Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

EMVS: The EM Approach to Bayesian Variable Selection

Veronika Ročková [a b] & Edward I. George [a b]

[a] Erasmus University and The University of Pennsylvania

[b] Department of Statistics , University of Pennsylvania , Philadelphia , PA , 19104

Accepted author version posted online: 13 Dec 2013.Published online: 13 Dec 2013.

To cite this article: Veronika Ročková & Edward I. George (2013): EMVS: The EM Approach to Bayesian Variable Selection, Journal of the American Statistical Association, DOI: 10.1080/01621459.2013.869223

To link to this article: http://dx.doi.org/10.1080/01621459.2013.869223

EMVS: The EM Approach to Bayesian Variable Selection

by

Veronika Ročková and Edward I. George [1]

Erasmus University and The University of Pennsylvania

Final Revision, November 2013

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

[1]Veronika Ročková is a postdoctoral researcher at the Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19104, vrockova@wharton.upenn.edu. Edward I. George is Professor of Statistics, Department of Statistics, University of Pennsylvania, Philadelphia, PA, 19104, edgeorge@wharton.upenn.edu.

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.

Keywords: Dynamic posterior exploration; High dimensionality; Regularization plots; Sparsity; SSVS.

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 paper 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.

An outline of the paper is as follows. In Section 2, we establish notation and describe the broad range of hierarchical prior formulations for EMVS. In Section 3, we derive the EM algorithm which underlies the main thrust of our approach. In Section 4, we illustrate the use of the new spike-and-slab regularization diagrams to select submodels from the solution set of models identified by the EM algorithm. In Section 5, we describe a deterministic annealing variant of EMVS, which can be used to mitigate posterior multimodality and enhance EM performance. In Section 6, we show how heavy tailed slab distributions are easily incorporated into our approach. In Section 7, we discuss how the structured priors on the model space can be integrated within the EMVS framework. In Section 8, we illustrate the potential of EMVS on a genetic data set which had previously required lengthly MCMC stochastic search for a Bayesian variable selection analysis. Section 9 concludes with a summary discussion and directions for future research.

2 Conjugate Spike-and-Slab Formulations for EMVS

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

$$f(\boldsymbol{y} \mid \alpha, \boldsymbol{\beta}, \sigma) = \mathrm{N}_n(\mathbf{1}_n\alpha + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n), \tag{2.1}$$

where $\mathbf{1}_n$ is a $n \times 1$ vector of 1's, α is an unknown scalar intercept, $\boldsymbol{\beta}$ 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 $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, $\gamma_i \in \{0, 1\}$, where $\gamma_i = 1$ indicates that x_i is to be included in the model. Combined with suitable prior distributions over $\alpha, \boldsymbol{\beta}$, σ and $\boldsymbol{\gamma}$, the induced posterior distribution $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ 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 $\boldsymbol{\beta}$,

$$\pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}, v_0, v_1) = \mathrm{N}_p(\mathbf{0}, \boldsymbol{D}_{\sigma,\boldsymbol{\gamma}}), \tag{2.2}$$

where $\boldsymbol{D}_{\sigma,\boldsymbol{\gamma}} = \sigma^2 \mathrm{diag}(a_1, \ldots, a_p)$ 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 $\boldsymbol{\beta}$ 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 $f(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma)$. This is equivalent to centering Y at 0 and treating it as a constrained multivariate Gaussian realization with mean $\boldsymbol{X\beta}$.

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

$$\pi(\sigma^2 \mid \gamma) = \text{IG}(\nu/2, \nu\lambda/2) \tag{2.3}$$

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 $\pi(\boldsymbol{\gamma})$ over the 2^p possible values of $\boldsymbol{\gamma}$. For this purpose, we shall be interested in hierarchical specifications of the form

$$\pi(\boldsymbol{\gamma}) = \mathsf{E}_{\pi(\boldsymbol{\theta})}\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta}) \tag{2.4}$$

where $\boldsymbol{\theta}$ 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 $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ is the i.i.d. Bernoulli prior form

$$\pi(\boldsymbol{\gamma} \mid \theta) = \theta^{|\gamma|}(1-\theta)^{p-|\gamma|}, \tag{2.5}$$

where $\theta \in [0, 1]$ and $|\gamma| = \sum_i \gamma_i$. With this form, any marginal $\pi(\boldsymbol{\gamma})$ in (2.4) will be exchangeable on the components of $\boldsymbol{\gamma}$. Of particular interest to us will be the exchangeable priors obtained with a beta prior $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$, $a, b > 0$, (2.5) which yields beta-binomial priors $\pi(\boldsymbol{\gamma})$ that can 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. The choice $a = b = 1$ yields the uniform hyperprior $\theta \sim U(0, 1)$ that will be used in later sections to illustrate EMVS. However, choosing a small and b large will be more effective for targeting sparse models in high-dimensions, as shown by Castillo and van der Vaart (2012) who recommend the choice $a = 1$ and $b = p$ in order to obtain optimal posterior concentration rates in sparse settings.

Beyond (2.5), when structural information about the predictors is available, more flexible priors can be used to transmit this information. In particular, two recent forms of $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ 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 consider these forms further in Section 7 and show how they can be folded into EMVS.

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 $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma|\boldsymbol{y})$ rather than simulating from the entire model posterior $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$, 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 $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma)$ to identify the associated high posterior loci of $\pi(\boldsymbol{\gamma}|\boldsymbol{y})$ when $v_0 = 0$.

Our implementation of the EM algorithm maximizes $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma|\boldsymbol{y})$ indirectly, proceeding iteratively in terms of the "complete-data" log posterior, $\log \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma, \boldsymbol{\gamma} \mid \boldsymbol{y})$, where the latent inclusion indicators $\boldsymbol{\gamma}$ 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 $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma)$. 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 $\pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{y})$.

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

$$Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = \mathsf{E}_{\boldsymbol{\gamma}|\cdot}\left[\log \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma, \boldsymbol{\gamma}|\boldsymbol{y}) \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}, \boldsymbol{y}\right] \tag{3.1}$$

where $\mathsf{E}_{\boldsymbol{\gamma}|\cdot}(\cdot)$ denotes the conditional expectation $\mathsf{E}_{\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)},\boldsymbol{y}}(\cdot)$. At the kth iteration, given $(\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)})$, an E-step is first applied, which computes the expectation of the right side of (3.1) to obtain Q. This is followed by an M-step, which maximizes Q over $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma)$ to yield the values of $(\boldsymbol{\beta}^{(k+1)}, \boldsymbol{\theta}^{(k+1)}, \sigma^{(k+1)})$.

For the conjugate spike-and-slab hierarchical prior formulations described in Section 2, the objective function Q in (3.1) is of the form

$$Q\left(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma \,|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = C + Q_1\left(\boldsymbol{\beta}, \sigma \,|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) + Q_2\left(\boldsymbol{\theta} \,|\, \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right), \tag{3.2}$$

where

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

Note that Q_2 above corresponds to the beta-binomial prior on γ. Different expressions for Q_2 will be described in Section 7 where we consider alternative forms for $\pi(\boldsymbol{\gamma} \mid \theta)$.

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 (3.1), the hierarchical posterior distribution of $\boldsymbol{\gamma}$ given $(\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}, \boldsymbol{y})$ depends on $\boldsymbol{y}$ only through the current estimates $(\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)})$, so that

$$\mathsf{E}_{\boldsymbol{\gamma}|\cdot}(\cdot) = \mathsf{E}_{\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)},\boldsymbol{y}}(\cdot) = \mathsf{E}_{\boldsymbol{\gamma}|\boldsymbol{\beta}^{(k)},\boldsymbol{\theta}^{(k)},\sigma^{(k)}}(\cdot). \tag{3.3}$$

Second, the separability of (3.2) into a pair of distinct functions, Q_1 of $(\boldsymbol{\beta}, \sigma)$ and Q_2 of $\boldsymbol{\theta}$, yields an M-step that is obtained by maximizing each of these functions separately.

We note that the rapidly computable forms for the E-step and the M-step described below resulted from proceeding conditionally on $\boldsymbol{\theta}$ and σ throughout the EM algorithm. Had we initially these margined out over their priors, the resulting algorithm would have been prohibitively expensive to carry out.

3.1 The E-step

The E-step proceeds by computing the conditional expectations $\mathsf{E}_{\boldsymbol{\gamma}|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right]$ and $\mathsf{E}_{\boldsymbol{\gamma}|\cdot}\gamma_i$ for Q_1 and Q_2, respectively. Considering the latter first, it follows from (3.3) that

$$\mathsf{E}_{\boldsymbol{\gamma}|\cdot}\gamma_i = \mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}) = p_i^\star, \tag{3.4}$$

where

$$p_i^\star = \frac{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)\mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\theta}^{(k)})}{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)\mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\theta}^{(k)}) + \pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 0)\mathsf{P}(\gamma_i = 0 \mid \boldsymbol{\theta}^{(k)})}. \tag{3.5}$$

Under (2.5), the conditional independence of the γ_i's leads to $\mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\theta}^{(k)}) = \theta^{(k)}$, greatly facilitating the computation of $p_i^\star$. Note that (3.5) is equivalent to the posterior update of mixing proportions for fitting a two-point Gaussian mixture to $\boldsymbol{\beta}^{(k)}$ 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 $\pi(\gamma_i|\boldsymbol{\beta}^{(k)}, \sigma^{(k)}, \boldsymbol{\theta})$, i.e.

$$\mathsf{E}_{\boldsymbol{\gamma}|\cdot}\left[\frac{1}{v_0(1-\gamma_i)+v_1\gamma_i}\right] = \frac{\mathsf{E}_{\boldsymbol{\gamma}|\cdot}(1-\gamma_i)}{v_0} + \frac{\mathsf{E}_{\boldsymbol{\gamma}|\cdot}\gamma_i}{v_1} = \frac{1-p_i^\star}{v_0} + \frac{p_i^\star}{v_1} \equiv d_i^\star. \tag{3.6}$$

3.2 The M-step

Maximization with respect to $(\boldsymbol{\beta}, \sigma)$ 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 $\boldsymbol{\beta}^{(k+1)}$ value that globally maximizes $Q_1\left(\boldsymbol{\beta}, \sigma|\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right)$, regardless of $\sigma^{(k+1)}$, is obtained quickly by the well-known solution to the ridge regression problem

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p}\{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \|\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}\|^2\}, \tag{3.7}$$

where $\|\cdot\|^2$ is the l_2 norm and $\boldsymbol{D}^{\star 1/2}$ denotes the square root of the $p \times p$ diagonal matrix $\boldsymbol{D}^\star = \text{diag}\{d_i^\star\}_{i=1}^p$ with diagonal entries $d_i^\star > 0$ from (3.6). The solution

$$\boldsymbol{\beta}^{(k+1)} = (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{D}^\star)^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{3.8}$$

is a generalized ridge estimator (GRR) with ridge matrix $\boldsymbol{D}^\star$ 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 (3.8) is that it is well defined even when $\boldsymbol{X}'\boldsymbol{X}$ is not invertible.

