

© 2016 by Jin Wang. All rights reserved.

SCALABLE ALGORITHMS FOR BAYESIAN VARIABLE SELECTION

BY

JIN WANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Associate Professor Feng Liang, Chair
Professor John I. Marden
Associate Professor Yuan Ji, University of Chicago
Assistant Professor Dave Zhao
Visiting Assistant Professor Trevor Park

Abstract

The innovation of modern technologies drives research and development on high-dimensional data analysis in diverse fields, where variable selection plays a pivotal role to ensure credible model estimation. We focus on scalable algorithms for variable selection that can handle large data sets.

Firstly, we propose an EM algorithm that returns the MAP estimate of the set of relevant variables. Due to its particular updating scheme, our algorithm can be implemented efficiently. We also show that the MAP estimate returned by our EM algorithm achieves variable selection consistency even when p diverges with n . In practice, EM algorithm tends to get stuck at local peaks. So we propose an ensemble version: repeatedly apply the EM algorithm on a subset of Bootstrap sample data and then aggregate the results. Empirical studies demonstrate the superior performance of this Bayesian Bootstrap EM algorithm.

Secondly, we propose a hybrid computation framework for Bayesian variable selection. This new algorithm SAB is a combination of the classical EM algorithm and the variational Bayes algorithm. It is very fast in handling high dimensional data with a large number of covariates. To address a critical biological problem, we apply SAB to a state-of-art cancer genomics data set with a goal to understand the complex regulatory relationship between miRNAs and mRNAs in cancer.

In the third part, we study the asymptotic behavior of the SAB algorithm in detail and prove that SAB achieves the selection consistency, Bayesian consistency and also an oracle property when the number of covariates p grows with the sample size n exponentially.

Lastly, we extend the hybrid framework of Bayesian variable selection to logistic models, where we adopt the Polya-Gamma specification and show that this specification is equivalent as the local approximation method in the variational Bayes framework.

Table of Contents

List of Tables	vi
List of Figures	vii
Chapter 1 Introduction	1
1.1 Penalized Likelihood Approach	1
1.2 Bayesian Regularization	3
1.3 Bayesian Hierarchical Formulation	4
1.3.1 The prior choices	5
1.3.2 Connection to regularization	6
1.4 Posterior Computation	8
1.4.1 Sampling approaches	8
1.4.2 MAP estimation	10
1.4.3 Approximation approaches	11
1.5 Extension to Generalized Linear Model	13
Chapter 2 EM and Ensemble	14
2.1 Introduction	14
2.2 Method	16
2.2.1 Prior specification	16
2.2.2 The EM algorithm	17
2.2.3 Computation cost	19
2.3 Asymptotic Analysis	20
2.4 Bayesian Bootstrap	21
2.5 Empirical Results	24
2.5.1 A widely used benchmark	25
2.5.2 A highly-correlated data	26
2.5.3 A Large- p small- n example	28
2.5.4 A real example	30
2.6 Discussion	32
Chapter 3 Variational Bayes	33
3.1 Introduction	33
3.2 Method	35
3.2.1 Model and priors	35

3.2.2	Posterior inference	37
3.2.3	A hybrid algorithm	38
3.3	Asymptotic Analysis	42
3.4	Empirical Results	44
3.4.1	Simulation studies	44
3.4.2	Cancer genomics data on miRNA-mRNA regulation . .	48
3.5	Discussion	55
Chapter 4	Asymptotic Analysis in High Dimensions	56
4.1	Introduction	56
4.2	Method	58
4.3	Set-up and Notations	60
4.4	Orthogonal Design	61
4.5	Full-rank Data	65
4.5.1	Selection consistency of one-step estimator	66
4.5.2	Bayesian consistency	67
4.5.3	Selection consistency of t -step estimator	68
4.5.4	Oracle property	69
4.6	High-dimensional Data	70
4.6.1	Selection consistency	71
4.6.2	Oracle property	72
Chapter 5	Logistic Model	73
5.1	Introduction	73
5.2	Method	75
5.2.1	Model and priors	75
5.2.2	The variational algorithm	76
5.2.3	Connection with local approximation	79
5.3	Empirical Results	80
Appendix A	Supplementary Material of Chapter 2	82
A.1	Proof of Theorem 2.1	82
Appendix B	Supplementary Material for Chapter 3	85
B.1	Proof of Proposition 3.1	85
B.2	Implementation of SAB	86
B.3	Conditions and Proofs of Theorems in Section 3.3	88
Appendix C	Supplementary Material for Chapter 4	95
C.1	Lemma and properties	95
C.2	General results	96
C.3	Proof of Theorem 4.1	99
C.4	Proof of Theorem 4.2	101
C.5	Proof of Theorem 4.3	102
C.6	Proof of Theorem 4.4	103

C.7	Proof of Theorem 4.5	105
C.8	Proof of Theorem 4.6	109
Appendix D	Supplementary Material for Chapter 5	111
D.1	Objective function	111
D.2	Local approximation	111
References		113

List of Tables

2.1	A widely used benchmark. The average number of zero-coefficients (i.e., no selection) out of 100 simulations for each types of variable (Signal or Noise) are shown. The results other than EM (Alg 1) are from (Fan and Li, 2001).	26
2.2	A widely used benchmark. The min, median, max number of being selected out of 100 simulations for each types of variable (Signal or Noise) are shown. The results other than EM (Alg 1) are from (Xin and Zhu, 2012).	27
2.3	A highly-correlated data. The min, median, max number of times being selected (i.e., no selection) out of 100 simulations for each type of variables (Signal and Noise) are shown. The results other than EM and BBEM are from (Xin and Zhu, 2012).	28
2.4	A large- p small- n example. The table shows the average number of signal and noise variables being selected out of 100 iterations. In BBEM, v_0 is either chosen by BIC or fixed at 0.03. EMVS is the algorithm proposed by Ročková and George (2014).	30
3.1	Study 1. The average number of 0 coefficients (i.e., no selection) out of 100 simulations for each type of variables (Signal or Noise). The ideal value is 5 for the 1st column and 0 for the 2nd column. The results of SCAD1, SCAD2 and LASSO are taken from Fan and Li (2001).	47
3.2	Study 2. The frequencies of being selected (out of 30 simulations) for the two types of variables (Signal or Noise), where EMVS is from (Ročková and George, 2014) and Oracle corresponds to the ideal case where we know the true relevant variables. The last column records the computation time on a Macbook Air (1.7 GHz Intel Core i5 and 4 GB 1333 MHz DDR3 memory).	47
5.1	Dense feature example. The mean and standard error (in parenthesis) of the 0/1 loss and negative log-likelihood from 50 simulations with different correlation ρ are shown for both LVB in Algorithm 4 and LASSO from the glmnet package.	81

List of Figures

1.1	A graphical display of the penalty function associated with the two-component normal mixture prior (1.3) for various choices of v_0 and v_1	7
1.2	A graphical display of the truncated posterior mean $\tilde{\beta}$ in y-axis versus the OLS estimator β_0 in x-axis with the two-component normal mixture prior (1.3) for various choices of v_0 and v_1 . . .	8
2.1	Highly-correlated data $n = 50$. A path-plot of the average selection frequency when v_0 varies in the logarithm scale of base 10. Top 6 lines represent the true variables $\mathbf{x}_{1:6}$ and the bottom 3 lines represent the maximum, median and minimum among the noise variables $\mathbf{x}_{7:40}$	29
2.2	Restaurant data. The path plot of selection frequency when v_0 varies in the logarithm scale of base 10. Only a subset of variables with high selection frequencies are displayed.	31
3.1	Left: The SAB path (averaged over 20 data sets) for the posterior mean of β_j for data from Simulation Study 1 with $n = 40$, $p = 8$, and $\sigma = 3$. Right: The averaged SAB path for the inclusion probability ϕ_j . In both plots, the paths of the 3 relevant variables, X_1 , X_2 and X_5 , are shown as solid lines and the paths for the other variables are shown as dashed lines. The x -coordinate denotes v_0 changed in the logarithm scale of base 10.	46
3.2	Study 3. The SAB path for the inclusion probability ϕ_j on the Zhao and Yu's example, where the x -coordinate denotes v_0 changed in the logarithm scale of base 10. In this example, the true variables are X_1 and X_2	48
3.3	Histogram of the numbers of miRNAs selected for all the mRNAs using the SAB algorithm.	52
3.4	Application of SAB to mRNA-miRNA regulation using TCGA data. Top 50 influential miRNAs are shown. The bar lengths represent the number of mRNAs that each miRNA significantly repressed.	53

Chapter 1

Introduction

The emergence of advance computational technologies drives the research of statistical methodologies to extract information from massive data sets. There are two aspects of interest, *learning* and *inference*. The former is referred to model estimation, while the latter is to predict using the model. With the presence of redundant features which is common for high-dimensional data, an important task is to identify a parsimonious model that has better statistical interpretability as well as prediction accuracy. The focus of this thesis is centered around variable selection, which is a special case of model selection, to identify a subset of features that can explain well the variation in the observed response.

Variable selection has been a long-standing problem in scientific studies. The classical idea is to select variables through either sequential searching (forward, backward or step-wise) or exhausting all possible subsets by minimizing the combination of residual sum of square (RSS) and a penalty term such as AIC/BIC. Although conceptually straightforward, the drawbacks of such searching algorithms are well recognized: computational too expensive for high-dimensional data and prone to stuck at local optima for step-wise searching. A variety of advanced tools have been developed and investigated recently to address these issues, among which two general frameworks are adopted: penalized likelihood approach and Bayesian method, the latter being the focus of my thesis. We devote the rest of this chapter to a review of these general frameworks of variable selection, leaving more detailed discussion of the recent work to the introduction sections of each chapter.

1.1 Penalized Likelihood Approach

Penalized likelihood approach or sometimes called the regularization method has been widely discussed and intensively used in many applications. It

identifies the relevant variable set and estimates their effects simultaneously by minimizing the combination of an empirical error (negative logarithm of the likelihood) plus a penalty term on the size of the effects.

$$\mathcal{L}(\theta) = -\log \text{Lik}(\mathbf{y} \mid \theta) + \text{Pen}_\lambda(\theta).$$

The goal is to find the optimal θ that minimizes $\mathcal{L}(\theta)$. In certain cases, the first term is also written as the residual sum of square (RSS) which alone leads to least square estimation. The second term is a penalty function indexed by a hyperparameter λ . For ease of exposition, we limit our analysis to the linear regression model with iid Gaussian errors, where the response variable $\mathbf{y}$ relates to a set of potential predictors $X_1, \dots, X_p$ in the form of a linear function. Let $\mathbf{X}_{n \times p}$ denote the design matrix and $\boldsymbol{\beta}_{p \times 1}$ the unknown regression coefficient vector. The response $\mathbf{y}$ is modeled by

$$\mathbf{y} \sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

where σ^2 denotes the error variance and $\mathbf{N}_n$ denotes a n -dimensional multivariate Normal distribution. For simplicity, set $\sigma^2 = 1$. Then the optimization problem of penalized likelihood approaches translates to

$$\mathcal{L}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \text{Pen}_\lambda(|\beta_j|) \quad (1.1)$$

A popular choice of the penalty function is the L_q penalty: $\text{Pen}_\lambda(|\beta_j|) = \lambda|\beta_j|^q$, where $q = 0$ corresponds to the classical criteria like AIC/BIC, $q = 1$ the popular LASSO penalty (Tibshirani, 1996) and $q = 2$ resulting in ridge regression (Hoerl and Kennard, 1970). While the L_0 penalty being the most essential sparsity measure strongly favors sparse solutions, it is NP-hard for optimization. The L_1 penalty which is a convex relaxation of the L_0 penalty, however, is more practical for implementation.

Fan and Li (2001) argued that a good penalty function should possess three properties: 1) unbiasedness, where the resulting estimator of the true effects are nearly unbiased, 2) sparsity, where the optimization procedure automatically zeros out unimportant variables and thereby leads to parsimonious solution, and 3) continuity, that is the estimation is robust towards small perturbation of the data to avoid prediction instability. The estima-

tor of the L_q penalty is continuous if and only if $q \geq 1$, but the sparsity condition is failed when $q > 1$. The L_1 penalty on the other hand, suffers from noticeable large model bias that shifts the estimator by a constant λ . Motivated by this observation, researchers have proposed penalty functions that fulfill these three conditions, among others are the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2001), the minimax convex penalty (MCP) (Zhang, 2010) and the adaptive LASSO (Zou, 2006). Apart from being the guideline of a good penalty function, these appealing properties also provide an insightful angle to evaluate Bayesian variable selection algorithms.

1.2 Bayesian Regularization

There is a natural connection between the penalized likelihood approach and the Bayesian framework: optimizing the penalized likelihood corresponds to the maximize-a-posteriori (MAP) process with certain prior. To be more specific, minimizing Equation (1.1) is equivalent as maximizing the following objective function:

$$\exp\{-\mathcal{L}(\boldsymbol{\beta})\} = \exp\{-\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2\} \exp\{-Pen_\lambda(\boldsymbol{\beta})\},$$

where the first term is proportional to the likelihood when variance σ^2 is fixed, and the second term can be viewed as a prior assigned for coefficients $\boldsymbol{\beta}$. The minus logarithm of the prior distribution can be interpreted as the penalty term,

$$Pen_\lambda(\boldsymbol{\beta}) = -\log \pi(\boldsymbol{\beta} \mid \lambda). \quad (1.2)$$

For example, the LASSO penalty corresponds to the double exponential (Laplace) prior for $\pi(\boldsymbol{\beta} \mid \lambda) = \exp\{-\lambda|\boldsymbol{\beta}|\}$, which is essentially a scale mixture of normals with an exponential mixing density. The Bayesian analogue to LASSO was proposed by (Park and Casella, 2008). While the frequentist LASSO chooses the shrinkage parameter λ by cross validation, (Park and Casella, 2008) suggested using either empirical Bayes through marginal maximum likelihood or a diffusing hyper-prior (for example Gamma prior on λ^2). Another method adapted from the recent proposed penalized likelihood approach is Bayesian Elastic net (Zou and Hastie, 2005; Li and Lin, 2010)

which uses a compromise of the Gaussian and Laplacian prior on β that is related to a mixture of the L_1 and L_2 penalty.

Despite the similarity, the sparsity feature of LASSO and elastic net is not transmitted to their Bayesian analogues. For variable selection, Bayesian regularization approaches need a hard shrinkage to zero out redundant variables, for example the credible interval criterion and the scaled neighborhood criterion suggested in (Li and Lin, 2010).

1.3 Bayesian Hierarchical Formulation

Bayesian hierarchical model is another framework of variable selection which introduces a p -dimensional binary vector $\gamma = (\gamma_1, \dots, \gamma_p)^T$ as a model index: $\gamma_j = 1$ if the j^{th} variable is included and 0 if not. Let $p_\gamma = \sum_j \gamma_j$ denote the size of the variable set γ , $\mathbf{X}_\gamma$ denote the $n \times p_\gamma$ design matrix which includes only the variables entering the model γ , and β_γ be the $p_\gamma \times 1$ vector that corresponds to the coefficients. Then the linear model becomes:

$$\mathbf{y} \sim \mathbf{N}_n(\mathbf{X}_\gamma \beta_\gamma, \sigma^2 \mathbf{I}_n).$$

The two key issues in this Bayesian approach are 1) the prior choice, and 2) computation of the posterior (Clyde and George, 2004). A popular choice of the prior on the model index $\pi(\gamma)$ is the product of independent Bernoulli distributions. The conjugate Zellner’s “ g -prior” (Zellner, 1986) and the “spike-and-slab” prior are widely-used for $\pi(\beta | \gamma)$ the prior of the coefficient given the model index, which is further elaborated in Section 1.3.1. After specifying the priors on all the unknowns, the Bayesian approaches make inference on the posterior distribution of γ (Zellner, 1971; Mitchell and Beauchamp, 1988; George, 2000; Liang et al., 2008) to select variables. Among a variety of posterior summary measures, two most common strategies are the median probability model (MPM) (Barbieri and Berger, 2004) that includes all variables with $p(\gamma_j | \mathbf{y}) > 0.5$, and the highest posterior density (HPD) model that selects the model with the highest posterior density $p(\gamma | \mathbf{y})$.

1.3.1 The prior choices

In the Bayesian hierarchical formulation, a popular prior choice is the independent Bernoulli distribution for each γ_j ,

$$\pi(\gamma_j) = \text{Bern}(\theta_j),$$

where $\theta = (\theta_1, \dots, \theta_p)$ is a $p \times 1$ hyper-parameter vector that contributes to small fraction of the sparsity and thereby can be used to incorporate the prior knowledge. Choosing small θ leads to a sparse solution. To give more flexibility, one can put another layer of hierarchy on the hyper-parameter θ , where the conjugate Beta prior is often used:

$$\pi(\theta) = \text{Beta}(a, b).$$

A more general prior of the model space is $\pi(\gamma) = w_{p_\gamma} / \binom{p}{p_\gamma}$ where w_{p_γ} is a prior weight put on models of size p_γ (Chipman et al., 2001). Given the model size, all possible models of the same size are equally likely. The weights w_{p_γ} reflect user's preference on sparsity. One can put more weight on smaller size if a parsimonious solution is favored, one example being the truncated Poisson prior (Denison et al., 1998).

The prior of the coefficient β given γ has the following popular choices.

Zellner's g -prior (Zellner, 1986) is a conjugate prior that attracted a lot attention.

$$p(\beta_\gamma \mid \gamma, g) = \text{N}_{p_\gamma}(0, g\sigma^2(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}),$$

where g is a tuning parameter that controls the expected size of the effects. Such prior allows efficient MH algorithms for computing the posterior (George and McCulloch, 1997; Smith and Kohn, 1996), however has the limitation of inconsistency in model comparison (Berger et al., 2001).

Spike-and-slab prior (Mitchell and Beauchamp, 1988) is another widely applied prior for the coefficients:

$$\pi(\beta_j \mid \gamma_j) = (1 - \gamma_j)\delta_0(\beta_j) + \gamma_j\pi_1(\beta_j),$$

where $\delta_0(\cdot)$ is the Kronecker delta function. The prior has two com-

ponents: the “spike” component that puts probability mass at 0 and shrinks the coefficients of all the irrelevant variables to 0, and the “slab” component that has a continuous prior distribution $\pi_1(\cdot)$ and induces soft shrinkage on the large effects. In the original form, the “slab” component is a uniform diffusive distribution although other alternatives are more common for example the Gaussian distribution.

Continuous spike-and-slab prior (George and McCulloch, 1993) is a continuous relaxation of the spike-and-slab prior to address the intractability issue, where the delta function is replaced by a spiky normal distribution.

$$\pi(\beta_j | \gamma_j) = (1 - \gamma_j) \mathbf{N}(\beta_j | 0, v_0 \tau^2) + \gamma_j \mathbf{N}(\beta_j | 0, v_1 \tau^2),$$

where v_0 and v_1 are two tuning parameters that control the relative size of signal and noise $0 < v_0 < v_1$. In its original form, τ^2 is a fixed hyper-parameter. (Ishwaran and Rao, 2003) suggested to assign an Inverse-Gamma (IG) prior on τ^2 and set $v_1 = 1$ for standardized $\mathbf{X}$ and $\mathbf{y}$ in the linear regression model.

1.3.2 Connection to regularization

The Bayesian regularization methods are different from the Bayesian hierarchical framework because they do not treat the model index as parameters nor assign prior over the model space, but make inference based on the posterior of coefficients. Albeit the difference, we can discuss the Bayesian regularization of the continuous spike-and-slab prior by translating it to a penalty function (1.2). Integrating out γ_j , we write the spike-and-slab prior (3.4) as a mixture of two normal distributions,

$$\pi(\beta_j | \theta, \sigma^2) = \theta \phi(\beta_j; 0, v_1 \sigma^2) + (1 - \theta) \phi(\beta_j; 0, v_0 \sigma^2), \quad (1.3)$$

where $\phi(\cdot; \mu, \tau^2)$ denotes the pdf of $\mathbf{N}(\mu, \tau^2)$. In Figure 1.1, we plot the negative logarithm of the prior in (1.3) with $\theta = 1/2$ and $\sigma^2 = 1$ for various choices of v_0 and v_1 ; we also shift the curves by $\log \pi(0 | \theta, \sigma^2)$ so that the penalty at $\beta_j = 0$ is zero. Compared with the popular L_1 penalty or equivalently the Double Exponential prior, the two-component normal mixture

prior (1.3) has an attractive feature from a regularization point of view: it reaches a plateau when β is large therefore less bias for estimation, and gives rise to a continuous approximation of the L_0 penalty, although not convex. The Bayesian penalties for g -prior and the spike-and-slab prior are similar to L_0 penalty like AIC/BIC.

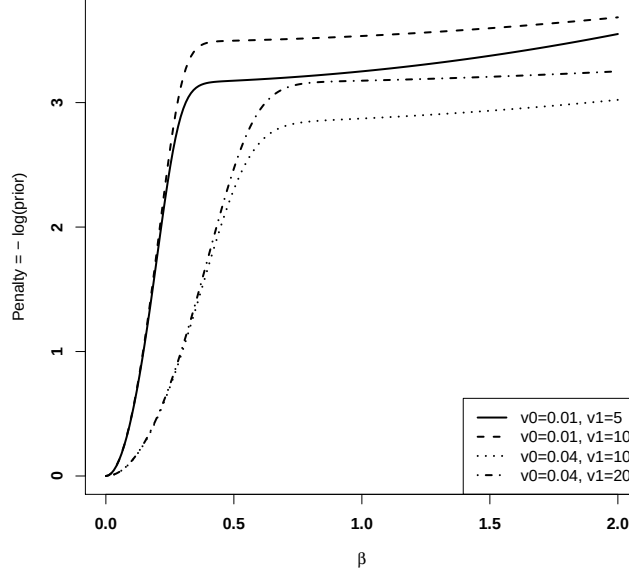


Figure 1.1: A graphical display of the penalty function associated with the two-component normal mixture prior (1.3) for various choices of v_0 and v_1 .

Let β_0 be the OLS estimator of the normal mean model with only one observation, $\beta_0 = y$. Then the posterior mean is

$$\begin{aligned}\hat{\beta} &= \mathbb{E}(\beta \mid \beta_0) = \phi \mathbb{E}(\beta \mid \gamma = 1, \beta_0) + (1 - \phi) \mathbb{E}(\beta \mid \gamma = 0, \beta_0) \\ &= \phi \frac{v_1}{1 + v_1} \beta_0 + (1 - \phi) \frac{v_0}{1 + v_0} \beta_0, s\end{aligned}$$

where $\phi = \pi(\gamma = 1 \mid \beta_0)$, and $\log \frac{\phi}{1-\phi} = \mathbf{N}(\beta_0; 0, 1 + v_1) / \mathbf{N}(\beta_0; 0, 1 + v_0)$. In Figure 1.2, we plot the truncated posterior mean $\tilde{\beta} = \hat{\beta} \cdot \mathbf{1}_{\{\phi \geq 1/2\}}$ versus the OLS estimator β_0 for various choices of v_0 and v_1 . The advantage of the Bayesian penalty is that it both returns a sparse solution and has diminishing bias, which are the desired properties discussed in (Fan and Li, 2001).

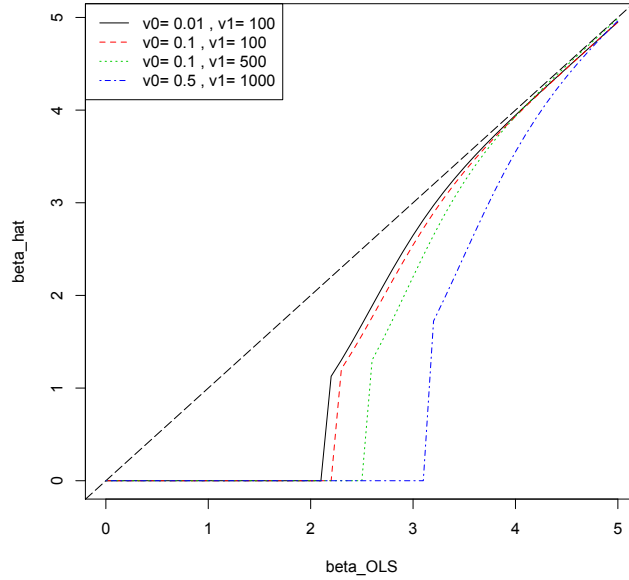


Figure 1.2: A graphical display of the truncated posterior mean $\tilde{\beta}$ in y-axis versus the OLS estimator β_0 in x-axis with the two-component normal mixture prior (1.3) for various choices of v_0 and v_1 .

1.4 Posterior Computation

1.4.1 Sampling approaches

To find out the most probable model, the exhaustive or step-wise search over the whole model space is impractical as the space size 2^p grows too rapidly with the dimension. Early efforts to address the scalability issue include importance sampling and branch-and-bound reduction strategies. Then the invention of MCMC techniques (including Gibbs sampler and Metropolis Hasting) brings about a proliferation of stochastic alternatives for searching the highest probability model, since a sequence of models $\gamma^{(1)}, \gamma^{(2)}, \gamma^{(3)}, \dots$ could be simulated such that their empirical distribution converges to the posterior $p(\gamma \mid \mathbf{y})$. One of the appealing features of MCMC is that it releases the close-form requirement for the posterior and thereby welcomes a wide variety of priors. However, the Metropolis Hasting using the conjugate priors is more practical, due to the advantage of rapid computation of marginal likelihood $p(\mathbf{y} \mid \gamma)$ in the following steps.

1. Given the Markov chain $\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(k)}$, propose a new model γ^*

from the proposal distribution $j(\gamma^* | \gamma^{(k)})$.

2. Let $\gamma^{(k+1)} = \gamma^*$ with probability below, otherwise $\gamma^{(k+1)}$ remains $\gamma^{(k)}$,

$$\min\left\{1, \frac{p(\mathbf{y} | \gamma^*)\pi(\gamma^*)}{p(\mathbf{y} | \gamma)\pi(\gamma)} \times \frac{j(\gamma | \gamma^*)}{j(\gamma^* | \gamma)}\right\}.$$

Markov chain Monte Carlo model composition (MC³) (Madigan et al., 1995) is one Metropolis Hasting algorithm that proposes a new model γ^* differing from the current γ by one variable and accept it with the rate $\min\{1, \frac{p(\gamma^*|\mathbf{y})}{p(\gamma|\mathbf{y})}\}$, which has the close-form in conjugate prior setup.

Marginalizing out the parameters β and searching over a discrete 2^p model space cause the model space approaches prone to be trapped within local peaks. Recent development to address this issue includes 1) adopting the “tempering” idea from physics that runs a population of tempered chains in parallel with flattened peaks such as Population-based MCMC (Jasra et al., 2007) and Evolutionary MCMC (Liang and Wong, 2000), and 2) combining Oscar’s Razor idea with Metropolis Hasting routine to determine a bigger set of best models such as Shotgun Stochastic Search (Hans et al., 2007).

Some other model space approaches explore the model space and the parameter space simultaneously, which eases the odds of being trapped around local optima, but the difficulty is that the dimension of parameter space changes at every step. Reversible Jump MCMC (Green and Hastie, 2009) can be used to adjust the dimension of β_γ dynamically. A Gibbs sampler method was proposed by (Carlin and Chib, 1995) and further extended to Gibbs variable selection (GVS) (Dellaportas et al., 2000), where a continuous spike-and-slab prior is used but the “spike” part is “pseudo-prior” meaning that the corresponding coefficients do not contribute to the predictor. Stochastic Search Variable Selection (SSVS) (George and McCulloch, 1993) is another Gibbs sampling approach on Bayesian hierarchical mixture model. Although both adopted the continuous version of spike-and-slab prior, SSVS is different from GVS in the sense that the coefficients β_j given $\gamma_j = 0$ of the former approach actually influence the posterior.

1.4.2 MAP estimation

Facing the increasing demand on scalability from various applications, the Bayesian approach for variable selection has one limitation, that is MCMC algorithms are computationally heavy and hard to scale with large data. In many real applications, computing the exact posterior distribution is time-consuming and unnecessary, so we turn our focus to the MAP estimation (EM algorithm) or the approximation approaches for fast solutions.

In the regularization framework, the key is to maximize a penalized likelihood which can be done based on the local quadratic approximation (LQA) (Fan and Li, 2001). Hunter and Li (2005) noted that LQA has a natural connection with the minorization-maximization (MM) algorithm and furthermore modified LQA algorithm by adding a perturbation to render the objective function differentiable. Griffin and Brown (2005) adopted the Bayesian shrinkage prior and proposed an EM algorithm corresponding to LQA. However, there is a gap in the Bayesian hierarchical framework for the EM-like variable selection algorithm until recently Ročková and George (2014) proposed an efficient EM algorithm EMVS which adopts the continuous “spike-and-slab” prior on β . The continuity of the prior on β turns out to be crucial for fast algorithms. To obtain the estimation of the model index γ , EMVS uses a two-stage approach: first use the EM algorithm to obtain the MAP estimator of β , then use a threshold to select the desired subset of variables.

Motivated by (Ročková and George, 2014), we propose an efficient EM algorithm that can return the MAP estimator of the model index γ directly, instead of a two-stage approach. Our EM algorithm is very fast and scales easily due to a novel computation trick that avoids inverting a large $p \times p$ matrix in each iteration. We further proposed an ensemble version of the EM algorithm based on Bayesian bootstrap BBEM. The main idea is to repeatedly apply a stochastic version of the EM algorithm on a subset of the data (the Bayesian bootstrap samples), and then aggregate the variable selection results across the bootstrap samples to improve the accuracy. The details are given in Chapter 2.

1.4.3 Approximation approaches

The MAP procedure is very fast in computation but only provides a point estimation. When the full posterior distribution is preferred, we consider the approximation approaches. Variational Bayes (VB) and Expectation propagation (EP) are two closely-related techniques for posterior approximation yet show advantages in different scenario. The approximation approaches introduce a factorized distribution Q and update each component in Q sequentially to minimize the Kullback-Leibler (KL) divergence between Q and the true posterior P until convergence. The KL divergence (Kullback and Leibler, 1951) is a non-negative, asymmetric metric defined as:

$$KL(P||Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx.$$

A fundamental difference between VB and EP is that the objective function of the former approach is $KL(Q||P)$ whereas the latter working on $KL(P||Q)$. Other extensions include power EP (Minka, 2004) that based on the α -divergence:

$$D_\alpha(P||Q) = \frac{1}{\alpha(1-\alpha)} \left(1 - \int P(x)^\alpha Q(x)^{(1-\alpha)} dx \right),$$

with $\alpha \rightarrow 0$ corresponding to VB, $\alpha \rightarrow 1$ corresponding to EP and $\alpha = 0.5$ corresponding to the Hellinger distance.

In the context of linear model, with the variance parameter σ^2 and the Bernoulli parameter θ fixed, the full posterior is

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}) \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) \pi(\boldsymbol{\gamma}).$$

The EP approach factorizes Q by:

$$Q(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{i=1}^n q_i(y_i, \boldsymbol{\beta}) \prod_{j=1}^p q_j(\beta_j, \boldsymbol{\gamma}) q(\boldsymbol{\gamma})$$

where $q_i(y_i, \boldsymbol{\beta})$, $q_j(\beta_j, \boldsymbol{\gamma})$ and $q(\boldsymbol{\gamma})$ correspond to the likelihood term $p(y_i \mid \boldsymbol{\beta})$, the prior $\pi(\beta_j)$ and $\pi(\boldsymbol{\gamma})$ respectively. Hernández-Lobato et al. (2010, 2013) introduced an efficient EP algorithm for multi-task feature selection and group feature selection that use a generalized spike-and-slab prior with

a point mass component. Recently, Andersen et al. (2014) proposed another EP framework for Bayesian feature selection that uses a structured spike-and-slab prior to incorporate the prior knowledge of the sparsity pattern. Carbonetto et al. (2012) introduced a VB approach for variable selection that adopts the original spike-and-slab prior and factorizes Q in the following form,

$$Q(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \prod_{j=1}^p q(\beta_j, \gamma_j).$$

The use of the point mass component demands the dependency of β_j and γ_j in Q due to the non-conjugate prior.

In Chapter 3, we propose a hybrid framework of VB and EM for scalable Bayesian variable selection called SAB that adopts the continuous spike-and-slab prior. Unlike (Carbonetto et al., 2012), we factorize the approximate function Q in the following way,

$$Q(\boldsymbol{\gamma}, \boldsymbol{\beta}) = q(\boldsymbol{\beta}) \prod_{j=1}^p q(\gamma_j).$$

The optimal $q(\boldsymbol{\beta})$ and $q(\gamma_j)$ are in the multivariate Normal family and the Bernoulli family respectively, which are conjugate to the priors. In this approach, we assume independence between $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for the approximate posterior, which entails Q cannot be as close as to P but the convergence is guaranteed due to convexity. Instead of reporting just a single model, the SAB algorithm returns the posterior probabilities of $\boldsymbol{\gamma}$, which provides more flexibility on model inference such as model ranking, FDR control and model averaging.

As a trade-off between computation cost and accuracy, we do not expect that the VB algorithm obtains the exact posterior distribution like MCMC methods. Instead we are looking for a good approximation of the posterior distribution. Here by “good”, we mean that the result from SAB, albeit an approximation, achieves the desired asymptotic properties, such as the selection consistency and the oracle property. The asymptotic analysis of the SAB algorithm for high dimensional data ($p > n$) is given in Chapter 4.

1.5 Extension to Generalized Linear Model

In Chapter 5, we discuss the extension of our Bayesian variable selection framework to the Generalized Linear Model (GLM). Logistic regression is the most commonly used method to model binary response data. But transforming the nicely-built algorithms from linear regression to logistic model is not always straightforward. In the Bayesian framework, the difficulty mainly comes from the non-conjugate likelihood. Several solutions have been proposed from one of the three basic ideas: the MAP estimation, sampling algorithms or approximation approaches. To be more specific, the MAP procedure provides a point estimator which is fast but surely misses out some information of the whole posterior. The sampling method uses a Metropolis Hasting algorithm to sample from the posterior, which is more accurate and informative but notorious for its slow mixing time. The approximation approach, on the other hand, uses different factorization functions (often connect to the exponential family) to approximate the posterior, for example the probit approximation or Laplace approximation.

A recent work by (Polson et al., 2013) introduced an auxiliary variable from Polya-Gamma distribution in the GLM specification. Given that auxiliary variable, the logistic likelihood is valued as Normal densities, which bypasses the non-conjugate issue so that the parameters and latent variables can be updated using an EM algorithm. But in the scenario of variable selection, this method has a drawback, that is for sparse data, if a variable with small coefficient estimation is deleted, it cannot reenter the model.

Based on the Polya-Gamma method, we propose a hybrid framework of VB and EM for logistic variable selection that adopts the continuous spike-and-slab prior in Chapter 5. We also show an equivalence between the Polya-Gamma method and the local approximation method (Bishop, 2006) for VB in logistic model.

Chapter 2

EM and Ensemble

2.1 Introduction

Consider a simple linear regression model with Gaussian noise:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (2.1)$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$ is the $n \times 1$ response, $\mathbf{e} = (e_1, \dots, e_n)^T$ is a vector of iid Gaussian random variables with mean 0 and variance σ^2 , and $\mathbf{X}$ is the $n \times p$ design matrix. The unknown parameters are the regression parameter $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ and the error variance σ^2 . In many real applications such as bioinformatics and image analysis, where linear regression models have been routinely used, the number of potential predictors (i.e., p) is large but only a small fraction of them is believed to be relevant. Therefore the linear model (2.1) is often assumed to be “sparse” in the sense that most of the coefficients β_j ’s are zero. Estimating the set of relevant variables, $S = \{j : \beta_j \neq 0\}$ is an important problem in modern statistical analysis.

The Bayesian approach to variable selection is conceptually simple and straightforward. First introduce a p -dimensional binary vector $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^T$ to index all the 2^p sub-models, where $\gamma_j = 1$ if the j th variable is included in this model and 0 if excluded. Usually γ_j ’s are modeled by independent Bernoulli distributions. Given $\boldsymbol{\gamma}$, a popular prior choice for $\boldsymbol{\beta}$ is the “spike and slab” prior (Mitchell and Beauchamp, 1988):

$$\pi(\beta_j \mid \gamma_j) = \begin{cases} \delta_0(\beta_j), & \text{if } \gamma_j = 0; \\ g(\beta_j), & \text{if } \gamma_j = 1, \end{cases} \quad (2.2)$$

where $\delta_0(\cdot)$ is the Kronecker delta function corresponding to the density function of a point mass at 0 and g is a continuous density function. After

specifying priors on all the unknowns, one needs to calculate the posterior distribution. Most algorithms for Bayesian variable selection rely on MCMC such as Gibbs or Metropolis Hasting to obtain the posterior distribution; for a review on recent developments in this area, see O’Hara and Sillanpää (2009). MCMC algorithms, however, are insufficient to meet the growing demand on scalability from real applications. Since the primary goal is variable selection, we focus on efficient algorithms that return the MAP estimate of γ , as an alternative to these MCMC-based sampling methods that return the whole posterior distribution on all the unknown parameters.

