Hastie, H., and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical LASSO. *Biostatistics*, 9:432–441.

Friedman, J., Hastie, T., and Tibshirani, R. (2010a). A note on the group LASSO and a sparse group LASSO. *Submitted manuscript*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010b). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 22:1–22.

Frühwirth-Schnatter, S. and Lopes, H. (2009). Parsimonious Bayesian factor analysis when the number of factors is unknown. *Technical report, University of Chicago Booth School of Business*.

Fu, W. J. (1998). Penalized regressions: The bridge versus the LASSO. *Journal of Computational and Graphical Statistics*, 7:397–416.

Geest, R. and Coffer, P. (2009). MAPK signaling pathways in the regulation of hematopoiesis. *Journal of Leukocyte Biology*, 86:237–250.

Gelfand, A. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4:11–15.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

George, E. and Foster, D. (1997). Calibration and empirical Bayes variable selection. *Biometrika*, 87:731–747.

George, E. I., McCulloch, R., and Tsay, R. (1996). Two approaches to Bayesian model selection with applications. In *Bayesian Analysis in Statistics and Econometrics*, pages 339–348. John Wiley & Sons.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Geweke, J. and Singleton, K. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, 85:133–137.

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.

Goldsmith, J., Huang, L., and Crainceanu, C. M. (2013). Smooth scalar-on-image regression via spatial Bayesian variable selection. *Journal of Computational and Graphical Statistics*, to appear.

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

Gradshteyn, I. and Ryzhik, E. (2000). *Table of Integrals Series and Products*. Academic Press.

Gramacy, R. and Polson, N. (2012). Simulation-based regularized logistic regression. *Bayesian Analysis*, 7:567–590.

Gramacy, R. B. and Pantaleo, E. (2010). Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. *Bayesian Analysis*, 5:237–262.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732.

Griffin, J. and Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report, University of Warwick, University of Kent.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5:17–188.

Griffin, J. E. and Brown, P. J. (2012). Bayesian hyper-LASSOS with non-convex penalization. *Australian & New Zealand Journal of Statistics*, 53:423–442.

Hans, C. (2009). Bayesian LASSO regression. *Biometrika*, 96:835–845.

Hans, C. (2010). Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing*, 20:221–229.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102:507–516.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hayashi, T. and Iwata, H. (2010). EM algorithm for Bayesian estimation of genomic breeding values. *BMC Genetics*, 11:1–9.

Hocking, R. (1976). The analysis and selection of variables in linear regression. *Annals of Statistics*, 32:1–49.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, 1:145–168.

Horvath, S., Zhang, B., Carlson, M., Lu, K., and et al. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a molecular target. *In Proceedings of National Academy of Sciences of United States of America*, 103:17402–17407.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15:651–674.

Huang, J., Liu, L., and Liu, N. (2007). Estimation of large covariance matrices of longitudinal data with basis function approximations. *Journal of Statistical Computation and Graphics*, 16:189–209.

Huang, J. and Morris, Q. (2007). Bayesian inference of MicroRNA targets from sequence and expression data. *Journal of Computational Biology*, 14:550–563.

Hunter, D. and Li, R. (2005). Variable selection using MM algorithms. *Annals of Statistics*, 33:1617–1642.

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.

Ishwaran, H. and Rao, J. S. (2003). Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association*, 98:438–455.

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

Jacob, L., Obozinski, G., and Vert, J. (2009). Group LASSO with overlap and graph LASSO. *In Proceedings of* 26^{th} *International Conference on Machine Learning*.

Jasra, A., Stephens, D., and Holmes, C. (2007). On population-based simulation for static inference. *Statistics and Computing*, 17:263–279.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, 32:1594–1649.

Jongen-Lavrencic, M., Sun, S., Dijkstra, M., and Valk, P. (2008). MicroRNA expression profiling in relation to the genetic heterogeneity of acute myeloid leukemia. *Blood*, 111:5078–5085.

Kanda, Y., Hamaki, T., Yamamoto, R., and et al. (2000). The clinical significance of CD34 expression in response to therapy of patients with acute myeloid leukemia: an overview of 2483 patients from 22 studies. *Cancer*, 88:2529–2533.

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

Karakas, T., Maurer, U., Weidmann, E., Miething, C., and et al. (1998). High expression of BCL-2 mRNA as a determinant of poor prognosis in acute myeloid leukemia. *Annals of Oncology*, 9:159–165.

Kearns, M. and Saul, L. (1998). Large deviation methods for approximate probabilistic inference. *In Proceedings of the 14th Annual Conference on Uncertainty in Artifficial Intel ligence*, pages 63–68.

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.

Kim, S., Dahly, D. B., and Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis*, 4:707–732.

Kirkpatrick, S., Gelatt, C., and Vecchi, M. (1983). Optimization by simulated annealing. *Science*, 220:671–680.