In problems where $p >> n$, the calculation cost of (3.8) can be substantially reduced by using the Sherman-Morrison-Woodbury formula to obtain

$$\boldsymbol{\beta}^{(k+1)} = \left[\boldsymbol{D}^{\star -1} - \boldsymbol{D}^{\star -1}\boldsymbol{X}'\left(\mathrm{I}_{n\times n} + \boldsymbol{X}\boldsymbol{D}^{\star -1}\boldsymbol{X}'\right)^{-1}\boldsymbol{X}\boldsymbol{D}^{\star -1}\right]\boldsymbol{X}'\boldsymbol{y}, \tag{3.9}$$

an expression which requires an $n \times n$ matrix inversion rather than a $p \times p$ matrix inversion. Alternatively, as described in George, Rockova and Lessafre (2013), the solution of (3.7) can be obtained even faster with the stochastic dual coordinate ascent algorithm of Shalev-Shwartz and Zhang (2012).

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

$$\sigma^{(k+1)} = \sqrt{\frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k+1)}||^2 + ||\boldsymbol{D}^{\star 1/2}\boldsymbol{\beta}^{(k+1)}||^2 + \nu\lambda}{n + p + \nu}}. \tag{3.10}$$

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

$$\theta^{(k+1)} = \arg\max_{\theta\in\mathbb{R}}\left\{\sum_{i=1}^{p} p_i^{\star}\log\left(\frac{\theta}{1-\theta}\right) + (a-1)\log(\theta) + (p+b-1)\log(1-\theta)\right\}, \tag{3.11}$$

namely

$$\theta^{(k+1)} = \frac{\sum_{i=1}^{p} p_i^{*} + a - 1}{a + b + p - 2}. \tag{3.12}$$

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.

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 $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma})$ 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 4.1. Section 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 $\mathrm{N}_p(\mathbf{0}, \Sigma)$ where $\Sigma = (\rho_{ij})_{i,j=1}^{p}$ with $\rho_{ij} = 0.6^{|i-j|}$. Response values were then generated according to the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\beta} = (1, 2, 3, 0, 0, \ldots, 0)'$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \mathrm{I}_n)$ with $\sigma^2 = 3$.

Beginning with an illustration of the EM algorithm from Section 3, we apply it to the simulated data using the spike-and-slab prior (2.2) with a single value $v_0 = 0.5$, $v_1 = 1\,000$, and the beta-binomial variable inclusion prior with $\theta \sim U(0, 1)$. The starting values for the EM algorithm were set to $\boldsymbol{\beta}^{(0)} = \mathbf{1}_p$ and $\sigma^{(0)} = 1$. After merely 4 iterations, the algorithm obtained the modal coefficient estimates $\widehat{\boldsymbol{\beta}}$ depicted in Figure 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 (2.5) under fixed $\theta = 0.5$ (Figure 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.

4.1 Thresholding the EM Output for Variable Selection

Looking at Figure 1(a), it seems intuitively reasonable that the submodel most closely associated with the EM estimate $\widehat{\boldsymbol{\beta}}$ is the one that includes only the variables corresponding to the three large estimates. This intuition is supported by defining the submodel $\widehat{\boldsymbol{\gamma}}$ associated with $(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma})$ to be

the most probable $\boldsymbol{\gamma}$ given $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma})$, namely

$$\widehat{\boldsymbol{\gamma}} = \arg\max_{\boldsymbol{\gamma}} \mathsf{P}(\boldsymbol{\gamma} \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}). \tag{4.1}$$

Obtaining $\widehat{\boldsymbol{\gamma}}$ is facilitated under full hierarchical prior formulations for which

$$\mathsf{P}(\boldsymbol{\gamma} \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) = \prod_{i=1}^{p} \mathsf{P}(\gamma_i \mid \widehat{\beta}_i, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}), \tag{4.2}$$

where the conditional component inclusion probabilities are given by

$$\mathsf{P}(\gamma_i \mid \widehat{\beta}_i, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) = \frac{\pi(\widehat{\beta}_i \mid \widehat{\sigma}, \gamma_i)\mathsf{P}(\gamma_i \mid \widehat{\boldsymbol{\theta}})}{\pi(\widehat{\beta}_i \mid \widehat{\sigma}, \gamma_i = 1)\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}}) + \pi(\widehat{\beta}_i \mid \widehat{\sigma}, \gamma_i = 0)\mathsf{P}(\gamma_i = 0 \mid \widehat{\boldsymbol{\theta}})}. \tag{4.3}$$

This is precisely the case when $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ in (2.4) is the i.i.d. Bernoulli prior form in (2.5) or the logistic regression form in (7.1), or when approximating the Markov random field priors with independent product forms, as described in Section 7.2. In all such cases, (4.1) is simply obtained by maximizing each component probability, namely

$$\widehat{\gamma}_i = 1 \iff \mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) \geq 0.5. \tag{4.4}$$

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

Selection of $\widehat{\boldsymbol{\gamma}}$ via (4.4) is equivalent to thresholding the $\widehat{\beta}_i$ values because $\mathsf{P}(\gamma_i = 1 \mid \widehat{\beta}_i, \widehat{\boldsymbol{\theta}}, \widehat{\sigma})$ 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 $\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}})$ weighted mixture of the spike-and-slab priors, namely

$$\pm\beta_i^*(v_0, v_1, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) = \pm\widehat{\sigma}\sqrt{2v_0 \log(\omega_i c)\, c^2/(c^2 - 1)}, \tag{4.5}$$

where $c^2 = v_1/v_0$ and $\omega_i = [1 - \mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}})]/\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}})$. Thus, (4.4) is equivalent to

$$\widehat{\gamma}_i = 1 \iff |\widehat{\beta}_i| \geq \beta_i^*(v_0, v_1, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}). \tag{4.6}$$

Applying this thresholding rule to the estimates in Figure 1(a) yields the correct three predictor submodel, in contrast to Figure 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 (4.5) may vary across variables with different inclusion probabilities, it will not vary under the beta-binomial prior, where $\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\theta}}) \equiv \widehat{\theta}$, the overall conditional probability of inclusion. Because the values of $\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma})$ 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 (4.4). Nonetheless, it may be useful to also consider larger threshold values to obtain sparser models.

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 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 $v_1 = 1\,000$ fixed and the same beta-binomial inclusion prior. Figure 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 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, 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 $\mathcal{S}_{v_0} = \{x_i : |\widehat{\beta}_i| > \beta_i^*(v_0, v_1, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}\}$. Letting $\widehat{\boldsymbol{\gamma}}_{v_0}$ denote the submodel identified by $\mathcal{S}_{v_0}$, the full procedure thus effectively generates a solution path $\{\widehat{\boldsymbol{\gamma}}_{v_0} : v_0 \in V\}$ through model space. To select the "best" $\boldsymbol{\gamma}$ from this solution path, a natural criterion is the marginal probability of $\boldsymbol{\gamma}$ under the prior with $v_0 = 0$, a marginal we denote by $\pi_0(\boldsymbol{\gamma}|\boldsymbol{y})$. The appeal of $\pi_0(\boldsymbol{\gamma}|\boldsymbol{y})$ is that it evaluates $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ 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 $\boldsymbol{\gamma}$ 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 $\pi_0(\boldsymbol{\gamma} \mid \boldsymbol{y})$, 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, a rapidly computable closed form

$$g_0(\boldsymbol{\gamma}) = C\, \pi_0(\boldsymbol{\gamma} \mid \boldsymbol{y}) \tag{4.7}$$

is available up to the unknown normalizing constant C. This $g_0(\boldsymbol{\gamma})$ serves our purposes perfectly since it suffices for identifying the $\boldsymbol{\gamma} \in \{\widehat{\boldsymbol{\gamma}}_{v_0} : v_0 \in V\}$ for which $g_0(\boldsymbol{\gamma})$ is largest. To illustrate how this would work on our simulated dats, Figure 2(b) plots $\log g_0(\boldsymbol{\gamma})$ 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.

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 $\boldsymbol{\gamma}$. To put EMVS and the MH sampler on equal footing in terms of initialization, we started the sampler at $\boldsymbol{\gamma}^{(0)} = \mathbf{0}_{1\,000}$, which is the local median probability model obtained by thresholding the EMVS initialization $\boldsymbol{\beta}^{(0)} = \mathbf{1}_{1\,000}$, $\sigma^{(0)} = 1$, $\theta^{(0)} = 0.5$ 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 $g_0(\boldsymbol{\gamma})$ 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 $\boldsymbol{\gamma}^{(0)} = \mathbf{1}_p$, (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 $g_0(\boldsymbol{\gamma})$ 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 should also be noted that, in contrast to the MH sample, the deterministic nature of the EMVS computation would always yield reproducible results.

