Electronic Journal of Statistics
Vol. 11 (2017) 3549–3594
ISSN: 1935-7524
DOI: 10.1214/17-EJS1332

A variational Bayes approach to variable selection

John T. Ormerod

School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia
ARC Centre of Excellence for Mathematical & Statistical Frontiers

Chong You

Department of Mathematics, University of Nottingham, Ningbo China

and

Samuel Müller

School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia

Abstract: We develop methodology and theory for a mean field variational Bayes approximation to a linear model with a spike and slab prior on the regression coefficients. In particular we show how our method forces a subset of regression coefficients to be numerically indistinguishable from zero; under mild regularity conditions estimators based on our method consistently estimate the model parameters with easily obtainable and (asymptotically) appropriately sized standard error estimates; and select the true model at an exponential rate in the sample size. We also develop a practical method for simultaneously choosing reasonable initial parameter values and tuning the main tuning parameter of our algorithms which is both computationally efficient and empirically performs as well or better than some popular variable selection approaches. Our method is also faster and highly accurate when compared to MCMC.

Keywords and phrases: Mean field variational Bayes, Bernoulli-Gaussian model, Markov Chain Monte Carlo.

Received June 2017.

1. Introduction

Variable selection is one of the key problems in statistics as evidenced by papers too numerous to mention all but a small subset. Major classes of model selection approaches include criteria based procedures [1, 43, 59], penalized regression [64, 17, 20] and Bayesian modeling approaches [7, 28, 38, 62]. Despite the amount of research in the area there is yet no consensus on how to perform model selection even in the simplest case of linear regression with more observations than predictors. One of the key forces driving this research is model selection for large scale problems where the number of candidate variables is large or where the model is nonstandard in some way. Good overviews of the latest approaches to model selection are [34, 19, 47, 9, 33].

Bayesian model selection approaches have the advantage of being able to easily incorporate simultaneously many sources of variation, including prior knowledge. However, care needs to be taken in specification of the priors. From a computational perspective a careful choice of priors leads to closed form expressions for the marginal likelihood for a given model. Zellner's g-prior for the regression coefficients is usually used for this purpose [74]. The hyperparameter g for Zellner's g-prior also needs to be carefully specified in order to avoid Bartlett's paradox [4] or the information paradox [39]. For careful choices of prior on g, closed form expressions are available for the marginal likelihood for a given model (see for example 39; or 46). If the size of the model space is small then all models can be enumerated and exact Bayesian inference can be performed, otherwise Markov Chain Monte Carlo (MCMC) methods are employed. MCMC for moderate to large scale problems can be computationally inefficient. For this reason an enormous amount of effort has been put into developing MCMC and similar stochastic search based methods which can be used to explore the model space in a computationally efficient manner [51, 28, 52, 7, 38, 62].

Despite this research, MCMC can still be deemed to be too slow in practice for sufficiently large scale problems. Further drawbacks to these methods include sensitivity to prior choices, and for models with discrete random variables there are no available diagnostics to determine whether the MCMC chain has either converged or explored a sufficient proportion of highest posterior probability models in the model space.

Recently Bayesian-like methods such as the empirical Bayes approach of [44, 45] and generalized fiducial inference [36] have also been proposed. The empirical Bayes approach of [44] explores the space of models using MCMC and so can still suffer the same aforementioned drawbacks. The approach of [36] uses a combination of the sure independence screening (SIS) procedure of [18] and the Lasso [64]. However, this can fail in situations when SIS fails (if the predictors are sufficiently correlated) or the Lasso path does not contain the true model.

Mean field variational Bayes (VB) is a computationally efficient but approximate alternative to MCMC for Bayesian inference [5, 53]. While fair comparison between MCMC and VB is difficult (for reasons discussed in Section 5.1), in general VB is typically a much faster, deterministic alternative to stochastic search algorithms. However, unlike MCMC, methods based on VB cannot achieve an arbitrary accuracy in its estimation of the posterior distribution. Nevertheless, VB has shown to be an effective approach to several practical problems including document retrieval [35], functional magnetic resonance imaging [23, 50], and cluster analysis for gene-expression data [63]. Furthermore, the speed of VB in such settings gives it an advantage for exploratory data analysis where many models are typically fit to gain some understanding of the data.

A criticism often leveled at VB methods is that they often fail to provide reliable estimates of various inferential quantities, particularly posterior variances. Such criticism can be made on empirical [68, 11], or theoretical grounds [71, 58]. However, as previously shown in [73] such criticism does not hold for VB methods in general, at least in an asymptotic sense. Furthermore, variational approximation has been shown to be useful in frequentist settings [26, 27].