Kwon, D., Landi, M., Vannucci, M., Issaw, H., Prieto, D., and Pfeiffer, R. (2011). An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors . *Computational Statistics and Data Analysis*, 55.

Kwon, D., Tadesse, M. G., Sha, N., Pfeiffer, R. M., and Vannucci, M. (2007). Identifying biomarkers from mass spectrometry data with ordinal outcome. *Cancer informatics*, 3:19–28.

Kyung, M., Gilly, J., Ghoshz, M., and Casella, G. (2010). Penalized regression, standard errors, and Bayesian Lassos. *Bayesian Analysis*, 5:369–412.

Langer, C., Marcucci, G., Holland, K., and et al. (2009). Prognostic importance of MN1-associated gene and microRNA expression signatures in cytogenetically normal acute myeloid leukemia. *Journal of Clinical Oncology*, 27:3198–3204.

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. John Wiley & Sons.

Leeb, H. and Potscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59.

Levis, M. and Small, D. (2003). FLT3:ITD does matter in leukemia. *Leukemia*, 17:1738–1752.

Li, C. and Li, H. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Biometrics*, 24:1175–1182.

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*, 105:1978–2002.

Li, Q. and Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, 5:151–170.

Li, Z., Huang, H., Chen, P., He, M., and Li, Y. e. a. (2012a). MiR-196b directly targets both HOXA9/MEIS1 oncogenes and FAS tumour suppressor in MLL-rearranged leukaemia. *Nature Communications*, 2:1–6.

Li, Z., Huang, H., Li, Y., Jiang, X., and et al. (2012b). Up-regulation of a HOXA-PBX3 homeobox-gene signature following down-regulation of miR-181 is associated with adverse prognosis in patients with cytogenetically abnormal AML. *Leukemia*, 119:2314–2324.

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*, 30:410–423.

Liang, F. and Wong, W. (2000). Evolutionary Monte Carlo: Applications to c_p model sampling and change point problem. *Statistica Sinica*, 10:317–342.

Liu, Y. and Wu, Y. (2007). Variable selection via a combination of the l_0 and l_1 penalties. *Journal of Computation and Graphical Statistics*, 16:782–798.

Lopes, H. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, 14:41–67.

Löwenberg, B., Boogaerts, M., Daenen, S., and et al. (1997). Value of different modalities of granulocyte-macrophage colony-stimulating factor applied during or after induction therapy of acute myeloid leukemia. *Journal of Clinical Oncology*, 115:3496–3506.

Lugthart, S., van Drunen, E., van Norden, Y., and et al. (2008). High EVI1 levels predict adverse outcome in acute myeloid leukemia: prevalence of evi1 overexpression and chromosome 3q26 abnormalities underestimated. *Blood*, 111:4329–4337.

Lunn, D. J., Best, N., and Whittaker, J. C. (2009). Generic reversible jump MCMC using graphical models. *Journal Statistics and Computing*, 19:395–408.

Madigan, D., Raftery, A., Wermuth, N., York, J., and Zucchini, W. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89:1535–1546.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review / Revue Internationale de Statistique*, 63:215–232.

Marcucci, G., Haferlach, T., and Döhner, H. (2011). Molecular genetics of adult acute myeloid leukemia: prognostic and therapeutic implications. *Journal of Clinical Oncology*, 29:475–486.

Marcucci, G., Maharry, K., Whitman, S., and et al. (2007). High expression levels of the ETS-related gene, ERG, predicts adverse outcome and improve molecular rusk-based classification of cytogenetically normal acute myeloid leukemia. *Journal of Clinical Oncology*, 25:3337–3343.

Maruyama, Y. and George, E. I. (2011). Fully Bayes factors with a generalized g-prior. *Annals of Statistics*, 39:2740–2765.

McDonald, J., Dunmire, V., Taylor, E. Sawaya, R., 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.

McLachlan, G. J. and Basford, K. (2004). *Mixture Models: Inference and Application to Clustering*. Marcel Dekker.

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

Meier, L., Van de Geer, S., and Bühlmann, P. (2008). The group LASSO for logistic regression. *Journal of the Royal Statistical Society. Series B*, 70:53–71.

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

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21:1087–1092.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032.

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.

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*, 203:779–787.

Nott, D. J. (2008a). Bayesian methods for highly correlated exposure data. *Epidemiology*, 28:199–207.

Nott, D. J. (2008b). Predictive performance of Dirichlet process shrinkage methods in linear regression. *Computational Statistics & Data Analysis*, 52:3658–3669.

Ntzoufras, I. (2002). Gibbs variable selection using BUGS. *Journal of Statistical Software*, 7:1–19.

Ntzoufras, I., Forster, J. J., and Dellaportas, P. (2000). Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation*, 68:23–37.