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 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

$$H_t(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) = \frac{1}{t} \log \sum_{\boldsymbol{\gamma}} \pi(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma, \boldsymbol{\gamma} \mid \boldsymbol{y})^t \quad \text{with} \quad 0 < t \leq 1, \tag{5.1}$$

which embeds the actual log incomplete posterior as a special case when $t = 1$. In (5.1), $1/t$ 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, (5.1) 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 (3.5) with

$$p_{i,t}^\star = \frac{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)^t \mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\theta}^{(k)})^t}{\pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 1)^t \mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\theta}^{(k)})^t + \pi(\beta_i^{(k)} \mid \sigma^{(k)}, \gamma_i = 0)^t \mathsf{P}(\gamma_i = 0 \mid \boldsymbol{\theta}^{(k)})^t}. \tag{5.2}$$

The deterministic annealing version of EMVS, which we shall refer to as DAEMVS, is obtained by making this substitution in (3.5). At high temperatures (t close to zero) the probabilities (5.2) become nearly uniform, as can be seen from the limiting behavior $\lim_{t\to 0} p_{i,t}^\star \equiv 0.5$. Thus tempering induces more equal penalties on all the coefficients through (3.7) regardless of their magnitude.

Finally, under (5.2) as $t \to 0$, the unique posterior mode $\widehat{\boldsymbol{\beta}}$ turns out to be a very promising general initialization value for EMVS. This mode is easily obtained as the M-step ridge regression solution (3.8) with equal penalties $\frac{v_0+v_1}{2v_0v_1}$, namely

$$\widehat{\boldsymbol{\beta}}_{t=0} = \left[\boldsymbol{X}'\boldsymbol{X} + \frac{v_0 + v_1}{2v_0v_1}\boldsymbol{I}_p\right]^{-1} \boldsymbol{X}'\boldsymbol{y}. \tag{5.3}$$

5.1 Simulated Example Revisited

In Section 4 we illustrated the EMVS procedure on a simple simulated example with a single set of starting values $\widehat{\boldsymbol{\beta}}^{(0)} = \mathbf{1}_p$. Here we apply EMVS and its tempered version DAEMVS on the

same data using a randomly generated starting vector $\boldsymbol{\beta}^{(0)} \sim \mathrm{N}_{1000}(\mathbf{0}, \mathrm{I})$ 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 for EMVS and DAEMVS 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 $\boldsymbol{\beta}^{(0)} \sim \mathrm{N}_{1000}(\mathbf{0}, 3\times\mathrm{I})$ and $\boldsymbol{\beta}^{(0)} \sim \mathrm{N}_{1000}(\mathbf{0}, 5 \times \mathrm{I})$. 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 9.

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 $\widehat{\boldsymbol{\beta}}_{t=0}$ in (5.3) 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 4(a). We recommend this starting value as a general choice for consideration in practice.

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 while providing enough support to keep them away from being

shrunk by the spike distribution. This can be achieved under our formulation by simply 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

$$\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2\sigma^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \sum_{i=1}^{p} pen_{v_0,v_1,\gamma_i}(|\beta_i|) \right\}, \tag{6.1}$$

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 $\boldsymbol{X}$ are orthonormal, the problem (6.1) can be solved component-wise (Fan and Li, 2001):

$$\widehat{\beta_i} = \text{argmin}_{\beta_i} \left\{ \frac{1}{2}(\widetilde{\beta_i} - \beta_i)^2 + \sigma^2 pen_{v_0,v_1,\gamma_i}(|\beta_i|) \right\}, \tag{6.2}$$

where $\widetilde{\beta} = \boldsymbol{X}'\boldsymbol{y}$. Taking the first derivative of (6.2) 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),

$$\pi(v_1) = \frac{v_1^b(1+v_1)^{-a-b-2}}{\text{B}(a+1, b+1)} \text{I}_{(0,\infty)}(v_1), \tag{6.3}$$

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 (6.3) 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 (6.3) can be written as

$$\pi(\beta_i|v_0, \sigma, \gamma) = (1-\gamma_i)\text{N}(0, \sigma^2 v_0) + \gamma_i \widetilde{\pi}_{a,b,\sigma}(\beta_i), \tag{6.4}$$

for $a > -1.5$ and $b > -0.5$ (Gradshteyn and Ryzhik, 2000, p. 362). Here $\widetilde{\pi}_{a,b,\sigma}(\beta_i)$ denotes a density function

$$\widetilde{\pi}_{a,b,\sigma}(\beta_i) = \frac{\Gamma\left(a+\frac{3}{2}\right)}{B(a+1,b+1)} \frac{\exp\left(\frac{\beta_i^2}{4\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \left(\frac{\beta_i^2}{2\sigma^2}\right)^{\frac{b}{2}-\frac{1}{4}} W_{-a-\frac{b}{2}-\frac{5}{4},-\frac{b}{2}-\frac{1}{4}}\left(\frac{\beta_i^2}{2\sigma^2}\right), \tag{6.5}$$

where $W_{\eta,\psi}$ is the Whittaker function (Abramowitz and Stegun, 1972, p. 505). When $b = 0$, (6.5) is equivalent to the normal-exponential-gamma (NEG) prior (Griffin and Brown, 2012) obtained by imposing the hierarchical exponential-gamma 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

$$\widetilde{\pi}_{a,0,\sigma}(\beta_i) = \frac{(a+1)2^{a+\frac{3}{2}}}{\sqrt{2\pi\sigma^2}} \Gamma\left(a+\frac{3}{2}\right) \exp\left(\frac{\beta_i^2}{4\sigma^2}\right) D_{-2\left(a+\frac{3}{2}\right)}\left(|\beta_i|/\sigma\right), \tag{6.6}$$

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 $\widetilde{pen}'_{v_0,v_1,\gamma_i}(|\beta_i|)$ for (6.4) 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 6.1 *Let* $\widetilde{\pi}_{a,b,\sigma}(\beta_i)$ *be the distribution given in* (6.5) *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*

$$\widetilde{pen}'_{a,b,\sigma}(|\beta_i|) = \frac{\sqrt{2}\left(a+\frac{3}{2}\right)}{\sigma} \frac{W_{-a-\frac{b}{2}-\frac{7}{4},-\frac{b}{2}-\frac{1}{4}}\left(\frac{\beta_i^2}{2\sigma^2}\right)}{W_{-a-\frac{b}{2}-\frac{5}{4},-\frac{b}{2}+\frac{1}{4}}\left(\frac{\beta_i^2}{2\sigma^2}\right)} \tag{6.7}$$

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

Proof: See the Appendix.

Remark 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* (6.5) *and* (A.3) *(in the appendix) it follows that* $\widetilde{\pi}_{a,b,\sigma}(\beta_i) = O\left[\left(\frac{\beta_i^2}{2\sigma^2}\right)^{-a-\frac{5}{4}}\right]$*. The tail behavior is therefore unaffected by b, a 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 6.1 diminishes faster for smaller a.*

Remark 6.2 *A 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 the Whittaker and parabolic cylinder function and the fact that* $W_{\eta,\psi} = W_{\eta,-\psi}$*. Since the asymptotic behavior in Theorem 6.1 is independent of b, it applies to the NEG prior as a special case.*

Margining out the parameter v_1 complicates the maximization with respect to β as the logarithm of the prior distribution (6.5) does not yield a tractable closed form. Instead, we proceed conditionally, by updating estimates of v_1 together with the remaining parameters. The E-step remains unchanged, but with the value v_1 implicit in the computation of (3.5) 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 $\boldsymbol{\beta}^{(k+1)}, \sigma^{(k+1)}$, we can find $v_1^{(k+1)}$ as the solution to

$$v_1^{(k+1)} = \text{argmax}_{v_1}\left\{-\frac{\|\boldsymbol{P}^{\star 1/2}\boldsymbol{\beta}\|^2}{2\sigma^{(k+1)}}\frac{1}{v_1} + (b - 1/2\sum_{i=1}^{p} p_i^{\star})\log(v_1) - (a+b+2)\log(1+v_1)\right\}, \tag{6.8}$$

where $\boldsymbol{P}^{\star} = \text{diag}\{p_1^{\star}, \ldots, p_p^{\star}\}$, a solution that can be obtained analytically.

Remark 6.3 *Because* $\sigma^2 v_1$ *is the common slab distribution variance for every component of* $\boldsymbol{\beta}$*, the prior on* v_1 *induces prior dependence across these components. However, a prior with independent heavy tailed distributions of the form* (6.5) *can be obtained by introducing distinct slab variances* $\sigma^2 v_{1i}$ *for each component* β_i *of* $\boldsymbol{\beta}$*, and then folding a version of* (6.8) *into a suitably modified EM algorithm at each step. Spike-and-slab priors with such independent heavy-tailed*

slab distributions have been shown to be essential to obtaining posteriors that concentrate around the true nonzero coefficients at optimal rates (Castillo and van der Vaart, 2012).

7 Structured Prior Information Forms for $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$

The beta-binomial prior based on the Bernoulli form (2.5) for $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ in (2.4) 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. These forms are incorporated naturally into the EMVS approach. As will be seen, choices of $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ that are log-linear in $\boldsymbol{\gamma}$ facilitate E-step calculations via closed forms, just as occurred for the beta-binomial case in Section 3.1.