In this paper we consider a Bernoulli-Gaussian model to select regression coefficients [see 60]. This entails using VB to approximate the posterior distribution of indicator variables to select which variables are to be included in the model. We consider this modification to be amongst the simplest such modifications to the standard Bayesian linear model (using conjugate priors) leading to variable selection. Our main new contributions are as follows:

(i) We show how VB, for our chosen model, induces sparsity upon the regression coefficients;
(ii) We show, under mild assumptions, that our estimators for the model parameters are consistent with easily obtainable and asymptotically appropriately sized standard error estimates;
(iii) Under these same assumptions we prove that our VB method selects the true model at an exponential rate in n; and
(iv) We develop a practical method for simultaneously choosing reasonable initial parameter values and tuning the main tuning parameter of our algorithms.

Contributions (i), (ii) and (iii) are the first results of their kind for VB approaches to model selection and suggest that our approach is promising and that extensions to more complicated settings should enjoy similar properties. The VB method used in (i) is standard. Result (ii) is in keeping with consistency results of Bayesian inferential procedures [12]. However, as VB methods are inexact these results are not applicable to VB-type approximations. Contribution (iii) gives the rate of convergence, but only holds for the case where $n > p$ and p is fixed (but still possibly very large). For situations where $p > n$, our method in (iv) is empirically competitive to other methods in the simulation settings we considered. We believe that our approach will have the greatest impact for cases where $n > p$, but where it is not computationally feasible to enumerate all possible models using exact Bayesian approaches. Our empirical results also suggest that this translates to at least some problems when $p > n$.

Most papers in the literature do not consider analysis of the rates of convergence of model inclusion indicator variables, but instead consider the rate of convergence for the probability that the true model dominates a given model selection criteria. For example, [13, 44, 3], and [14] showed that the convergence to the true model is power of p or at an exponential rate in $\log p$, where the number of variables p is allowed to grow exponentially in n. [49] is one exception who also showed that model inclusion indicator variables approach their true values at an exponential rate with the sample size n and where again the number of variables p is allowed to grow exponentially in n. However, all these methods rely on a MCMC search over the parameter space and may suffer from the drawbacks of MCMC, particularly for large scale problems.

We are by no means the first to consider model selection via the use of model indicator variables within the context of variational approximation. Earlier papers which use either expectation maximization (which shares similarities with VB) or VB include [32], [55], [40], [11], [67] and [57]. However, apart from [57], there was no contribution to understand how sparsity was achieved and the later reference did not analyze the rates of convergence for their estimators. Furthermore, each of these papers considered slightly different models and tuning parameter selection approaches to those here.

Perhaps the most promising aspect of VB methodology in practice is the potential to handle non-standard complications. Examples of the flexibility of VB methods to handle such complications are contained in [42]. For example, it is not difficult to extend the methodology developed here to handle responses from elaborate distributions [68], missing data [16] or measurement error [54]. This contrasts with criteria based procedures, penalized regression and some Bayesian procedures [for example 39, 46, where the models are chosen carefully so that an exact expression for the marginal likelihood is obtainable]. For these approaches their handling of such complications will have a large computational overhead. Here we will consider the simple extension of our VB method from using Gaussian errors to Laplace distributed errors to demonstrate the flexibility of the approach.

The paper is organized as follows. Section 2 considers model selection for a linear model using a spike and slab prior on the regression coefficients and provides a motivating example from real data. Section 3 summarizes our main results which are proved in Appendix A. Section 4 discusses initialization and hyperparameter selection. Numerical examples are shown in Section 5 and illustrate the good empirical properties of our methods. We discuss our results and conclude in Section 7.

2. Bayesian linear model selection

Suppose that we have observed data $(y_i, \mathbf{x}_i)$, $1 \le i \le n$, and hypothesize that $y_i \overset{\text{ind.}}{\sim} N(\mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2)$, $1 \le i \le n$ for some coefficients $\boldsymbol{\beta}$ and noise variance σ^2 where $\mathbf{x}_i$ is p-vector of predictors. Using a binary mask, a Bayesian version of the linear regression model with conjugate prior on σ^2 may be written as,

$$\begin{array}{c} \mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim N(\mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \quad \sigma^2 \sim \text{Inverse-Gamma}(A, B), \\ \beta_j \sim N(0, \sigma_\beta^2) \quad \text{and} \quad \gamma_j \sim \text{Bernoulli}(\rho),\ j = 1, \ldots, p, \end{array} \tag{1}$$

where $\mathbf{X}$ is a $n \times p$ design matrix whose ith row is $\mathbf{x}_i^T$ (possibly including an intercept), $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ is a p-vector of regression coefficients, $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_p)$, and Inverse-Gamma$(A, B)$ is the inverse Gamma distribution with shape parameter A and scale parameter B. The parameters σ_β^2, A and B are fixed prior hyperparameters, and $\rho \in (0, 1)$ is also a hyperparameter which controls sparsity. Contrasting with [57] we use ρ rather than σ_β^2 as a tuning parameter to control sparsity. The selection of ρ (or σ_β^2 for that matter) is particularly important and is a point which we will discuss later.