Recently, Ročková and George (2014) proposed a simple, elegant EM algorithm for Bayesian variable selection. They adopted a continuous version of the “spike and slab” prior—the spike component in (2.2) is replaced by a normal distribution with a small variance (George and McCulloch, 1993), and proposed an EM algorithm to obtain the MAP estimate of the regression coefficient β . The MAP estimate $\hat{\beta}_{\text{MAP}}$, however, is not sparse, so an additional thresholding step is needed to estimate γ .

In this chapter, we develop an EM algorithm that directly returns the MAP estimate of γ , so no further thresholding is needed. We adopt the same continuous “spike and slab” prior. Different from the algorithm by Ročková and George (2014) that returns $\hat{\beta}_{\text{MAP}}$ by treating γ as latent, our algorithm returns the MAP estimate of the model index, $\hat{\gamma}_{\text{MAP}}$, by treating β as latent. The special structure of our EM algorithm allows us to use a computational trick to avoid inverting a big matrix at each iteration, which seems unavoidable in the algorithm by Ročková and George (2014). Further we can show that the $\hat{\gamma}_{\text{MAP}}$ achieves asymptotic consistency even when p diverges to infinity with the sample size n .

Although shown to achieve selection consistency, in practice, our EM algorithm could get stuck at a local mode due to the large discrete space in which γ lies. Borrowing the idea of bagging, we propose an ensemble version of our EM algorithm (which we call BBEM): apply the algorithm on multiple Bayesian bootstrap (BB) copies of the data, and then aggregate the variable selection results. Bayesian bootstrap for variable selection was explored before by Clyde and Lee (2001) for the purpose of prediction, where models built on different bootstrap copies are combined to predict the response. But the focus of our approach is to summarize the evidence for variable relevance from multiple BB copies, which is similar in nature to several frequentist

ensemble methods for variable selection, such as the AIC ensemble (Zhu and Chipman, 2006), stability selection (Meinshausen and Bühlmann, 2010), and random Lasso (Wang et al., 2011).

The remaining of the paper is organized as follows. Section 2 describes the EM algorithm in detail, Section 3 presents the asymptotic results, and Section 4 describes the BBEM algorithm. Empirical studies are presented in Section 5 and conclusions and remarks in Section 6.

2.2 Method

2.2.1 Prior specification

We adopt the continuous version of “spike and slab” prior for β , i.e. a mixture of two normal components with mean zero and different variances:

$$\pi(\beta_j \mid \sigma, \gamma_j) = \begin{cases} \mathbf{N}(0, \sigma^2 v_0), & \text{if } \gamma_j = 0; \\ \mathbf{N}(0, \sigma^2 v_1), & \text{if } \gamma_j = 1, \end{cases} \quad (2.3)$$

where $v_1 > v_0 > 0$. Alternatively, we can write the prior on β as

$$\pi(\beta_j \mid \sigma^2, \gamma_j) = \mathbf{N}(0, \sigma^2 d_{\gamma_j}),$$

where

$$d_{\gamma_j} = \gamma_j v_1 + (1 - \gamma_j) v_0.$$

The choice of tuning parameters v_0 and v_1 is discussed in Section 2.5.

For the remaining parameters, we specify independent Bernoulli priors on elements of γ , and conjugate priors like Beta and Inverse Gamma on θ and σ^2 , respectively:

$$\begin{aligned} \pi(\gamma \mid \theta) &= \text{Bern}(\theta), \\ \pi(\theta) &= \text{Beta}(a_0, b_0), \\ \pi(\sigma^2) &= \text{IG}(\nu_0/2, \nu_0 \lambda_0/2). \end{aligned}$$

For hyper-parameters $(a_0, b_0, \nu_0, \lambda_0)$, we suggest the following non-informative

choices unless prior knowledge is available:

$$a_0 = b_0 = 1.1, \quad \nu_0 = \lambda_0 = 1. \quad (2.4)$$

2.2.2 The EM algorithm

With the Gaussian model and prior distributions specified above, we can write down the full posterior distribution:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}, \theta, \sigma^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) \times \pi(\boldsymbol{\beta} \mid \sigma, \boldsymbol{\gamma}) \times \pi(\boldsymbol{\gamma} \mid \theta) \times \pi(\theta) \times \pi(\sigma^2).$$

Treating $\boldsymbol{\beta}$ as the latent variable, we derive an EM algorithm that returns the MAP estimation of parameters $\Theta = (\boldsymbol{\gamma}, \sigma^2, \theta)$, whereas the roles of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are switched in Ročková and George (2014).

E Step

The objective function Q at the $(t+1)$ -th iteration in an EM algorithm is defined as the integrated logarithm of the full posterior with respect to $\boldsymbol{\beta}$ given $\mathbf{y}$ and the parameter values from the previous iteration $\Theta^{(t)} = (\boldsymbol{\gamma}^{(t)}, \sigma_{(t)}^2, \theta^{(t)})$, i.e.,

$$\begin{aligned} Q(\Theta \mid \Theta^{(t)}) &= \mathbb{E}_{\boldsymbol{\beta} \mid \Theta^{(t)}, \mathbf{y}} \log \pi(\Theta, \boldsymbol{\beta} \mid \mathbf{y}) \\ &= -\frac{1}{2\sigma^2} \mathbb{E}_{\boldsymbol{\beta} \mid \Theta^{(t)}, \mathbf{y}} \left[\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \frac{\beta_j^2}{d_{\gamma_j}} \right] + F(\Theta), \end{aligned} \quad (2.5)$$

where

$$\begin{aligned} F(\Theta) &= -\frac{n+p}{2} \log \sigma^2 - \frac{1}{2} \sum_{j=1}^p \log d_{\gamma_j} + \log \pi(\boldsymbol{\gamma} \mid \theta) \\ &\quad + \log \pi(\theta) + \log \pi(\sigma^2) + \text{Constant} \end{aligned}$$

is a function of Θ not depending on $\boldsymbol{\beta}$.

It is easy to show that $\boldsymbol{\beta}$ follows a Normal distribution with mean $\mathbf{m}$ and covariance matrix $\sigma_{(t)}^2 \mathbf{V}$, given $\Theta^{(t)}$ and $\mathbf{y}$, where

$$\begin{aligned} \mathbf{m} &= \mathbf{V}^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{V} = (\mathbf{X}^T \mathbf{X} + D_{\boldsymbol{\gamma}^{(t)}}^{-1})^{-1}, \\ D_{\boldsymbol{\gamma}^{(t)}} &= \text{diag} \left(d_{\gamma_j^{(t)}} \right)_{j=1}^p = \text{diag} \left(\gamma_j^{(t)} v_1 + (1 - \gamma_j^{(t)}) v_0 \right)_{j=1}^p. \end{aligned} \quad (2.6)$$

Then the two expectation terms in (2.5) can be expressed as:

$$\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \|\mathbf{y} - \mathbf{X}\beta\|^2 = \sigma_{(t)}^2 \text{tr}(\mathbf{XVX}^T) + \|\mathbf{y} - \mathbf{Xm}\|^2, \quad (2.7)$$

$$\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \sum_{j=1}^p \frac{\beta_j^2}{d_{\gamma_j}} = \sum_{j=1}^p \frac{\sigma_{(t)}^2 V_{jj} + m_j^2}{(1 - \gamma_j^{(t)})v_0 + \gamma_j^{(t)}v_1}. \quad (2.8)$$

M Step

We sequentially update parameters (γ, θ, σ) to maximize the objective function Q .

1. **Update γ_j 's.** The terms involving γ_j in (2.5) are

$$-\frac{1}{2\sigma_{(t)}^2} \mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \left[\frac{\beta_j^2}{d_{\gamma_j}} \right] - \frac{1}{2} \log d_{\gamma_j} + \log \pi(\gamma_j | \theta^{(t)}). \quad (2.9)$$

Plug in $\gamma_j = 0$ and $\gamma_j = 1$ to (2.9) respectively, then we have

$$\gamma_j^{(t+1)} = 1, \quad \text{if } \mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} [\beta_j^2] > r^{(t)}, \quad (2.10)$$

where

$$r^{(t)} = \frac{\sigma_{(t)}^2}{1/v_0 - 1/v_1} \left(\log \frac{v_1}{v_0} - 2 \log \frac{\theta^{(t)}}{1 - \theta^{(t)}} \right).$$

2. **Update (σ^2, θ) .** Given $\gamma^{(t+1)}$, the updating equations for the other two parameters are given by

$$\sigma_{(t+1)}^2 = \frac{\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \left[\|\mathbf{y} - \mathbf{X}\beta\|^2 + \sum_{j=1}^p \beta_j^2 / d_{\gamma_j^{(t+1)}} \right] + v_0 \lambda_0}{n + p + v_0}, \quad (2.11)$$

$$\theta^{(t+1)} = \frac{\sum_{j=1}^p \gamma_j^{(t+1)} + a_0 - 1}{p + a_0 + b_0 - 2}. \quad (2.12)$$

Stopping Rule

The EM algorithm alternates between the E-step and M-step until convergence. A natural stopping criterion is to check whether the change of the objective function Q is small. To reduce the computation cost for evaluating the Q function, we adopt a different stopping rule as our main focus is γ : we stop our algorithm when the estimate $\gamma^{(t)}$ stays the same for k_0 iterations.

In practice, we suggest to set $k_0 = 3$. The pseudo code of this EM algorithm is summarized in Algorithm 1.

Algorithm 1: EM Algorithm Input: $\mathbf{X}, \mathbf{y}, v_0, v_1, a_0, b_0, \nu_0, \lambda_0$ Initialize $\Theta^{(0)}$; E-step: Calculate the two expectations in (2.7) and (2.8), denoted as $EE^{(0)}$; for $t = 1 : \text{maxIter}$ do M-step: Update $\Theta^{(t)}$ from Eq (2.10, 2.11, 2.12); E-step: Update $EE^{(t)}$ from Eq (2.7, 2.8); if $\gamma^{(t)}$ stays the same for $k_0 = 3$ iterations then break; end end Return $\gamma, \mathbf{m}, \mathbf{V}$;
--

2.2.3 Computation cost

At each E-step, updating the posterior of β given other parameters in (2.6) requires inverting a $p \times p$ matrix

$$\mathbf{V}_{(t)} = (\mathbf{X}^T \mathbf{X} + D_{\gamma^{(t)}}^{-1})^{-1}, \quad (2.13)$$

which is the major computational burden of the algorithm. When $p > n$, we can use the Sherman-Morrison-Woodbury formula to compute the inverse of an $n \times n$ matrix. So the computation cost at each iteration is of order $O(\min(n, p)^3)$. It is, however, still time-consuming when both n and p are large.

Note that the only thing that changes in (2.13) from iteration to iteration is $D_{\gamma^{(t)}}$, a diagonal matrix depending on the binary vector $\gamma^{(t)}$. From our experience, only a small fraction of $\gamma_j^{(t)}$'s are changed at each iteration after the first a couple of iterations. So the idea is to use the following recursive formula to compute $\mathbf{V}_{(t)}$:

$$\begin{aligned} \mathbf{V}_{(t)} &= (\mathbf{X}^T \mathbf{X} + D_{\gamma^{(t-1)}}^{-1} + D_{\gamma^{(t)}}^{-1} - D_{\gamma^{(t-1)}}^{-1})^{-1} \\ &= (\mathbf{V}_{(t-1)}^{-1} + D_{\gamma^{(t)}}^{-1} - D_{\gamma^{(t-1)}}^{-1})^{-1} \end{aligned} \quad (2.14)$$

where $D_{\gamma^{(t)}}^{-1} - D_{\gamma^{(t-1)}}^{-1}$ is a diagonal matrix with the j -th diagonal entry being non-zero only if the inclusion/exclusion status, i.e., the value of γ_j , is changed from the last iteration. Let l denote the number of variables whose γ_j values are changed from iteration $(t-1)$ to t . Then $D_{\gamma^{(t)}}^{-1} - D_{\gamma^{(t-1)}}^{-1}$ is a rank l matrix. We can apply the Woodbury formula on (2.14) to reduce the computation complexity from $O(\min(n, p)^3)$ to $O(l^3)$.

For example, without loss of generality, suppose only the first l covariates have their γ_j values changed. Then, we can write

$$D_{\gamma^{(t)}}^{-1} - D_{\gamma^{(t-1)}}^{-1} = U_{p \times l} A_{l \times l} U^T,$$

where $A = (\frac{1}{v_0} - \frac{1}{v_1}) \text{diag}(2\gamma_j^{(t)} - 1)_{j=1}^l$ and U consists of the first l columns from $\mathbf{I}_p$. Applying the Woodbury formula, we have

$$\mathbf{V}_{(t)} = \mathbf{V}_{(t-1)} - \mathbf{V}_{(t-1)} U (A^{-1} + U^T \mathbf{V}_{(t-1)} U)^{-1} U^T \mathbf{V}_{(t-1)}.$$

2.3 Asymptotic Analysis

In this section, we study the asymptotic property of $\hat{\gamma}_n$, the MAP estimate of model index returned by our EM algorithm. Assume the data $\mathbf{y}_n$ are generated from a Gaussian regression model:

$$\mathbf{y}_n \sim \mathbf{N}_n(\mathbf{X}_n \beta_n^*, \sigma^2 \mathbf{I}_n).$$

Here we consider a triangular array set up: the dimension $p = p_n$ diverges with n and the true coefficient β_n^* also vary with n . Suppose the true model is indexed by γ_n^* , where $\gamma_{nj}^* = 1$ if $\beta_{nj}^* \neq 0$ and $\gamma_{nj}^* = 0$ if $\beta_{nj}^* = 0$. We show that our EM algorithm has the following selection consistency property:

$$\mathbb{P}(\hat{\gamma}_n = \gamma_n^*) \rightarrow 1, \quad \text{as } n \rightarrow \infty.$$

First we list some regularity conditions needed in our proof. Let $\lambda_{\min}(A)$ denote the smallest eigenvalue of matrix A . We assume

(A1) On collinearity: let $\lambda_{\min}(A)$ denote the smallest eigenvalue of matrix A ,

$$\lambda_{\min}(\mathbf{X}_n^T \mathbf{X}_n)^{-1} = O(n^{-\eta_1}), \quad 0 < \eta_1 \leq 1.$$

In the traditional asymptotic setting where p is fixed, we have $\eta_1 = 1$.

(A2) On sparsity:

$$\|\beta_n^*\|_2 = O(n^{\eta_2}), \quad 0 < \eta_2 < \eta_1,$$

which controls the L_2 norm of the true regression coefficient vector.

(A3) The beta-min condition:

$$\liminf_n \frac{\min \{|\beta_{nj}^*|, \gamma_{nj}^* = 1\}}{n^{(\eta_3-1)/2}} \geq M, \quad 0 \leq \eta_3 < 1,$$

where M is a positive constant. This condition requires that the minimal non-zero coefficient cannot go to zero at a rate faster than $1/\sqrt{n}$. In the traditional asymptotic setting where β_n^* is fixed, we have $\eta_3 = 0$.

(A4) On hyper-parameters: assume $\log[\hat{\theta}_n/(1 - \hat{\theta}_n)]$ and $\hat{\sigma}_n^2$ are bounded.

This condition is purely technical for the simplicity of the proof. To satisfy this condition, we could fix $\hat{\theta}_n$ and $\hat{\sigma}_n$ at constant or set a threshold to avoid them going to extreme. In simulations, we recommend (2.4) as the choice for hyper-parameters unless p is large.

(A5) On tuning parameters: assume v_1 is fixed at constant and v_0 satisfies

$$0 < v_0 = O(n^{-r_0}), \quad 1 - \eta_3 < r_0 < \min \left\{ \eta_1 - \alpha, \frac{2}{3}(\eta_1 - \eta_2) \right\},$$

where $0 < \alpha < 1$ is the rate of the dimension $p = O(n^\alpha)$.

Theorem 2.1. *Assume (A1-A5) and $p = O(n^\alpha)$ where $0 \leq \alpha < 1$, then the model returned by our EM algorithm, $\hat{\gamma}_n$, achieves the following selection consistency,*

$$\mathbb{P}(\hat{\gamma}_n = \gamma_n^*) \rightarrow 1, \quad \text{as } n \rightarrow \infty. \quad (2.15)$$

Proof. See Appendix. □

2.4 Bayesian Bootstrap

A common issue with EM algorithms is that they could be trapped at a local maximum. There are some standard remedies available for dealing with this issue, for instance, trying a set of different initial values or utilizing some more

advanced optimization procedures at the M-step. Since our EM algorithm is searching for the optimal γ over a big discrete space, all p -dimensional binary vectors, these remedies are less useful when p is large.

When doing optimization with γ , a discrete vector, the resulting solution is often not stable, i.e., has a large variance. Bagging is an easy but powerful method (Breiman, 1996) for variance reduction, which applies the same algorithm on multiple bootstrap copies of the data, and then aggregates the final results. We proposed the following ensemble EM algorithm, in which we repeatedly run the EM variable selection algorithm, Algorithm 1 from Section 2.2.2, on Bayesian bootstrap replicates.

The original bootstrap repeatedly samples data from the original data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with replacement, i.e., each observation $(\mathbf{x}_i, y_i)$ is sampled with probability $1/n$. In Bayesian bootstrap (Rubin, 1981), instead of sampling a subset of the data, we assign a random weight w_i to the i -th observation and then fit a weighted least squares regression model on the whole data set. In particular, following Rubin (1981), we generate the weights $\mathbf{w} = (w_1, \dots, w_n)$ from a n -category Dirichlet distribution:

$$\mathbf{w}_{n \times 1} \sim \text{Dir}(1, \dots, 1). \quad (2.16)$$

When applying Algorithm 1 on a weighted linear regression model, all the updating equations stay the same, except the updating equation (2.6) for the posterior of β , which should be changed to:

$$\mathbf{m} = \mathbf{V}\mathbf{X}^T \text{diag}(\mathbf{w})\mathbf{y}, \quad \mathbf{V} = (\mathbf{X}^T \text{diag}(\mathbf{w})\mathbf{X} + D_{\gamma^{(t)}}^{-1})^{-1}. \quad (2.17)$$

Eq (2.7), the expectation of the weighted residual sum of squares, should also be changed accordingly:

$$\mathbb{E}_{\beta|\Theta^{(t)}, \mathbf{y}} \|\mathbf{y} - \mathbf{X}\beta\|_{\mathbf{w}}^2 = \sigma_{(t)}^2 \text{tr}(\text{diag}(\mathbf{w})\mathbf{X}\mathbf{V}\mathbf{X}^T) + (\mathbf{y} - \mathbf{X}\mathbf{m})^T \text{diag}(\mathbf{w})(\mathbf{y} - \mathbf{X}\mathbf{m}). \quad (2.18)$$

It is well-known that in order to make the aggregation work, we should control the correlation among estimates from bootstrap replicates. For example, in random forest (Breiman, 2001), the number of variables used for choosing the optimal split of a tree is restricted to a subset of the variables, instead of using all p variables. A similar idea was implemented in Random

Lasso (Wang et al., 2011), an ensemble algorithm for variable selection. In the same spirit, we apply the EM algorithm only on a subset of the variables at each Bayesian bootstrap iteration. A naive way is to randomly pick a subset from the p variables. This, however, will be inefficient when p is large and the true model is sparse, since it is likely most random subsets will not contain any relevant variables. So we employ a biased sampling procedure: sample the p variables based on a weight vector $\tilde{\pi}$ that is defined as

$$\tilde{\pi}_{p \times 1} \propto |\mathbf{X}^T \mathbf{y}| / \text{diag}(\mathbf{X}^T \mathbf{X}), \quad (2.19)$$

that is, variables are sampled based on their marginal effect in a simple linear regression.

The ensemble EM algorithm operates as follows. First we sample a random set of L variables according to the probability vector $\tilde{\pi}$, and draw a $n \times 1$ bootstrap weight vector $\mathbf{w}$ from (2.16). Let $\tilde{\mathbf{X}}$ be the new data matrix with the L columns. Then apply the EM algorithm on the bootstrap replicate $\tilde{\mathbf{X}}$ with weight $\mathbf{w}$. Let $\boldsymbol{\gamma}_k$ denote the model returned by the k -th Bayesian bootstrap iteration, where the j -th position of $\boldsymbol{\gamma}_k$ is 1 if the j -th variable is selected and zero otherwise; of course, the j -th position is zero if the j -th variable is not included in the initial L variables. Define the final variable selection frequency for the p variables as

$$\boldsymbol{\phi}_{p \times 1} = \frac{1}{K} \sum_{k=1}^K \boldsymbol{\gamma}_k. \quad (2.20)$$

We can report the final variable selection result by thresholding ϕ_j 's at some fixed number, for example, a half. Or we can produce a path-plot of $\boldsymbol{\phi}$ as v_0 varies, which could be a useful tool to investigate the importance of each variable. We illustrate this in our simulation study in Section 2.5.

As for the computational cost, the inversion of the $L \times L$ matrix in (2.17) is a big improvement compared with that of a $p \times p$ matrix. By the same logic as in Section 2.2.3, it can be further simplified to inverting a l -by- l matrix, where l is the number of variables that changes their inclusion status, and $l \leq L$. The complete BBEM algorithm is summarized in Algorithm 2.

Algorithm 2: BBEM Algorithm

Input: $\mathbf{X}, \mathbf{y}, v_0, v_1, a_0, b_0, \nu_0, \lambda_0, K, L$
 Compute the variable weight $\tilde{\pi}$ from (2.19);
for $k = 1 : K$ **do**
 Generate a subset of L variables according to $\tilde{\pi}$;
 Make the replicate $\tilde{\mathbf{X}}^k$ with the L variables;
 Initialize $\Theta_k^{(0)}$;
 Generate bootstrap weight $\mathbf{w}$ from (2.16);
 E-step: Calculate the two expectations in (2.8), denoted as $EE_k^{(0)}$;
 for $t = 1 : \text{maxIter}$ **do**
 M-step: Update $\Theta_k^{(t)}$ from Eq (2.10, 2.11, 2.12);
 E-step: Update $EE_k^{(t)}$ from Eq (2.17, 2.18);
 if $\gamma_k^{(t)}$ *stays the same for* $k_0 = 3$ *iterations* **then**
 break;
 end
 end
 Record $\gamma_k^{(t)}$;
end
 Return ϕ from Eq (2.20);

2.5 Empirical Results

In this section, we first compare the proposed EM algorithm (Algorithm 1) with other popular methods on a widely used benchmark data set. Then we compare BBEM (Algorithm 2) with other methods on two more challenging data sets of larger dimensions. Finally, we applied BBEM on a restaurant revenue data from a Kaggle competition, and showed that our algorithm outperforms the benchmark from random forest.

For the hyper-parameters v_0 and v_1 , we set $v_1 = 100$ as fixed and tune an appropriate value for v_0 either based on 5-fold cross-validation or BIC. For the initial value of θ , we suggest to use $1/2$ for ordinary problems, but $\sqrt{n}/p$ for large- p problems. Given $\theta^{(0)}$, the initial value of the binary vector $\gamma^{(0)}$ is randomly generated from Bernoulli distribution with parameter $\theta^{(0)}$. The initial value of σ^2 is set as 1. In addition, there are two bootstrap parameters: the total number of replicates K and the number of variables used in each bootstrap L . For efficiency, the number of variables in each bootstrap replicate should not exceed the sample size n . We use $K = 100$, and $L = n/2$ if p is large and $L = p$ if p is small.

2.5.1 A widely used benchmark

First we apply our EM algorithm on a widely used benchmark data set (Tibshirani, 1996), which has $p = 8$ variables, each from a standard normal distribution with pairwise correlation $\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$. The response variable is generated from

$$\mathbf{y} = 3\mathbf{x}_1 + 1.5\mathbf{x}_2 + 2\mathbf{x}_5 + \epsilon$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Following Fan and Li (2001), we repeat the experiment 100 times under two scenarios: (1) $n = 40, \sigma = 3$ and (2) $n = 60, \sigma = 1$. The result is shown in Table 2.1, which reports the average number of zero-coefficients (i.e., no selection) among signal variables ($\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_5$) and among noise variables, respectively. The results for SCAD1 (tuning parameter selected by cross-validation), SCAD2 (tuning parameter fixed) and LASSO are taken from (Fan and Li, 2001). In the first “small sample-size high noise” scenario, our EM algorithm has the highest number of zero-coefficients among noise variables, i.e., the lowest type I error. The average number of signal variables missed by EM is slightly higher than SCAD1 (where the tuning parameter is chosen by cross-validation) but less than SCAD2 (where the tuning parameter is pre-fixed). But overall, our EM algorithm and the two SCAD methods perform the best. In the second “large sample-size low noise” scenario, no signal variables are missed by any method, but EM has the lowest type I error.

Following (Wang et al., 2011) and (Xin and Zhu, 2012), we repeat the experiment 100 times with the same sample size $n = 50$ but two different noise levels: low noise level ($\sigma = 3$) and high noise level ($\sigma = 6$). Table 2.2 reports the minimum, median, maximum of being selected out of 100 simulations for the signal and the noise variables, respectively. Both Lasso and random Lasso have a higher chance of selecting the signal variables, but at the price of mistakenly including many noise variables. Overall, our EM algorithm performs the best, along with PGA and stability selection, two frequentist ensemble methods for variable selection.

Method	$\mathbf{x}_j \in \text{Noise}$ ($j=3,4,6,7,8$)	$\mathbf{x}_j \in \text{Signal}$ ($j=1,2,5$)
$n = 40, \sigma = 3$		
EM	4.55	0.24
SCAD1	4.20	0.21
SCAD2	4.31	0.27
LASSO	3.53	0.07
Oracle	5.00	0.00
$n = 60, \sigma = 1$		
EM	4.72	0.00
SCAD1	4.37	0.00
SCAD2	4.42	0.00
LASSO	3.56	0.00
Oracle	5.00	0.00

Table 2.1: A widely used benchmark. The average number of zero-coefficients (i.e., no selection) out of 100 simulations for each types of variable (Signal or Noise) are shown. The results other than EM (Alg 1) are from (Fan and Li, 2001).

2.5.2 A highly-correlated data

Next we demonstrate our two algorithms on a highly-correlated example from (Wang et al., 2011). The data has $p = 40$ variables and the response $\mathbf{y}$ is generated from

$$\mathbf{y} = 3\mathbf{x}_1 + 3\mathbf{x}_2 - 2\mathbf{x}_3 + 3\mathbf{x}_4 + 3\mathbf{x}_5 - 2\mathbf{x}_6 + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\sigma = 6$. Each $\mathbf{x}_i$ is generated from a standard normal with the following correlation structure among the first six signal variables: the signal variables are divided into two groups, $V_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3\}$ and $V_2 = \{\mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6\}$; the within group correlation is 0.9 and the between-group correlation is 0.

We repeat the simulation 100 times with $n = 50$ and $n = 100$, and the results are summarized in Table 2.3. For this example, due to the high correlation among features we expect ensemble methods to perform better. Indeed, BBEM has the best performance in terms of selecting true signal variables while controlling the error of including noise variables. The performance of EM algorithm, although not the best, is also comparable with other top ensemble methods like random Lasso from (Wang et al., 2011), and

Method	$\mathbf{x}_j \in \text{Signal } (j=1,2,5)$			$\mathbf{x}_j \in \text{Noise } (j=3,4,6,7,8)$		
	Min	Median	Max	Min	Median	Max
$n = 50, \sigma = 3$						
EM	91	97	100	3	6	12
Lasso	99	100	100	48	55	61
Random Lasso	95	99	100	33	40	48
ST2E	89	96	100	4	12	20
PGA	82	98	100	4	7	11
Stability selection						
$\lambda_{min} = 1$	81	83	100	0	2	9
$\lambda_{min} = 0.5$	90	98	100	4	8	22
$n = 50, \sigma = 6$						
EM	53	67	91	6	10	14
Lasso	76	85	99	47	49	53
Random Lasso	92	94	100	40	48	58
ST2E	68	69	96	9	13	21
PGA	54	76	94	9	14	16
Stability selection						
$\lambda_{min} = 1$	59	61	92	4	8	18
$\lambda_{min} = 0.5$	76	84	100	30	42	50

Table 2.2: A widely used benchmark. The min, median, max number of being selected out of 100 simulations for each types of variable (Signal or Noise) are shown. The results other than EM (Alg 1) are from (Xin and Zhu, 2012).

T2E and PGA from (Xin and Zhu, 2012).

For illustration purpose, we apply BBEM on a data set with $n = 50$ and v_0 varying from 10^{-4} to 1. Figure 2.1 shows the path-plot of the selection frequency from BBEM. There is clearly a gap between the signal variables and the noise ones. For a range of v_0 , from 0.001 to 0.02, BBEM can successfully select the six true variables $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_6\}$ if we threshold the selection frequency ϕ_j at 0.5.

Unlike the trumpet shape in LASSO, Figure 2.1 shows a hat shape for BBEM. When v_0 is extremely small, the “spike” component is close to point mass at 0. Intuitively, any nonzero coefficient should not be clustered to a point mass component. But in the update equation (2.6), we find that the posterior mean and variance of the coefficient β shrinkage to 0 very quickly for small v_0 . In the update equation (2.10) for γ_j , the second moment of β_j (left hand side) goes to zero faster than the threshold (right hand side). This

Method	$\mathbf{x}_j \in \text{Signal } (j= 1:6)$			$\mathbf{x}_j \in \text{Noise}$		
	Min	Median	Max	Min	Median	Max
$n = 50, \sigma = 6$						
Lasso	11	70	77	12	17	25
Random Lasso	84	96	97	11	21	30
ST2E	85	96	100	18	25	34
PGA	55	87	90	14	23	32
EM	65	85.5	89	4	10	13
BBEM	89	96	100	4	8	15
$n = 100, \sigma = 6$						
Lasso	8	84	88	12	22	31
Random Lasso	89	99	99	8	14	21
ST2E	93	100	100	14	21	27
PGA	40	85	92	13	22	33
EM	84	91	95	1	7	16
BBEM	95	99	100	4	9	14

Table 2.3: A highly-correlated data. The min, median, max number of times being selected (i.e., no selection) out of 100 simulations for each type of variables (Signal and Noise) are shown. The results other than EM and BBEM are from (Xin and Zhu, 2012).

explains why the BBEM algorithm labels every variable as “not included” when v_0 goes too small.

2.5.3 A Large- p small- n example

Finally we apply BBEM on a large- p small- n example from (Ročková and George, 2014), where $p = 1000$ and $n = 100$. Each of the p features is generated from a standard normal with pairwise correlation to be $0.6^{|i-j|}$ and the response $\mathbf{y}$ is generated from the following linear model:

$$\mathbf{y} = \mathbf{x}_1 + 2\mathbf{x}_2 + 3\mathbf{x}_3 + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 3)$.

For this large p example, we set the parameters in the BBEM algorithm as follows: the initial value of θ is $\sqrt{n}/p$, the number of variables used in each bootstrap iteration $L = n/2 = 50$ and the total number of replicates $K = 100$. It is shown that cross-validation based on prediction accuracy tends to include more noise variables (Wang et al., 2007). So, for this example where the true model is known to be sparse, we choose to tune v_0 via BIC. For

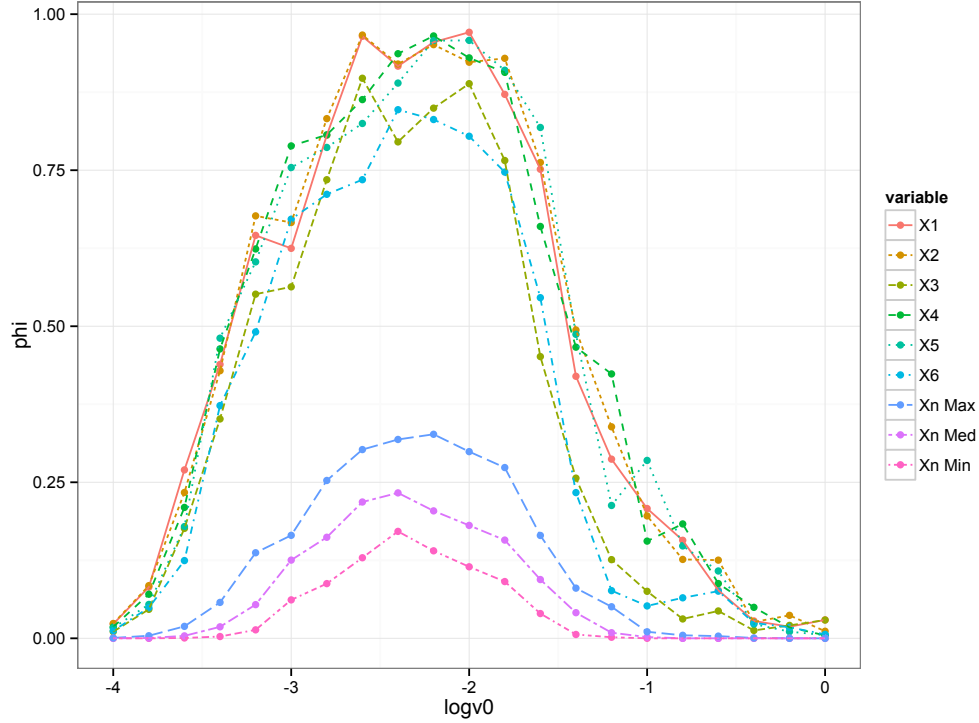


Figure 2.1: Highly-correlated data $n = 50$. A path-plot of the average selection frequency when v_0 varies in the logarithm scale of base 10. Top 6 lines represent the true variables $\mathbf{x}_{1:6}$ and the bottom 3 lines represent the maximum, median and minimum among the noise variables $\mathbf{x}_{7:40}$.

illustration purpose, we also include BBEM with a fixed tuning parameter $v_0 = 0.03$ in the comparison group. We compare BBEM with the EMVS algorithm from (Ročková and George, 2014), which is implemented by us using the annealing technique for β 's initialization, and fixed $v_0 = 0.5, v_1 = 1000$ as suggested in (Ročková and George, 2014).

Table 2.4 reports the average number of signal and noise variables being selected over 100 iterations for each method. BBEM with BIC tuning performs the best: it selects 2.99 signal variables out of 3 on average (i.e., only miss one variable, the weakest signal $\mathbf{x}_1$, once in all 100 iterations) and meanwhile has the smallest type I error. The BBEM algorithm with a fixed tuning parameter has a similar result as EMVS but is much faster. The computation advantage for BBEM comes from two aspects: the computation trick that reduces the computation cost on matrix inversion and the sub-sampling step in Bayesian bootstrap which allows us to deal with just a subset of variables of size smaller than p .

	$x_j \in \text{Signal}$	$x_j \in \text{Noise}$
BBEM (BIC)	2.99	0.24
BBEM ($v_0 = 0.03$)	2.96	0.27
EMVS	2.97	0.29
Oracle	3	0

Table 2.4: A large- p small- n example. The table shows the average number of signal and noise variables being selected out of 100 iterations. In BBEM, v_0 is either chosen by BIC or fixed at 0.03. EMVS is the algorithm proposed by Ročková and George (2014).

2.5.4 A real example

For TFI, a company that owns some of the world’s most well-known brands like Burger King and Arby’s, decisions on where to open new restaurants are crucial. It usually takes a big investment of both time and capital at the beginning to set up a new restaurant. If a wrong location is chosen, likely the restaurant will soon be closed and all the initial investment will be lost. TFI hosted a prediction competition on Kaggle¹, where the goal is to build a mathematical model to predict the revenue of a restaurant based on a set of demographic, real estate, and commercial information. The data contains 137 restaurants in the training set and 1000 restaurants in the test set. Features include the Open Date, City, City Group, Restaurant Type, and three categories of obfuscated data (P1-P37, numeric): demographic data, real estate data, and commercial data. The response is the transformed restaurant revenue in a given year.