Ozeki, K., Kiyoi, H., Hirose, Y., and et al. (2004). Biologic and clinical significance of the FLT3 transcript level in acute myeloid leukemia. *Blood*, 103:1901–1908.

Pan, W., Benhuai, X., and Xiaotong, S. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, 66:474–484.

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

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

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

Peng, J., Wang, P., Zhou, N., and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104:735–746.

Polson, N. and Scott, J. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–539.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*. Springer-Verlag.

Rockova, V. and George, E. (2013). EMVS: The EM approach to Bayesian variable selection. *Tentatively accepted by the Journal of the American Statistical Association*.

Rothman, A. J., Bickel, P. J., Levina, E., , and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515.

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.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Sciumč, 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*, 111:3626–3634.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38:2587–2619.

Sha, N., Tadesse, M. G., and Vannucci, M. (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics*, 22:2262–2268.

Sha, N., Vannucci, M., Tadesse, M. G., Brown, P. J., Dragoni, I., Davies, N., Roberts, T. C., Contestabile, A., Salmon, M., Buckley, C., and Falciani, F. (2004). Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60:812–819.

Shalev-Shwartz, S. and Zhang, T. (2013). Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14:567–599.

Smith, M. and Fahrmeir, L. (2007). Spatial Bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102:417–431.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–344.

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*, 97:1141–1153.

Smith, S. M., Putz, B., Auer, D., and Fahrmeir, L. (2003). Assessing brain activity through spatial Bayesian variable selection. *NeuroImage*, 20:802–815.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9:3273–3297.

Stegle, O., Sharp, K., and Winn, J. (2000). A comparison of inference in sparse factor analysis. *Journal of Machine Learning Research*, 1:1–48.

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. *Annals of Applied Statistics*, 5:1202–1214.

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.

Stingo, F. and Vannucci, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics*, 27:495–501.

Strawderman, W. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics*, 42:385–388.

Sun, S., Rockova, V., Bullinger, L., Dijkstra, M., Döhner, H., Löwenberg, B., and Jongen-Lavrencic, M. (2013a). The prognostic relevance of miR-212 expression with survival in cytogenetically and molecularly heterogeneous AML. *Leukemia*, 27:100–106.

Sun, Y., Lin, K., and Chen, Y. (2013b). Diverse functions of miR-125 family in different cell contexts. *Journal of Hematology & Oncology*, 6:1–6.

Tadesse, M., Vannucci, M., and Lio, P. (2004). Identification of DNA regulatory motifs using Bayesian variable selection. *Bioinformatics*, 20:2553–2561.

Tibshirani, R. (1994). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B*, 58:267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B*, 67:91–108.

Tipping, M. and Bishop, C. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society. Series B*, 61:611–622.

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

Valk, P., Verhaak, R., Beijen, M., Erpelinck, C., van Waalwijk, B., van Doorn-Khosrovani, S., Boer, J., Beverloo, H., Moorhouse, M., van der Spek, P., Löwenberg, B., and Delwel, R. (2004). Prognostically useful gene-expression profiles in acute myeloid leukemia. *New England Journal of Medicine*, 350:1617–1628.

van den Heuvel-Elbrink, M., van der Holt, B., te Boekhorst, P., and et al. (1997). MDR1 expression is an independent prognostic factor for response and survival in de novo acute myeloid leukemia. *British Journal of Hematology*, 99:76–83.

Verhaak, R., Goudswaard, C., van Putten, W., and et al. (2005). Mutations in nucleophosmin NPM1 in acute myeloid leukemia (AML): association with other gene abnormalities and previously established gene expression signatures and their favorable prognostic significance. *Blood*, 106:3747–3754.

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

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74:646–648.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons.

Wouters, B., Löwenberg, B., Erpelinck-Verschueren, C., van Putten, W., Valk, P., and Delwel, R. (2009). Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood*, 113:3088–3091.

Wu, W. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, 90:831–844.

Yang, A.-J. and Song, X.-Y. (2010). Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, 26:215–222.

Yoshida, R. and West, M. (2010). Bayesian learning in sparse graphical factor models via variational mean-field annealing. *Journal of Machine Learning Research*, 11:1771–1798.

Yuan, M., Joseph, R., and Zou, H. (2009). Structured variable selection and estimation. *Annals of Applied Statistics*, 3:1738–1757.

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.

Zacher, B., Abnaof, K., Gade, S., Younesi, E., Tresch, A., and Fröhlich, M. (2012). Joint Bayesian inference of condition-specific miRNA and transcription factor activities from combined gene and miRNA expression data. *Bioinformatics*, 28:1714–1720.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, 38:894–942.

Zhou, X., Liu, K. Y., and Wong, S. T. (2004). Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics*, 37:249–259.

Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

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.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36:1509–1533.