7.1 The Independent Logistic Regression Prior

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

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta}) = \prod_{i=1}^{p} \left(\frac{\exp(\boldsymbol{Z}_i' \boldsymbol{\theta})}{1 + \exp(\boldsymbol{Z}_i' \boldsymbol{\theta})} \right)^{\gamma_i} \left(\frac{1}{1 + \exp(\boldsymbol{Z}_i' \boldsymbol{\theta})} \right)^{1-\gamma_i}, \tag{7.1}$$

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 (7.1), $\boldsymbol{Z}_i$ is a $q \times 1$ vector of covariates which may influence the model inclusion probability of x_i, and $\boldsymbol{\theta}$ is a $q \times 1$ vector of regression coefficients. Letting $\boldsymbol{Z} = [\boldsymbol{Z}^1, \ldots, \boldsymbol{Z}^q]$ be the $p \times q$ matrix whose i^{th} row is equal to $\boldsymbol{Z}_i'$, θ_j is the weight assigned to $\boldsymbol{Z}^j$, the jth column of Z. As will be illustrated in Section 7.3 below, the columns of $\boldsymbol{Z}$ can conveniently be used to represent potential variable inclusion groupings by using dummy variables to represent potential inclusion. With the addition of a prior $\pi(\boldsymbol{\theta})$, posterior estimates of $\boldsymbol{\theta}$ can yield additional information about the relative influence of the $\boldsymbol{Z}_j^{th}$ grouping.

The choice of a prior for $\boldsymbol{\theta}$ is motivated by observing that when the $\boldsymbol{Z}_i$ identify nonoverlapping groupings, (7.1) 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. $\boldsymbol{Z}_i \equiv 1$, the prior (7.1) simplifies to the exchangeable Bernoulli form (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

$$\pi(\theta) = \frac{1}{B(a,b)}\left[\frac{\exp(\theta)}{1+\exp(\theta)}\right]^a\left[\frac{1}{1+\exp(\theta)}\right]^b, \tag{7.2}$$

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

$$\pi(\boldsymbol{\theta}) = \frac{1}{B(a,b)}\left[\frac{\exp(\mathbf{1}'\boldsymbol{\theta})}{1+\exp(\mathbf{1}'\boldsymbol{\theta})}\right]^a\left[\frac{1}{1+\exp(\mathbf{1}'\boldsymbol{\theta})}\right]^b. \tag{7.3}$$

The EMVS algorithm under the form (7.1) with the prior (7.3), is then obtained by replacing Q_2 in (3.2) with

$$Q_2^{LR}\left(\boldsymbol{\theta}|\boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}\right) = \sum_{i=1}^{p}\left\{\boldsymbol{Z}_i'\boldsymbol{\theta}\,\mathsf{E}_{\boldsymbol{\gamma}|\cdot}\gamma_i - \log[1+\exp(\boldsymbol{Z}_i'\boldsymbol{\theta})]\right\} + \{a\mathbf{1}'\boldsymbol{\theta} - (a+b)\log[1+\exp(\mathbf{1}'\boldsymbol{\theta})]\} \tag{7.4}$$

Using the fact that

$$E_{\boldsymbol{\gamma}|\cdot}\gamma_i = \mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\theta}^{(k)}) = \frac{\exp(\boldsymbol{Z}_i'\boldsymbol{\theta}^{(k)})}{1+\exp(\boldsymbol{Z}_i'\boldsymbol{\theta}^{(k)})}, \tag{7.5}$$

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

7.2 The Markov Random Field Prior

The second structured prior form we consider for $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ 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 $\mathcal{E} = \{(i,j) : 1 \le i \ne j \le p\}$, they proposed the MRF prior

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta}) = \exp\left[\boldsymbol{\theta}_1'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\boldsymbol{\theta}_2\boldsymbol{\gamma} - \psi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\right], \tag{7.6}$$

where $\boldsymbol{\theta}_1 = (\theta_1, \ldots, \theta_p)'$ is a vector of sparsity parameters, $\boldsymbol{\theta}_2 = \left(\theta_{ij}\right)_{i,j=1}^{p}$ is a symmetric matrix of real numbers with $\theta_{ij} = 0 \Leftrightarrow (i,j) \ne \mathcal{E}$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. The matrix $\boldsymbol{\theta}_2$ regulates the smoothness of

the distribution (7.6) by controlling the inclusion probability of a variable based on the selection status of its neighbors. If all the genes are disconnected, so that $\boldsymbol{\theta}_2 = \mathbf{0}$, then the prior (7.6) reduces to an independent product of Bernoulli distributions with parameters $p_i = \exp(\theta_i)/[1 + \exp(\theta_i)]$. The normalizing constant $\psi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, known as the partition function, is typically intractable due to the many combinatorial possibilities when summing over all 2^p model configurations.

Letting $\boldsymbol{\gamma}_{\setminus i} = \{\gamma_j : j \neq i\}$ denote the subvector containing all but the i^{th} inclusion indicator, the distribution (7.6) implies a simple form for the conditional distributions

$$\pi(\gamma_i \mid \boldsymbol{\gamma}_{\setminus i}) = \frac{\exp(\theta_i + \sum_{j \neq i} \theta_{ij}\gamma_j)}{1 + \exp(\theta_i + \sum_{j \neq i} \theta_{ij}\gamma_j)}, \tag{7.7}$$

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 (7.6), the key calculation for the E-step is the evaluation of $\mathsf{E}_{\boldsymbol{\gamma}\mid\cdot}\gamma_i = \mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}) = p_i^*$ in (3.5). Because this evaluation is complicated by the dependence among the components in $\boldsymbol{\gamma}$ under the MRF prior, we approximate it as follows. To begin with, note that the $\mathsf{P}(\gamma_i = 1 \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)})$ values here arise as marginal means under the joint conditional distribution

$$\pi(\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}) \propto \exp\left[\left(\frac{1}{2}\log(v_0/v_1)\,\mathbf{1} - \frac{v_0 - v_1}{2\sigma^{(k)2}v_1v_0}\boldsymbol{\beta}^{(k)}\mathrm{diag}\{\beta_i^{(k)}\}_{i=1}^p + \boldsymbol{\theta}_1^{(k)}\right)'\boldsymbol{\gamma} + \boldsymbol{\gamma}'\boldsymbol{\theta}_2^{(k)}\boldsymbol{\gamma}\right]. \tag{7.8}$$

The first two terms in the exponent follow directly from the prior distribution $\pi(\boldsymbol{\beta}|\sigma, \boldsymbol{\gamma}) = \mathrm{N}_p(\mathbf{0}, \boldsymbol{D}_{\sigma,\boldsymbol{\gamma}})$, by rewriting the determinant of the matrix $\boldsymbol{D}_{\sigma,\boldsymbol{\gamma}}^{-1/2} = \frac{\mathrm{diag}\{\gamma_i/\sqrt{v_1} + (1-\gamma_i)/\sqrt{v_0}\}_{i=1}^p}{\sigma}$ as

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

The conditional distribution in (7.8) can be regarded as an MRF distribution with adjusted parameters $\boldsymbol{\theta}_1^\star = (\theta_1^\star, \ldots, \theta_p^\star)'$ and $\boldsymbol{\theta}_2^\star = \boldsymbol{\theta}_2$, where $\theta_i^\star = \frac{1}{2}\log(v_0/v_1) - \frac{v_0 - v_1}{2\sigma^{(k)2}v_1v_0}\beta_i^{(k)2} + \theta_i$.

Because the partition function $\psi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ is a normalizing factor in the exponential family, it follows that $\mathsf{E}(\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}) = \frac{\partial\psi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\partial\boldsymbol{\theta}_1}|_{\boldsymbol{\theta}_1 = \boldsymbol{\theta}_1^\star}$. 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 $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)})$, with a simpler distribution for which it is feasible to do exact inference. Here we make use of the naive mean field method, which uses tractable approximating distributions for completely disconnected graphs. Thus, we employ approximating distributions of the form $q(\boldsymbol{\gamma} \mid \boldsymbol{\mu}) = \prod_i \mu_i^{\gamma_i}(1-\mu_i)^{1-\gamma_i}$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)' \in [0,1]^p$ denotes the vector of mean parameters.

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

$$\widehat{\mu}_i = \frac{\exp(\theta_i^\star + \sum_{j \neq i} \theta_{ij} \widehat{\mu}_j)}{1 + \exp(\theta_i^\star + \sum_{j \neq i} \theta_{ij} \widehat{\mu}_j)}, \quad 1 \leq i \leq p. \tag{7.9}$$

Each of the equations (7.9) can be regarded as an averaged version of the expression in (7.7). The solution can be found by iteratively updating (7.9), 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 (3.5).

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 $\boldsymbol{\theta}_1$ to be unknown (arising from a prior distribution $\pi(\boldsymbol{\theta}_1)$). In what follows, we restrict attention to vectors of type $\boldsymbol{\theta}_1 = \theta(1, \ldots, 1)'$. A natural candidate prior distribution $\pi(\theta)$, which corresponds to the beta-binomial prior in case $\boldsymbol{\theta}_2 = \mathbf{0}$, is the logistic-beta distribution (7.2). The M-step of the algorithm then requires the additional step of updating the parameter θ by finding the maximum of the function

$$Q_2^{MRF}(\theta \mid \boldsymbol{\beta}^{(k)}, \boldsymbol{\theta}^{(k)}, \sigma^{(k)}) = \theta \left(\sum_{i=1}^{p} p_i^\star + a \right) + \psi(\theta, \boldsymbol{\theta}_2) - (a+b) \log[1 + \exp(\theta)].$$

Maximizing Q_2^{MRF} w.r.t. θ is complicated by the unavailability of the partition function in a closed form. 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

$$\psi(\theta, \boldsymbol{\theta}_2) \approx \theta \sum_{i=1}^{p} \mu_i + \boldsymbol{\mu}' \boldsymbol{\theta}_2 \boldsymbol{\mu} - \psi^\star(\boldsymbol{\mu}), \tag{7.10}$$

where $\mu_i = \mathsf{E}_{\theta,\boldsymbol{\theta}_2}\gamma_i$ and $\psi^\star$ denotes the conjugate dual function to ψ, which has an explicit form for the approximating product distributions, i.e.