We first transform the “Open Date” to a numeric feature called “Year Since 1900” and merge the “City” column into the “City Group” column which now contains four categories: Istanbul, Izmir, Ankara, and others (small cities). Then we create dummy variables for the categorical features like “City Group” and “Restaurant Type” and keep all the obfuscated numeric columns P1-P37. The final training set has 43 features and 137 samples.

After standardizing the data, we fix v_1 at 100 and tune v_0 from $10^{-4.5}$ to $10^{-0.5}$ for the BBEM algorithm, where each bootstrap sample uses $L = 15$ variables, and the total number of replicates is $K = 300$. The path-plot of selection frequency for important features is shown in Figure 2.2. It is not surprising that “City Group”, “Years Since 1900” and “Restaurant Type” are

¹<https://www.kaggle.com/c/restaurant-revenue-prediction>

some important predictors for the revenue. Quite a few obfuscated features are also selected as important predictors. Although we do not know their meanings, they should provide valuable information for TFI to choose their next restaurant’s location.

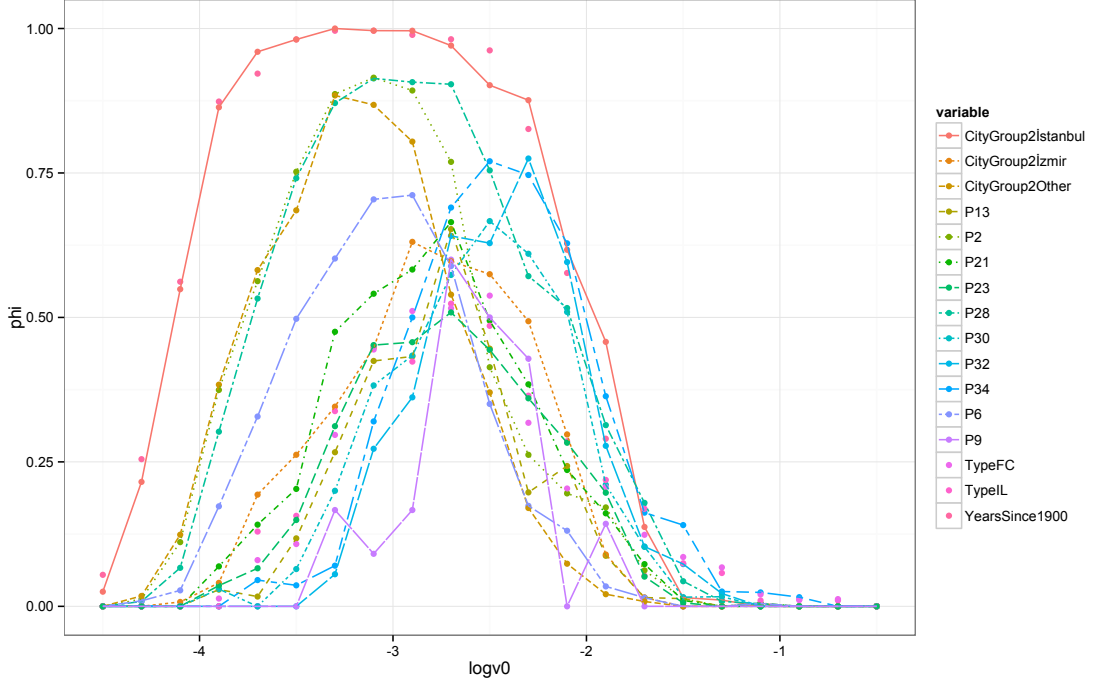


Figure 2.2: Restaurant data. The path plot of selection frequency when v_0 varies in the logarithm scale of base 10. Only a subset of variables with high selection frequencies are displayed.

Since the evaluation metric for this specific competition is based on the rooted mean square error (RMSE), we use the same metric in our 5-fold cross-validation. We tuned v_0 from the set $\{0.0001, 0.0002, 0.0005, 0.001, 0.002, 0.005, 0.01\}$, and found $v_0 = 0.002$ has the smallest RMSE score. Then we fix v_0 at 0.002, and re-run BBEM on the whole training data. Let $\mathbf{m}$ denote the averaged posterior mean of $\boldsymbol{\beta}$ from L bootstrap iterations, and $\boldsymbol{\gamma}$ the averaged selection frequency for p variables. We then use $\mathbf{m} * \boldsymbol{\gamma}$ (where $*$ denotes element-wise product) for prediction in the same spirit as the Bayesian model averaging. Our final Kaggle score is 1989762.52, which outperforms the random forest benchmark (RMSE=1998014.94) provided by Kaggle². It is remarkable for

²At Kaggle, each team can submit their prediction and see the corresponding performance on the test data many times, so one can easily obtain a good score by keep tweaking the model to overfit the test data. For this reason, we did not compare our result with

BBEM to outperform random forest considering that BBEM does not use any nonlinear features but random forest does.

2.6 Discussion

Variable selection is an important problem in modern statistics. In this chapter, we study the Bayesian approach to variable selection in the context of multiple linear regression. We proposed an EM algorithm that returns the MAP estimate of the set of relevant variables. The algorithm can be operated very efficiently and therefore can scale up with big data. In addition, we have shown that the MAP estimate from our algorithm provides a consistent estimator of the true variable set even when the model dimension diverges with the sample size. Further, we propose an ensemble version of our EM algorithm based on Bayesian bootstrap, which, as demonstrated via real and simulated examples, can substantially increase accuracy while maintaining the computation efficiency.

Although we restrict our discussion for the linear model, the two algorithm we proposed can be easily extended to other generalized linear models by using latent variables (Polson and Scott, 2013), an interesting topic for future research.

those “low” scores on the leaderboard provided by individual teams.

Chapter 3

Variational Bayes

3.1 Introduction

Consider the following canonical setup for a linear regression model: given a dependent variable Y and p independent predictor variables $X_1, \dots, X_p$, we model Y as

$$Y = X_1\beta_1 + X_2\beta_2 + \dots + X_p\beta_p + \epsilon \quad (3.1)$$

where the β_j 's are the unknown regression coefficients and the error term ϵ is often assumed to follow a normal distribution. A long-standing problem in regression analysis is to identify variables that are truly (if the assumed model is correct) or approximately relevant to the response Y , i.e., variables whose coefficients β_j 's are non-zero in equation (3.1).

Many variable selection methods are based on the framework of penalized likelihood. In the context of linear regression, the penalized likelihood usually takes the following form

$$\text{RSS}(\boldsymbol{\beta}) + \sum_{j=1}^p p_\lambda(\beta_j), \quad (3.2)$$

where RSS is the residual sum of squares and $p_\lambda(\cdot)$ is a penalty function that depends on a tuning parameter λ . With a proper choice of the penalty function p_λ , the resulting estimate $\hat{\boldsymbol{\beta}}$ that minimizes the objective function (3.2) will have some components to be exactly 0. That is, variables are automatically selected through the estimation of their regression coefficients. Examples of such a framework include classical model selection procedures, such as AIC and BIC, and modern procedures, such as Lasso (Tibshirani, 1996) and SCAD (Fan and Li, 2001).

There is a natural Bayesian interpretation of the penalized likelihood approach. The minimizer of the objective function (3.2) is the Maximum a

Posteriori (MAP) estimator with respect to a prior distribution $\pi(\beta_j) \propto \exp\{-p_\lambda(\beta_j)\}$. For example, Lasso, which uses L_1 penalty, corresponds to a Double Exponential prior on each β_j .

In this chapter, however, we specify priors not only on β_j 's but also on the model index $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$, a p -dimensional binary vector, where $\gamma_j = 1$ means the j -th variable is included, and 0, otherwise. It is appealing to obtain a posterior distribution over $\boldsymbol{\gamma}$, which will be useful not only for selecting a single model (i.e., a point estimate of $\boldsymbol{\gamma}$), but also for a variety of other inferences related to model uncertainty, such as model ranking, false discovery rate (FDR) control, and model averaging when the goal is prediction (Draper, 1995; Raftery et al., 1997).

In our hierarchical prior specification, we start with assigning prior probabilities $\pi(\boldsymbol{\gamma})$ to $\boldsymbol{\gamma}$ and then prior distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma})$ to the parameters given $\boldsymbol{\gamma}$. Popular choices of priors include the Bernoulli families for γ_j and the spike-and-slab prior (Mitchell and Beauchamp, 1988) for β_j , i.e.,

$$\pi(\beta_j | \gamma_j) = \gamma_j g(\beta_j) + (1 - \gamma_j) \delta_0(\beta_j) \quad (3.3)$$

where $g(\cdot)$ is usually the density function of a symmetric distribution like normal distributions with mean 0, and $\delta_0(\cdot)$ denotes the distribution function of a point mass at 0. The likelihood and priors induce a joint distribution over the data, the parameters, and the model space. Then, any inference on variable sets (or models) can be made based on the posterior distribution over $\boldsymbol{\gamma}$ given the observed data.

The posterior distribution on all the unknown parameters is usually not in closed form except for some special cases, for example, when the design matrix is orthogonal (Clyde et al., 1996). Instead, most Bayesian variable selection algorithms are implemented through Markov chain Monte Carlo (MCMC) algorithms (George and McCulloch, 1997; O'Hara and Sillanpää, 2009). The high computation cost of MCMC prohibits explorations over a large number of predictors. In contrast, many fast algorithms are available for variable selection in the penalized likelihood framework, such as the least-angle regression (Efron et al., 2004), Stagewise Lasso (Zhao and Yu, 2007), and the one-step local linear approximation (Zou and Li, 2008).

A breakthrough is made by (Ročková and George, 2014) recently for Bayesian variable selection. In their work, they adopt the continuous version of spike-

and-slab prior (George and McCulloch, 1993), in which the point mass $\delta_0(\cdot)$ in (3.3) is replaced by a normal density with a small variance. Replacing an indicator function by a continuous density turns out to be crucial for the development of a fast algorithm. Under such a prior specification, (Ročková and George, 2014) developed an efficient EM algorithm for the MAP estimate of $\boldsymbol{\beta}$. To obtain the inference on the variable set $\boldsymbol{\gamma}$, they propose a two-stage approach: first use the EM algorithm to obtain an estimate of $\boldsymbol{\beta}$, and then use a threshold to select the desired subset of variables.

In this chapter, we develop an efficient algorithm that returns the posterior distribution over the variable set $\boldsymbol{\gamma}$ directly, instead of a two-stage approach. As a trade-off between the computation cost and accuracy, we do not expect our new algorithm to return the exact posterior distribution like those traditional MCMC algorithms, but a good approximation of the posterior distribution. Here by “good”, we mean that what’s returned by our algorithm, albeit an approximation, should still have the desired asymptotic properties, such as asymptotic consistency.

The remaining of this chapter is arranged as follows. We close the introduction with a motivating example in mRNA-miRNA regulations. Section 2 gives necessary background and setup for Bayesian variable selection and introduces the SAB algorithm. Section 3 investigates the performance of SAB using simulated and real-world examples. An application of SAB to the mRNA-miRNA regulations is presented in Section 4. We conclude with a brief discussion in Section 5.

3.2 Method

3.2.1 Model and priors

Let $\mathbf{y} = (y_1, \dots, y_n)^T$ denote n observations for the response variable and assume the following normal linear regression model for $\mathbf{y}$,

$$\mathbf{y} \sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

where $\mathbf{N}_n$ denotes an n -variate normal distribution, $\mathbf{X}$ is an $n \times p$ design matrix, and $\boldsymbol{\beta}$ and σ^2 are the $p \times 1$ unknown regression coefficients and error variance, respectively. Without loss of generality, assume the data vector $\mathbf{y}$

and each column of $\mathbf{X}$ have mean 0 and sample variance 1; so we can drop the intercept in the regression model.

The hierarchical priors are specified as follows:

$$\beta_j \mid \sigma^2, \gamma_j \begin{cases} \sim \mathbf{N}(0, v_1 \sigma^2), & \text{if } \gamma_j = 1, \\ \sim \mathbf{N}(0, v_0 \sigma^2), & \text{if } \gamma_j = 0, \end{cases} \quad (3.4)$$

$$\begin{aligned} \gamma_j \mid \theta &\sim \mathbf{Bern}(\theta), & j = 1, 2, \dots, p, \\ \theta &\sim \mathbf{Beta}(a_0, b_0), \\ \sigma^2 &\sim \mathbf{IG}(\nu_0/2, \nu_0 \lambda_0/2), \end{aligned}$$

where $\mathbf{Bern}(\theta)$ is a Bernoulli distribution with mean θ , $\mathbf{Beta}(a, b)$ is a beta distribution with mean $a/(a+b)$, and $\mathbf{IG}(a, b)$ represents an inverse gamma distribution with mean $b/(a-1)$. In practice, we recommend the following values for these hyper-parameters $a_0 = b_0 = 1$, and $\nu_0 = \lambda_0 = 1$.

Prior (3.4) on β_j is the same as the one used in (Ročková and George, 2014). The two parameters, v_1 and v_0 , describe the variability in β_j (relative to the error variance σ^2) when variable j is significant and not significant, respectively. Therefore we restrict $0 < v_0 < v_1$. The normal component $\mathbf{N}(0, v_0 \sigma^2)$, corresponding to $\gamma_j = 0$, can be viewed as a continuous relaxation of the point mass at zero used in the original spike-and-slab prior. The motivation behind this continuous relaxation is twofold: computation and applications. The former will be made clear in the upcoming discussion. For the latter, in many applications, the dependence among the predictors X_j 's and the dependence between X_j 's and the response variable Y are complicated and usually nonlinear. If we approximate the dependence between X_j 's and Y by a linear function, then most likely all the coefficients are non-zero. So the goal of variable selection is to differentiate variables that have a strong effect (i.e., $|\beta_j|$ is large) from variables have a weak effect (i.e., their $|\beta_j|$ is small), which justifies the use of two normal components with non-zero variances, instead requiring the non-significant variables have an exactly zero coefficient.

3.2.2 Posterior inference

When the primary goal is variable selection, we can integrate over β and obtain the following marginal likelihood for each model:

$$\begin{aligned} p(\mathbf{y} \mid \gamma, \sigma^2) &= \int p(\mathbf{y} \mid \gamma, \beta, \sigma^2) \pi(\beta \mid \gamma) d\beta \\ &= \frac{1}{(2\pi)^{n/2}} |\mathbf{X}D_\gamma\mathbf{X}^T + \mathbf{I}|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2} \mathbf{y}^T (\mathbf{X}D_\gamma\mathbf{X}^T + \mathbf{I})^{-1} \mathbf{y} \right\}, \end{aligned}$$

where D_γ is a diagonal matrix given by

$$D_\gamma = \text{diag} \left\{ (1 - \gamma_j)v_0 + \gamma_j v_1 \right\}_{j=1}^p.$$

Then the joint posterior distribution over $(\gamma, \sigma^2, \theta)$ is then given by

$$\pi(\gamma, \sigma^2, \theta \mid \mathbf{y}) \propto p(\mathbf{y} \mid \gamma, \sigma^2) \pi(\gamma \mid \theta) \pi(\theta) \pi(\sigma^2).$$

Although we can evaluate the posterior distribution for any fixed value of $(\gamma, \sigma^2, \theta)$, any inference on the full posterior would need to rely on computational algorithms. MCMC is one such algorithm although it is time-consuming and difficult to scale to large data.

For variable selection, there is no need to obtain the joint posterior on all the unknown parameters. What is relevant is the posterior distribution of the variable subsets γ 's. To reduce the computation cost, we consider plugging in point estimates of the nuisance parameters (θ, σ^2) , for example, the MAP estimates:

$$(\hat{\theta}, \hat{\sigma}^2) = \arg \max \pi(\theta, \sigma^2 \mid \mathbf{y}). \quad (3.5)$$

Then inference on variable selection can be made based on the following conditional posterior probability on each model γ :

$$\pi(\gamma \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2) = p(\mathbf{y} \mid \gamma, \hat{\sigma}^2) \prod_{j=1}^p \hat{\theta}^{\gamma_j} (1 - \hat{\theta})^{(1-\gamma_j)}. \quad (3.6)$$

For example, the corresponding MAP estimate of γ is given by

$$\hat{\gamma}^{\text{MAP}} = \arg \max_{\gamma} \pi(\gamma \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2), \quad (3.7)$$

and the Barbieri-Berger (BB) median model (Barbieri and Berger, 2004) is

given by

$$\hat{\gamma}_j^{\text{BB}} = 1, \text{ if } \pi(\gamma_j = 1 \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2) > 0.5. \quad (3.8)$$

When the dimension p is high, however, it is not computationally feasible to solve (3.7) nor (3.8) by visiting all 2^p models. This is further complicated by the need to obtain $\hat{\theta}$ and $\hat{\sigma}^2$, the maximum a posteriori (MAP) estimates of θ and σ^2 .

We propose to take a variational approach to solve for $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2)$ instead. We first establish a connection between the maximizer of an objective function and $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2)$ in Proposition 3.1 (Proof is given in Web Appendix A), and then introduce the SAB algorithm in Section 2.3 to solve for the maximizer. To start, let $q(\boldsymbol{\gamma}, \boldsymbol{\beta})$ denote any joint distribution over $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Consider the following optimization problem in which we want to find the optimal distribution $q(\cdot, \cdot)$ and parameters (θ, σ^2) that maximize the following objective function:

$$J_0(q, \theta, \sigma^2) = \int q(\boldsymbol{\gamma}, \boldsymbol{\beta}) \log \frac{\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}, \theta, \sigma^2 \mid \mathbf{y})}{q(\boldsymbol{\gamma}, \boldsymbol{\beta})} d\boldsymbol{\gamma} d\boldsymbol{\beta}. \quad (3.9)$$

Proposition 3.1. *Suppose $q^*(\boldsymbol{\gamma}, \boldsymbol{\beta})$ is a maximizer of (3.9). Then the corresponding marginal distribution over $\boldsymbol{\gamma}$, i.e., $q^*(\boldsymbol{\gamma}) = \int q^*(\boldsymbol{\gamma}, \boldsymbol{\beta}) d\boldsymbol{\beta}$ is the same as the posterior distribution $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2)$ defined in (3.6), where $\hat{\theta}$ and $\hat{\sigma}^2$ are MAP estimates defined in (3.5).*

3.2.3 A hybrid algorithm

Proposition 3.1 implies that it suffices to solve (3.9), in order to estimate the posterior probability $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2)$. In practice, to enable the tractability of this optimization problem, we further approximate the joint posterior by a product form:

$$q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = q_1(\boldsymbol{\beta})q_2(\boldsymbol{\gamma}). \quad (3.10)$$

The objective function (3.9) becomes

$$J(q_1, q_2, \theta, \sigma^2) = \mathbb{E}_{\boldsymbol{\beta}}^{q_1} \mathbb{E}_{\boldsymbol{\gamma}}^{q_2} \log \frac{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2) \pi(\boldsymbol{\gamma} \mid \theta) \pi(\theta) \pi(\sigma^2)}{q_1(\boldsymbol{\beta})q_2(\boldsymbol{\gamma})}, \quad (3.11)$$

where $\mathbb{E}_W^q$ denotes taking expectation of a random variable W with respect to distribution $q(w)$. Then we optimize $(q_1, q_2, \theta, \sigma^2)$ iteratively, until con-

vergence.

Although the algorithm operates with the factorization restriction (3.10), the two distributions $q_1(\boldsymbol{\beta})$ and $q_2(\boldsymbol{\gamma})$ are not independent of each other, instead they are iteratively estimated such that their product can approximate the joint posterior $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2)$. As shown in the updating steps given below, the parameters for $q_1(\boldsymbol{\beta})$ and the one for $q_2(\boldsymbol{\gamma})$ are clearly not estimated independently.

We update the approximate posterior distributions $q_1(\boldsymbol{\beta})$ and $q_2(\boldsymbol{\gamma})$ based on the following Lemma. Its proof follows as a known result for variational methods, e.g., in (Bishop, 2006).

Lemma 3.1. *Let $p(\mathbf{y}, \Theta) = p(\mathbf{y}|\Theta)\pi(\Theta)$ be the joint distribution function over the data $\mathbf{y}$ and the parameter $\Theta = (\theta_1, \dots, \theta_m)$. For any parameter θ_j , the optimal distribution function $q_j(\theta_j)$ that maximizes*

$$\mathbb{E}_{\theta_1}^{q_1} \dots \mathbb{E}_{\theta_m}^{q_m} \frac{p(\mathbf{y}, \Theta)}{q_1(\theta_1) \dots q_m(\theta_m)},$$

is given by

$$q_j(\theta_j) = \frac{\exp \{ \mathbb{E}_{[-j]} \log p(\mathbf{y}, \Theta) \}}{\int \exp \{ \mathbb{E}_{[-j]} \log p(\mathbf{y}, \Theta) \} d\theta_j}$$

where $\mathbb{E}_{[-j]}$ means taking expectation with respect to $\prod_{l \neq j} q_l(\theta_l)$, which are given, provided that the denominator above is finite.

[Update q_1] Applying Lemma 3.1 we find the optimal $q_1(\boldsymbol{\beta})$ that maximizes (3.11) given $(q_2, \hat{\theta}, \hat{\sigma}^2)$ as

$$\begin{aligned} \log q_1(\boldsymbol{\beta}) - \text{const} &= \mathbb{E}_{\boldsymbol{\gamma}}^{q_2} [\log p(\mathbf{y} \mid \boldsymbol{\beta}, \hat{\sigma}^2) \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \hat{\sigma}^2)] \\ &= -\frac{1}{2\hat{\sigma}^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \frac{1}{2\hat{\sigma}^2} \boldsymbol{\beta}^t \boldsymbol{\Delta} \boldsymbol{\beta} \\ &= -\frac{1}{2\hat{\sigma}^2} [\boldsymbol{\beta}^T (\mathbf{X}^T \mathbf{X} + \boldsymbol{\Delta}) \boldsymbol{\beta} - 2\mathbf{y}^T \mathbf{X} \boldsymbol{\beta}]. \end{aligned} \quad (3.12)$$

Here $\boldsymbol{\Delta} = \mathbb{E}_{\boldsymbol{\gamma}}^{q_2} D_{\boldsymbol{\gamma}}^{-1}$ is a diagonal matrix with entry

$$\Delta_j = \frac{1 - \phi_j}{v_0} + \frac{\phi_j}{v_1},$$

where $\phi_j = \mathbb{P}(\gamma_j = 1)$ is the marginal probability of $\gamma_j = 1$ with respect to the distribution $q_2(\boldsymbol{\gamma})$. The quadratic form in (3.12) suggests that $q_1(\boldsymbol{\beta})$

belongs to a normal family $\mathbf{N}_p(\mathbf{m}, \mathbf{D})$ and the mean and variance are given by

$$\mathbf{m} = (\mathbf{X}^T \mathbf{X} + \mathbf{\Delta})^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D} = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X} + \mathbf{\Delta})^{-1}. \quad (3.13)$$

[Update q_2] Following the same updating rule in Lemma 3.1, the optimal $q_2(\boldsymbol{\gamma})$ given $(q_1, \hat{\theta}, \hat{\sigma}^2)$ has the following explicit form:

$$\begin{aligned} \log q_2(\boldsymbol{\gamma}) - \text{const} &= \mathbb{E}_{\boldsymbol{\beta}}^{q_1} \log \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \hat{\sigma}^2) \pi(\boldsymbol{\gamma} \mid \hat{\theta}) \\ &= \sum_{j=1}^p \left\{ \gamma_j \left[-\frac{1}{2} \log v_1 - \frac{1}{2\hat{\sigma}^2} \frac{\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \beta_j^2}{v_1} + \log \hat{\theta} \right] \right. \\ &\quad \left. + (1 - \gamma_j) \left[-\frac{1}{2} \log v_0 - \frac{1}{2\hat{\sigma}^2} \frac{\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \beta_j^2}{v_0} + \log(1 - \hat{\theta}) \right] \right\}. \end{aligned}$$

The equation above implies that $q_2(\boldsymbol{\gamma})$ is a product of independent $\text{Bern}(\phi_j)$ and the Bernoulli parameter ϕ_j satisfies

$$\log \frac{\phi_j}{1 - \phi_j} = \frac{1}{2} \log \frac{v_0}{v_1} + \frac{m_j^2 + D_{jj}}{2\hat{\sigma}^2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) + \log \frac{\hat{\theta}}{1 - \hat{\theta}}, \quad (3.14)$$

where m_j is the j -th component of vector $\mathbf{m}$ and D_{jj} is the j -th diagonal value of matrix $\mathbf{D}$. Although in (3.10) we do not assume $q_2(\boldsymbol{\gamma})$ can be factorized into independent distributions over γ_j , it turns out that the optimal approximation is achieved by independent Bernoulli distributions over each γ_j .

[Update θ and σ^2] A general updating rule for the expectation-maximization (EM) algorithm (Dempster et al., 1977) can be applied for the update of the the MAP estimates of θ and σ . Specifically,

$$\begin{aligned} \hat{\theta} &= \arg \max_{\theta} \left[\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \mathbb{E}_{\boldsymbol{\gamma}}^{q_2} \log \pi(\boldsymbol{\gamma} \mid \theta) \pi(\theta \mid a_0, b_0) \right] \\ &= \frac{a_0 + \sum_{j=1}^p \phi_j}{p + a_0 + b_0}, \end{aligned} \quad (3.15)$$

and

$$\begin{aligned}\hat{\sigma}^2 &= \arg \max_{\sigma^2} \left[\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \mathbb{E}_{\boldsymbol{\gamma}}^{q_2} \log p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2) \pi(\sigma^2 \mid \nu_0, \lambda_0) \right] \\ &= \frac{\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \Delta_j (m_j^2 + D_{jj}) + \nu_0 \lambda_0}{n + p + v_0},\end{aligned}\quad (3.16)$$

where

$$\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{y} - \mathbf{X}\mathbf{m}\|^2 + \text{tr}(\mathbf{X}^T \mathbf{D} \mathbf{X}).$$

The resulting algorithm is a *hybrid* of variational Bayesian (VB) algorithm, using which we update q_1 and q_2 , and an EM algorithm, using which we update θ and σ^2 . We call our algorithm, Scalable Approximate Bayesian (SAB) algorithm for variable selection. The entire SAB algorithm is given in Algorithm 3 as psuedo code. Some details on the implementation of SAB, such as the stopping criteria, choice for v_0 and v_1 , and fast computation tricks are given in Wed Appendix B.

Algorithm 3: SAB for Variable Selection.

Input: $\mathbf{X}, \mathbf{y}, v_0, v_1, a_0, b_0, \nu_0, \lambda_0$
Initialization $\boldsymbol{\phi}^{(0)}, \mathbf{m}_0, D_0, \theta_0, \sigma_0$;
for $t = 1 : \text{maxIter}$ **do**
 Update the approximate posterior parameters $\mathbf{m}^{(t)}, D^{(t)}, \boldsymbol{\phi}^{(t)}$ from equations (3.13) and (3.14);
 Update the MAP estimator $\hat{\theta}^{(t)}, \hat{\sigma}^{(t)}$ from equations (3.15) and (3.16);
 if *The stopping criterion (Appendix B.2) is satisfied* **then**
 | break;
 end
end

Remark 1. The major computation cost in SAB lies in the inversion of the $p \times p$ matrix $(\mathbf{X}^T \mathbf{X} + \boldsymbol{\Delta})^{-1}$ in (3.13). The computation complexity is of order $O(\min(n, p)^3)$ for each iteration. When $p > n$ we can use the Sherman-Morrison-Woodbury formula to compute the inverse of an $n \times n$ matrix instead. Note that the only thing in $(\mathbf{X}^T \mathbf{X} + \boldsymbol{\Delta})$ that changes from iteration to iteration is $\boldsymbol{\Delta}$. From our experience, the change of some ϕ_j 's is negligible after a couple of iterations. In order to scale up the computation, we update only those ϕ_j 's whose change is greater than a threshold (e.g., 0.01). With this modification, the objective function is not fully optimized

but is still decreasing monotonically in each iteration. The advantage of this sub-optimal updating rule is that we only need to invert an $l \times l$ matrix with a computation cost of order $O(l^3)$, where l is the number of ϕ_j 's that are updated in that iteration, a number usually much smaller than $\min(n, p)$ after a couple of initial iterations (see Web Appendix B for details).

Remark 2. SAB shares some similarities with the EM algorithm proposed by (Ročková and George, 2014). They are, however, different in the following aspects. First, the EM algorithm returns a point estimate of β . Therefore, in each iteration, only the first moment of β is used in the EM algorithm. In contrast, SAB also utilizes the information from the second moment: updating ϕ_j in (3.14) uses both the posterior mean $\mathbf{m}$ and the posterior covariance matrix $\mathbf{D}$. Second, the underlying objective functions are different. While SAB maximizes (3.11), the EM algorithm maximizes

$$G(\beta, q_2, \theta, \sigma^2) = \mathbb{E}_{\gamma}^{q_2} \log \frac{p(\mathbf{y} \mid \beta, \sigma^2) \pi(\beta \mid \gamma, \sigma^2) \pi(\gamma \mid \theta) \pi(\theta) \pi(\sigma^2)}{q_2(\gamma)},$$

which returns points estimates $(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$ with the optimal $q_2(\gamma) = \pi(\gamma \mid \hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$ being the posterior distribution over the γ conditioning on $(\hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$. Last, for variable selection, the EM algorithm is based on $\pi(\gamma \mid \hat{\beta}, \hat{\theta}, \hat{\sigma}^2)$, which is essential a two-stage approach: obtain the MAP estimate $\hat{\beta}$ via EM and then test $\gamma_j = 1$ or 0 based on $\hat{\beta}_j$. In contrast, SAB simultaneously estimates the distribution of β and γ_j 's. The updating scheme can be justified as an iterative optimization procedure.

3.3 Asymptotic Analysis

In this Section, we investigate the asymptotic property of SAB regarding variable selection. Assume $\mathbf{y}_n$ are generated from the following linear regression model

$$\mathbf{y}_n = \mathbf{X}_n \beta_n^* + \mathbf{e}_n, \quad (3.17)$$

where $\mathbf{e}_n$ is an n -dimensional random vector with each element independently distributed as $\mathbf{N}(0, \sigma^2)$, and $\beta_n^* = (\beta_{n1}^*, \dots, \beta_{np}^*)^T$ denotes the true regression coefficients. Define the true model index $\gamma^* = (\gamma_1^*, \dots, \gamma_p^*)^T$ as $\gamma_j^* = 0$, if $\beta_j^* = 0$, and 1, otherwise. Throughout, the number of parameters $p = p_n$ is

allowed to diverge to infinity when $n \rightarrow \infty$.

We examine two types of variable selection consistency: the usual definition of consistency that concerns whether the selected variables, as a set, form a consistent estimator of the true variable set $S_n^* = \{j : \gamma_{nj}^* = 1\}$, and the Bayesian consistency that concerns whether the posterior distribution on γ converges to a point mass at the true model γ^* . The main results are stated below with details on the conditions and proofs in the Web Appendix C.

Estimation consistency

Let $\hat{S}_n$ denotes the estimate of S_n^* based on SAB:

$$\hat{S}_n = \{j : \phi_{nj} \geq 0.5, j = 1, \dots, n.\}$$

where ϕ_{nj} is the inclusion probability returned by SAB. Here we set the cutoff value for ϕ_{nj} to be 0.5, that is, $\hat{S}_n$ contains variables selected by the Barbieri-Berger (BB) model (Barbieri and Berger, 2004). As shown in our proof, the consistency result holds true with any other cutoff values between 0 and 1.

It can be shown that ϕ_{nj} 's returned by SAB satisfy: $\min_{j \in S_n^*} \phi_{nj} \xrightarrow{P} 1$ and $\max_{j \notin S_n^*} \phi_{nj} \xrightarrow{P} 0$, where $\xrightarrow{P}$ denotes convergence in probability, and the probability measure corresponds to the true sampling distribution (3.17). Therefore $P(\hat{S}_n = S_n^*) \rightarrow 1$, that is, SAB achieves estimation consistency for variable selection.

Theorem 3.1. *Assume conditions (C1-C4), and $p = O(n^\alpha)$ where $0 \leq \alpha < 1$. With v_1 fixed at constant and $v_0 = O(n^{-r_0})$ with $1 - \eta_3 < r_0 < \min\{\eta_1 - \alpha, 2(\eta_1 - \eta_2)/3\}$, we have*

$$\mathbb{P}(\hat{S}_n = S_n^*) = 1 - o(e^{-Mn^{\eta_1 - r_0 - \alpha}}) \rightarrow 1. \quad (3.18)$$

Bayesian consistency

A Bayesian variable selection procedure is said to be consistent, if the posterior distribution over γ converges to a point mass at the true model index γ^* (Liang et al., 2008; Fernández et al., 2001). Recall that SAB approximates the posterior distribution over γ by a product of independent $\text{Bern}(\phi_{nj})$,

namely,

$$q_2(\boldsymbol{\gamma}) = \prod_{j=1}^p \phi_{nj}^{\gamma_j} \times (1 - \phi_{nj})^{(1-\gamma_j)}.$$

So SAB achieves *Bayesian consistency* for variable selection if

$$q_2(\boldsymbol{\gamma}^*) = \left[\prod_{j:\gamma_j^*=1} \phi_{nj} \right] \left[\prod_{j:\gamma_j^*=0} (1 - \phi_{nj}) \right] \xrightarrow{P} 1.$$

or equivalently $1 - q_2(\boldsymbol{\gamma}^*) \xrightarrow{P} 0$.

Note that the point estimator $\hat{\boldsymbol{\gamma}}_n$, defined as $\hat{\gamma}_{nj} = 1$ if $\phi_{nj} \geq 0.5$ and 0 otherwise, is also the MAP estimator under q_2 . Theorem 3.1 has shown that asymptotically the true model $\boldsymbol{\gamma}^*$ receives the highest posterior probability, but Bayesian consistency requires a stronger result: the true model $\boldsymbol{\gamma}^*$ receives not the highest, but all posterior probability mass. Next we show that SAB can achieve Bayesian consistency if v_1 grows to infinity at a certain rate. A similar condition on v_1 arises in the asymptotic analysis in (Narisetty and He, 2014) where they study the Bayesian consistency based on the true posterior (i.e., MCMC is required).

Theorem 3.2. *With the same conditions for Theorem 3.1 and*

$$v_1 = O(n^{r_1}), \quad \text{with } r_1 > 0, \quad r_0 + r_1 > 2\alpha.$$

Then $q_2(\boldsymbol{\gamma}_n^) \xrightarrow{P} 1$.*

3.4 Empirical Results

3.4.1 Simulation studies

In all the simulation studies, we center and scale the response vector $\mathbf{y}$ and each column of the design matrix $\mathbf{X}$ to have sample mean 0 and sample variance 1. We recommend to set $v_1 = 1$ and $v_0 = 1/\sqrt{10np}$, and initialize ϕ_j 's as 0.5.

[Study 1] A widely used benchmark ($n = 40$, $p = 8$, $p_0 = 3$).

This data set was first introduced by (Tibshirani, 1996), and has been used as a benchmark data set in the variable selection literature. There are $p = 8$

variables, $X_1, \dots, X_8$, each of which is from a standard normal with pairwise correlation $\rho(X_i, X_j) = 0.5^{|i-j|}$. The response variable Y is generated as

$$Y = 3X_1 + 1.5X_2 + 2X_5 + \epsilon$$

where ϵ 's are iid samples from $\mathbf{N}(0, \sigma^2)$.

We first apply SAB on data sets generated from this model with $n = 40$ and $\sigma = 3$. In order to study how the tuning parameter v_0 affects our result, we vary v_0 from 10^{-4} to 10^0 . The corresponding path for the inclusion probability ϕ_j and the path for the mean coefficient m_j (averaged over 20 iterations) are displayed in Figure 3.1. The path for m_j is similar to the one from LASSO. But unlike the trumpet shape for m_j , the path for ϕ_j has the shape of a tilted hat. When v_0 is very small or equivalently $1/v_0$ is very large, the normal component $\mathbf{N}(0, v_0\sigma^2)$ shrinks the posterior mean and variance of each coefficient β_j to almost zero, consequently ϕ_j goes to 0, and therefore no variable is selected. This is the same as LASSO with a very large penalty parameter. When v_0 gets larger, important variables are selected one-by-one. But when v_0 gets close to 1, the two normal components are no longer distinguishable, so all ϕ_j 's converge to a constant. In this example, for a wide range of v_0 , SAB can successfully select the three true variables, whose paths are displayed in solid lines in Figure 3.1.