In the signal processing literature this is sometimes called the Bernoulli-Gaussian [60] and is closely related to ℓ_0 regularization (see 48, Section 13.2.2) and the spike and slab prior [67]. [67] also considered what they call the Laplace-zero model where the normal distributed slab in the spike and slab is replaced with a Laplace distribution. Using their naming convention this model might also be called a normal-zero or Gaussian-zero model.

Note that the Bernoulli-Gaussian model is slightly different than the linear model with spike and slab prior. The key difference is that if $\gamma_j = 0$ for the Bernoulli-Gaussian model then $\beta_j | \mathbf{y}, \gamma_j \sim N(0, \sigma_\beta^2)$. In contrast for the spike and slab prior if $\gamma_j = 0$ then $\beta_j | \mathbf{y}, \gamma_j$ is a point mass at zero.

VB methodology is based on minimizing the Kullback-Leibler distance between the true posterior distribution and a factorized approximation to the posterior. Let $\boldsymbol{\theta}$ be the set of all model parameters and $\mathbf{d}$ be a vector of data then $p(\boldsymbol{\theta}|\mathbf{d})$ is approximated by $q(\boldsymbol{\theta}) = \prod_{k=1}^K q_k(\boldsymbol{\theta}_k)$ where $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ is a partition of $\boldsymbol{\theta}$. Then it can be shown that the optimal q_k densities, called q-densities, satisfy

$$q_k(\boldsymbol{\theta}_k) \propto \exp[\mathbb{E}_{-q_k(\boldsymbol{\theta}_k)}\{\log p(\mathbf{d}, \boldsymbol{\theta})\}], \tag{2}$$

where $\mathbb{E}_{-q_k(\boldsymbol{\theta}_k)}$ is the expectation with respect to all densities except $q_k(\boldsymbol{\theta}_k)$. For any choice of the q-densities a lower bound for the marginal likelihood for $\mathbf{d}$ can be obtained by

$$\log \underline{p}(\mathbf{d}) = \mathbb{E}_q \left[\log \left\{ \frac{p(\mathbf{d}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} \right],$$

where the underline is used to indicate the quantity is a lower bound. It can be shown that updating q_k via (2) for fixed values of the remaining q-densities results in an increase in the lower bound $\log \underline{p}(\mathbf{d})$. Cycling through the update for each k results in a monotonic increase in $\log \underline{p}(\mathbf{d})$. The resulting scheme can be interpreted as a coordinate ascent method which, under mild regularity conditions, will converge to a local maximizer of the lower bound [41].

Thus the VB approximation depends on the choice of factorization which we will now discuss. A non-exhaustive list of choices for the factorization of $q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma})$ include:

(A) $q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) = q(\boldsymbol{\beta}, \boldsymbol{\gamma}) q(\sigma^2)$;

(B) $q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) = q(\sigma^2) \displaystyle\prod_{j=1}^p q(\beta_j, \gamma_j)$; and

(C) $q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}) = q(\boldsymbol{\beta}) q(\sigma^2) \displaystyle\prod_{j=1}^p q(\gamma_j)$.

We have dropped subscripts from the q's for ease of reading.

Choice (A) leads to

$$q(\boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \exp\left[\lambda \mathbf{1}^T \boldsymbol{\gamma} - \tfrac{1}{2} \boldsymbol{\beta}^T \left(\tau \boldsymbol{\Gamma} \mathbf{X}^T \mathbf{X} \boldsymbol{\Gamma} \boldsymbol{\beta} + \sigma_\beta^{-2} \mathbf{I} \right) \boldsymbol{\beta} + \tau \boldsymbol{\beta}^T \boldsymbol{\Gamma} \mathbf{X}^T \mathbf{y} \right],$$

where $\lambda = \text{logit}(\rho) = \log(\rho/(1-\rho))$ and $\tau = \mathbb{E}_q(1/\sigma^2)$. Hence