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

The mean values $\mu_i = \mathsf{E}_{\theta,\boldsymbol{\theta}_2}\gamma_i$ for each specific value θ can be obtained from (7.9).

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 $f(\theta, \boldsymbol{\theta}_2) = \sum_{i=1}^{p} \mathsf{E}_{\theta,\boldsymbol{\theta}_2}\gamma_i$ exhibits rapid growth or even a jump. There may be multiple transition points in situations when the matrix $\boldsymbol{\theta}_2$ has complicated structural patterns.

In order to visually assess the quality of the approximation to the partition function, Figure 5(a) plots the approximated and true function $\psi(\theta, \boldsymbol{\theta}_2)$ for varying θ with $\boldsymbol{\theta}_2$ 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 (7.10). The values of $p_i^\star$ are set to one and $a = b = 1$.

We observe that the approximation (7.10) with imputed approximated mean values loses the convexity property (Figure 5(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 $f(\theta, \boldsymbol{\theta}_2)$ in Figure 5(b). The plots of the true and approximated $Q_2^{MRF}(\cdot)$ function in Figure 5(c) 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. Under the MRF prior, the E-step is performed with parameters $\boldsymbol{\theta}_1^\star$ and $\boldsymbol{\theta}_2^\star$ multiplied by an inverse temperature parameter.

7.3 Simulated Example for Structured Priors

To illustrate the potential of the structured variable inclusion priors from Sections 7.1 and 7.2, we compare them with the benchmark beta-binomial prior on simulated data with substantial grouped structure. For this purpose, we simulated $Y \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathrm{I}_n)$ with $n = 100$ and $\sigma^2 = 5$. The $n \times p$ predictor matrix $\boldsymbol{X}$ 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 $\pi(\boldsymbol{\gamma} \mid \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ to reflect varies degrees of prior knowledge: (a) The Bernoulli form (2.5) coupled with the uniform prior on θ which yields the beta-binomial prior with $a = b = 1$ on $\boldsymbol{\gamma}$. This unstructured exchangeable choice ignores the potential grouping information; (b) The independent logistic regressions form (7.1) with the three grouping vector choices $\boldsymbol{Z}^1, \boldsymbol{Z}^2, \boldsymbol{Z}^3$ where the ith component of $\boldsymbol{Z}^j$ is given by $z_{ij} = \mathbb{I}_{[33\,(j-1)+1;\,33\,j]}(i)$ for $1 \le i \le p$. To this we add the logistic-beta prior (7.3) with $a = b = 1$ on $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$. 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 (7.6) with sparsity parameter $\boldsymbol{\theta}_1 = \theta \mathbf{1}$, where θ is assigned prior (7.2) with $a = b = 1$, and with fixed $\boldsymbol{\theta}_2 = (\mathbf{1}_{33\times 33} - \mathrm{I}_{33}) \otimes \mathrm{I}_3$, where $\mathbf{1}_{33\times 33}$ 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 $\boldsymbol{Z}^1 = \mathbf{1}_p, \boldsymbol{Z}^2 = \mathbf{0}, \boldsymbol{Z}^3 = \mathbf{0}$ and $\boldsymbol{\theta}_2 = \mathbf{0}$.

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 (6.3) 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 $\boldsymbol{\beta}^{(0)}$ were selected according to (5.3) with $v_0 = 1$ and $v_1 = 1\,000$, $\sigma^{(0)} = 1$, $\boldsymbol{\theta}^{(0)} = \mathbf{1}_3$ for the logistic prior and $\theta^{(0)} = \theta_{trans}$ for the MRF prior. To evaluate the solution path of models $\{\widehat{\boldsymbol{\gamma}}_{v_0} : v_0 \in V\}$ generated under each prior, we

used the same g_0 function from (4.7) 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 6. 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 7. 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 positives. The posterior estimates $\widehat{\boldsymbol{\theta}} = (0.93, -3.35, -3.44)'$, 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 8. 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 $f(\theta, \boldsymbol{\theta}_2)$ function for varying values of θ (Figure 9(a)). 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 9(b)) and $p_i^\star = 1$, $(i = 1, \ldots, p)$ (Figure 9(c)) 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.

8 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 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 $\boldsymbol{Y}$ 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 $\boldsymbol{X}$ 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 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 $\boldsymbol{\theta}_2$ 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 (6.3) 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 $g_0(\boldsymbol{\gamma})$ 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 (5.3), 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 10. As v_0 increases, $g_0(\boldsymbol{\gamma})$ 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 can be found corresponding to more moderate v_0 values. For $v_0 = 9.001$, 18 motifs were screened out (Table 2), 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 `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 $\boldsymbol{\theta}_2$ equal to 0.83. We assume $\boldsymbol{\theta}_1 = \theta(1, \ldots, 1)'$ and specify an appropriate distribution $\pi(\theta)$ according to (7.2), which locates the majority of its mass in the phase transition region. The plot of the transition function in Figure 11(a) indicates multiple transition points, a consequence of the structure $\boldsymbol{\theta}_2$ which allows for overlapping components. Selecting the hyperparameter values $a_\theta = 5$ and $b_\theta = 10\,000$ guarantees accumulation of the prior distribution within boundaries $[-9, -6]$ (Figure 11(b)). The corresponding $Q_2(\cdot)$ function for $p_i^\star \equiv 1$ is plotted in Figure 11(c). 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 $g_0(\boldsymbol{\gamma})$ is displayed in Figure 12. Among the explored models, the highest value of $g_0(\boldsymbol{\gamma})$ was obtained for a model with 4 motifs (`ACGCGTT, CGCGTTT, GACGCGT, TTTCGCG`). Table 2 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 $\max_{1 \le i \le p}\{|\boldsymbol{\beta}_i^{(k+1)} - \boldsymbol{\beta}_i^{(k)}|\} < 10^{-4}$ 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 $v_1 = 1\,000$). 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].

9 Discussion and Future Directions

The main thrust of this paper 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 (Tong et al) 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 al-

gorithm 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 (4.7) 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 and C++. Both are available from the authors upon request.

Acknowledgements

The authors would like to thank Nancy Zhang for fruitful discussions and for kindly providing the dataset. This work resulted from a collaboration initiated thanks to Emmanuel Lesaffre.

References

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

Bar, H., Booth, J., and Wells, M. (2010), "An Empirical Bayes Approach to Variable Selection and

QTL Analysis," *In the Proceedings of the 25th International Workshop on Statistical Modelling, Glasgow, Scotland*, pages 63–68.

Berger, J., Pericchi, L., and Varshavsky, J. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations," *Sankhy$\bar{a}$ Ser. A*, 60, 307–321.

Bottolo, L. and Richardson, S. (2010), "Evolutionary Stochastic Search for Bayesian Model Exploration," *Bayesian Analysis*, 5, 583–618.

Bussemaker, H., Li, H., and Siggia, E. (2001), "Regulatory Elements Detection Using Correlation With Expression," *Nature Genetics*, 27, 167–171.

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

Castillo, I. and van der Vaart, A. (2012), "Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences," *Annals of Statistics*, 40, 2069–2101.

Cui, W. and George, E. I. (2008), "Empirical Bayes vs. Fully Bayes Variable Selection," *Journal of Statistical Planning and Inference*, 138, 888–900.

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.

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

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.

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

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. (2012), "Bayesian Hyper-LASSOS with Non-convex Penalization," *Australian & New Zealand Journal of Statistics*, 53, 423–442.

Hans, C., Dobra, A., and West, M. (2007), "Shotgun Stochastic Search for "large p" regression," *Journal of the American Statistical Association*, 102, 507–516.

Hayashi, T. and Iwata, H. (2010), "EM Algorithm for Bayesian Estimation of Genomic Breeding Values," *BMC Genetics*, 11, 1–9.

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

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.

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.

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*, pages 410–423.

Maruyama, Y. and George, E. I. (2011), "Fully Bayes Factors with a Generalized g-prior," *Annals of Statistics*, 39, 2740–2765.

McLachlan, G. J. and Basford, K. (2004), *Mixture Models: Inference and Application to Clustering*, New York and Basel: Marcel Dekker.

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.

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.

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.

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.

Ueda, N. and Nakano, R. (1998), " Deterministic Annealing EM Algorithm," *Neural Networks*, 11, 271–282.

Wainwright, M. and Jordan, M. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends in Machine Learning*, 1, 1–305.

Appendix

Proof of Theorem 6.1

The proof of the expression (6.7) is facilitated by noting $\widetilde{pen}'_{a,b,\sigma}(|\beta_i|) = \frac{\partial \widetilde{\pi}_{a,b,\sigma}(\beta_i)/\partial|\beta_i|}{\widetilde{\pi}_{a,b,\sigma}(\beta_i)}$. The denominator can be for $b > -\frac{1}{2}$ and $a > -\frac{3}{2}$ rewritten using the expression for marginal prior distribution

in (6.5). The numerator can be expressed as

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

The identity (A.1) follows from the Leibnitz integral rule, which is justified since the integrand in (A.1) 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

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

The identity (A.2) together with the expression for the marginal distribution in (6.5) then completes the proof of the equation (6.7).