Next we fix $v_0 = 1/\sqrt{10np} = 0.0177$ as the recommended value for SAB and compare its performance with other methods regarding variable selection accuracy. Like in (Fan and Li, 2001), we consider two cases: a challenge case with a smaller sample size ($n = 40$) and a large error variance ($\sigma^2 = 3^2$), and an easy case with a larger sample size ($n = 60$) and a smaller error variance ($\sigma^2 = 1$). The results of 100 simulations are summarized in Table 3.1, where results for methods other than SAB are taken from (Fan and Li, 2001). The 1st column of Table 3.1 shows the average number of 0 coefficients among noise variables: the higher, the better, with the ideal value 5. Compared with SCAD1 (tuning parameter selected by cross-validation), SCAD2 (tuning parameter fixed) and LASSO, SAB has the highest number in the 1st column, that is, achieving the best performance in excluding noise variables. The 2nd column of Table 3.1 shows the average number of 0 coefficients among correct, signal variables: the smaller, the better, with the ideal value 0. When the signal-to-noise ratio is high (i.e., the 2nd case), both

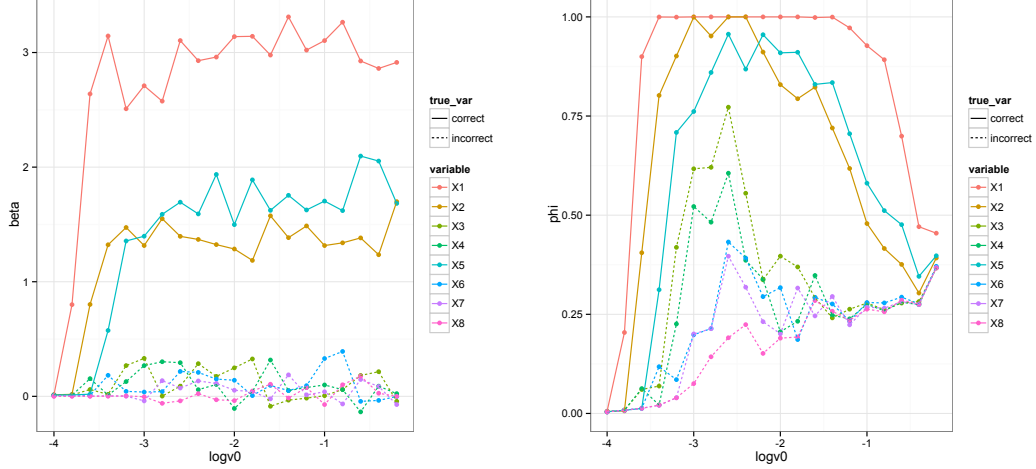


Figure 3.1: **Left:** The SAB path (averaged over 20 data sets) for the posterior mean of β_j for data from Simulation Study 1 with $n = 40$, $p = 8$, and $\sigma = 3$. **Right:** The averaged SAB path for the inclusion probability ϕ_j . In both plots, the paths of the 3 relevant variables, X_1 , X_2 and X_5 , are shown as solid lines and the paths for the other variables are shown as dashed lines. The x -coordinate denotes v_0 changed in the logarithm scale of base 10.

SAB and the other methods do well. When the signal-to-noise ratio is low (i.e., the first case), the average number of correct variables missed by SAB is less than SCAD but higher than LASSO. This is because LASSO uses cross-validation to tune the regularization parameter and it is less likely to miss important signal variables. But due to the same reason, LASSO tends to include more noise variables—it has the smallest value for the 1st column.

[Study 2] A large p small n example ($n = 100$, $p = 1000$, $p_0 = 3$)

We consider the following large p small n example from (Ročková and George, 2014). There is a total of $p = 1000$ variables and each is generated from a standard normal with pairwise correlation $\rho(X_i, X_j) = 0.6^{|i-j|}$. The response variable Y is generated according to a linear model involving only three variables,

$$Y = X_1 + 2X_2 + 3X_3 + \epsilon,$$

where ϵ 's are iid samples from $\mathbf{N}(0, 3)$.

We set $v_0 = 0.001$ and $v_1 = 1$ at the recommended values and compare SAB with EMVS from (Ročková and George, 2014) over 30 iterations. The result on variable selection is summarized Table 3.2. Regarding the accu-

Method	$\mathbf{x}_j \in \text{Noise}$ (j=3,4,6,7,8)	$\mathbf{x}_j \in \text{Signal}$ (j=1,2,5)
$n = 40, \sigma = 3$		
SAB	4.43	0.13
SCAD1	4.20	0.21
SCAD2	4.31	0.27
LASSO	3.53	0.07
$n = 60, \sigma = 1$		
SAB	4.9	0.00
SCAD1	4.37	0.00
SCAD2	4.42	0.00
LASSO	3.56	0.00

Table 3.1: Study 1. The average number of 0 coefficients (i.e., no selection) out of 100 simulations for each type of variables (Signal or Noise). The ideal value is 5 for the 1st column and 0 for the 2nd column. The results of SCAD1, SCAD2 and LASSO are taken from Fan and Li (2001).

racy for variable selection, SAB and EMVS perform similarly. However, the computation time for SAB is much less.

	$x_j \in \text{Signal}$	$x_j \in \text{Noise}$	Computation Time (s)
SAB	28.3	0.003	83.4
EMVS	28.0	0.002	278.7
Oracle	30.0	0.000	—

Table 3.2: Study 2. The frequencies of being selected (out of 30 simulations) for the two types of variables (Signal or Noise), where EMVS is from (Ročková and George, 2014) and Oracle corresponds to the ideal case where we know the true relevant variables. The last column records the computation time on a Macbook Air (1.7 GHz Intel Core i5 and 4 GB 1333 MHz DDR3 memory).

[Study 3] The Zhao and Yu’s example ($n = 1000, p = 3, p_0 = 2$)

This is the example from (Zhao and Yu, 2006) where the sample size n is large and the number of features p is relatively small but LASSO fails due to the dependence among the predictors. Following (Zhao and Yu, 2006), the three variables (X_1, X_2, X_3) are generated as follows: X_1, X_2 are independent $\mathbf{N}(0, 1)$ random variables, and

$$X_3 \mid X_1, X_2 \sim \mathbf{N}\left(\frac{2}{3}X_1 + \frac{2}{3}X_2, \frac{1}{3^2}\right).$$

The response variable Y is generated using

$$Y = 2X_1 + 3X_2 + \epsilon$$

where ϵ 's are iid samples from $\mathbf{N}(0, 1)$.

In this example, the true variables are X_1 and X_2 . However, as shown in (Zhao and Yu, 2006), no matter how one sets the regularization parameter, LASSO either misses some true predictors or includes the noise predictor X_3 , due to the correlation between X_3 and the true predictors.

We apply SAB on this data set. We fix $v_1 = 1$ and vary v_0 to produce the SAB path for the inclusion probability ϕ_j 's. As shown in Figure 3.2, with a wide range of v_0 , SAB can identify the true variable set which contains just X_1 and X_2 .

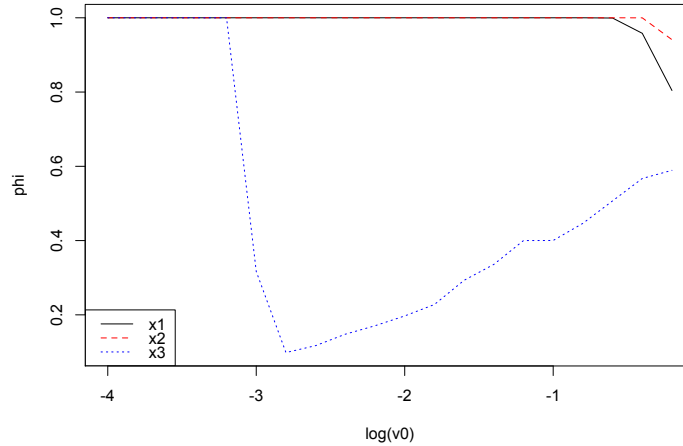


Figure 3.2: Study 3. The SAB path for the inclusion probability ϕ_j on the Zhao and Yu's example, where the x -coordinate denotes v_0 changed in the logarithm scale of base 10. In this example, the true variables are X_1 and X_2 .

3.4.2 Cancer genomics data on miRNA-mRNA regulation

The Cancer Genome Atlas (TCGA, <https://tcga-data.nci.nih.gov/tcga>) is supported by the National Cancer Institute and the National Human Genome Research Institute to delineate the molecular landscape of tumor

samples for more than 20 types of cancer (The Cancer Genome Atlas Network, 2012, 2008). TCGA has been generating multi-layer genomics, epigenomics, and proteomics data for thousands of cancer patients. There are different levels of data access for TCGA; while the access to most of the level-1 and -2 data is restricted to protect patient privacy, the entire level-3 TCGA data as well as some patient clinical information (e.g., survival and drug treatments) are publicly available. Level-3 TCGA data are normalized measurements of genomics and epigenomics features, including mRNA gene expression and miRNA expression. An miRNA (called “micro-RNA”) is a small non-coding RNA of about 22 base pairs long. The main function of miRNAs is to regulate the expression level of mRNAs (called “messenger RNAs”), which are protein coding RNAs. The abundance level of mRNA is important since it determines how many protein molecules are produced in the human body. There are tens of thousands of different mRNAs and hundreds of miRNAs in human cells. It is believed that multiple miRNAs can bind to multiple mRNAs to either degrade the mRNAs into “junks” or to block the translation from mRNAs to proteins, thus regulating the expression mRNAs and proteins (Bartel, 2004, 2009) The exact mechanism of regulation is still an open research topic. In this chapter, we consider one open question: *which and how many miRNAs regulate which and how many mRNAs?* We plan to treat expression of each mRNA as a response variable Y and miRNAs as predictors X , and analyze TCGA data retrieved from our own software, TCGA-Assembler (Zhu et al., 2014) as a first analysis of the most up-to-date cancer genomics data.

We plan to combine existing biological knowledge in our investigation to improve the power of analysis. In particular, we will prescreen the miRNAs and select the ones that potentially regulate the target gene using existing databases of miRNA-mRNA regulation. These databases are produced by biological and bioinformatics researchers based on either experimental investigation or literature mining. We will use information from two categories of databases. The first category includes miRNA-mRNA regulation validated by experimental results curated either manually or through text mining tools, such as TarBase (Papadopoulos et al., 2009), miRWalk (Dweep et al., 2011), and miRtarbase (Hsu et al., 2011). Because there is no high-throughput and low-cost miRNA target screening technique available, the number of experimentally validated targets is low (Muniategui et al., 2013). The second

category of databases collects miRNA-mRNA regulations based on the complementary match between the miRNA seed region and the binding site in the 3'UTR of mRNA. Representative examples include TargetScan (Friedman et al., 2009), miRBase (Griffiths-Jones et al., 2008), and PicTar (Krek et al., 2005).

Background As the main application, we apply SAB to answer an important question in cancer genomics: how miRNAs regulate mRNA expression in a post transcriptional fashion. The central dogma of genetics states that DNA transcribe to RNA and RNA translate to protein. The expression level of messenger RNAs (mRNAs) affects the abundance of the corresponding proteins in the cellular environment. Since proteins are building blocks of all the tissues and organs, the proper expression of mRNA is important for maintaining healthy status of a human being. A person having cancer typically has abnormal expression of certain mRNAs. For example, the expression of mRNAs that suppress tumor growth might be reduced thus promoting tumor growth and cancer. MicroRNAs (miRNAs) are non-coding, meaning they do not translate to proteins. Their main function is to regulate the expression of mRNAs. Imaging an mRNA being a straight line, miRNAs can bind to the end of the line, called 3'UTR, and cause a cascade of biochemical events that could lead to the disintegration of the mRNA. Therefore, the abundance of miRNAs could negatively link to the abundance of mRNAs. It is known that many miRNAs regulate many mRNAs. There are at least 20,000 different mRNAs, each corresponding to a gene; there are at least 600 miRNAs known to human. Scientists do not know the exact mapping between mRNA and miRNA. For example, in most cases, we do not know how many and which miRNA can bind to a given mRNA and down-regulate its expression. Using a state-of-art cancer genomics data, we attempt to investigate the open question.

Data Preparation We resort to a public big data set produced by the National Institute of Health in the United States. The Cancer Genome Atlas (TCGA) is a NIH-led pilot study aiming to study cancer using a systemic approach. Among many different genomic features, TCGA measures mRNA and miRNA expression across the genome on thousands of cancer patients. Having matched miRNA and mRNA expression values on the same set of

patients provides a golden opportunity to study their interactions in cancer. We conduct an investigation. To obtain the data, we use TCGA-Assembler (Zhu et al., 2014) to download and process level-3 TCGA data of miRNA expressions and mRNA expressions for ovarian cancer patients. We selected ovarian cancer since there is a large number of patients in this cancer type. The data set includes expression measurements of 618 miRNAs and 19,830 mRNAs for 391 patients. Both mRNA expression values and miRNA expression values are generated by the next generation sequencing platforms. After preprocessing, we obtain one continuous value as expression for an mRNA or an miRNA for one sample.

The final data set contains $p = 618$ miRNAs and $P = 19,830$ mRNAs from the same $n = 391$ samples. We perform a separate linear regression treating each mRNA as a $\mathbf{y}$ response variable, and all the miRNAs as candidate $\mathbf{X}$ variables. Each variable is represented by an $n \times 1$ vector containing measurements across the $n = 391$ patient samples. Thus, we perform 19,830 separate linear regressions and apply the SAB algorithm for each regression to select important miRNAs for the corresponding mRNA. As suggested in simulation studies, we first normalize the data and then use hyperparameters $v_1 = 1$ and $v_0 = 1/\sqrt{10np} = 0.00064$.

Results After running the SAB algorithm, a set of miRNAs will be selected we use the posterior false discovery rate (FDR) to threshold the selected miRNA's. FDR is a used in multiple testing that estimate the fraction of false rejections. Here, treating the selection of each variable as a hypothesis testing, we apply the same Bayesian FDR approach (Newton et al., 2004) that has been widely used (Ji et al., 2008; Mitra et al., 2013). This procedure allows an additional selection so that biologists can rank and focus on the top selected variables for further investigation. Suppose the vector of posterior inclusion probabilities returned from SAB is $\phi = (\phi_1, \dots, \phi_p)$. The Bayesian FDR of selecting k variables is defined as follows. For an positive integer k , calculate

$$\alpha_k = 1 - \sum_{j=1}^k \phi_{(j)} / k$$

where $\phi_{(1)} \geq \dots \geq \phi_{(p)}$ are the order statistic. Then choose the largest k^* such that $\alpha_{k^*} < f_0$, where f_0 is a desired FDR rate, say 0.01. Select the top k^* variables with the highest inclusion probabilities.

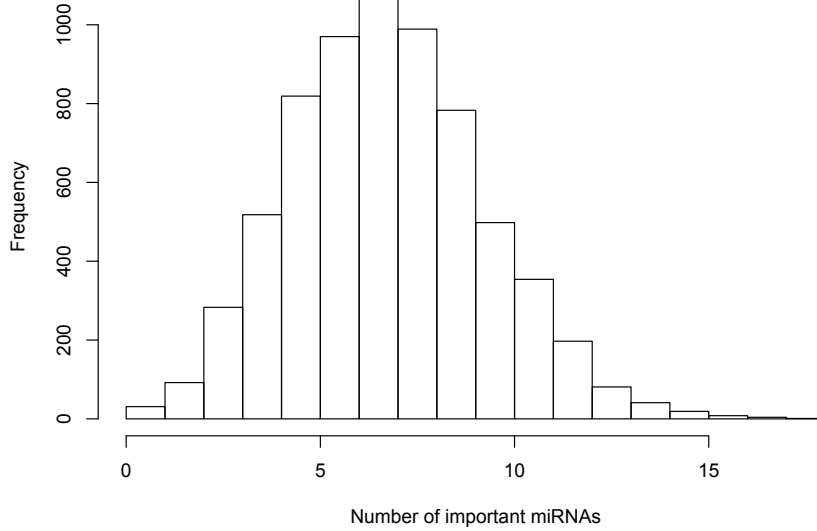


Figure 3.3: Histogram of the numbers of miRNAs selected for all the mRNAs using the SAB algorithm.

We carry out the regression analysis for all the mRNAs. For each regression, we apply SAB and threshold the miRNAs based on the Bayesian FDR procedure with an FDR cutoff $f_0 = 0.01$. Figure 3.3 shows a histogram of the numbers of miRNAs selected across all the mRNAs. The mode is 6 and closely followed by 5 and 7, implying most frequently around 6 miRNAs potentially regulate an mRNA. The histogram is slightly skewed to the right, and ranges from 0 to 17. These numbers provide an overview of the complexity of miRNA-mRNA regulation in ovarian cancer. We also provide a $P \times p$ binding matrix of potential miRNA-mRNA regulations (Online Materials) based on our analysis results. A value of 1 or 0 indicates if the corresponding miRNA is selected or not for the corresponding mRNA in our analysis, specifically.

We did post-hoc analysis including cross reference with existing biological databases. We compare our inferred $P \times p$ miRNA-mRNA binding matrix with the putative regulations identified in an existing database, TargetScan (Friedman et al., 2009). TargetScan consists of miRNA-mRNA pairs predicted based on DNA sequences. Since an miRNA binds to an mRNA following the base-pairing rule, A — T and G — C, an miRNA cannot

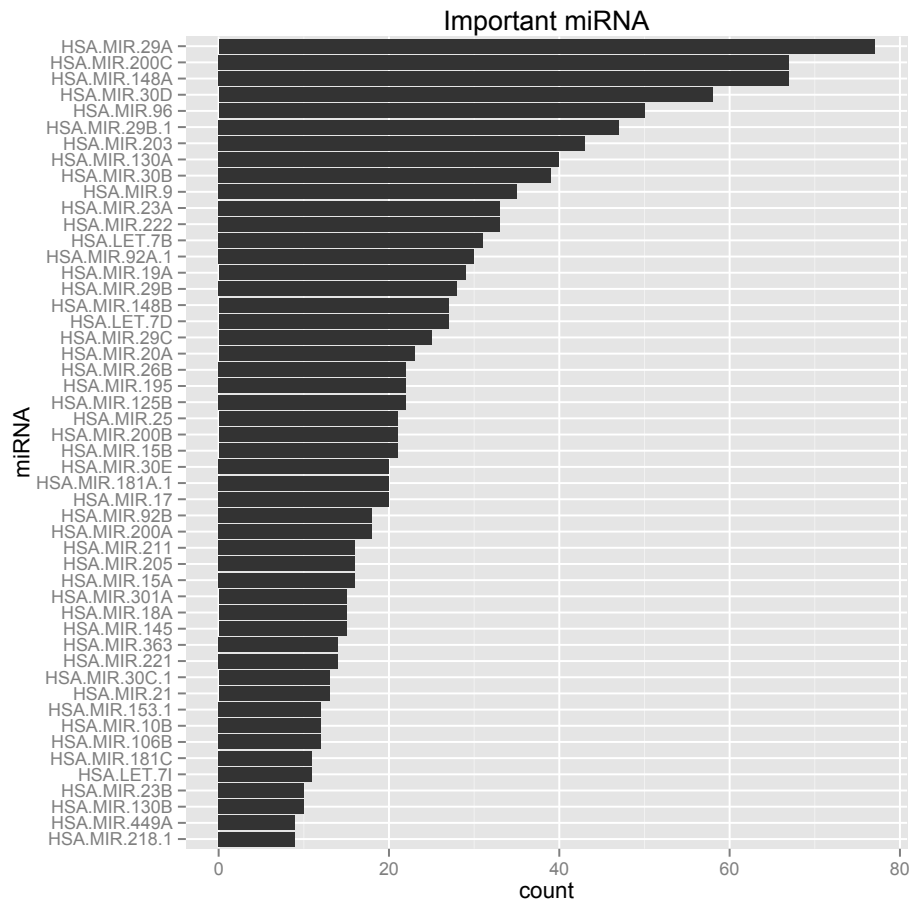


Figure 3.4: Application of SAB to mRNA-miRNA regulation using TCGA data. Top 50 influential miRNAs are shown. The bar lengths represent the number of mRNAs that each miRNA significantly repressed.

regulate an mRNA if their sequences do not pair. Based on this simple restriction, TargetScan predicts a pair of miRNA and mRNA based on the seed sequences of miRNA and the binding site sequences in the 3'UTR of the mRNA. We identify the overlap between SAB-predicted miRNA/mRNA pairs and those in TargetScan. We also require that the overlapping pairs must have negative regression coefficients to explicitly supporting the repressor roles of miRNAs. We summarize the top miRNAs identified from the cross reference in Figure 3.4. Bar plots are provided for the top influential miRNAs ranked in a decreasing order of the number of mRNAs they repress. Next, we perform some annotation on these influential miRNAs and highlight a few examples.

MiR-29a is the most influential miRNA in Figure 3.4, predicted to repress the expressions of 78 mRNAs. It acts as a tumor suppressor and its expression has been shown anti-correlated with the expressions of DNA methyltransferases including DNMT3A and DNMT3B in ovarian cancer (Creighton et al., 2012). Consistently, we observe significant negative miR-29a regulations on DNMT3A and DNMT3B. This supports a potential mechanism of the tumor suppressive role of miR-29a, in which miR-29a reverts aberrant DNA methylations that cause tumorigenesis by repressing DNA methyltransferases (Creighton et al., 2012).

MiR-200b and miR-200c are among the top influential miRNAs. They significantly repress the expression of ZEB1, a transcription factor involved in the mediation of the epithelial to mesenchymal transition (EMT), a process in which differentiated cells with an epithelial phenotype detach from one another, acquire mesenchymal properties, and become mobile (Leva and Croce, 2013). Our analysis confirms the repression of ZEB1 by miR-200b and miR-200c in ovarian cancer and explains the observation that in aggressive ovarian cancer miR-200b and miR-200c are down-regulated and the tumor undergoes EMT and becomes more invasive (Leva and Croce, 2013).

Let-7b is a tumor suppressor and serves as one of the global regulators of cellular differentiation (Li et al., 2010). Down-regulation of let-7b and loss of its gene copy have been observed in ovarian cancer (Wang et al., 2012). Our analysis indicates that let-7b represses the expression level of HMGA2, an important mRNA for cancer. The relationship has been identified in the literature (Mahajan et al., 2010). Our finding serves as an independent validation. In particular, HMGA2 is an oncogene highly expressed in ovarian

cancer and associated with advanced tumor and poor prognosis (Li et al., 2010).

mRNA CDC7 encodes a protein kinase involved in the regulation of cell cycle at the point of chromosomal DNA replication, whose expression level predicts patient disease-free survival in epithelial ovarian cancer (Kulkarni et al., 2009). MiR-29a has been observed to bind to and regulate CDC7 expression level in lung cancer cells (Barkley and Santocanale, 2013). Our analysis indicates that miR-29c, another member of the miR-29 family that shares the same seed sequence as miR-29a, significantly represses the expression of CDC7 in ovarian cancer, which can be further tested and validated in experiments.

3.5 Discussion

We have proposed a scalable algorithm for Bayesian variable selection. This fast algorithm is shown to perform well both empirically and theoretically. In our simulation studies, the accuracy of SAB is comparable and often superior in both small- p and large- p examples, when compared to existing variable selection methods. Theoretically, we have shown that SAB is asymptotically consistent. Although in most big-data applications, asymptotics is not relevant as the sample size is typically much smaller than the number of unknown parameters, methods having good large-sample properties are still desirable as they tend to be more theoretically sound.

We apply SAB to address an important biological problem in cancer genomics, and identify a list of miRNAs that are potentially important in cancer development due to their regulatory roles for a large number of genes. We note that the same type of Bayesian modeling approaches as SAB can be easily extended to address other types of biological problems, in which a massive covariates must be screened for prediction of certain outcomes. Future directions include extending SAB to accommodate discrete or survival outcomes.

Chapter 4

Asymptotic Analysis in High Dimensions

4.1 Introduction

The innovation of modern technologies drives research and development on high-dimensional data analysis in diverse fields. As the dimension grows, variable selection plays a pivotal role to ensure credible model estimation. Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad \mathbf{e} \sim \mathbf{N}(0, \sigma^2 \mathbf{I}_n),$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$ denotes the response variable of n observations, $\mathbf{X}$ the $n \times p$ design matrix, $\boldsymbol{\beta}$ the p -dimensional regression coefficient, and $\mathbf{e}$ the n -dimensional error term. The problem of variable selection is to identify the nonzero component in $\boldsymbol{\beta}$.

There is rich literature on variable selection from both frequentist and Bayesian framework. One direction is the penalized likelihood methods, such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001). Recent studies in asymptotic theory show that some nonconcave penalized likelihood methods have nice asymptotic properties when the dimension p diverges with sample size n in polynomial order $p = O(n^l)$, $l > 0$ (Fan et al., 2004; Huang and Xie, 2007) or in exponential order $p = O(e^{n^v})$, $v \in (0, 1)$ (ultra-high dimensions)(Shen et al., 2012). When X_i 's are random covariates, under certain regularity conditions, some penalized likelihood methods have been shown to achieve selection consistency, that is, with probability going to 1 with respect to the joint distribution (y_i, X_i) , the set of selected variables are exactly those in the true model, as $n \rightarrow \infty$. When X_i 's are deterministic covariates, the identifiability of the regression coefficient $\boldsymbol{\beta}$ becomes an issue when $p > n$ (Fan and Lv, 2008). As Shao and Deng (2012) pointed out, for deterministic X_i 's, the coefficient $\boldsymbol{\beta}$ is identifiable if and only if it is in

the row space of $\mathbf{X}$, denoted as $\mathcal{R}(\mathbf{X})$. Instead of estimating β , Shao and Deng (2012) showed that thresholding the estimator of θ from ridge regression achieves the selection consistency and oracle property, where θ is the projection of the true coefficient β onto the row space $\mathcal{R}(\mathbf{X})$.

Another direction of variable selection methods is the Bayesian hierarchical framework, which introduces a $p \times 1$ vector γ , such that $\gamma_j = 1$ if the j -th variable is included and $\gamma_j = 0$ otherwise. Given γ_j , a popular choice for the prior of β_j is the “spike-and-slab” prior (Mitchell and Beauchamp, 1988): a spike component (usually a point mass at zero) when $\gamma_j = 0$, and a slab component (uniform or normal distribution) when $\gamma_j = 1$. Some Bayesian variable selection methods including stochastic search variable selection (George and McCulloch, 1993) and Gibbs variable selection (Dellaportas et al., 2000) adopt a continuous version “spike-and-slab” prior where the point mass component is replaced by a spiky normal distribution with small variance, as in 3.4. With this Gaussian continuous “spike-and-slab” prior, a Gibbs sampler can be carried out to compute the posterior distribution of the model index γ . Narisetty and He (2014) showed that this Gibbs sampling method with the Gaussian “spike-and-slab” prior achieves the selection consistency and oracle property if we allow the prior variances shrinking and diffusing at a certain rate as $n \rightarrow \infty$, when p diverges with n at a nearly exponential rate.

In this chapter, we investigate the asymptotic properties of a Bayesian variable selection approach called SAB, which adopts the continuous “spike-and-slab” prior and uses the variational Bayes idea for fast computation. The model and algorithm details are given in Section 4.2. The rest of the paper is arranged as follows. We introduce some notations in Section 4.3, and then use the orthogonal design to provide insights on the shrinkage and diffusion behavior induced by the hyperparameters in Section 4.4. General results are given in Section 4.5 and 4.6. In Section 4.5, we assume $p \leq n$ and the true model β_n^* is identifiable. Section 4.6 focuses on the high dimensional data when $p > n$ with additional constraints to resolve the identifiability issue since the true model β_n^* may not be uniquely determined.

4.2 Method

The SAB algorithm adopts the continuous “spike and slab” prior for the coefficient $\boldsymbol{\beta}$, i.e. a mixture of two normal priors with different variances given the model index $\boldsymbol{\gamma}$. The hierarchical model is summarized as follows:

$$\begin{aligned} \mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathbf{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ \beta_j \mid \sigma^2, \gamma_j = 1 &\sim \mathbf{N}(0, v_1 \sigma^2) \\ \beta_j \mid \sigma^2, \gamma_j = 0 &\sim \mathbf{N}(0, v_0 \sigma^2) \\ \gamma_j \mid \theta &\sim \text{Bern}(\theta), \quad \theta \sim \text{Beta}(a_0, b_0) \\ \sigma^2 &\sim \text{IG}\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right) \end{aligned}$$

where $j = 1, \dots, p$, and v_0, v_1 are the two tuning parameters that control the shrinkage and diffusion behavior of the “spike-and-slab” prior. The prior of model index γ_j is $\text{Bern}(\theta)$, the independent Bernoulli distribution with mean θ , where θ has a Beta prior $\text{Beta}(a_0, b_0)$ with mean $a_0/(a_0 + b_0)$, and σ^2 has the inverse gamma prior $\text{IG}(\nu/2, \nu\lambda/2)$ with the mean $\mathbb{E}\sigma^{-2} = \lambda^{-1}$. In practice, after standardization, we recommend the following values for the hyper-parameters: $a_0 = b_0 = 1$, and $\nu = \lambda = 1$.

The goal of variable selection is to estimate the coefficient $\boldsymbol{\beta}$ and simultaneously identify the set of relevant variables, i.e. estimate the model index $\boldsymbol{\gamma}$. Due to the hierarchical structure of Bayesian model, we cannot directly obtain a tractable form of the posterior probability $p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y})$. The SAB algorithm uses the variational Bayes approach to compute an approximate posterior of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. More specifically, it minimizes the KL-distance between the true posterior $p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y})$ and an approximate distribution $q(\boldsymbol{\beta}, \boldsymbol{\gamma})$ that is factorized in the following form

$$q(\boldsymbol{\beta}, \boldsymbol{\gamma}) = q_1(\boldsymbol{\beta})q_2(\boldsymbol{\gamma}). \quad (4.1)$$

For the nuisance parameters θ and σ^2 , there is no need to compute their posterior, so they are updated repeatedly from their MAP estimators. Therefore the objective function of SAB is formulated as

$$J(q_1, q_2, \theta, \sigma^2) = \mathbb{E}_{\boldsymbol{\beta}}^{q_1} \mathbb{E}_{\boldsymbol{\gamma}}^{q_2} \log \frac{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2) \pi(\boldsymbol{\gamma} \mid \theta) \pi(\theta) \pi(\sigma^2)}{q_1(\boldsymbol{\beta}) q_2(\boldsymbol{\gamma})}, \quad (4.2)$$

where $\mathbb{E}_W^q$ denotes the expectation of a random variable W with respect to the distribution q . At each iteration, we optimize J by updating the distributions q_1 , q_2 and the point estimates $\hat{\theta}$, $\hat{\sigma}^2$, respectively, until they converge.

[Update q_1] Given $q_2, \hat{\sigma}^2$, the optimal $q_1(\boldsymbol{\beta})$ that maximizes the objective function is a normal distribution $\mathbf{N}_p(\mathbf{m}, \mathbf{D})$, with the mean and variance,

$$\mathbf{m} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D} = \hat{\sigma}^2 (\mathbf{X}^T \mathbf{X} + \Delta)^{-1}, \quad (4.3)$$

where Δ is a diagonal matrix with entry $\Delta_j = (1 - \phi_j)/v_0 + \phi_j/v_1$, and ϕ_j is the parameter in q_2 .

[Update q_2] Given $q_1, \hat{\theta}, \hat{\sigma}^2$, the optimal $q_2(\boldsymbol{\gamma})$ is a product of independent Bernoulli distributions, $\prod_{j=1}^p \text{Bern}(\phi_j)$, and the parameter ϕ_j 's satisfy

$$\log \frac{\phi_j}{1 - \phi_j} = \frac{1}{2} \log \frac{v_0}{v_1} + \frac{m_j^2 + D_{jj}}{2\hat{\sigma}^2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) + \log \frac{\hat{\theta}}{1 - \hat{\theta}}, \quad (4.4)$$

where m_j is the j -th component of the mean vector $\mathbf{m}$ and D_{jj} is the j -th diagonal entry of the covariance matrix $\mathbf{D}$.

[Update θ, σ^2] Given q_1, q_2 , the optimal point estimates of θ and σ^2 are,

$$\hat{\theta} = \frac{\sum_{j=1}^p \phi_j + a_0}{p + a_0 + b_0}, \quad (4.5)$$

$$\hat{\sigma}^2 = \frac{\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \Delta_j (m_j^2 + D_{jj}) + \nu\lambda}{n + p + \nu}, \quad (4.6)$$

where $\mathbb{E}_{\boldsymbol{\beta}}^{q_1} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{y} - \mathbf{X}\mathbf{m}\|^2 + \text{tr}(\mathbf{X}^T \mathbf{D} \mathbf{X})$.

The SAB algorithm is summarized as Algorithm 3 in Chapter 3.

Partial updating scheme The updating formula of q_1 involves inverting a $p \times p$ matrix $(\mathbf{X}^T \mathbf{X} + \Delta)$, where Δ depends on $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)^T$ that changes in every iteration, which is especially time-consuming when p is large. However, in practice we notice that the change for most ϕ_j 's is negligible after the first 2-3 iterations. Therefore, we introduce a new updating scheme to reduce the computation: stop updating ϕ_j once it hits the boundary (close to 0 or 1), i.e. when we are confident about its status. Suppose c_1 and c_2 are

two small constants that both go to 0 with $n \rightarrow \infty$. After the first iteration, we update $q_2 = \prod_{j=1}^p \text{Bern}(\phi_j)$ with the partial updating scheme,

$$\phi_j^{\text{par}} = \begin{cases} c_1, & \phi_j \leq c_1 \\ \phi_j, & c_1 < \phi_j < 1 - c_2 \\ 1 - c_2, & \phi_j \geq 1 - c_2. \end{cases} \quad (4.7)$$

4.3 Set-up and Notations

In the asymptotic study, we allow the number of covariates p diverging with $n \rightarrow \infty$, so that the data $\mathbf{y}_n$, the design matrix $\mathbf{X}_n$ and the coefficients $\boldsymbol{\beta}_n$ form triangular arrays. Assume that $\mathbf{y}_n$'s are generated from the following linear model

$$\mathbf{y}_n = \mathbf{X}_n \boldsymbol{\beta}_n^* + \mathbf{e}_n, \quad \mathbf{e}_n \sim \mathbf{N}_n(0, \sigma^2 \mathbf{I}_n), \quad (4.8)$$

where $\mathbf{X}_n$'s are given deterministic design matrices, $\boldsymbol{\beta}_n^* = (\beta_{n1}^*, \dots, \beta_{np}^*)^T$ denotes the true regression coefficients. We denote the true model index $\boldsymbol{\gamma}_n^* = (\gamma_{n1}^*, \dots, \gamma_{nj}^*)$ where $\gamma_{nj}^* = 1$ if $\beta_{nj}^* \neq 0$ and 0 otherwise. Let $S_n^* = \{j : \gamma_{nj}^* = 1\}$ be the set of true relevant variables and $\hat{S}_n$ denote the estimated model returned from the SAB algorithm,

$$\hat{S}_n = \{j : \phi_{nj} \geq c, j = 1, \dots, n.\}$$

where ϕ_{nj} 's are the estimated inclusion probabilities from SAB. Here we choose $c = 0.5$ as the threshold but in general it could be any number between 0 and 1 since the estimated inclusion probabilities tend to be sparse.

The goals of our asymptotic study are explained as follows. First we aim to show the selection consistency,

$$\min_{j \in S_n^*} \phi_{nj} \xrightarrow{P} 1, \quad \max_{j \notin S_n^*} \phi_{nj} \xrightarrow{P} 0. \quad (\text{i})$$

Consequently, $\mathbb{P}(\hat{S}_n = S_n^*) \rightarrow 1$, that is, the model selected by marginal posterior from SAB converges to the true model in probability. Then we will discuss the Bayesian consistency that the posterior probability on the true model converges to 1. For SAB, we aim to show that the approximate

posterior distribution of γ converges to 1,

$$q_2(\gamma_n^*) \xrightarrow{P} 1. \quad (\text{ii})$$

The Bayesian consistency, which implies the selection consistency in (i), is also called the strong consistency. At last, we aim to show that SAB achieves the oracle property,

$$\mathbb{E}\|\mathbf{X}_n\hat{\beta}_n - \mathbf{X}_n\beta_n^*\|^2 = O(q_n), \quad (\text{iii})$$

where $\hat{\beta}_n$ is the coefficient estimator from SAB, and q_n is the true number of relevant variables.

Before presenting the results, we need the following notations that are used throughout the paper.