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

$$W_{\eta,\psi}(z) \sim \exp\left(-\frac{z}{2}\right)z^\eta\left(1+\sum_{k=1}^{\infty}\frac{[\psi^2-(\eta-\frac{1}{2})^2]\ldots[\psi^2-(\eta-k+\frac{1}{2})^2]}{k!z^k}\right), \tag{A.3}$$

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

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

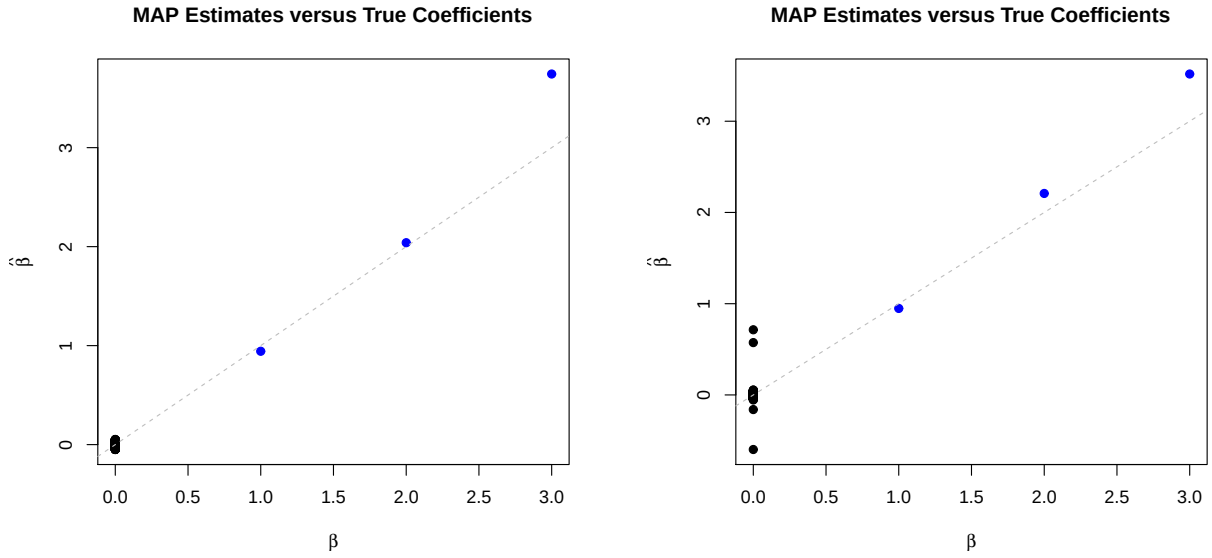
which was to be demonstrated.

	$\boldsymbol{\beta^{(0)} = \mathbf{1}_{1000}}$			$\boldsymbol{\beta}^{(0)} \sim \mathrm{N}_{1000}(\mathbf{0}, \mathrm{I})$			$\boldsymbol{\beta}^{(0)} \sim \mathrm{N}_{1000}(\mathbf{0}, 3 \times \mathrm{I})$			$\boldsymbol{\beta}^{(0)} \sim \mathrm{N}_{1000}(\mathbf{0}, 5 \times \mathrm{I})$		
	#Iter	#Var	$\log g_0(\boldsymbol{\gamma})$	#Iter	#Var	$\log g_0(\boldsymbol{\gamma})$	#Iter	#Var	$\log g_0(\boldsymbol{\gamma})$	#Iter	#Var	$\log g_0(\boldsymbol{\gamma})$
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 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 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.

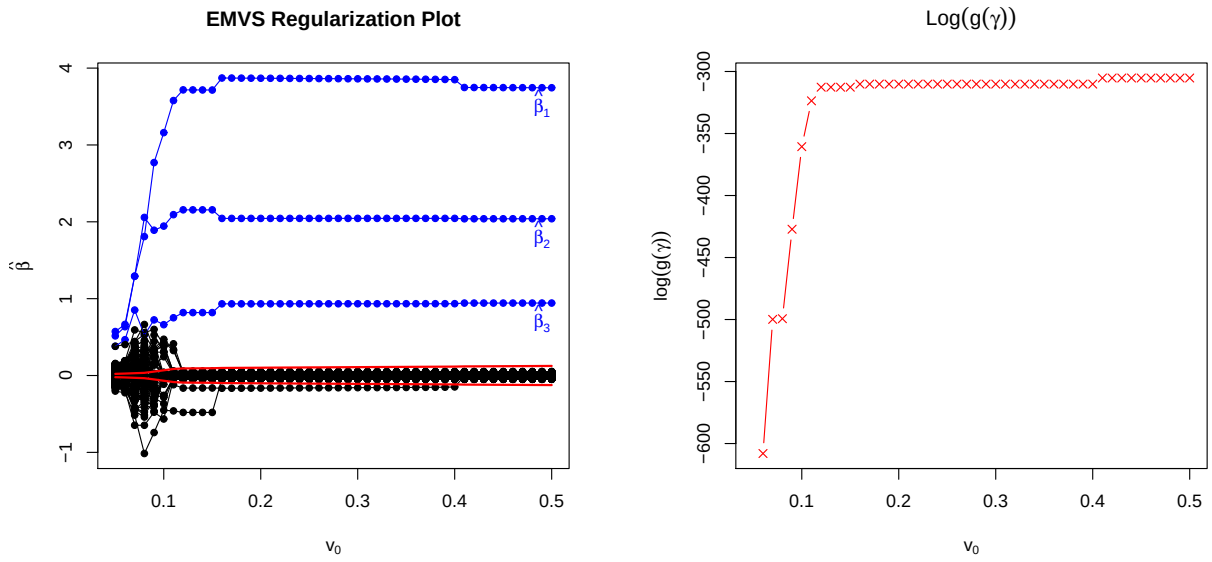
18 **Selected Motifs**		7 **Selected Motifs**		
BB	**MRF**	**BB**	**MRF**	**Known**
GACGCGT[1]	*GACGCGT*[1]	*GACGCGT*[1]	*GACGCGT*[1]	×
TACGCGT[1]	*TACGCGT*[1]		*TACGCGT*[1]	×
TTCGCGT[1]	*TTCGCGT*[1]	*TTCGCGT*[1]	*TTCGCGT*[1]	×
	TTACGCG[2]			
TTTCGCG[2]	*TTTCGCG*[2]	*TTTCGCG*[2]	*TTTCGCG*[2]	×
	TGACGCG[2]			
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 2: 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)



(a) $v_0 = 0.5, v_1 = 1\,000, \theta \sim U(0,1)$ (b) $v_0 = 0.5, v_1 = 1\,000, \theta = 0.5$

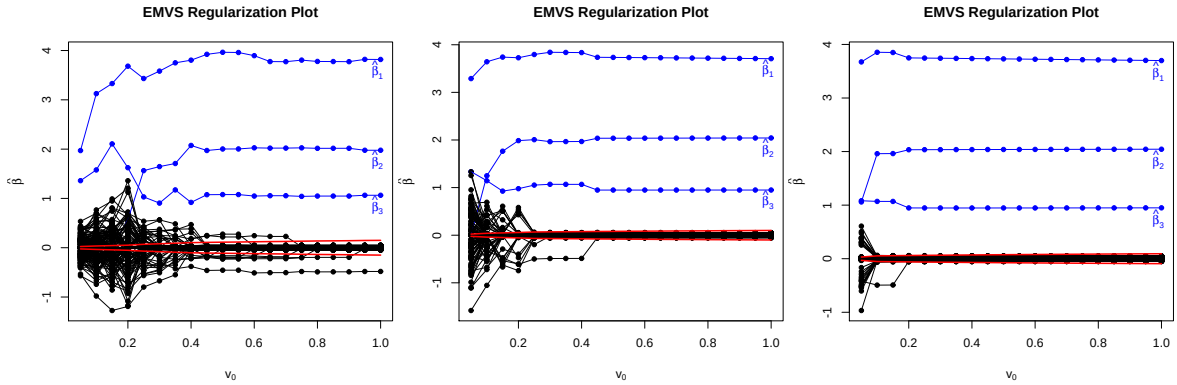
Figure 1: Modal estimates of the regression coefficients; (a) beta binomial prior, (b) Bernoulli prior with fixed $\theta = 0.5$.



(a) EMVS Regularization Plot

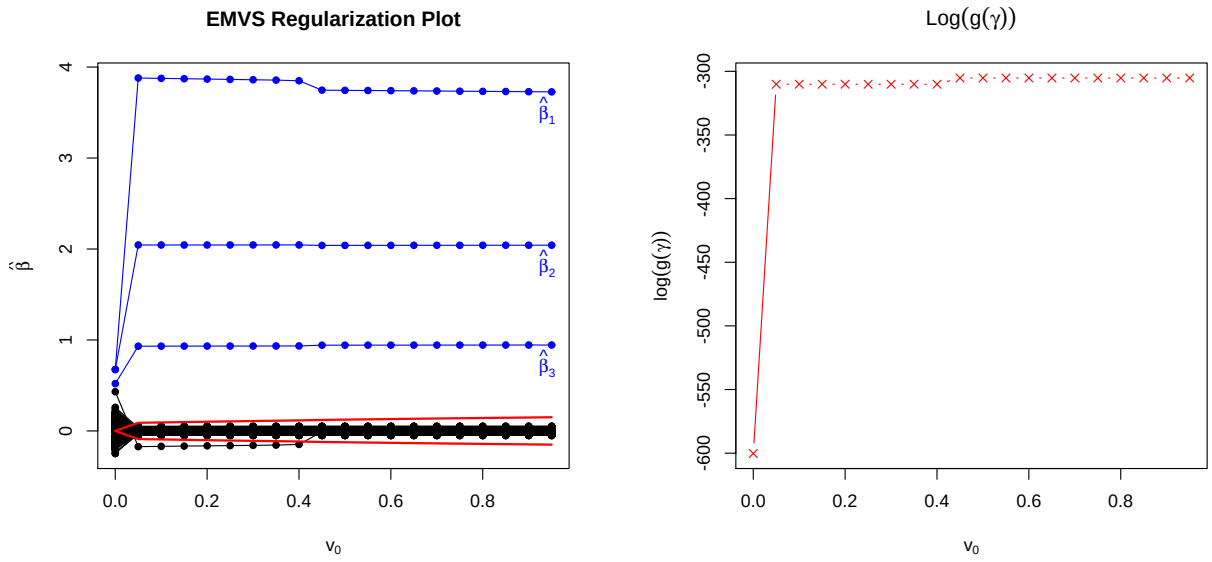
(b) Submodel Evaluations

Figure 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.



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

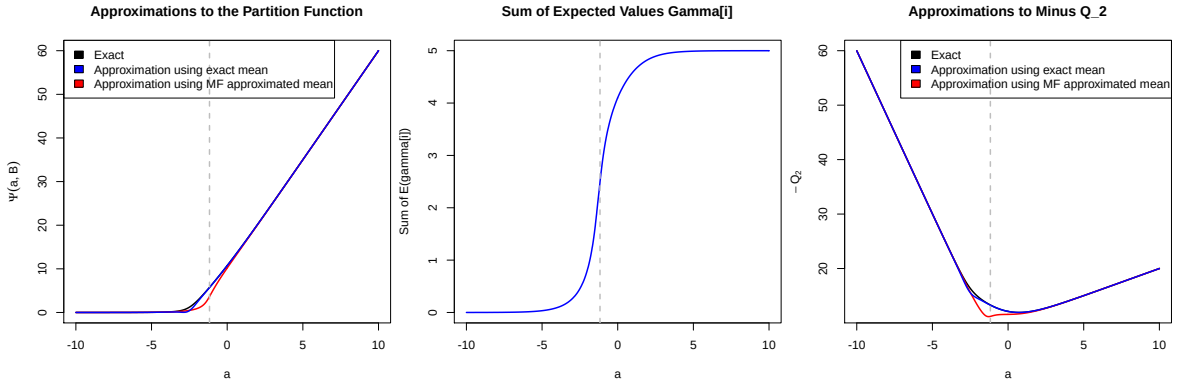
Figure 3: Regularization plots for the simulated example from Section 4 using EMVS and the deterministic annealing EMVS (DAEMVS) considering randomly generated starting vector $\boldsymbol{\beta}^{(0)} \sim \mathrm{N}_{1000}(\mathbf{0}, \mathrm{I})$



(a) EMVS Regularization Plot

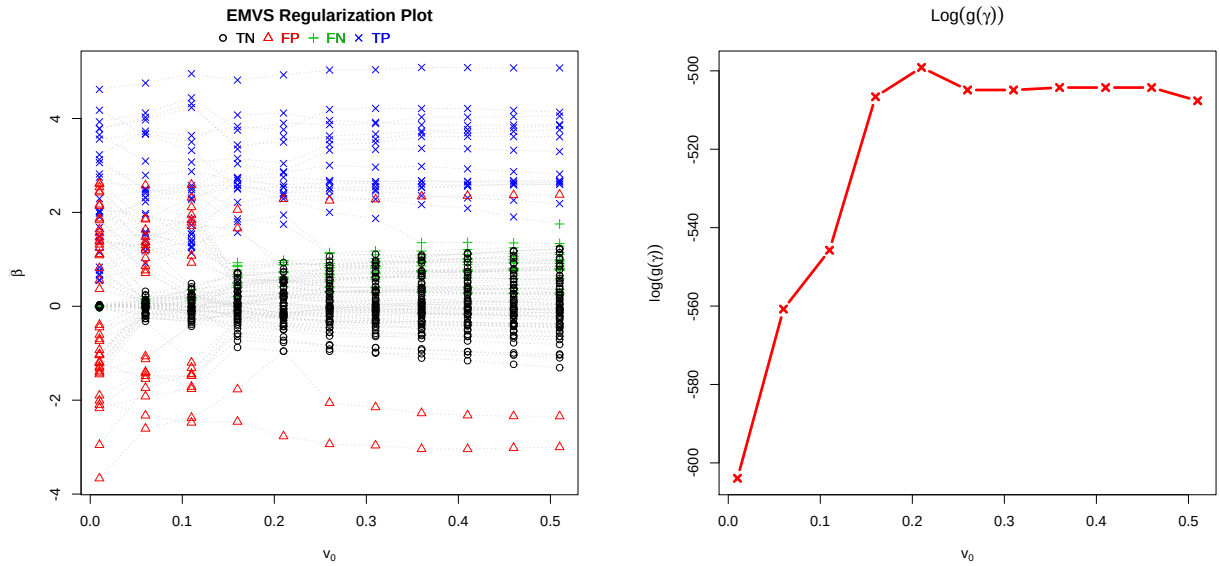
(b) Submodel Evaluations

Figure 4: (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.



(a) $\psi(\theta, \boldsymbol{\theta}_2) \sim \theta$ (b) $f(\theta, \boldsymbol{\theta}_2) = \sum_{i=1}^{p} \mathsf{E}_{\theta,\boldsymbol{\theta}_2} \gamma_i \sim \theta$ (c) $Q_2^{MRF}(\theta, \boldsymbol{\theta}_2) \sim \theta$

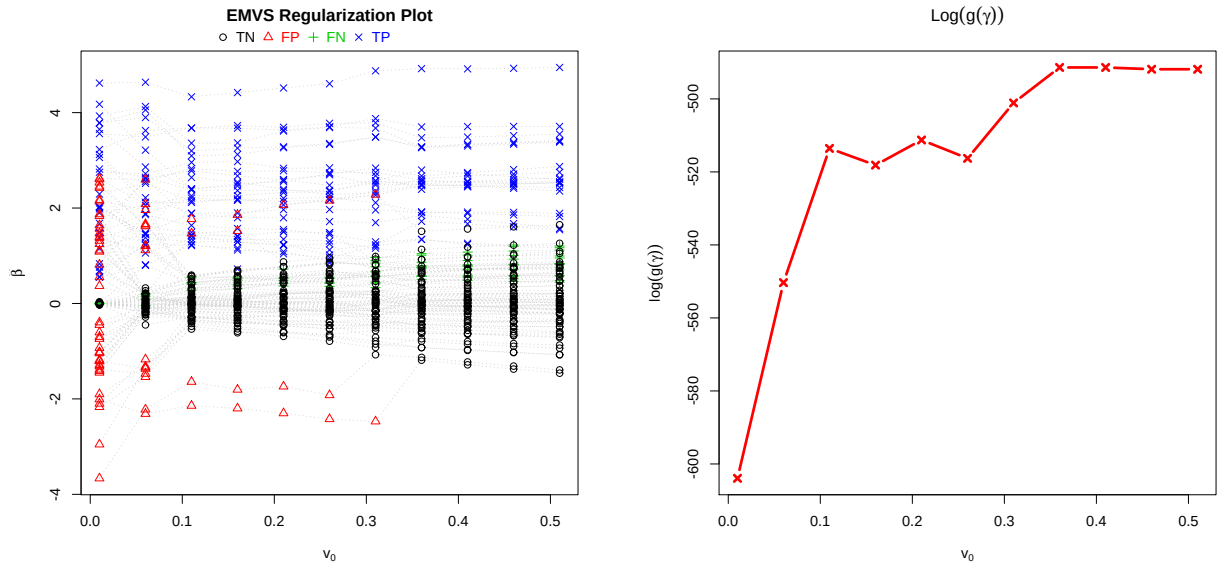
Figure 5: Plots of the approximated and true partition functions $\psi(\theta, \boldsymbol{\theta}_2)$, the phase transition function $f(\theta, \boldsymbol{\theta}_2)$ and $Q_2^{MRF}(\theta, \boldsymbol{\theta}_2)$ in relation to parameter θ.



(a) EMVS Regularization Plot

(b) Submodel Evaluations

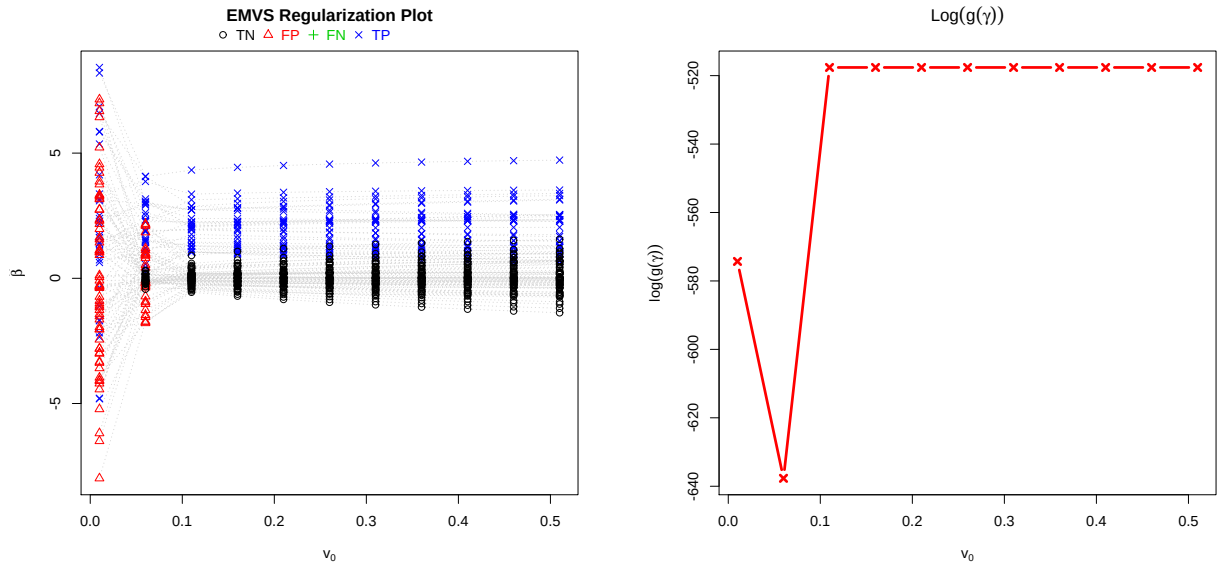
Figure 6: (a) plot of estimated regression coefficients for varying choices of v_0, TN/FP/FN/TP stand for true negatives/false positives/false negatives/true positives; (b) $\log g_0(\gamma)$ for models selected with $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) > 0.5$.