Rate With two sequences a_n and b_n , we write $a_n \prec b_n$ if $a_n/b_n \rightarrow 0$; $a_n \sim b_n$ if $a_n/b_n \rightarrow c > 0$; and $a_n \simeq b_n$ if $a_n/b_n \rightarrow 1$. Besides, we write $a_n = O(b_n)$ if there exist a constant $M > 0$ such that $a_n/b_n \leq M$, and $a_n = o(b_n)$ if $a_n/b_n \rightarrow 0$.

Eigenvalue For a symmetric matrix A , we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the largest and smallest *positive* eigenvalue of A respectively. For a non-symmetric matrix A , we denote $\sigma_{\max}(A)$ and $\sigma_{\min}(A)$ as the largest and smallest non-zero singular value.

4.4 Orthogonal Design

In this section, we consider the special case when the design matrix $\mathbf{X}$ is orthogonal, i.e. $\mathbf{X}^T\mathbf{X} = n\mathbf{I}_p$. For the sake of conciseness, we omit the subscript n hereafter.

Selection consistency Recall that,

$$\log \frac{\phi_j}{1 - \phi_j} = \frac{1}{2} \log \frac{v_0}{v_1} + \frac{m_j^2 + D_{jj}}{2\hat{\sigma}^2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) + \log \frac{\hat{\theta}}{1 - \hat{\theta}}. \quad (4.4)$$

For simplicity, we assume that $\frac{1}{\sigma^2}$ and $\log \frac{\hat{\theta}}{1-\hat{\theta}}$ are bounded by constants. Let $v_0 \sim n^{-r_0}$, $r_0 > 0$ and fix v_1 at constant, then we have

$$\log \frac{\phi_j}{1 - \phi_j} = -O(\log n) + (m_j^2 + D_{jj}) O(n^{r_0}).$$

Thus the event $\{\hat{S}_n = S_n^*\}$ in (i) is equivalent to

$$A_n = \left\{ \min_{j \in S_n^*} [m_j^2 + D_{jj}] > r^{(0)} \right\} \cap \left\{ \max_{j \notin S_n^*} [m_j^2 + D_{jj}] < r^{(0)} \right\},$$

where the threshold $r^{(0)} \sim n^{-r_0} \log n$.

For orthogonal design, the posterior mean $\mathbf{m}$ and covariance $\mathbf{D}$ can be simplified as,

$$\begin{aligned} \mathbf{D} &= (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} = \text{diag}\left(\frac{1}{n + \Delta_j}\right) \\ \mathbf{m} &= (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T (\mathbf{X} \boldsymbol{\beta}^* + \mathbf{e}) \\ &= \text{diag}\left(\frac{n}{n + \Delta_j}\right) \boldsymbol{\beta}^* + \text{diag}\left(\frac{1}{n + \Delta_j}\right) \mathbf{X}^t \mathbf{e}, \end{aligned} \quad (4.9)$$

where $\Delta_j = \frac{\phi_j}{v_1} + \frac{1-\phi_j}{v_0} = O(n^{r_0})$. Assuming $r_0 < 1$, the posterior variance is bounded by,

$$D_{jj} = \frac{1}{n + \Delta_j} \sim n^{-1} \prec r^{(0)}. \quad (4.10)$$

Let $\mathbf{W}$ denote the second term in expression (4.9), which is a multivariate normal random variable and $\text{Cov}(\mathbf{W}) = \text{diag}(n\sigma^2/(n+\Delta_j)^2)$. So the variance of each W_j is

$$\text{Var}(W_j) = \frac{n\sigma^2}{(n + \Delta_j)^2} \sim n^{-1}.$$

Apply the upper bound property for the tail probability of Normal distribution,

$$\mathbb{P}(|W_j| > \sqrt{r^{(0)}}) \leq \sqrt{\frac{2}{\pi}} \frac{\tau}{\sqrt{r^{(0)}}} \exp\left\{-\frac{r^{(0)}}{2\tau^2}\right\} = o\left(\exp\{-n^{1-r_0}\}\right),$$

where $\tau^2 = \text{Var}(W_j)$. Using Bonferroni's inequality,

$$\mathbb{P}(\max_j W_j^2 > r^{(0)}) \leq p \cdot \mathbb{P}(|W_j| > \sqrt{r^{(0)}}) = o\left(p \exp\{-n^{1-r_0}\}\right). \quad (4.11)$$

From the two inequality results (4.10) and (4.11), we can now prove the selection consistency.

For $j \notin S_n^*$, $\beta_j^* = 0$, we have $m_j = W_j$, so,

$$\mathbb{P} \left(\max_{j \notin S_n^*} [m_j^2 + D_{jj}] > r^{(0)} \right) \leq \mathbb{P} \left(\max_{j \notin S_n^*} W_j^2 + \max_{j \in S_n^*} D_{jj} > r^{(0)} \right) = o \left(p \exp\{-n^{1-r_0}\} \right).$$

For $j \in S_n^*$, we have $m_j = \frac{n}{n+\Delta_j} \beta_j^* + W_j$, so

$$\begin{aligned} \mathbb{P} \left(\min_{j \in S_n^*} [m_j^2 + D_{jj}] < r^{(0)} \right) &\leq \mathbb{P} \left(\min_{j \in S_n^*} \frac{1}{(1 + \frac{\Delta_j}{n})^2} |\beta_j^*|^2 - \max_{j \in S_n^*} W_j^2 < r^{(0)} \right) \\ &\leq \mathbb{P} \left(\max_{j \in S_n^*} W_j^2 > r^{(0)} \right) = o \left(p \exp\{-n^{1-r_0}\} \right), \end{aligned}$$

where the second inequality requires the Beta-min condition

$$\min_{j \in S_n^*} |\beta_j^*| \sim n^{-\eta} \succ \sqrt{r^{(0)}}, \quad (4.12)$$

i.e. $2\eta < r_0$, which is sensible because the magnitude of the true effect should be greater than the noise level $n^{-1/2}$, which means $\eta < 1/2$.

To summarize, assume v_1 fixed, $v_0 \sim n^{-r_0}$, $\min_{j \in S_n^*} |\beta_j^*| \sim n^{-\eta}$, and $2\eta < r_0 < 1$, $p \prec \exp\{n^{1-r_0}\}$, then SAB achieves the selection consistency for orthogonal design, i.e.

$$\mathbb{P}(\hat{S}_n = S_n^*) = 1 - o \left(p \exp\{-n^{1-r_0}\} \right) \rightarrow 1.$$

Bayesian consistency The approximate posterior q_2 is a product of independent Bernoulli's, so the goal is to prove

$$q_2(\gamma_n^*) = \left[\prod_{j \in S_n^*} \phi_j \right] \left[\prod_{j \notin S_n^*} (1 - \phi_j) \right] \xrightarrow{P} 1.$$

Applying the inequality $\prod_j (1 - p_j) \geq 1 - \sum_j p_j$ for any $0 < p_j < 1$, we have

$$1 - q_2(\gamma_n^*) \leq \sum_{j \in S_n^*} (1 - \phi_j) + \sum_{j \notin S_n^*} \phi_j. \quad (4.13)$$

It means that the sum of probabilities on the incorrect models is upper bounded by the sum of the marginal probabilities of either including noise

variables or excluding relevant variables. When p is fixed, the selection consistency implies the Bayesian consistency since each of the p terms on the right side of (4.13) goes to 0. When p diverges with n , each term going to 0 does not guarantee the sum going to 0.

By the results (4.10), (4.11) and the condition (4.12), we can see that the leading terms in (4.4) are

$$\log \frac{\phi_j}{1 - \phi_j} \simeq -\frac{1}{2} \log \frac{v_1}{v_0}, \quad j \notin S_n^*; \quad \log \frac{\phi_j}{1 - \phi_j} \simeq \frac{|\beta_j^*|^2}{2\hat{\sigma}^2 v_0}, \quad j \in S_n^*.$$

Let C_n be a sequence that is smaller than the leading term, for example,

$$C_n = \frac{1}{2} \log \frac{v_1}{v_0 n^\alpha}, \quad 0 < \alpha < r_0. \quad (4.14)$$

Applying the new cutoff C_n to the logit function of the inclusion probabilities, the probability of selecting the true model still goes to 1, i.e.

$$\mathbb{P}\left(\max_{j \notin S_n^*} \log \frac{\phi_j}{1 - \phi_j} < -C_n\right) \rightarrow 1, \quad \mathbb{P}\left(\min_{j \in S_n^*} \log \frac{\phi_j}{1 - \phi_j} > C_n\right) \rightarrow 1.$$

Then we have

$$\begin{aligned} 1 - q_2(\gamma_n^*) &\leq p \times \max \left\{ \max_{j \in S_n^*} (1 - \phi_{nj}), \max_{j \notin S_n^*} \phi_{nj} \right\} \\ &= O_P(p e^{-C_n}) = O_P\left(p \sqrt{\frac{v_0 n^\alpha}{v_1}}\right), \end{aligned}$$

which diminishes to zero when $p \prec n^{(r_0 - \alpha)/2}$.

To achieve the Bayesian consistency when p grows more rapidly, we need to let v_1 diverging with n . Suppose $v_1 \sim n^{r_1}$ with $r_1 > 0$ such that $p \prec n^{(r_0 + r_1 - \alpha)/2}$, then

$$1 - q_2(\gamma_n^*) \leq O_P\left(\frac{p}{n^{(r_1 + r_0 - \alpha)/2}}\right) \xrightarrow{P} 0.$$

This special case of orthogonal design sheds light on the functions of the tuning parameters v_0 and v_1 on the asymptotic behavior of SAB. To achieve the selection consistency, we need v_0 going to 0 at a proper rate, slower than $1/n$. If we let v_1 going to infinity with n , SAB achieves the Bayesian consistency. Similar results were shown in (Narisetty and He, 2014).

4.5 Full-rank Data

In this section, we show the general results for full-rank data, which implies $p < n$. First we discuss the following regularity conditions.

- (C1) On colliearity. Assume λ_{n1} is the smallest eigenvalue of the gram matrix $\mathbf{X}^T \mathbf{X}$, and

$$0 < \lambda_{n1} \sim n^{\eta_1}, \quad \eta_1 \leq 1.$$

This condition is imposed to control the collinearity among the $\mathbf{X}$ columns. For finite p , (C1) is often written as $\mathbf{X}^T \mathbf{X}/n$ converging to a non-degenerate matrix, same as $\eta_1 = 1$. However for diverging p , the minimal eigenvalue of $\mathbf{X}^T \mathbf{X}$ is usually $o(n)$, so that $\eta_1 < 1$.

- (C2) The sparsity condition. Suppose the L_2 norm of the true coefficient β^* is bounded by

$$\|\beta^*\| = O(n^{\eta_2}), \quad \eta_1 > \eta_2 \geq 0.$$

For finite p and fixed β^* , (C2) is satisfied automatically with $\eta_2 = 0$. In general, we require η_2 is upper bounded by η_1 . Different from the unbiased coefficients estimated from ordinary least square, SAB estimates β similar to ridge regression, where encounters a bias term. This L_2 condition is to control the magnitude of the bias. Similar condition is assumed in (Shao and Deng, 2012).

- (C3) The beta-min condition. Assume the smallest non-zero coefficient satisfies

$$\liminf_n \frac{\min_{j \in S_n^*} |\beta_j^*|}{n^{-\eta_3}} \geq M, \quad \eta_3 < 1/2$$

for some constant $M > 0$. This condition ensures a gap between the smallest signal coefficient of $O(n^{-\eta_3})$ and the noise term which is $O(n^{-1/2})$, so that we could asymptotically detect even the weakest signal. Similar condition is assumed in the selection consistency of LASSO (Zhao and Yu, 2006).

- (C4) On hyperparameters. Assume the point estimates of $\hat{\sigma}^2$ and $\hat{\theta}$ are bounded such that

$$\frac{1}{\hat{\sigma}^2} \sim 1, \quad \log \frac{\hat{\theta}}{1 - \hat{\theta}} \sim 1.$$

This condition is purely technical for the simplicity of the proof. Another way is to set thresholds for $\hat{\sigma}^2$ and $\hat{\theta}$ to avoid them going to extreme values.

4.5.1 Selection consistency of one-step estimator

Now consider a one-step SAB algorithm. Initialize the algorithm with everybody-in, i.e. $\phi^{(0)} = (1, \dots, 1)^t$, $\theta^{(0)} = 1/2$ and $\sigma^{(0)} = 1$. Compute $\mathbf{m}^{(0)}$ and $\mathbf{D}^{(0)}$ from (4.3),

$$\mathbf{m}^{(0)} = (\mathbf{X}^T \mathbf{X} + \mathbf{I}/v_1)^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D}^{(0)} = (\mathbf{X}^T \mathbf{X} + \mathbf{I}/v_1)^{-1}.$$

Update $\phi^{(1)}$ from (4.4),

$$\log \frac{\phi_j^{(1)}}{1 - \phi_j^{(1)}} = \frac{1}{2} \log \frac{v_0}{v_1} + \frac{m_j^{(0)2} + D_{jj}^{(0)}}{2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) \quad (4.15)$$

Denote $\hat{S}_n = \{j : \phi_j^{(1)} > 0.5\}$ as the model selected by the one-step SAB algorithm. We aim to show that $\hat{S}_n$ converges to S_n^* in probability. Assume $v_0 \sim n^{-r_0}$, and $v_1 \sim 1$. The event $\{\hat{S}_n = S_n^*\}$ is equivalent to

$$\left\{ \min_{j \in S_n^*} [m_j^2 + D_{jj}] > r^{(0)} \right\} \cap \left\{ \max_{j \notin S_n^*} [m_j^2 + D_{jj}] < r^{(0)} \right\},$$

where the threshold $r^{(0)} = O(n^{-r_0} \log n)$. Rewrite $\mathbf{m}^{(0)}$ as three terms,

$$\begin{aligned} \mathbf{m}^{(0)} &= (\mathbf{X}^T \mathbf{X} + \mathbf{I}/v_1)^{-1} \mathbf{X}^T (\mathbf{X} \boldsymbol{\beta}^* + \mathbf{e}) \\ &= \boldsymbol{\beta}^* - (v_1 \mathbf{X}^T \mathbf{X} + \mathbf{I})^{-1} \boldsymbol{\beta}^* + (\mathbf{X}^T \mathbf{X} + \mathbf{I}/v_1)^{-1} \mathbf{X}^t \mathbf{e} \\ &= \boldsymbol{\beta}^* - \mathbf{b}^{(0)} + \mathbf{W}^{(0)}, \end{aligned}$$

Some key results on the order of $[m_j^2 + D_{jj}]$ are listed below.

- The diagonal entries of $\mathbf{D}^{(0)}$ are upper bounded by,

$$D_{jj}^{(0)} \leq \lambda_{\max} ((\mathbf{X}^T \mathbf{X} + \mathbf{I}/v_1)^{-1}) \leq \frac{1}{\lambda_{n1} + 1/v_1} = O(n^{-\eta_1}). \quad (4.16)$$

- The bias term $b_j^{(0)}$ is bounded by

$$\max_j |b_j^{(0)}| \leq \|\mathbf{b}^{(0)}\| \leq \frac{1}{v_1 \lambda_{n1} + 1} \|\boldsymbol{\beta}^*\| = O(n^{\eta_2 - \eta_1}). \quad (4.17)$$

- $W_j^{(0)}$ is a Gaussian random variable with mean zero and variance:

$$\text{Var}(W_j^{(0)}) \leq \sigma^2 \lambda_{\max} ((\mathbf{X}^T \mathbf{X} + \mathbf{I}/v_1)^{-1}) = O(n^{-\eta_1}).$$

- By Bonferroni's inequality and the Gaussian upper tail property,

$$\mathbb{P}\left(\max_j |W_j^{(0)}|^2 > r^{(0)}\right) = o(p \exp\{-n^{\eta_1 - r_0}\}). \quad (4.18)$$

Combining the results (4.16- 4.18), we conclude the following theorem of the selection consistency for one-step SAB algorithm.

Theorem 4.1. *Assume conditions (C1-C4), and $v_1 \sim 1$, $v_0 \sim n^{-r_0}$ with $2\eta_3 < r_0 < \min\{\eta_1, 2(\eta_1 - \eta_2)\}$. Then the one-step SAB algorithm for full-rank data ($p < n$) achieves selection consistency,*

$$\mathbb{P}(\hat{S}_n = S_n^*) = 1 - o(p \exp\{-n^{\eta_1 - r_0}\}).$$

4.5.2 Bayesian consistency

Theorem 1 shows the consistency of a point estimate of S_n^* . In the Bayesian framework, we obtain the posterior distribution over the model space. The *Bayesian consistency* for variable selection requires that the posterior mass on the true model $\boldsymbol{\gamma}_n^*$ converges to 1, which is stronger than the selection consistency.

Following the same argument as in orthogonal design, we conclude that SAB achieves strong Bayesian consistency for variable selection, i.e.

$$q_2(\boldsymbol{\gamma}_n^*) = \left[\prod_{j: \gamma_{nj}^* = 1} \phi_{nj} \right] \left[\prod_{j: \gamma_{nj}^* = 0} (1 - \phi_{nj}) \right] \xrightarrow{P} 1,$$

if we let $v_1 \sim n^{r_1}$ where r_1 satisfies $r_1 > 2 - r_0$, or more generally $p \prec n^{(r_0 + r_1)/2}$. The key idea is that for any C_n defined at (4.14) with small

enough $\alpha > 0$, we have

$$\begin{aligned} 1 - q_2(\gamma_n^*) &\leq p \times \max \left\{ \max_{j \in S_n^*} (1 - \phi_{nj}), \max_{j \notin S_n^*} \phi_{nj} \right\} \\ &\leq O_P(p e^{-C_n}) = O_P\left(\frac{p}{n^{(r_1+r_0-\alpha)/2}}\right) \xrightarrow{P} 0. \end{aligned}$$

Theorem 4.2. *Assume conditions (C1-C4), and $v_1 \sim n^{r_1}$, $v_0 \sim n^{-r_0}$ with*

$$2\eta_3 < r_0 < \min\{\eta_1, 2(\eta_1 - \eta_2)\}, \quad r_1 > 2 - r_0.$$

Then SAB achieves Bayesian consistency for variable selection, $q_2(\gamma_n^) \xrightarrow{P} 1$.*

4.5.3 Selection consistency of t -step estimator

Now we consider the model returned by SAB after finite $t > 1$ iterations. Note that at the first step, $\Delta = \mathbf{I}/v_1$, so the derivation is similar to ridge regression. After t iterations, however, the diagonal entries of Δ are random between $1/v_0$ and $1/v_1$, and correlated with the error term. Therefore the bound for the variance term $\mathbf{W}$ which is not Gaussian is more complicated now.

A simple remedy is the partial updating scheme. Theorem 4.2 shows that with probability $1 - o(p \exp\{-n^{\eta_1-r_0}\})$,

$$\max \left\{ \max_{j \in S_n^*} (1 - \phi_j^{(1)}), \max_{j \notin S_n^*} \phi_j^{(1)} \right\} < e^{-C_n} = O(n^{-(r_1+r_0-\alpha)/2}),$$

for any small constant $\alpha > 0$. So if we let $c_1 = c_2 = e^{-C_n}$, and apply the partial updating scheme defined in (4.7), with probability going to 1, we have

$$\phi_j^{\text{par}} = 1 - c_2, \quad j \in S_n^*; \quad \phi_j^{\text{par}} = c_1, \quad j \notin S_n^*.$$

After the first iteration, the partial inclusion probability ϕ^{par} is a bi-valued vector with probability $1 - o(p \exp\{-n^{\eta_1-r_0}\})$. Therefore the new Δ is not correlated with the error term $\mathbf{e}$. With a similar outline in the proof of Theorem 4.1, the selection consistency of the estimator from the T -th iteration follows.

Theorem 4.3. Assume conditions (C1-C4), and $v_1 \sim n^{r_1}, v_0 \sim n^{-r_0}$, with

$$2\eta_3 < r_0 < \frac{2}{3}(\eta_1 - \eta_2), \quad r_1 > r_0 + \alpha.$$

If we start from all variables included and run SAB for finite T iterations, applying the partial updating scheme with fixed $c_1 = c_2 = e^{-C_n} \sim n^{-(r_0+r_1-\alpha)/2}$, then the estimated model $\hat{S}_n = \{j : \phi_j^{(t)} > 0.5\}$ is a consistent estimation of the true model S_n^* , that is,

$$\mathbb{P}(\hat{S}_n = S_n^*) = (1 - o(p \exp\{-n^{\eta_1-r_0}\}))^T.$$

4.5.4 Oracle property

As pointed out by (Shao and Deng, 2012), the estimator of β from ridge regression, for example the posterior mean of the one-step SAB algorithm, $\hat{\beta} = (\mathbf{X}^T \mathbf{X} + \mathbf{I}/v_1)^{-1} \mathbf{X} \mathbf{y}$, has the prediction error $\mathbb{E} \|\mathbf{X} \hat{\beta} - \mathbf{X} \beta^*\|^2 = O(p)$. However, given the true set of relevant variables, the estimator of the q -dimensional coefficient vector from ordinary least square has the oracle property that the prediction error is of $O(q)$, much smaller than $O(p)$, where $q = \sum_j \gamma_j^*$ denotes the size of the true model. Let $\tilde{\beta} = \mathbf{m} \cdot \hat{\gamma}$, where $\hat{\gamma}_j = \mathbf{1}(\hat{\phi}_j^{\text{par}} > 0.5)$, the truncated posterior mean returned from SAB algorithm. We aim to show that the prediction based on $\tilde{\beta}$ achieves the oracle property, i.e.

$$\mathbb{E} \|\mathbf{X} \tilde{\beta} - \mathbf{X} \beta^*\|^2 = O(q).$$

In the proof, we need an additional condition on model identifiability.

- (C5) The identifiability condition. Denote $\mathbf{X}_1$ as the $n \times q$ signal covariates and $\mathbf{X}_2$ the $n \times (p - q)$ noise covariates. Assume that there exists some number $c \sim n^{-\eta_4}$, $\eta_4 \geq 0$,

$$\lambda_{\max}(P_{\mathbf{X}_1} P_{\mathbf{X}_2}) < 1 - c,$$

where $P_{\mathbf{X}_1}$ and $P_{\mathbf{X}_2}$ are the projection matrices of $\mathbf{X}_1$ and $\mathbf{X}_2$. This condition assumes that the two sub-column spaces are disjoint, which guarantees the model identifiability. On the contrary, if some part in $\mathbf{X}_1$ can be linearly explained by $\mathbf{X}_2$, i.e. there exists non-zero vectors u_1, u_2 such that $\mathbf{X}_1 u_1 = \mathbf{X}_2 u_2$, then we would have $\lambda_{\max}(P_{\mathbf{X}_1} P_{\mathbf{X}_2}) = 1$.

Theorem 4.4. *Assume conditions (C1-C5), and $v_1 \sim n^{r_1}$, $v_0 \sim n^{-r_0}$ with $r_1 - r_0 > 2\eta_2 - \eta_1$, and $2\eta_3 < r_0 < \min\{\eta_1, 2(\eta_1 - \eta_2)\}$. Then the estimator $\tilde{\beta} = \mathbf{m} \cdot \hat{\gamma}$ has the following oracle property,*

$$\mathbb{E}\|\mathbf{X}\tilde{\beta} - \mathbf{X}\beta^*\|^2 = O(qn^{\eta_4}).$$

4.6 High-dimensional Data

In the previous section, we proved the asymptotic properties of SAB for $p < n$ when the model is uniquely determined. While, the coefficient β^* is typically not identifiable for high dimensional data. The true coefficient β^* is identifiable if and only if β^* is in the subspace spanned by the row vectors of $\mathbf{X}$, denoted as $\mathcal{R}(\mathbf{X})$. Assume $\mathbf{X} = P_{n \times r} D_{r \times r} Q_{p \times r}^T$ is the singular value decomposition with rank r . Consider the projection of β^* onto $\mathcal{R}(\mathbf{X})$,

$$\theta = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-}\mathbf{X}\beta^* = QQ^T\beta^*.$$

Clearly $\mathbf{X}\theta = \mathbf{X}\beta^*$. Since $\|QQ^T\| \leq 1$, the projection θ has the smallest L_2 norm among the equivalent set $\mathcal{B} = \{\beta : \mathbf{X}\beta = \mathbb{E}\mathbf{y}\}$. So another way to define θ is,

$$\theta = \arg \min_{\beta \in \mathcal{B}} \|\beta\|_2.$$

Even if the true coefficient β^* is sparse, its projection θ may not have any zero component. However its nonzero components must be negligible since its L_2 norm is bounded. To bypass the non-identifiable issue, (Shao and Deng, 2012) estimated the projection θ instead of the true coefficient β to discuss the asymptotic properties of ridge regression. But there was no guarantee that the variables selected by thresholding θ are the same set of signals in the true β^* .

However, when assuming condition (C5) which implies that the signal terms $\mathbf{X}_1$ cannot be linearly explained by the noise terms $\mathbf{X}_2$, the signal part of the projection coefficient θ is the same as that of the true coefficient β^* . Moreover, when β^* is sparse, $\theta = \beta^*$. For this reason, together with condition (C5), we use $\beta^* = \theta$ exchangeably to denote the “ideal” coefficient for estimation.

When the true β^* is not sparse, we need to put a threshold to define the

true set of variables. So the beta-min condition on weak signals should be replaced by the following condition.

(C3*) Define the set of signals S_n^* and the set of noises S_n^{*c} as,

$$S_n^* = \{j : |\beta_{nj}^*| \geq b_n\}, \quad S_n^{*c} = \{j : |\beta_{nj}^*| \leq a_n\} \quad (4.19)$$

where a_n and b_n are two sequences depending on n , which satisfy $a_n \leq b_n$ and $a_n/b_n \rightarrow 1$. For example, let $a_n \sim \sqrt{r^{(0)}/u_n}$, $b_n \sim \sqrt{r^{(0)}u_n}$, where $u_n = 1 + 1/\log \log n$ and $r^{(0)} \sim n^{-r_0} \log n$. This condition guarantees a gap between the magnitudes of the signal and noise variables. The coefficients with a magnitude between a_n and b_n are not identifiable and thereby omitted in this discussion. Similar condition is used in (Shao and Deng, 2012).

In the high-dimensional setting, the $p \times p$ gram matrix $\mathbf{X}^T \mathbf{X}$ is not full rank, so the condition on the smallest eigenvalue of $\mathbf{X}^T \mathbf{X}$ should be replaced by the smallest positive eigenvalue.

(C1*) Assume λ_{n1} is the smallest *positive* eigenvalue of the gram matrix $\mathbf{X}_n^T \mathbf{X}_n$ which satisfies

$$\lambda_{n1} \sim n^{\eta_1}, \quad \eta_1 \leq 1.$$

Now with the modified conditions, we have the following asymptotic results for the SAB estimator with high-dimensional data.

4.6.1 Selection consistency

Theorem 4.5. Assume the conditions (C1*, C2, C3*, C4-C5), and $v_1 \sim n^{r_1}$, $v_0 \sim n^{-r_0}$, with

$$2\eta_3 < r_0 < \min\{2(\eta_1 - \eta_2 - \eta_4)/3, \eta_1 - 2\eta_4\}, \quad r_1 \geq 0.$$

If we start from all variables out, i.e. $\phi^{(0)} = (0, \dots, 0)'$, and run SAB for finite T iterations, applying the partial updating scheme with fixed $c_1 = c_2 = e^{-C_n} \sim n^{-(r_0+r_1-\alpha)/2}$, then the estimator from SAB $\hat{S}_n = \{j : \phi_j^{(t)} > 0.5\}$ is a

consistent estimation of the true model S_n^* , more specifically:

$$\mathbb{P}(\hat{S}_n = S_n^*) = (1 - o(p \exp\{-n^{\eta_1 - 2\eta_4 - r_0}\}))^T. \quad (4.20)$$

This theorem shows that under certain conditions the SAB estimator achieves selection consistency, when the number of covariates $p = O(e^{n^v})$ and $v > 0$. The proof is similar to Theorem 4.1 and 4.3 but with three main differences. First, to bound the variance term, we need to initialize from everybody excluded since $\mathbf{X}^T \mathbf{X}$ is not full rank. Second, at the t -th iteration ($t > 1$), we need to consider the order of signals and noises separately, with details given in the appendix. Finally, due to the non-sparseness of the true β^* , we need to compare the cutoffs a_n and b_n in (C3*) with the threshold $r^{(0)}$. The details of the proof is given in the appendix.

4.6.2 Oracle property

Theorem 4.6. Assume the conditions (C1*, C2, C3*, C4-C5), and $v_1 \sim n^{r_1}, v_0 \sim n^{-r_0}$, with

$$2\eta_3 < r_0 < \min\{2(\eta_1 - \eta_2 - \eta_4)/3, \eta_1 - 2\eta_4\}, \quad r_1 > \max\{\eta_2 - r_0/2, 0\}.$$

The prediction based on $\tilde{\beta} = \mathbf{m} \cdot \hat{\gamma}$ has the following oracle property:

$$\mathbb{E}\|\mathbf{X}\tilde{\beta} - \mathbf{X}\beta^*\|^2 = O(qn^{\eta_4}) + O(n^{\alpha + r_0 - r_1 + 2\eta_2 - \eta_1 + 2\eta_4}) = O(qn^{\eta_4}),$$

where $\mathbf{m}$ is the posterior mean, $\hat{\gamma}$ is the binary vector $\hat{\gamma}_j = \mathbf{1}(\hat{\phi}_j^{par} > 0.5)$ and $\hat{\phi}_j^{par}$'s are the inclusion probabilities returned from SAB with the partial updating scheme, and $q = q_n = \sum_j \gamma_j^*$ denote the size of the true model.

This theorem shows that under certain conditions the truncated posterior mean returned from SAB achieves the oracle property.

Chapter 5

Logistic Model

5.1 Introduction

In statistics, logistic regression is the most commonly used method to model binary response data. However, transforming those nicely-built theorems and algorithms from linear regression to logistic model is not always straightforward. In the Bayesian framework, this difficulty is mainly due to the non-conjugate character of the likelihood, so that the posterior does not have an explicit form. But researchers have developed multiple solutions to bypass this problem, which come from either one or a combination of the three basic ideas: MAP (maximize a posteriori) estimation, sampling algorithm or approximation methods. To be more specific, MAP procedure provides only a point estimation so that we don't need to compute the whole posterior, but this will surely miss out some information. Sampling methods use the general Metropolis Hasting Algorithm to sample from the posterior, which is more accurate and informative but also notorious for its slow mixing time. Approximation methods, on the other hand, use different probability models (often connecting to exponential family) to approximate the posterior, for example Probit approximation or Laplace approximation, etc. In this chapter, we focus on the third one, approximation methods.

There are several basic equivalent expressions of the Logistic model, of which the Generalized Linear Model (GLM) and the Latent Variable Model (LVM) are the most common ones. Other specifications like Log-linear Model and Single-layer Perceptron Model are beyond the scope of this thesis.

- GLM

$$\begin{aligned} P(y_i = 1) &= p_i \\ \text{logit}(p_i) &\triangleq \log \frac{p_i}{1 - p_i} = \boldsymbol{\beta}^T X_i \end{aligned}$$

- LVM

$$y_i = \begin{cases} 1, & \text{if } Y_i^* > 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$Y_i^* = \beta^T X_i + \epsilon, \quad \epsilon \sim \text{Logistic}(0, 1)$$

In the LVM specification, replacing the Logistic prior by a standard normal distribution leads to the Probit approximation, which makes the computation easier but the prediction much less accurate. A recent work by Holmes et al. (2006) first introduced a hierarchical structure for the error term. Instead of directly applying a Logistic prior, it introduced a latent variable that follows Kolmogorov-Smirnov (KS) distribution, so that the error term has a prior from the Normal family with that latent variable being its standard deviation.

$$Y_i^* = \beta^T X_i + \epsilon_i$$

$$\epsilon_i \sim \text{N}(0, (2\phi_i)^2), \quad \phi_i \sim \text{KS}$$

This hierarchical structure makes many Bayesian algorithms applicable because the logistic likelihood is transformed into the conjugate Gaussian family, and sampling from the KS distribution is simple.

In the GLM specification, Polson et al. (2013) introduced an auxiliary variable from Polya-Gamma distribution, given which the likelihood is valued as a certain Normal density. Then they use an EM algorithm to estimate the parameters for variable selection. But in terms of sparse data, if we delete variables with a small point estimation, then those variables cannot reenter the model.

Based on the Polya-Gamma setting, we propose a hybrid framework that combines the Variational Bayes technique and MAP estimation, using the continuous spike-and-slab prior on β . In section 5.2.3, we prove that the Polya-Gamma method is equivalent to the local approximation method (Bishop, 2006) for Variational Bayes algorithms.

5.2 Method

5.2.1 Model and priors

We use the GLM specification and code y_i 's as ± 1 . Let $z_i = y_i X_i^T \boldsymbol{\beta}$ to be the working response. Then the likelihood function can be written as,

$$p(z_i | \boldsymbol{\beta}) = \frac{1}{1 + e^{-z_i}}.$$

Note that given $\boldsymbol{\beta}$, z_i only has two possible values $\pm X_i^T \boldsymbol{\beta}$. This likelihood is proportional to an integral of an exponential family of z_i over latent variable w_i ,

$$p(z_i | \boldsymbol{\beta}) = \frac{1}{2} \int_0^\infty \exp\left\{-\frac{w_i}{2} z_i^2 + \frac{1}{2} z_i\right\} dP_{\mathcal{PG}}(w_i),$$

where $P_{\mathcal{PG}}$ is the Polya-Gamma distribution $\mathbf{PG}(1, 0)$.

The Polya-Gamma distribution $P_{\mathcal{PG}}(v | b, c)$ is defined as the sum of infinite many independent Gamma random variables,

$$V \triangleq \frac{1}{2\pi^2} \sum_{k=1}^{\infty} \frac{g_k}{(k - 1/2)^2 + c^2/(4\pi^2)}, \quad g_k \sim \text{Gamma}(b, 1),$$

where the Gamma distribution uses a shape and scale parameterization. The main result is based on the following integral identity:

$$\frac{1}{1 + e^{-\theta}} = \frac{1}{2} \int_0^\infty \exp\left\{-\frac{\theta^2}{2} v + \frac{\theta}{2}\right\} P_{\mathcal{PG}}(v | 1, 0) dv,$$

If $v \sim \mathbf{PG}(b, c)$, $c > 0$, we have

$$\mathbb{E}v = \frac{b}{2c} \frac{e^c - 1}{e^c + 1} = 2b\lambda(c), \quad \text{where } \lambda(c) \triangleq \frac{1}{4c} \frac{e^c - 1}{e^c + 1}, \quad (5.1)$$

$$p(v) \propto \exp\left\{-\frac{c^2}{2} v\right\} P_{\mathcal{PG}}(v | b, 0). \quad (5.2)$$

When $c = 0$, $\mathbb{E}v = b/4$.

We adopt the continuous spike-and-slab prior for the coefficient $\boldsymbol{\beta}$,

$$\pi(\beta_j | \gamma_j, \sigma) = \mathbf{N}(0, \sigma^2 \alpha_j), \quad \alpha_j = v_0(1 - \gamma_j) + v_1 \gamma_j,$$

where $0 < v_0 < v_1$, and γ_j is the indicator of inclusion for the j^{th} variable.

When $\gamma_j = 1$, β_j is from a flat normal prior with a large variance; when $\gamma_j = 0$, β_j has a spiky normal prior. The remaining prior specification is standard,

$$\begin{aligned}\pi(\boldsymbol{\gamma} \mid \theta) &= \text{Bern}(\theta) \\ \pi(\theta) &= \text{Beta}(a_0, b_0) \\ \pi(\sigma^2) &= \text{IG}(\nu_0/2, \nu_0\lambda_0/2)\end{aligned}$$

Given latent variables w , the likelihood becomes a normal density measured at $\pm \mathbf{X}_i^T \boldsymbol{\beta}$, conjugate to the priors. However, the full posterior below is not available in close-form due to the dependency among the parameters,

$$\pi(\boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\gamma}, \theta, \sigma \mid \mathbf{y}) = p(\mathbf{z} \mid \boldsymbol{\beta}, \mathbf{w})\pi(\mathbf{w})\pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma)\pi(\boldsymbol{\gamma} \mid \theta)\pi(\theta)\pi(\sigma).$$

One can use MCMC to get an exact inference on the full posterior, but for big data problems, the computation is sometimes too time-consuming. In this chapter, we use the Variational Bayes approach to approximate the posterior.