(a) EMVS Regularization Plot

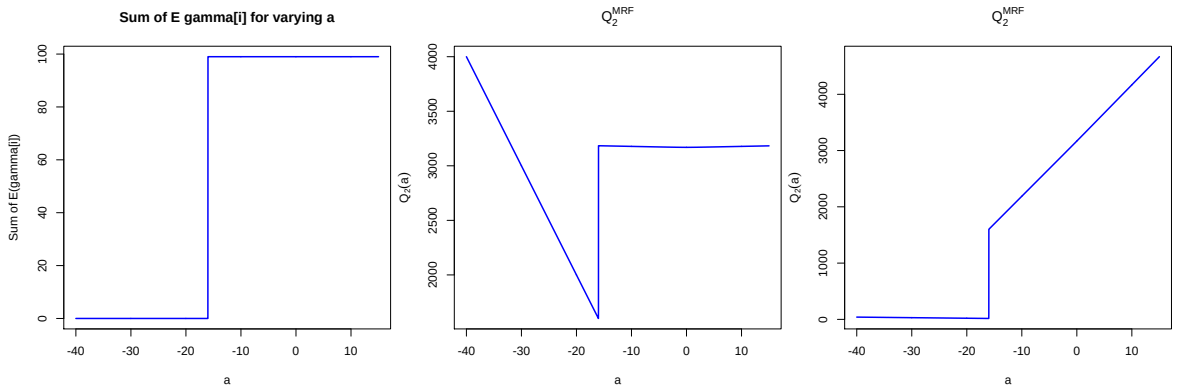
(b) Submodel Evaluations

Figure 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 $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) > 0.5$.



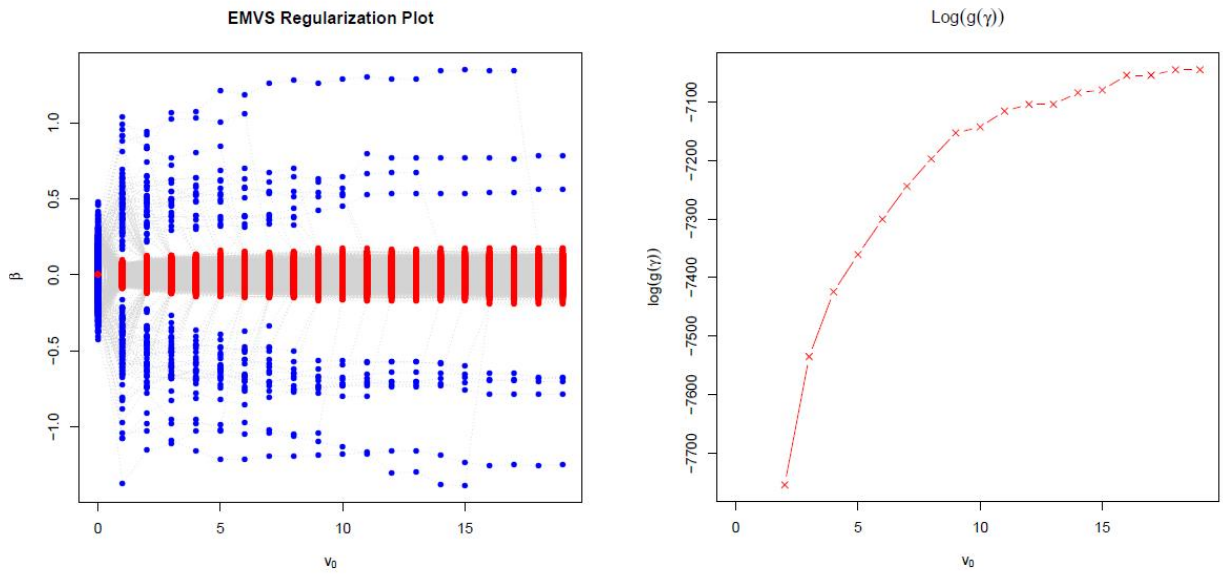
(a) EMVS Regularization Plot (b) Submodel Evaluations

Figure 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 $\mathsf{P}(\gamma_i = 1 \,|\, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) > 0.5$.



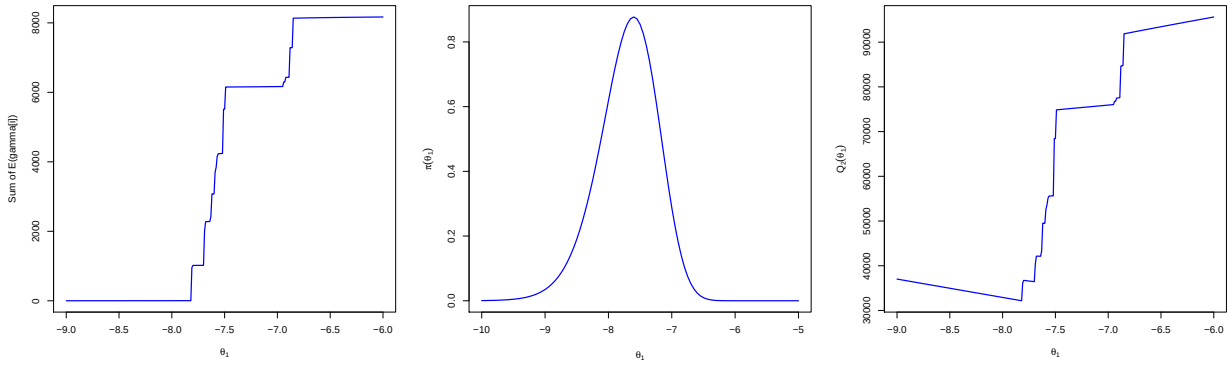
(a) $f(\theta, \boldsymbol{\theta}_2) = \sum_{i=1}^{p} \mathsf{E}_{\theta,\boldsymbol{\theta}_2} \gamma_i \sim \theta$ (b) $Q_2^{MRF}(\theta, \boldsymbol{\theta}_2) \sim \theta$ (c) $Q_2^{MRF}(\theta, \boldsymbol{\theta}_2) \sim \theta$

Figure 9: Approximated Q_2 function together with the phase transition function for the simulated data example



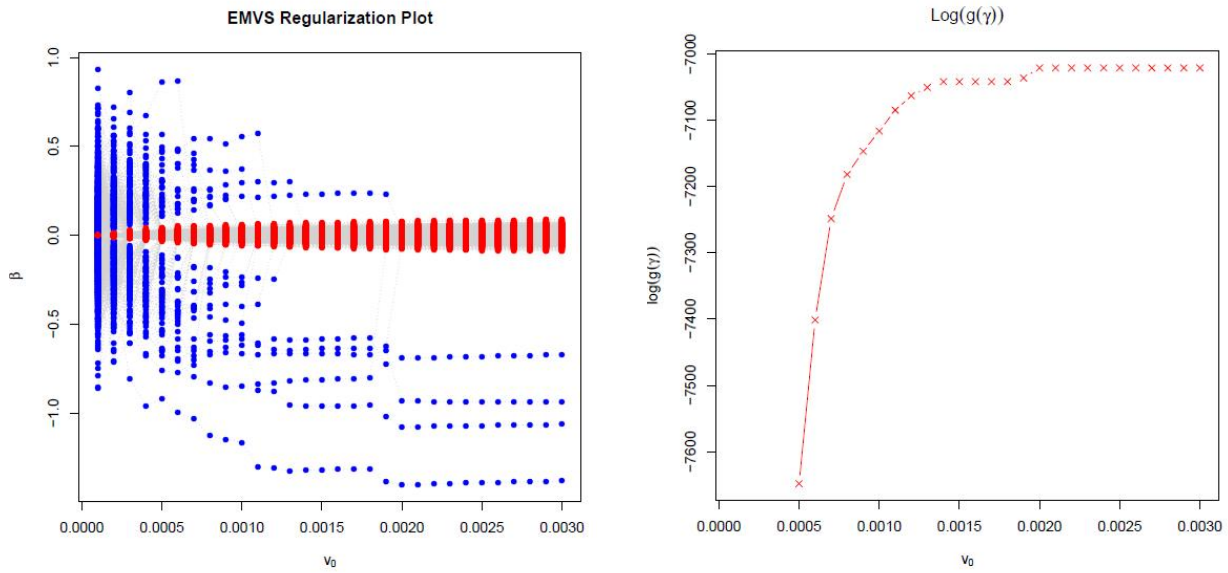
(a) EMVS Regularization Plot (b) Submodel Evaluations

Figure 10: (a) plot of estimated regression coefficients for varying choices of v_0, estimates for variables with conditional posterior inclusion probability $\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma})$ above (below) 0.5 depicted in blue (red); (b) $\log g_0(\gamma)$ for models selected with $\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) > 0.5$.



(a) $f(\theta, \boldsymbol{\theta}_2) = \sum_{i=1}^{p} \mathsf{E}_{\theta,\boldsymbol{\theta}_2} \gamma_i$ (b) $\pi(\theta)$ (c) $Q_2^{MRF}(\theta)$

Figure 11: Approximated $Q_2(\cdot)$ function together with the phase transition function and prior distribution $\pi(\theta)$ for the Spellman data



(a) EMVS Regularization Plot (b) Submodel Evaluations

Figure 12: (a) plot of estimated regression coefficients for varying choices of v_0, estimates for variables with conditional posterior inclusion probability $\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma})$ above (below) 0.5 depicted in blue (red); (b) logarithm of $g(\gamma)$ for models with selected variables with $\mathsf{P}(\gamma_i = 1 \mid \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}}, \widehat{\sigma}) > 0.5$.