5.2.2 The variational algorithm

To reduce computation, we use point estimation for the nuisance parameters θ and σ . Note that the posterior probability $P = \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{w} \mid \mathbf{z}, \hat{\theta}, \hat{\sigma}^2)$ together with the MAP estimators $\hat{\theta}$ and $\hat{\sigma}^2$ is the optimal solution to maximize the following objective function:

$$\mathcal{L}(Q, \theta, \sigma^2) = \mathbb{E}^Q \log \frac{\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{w}, \theta, \sigma^2 \mid \mathbf{y})}{Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{w})} \quad (5.3)$$

which is the negative KL divergence between Q and the true posterior. But the true posterior P does not have an explicit form because of the hierarchical structure. The variational idea is to restrict Q to a factorized form, such that the solution of the optimization problem is in the conjugate family. Assume the approximate posterior Q factorized in the form,

$$Q(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{w}) = q(\boldsymbol{\beta}) \prod_{j=1}^p q(\gamma_j) \prod_{i=1}^n q(w_i).$$

Now we can update the approximate posterior functions q_{β} , q_{γ} , $q_{\mathbf{w}}$, as well as the point estimators $\hat{\theta}$ and $\hat{\sigma}^2$ sequentially until the objective function is converged and optimized.

[Update of q_{β}] The optimal q_{β} that maximize (5.3) is

$$\begin{aligned}\log q(\beta) &= \mathbb{E}_{-\beta} \log [p(z_i | \beta, \sigma, w_i) \pi(\beta | \gamma, \sigma)] + C. \\ &= \sum_{i=1}^n -\frac{z_i^2}{2} \mathbb{E} w_i + \frac{z_i}{2} - \frac{1}{2\sigma^2} \beta^T \text{diag}(\Delta_i) \beta \\ &= -\frac{1}{2} \beta^T \left(\sum_{i=1}^n \mathbb{E} w_i X_i X_i^T + \text{diag}(\Delta_j)/\sigma^2 \right) \beta + \beta^T \sum_{i=1}^n \frac{1}{2} y_i X_i\end{aligned}$$

where $\Delta_j = \mathbb{E} \frac{1}{\alpha_{\gamma_j}} = \frac{(1-\phi_j)}{v_0} + \frac{\phi_j}{v_1}$. Therefore, $q(\beta)$ is a multivariate normal distribution $\mathbf{N}(\mathbf{m}, D)$ with,

$$D^{-1} \mathbf{m} = \sum_{i=1}^n \frac{1}{2} y_i X_i, \quad D^{-1} = \sum_{i=1}^n \mathbb{E} w_i X_i X_i^T + \text{diag}(\Delta_j)/\sigma^2. \quad (5.4)$$

Note that even though the prior of $\mathbf{w}$ is a complex distribution, the only statistic that we need for this algorithm is the mean. From the property of Polya-Gamma distribution in (5.1), we get $\mathbb{E}_q w_i = 2\lambda(c_i)$.

[Update of q_{γ}] The optimal q_{γ} that maximize (5.3) is standard:

$$\begin{aligned}\log q(\gamma_j) &= \mathbb{E}_{-\gamma} \log [\pi(\beta | \gamma, \sigma) \pi(\gamma | \theta)] + C. \\ &= -\frac{\mathbb{E} \beta_j^2}{2\sigma^2} \left(\frac{(1-\gamma_j)}{v_0} + \frac{\gamma_j}{v_1} \right) + \gamma_j \log \theta + (1-\gamma_j) \log(1-\theta) \\ &= \gamma_j \left[\frac{\mathbb{E} \beta_j^2}{2\sigma^2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) + \log \frac{\theta}{1-\theta} \right]\end{aligned}$$

So $q(\gamma_j) = \text{Bern}(\phi_j)$ with parameter ϕ_j computed from,

$$\log \frac{\phi_j}{1-\phi_j} = \frac{\mathbb{E} \beta_j^2}{2\sigma^2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) + \log \frac{\theta}{1-\theta}. \quad (5.5)$$

[Update of $\hat{\theta}$ and $\hat{\sigma}^2$] The point estimator of θ and σ^2 that maximize (5.3) are

$$\hat{\theta} = \frac{a_0 + \sum \phi_j}{a_0 + b_0 + p}, \quad \hat{\sigma}^2 = \frac{\sum \mathbb{E} \beta_j^2 \Delta_j + \nu_0 \lambda_0}{p + \nu_0}. \quad (5.6)$$

[Update of q_w]

$$\begin{aligned}
q(w_i) &\propto \exp\{\mathbb{E}_{z_i} \log p(w_i | y_i)\} \\
&\propto \exp\left\{-\frac{\mathbb{E}z_i^2}{2}w_i + \frac{\mathbb{E}z_i}{2} + \log P_{\mathcal{PG}}(w_i | 1, 0)\right\} \\
&\propto \exp\left\{-\frac{\mathbb{E}z_i^2}{2}w_i\right\} P_{\mathcal{PG}}(w_i | 1, 0)
\end{aligned}$$

From the property (5.2), $q(w_i) = \text{PG}(1, c_i)$ with $c_i^2 = \mathbb{E}z_i^2 = X_i^T(D + \mathbf{m}\mathbf{m}^T)X_i$.

As we sequentially update the approximate function Q and the point estimators, the object function $\mathcal{L}$ keeps increasing, but cannot exceed the upper bound $\log p(\mathbf{y})$, which can be used as a stopping criteria: stop the algorithm when the increase of $\mathcal{L}$ is below a threshold. However computing $\mathcal{L}$ involves the determinant of a $p \times p$ matrix D (as shown in Appendix), which is time-consuming. Since γ is the main interest for variable selection, we use the entropy of γ with respect to the distribution q_γ as our stopping criteria,

$$Ent_q(\gamma) = -\sum \phi_j \log \phi_j - \sum (1 - \phi_j) \log(1 - \phi_j). \quad (5.7)$$

Algorithm 4: LVB Algorithm

Input: $\mathbf{X}, \mathbf{y}, v_0, v_1, a_0, b_0, \nu_0, \lambda_0$

Output: $\hat{\phi}, \hat{\mathbf{m}}, \hat{D}, \hat{\theta}, \hat{\sigma}^2$

Initialization $\phi^{(0)}, \mathbf{m}_0, D_0, \theta_0, \sigma_0$;

for $t = 1 : \text{maxIter}$ **do**

 Update the approximate posterior parameters $\mathbf{m}^{(t)}, D^{(t)}, \phi^{(t)}$ from equations (5.4, 5.5);

 Update the point estimators $\hat{\theta}^{(t)}, \hat{\sigma}^{(t)}$ from equation (5.6);

 Calculate the entropy $Ent_q(\gamma)$ from equation (5.7);

if *The difference of the entropy is below a threshold* **then**

 break;

end

end

5.2.3 Connection with local approximation

Local approximation is another variational method for the logistic model (Bishop, 2006), which uses a lower bound in the exponential form to approximate the likelihood. Recall that $y_i = \pm 1$, for all ξ ,

$$p(y_i | \boldsymbol{\beta}) = e^{a(y_i+1)/2} \zeta(-a) \geq e^{ay_i/2} \zeta(\xi) \exp \{-\xi/2 - \lambda(\xi)(a^2 - \xi^2)\}$$

where $\zeta(a) = 1/(1 + e^{-a})$, $a = \boldsymbol{\beta}^T X_i$, $\lambda(\xi) = \frac{1}{2\xi} [\zeta(\xi) - \frac{1}{2}]$. More detail of this inequality is given in the Appendix. The lower bound of the likelihood is

$$p(\mathbf{y} | \boldsymbol{\beta}) \geq h(\boldsymbol{\beta}, \xi) \stackrel{\text{def}}{=} \prod_{i=1}^n \zeta(\xi_i) \exp \{ \boldsymbol{\beta}^T X_i y_i / 2 - \xi_i / 2 - \lambda(\xi_i) ((\boldsymbol{\beta}^T X_i)^2 - \xi_i^2) \}.$$

Adopt the same priors as in the previous section, and replace the likelihood by its lower bound $h(\boldsymbol{\beta}, \xi)$, then the new objective function becomes,

$$\begin{aligned} \log p(\mathbf{y}) &\geq \tilde{\mathcal{L}}(q_{\boldsymbol{\beta}}, q_{\boldsymbol{\gamma}}, \theta, \sigma, \xi) \\ &= \int \int q(\boldsymbol{\beta}) q(\boldsymbol{\gamma}) \log \left\{ \frac{h(\boldsymbol{\beta}, \xi) \pi(\boldsymbol{\beta} | \boldsymbol{\gamma}, \sigma) \pi(\boldsymbol{\gamma} | \theta) \pi(\theta) \pi(\sigma)}{q(\boldsymbol{\beta}) q(\boldsymbol{\gamma})} \right\} d\boldsymbol{\beta} d\boldsymbol{\gamma}. \end{aligned}$$

The goal is to maximize the objective function $\tilde{\mathcal{L}}$:

[Update of $q_{\boldsymbol{\beta}}$]

$$\begin{aligned} \log q(\boldsymbol{\beta}) &= \mathbb{E}_{-\boldsymbol{\beta}} \log [h(\boldsymbol{\beta}, \xi) \pi(\boldsymbol{\beta} | \boldsymbol{\gamma}, \sigma)] + C \\ &= \sum_{i=1}^n \boldsymbol{\beta}^T \frac{1}{2} y_i X_i - \sum_{i=1}^n \lambda(\xi_i) (\boldsymbol{\beta}^T X_i)^2 - \frac{1}{2\sigma^2} \boldsymbol{\beta}^T \text{diag}(\Delta_i) \boldsymbol{\beta} \\ &= -\frac{1}{2} \boldsymbol{\beta}^T (2 \sum_{i=1}^n \lambda(\xi_i) X_i X_i^T + \text{diag}(\Delta_j)/\sigma^2) \boldsymbol{\beta} + \boldsymbol{\beta}^T \sum_{i=1}^n \frac{1}{2} y_i X_i \end{aligned}$$

So $q(\boldsymbol{\beta}) = \mathbf{N}(\mathbf{m}, D)$ with

$$D^{-1} \mathbf{m} = \sum_{i=1}^n \frac{1}{2} y_i X_i, \quad D^{-1} = 2 \sum_{i=1}^n \lambda(\xi_i) X_i X_i^T + \text{diag}(\Delta_j)/\sigma^2,$$

which is equivalent as the updating formula in Section 5.2.2 if we let $w_i = 2\lambda(\xi_i)$.

[Update of ξ_i]

$$\begin{aligned}
\tilde{\mathcal{L}}(\cdots, \xi_i) &= \mathbb{E}_{\boldsymbol{\beta}} \log h(\boldsymbol{\beta}, \xi) + C \\
&= \log \zeta(\xi_i) - \xi_i/2 - \lambda(\xi_i)(\mathbb{E}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^T X_i)^2 - \xi_i^2) \\
&= -\log(1 + e^{-\xi_i}) - \frac{\xi_i}{2} - \frac{1}{2\xi_i} \left[\frac{1}{1 + e^{-\xi_i}} - \frac{1}{2} \right] (\mathbb{E}_{\boldsymbol{\beta}}(\boldsymbol{\beta}^T X_i)^2 - \xi_i^2)
\end{aligned}$$

After some calculation,

$$\xi_i^2 = X_i^T (D + \mathbf{m}\mathbf{m}^T) X_i,$$

which is the same as $\mathbb{E}z_i^2$ in Section 5.2.2.

To conclude, for the variational algorithm, the local approximation method with a local parameter ξ_i is equivalent to the Polya-Gamma method with a latent variable w_i , where the local parameter ξ_i plays the same role as the second Polya-Gamma parameter c_i .

5.3 Empirical Results

Dense feature example (glmnet).

We use the example from the glmnet paper (Hastie and Qian, 2014) to compare our method with LASSO on the logistic model. Suppose there are n samples and p random variables $\mathbf{x}_1, \dots, \mathbf{x}_p$ that are generated from the standard normal distribution, with the same correlation $\rho(\mathbf{x}_i, \mathbf{x}_j) = \rho$. The coefficients in $\boldsymbol{\beta}$ are exponentially decreasing in magnitude:

$$\beta_j = (-1)^j \exp\{-2(j-1)/20\}.$$

Let $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\epsilon \sim \mathbf{N}(0, 3\mathbf{I})$ is the noise term added to the model. The response variable is generated from a Binomial distribution with probability:

$$P(\mathbf{y} = 1) = 1/(1 + e^{-\mathbf{z}})$$

In this example data, the absolute values of the first 30 coefficients drop exponentially from 1 to 0.05. Because of the property of dense feature, we focus more on the prediction efficiency rather than model accuracy. We compared our method with LASSO using glmnet R package. Two criteria

are used here to measure the prediction accuracy: the 0/1 loss function and the negative log-likelihood.

Method	0/1 Loss	Neg-log-likelihood	time (s)
$\rho = 0$:			
LVB	0.3551 (0.0173)	-629.44 (12.54)	124.10
glmnet	0.3420 (0.0159)	-629.92 (26.65)	69.05
$\rho = 0.5$:			
LVB	0.4065 (0.0248)	-666.55 (13.63)	121.68
glmnet	0.3877 (0.0157)	-673.62 (23.37)	85.90

Table 5.1: Dense feature example. The mean and standard error (in parenthesis) of the 0/1 loss and negative log-likelihood from 50 simulations with different correlation ρ are shown for both LVB in Algorithm 4 and LASSO from the glmnet package.

Appendix A

Supplementary Material of Chapter 2

A.1 Proof of Theorem 2.1

Proof. Recall the EM algorithm returns

$$\hat{\gamma}_{nj} = 1, \quad \text{if} \quad \mathbb{E}_{\beta_n | \Theta^{(t)}, \mathbf{y}} [\beta_{nj}^2] > r_n,$$

where the threshold

$$r_n = \frac{\hat{\sigma}_n^2}{1/v_0 - 1/v_1} \left(\log \frac{v_1}{v_0} - 2 \log \frac{\hat{\theta}_n}{1 - \hat{\theta}_n} \right) = O(n^{-r_0} \log n)$$

and the conditional second moment of β_{nj} is equal to $m_j^2 + \hat{\sigma}_n^2 V_{jj}$ with

$$\begin{aligned} \mathbf{m} &= (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} \mathbf{X}_n^T (\mathbf{X}_n \beta_n^* + \mathbf{e}_n) \\ &= \beta_n^* - (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} D^{-1} \beta_n^* + (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} \mathbf{X}_n^T \mathbf{e}_n \\ &= \beta_n^* - \mathbf{b}_n + \mathbf{W}_n \\ \mathbf{V} &= (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1}, \quad D^{-1} = \text{diag} \left(\frac{1 - \hat{\gamma}_{nj}}{v_0} + \frac{\hat{\gamma}_{nj}}{v_1} \right). \end{aligned}$$

Here we represent the posterior mean of β_n as three separate terms: the true coefficient vector β_n^* , the bias term $\mathbf{b}_n$ and the random error term $\mathbf{W}_n$. So the event $\{\hat{\gamma}_n = \gamma_n^*\}$ is equivalent to

$$\left\{ \min_{j: \gamma_{nj}^* = 1} m_j^2 + \hat{\sigma}_n^2 V_{jj} > r_n \right\} \cap \left\{ \max_{j: \gamma_{nj}^* = 0} m_j^2 + \hat{\sigma}_n^2 V_{jj} < r_n \right\}. \quad (\text{A.1})$$

First we prove the following results that quantify m_j^2 and V_{jj} .

(R1) V_{jj} is upper bounded by the largest eigenvalue of $\mathbf{V}$,

$$V_{jj} \leq \frac{1}{\lambda_{n1} + 1/v_1} = O(n^{-\eta_1}) \prec O(n^{-r_0} \log n) = r_n, \quad (\text{A.2})$$

where for two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \prec b_n$ if $a_n/b_n \rightarrow 0$.

(R2) The bias term $\mathbf{b}_n$ is bounded by

$$\begin{aligned} \max_j |b_{nj}| \leq \|\mathbf{b}_n\|_2 &\leq \|(\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1}\|_2 \cdot \|D^{-1} \boldsymbol{\beta}_n^*\|_2 \\ &\leq \frac{1/v_0}{\lambda_{n1} + 1/v_1} \|\boldsymbol{\beta}_n^*\|_2 = O(n^{r_0 - \eta_1 + \eta_2}). \quad (\text{A.3}) \end{aligned}$$

When $r_0 < 2(\eta_1 - \eta_2)/3$, $\max_j |b_{nj}|^2 \prec O(n^{-r_0} \log n) = r_n$.

The matrix L_2 norm is defined as $\|A\|_2 = \sup_{\|v\|=1} \|Av\|_2$, which is equal to its largest eigenvalue (singular value) when A is symmetric (non-symmetric).

(R3) Note that $\mathbf{W}_n$ is not a Gaussian random vector due to the dependence between D and $\mathbf{e}_n$, but it can be rewritten as

$$\mathbf{W}_n = (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} (\mathbf{X}_n^T \mathbf{X}_n) (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{e}_n = A \tilde{\mathbf{W}}_n.$$

where $A = (\mathbf{X}_n^T \mathbf{X}_n + D^{-1})^{-1} (\mathbf{X}_n^T \mathbf{X}_n)$ and $\tilde{\mathbf{W}}_n = (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{e}_n$. Since A is a matrix with norm bounded by 1, we have

$$\max_j |W_{nj}| \leq \|A\|_\infty \max_j |\tilde{W}_{nj}| \leq \sqrt{p} \|A\|_2 \max_j |\tilde{W}_{nj}| \leq \sqrt{p} \max_j |\tilde{W}_{nj}|.$$

(R4) $\tilde{\mathbf{W}}_n = (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^T \mathbf{e}_n$ is a Gaussian random vector with covariance $\sigma^2 (\mathbf{X}_n^T \mathbf{X}_n)^{-1}$ and mean $\mathbf{0}$. So the variance for W_{nj} is upper bounded by $\sigma^2 \lambda_{n1}^{-1}$.

Recall the tail bound for Gaussian variables: for any $Z \sim \mathbf{N}(0, \tau^2)$,

$$\mathbb{P}(|Z| > t) = \mathbb{P}(|Z|/\tau > t/\tau) \leq \frac{\tau}{t} e^{-\frac{t^2}{2\tau^2}}.$$

With Result (R3) and Bonferroni's inequality, we can find a constant $M > 0$ such that

$$\begin{aligned} \mathbb{P}(\max_j |W_{nj}| > \sqrt{r_n}) &\leq \mathbb{P}(\max_j |\tilde{W}_{nj}| > \sqrt{r_n/p}) \\ &\leq p \cdot \mathbb{P}(|\tilde{W}_{nj}| > \sqrt{r_n/p}) \\ &\leq \frac{p\sqrt{p}\sigma}{\sqrt{r_n\lambda_{n1}}} e^{-\frac{r_n\lambda_{n1}}{2p\sigma^2}} = O(e^{-Mn^{\eta_1 - r_0 - \alpha}}), \end{aligned}$$

which goes to 0 when $r_0 < \eta_1 - \alpha$. So with probability going to 1, $\max_j |W_{nj}|$ is upper bounded by $\sqrt{r_n}$.

(R5) When $1 - \eta_3 < r_0$, $\min_{j:\gamma_j^*=1} |\beta_{nj}^*|^2 \sim n^{\eta_3-1} \succ O(n^{-r_0} \log n) = r_n$.

Now we prove (A.1). Given $1 - \eta_3 < r_0 < \min\{\eta_1 - \alpha, 2(\eta_1 - \eta_2)/3\}$, we have

$$\begin{aligned} \mathbb{P}\left(\max_{j:\gamma_{nj}^*=0} (m_j^2 + \hat{\sigma}_n^2 V_{jj}) > r_n\right) &\leq \mathbb{P}\left(\left(\max_j |b_{nj}| + \max_j |W_{nj}|\right)^2 + \hat{\sigma}_n^2 \max_j V_{jj} > r_n\right) \\ &\leq \mathbb{P}\left(\max_j |W_{nj}| > \sqrt{r_n}\right) = O(e^{-Mn^{\eta_1-r_0-\alpha}}), \\ \mathbb{P}\left(\min_{j:\gamma_{nj}^*=1} (m_j^2 + \hat{\sigma}_n^2 V_{jj}) < r_n\right) &\leq \mathbb{P}\left(\min_{j:\gamma_{nj}^*=1} |\beta_{nj}^*|^2 - \left(\max_j |b_{nj}| + \max_j |W_{nj}|\right)^2 < r_n\right) \\ &\leq \mathbb{P}\left(\max_j |W_{nj}| > \sqrt{r_n}\right) = O(e^{-Mn^{\eta_1-r_0-\alpha}}). \end{aligned}$$

So (A.1) holds with probability $1 - O(e^{-Mn^{\eta_1-r_0-\alpha}}) \rightarrow 1$. $\square$

Appendix B

Supplementary Material for Chapter 3

B.1 Proof of Proposition 3.1

Rewrite the objective function J_0 in (3.9) as

$$\begin{aligned} J_0(q, \theta, \sigma^2) &= \int q(\boldsymbol{\gamma}, \boldsymbol{\beta}) \log \frac{\pi(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \mathbf{y}, \theta, \sigma^2) \pi(\theta, \sigma^2 \mid \mathbf{y})}{q(\boldsymbol{\gamma}, \boldsymbol{\beta})} d\boldsymbol{\gamma} d\boldsymbol{\beta} \\ &= \int q(\boldsymbol{\gamma}, \boldsymbol{\beta}) \log \frac{\pi(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \mathbf{y}, \theta, \sigma^2)}{q(\boldsymbol{\gamma}, \boldsymbol{\beta})} d\boldsymbol{\gamma} d\boldsymbol{\beta} + \log \pi(\theta, \sigma^2 \mid \mathbf{y}). \end{aligned}$$

Note that the first term above, a negative Kullback-Leibler distance between two distributions, is bounded above by 0, and equal to 0 if and only if the two distributions are the same (almost surely). So

$$\begin{aligned} J_0(q, \theta, \sigma^2) &\leq \log \pi(\theta, \sigma^2 \mid \mathbf{y}) \\ &\leq \max_{\theta, \sigma^2} \log \pi(\theta, \sigma^2 \mid \mathbf{y}) \\ &= \log \pi(\hat{\theta}, \hat{\sigma}^2 \mid \mathbf{y}) \end{aligned}$$

where $\hat{\theta}$ and $\hat{\sigma}^2$ are the MAP estimates defined in (3.5). Define

$$q^*(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \pi(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2),$$

and it is easy to check that

$$J_0(q^*, \hat{\theta}, \hat{\sigma}^2) = \log \pi(\hat{\theta}, \hat{\sigma}^2 \mid \mathbf{y}).$$

That is, $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2)$, along with the MAP estimates $\hat{\theta}$ and $\hat{\sigma}^2$, maximizes $J_0(q, \theta, \sigma^2)$. Therefore $\int q^*(\boldsymbol{\gamma}, \boldsymbol{\beta}) d\boldsymbol{\beta}$ is the same as $\pi(\boldsymbol{\gamma} \mid \mathbf{y}, \hat{\theta}, \hat{\sigma}^2)$.

B.2 Implementation of SAB

Stopping criteria

SAB iteratively updates $(\mathbf{m}, \mathbf{D}, \phi_{1:p}, \theta, \sigma^2)$ to minimize the objective function (3.11) until convergence. A natural stopping criterion would be based on the change of the objection function. However, evaluating the objective function in each iteration requires some additional computation that is irrelevant to the updating procedure, such as computing the determinant of $\mathbf{D}$, a $p \times p$ matrix, which increases the overall computation cost. Since our primary interest is variable selection, what matters is the convergence of the variable inclusion probabilities ϕ_j 's. Therefore we monitor the entropy of the parameter $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)$:

$$\text{Ent}(\boldsymbol{\phi}) = - \sum \phi_j \log \phi_j - \sum (1 - \phi_j) \log(1 - \phi_j), \quad (\text{B.1})$$

and stop the iteration when the change of $\text{Ent}(\boldsymbol{\phi})$ is below some pre-specified threshold.

Fast computation of matrix inverse

In SAB, at each iteration, we need to compute the inverse of a $p \times p$ matrix $\mathbf{V}_t = (\mathbf{X}^T \mathbf{X} + \boldsymbol{\Delta}_t)^{-1}$, where

$$\boldsymbol{\Delta}_t = \text{diag} \left(\frac{1 - \phi_j^{(t)}}{v_0} + \frac{\phi_j^{(t)}}{v_1} \right)_{j=1}^p$$

and $\phi_j^{(t)}$ denotes the value of ϕ_j updated by formula (3.14) at the t -th iteration.

When $p > n$, we can use the Sherman-Morrison-Woodbury formula to compute the inverse of an $n \times n$ matrix instead, namely,

$$(\mathbf{X}^T \mathbf{X} + \boldsymbol{\Delta}_t)^{-1} = \boldsymbol{\Delta}_t^{-1} - \boldsymbol{\Delta}_t^{-1} \mathbf{X}^T (\mathbf{I}_n + \mathbf{X} \boldsymbol{\Delta}_t^{-1} \mathbf{X}^T)^{-1} \mathbf{X} \boldsymbol{\Delta}_t^{-1}. \quad (\text{B.2})$$

So at each iteration, the computation cost for inverting this matrix is of the order $O(\min(n, p)^3)$, which is still time-consuming when both n and p are large.

Next we describe a sub-optimal updating rule to reduce the computation

cost. Note

$$\mathbf{V}_{t+1} = (\mathbf{X}^T \mathbf{X} + \mathbf{\Delta}_{t+1})^{-1} = (\mathbf{V}_t^{-1} + \mathbf{\Delta}_{t+1} - \mathbf{\Delta}_t)^{-1}. \quad (\text{B.3})$$

Suppose we only update those ϕ_j 's whose change is greater than a threshold (e.g., 0.01). Let l be the number of ϕ_j 's that are updated at iteration $(t+1)$. Then the matrix $\mathbf{\Delta}_{t+1} - \mathbf{\Delta}_t$ is a rank- l diagonal matrix with entries equal to zero except the l diagonal entries where $\phi_j^{t+1} \neq \phi_j^{(t)}$. So we can write

$$\mathbf{\Delta}_{t+1} - \mathbf{\Delta}_t = J_t G_l J_t^T$$

where G_l is an $l \times l$ diagonal matrix and J_t is a $p \times l$ matrix with entries 0/1.

Applying the Woodbury formula, we have

$$(\mathbf{X}^T \mathbf{X} + \mathbf{\Delta}_{t+1})^{-1} = \mathbf{V}_t - \mathbf{V}_t J_t [G_l^{-1} + J_t^T \mathbf{V}_t J_t]^{-1} J_t^T \mathbf{V}_t. \quad (\text{B.4})$$

That is, we only need to inverse an $l \times l$ matrix, so the computation cost is just $O(l^3)$.

To summarize, when the dimension n is small, we can use (B.2) to reduce the computation from $O(p^3)$ to $O(n^3)$. When both n and p are large, we can use the sub-optimal updating rule, i.e., only update ϕ_j 's whose changes between two adjacent iterations are bigger than some threshold, then we can further reduce the computation to $O(l^3)$ where l is the number of ϕ_j 's that are updated in the iteration.

Choice of v_1 and v_0

The variances of the spike-and-slab prior, v_0 and v_1 , reflect the magnitude of coefficients in the noise and the signal groups, respectively. Their values may affect the results of SAB. Like in most other variable selection algorithms, we suggest the following pre-processing procedure: center and scale the response vector $\mathbf{y}$ and each column of the design matrix $\mathbf{X}$ to have sample mean zero and sample variance one.

We suggest to choose $v_1 = 1$ as a default value. Regarding the choice of v_0 , we suggest several alternatives. The asymptotic result presented in the next section suggests to set v_0 of an order $O(n^{-\alpha})$ where $0 < \alpha < 2/3$ when p is fixed. In practice we have found that v_0 should also change with p . Based

on our experience, we suggest to choose $v_0 = 1/\sqrt{10np}$, which is the value we will use through our simulation studies. Instead of using a fixed value for v_0 , we also suggest to try a range of v_0 values and then tune v_0 via cross validation. In addition, one can plot the path for ϕ_j and m_j , the posterior probability of being non-zero and the posterior mean for the j coefficient, respectively. See Figure 3.1 and the discussions therein.

B.3 Conditions and Proofs of Theorems in Section 3.3

Conditions

Recall that $\mathbf{y}_n$ are generated from the following linear regression model

$$\mathbf{y}_n = \mathbf{X}_n \boldsymbol{\beta}_n^* + \mathbf{e}_n, \quad (\text{B.5})$$

where $\mathbf{e}_n$ is an n -dimensional random vector with each element independently distributed as $\mathbf{N}(0, \sigma^2)$, and $\boldsymbol{\beta}_n^* = (\beta_{n1}^*, \dots, \beta_{np}^*)^T$ denotes the true regression coefficients. Define the true model index $\boldsymbol{\gamma}^* = (\gamma_1^*, \dots, \gamma_p^*)^T$ as $\gamma_j^* = 0$, if $\beta_j^* = 0$, and 1, otherwise.

We focus on the case where the design matrices are given deterministic matrices. So throughout, we assume $p = p_n \leq n$, otherwise model (3.17) is not identifiable i.e., the true coefficient $\boldsymbol{\beta}_n^*$ is not uniquely determined (Shao and Deng, 2012; Fan and Lv, 2010). But the number of parameters p is still allowed to diverge to infinity when $n \rightarrow \infty$. The data $\mathbf{y}_n$, the design matrix $\mathbf{X}_n$ and the regression coefficient $\boldsymbol{\beta}_n$ form triangular arrays.

We assume the following conditions. Note that some conditions could be relaxed and we do not intend to make them minimal.

(C1) Assume λ_{n1} , the smallest eigenvalue of matrix $\mathbf{X}_n^t \mathbf{X}_n$ satisfies

$$\lambda_{n1}^{-1} = O(n^{-\eta_1}) > 0, \quad \eta_1 \leq 1.$$

This condition is imposed to control collinearity among the columns of $\mathbf{X}_n$. For example, if some columns of $\mathbf{X}_n$ are co-linear, then λ_{n1} is zero. In the finite p case, we often assume that $(\mathbf{X}_n^T \mathbf{X}_n)/n$ converges to a non-degenerate $p \times p$ matrix, so that (C1) is satisfied with $\eta_1 = 1$.

But when p diverges, the minimal eigenvalue of $(\mathbf{X}_n^T \mathbf{X}_n)$ is usually of order $o(n)$, so $\eta_1 \leq 1$.

- (C2) The L_2 norm of the true regression coefficient β_n^* satisfies the following *sparsity* condition

$$\|\beta_n^*\| = O(n^{\eta_2}), \quad \eta_1 > \eta_2 \geq 0.$$

For example, if the number of non-zero coefficients is of order $O(n^{2\eta_2})$ and each non-zero coefficient is upper bounded by some constant, then condition (C2) is satisfied. For finite p and fixed β^* , (C2) is satisfied automatically with $\eta_2 = 0$.

We require η_2 is upper bounded by η_1 . A similar condition is also imposed by Shao and Deng (2012). Different from the ordinary least square where the estimate of the regression coefficient is unbiased, SAB makes its inference on β based on ridge regression, so will encounter a bias term. This condition is needed to control the magnitude of the bias.

- (C3) The minimal non-zero coefficient satisfies

$$\liminf_n \frac{\min_{j \in S_n^*} |\beta_{nj}^*|}{n^{-(1-\eta_3)/2}} \geq M,$$

where $1 > \eta_3 \geq 0$ and $M > 0$ are two constants not depending on n .

This condition ensures there exists a gap between the magnitude of the smallest non-zero coefficient and the noise (of order $n^{-1/2}$), so we could asymptotically detect even the weakest signal. Similar conditions are also assumed for selection consistency of LASSO.

- (C4) The hyper-parameters in our priors specification are (a_0, b_0) in the Beta prior for θ and (λ, ν) in the InvGamma prior for σ^2 . Assume

$$a_0 = O(p), \quad b_0 = O(p), \quad \lambda_0 = O(n), \quad \nu_0 = O(1). \quad (\text{B.6})$$

The choice of hyper-parameters suggested at (B.6) is purely technical, which ensures that $\hat{\sigma}_n^2$ and $\hat{\theta}_n$ behave like $O_P(1)$ when both p and n go to infinity. In fact, we can plug-in any fixed constant for θ and σ^2 ,

which does not affect our proof for consistency. So this condition does not imply that (B.6) should be used as the default choice for hyper-parameters. Other choices of hyper-parameters should be fine for real applications as long as the MAP estimates do not diverge to extreme values.

Proof of Theorem 3.1

Theorem B.1. *Assume conditions (C1-C4), and $p = O(n^\alpha)$ where $0 \leq \alpha < 1$. With v_1 fixed at constant and $v_0 = O(n^{-r_0})$ with,*

$$1 - \eta_3 < r_0 < \min\{\eta_1 - \alpha, 2(\eta_1 - \eta_2)/3\}. \quad (\text{B.7})$$

Then for any constant $C_n \geq 0$ and $C_n = o(\log n)$, with probability $1 - o(e^{-Mn^{\eta_1 - r_0 - \alpha}})$, we have

$$\max_{j:\gamma_j^*=0} \log \frac{\phi_{nj}}{1 - \phi_{nj}} < -C_n \quad \text{and} \quad \min_{j:\gamma_j^*=1} \log \frac{\phi_{nj}}{1 - \phi_{nj}} > C_n, \quad (\text{B.8})$$

where M is a positive constant.

It is easy to show that (3.18) holds as a consequence of Theorem B.1 with $C_n = 0$.

Proof for Theorem B.1 Recall

$$\begin{aligned} \log \frac{\phi_{nj}}{1 - \phi_{nj}} &= \frac{1}{2} \log \frac{v_0}{v_1} + \frac{m_{nj}^2 + D_{njj}}{2\hat{\sigma}_n^2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) + \log \frac{\hat{\theta}_n}{1 - \hat{\theta}_n} \\ &= -\frac{1}{2} \log \frac{v_1}{v_0} + \left[\left(\frac{m_{nj}}{\sqrt{v_0}} \right)^2 + \frac{D_{njj}}{v_0} \right] \left(\frac{1 - v_0/v_1}{2\hat{\sigma}_n^2} \right) + \log \frac{\hat{\theta}_n}{1 - \hat{\theta}_n}, \end{aligned}$$

where m_{nj} denotes the j th element of $\mathbf{m}_n$ and D_{njj} denotes the j th diagonal entry of matrix $\mathbf{D}_n$. The updating formulae for $\mathbf{m}_n$ and $\mathbf{D}_n$ are

$$\begin{aligned} \mathbf{m}_n &= (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} \mathbf{X}_n^T (\mathbf{X}_n \boldsymbol{\beta}^* + \mathbf{e}_n) \\ &= \boldsymbol{\beta}^* - (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} \Delta_n \boldsymbol{\beta}^* + (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} \mathbf{X}_n^t \mathbf{e}_n \\ &= \boldsymbol{\beta}^* - \mathbf{b}_n + \mathbf{W}_n, \end{aligned} \quad (\text{B.9})$$

$$\mathbf{D}_n = \hat{\sigma}_n^2 (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1}, \quad \Delta_n = \text{diag} \left(\frac{\phi_{nj}}{v_1} + \frac{1 - \phi_{nj}}{v_0} \right).$$

In (B.9), we express $\mathbf{m}_n$ into three terms: the true coefficient $\boldsymbol{\beta}^*$, the bias term $\mathbf{b}_n$, and the projection of the error $\mathbf{W}_n$. Note that $\mathbf{W}_n$ no longer follows a Gaussian distribution since Δ_n is random depending on $\mathbf{e}_n$.

First we list some key results which will be used in the proof.

(R1) With condition (C4), we have $\hat{\sigma}_n^2 = 1$ and $|\log \hat{\theta}_n / (1 - \hat{\theta}_n)|$ bounded by a constant.

(R2) The diagonal entries of $\mathbf{D}_n$ can be bounded from above by its largest eigenvalue, which is further bounded above by $(\lambda_{n1} + 1/v_1)^{-1}$,

$$\frac{D_{njj}}{v_0} \leq \frac{1}{v_0} \frac{1}{\lambda_{n1} + 1/v_1} = O(n^{r_0 - \eta_1}).$$

(R3) The bias term b_{nj} is bounded by

$$\begin{aligned} \max_j \frac{|b_{nj}|}{\sqrt{v_0}} &\leq \frac{\|\mathbf{b}_n\|_2}{\sqrt{v_0}} = \frac{1}{\sqrt{v_0}} \left\| (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} \Delta_n \boldsymbol{\beta}^* \right\| \\ &\leq \frac{1}{\sqrt{v_0}} \frac{1/v_0}{\lambda_{n1} + 1/v_1} \|\boldsymbol{\beta}^*\| \\ &= O(n^{3r_0/2 - \eta_1 + \eta_2}), \end{aligned}$$

which goes to 0 when $r_0 < \frac{2}{3}(\eta_1 - \eta_2)$.

(R4) With probability tending to 1 exponentially, the maximum of $\mathbf{W}_n / \sqrt{v_0}$ is bounded above by a constant.

Although $\mathbf{W}_n$ is not a Gaussian random vector due to the dependence between Δ_n and $\mathbf{e}_n$, we can write it as an (random) affine transformation of a multivariate Gaussian random variable:

$$\begin{aligned} \mathbf{W}_n &= (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} \mathbf{X}_n^t \mathbf{e}_n \\ &= (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} (\mathbf{X}_n^T \mathbf{X}_n) (\mathbf{X}_n^T \mathbf{X}_n)^{-1} \mathbf{X}_n^t \mathbf{e}_n \\ &= (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} (\mathbf{X}_n^T \mathbf{X}_n) \tilde{\mathbf{W}}_n, \end{aligned}$$

where each $\tilde{W}_{nj}$ is a Gaussian random variable with mean zero and variance bounded by λ_{n1}^{-1} . Let $A = (\mathbf{X}_n^T \mathbf{X}_n + \Delta_n)^{-1} (\mathbf{X}_n^T \mathbf{X}_n)$. Using the result on $\|A\|_\infty$ (the infinity norm) and $\|A\|_2$ (the spectral norm),

we have

$$\max_j |W_{nj}| \leq \|A\|_\infty \max_j |\tilde{W}_{nj}| \leq \sqrt{p} \|A\|_2 \max_j |\tilde{W}_{nj}|$$

Note that $\|A\|_2$ is equal to its largest singular value which is the same as the largest eigenvalue of $(\mathbf{I}_p + (\mathbf{X}_n^T \mathbf{X}_n)^{-1/2} \Delta_n (\mathbf{X}_n^T \mathbf{X}_n)^{-1/2})^{-1}$ denoted as $(\mathbf{I}_p + L)^{-1}$. Since L is positive definite, any eigenvalue of $(\mathbf{I}_p + L)$ should be greater than 1. Therefore we have

$$\|A\|_2 = \|(\mathbf{I}_p + L)^{-1}\|_2 \leq 1.$$

Note that a Gaussian random variable $Z \sim \mathbf{N}(0, \tau^2)$ has its tail probability bounded by

$$\mathbb{P}(|Z| > t) = \mathbb{P}(|Z|/\tau > t/\tau) \leq \sqrt{\frac{2}{\pi}} \frac{\tau}{t} e^{-\frac{1}{2} \frac{t^2}{\tau^2}},$$

for any positive t . Recall that $\tilde{W}_{nj}$ is a Gaussian random variable with variance upper bounded by λ_{n1}^{-1} . So we have that for any constant $c > 0$, there exists another constant $M > 0$ such that

$$\mathbb{P}\left(\frac{\sqrt{p}|\tilde{W}_{nj}|}{\sqrt{v_0}} > c\right) \leq O\left(\sqrt{\frac{p}{\lambda_{n1}v_0}} \exp\left\{-\frac{c^2}{2} \frac{\lambda_{n1}v_0}{p}\right\}\right).$$

Apply the Bonferonni's inequality, and the fact that $\|A\|_2 \leq 1$, for any $c > 0$,

$$\begin{aligned} \mathbb{P}\left(\max_j \frac{|W_{nj}|}{\sqrt{v_0}} > c\right) &\leq \mathbb{P}\left(\max_j \frac{\sqrt{p}|\tilde{W}_{nj}|}{\sqrt{v_0}} > c\right) \\ &\leq O\left(p \sqrt{\frac{p}{\lambda_{n1}v_0}} \exp\left\{-\frac{c^2}{2} \frac{\lambda_{n1}v_0}{p}\right\}\right) \\ &= o\left(\exp\{-Mn^{\eta_1-r_0-\alpha}\}\right), \end{aligned} \tag{B.10}$$

for some constant $M < c^2/2$, which goes to 0 when $r_0 \leq \eta_1 - \alpha$.

Now we prove (B.8). For $\gamma_j^* = 0$, we have $\beta_j^* = 0$ and by results (R2) and

(R4),

$$\begin{aligned} \max_{j:\gamma_j^*=0} \left(\frac{m_{nj}^2}{v_0} + \frac{D_{njj}}{v_0} \right) &\leq \left(\max_j \frac{|b_{nj}|}{\sqrt{v_0}} + \max_j \frac{|W_{nj}|}{\sqrt{v_0}} \right)^2 + \max_j \frac{D_{njj}}{v_0} \\ &\leq O(n^{3r_0-2\eta_1+2\eta_2}) + \left(\max_j \frac{|W_{nj}|}{\sqrt{v_0}} \right)^2 + O(n^{r_0-\eta_1}) \end{aligned}$$

Under the condition $r_0 < 2(\eta_1 - \eta_2)/3$, both the squared bias term and the variance term go to zero, so there exists a constant $C > 0$ that for large enough n ,

$$\mathbb{P} \left(\max_{j:\gamma_j^*=0} \left(\frac{m_{nj}^2}{v_0} + \frac{D_{njj}}{v_0} \right) > c^2 \right) \leq \mathbb{P} \left(\max_j \frac{|W_{nj}|}{\sqrt{v_0}} > c \right) = o(\exp\{-Mn^{\eta_1-r_0-\alpha}\})$$

By results (R1), with probability $o(e^{-Mn^{\eta_1-r_0-\alpha}})$,

$$\max_{j:\gamma_j^*=0} \log \frac{\phi_{nj}}{1 - \phi_{nj}} = -\frac{1}{2} \log \frac{v_1}{v_0} + O(1) = -O(\log n) < -C_n.$$

For $\gamma_j^* = 1$, by results (R2-R4) we have

$$\begin{aligned} \min_{j:\gamma_j^*=1} \left(\frac{m_{nj}^2}{v_0} + \frac{D_{njj}}{v_0} \right) &\geq \left(\min_{j:\gamma_j^*=1} \frac{|\beta_j^*|}{\sqrt{v_0}} - \max_j \frac{|b_{nj}|}{\sqrt{v_0}} - \max_j \frac{|W_{nj}|}{\sqrt{v_0}} \right)^2 \\ &\geq O(n^{r_0-2\eta_3}) - O(n^{3r_0-2\eta_1+2\eta_2}) - \left(\max_j \frac{|W_{nj}|}{\sqrt{v_0}} \right)^2 \end{aligned}$$

Under the condition $r_0 > 2\eta_3$, the dominate order of the left hand side should be $O(n^{r_0-2\eta_3})$ which is greater than $-\frac{1}{2} \log \frac{v_1}{v_0} = -O(\log n)$, with probability $o(e^{-Mn^{\eta_1-r_0-\alpha}})$. Therefore

$$\max_{j:\gamma_j^*=1} \log \frac{\phi_{nj}}{1 - \phi_{nj}} = O(n^{r_0-2\eta_3}) > C_n.$$

Combining the two inequalities, (B.8) holds for $C_n = o(\log n)$, and we have

$$\mathbb{P}(\hat{S}_n = S_n^*) = 1 - o(p \exp\{-n^{\eta_1-r_0}\}) \rightarrow 1.$$

Proof for Theorem 3.2

Using the result that $\prod_j (1 - p_j) \geq 1 - \sum_j p_j$ where $0 < p_j < 1$, we have

$$1 - q_2(\gamma^*) \leq \sum_{j:\gamma_j^*=1} (1 - \phi_{nj}) + \sum_{j:\gamma_j^*=0} \phi_{nj}.$$

That is, the sum of probabilities on the incorrect models is upper bounded by the sum of the marginal probability of including a noise variable j or excluding a relevant variable j . When p is fixed, Theorem 1 implies strong consistency since each of the p terms on the right side goes to 0. But when $p = p_n$ diverges with n , each term going to 0 does not guarantee that their sum also goes to 0.

Theorem B.1 allows the constant C_n to vary with n , and it still holds if we let

$$C_n = \frac{1}{2} \log \frac{v_1}{v_0 n^\delta}, \quad \delta < r_0. \quad (\text{B.11})$$

Then we have

$$\begin{aligned} 1 - q_2(\gamma_n^*) &\leq p \times \max \left\{ \max_{j \in S_n^*} (1 - \phi_{nj}), \max_{j \notin S_n^*} \phi_{nj} \right\} \\ &\leq p \left(\frac{v_0 n^\delta}{v_1} \right)^{1/2} = O\left(\frac{p}{v_1 n^{(r_0 - \delta)/2}} \right), \end{aligned}$$

which only diminishes to zero for $p = O(n^\alpha) = o(n^{(r_0 - \delta)/2})$, that is $\alpha \leq (r_0 - \delta)/2$. Apparently, if we just require $p < n$, then SAB does not necessarily achieve Bayesian consistency. To fix this problem, we need to let v_1 grow with n : if $v_1 = O(n^{r_1})$ where r_1 is a constant satisfying $r_1 > 2\alpha - r_0$, then Theorem 1 still holds for C_n given at (B.11) with $\delta < r_0 + r_1 - 2\alpha$, and we have

$$\begin{aligned} 1 - q_2(\gamma^*) &\leq p \left(\frac{v_0 n^\delta}{v_1} \right)^{1/2} \\ &= O(n^{\alpha - (r_1 + r_0 - \delta)/2}) \rightarrow 0. \end{aligned}$$

Appendix C

Supplementary Material for Chapter 4

C.1 Lemma and properties

In this section we list some lemmas and properties that will be used in later proofs. Suppose $\mathbf{A}$ is a $n \times n$ matrix, $\lambda_1 \leq \dots \leq \lambda_n$ are the n eigenvalues of $\mathbf{A}$ in non-decreasing order.

- If $\mathbf{A}$ is symmetric, $\mathbf{x}$ is any unit vector, then

$$\lambda_1 \leq \mathbf{x}^T \mathbf{A} \mathbf{x} \leq \lambda_n.$$

In particular, $\lambda_1 \leq a_{ii} \leq \lambda_n$ where a_{ii} 's are the diagonal entries of $\mathbf{A}$.

- If $\mathbf{A}$ is symmetric, then for any vector $\mathbf{b}$,

$$\|\mathbf{A}\mathbf{b}\| \leq \|\mathbf{A}\| \cdot \|\mathbf{b}\| = \lambda_n \|\mathbf{b}\|.$$

Here the norm of a matrix is defined as $\|\mathbf{A}\| = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$.

- For any $t > 0$, a Gaussian random variable $Z \sim \mathcal{N}(0, \tau^2)$ has its tail probability bounded by

$$\mathbb{P}(|Z| > t) = \mathbb{P}(|Z|/\tau > t/\tau) \leq \sqrt{\frac{2}{\pi}} \frac{\tau}{t} e^{-\frac{1}{2} \frac{t^2}{\tau^2}}.$$

- Weyl's inequality. If $\mathbf{A}$ and $\mathbf{B}$ are two symmetric matrix, and $\lambda_j(\cdot)$ denote the function that returns the j^{th} smallest eigenvalue from a matrix, then

$$\begin{aligned} \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}) &\leq \lambda_{i+j-1}(\mathbf{A} + \mathbf{B}) \\ \lambda_i(\mathbf{A}) + \lambda_j(\mathbf{B}) &\geq \lambda_{i+j-n}(\mathbf{A} + \mathbf{B}). \end{aligned}$$

In particular

$$\lambda_i(\mathbf{A}) + \lambda_1(\mathbf{B}) \leq \lambda_i(\mathbf{A} + \mathbf{B}) \leq \lambda_i(\mathbf{A}) + \lambda_n(\mathbf{B}).$$

The inequalities also hold for singular values.

- Multiplicity version of Horn's inequality (Bhatia, 2001).

$$\lambda_j(\mathbf{A})\lambda_1(\mathbf{B}) \leq \lambda_j(\mathbf{AB}) \leq \lambda_j(\mathbf{A})\lambda_n(\mathbf{B}).$$

C.2 General results

Recall the update rule for the inclusion probability ϕ_j :

$$\log \frac{\phi_j}{1 - \phi_j} = \frac{1}{2} \log \frac{v_0}{v_1} + \frac{m_j^2 + D_{jj}}{2\hat{\sigma}^2} \left(\frac{1}{v_0} - \frac{1}{v_1} \right) + \log \frac{\hat{\theta}}{1 - \hat{\theta}} \quad (4.4)$$

For the sake of simplicity, assume $\hat{\sigma} = 1$, and $\hat{\theta} = 1/2$. The event $\{\hat{S}_n = S_n^*\}$ is equivalent to

$$A_n = \left\{ \min_{j \in S_n^*} [m_j^2 + D_{jj}^{(0)}] > r^{(0)} \right\} \cap \left\{ \max_{j \notin S_n^*} [m_j^2 + D_{jj}] < r^{(0)} \right\} \quad (C.1)$$

where the threshold $r^{(0)} \simeq v_0 \log \frac{v_1}{v_0} \sim n^{-r_0} \log n$.

The posterior mean and variance of the coefficient β are

$$\mathbf{m} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1},$$

where Δ is a diagonal matrix with $\Delta_j = \frac{\phi_j}{v_1} + \frac{1 - \phi_j}{v_0}$. We discuss the following scenarios in later proofs:

1. At the first iteration, initialize with everybody included, i.e. $\phi^{(0)} = (1, \dots, 1)'$, then we have $\Delta_j = 1/v_1$, which is used in Theorem 4.1 for full-rank data.
2. At the first iteration, initialize with everybody excluded, i.e. $\phi^{(0)} = (0, \dots, 0)'$, then we have $\Delta_j = 1/v_0 \sim n^{r_0}$, which is used in Theorem 4.5 for high-dimensional data.
3. At later iterations, without loss of generality assume the first q variables

are the signals. If we let $c_1 = e^{-C_n} \sim n^{(\alpha-r_0-r_1)/2}$ as the cutoff value for the partial updating scheme where α is any small positive number, then the inclusion probability $\boldsymbol{\phi}^{\text{par}} = ((1 - c_1)\mathbf{1}'_q, c_1\mathbf{1}'_{p-q})'$, with probability going to 1. So $\Delta = \text{diag}(c_t\mathbf{1}'_q, c_n\mathbf{1}'_{p-q})$, where

$$c_t = \frac{1 - c_1}{v_1} + \frac{c_1}{v_0} \sim n^{(\alpha+r_0-r_1)/2}, \quad c_n = \frac{c_1}{v_1} + \frac{1 - c_1}{v_0} \sim n^{r_0}. \quad (\text{C.2})$$

We rewrite $\mathbf{m}$ as

$$\begin{aligned} \mathbf{m} &= (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T (\mathbf{X} \boldsymbol{\beta}^* + \mathbf{e}) \\ &= \boldsymbol{\beta}^* - (\Delta^{-1} \mathbf{X}^T \mathbf{X} + \mathbf{I})^{-1} \boldsymbol{\beta}^* + (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^t \mathbf{e} \\ &= \boldsymbol{\beta}^* - \mathbf{b} + \mathbf{W}. \end{aligned} \quad (\text{C.3})$$

Here $\mathbf{m}$ is decomposed to three terms: the true coefficient $\boldsymbol{\beta}^*$, the bias term $\mathbf{b}$, and the error projection term $\mathbf{W}$. Then we have

$$\begin{aligned} \max_{j \notin S_n^*} [m_j^2 + D_{jj}] &\leq 2 \max_j |b_j|^2 + 2 \max_j |W_j|^2 + \max_j D_{jj} \\ \min_{j \in S_n^*} [m_j^2 + D_{jj}] &\geq \frac{1}{3} \min_{j \in S_n^*} |\beta_j^*|^2 - \max_j |b_j|^2 - \max_j |W_j|^2 \end{aligned} \quad (\text{C.4})$$

The idea of proving selection consistency is to find the probability of the event (C.1) by comparing the relative order of the four terms: $\max D_{jj}$, $\max |b_j|^2$, $\max |W_j|^2$, $\min_{j \in S_n^*} |\beta_j^*|^2$, with the threshold $r^{(0)}$, under different scenarios of Δ .

For high-dimensional data, assume the signal terms $\mathbf{X}_1$ is rank q . Write

$$(\mathbf{X}^T \mathbf{X} + \Delta)^{-1} = \begin{pmatrix} \mathbf{X}_1^T \mathbf{X}_1 + c_t I_q & \mathbf{X}_1^T \mathbf{X}_2 \\ \mathbf{X}_2^T \mathbf{X}_1 & \mathbf{X}_2^T \mathbf{X}_2 + c_n I_{n-q} \end{pmatrix}^{-1} = \begin{pmatrix} U_1 & U_2 \\ U_2^T & U_3 \end{pmatrix},$$

Apply the formula of block-wise inversion, we have

$$\begin{aligned} U_3 &= (\mathbf{X}_2^T (I_q - \mathbf{X}_1 A \mathbf{X}_1^T) \mathbf{X}_2 + c_n I_{p-q})^{-1} \\ U_2 &= -A \mathbf{X}_1^T \mathbf{X}_2 U_3 \\ U_1 &= A + A \mathbf{X}_1^T \mathbf{X}_2 U_3 \mathbf{X}_2^T \mathbf{X}_1 A \end{aligned}$$

where $A = (\mathbf{X}_1^T \mathbf{X}_1 + c_t I_q)^{-1}$. Then we have the following lemma.

Lemma C.1. Assume $\lambda_{\max}(P_{\mathbf{X}_1}P_{\mathbf{X}_2}) < 1-c$. Denote the largest and smallest positive eigenvalues of $\mathbf{X}^T\mathbf{X}$ are λ_{nn} and λ_{n1} respectively. Then we have

$$\begin{aligned}\lambda_{\max}(\mathbf{X}_1 A \mathbf{X}_1^T) &= \frac{\lambda_{nn}}{\lambda_{nn} + c_t} \simeq 1, \quad \sigma_{\max}(\mathbf{X}_1 A) = \frac{\sqrt{\lambda_{n1}}}{\lambda_{n1} + c_t} \simeq \frac{1}{\sqrt{\lambda_{n1}}} \\ \lambda_{\max}(\mathbf{X}_2 U_3 \mathbf{X}_2^T) &\leq \frac{\lambda_{nn}}{c\lambda_{nn} + c_n} \simeq \frac{1}{c}, \quad \sigma_{\max}(\mathbf{X}_2 U_3) \leq \frac{\sqrt{\lambda_{n1}}}{c\lambda_{n1} + c_n} \simeq \frac{1}{c\sqrt{\lambda_{n1}}} \\ \lambda_{\max}(\mathbf{X}_1 U_1 \mathbf{X}_1^T) &\leq \frac{1}{c}, \quad \sigma_{\max}(\mathbf{X}_1 U_1) \leq \frac{1}{\sqrt{\lambda_{n1}}} + \frac{1}{c\sqrt{\lambda_{n1}}} \simeq \frac{1}{c\sqrt{\lambda_{n1}}}\end{aligned}$$

Proof. First, $\lambda_{\max}(A) = (\lambda_{n1} + c_t)^{-1} \simeq \lambda_{n1}^{-1}$.

Write the singular value decomposition $\mathbf{X}_1 = P_1 D_1 Q_1^T$, and $\mathbf{X}_2 = P_2 D_2 Q_2^T$, where D_1 and D_2 are the $q \times q$ and $r \times r$ diagonal matrices respectively and r is the rank of $\mathbf{X}_2$, then

$$\begin{aligned}\mathbf{X}_1 A \mathbf{X}_1^T &= \mathbf{X}_1 (\mathbf{X}_1^T \mathbf{X}_1 + c_t I_q)^{-1} \mathbf{X}_1^T \\ &= P_1 D_1 Q_1^T (Q_1 D_1^2 Q_1^T + c_t I_q)^{-1} Q_1 D_1 P_1^T \\ &= P_1 (I_q + c_t D_1^{-2})^{-1} P_1^T\end{aligned}$$

So the matrix $\mathbf{X}_1 A \mathbf{X}_1^T$ has exactly q non-zero eigenvalues and $\lambda_{\max}(\mathbf{X}_1 A \mathbf{X}_1^T) = (1 + c_t/\lambda_{nn})^{-1} \leq 1$. By similar derivation, we can prove $\sigma_{\max}(\mathbf{X}_1 A) = \sqrt{\lambda_{n1}}/(\lambda_{n1} + c_t) \simeq \lambda_{n1}^{-1/2}$, while the details are omitted.

From this result, we have $\mathbf{X}_1 A \mathbf{X}_1^T \leq I_q$, therefore

$$U_3 = (\mathbf{X}_2^T (I_q - \mathbf{X}_1 A \mathbf{X}_1^T) \mathbf{X}_2 + c_n I_{p-q})^{-1} \leq c_n^{-1} I_{p-q},$$

so that $\lambda_{\max}(U_3) = c_n^{-1}$.

$$\begin{aligned}\mathbf{X}_2 U_3 \mathbf{X}_2^T &= \mathbf{X}_2 (\mathbf{X}_2^T (I - \mathbf{X}_1 A \mathbf{X}_1^T) \mathbf{X}_2 + c_n I_{n-q})^{-1} \mathbf{X}_2^T \\ &= P_2 D_2 Q_2^T (Q_2 D_2 P_2^T (I - \mathbf{X}_1 A \mathbf{X}_1^T) P_2 D_2 Q_2^T + c_n I)^{-1} Q_2 D_2 P_2^T \\ &= P_2 (P_2^T (I_n - \mathbf{X}_1 A \mathbf{X}_1^T) P_2 + c_n D_2^{-2})^{-1} P_2^T \\ &= P_2 (I_r - P_2^T P_1 (I_q + c_t D_1^{-2})^{-1} P_1^T P_2 + c_n D_2^{-2})^{-1} P_2^T \\ &\leq P_2 (I_r - P_2^T P_1 P_1^T P_2 + c_n D_2^{-2})^{-1} P_2^T\end{aligned}$$

From condition (C5), we have $\lambda_{\max}(P_2^T P_1 P_1^T P_2) = \lambda_{\max}(P_{\mathbf{X}_1} P_{\mathbf{X}_2}) < 1 - c$ for some constant $c > 0$, so that $\lambda_{\max}(\mathbf{X}_2 U_3 \mathbf{X}_2^T) \leq (c + c_n/\lambda_{nn})^{-1} \simeq c^{-1}$. The proof on $\sigma_{\max}(\mathbf{X}_2 U_3)$ is similar and thereby omitted here for conciseness.

From $\mathbf{X}_1 U_1 \mathbf{X}_1^T = A_1 + A_1 \mathbf{X}_2 U_3 \mathbf{X}_2^T A_1$ where $A_1 = \mathbf{X}_1 A \mathbf{X}_1^T$, and $\lambda_{\max}(A_1) \simeq 1$, we have

$$\lambda_{\max}(\mathbf{X}_1 U_1 \mathbf{X}_1^T) \leq 1 + 1 \times \lambda_{\max}(\mathbf{X}_2 U_3 \mathbf{X}_2^T) \simeq \frac{1}{c}.$$

From $\mathbf{X}_1 U_1 = \mathbf{X}_1 A - A_1 \mathbf{X}_2 U_3 \mathbf{X}_2^T \mathbf{X}_1 A$, we have

$$\sigma_{\max}(\mathbf{X}_1 U_1) \leq \sigma_{\max}(\mathbf{X}_1 A) + \sigma_{\max}(A_1 \mathbf{X}_2 U_3 \mathbf{X}_2^T \mathbf{X}_1 A) \leq \frac{1}{\sqrt{\lambda_{n1}}} + \frac{1}{c\sqrt{\lambda_{n1}}} \simeq \frac{1}{c\sqrt{\lambda_{n1}}}$$

□

C.3 Proof of Theorem 4.1

In this theorem, we consider the selection consistency for the first iteration with everybody-in. So the posterior mean and variance of the coefficient $\boldsymbol{\beta}$ are

$$\mathbf{m} = (\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D} = (\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1},$$

where $c_0 = 1/v_1$ is a constant. First we decompose $\mathbf{m}$ as in (C.3) and present the following key results for the proof.

- The diagonal entries of $\mathbf{D}^{(0)}$ can be upper bounded by its largest eigenvalue,

$$D_{jj}^{(0)} \leq \lambda_{\max}((\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1}) \leq \frac{1}{\lambda_{n1} + c_0} = O(n^{-\eta_1}).$$

- Bound for the maximal of the bias term $\mathbf{b}^{(0)} = (c_0^{-1} \mathbf{X}^T \mathbf{X} + \mathbf{I}_p)^{-1} \boldsymbol{\beta}^*$:

$$\max_j |b_j^{(0)}| \leq \|\mathbf{b}^{(0)}\| \leq \frac{1}{\lambda_{n1}/c_0 + 1} \|\boldsymbol{\beta}^*\| = O(n^{\eta_2 - \eta_1})$$

- $\mathbf{W}^{(0)}$ is a Gaussian random variable with mean zero and covariance:

$$\begin{aligned} \text{Var}(\mathbf{W}^{(0)}) &= \sigma^2 (\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1} \\ &\leq \sigma^2 (\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1} \end{aligned}$$

So the variance of each $W_j^{(0)}$ can be bounded by its largest eigenvalue:

$$\text{Var}(W_j^{(0)}) \leq \sigma^2 \lambda_{\max} ((\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1}) = O(n^{-\eta_1}).$$

Using Bonferroni's inequality and the upper tail bound for Gaussian random variable, we have that for any $t > 0$,

$$\mathbb{P}\left(\max_j |W_j^{(0)}| > t\right) \leq \sum_j \mathbb{P}\left(|W_j^{(0)}| > t\right) \leq p \sqrt{\frac{2}{\pi}} \frac{\tau}{t} e^{-\frac{t^2}{2\tau^2}} \quad (\text{C.5})$$

where $\tau^2 = \text{Var}(W_j^{(0)}) = O(n^{-\eta_1})$. Let $t^2 = r^{(0)} \sim n^{-r_0} \log n$, then

$$\mathbb{P}\left(\max_j |W_j^{(0)}|^2 > r^{(0)}\right) = o(p \exp\{-n^{\eta_1 - r_0}\}).$$

With the results above and (C.4), we have

$$\begin{aligned} \max_{j \notin S_n^*} [m_j^{(0)2} + D_{jj}^{(0)}] &\leq 2 \max_j |b_j^{(0)}|^2 + 2 \max_j |W_j^{(0)}|^2 + \max_j D_{jj}^{(0)} \\ &\leq O(n^{2\eta_2 - 2\eta_1}) + \max_j |W_j^{(0)}|^2 + O(n^{-\eta_1}) \end{aligned}$$

Under the condition $r_0 < \min\{\eta_1, 2(\eta_1 - \eta_2)\}$, both the squared bias term and the variance term are smaller than the order of threshold $r^{(0)} \sim n^{-r_0} \log n$, so for large enough n ,

$$\mathbb{P}\left(\max_{j \notin S_n^*} [m_j^{(0)2} + D_{jj}^{(0)}] > r^{(0)}\right) \leq \mathbb{P}\left(\max_j |W_j^{(0)}|^2 > r^{(0)}\right) = o(p \exp\{-n^{\eta_1 - r_0}\})$$

On the other hand, with (C.4) we have

$$\begin{aligned} \min_{j \in S_n^*} [m_j^{(0)2} + D_{jj}^{(0)}] &\geq \frac{1}{3} \min_{j \in S_n^*} |\beta_j^*|^2 - \max_j |b_j^{(0)}|^2 - \max_j |W_j^{(0)}|^2 \\ &\geq O(n^{-2\eta_3}) - O(n^{2\eta_2 - 2\eta_1}) - \max_j |W_j^{(0)}|^2 \end{aligned}$$

Under the condition $r_0 > 2\eta_3$, the dominate order of the left hand side should be $O(n^{-2\eta_3})$ which is greater than the threshold $r^{(0)}$. Therefore,

$$\mathbb{P}\left(\min_{j \in S_n^*} [m_j^{(0)2} + D_{jj}^{(0)}] < r^{(0)}\right) \leq \mathbb{P}\left(\max_j |W_j^{(0)}|^2 > r^{(0)}\right) = o(p \exp\{-n^{\eta_1 - r_0}\})$$

Combine the two inequalities,

$$\mathbb{P}(\hat{S}_n = S_n^*) = 1 - o(p \exp\{-n^{\eta_1 - r_0}\}) \rightarrow 1.$$

C.4 Proof of Theorem 4.2

Recall that SAB approximates the posterior distribution over γ_n by q_2 , a product of independent $\text{Bern}(\phi_{nj})$. SAB achieves strong Bayesian consistency for variable selection if:

$$q_2(\gamma_n^*) = \left[\prod_{j: \gamma_{nj}^* = 1} \phi_{nj} \right] \left[\prod_{j: \gamma_{nj}^* = 0} (1 - \phi_{nj}) \right] \rightarrow 1$$

or equivalently $1 - q_2(\gamma_n^*) \rightarrow 0$. Using the result that $\prod_j (1 - p_j) \geq \sum_j p_j$ where $0 < p_j < 1$, we have

$$1 - q_2(\gamma^*) \leq \sum_{j \in S_n^*} (1 - \phi_{nj}) + \sum_{j \notin S_n^*} \phi_{nj}. \quad (\text{C.6})$$

That is, the sum of probabilities on the incorrect models is upper bounded by the sum of the marginal probability of including a noise variable j or excluding a relevant variable j . When p is fixed, Theorem 1 implies strong consistency since each of the p terms on the right side of (C.6) goes to 0. But when p diverges with n , each term going to 0 does not guarantee that their sum also goes to 0.

In equation (C.5), let $t = \tau * \sqrt{2 \log p}$, then the upper tail bound becomes $O(1/\sqrt{\log p})$ which goes to 0 as n goes to infinity. That means

$$\max_j |W_j^{(0)}| = O_P(n^{-\eta_1/2} \sqrt{\log p}), \quad (\text{C.7})$$

which is smaller than $\sqrt{r_0} \sim n^{-r_0/2} \sqrt{\log n}$. So the leading term in (4.15) is

$$\log \frac{\phi_j^{(1)}}{1 - \phi_j^{(1)}} \simeq \frac{1}{2} \log \frac{v_0}{v_1}, \quad j \notin S_n^*; \quad \log \frac{\phi_j^{(1)}}{1 - \phi_j^{(1)}} \simeq \frac{|\beta_j^*|^2}{2v_0}, \quad j \in S_n^*.$$

Let

$$C_n = \frac{1}{2} \log \frac{v_1}{v_0 n^\alpha}, \quad 0 < \alpha < r_0. \quad (\text{C.8})$$

Instead of using 0, if we use the new cutoff C_n for the logit of inclusion probability, the selection consistency in Theorem 4.1 still holds, i.e.

$$\mathbb{P}\left(\max_{j \notin S_n^*} \log \frac{\phi_j^{(1)}}{1 - \phi_j^{(1)}} < -C_n\right) \rightarrow 1, \quad \mathbb{P}\left(\min_{j \in S_n^*} \log \frac{\phi_j^{(1)}}{1 - \phi_j^{(1)}} > C_n\right) \rightarrow 1. \quad (\text{C.9})$$

Then we have

$$\begin{aligned} 1 - q_2(\gamma_n^*) &\leq p \times \max \left\{ \max_{j \in S_n^*} (1 - \phi_{nj}), \max_{j \notin S_n^*} \phi_{nj} \right\} \\ &\leq O_P \left(p \sqrt{\frac{v_0 n^\alpha}{v_1}} \right) = O_P \left(\frac{p}{v_1 n^{(r_0 - \alpha)/2}} \right), \end{aligned}$$

which only diminishes to zero for $p = p_n = o(n^{(r_0 - \alpha)/2})$. For example, if we just require $p \leq n$, then SAB does not achieve strong consistency. As pointed out in (Narisetty and He, 2014), to achieve strong consistency, we need to let v_1 increase with n . If we set $v_1 \sim n^{r_1}$ where r_1 is a constant satisfying $r_1 > 2 - r_0$, or more generally $p = o(n^{(r_0 + r_1)/2})$, then Theorem 4.1 still holds for C_n given at (4.14) with small enough α , and we now have

$$\begin{aligned} 1 - q_2(\gamma^*) &\leq p \times \max \left\{ \max_{j \in S_n^*} (1 - \phi_{nj}), \max_{j \notin S_n^*} \phi_{nj} \right\} \\ &= O_P \left(\frac{p}{n^{(r_1 + r_0 - \alpha)/2}} \right) \xrightarrow{P} 0. \end{aligned}$$

C.5 Proof of Theorem 4.3

We only need to prove for $t = 2$ case and then the others are followed. Without loss of generality, we assume the first q variables are the true ones. As shown in the proof for $t = 1$, the partial inclusion probability ϕ^{par} is a bi-valued vector with probability $1 - o(p \exp\{-n^{\eta_1 - r_0}\})$. Conditioning on the bi-valued ϕ^{par} , the new posterior mean and variance for β are

$$\mathbf{m}^{(1)} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D}^{(1)} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1},$$

where $\Delta = \text{diag}(c_t \mathbf{1}'_q, c_n \mathbf{1}'_{p-q})$, $c_t = \frac{1-c_1}{v_1} + \frac{c_1}{v_0} = O(n^{(\alpha + r_0 - r_1)/2})$, and $c_n = \frac{c_1}{v_1} + \frac{1-c_1}{v_0} = O(n^{r_0})$.

If we assume $r_1 > r_0 + \alpha$, then $c_t = o(1)$. The key results listed in the proof of Theorem 4.1 still hold:

- Since $(\mathbf{X}^T \mathbf{X} + \Delta)^{-1} < (\mathbf{X}^T \mathbf{X} + c_t I)^{-1}$, the diagonal entries of $\mathbf{D}^{(1)}$ can be bounded by its largest eigenvalue,

$$D_{jj}^{(1)} \leq \frac{1}{\lambda_{n1} + c_t} = O(n^{-\eta_1}).$$

- From $\Delta \boldsymbol{\beta}^* = c_t \boldsymbol{\beta}^*$, the bound for the bias term $\mathbf{b}^{(1)} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \Delta \boldsymbol{\beta}^*$ is

$$\max_j |b_j^{(1)}| \leq \|\mathbf{b}^{(1)}\| \leq \frac{c_t}{\lambda_{n1} + c_t} \|\boldsymbol{\beta}^*\| = O(n^{\eta_2 - \eta_1})$$

- $\mathbf{W}^{(1)}$ is a Gaussian random variable with mean zero and covariance:

$$\begin{aligned} \text{Var}(\mathbf{W}^{(1)}) &= \sigma^2 (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \\ &\leq \sigma^2 (\mathbf{X}^T \mathbf{X} + \Delta)^{-1}. \end{aligned}$$

So the variance of each $W_j^{(1)}$ can be bounded by its largest eigenvalue:

$$\text{Var}(W_j^{(1)}) \leq \lambda_{\max} (\sigma^2 (\mathbf{X}^T \mathbf{X} + \Delta)^{-1}) = \frac{\sigma^2}{\lambda_{n1} + c_t} = O(n^{-\eta_1}).$$

Using the similar inference as in Theorem 4.1, we can prove that after the second iteration we still stay the correct model with probability,

$$\mathbb{P}(\{\phi_j^{(2)} < c_1, j \notin S_n^*\} \wedge \{\phi_j^{(2)} > 1 - c_1, j \in S_n^*\}) = (1 - o(p \exp\{-n^{\eta_1 - r_0}\}))^2.$$

C.6 Proof of Theorem 4.4

To discuss the prediction error, we use the estimate $\tilde{\boldsymbol{\beta}} = \mathbf{m} \cdot \hat{\boldsymbol{\gamma}}$ from SAB, where the posterior mean $\mathbf{m} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^t \mathbf{y}$, and $\Delta = \text{diag}(c_t \mathbf{1}'_q, c_n \mathbf{1}'_{p-q})$, as in (C.2). From Theorem 4.3, $\boldsymbol{\phi}^{\text{par}} = ((1 - c_1) \mathbf{1}'_q, c_1 \mathbf{1}'_{p-q})'$ for the event A_n in (C.1) with probability $(1 - o(p \exp\{-n^{\eta_1 - r_0}\}))^t$ after t -th iteration. We consider the two terms of the prediction error,

$$\mathbb{E} \|\mathbf{X} \tilde{\boldsymbol{\beta}} - \mathbf{X} \boldsymbol{\beta}^*\|^2 = \mathbb{E} \|\mathbf{X} \tilde{\boldsymbol{\beta}} - \mathbf{X} \boldsymbol{\beta}^*\|^2 \mathbf{1}_{A_n} + \mathbb{E} \|\mathbf{X} \tilde{\boldsymbol{\beta}} - \mathbf{X} \boldsymbol{\beta}^*\|^2 \mathbf{1}_{A_n^c}. \quad (\text{C.10})$$

Since the probability of A_n^c going to zero exponentially fast, while $\mathbb{E} \|\mathbf{X} \tilde{\boldsymbol{\beta}} - \mathbf{X} \boldsymbol{\beta}^*\|^2$ can be bounded by polynomial of n , the second term above is bounded

by $o(1)$.

Denote the top $q \times 1$ subvector of $\mathbf{m}$ as $\hat{\boldsymbol{\beta}}_1$ and the rest $\hat{\boldsymbol{\beta}}_2$. The design matrix $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ where $\mathbf{X}_1$ represents the $n \times q$ signal variables. Then the prediction error in A_n can be written as,

$$\mathbb{E}\|\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|^2 \mathbf{1}_{A_n} = \mathbb{E}\|\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1\boldsymbol{\beta}_1^*\|^2$$

Then we work on $\mathbb{E}\|\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1\boldsymbol{\beta}_1^*\|^2$, which can be decomposed into,

$$\mathbb{E}\|\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 - \mathbf{X}_1\boldsymbol{\beta}_1^*\|^2 = \text{tr}(\mathbf{X}_1 \text{Var}(\hat{\boldsymbol{\beta}}_1) \mathbf{X}_1^T) + \|\mathbf{X}_1 \text{bias}(\hat{\boldsymbol{\beta}}_1)\|^2. \quad (\text{C.11})$$

From (C.10) and (C.11), we have

$$\mathbb{E}\|\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|^2 = \text{tr}(\mathbf{X}_1 \text{Var}(\hat{\boldsymbol{\beta}}_1) \mathbf{X}_1^T) + \|\mathbf{X}_1 \text{bias}(\hat{\boldsymbol{\beta}}_1)\|^2 + o(1).$$

For the variance part,

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \leq (\mathbf{X}^T \mathbf{X} + \Delta)^{-1}.$$

Write the first $q \times q$ submatrix of $(\mathbf{X}^T \mathbf{X} + \Delta)^{-1}$ as U_1 , so that

$$\text{tr}(\mathbf{X}_1 \text{Var}(\hat{\boldsymbol{\beta}}_1) \mathbf{X}_1^T) = \text{tr}(\mathbf{X}_1 U_1 \mathbf{X}_1^T).$$

From Lemma C.1, the variance term is bounded by,

$$\text{tr}(\mathbf{X}_1 \text{Var}(\hat{\boldsymbol{\beta}}_1) \mathbf{X}_1^T) \leq q \times \lambda_{\max}(\mathbf{X}_1 U_1 \mathbf{X}_1^T) = O(qn^{\eta_4}).$$

The bias term $\text{bias}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \Delta \boldsymbol{\beta}^*$, with $\Delta \boldsymbol{\beta}^* = c_t \boldsymbol{\beta}^*$, so we have

$$\|\mathbf{X}_1 \text{bias}(\hat{\boldsymbol{\beta}}_1)\|^2 = \|\mathbf{X}_1 U_1 c_t \boldsymbol{\beta}_1^*\|^2 \leq c_t^2 \sigma_{\max}^2(\mathbf{X}_1 U_1) \|\boldsymbol{\beta}^*\|^2 = O(n^{\alpha+r_0-r_1-\eta_1+2\eta_2}),$$

which is of $o(1)$ if we choose r_1 large enough such that $r_1 - r_0 > 2\eta_2 - \eta_1$.

C.7 Proof of Theorem 4.5

First we aim to prove that after the first iteration, i.e. with probability $1 - o(p \exp\{-n^{\eta_1 - r_0}\})$,

$$\max \left\{ \max_{j \in S_n^*} (1 - \phi_j^{(1)}), \max_{j \notin S_n^*} \phi_j^{(1)} \right\} < e^{-C_n} = O(n^{-(r_1 + r_0 - \alpha)/2}). \quad (\text{C.12})$$

So the partial inclusion probability after the first iteration is $\phi_j^{\text{par}} = 1 - c_2$, for $j \in S_n^*$, and $\phi_j^{\text{par}} = c_1$, for $j \notin S_n^*$ with probability going to 1. Then we only need to prove (4.20) for $t = 2$ case and the others are followed. Without loss of generality, we assume the first q variables are the true ones.

To prove (C.12), since we initialize with everybody excluded, the posterior mean and variance are

$$\mathbf{m}^{(0)} = (\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D}^{(0)} = (\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1},$$

where $c_0 = 1/v_0 \sim n^{r_0}$.

The similar results as in Theorem 4.1 can be obtained, but with some modifications.

- The diagonal entries of $\mathbf{D}^{(0)}$ can be upper bounded by its largest eigenvalue,

$$D_{jj}^{(0)} \leq \lambda_{\max}((\mathbf{X}^T \mathbf{X} + c_0 \mathbf{I})^{-1}) \leq c_0^{-1} = O(n^{-r_0}).$$

- Recall the singular value decomposition of $\mathbf{X} = PDQ^T$ with rank r and the true coefficient is defined as $\boldsymbol{\beta}^* = QQ^T \boldsymbol{\beta}^*$. The bias term can be written as,

$$\begin{aligned} \mathbf{b}^{(0)} &= (c_0^{-1} \mathbf{X}^T \mathbf{X} + \mathbf{I}_p)^{-1} \boldsymbol{\beta}^* \\ &= (c_0^{-1} Q D^2 Q^T + \mathbf{I}_p)^{-1} Q Q^T \boldsymbol{\beta}^* \\ &= Q (c_0^{-1} D^2 + \mathbf{I}_r)^{-1} Q^T \boldsymbol{\beta}^* \end{aligned}$$

So the bound for the bias term b_j is:

$$\max_j |b_j^{(0)}| \leq \|\mathbf{b}^{(0)}\| \leq \frac{1}{\lambda_{n1}/c_0 + 1} \|\boldsymbol{\beta}^*\| = O(n^{r_0 + \eta_2 - \eta_1})$$

- $\mathbf{W}^{(0)}$ is a Gaussian random variable with mean zero and covariance:

$$\begin{aligned}
\text{Var}(\mathbf{W}^{(0)}) &= \sigma^2(\mathbf{X}^T\mathbf{X} + c_0\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + c_0\mathbf{I})^{-1} \\
&= \sigma^2(QD^2Q^T + c_0\mathbf{I}_p)^{-1}QD^2Q^T(QD^2Q^T + c_0\mathbf{I}_p)^{-1} \\
&= \sigma^2Q(D^2 + c_0\mathbf{I}_r)^{-1}D^2(D^2 + c_0\mathbf{I}_r)^{-1}Q^T \\
&= \sigma^2Q\text{diag}\left(\frac{d_j^2}{(d_j^2 + c_0)^2}\right)Q^T
\end{aligned}$$

So the variance of each $W_j^{(0)}$ can be bounded by its largest eigenvalue:

$$\text{Var}(W_j^{(0)}) \leq \sigma^2 \frac{\lambda_{n1}}{(\lambda_{n1} + c_0)^2} = O(n^{-\eta_1}).$$

Similar to that in Theorem 4.1, using Bonferroni's inequality and the upper tail bound for Gaussian random variable, we have

$$\mathbb{P}\left(\max_j |W_j^{(0)}|^2 > r^{(0)}\right) = o(p \exp\{-n^{\eta_1 - r_0}\}).$$

From the result above and similar to (C.4), we have,

$$\begin{aligned}
\max_{j \notin S_n^*} [m_j^{(0)2} + D_{jj}^{(0)}] &\leq \left(\max_{j \notin S_n^*} |\beta_j^*| + \max_j |b_j^{(0)}| + \max_j |W_j^{(0)}|\right)^2 + \max_j D_{jj}^{(0)} \\
&\leq a_n^2 + O(n^{2r_0 + 2\eta_2 - 2\eta_1}) + \max_j |W_j^{(0)}|^2 + O(n^{-r_0})
\end{aligned}$$

Under the condition $r_0 < 2(\eta_1 - \eta_2)/3$, and $a_n^2 \sim r^{(0)}/(1 + 1/\log \log n)$, we have $r^{(0)} - a_n^2 = r^{(0)}/(1 + \log \log n) \geq n^{-r_0}$, and thus

$$\mathbb{P}\left(\max_{j \notin S_n^*} [m_j^{(0)2} + D_{jj}^{(0)}] > r^{(0)}\right) \leq \mathbb{P}\left(\max_j |W_j^{(0)}|^2 > \frac{r^{(0)}}{1 + \log \log n}\right) = o(p \exp\{-n^{\eta_1 - r_0}\}).$$

On the other hand, we have

$$\begin{aligned}
\min_{j \in S_n^*} [m_j^{(0)2} + D_{jj}^{(0)}] &\geq \left(\min_{j \in S_n^*} |\beta_j^*| - \max_j |b_j^{(0)}| - \max_j |W_j^{(0)}|\right)^2 \\
&\geq b_n^2 - O(n^{2r_0 + 2\eta_2 - 2\eta_1}) - \max_j |W_j^{(0)}|^2
\end{aligned}$$

Under the condition $b_n^2 \sim r^{(0)}(1 + 1/\log \log n)$, we have $b_n^2 - r^{(0)} = r^{(0)}/\log \log n \geq$

n^{-r_0} , thus

$$\mathbb{P}\left(\min_{j:\gamma_j^*=1} [m_j^{(0)2} + D_{jj}^{(0)}] < r^{(0)}\right) \leq \mathbb{P}\left(\max_j |W_j^{(0)}|^2 > \frac{r^{(0)}}{\log \log n}\right) = o(p \exp\{-n^{\eta_1 - r_0}\})$$

Combining the two inequalities, we have proved (C.12).

Next we show the selection consistency for $t = 2$. Conditioning on the bivalued ϕ^{par} , the new posterior mean and variance for β are

$$\mathbf{m}^{(1)} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{y}, \quad \mathbf{D}^{(1)} = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1},$$

where $\Delta = \text{diag}(c_t \mathbf{1}'_q, c_n \mathbf{1}'_{p-q})$, $c_t = \frac{1-c_2}{v_1} + \frac{c_2}{v_0} = O(n^{(\alpha+r_0-r_1)/2})$, $c_n = \frac{c_1}{v_1} + \frac{1-c_1}{v_0} = O(n^{r_0})$.

From Lemma C.1 we get the following key results for the proof.

- The diagonal entries of $\mathbf{D}^{(1)}$ can be upper bounded by its largest eigenvalue. Here we separate the first q diagonal entries from the rest.

$$\begin{aligned} D_{jj}^{(1)} &\leq \lambda_{\max}(U_1) \leq \frac{1}{\lambda_{n1}} + \frac{1}{\lambda_{n1}c} = O(n^{\eta_4 - \eta_1}) \leq O(n^{-r_0}), \quad 1 \leq j \leq q \\ D_{jj}^{(1)} &\leq \lambda_{\max}(U_3) = \frac{1}{c_n} = O(n^{-r_0}), \quad q < j \leq p. \end{aligned}$$

- Note that $\Delta \beta^* = c_n \beta^* + (c_t - c_n)(\beta_1^{*T}, \mathbf{0}_{p-q}^T)^T$, so the bias term is

$$\begin{aligned} \mathbf{b}^{(1)} &= (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \Delta \beta^* \\ &= c_n (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \beta^* + (c_t - c_n) (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} (\beta_1^{*T}, \mathbf{0}_{p-q}^T)^T \\ &= c_n (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \beta^* + (c_t - c_n) (\beta_1^{*T} U_1, \beta_1^{*T} U_2)^T \end{aligned}$$

where the first term can be bounded by

$$\begin{aligned} \|c_n (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \beta^*\| &\leq \|c_n (\mathbf{X}^T \mathbf{X} + c_t I_p)^{-1} Q Q^T \beta^*\| \\ &= \|c_n Q (D^2 + c_t I_r)^{-1} Q^T \beta^*\| \\ &= \frac{c_n}{\lambda_{n1} + c_t} \|\beta^*\| = O(n^{r_0 + \eta_2 - \eta_1}) \end{aligned}$$

The second term can be divided into two cases,

$$\begin{aligned}\max_{1 \leq j \leq q} |b_j^{(1)}| &\leq \|(c_t - c_n)\boldsymbol{\beta}_1^{*T} U_1\| \leq \frac{c_n}{c\lambda_{n1}} \|\boldsymbol{\beta}^*\| = O(n^{r_0 + \eta_2 + \eta_4 - \eta_1}) \\ \max_{q < j \leq p} |b_j^{(1)}| &\leq \|(c_t - c_n)\boldsymbol{\beta}_1^{*T} U_2\| \leq \frac{c_n}{c\lambda_{n1}} \|\boldsymbol{\beta}^*\| = O(n^{r_0 + \eta_2 + \eta_4 - \eta_1})\end{aligned}$$

- $\mathbf{W}^{(1)}$ is a Gaussian random variable with mean zero and covariance:

$$\text{Var}(\mathbf{W}^{(1)}) = \sigma^2(\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \Delta)^{-1}$$

So the variance of each $W_j^{(1)}$ can be bounded by its largest eigenvalue or equivalently the largest singular value of $\mathbf{X}(\mathbf{X}^T \mathbf{X} + \Delta)^{-1}$ square. From Lemma C.1,

$$\begin{aligned}\text{Var}(W_j^{(1)})_{1 \leq j \leq q} &\leq \sigma_{\max}^2 (\mathbf{X}_1 U_1 + \mathbf{X}_2 U_2^T) \\ &= \sigma_{\max}^2 (\mathbf{X}_1 A + (\mathbf{X}_1 A \mathbf{X}_1^T - I) \mathbf{X}_2 U_3 \mathbf{X}_2^T \mathbf{X}_1 A) \\ &\leq \frac{1}{c^2 \lambda_{n1}} = O(n^{2\eta_4 - \eta_1}). \\ \text{Var}(W_j^{(1)})_{q < j \leq p} &\leq \sigma_{\max}^2 (\mathbf{X}_1 U_2 + \mathbf{X}_2 U_3) \\ &= \sigma_{\max}^2 ((I - \mathbf{X}_1 A \mathbf{X}_1^T) \mathbf{X}_2 U_3) \\ &\leq \frac{1}{c^2 \lambda_{n1}} = O(n^{2\eta_4 - \eta_1}).\end{aligned}$$

Using Bonferroni's inequality and the upper tail bound for Gaussian random variable,

$$\mathbb{P}\left(\max_j |W_j^{(1)}|^2 > r^{(0)}\right) = O(p \exp\{-n^{\eta_1 - 2\eta_4 - r_0}\}),$$

By similar inference as in Theorem 4.3, under the condition $r_0 < 2(\eta_1 - \eta_2 - \eta_4)/3$, we have proved that after the second iteration we still stay the correct model with probability,

$$\mathbb{P}(\{\phi_j^{(2)} < c_1, j \notin S_n^*\} \wedge \{\phi_j^{(2)} > 1 - c_1, j \in S_n^*\}) = (1 - o(p \exp\{-n^{\eta_1 - 2\eta_4 - r_0}\}))^2,$$

which goes to 0 when $r_0 < \eta_1 - 2\eta_4$ and $p = O(e^{n^{\eta_1 - 2\eta_4 - r_0}})$.

C.8 Proof of Theorem 4.6

From (C.10) and (C.11), we have

$$\mathbb{E}\|\mathbf{X}\tilde{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|^2 = \text{tr}(\mathbf{X}_1 \text{Var}(\hat{\boldsymbol{\beta}}_1) \mathbf{X}_1^T) + \|\mathbf{X}_1 \text{bias}(\hat{\boldsymbol{\beta}}_1)\|^2 + o(1).$$

First for the variance part, write

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \mathbf{X}^T \mathbf{X} (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} = B^T B$$

where $B = (\mathbf{X}_1, \mathbf{X}_2)(\mathbf{X}^T \mathbf{X} + \Delta)^{-1} = (B_1, B_2)$, and $B_1 = \mathbf{X}_1 U_1 + \mathbf{X}_2 U_2$.

$$\text{tr}(\mathbf{X}_1 \text{Var}(\hat{\boldsymbol{\beta}}_1) \mathbf{X}_1^T) = \text{tr}(\mathbf{X}_1 B_1^T B_1 \mathbf{X}_1^T) = \|\mathbf{X}_1 B_1^T\|_F^2 = \sum_j \sigma_j(\mathbf{X}_1 B_1^T)^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and $\sigma_j(\cdot)$ denotes the j -th smallest singular value. The key is to know the behavior of the singular values of matrix $\mathbf{X}_1 B_1^T$. From Lemma C.1,

$$\mathbf{X}_1 B_1^T = \mathbf{X}_1 U_1 \mathbf{X}_1^T + \mathbf{X}_1 U_2 \mathbf{X}_2^T = A_1 - A_1 \mathbf{X}_2 U_3 \mathbf{X}_2^T (I - A_1),$$

where $A_1 = \mathbf{X}_1 A \mathbf{X}_1^T$ that has q non-zero eigenvalues of $O(1)$. Next we show the largest singular value of $\mathbf{X}_2 U_3 \mathbf{X}_2^T (I - A_1)$ is $1/\sqrt{c}$.

$$\begin{aligned} \mathbf{X}_2 U_3 \mathbf{X}_2^T (I - A_1) &= \mathbf{X}_2 (\mathbf{X}_2^T (I - A_1) \mathbf{X}_2 + c_n I_{n-q})^{-1} \mathbf{X}_2^T (I - A_1) \\ &= P_2 (P_2^T (I - A_1) P_2 + c_n D_2^{-2})^{-1} P_2^T (I - A_1) \end{aligned}$$

Denote $P_0 = P_2^T (I - A_1) P_2$ so that by condition (C5) $P_0 \geq I - P_2^T P_1 P_1^T P_2 \geq c I_r$. Since $(I - A_1)^2 \leq I - A_1$, the singular value of $\mathbf{X}_2 U_3 \mathbf{X}_2^T (I - A_1)$ is bounded by

$$\begin{aligned} \sigma_{\max}^2(\mathbf{X}_2 U_3 \mathbf{X}_2^T (I - A_1)) &= \lambda_{\max}(P_2 (P_0 + c_n D_2^{-2})^{-1} P_2^T (I - A_1)^2 P_2 (P_0 + c_n D_2^{-2})^{-1} P_2^T) \\ &\leq \lambda_{\max}((P_0 + c_n D_2^{-2})^{-1} P_0 (P_0 + c_n D_2^{-2})^{-1}) \\ &\leq \lambda_{\max}(P_0^{-1}) \leq 1/c \sim n^{\eta_4} \end{aligned}$$

Therefore the variance term is of order $O(qn^{\eta_4})$.

Next look at the bias term. Since $\text{bias}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{X} + \Delta)^{-1} \Delta \boldsymbol{\beta}^*$, we can

write

$$\|\mathbf{X}_1 \text{bias}(\hat{\boldsymbol{\beta}}_1)\| = \|\mathbf{X}_1 U_1 c_t \boldsymbol{\beta}_1^*\| = c_t \|\mathbf{X}_1 U_1\| \|\boldsymbol{\beta}_1^*\| = c_t \sigma_{\max}(\mathbf{X}_1 U_1) \|\boldsymbol{\beta}_1^*\|.$$

By Lemma C.1, we have $\sigma_{\max}(\mathbf{X}_1 U_1) \leq 1/(c\sqrt{\lambda_{n1}})$. So the bias term:

$$\|\mathbf{X}_1 \text{bias}(\hat{\boldsymbol{\beta}}_1)\|^2 = O\left(\frac{c_t^2}{c^2 \lambda_{n1}} \|\boldsymbol{\beta}_1^*\|^2\right) = O(n^{\alpha+r_0-r_1+2\eta_2-\eta_1+2\eta_4}),$$

which is of $o(n^{\eta_4})$ if $r_1 > \eta_2 - r_0/2$.

Therefore, the prediction error of the truncated posterior mean is of $O(qn^{\eta_4})$.

Appendix D

Supplementary Material for Chapter 5

D.1 Objective function

$$\begin{aligned}
\tilde{\mathcal{L}} &= \mathbb{E} [\log h(\boldsymbol{\beta}, \xi) + \log \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \sigma^2) + \log \pi(\boldsymbol{\gamma} \mid \theta) - \log q(\boldsymbol{\beta}) - \log q(\boldsymbol{\gamma})] \\
&= \sum_{i=1}^n \log \zeta(\xi_i) + \mathbb{E} \sum_{i=1}^n [\boldsymbol{\beta}^T X_i y_i - (\boldsymbol{\beta}^T X_i + \xi_i)/2 - \lambda(\xi_i)((\boldsymbol{\beta}^T X_i)^2 - \xi_i^2)] \\
&\quad - \frac{1}{2} \sum_{j=1}^p \mathbb{E} \log \sigma^2 \alpha_{\gamma_j} - \frac{1}{2\sigma^2} \sum_{j=1}^p \mathbb{E} \boldsymbol{\beta}^T \text{diag}(\frac{1}{\alpha_{\gamma_j}}) \boldsymbol{\beta} + \sum_{j=1}^p \phi_j \log \theta + \sum_{j=1}^p (1 - \phi_j) \log(1 - \theta) \\
&\quad + \frac{1}{2} \log |D| - \sum_{j=1}^p \phi_j \log \phi_j - \sum_{j=1}^p (1 - \phi_j) \log(1 - \phi_j) \\
&= - \sum_{i=1}^n \log(1 + e^{-\xi_i}) + \mathbf{m}^T \sum_{i=1}^n (y_i - \frac{1}{2}) X_i - \sum_{i=1}^n \frac{\xi_i}{2} - \sum_{i=1}^n \lambda(\xi_i) (X_i^T (D + \mathbf{m} \mathbf{m}^T) X_i - \xi_i^2) \\
&\quad - \frac{p}{2} \log \sigma^2 - \frac{1}{2} \sum_{j=1}^p \mathbb{E} \log \alpha_{\gamma_j} - \frac{1}{2\sigma^2} \sum_{j=1}^p \mathbb{E} \beta_j^2 \Delta_j + \sum_{j=1}^p \phi_j \log \frac{\theta}{\phi_j} + \sum_{j=1}^p (1 - \phi_j) \log \frac{1 - \theta}{1 - \phi_j} \\
&\quad + \frac{1}{2} \log |D|
\end{aligned}$$

where $|D|$ is computed through the matrix determinant formula when $p > n$.

D.2 Local approximation

Write $\log \zeta(z) = \frac{z}{2} - \log \{e^{z/2} + e^{-z/2}\} = \frac{z}{2} + f(z)$. Note that $f(z) = -\log(e^{z/2} + e^{-z/2})$ is a convex function of variable z^2 . Since a tangent surface of convex function is a global bound for the function, we can bound $f(z)$ by

its first order Taylor expansion on z^2 :

$$\begin{aligned} f(z) &\geq f(\xi) + \frac{\partial f(\xi)}{\partial \xi^2} (z^2 - \xi^2) \\ &= -\xi/2 + \log \zeta(\xi) - \frac{1}{4\xi} \tanh(\xi/2) (z^2 - \xi^2) \end{aligned}$$

Write $\lambda(\xi) = \frac{1}{4\xi} \tanh(\xi/2) = \frac{1}{2\xi} [\zeta(\xi) - \frac{1}{2}]$, then we get

$$\zeta(z) \geq \zeta(\xi) \exp\{(z - \xi)/2 - \lambda(\xi)(z^2 - \xi^2)\}.$$

References

- Andersen, M. R., O. Winther, and L. K. Hansen (2014). Bayesian inference for structured spike and slab priors. In *Advances in Neural Information Processing Systems*, pp. 1745–1753. 12
- Barbieri, M. M. and J. O. Berger (2004). Optimal predictive model selection. *The Annals of Statistics* 32(3), 870–897. 4, 37, 43
- Barkley, L. and C. Santocanale (2013). MicroRNA-29a regulates the benzo[a]pyrene dihydrodiol epoxide-induced DNA damage response through Cdc7 kinase in lung cancer cells. *Oncogenesis* 2(7), e57. 55
- Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116(2), 281–297. 49
- Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell* 136(2), 215–233. 49
- Berger, J. O., L. R. Pericchi, J. Ghosh, T. Samanta, F. De Santis, J. Berger, and L. Pericchi (2001). Objective bayesian methods for model selection: introduction and comparison. *Lecture Notes-Monograph Series* 38, 135–207. 5
- Bhatia, R. (2001). Linear algebra to quantum cohomology: the story of alfred horn’s inequalities. *The American Mathematical Monthly* 108(4), 289–318. 96
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. 13, 39, 74, 79
- Breiman, L. (1996). Bagging predictors. *Machine Learning* 24(2), 123–140. 22
- Breiman, L. (2001). Random forests. *Machine Learning* 45(1), 5–32. 22
- Carbonetto, P., M. Stephens, et al. (2012). Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian analysis* 7(1), 73–108. 12

- Carlin, B. P. and S. Chib (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 473–484. 9
- Chipman, H., E. I. George, R. E. McCulloch, M. Clyde, D. P. Foster, and R. A. Stine (2001). The practical implementation of bayesian model selection. *Lecture Notes-Monograph Series*, 65–134. 5
- Clyde, M., H. Desimone, and G. Parmigiani (1996). Prediction via orthogonalized model mixing. *Journal of the American Statistical Association* 91(435), 1197–1208. 34
- Clyde, M. and E. I. George (2004). Model uncertainty. *Statistical Science* 19(1), 81–94. 4
- Clyde, M. A. and H. K. H. Lee (2001). Bagging and Bayesian bootstrap. In T. Richardson and T. Jaakkola (Eds.), *Artificial Intelligence and Statistics*, pp. 169–174. 15
- Creighton, C. J., A. Hernandez-Herrera, A. Jacobsen, P. M. Douglas A. Levine, N. Schultz, Y. Du, Y. Zhang, E. Larsson, R. Sheridan, W. Xiao, P. T. Spellman, G. Getz, D. A. Wheeler, C. M. Perou, R. A. Gibbs, C. Sander, D. N. Hayes, and P. H. Gunaratne (2012). Integrated analyses of microRNAs demonstrate their widespread influence on gene expression in high-grade serous ovarian carcinoma. *PLoS ONE* 7(3), e34546. 54
- Dellaportas, P., J. J. Forster, and I. Ntzoufras (2000). Bayesian variable selection using the gibbs sampler. *BIostatistics-BASEL-* 5, 273–286. 9, 57
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1), 1–38. 40
- Denison, D., B. Mallick, and A. Smith (1998). Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(2), 333–350. 5
- Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 45–97. 34
- Dweep, H., C. Sticht, P. Pandey, and N. Gretz (2011). miRWalkdatabase: prediction of possible miRNA binding sites by “walking” the genes of three genomes. *Journal of Biomedical Informatics* 44(5), 839–847. 49
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics* 32(2), 407–451. 34

- Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456), 1348–1360. vi, 2, 3, 7, 10, 25, 26, 33, 45, 56
- Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(5), 849–911. 56
- Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20(1), 101–148. 88
- Fan, J., H. Peng, et al. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* 32(3), 928–961. 56
- Fernández, C., E. Ley, and M. F. Steel (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* 100(2), 381–427. 43
- Friedman, R., K. Farh, C. Burge, and D. Bartel (2009). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research* 19(1), 92–105. 50, 52
- George, E. I. (2000). The variable selection problem. *Journal of the American Statistical Association* 95(452), 1304–1308. 4
- George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association* 88(423), 881–889. 6, 9, 15, 35, 57
- George, E. I. and R. E. McCulloch (1997). Approaches for Bayesian variable selection. *Statistica Sinica* 7(2), 339–373. 5, 34
- Green, P. J. and D. I. Hastie (2009). Reversible jump mcmc. *Genetics* 155(3), 1391–1403. 9
- Griffin, J. E. and P. J. Brown (2005). Alternative prior distributions for variable selection with very many more variables than observations. 10
- Griffiths-Jones, S., H. Saini, S. van Dongen, and A. Enright (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Research* 36(Database issue), D154–D158. 50
- Hans, C., A. Dobra, and M. West (2007). Shotgun stochastic search for large p regression. *Journal of the American Statistical Association* 102(478), 507–516. 9
- Hastie, T. and J. Qian (2014). Glmnet vignette. Technical report, Technical report, Stanford. 80

- Hernández-Lobato, D., J. M. Hernández-Lobato, and P. Dupont (2013). Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *The Journal of Machine Learning Research* 14(1), 1891–1945. 11
- Hernández-Lobato, D., J. M. Hernández-Lobato, T. Helleputte, and P. Dupont (2010). Expectation propagation for bayesian multi-task feature selection. In *Machine Learning and Knowledge Discovery in Databases*, pp. 522–537. Springer. 11
- Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12(1), 55–67. 2
- Holmes, C. C., L. Held, et al. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 1(1), 145–168. 74
- Hsu, S., F. Lin, W. Wu, C. Liang, W. Huang, W. Chan, W. Tsai, G. Chen, C. Lee, C. Chiu, C. Chien, M. Wu, C. Huang, A. Tsou, and H. Huang (2011). miRTarBase: a database curates experimentally validated microRNA- target interactions. *Nucleic Acids Research* 39(Database issue), D163–D169. 49
- Huang, J. and H. Xie (2007). Asymptotic oracle properties of scad-penalized least squares estimators. *Lecture Notes-Monograph Series*, 149–166. 56
- Hunter, D. R. and R. Li (2005). Variable selection using mm algorithms. *Annals of statistics* 33(4), 1617. 10
- Ishwaran, H. and J. S. Rao (2003). Detecting differentially expressed genes in microarrays using bayesian model selection. *Journal of the American Statistical Association* 98(462), 438–455. 6
- Jasra, A., D. A. Stephens, and C. C. Holmes (2007). Population-based reversible jump markov chain monte carlo. *Biometrika* 94(4), 787–807. 9
- Ji, Y., Y. Lu, and G. B. Mills (2008). Bayesian models based on test statistics for multiple hypothesis testing problems. *Bioinformatics* 24(7), 943–949. 51
- Krek, A., D. Grun, M. Poy, R. Wolf, L. Rosenberg, E. Epstein, P. MacMenamin, I. da Piedade, K. Gunsalus, M. Stoffel, and N. Rajewsky (2005). Combinatorial microRNA target predictions. *Nature Genetics* 37(5), 495–500. 50
- Kulkarni, A. A., S. R. Kingsbury, S. Tudzarova, H.-K. Hong, M. Loddo, M. Rashid, S. Rodriguez-Acebes, A. T. Prevost, J. A. Ledermann, K. Stoeber, and G. H. Williams (2009). Cdc7 kinase is a predictor of survival and a novel therapeutic target in epithelial ovarian carcinoma. *Clinical Cancer Research* 15(7), 2417–2425. 55

- Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The annals of mathematical statistics* 22(1), 79–86. 11
- Leva, G. D. and C. M. Croce (2013). The role of microRNAs in the tumorigenesis of ovarian cancer. *Frontiers in Oncology* 3, 153. 54
- Li, Q. and N. Lin (2010). The bayesian elastic net. *Bayesian Analysis* 5(1), 151–170. 3, 4
- Li, S., J. Zhang, Y. Wang, and X. Wan (2010). The role of microRNAs in ovarian cancer initiation and progression. *Journal of Cellular and Molecular Medicine* 14(9), 2240–2249. 54, 55
- Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for Bayesian variable selection. *Journal of American Statistical Association* 103(481), 410–423. 4, 43
- Liang, F. and W. H. Wong (2000). Evolutionary monte carlo: Applications to c p model sampling and change point problem. *Statistica sinica*, 317–342. 9
- Madigan, D., J. York, and D. Allard (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique*, 215–232. 9
- Mahajan, A., Z. Liu, L. Gellert, X. Zou, G. Yang, P. Lee, X. Yang, and J.-J. Wei (2010). HMGA2: a biomarker significantly overexpressed in high-grade ovarian serous carcinoma. *Modern Pathology* 23(5), 673–681. 54
- Meinshausen, N. and P. Bühlmann (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(4), 417–473. 16
- Minka, T. (2004). Power ep. Technical report, Technical report, Microsoft Research, Cambridge. 11
- Mitchell, T. J. and J. J. Beauchamp (1988). Bayesian variable selection in linear regression. *Journal of American Statistical Association* 83(404), 1023–1032. 4, 5, 14, 34, 57
- Mitra, R., P. Müller, S. Liang, L. Yue, and Y. Ji (2013). A Bayesian graphical model for ChIP-Seq data on histone modifications. *Journal of the American Statistical Association* 108(501), 69–80. 51
- Muniategui, A., J. Pey, F. Planes, and A. Rubio (2013). Joint analysis of miRNA and mRNA expression data. *Briefings in Bioinformatics* 14(3), 263–278. 49

- Narisetty, N. N. and X. He (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* 42(2), 789–817. 44, 57, 64, 102
- Newton, M. A., A. Noueir, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5(2), 155–176. 51
- O’Hara, R. B. and M. J. Sillanpää (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* 4(1), 85–118. 15, 34
- Papadopoulos, G., M. Reczko, V. Simossis, P. Sethupathy, and A. Hatzigeorgiou (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Research* 37(Database issue), D155–D158. 49
- Park, T. and G. Casella (2008). The bayesian lasso. *Journal of the American Statistical Association* 103(482), 681–686. 3
- Polson, N. G. and J. G. Scott (2013). Data augmentation for non-gaussian regression models using variance-mean mixtures. *Biometrika* 100(2), 549–71. 32
- Polson, N. G., J. G. Scott, and J. Windle (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association* 108(504), 1339–1349. 13, 74
- Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* 92(437), 179–191. 34
- Ročková, V. and E. I. George (2014). EMVS: The EM approach to bayesian variable selection. *Journal of the American Statistical Association* 109(506), 828–847. vi, 10, 15, 17, 28, 29, 30, 34, 35, 36, 42, 46, 47
- Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics* 9(1), 130–134. 22
- Shao, J. and X. Deng (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics* 40(2), 812–831. 56, 57, 65, 69, 70, 71, 88
- Shen, X., W. Pan, and Y. Zhu (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* 107(497), 223–232. 56

- Smith, M. and R. Kohn (1996). Nonparametric regression using bayesian variable selection. *Journal of Econometrics* 75(2), 317–343. 5
- The Cancer Genome Atlas Network (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455(7216), 1061–1068. 49
- The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490(7418), 61–70. 49
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 58(1), 267–288. 2, 25, 33, 44, 56
- Wang, H., R. Li, and C.-L. Tsai (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* 94(3), 553–568. 28
- Wang, S., B. Nan, S. Rosset, and J. Zhu (2011). Random Lasso. *The Annals of Applied Statistics* 5(1), 468–485. 16, 23, 25, 26
- Wang, Y., X. Hu, J. Greshock, L. Shen, X. Yang, Z. Shao, S. Liang, J. L. Tanyi, A. K. Sood, and L. Zhang (2012). Genomic DNA Copy-Number Alterations of the let-7 Family in Human Cancers. *PLoS ONE* 7(9), e44399. 54
- Xin, L. and M. Zhu (2012). Stochastic stepwise ensembles for variable selection. *Journal of Computational and Graphical Statistics* 21(2), 275–294. vi, 25, 27, 28
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. J. Wiley. 4
- Zellner, A. (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association* 81(394), 446–451. 4, 5
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 894–942. 3
- Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research* 7, 2541–2563. 47, 48, 65
- Zhao, P. and B. Yu (2007). Stagewise Lasso. *Journal of Machine Learning Research* 8, 2701–2726. 34
- Zhu, M. and H. A. Chipman (2006). Darwinian evolution in parallel universes: A parallel genetic algorithm for variable selection. *Technometrics* 48(4), 491–502. 16

- Zhu, Y., P. Qiu, and Y. Ji (2014). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature Methods* 11(6), 599–600. 49, 51
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* 101(476), 1418–1429. 3
- Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2), 301–320. 3
- Zou, H. and R. Li (2008). One-step sparse estimates in nonconcave penalized likelihood models (with discussion). *The Annals of Statistics* 36(4), 1509–1566. 34