$$\begin{aligned} q(\boldsymbol{\beta}|\boldsymbol{\gamma}) &\sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{\gamma}), \boldsymbol{\Sigma}(\boldsymbol{\gamma})) \quad \text{and} \\ q(\boldsymbol{\gamma}) &\propto \int q(\boldsymbol{\beta}, \boldsymbol{\gamma}) d\boldsymbol{\beta} = |\boldsymbol{\Sigma}(\boldsymbol{\gamma})|^{1/2} \exp\left[\lambda \mathbf{1}^T \boldsymbol{\gamma} + \tfrac{1}{2} \boldsymbol{\mu}(\boldsymbol{\gamma})^T \boldsymbol{\Sigma}(\boldsymbol{\gamma})^{-1} \boldsymbol{\mu}(\boldsymbol{\gamma}) \right] \\ &= f(\boldsymbol{\gamma}), \end{aligned}$$

where $\mathbf{\Sigma}(\boldsymbol{\gamma}) = (\tau\mathbf{\Gamma}\mathbf{X}^T\mathbf{X}\mathbf{\Gamma}\boldsymbol{\beta} + \sigma_\beta^{-2}\mathbf{I})^{-1}$ and $\boldsymbol{\mu}(\boldsymbol{\gamma}) = \mathbf{\Sigma}(\boldsymbol{\gamma})\mathbf{\Gamma}\mathbf{X}^T\mathbf{y}$. It follows that $q(\boldsymbol{\gamma}) = f(\boldsymbol{\gamma})/\sum_{\widetilde{\boldsymbol{\gamma}}\in\{0,1\}^p} f(\widetilde{\boldsymbol{\gamma}})$, where $\sum_{\widetilde{\boldsymbol{\gamma}}\in\{0,1\}^p}$ is a combinatorial sum over all 2^p possible values of $\widetilde{\boldsymbol{\gamma}}$. Then $q(\boldsymbol{\beta})$ is a mixture of normals with 2^p components given by

$$q(\boldsymbol{\beta}) = \sum_{\boldsymbol{\gamma}\in\{0,1\}^p} q(\boldsymbol{\gamma})\ \phi(\boldsymbol{\beta};\boldsymbol{\mu}(\boldsymbol{\gamma}),\mathbf{\Sigma}(\boldsymbol{\gamma})),$$

where $\phi(\boldsymbol{\beta};\boldsymbol{\mu},\mathbf{\Sigma})$ is a multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$. Calculating a combinatorial sum over 2^p terms is not computationally feasible for large p. If the sum were computationally feasible, exact Bayesian methods such as those proposed by [39] or [46] would be feasible and there would be no point in using VB. For this reason we do not pursue choice (A) here.

Choice (B) has been used by [11] who used spike and slab priors for the regression coefficients. This choice is computationally feasible for large p but underestimates the posterior variances for the regression coefficients. For this reason we will not pursue this choice of approximation here.

Choice (C) does not involve a computational sum over 2^p terms but will do better job at estimating the posterior variances of the regression coefficients by keeping all of the regression coefficients in the same partition. The remainder of the paper will explore this choice. For choice (C) the optimal q-densities are of the form

$$\begin{aligned} q^*(\boldsymbol{\beta}) &\ \text{ is a }\ N(\boldsymbol{\mu},\mathbf{\Sigma})\ \text{ density,} \\ q^*(\sigma^2) &\ \text{ is a }\ \text{Inverse-Gamma}(A+n/2,s)\ \text{ density} \\ \text{and}\quad q^*(\gamma_j) &\ \text{ is a }\ \text{Bernoulli}(w_j)\ \text{ density for } j=1,\ldots,p, \end{aligned}$$

where a necessary (but not sufficient) condition for optimality is that the following system of equations hold:

$$\mathbf{\Sigma} = \left[\tau(\mathbf{X}^T\mathbf{X})\odot\mathbf{\Omega} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1} = \left(\tau\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W} + \mathbf{D}\right)^{-1}, \tag{3}$$

$$\boldsymbol{\mu} = \tau\left(\tau\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W} + \mathbf{D}\right)^{-1}\mathbf{W}\mathbf{X}^T\mathbf{y}, \tag{4}$$

$$s = B + \frac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\mathbf{X}\mathbf{W}\boldsymbol{\mu} + \operatorname{tr}\left\{(\mathbf{X}^T\mathbf{X}\odot\mathbf{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \mathbf{\Sigma})\right\}\right] \tag{5}$$

$$\tau = \frac{2A+n}{2s} \tag{6}$$

$$\begin{aligned} \eta_j = {} & \lambda - \tfrac{\tau}{2}(\mu_j^2 + \Sigma_{j,j})\|\mathbf{X}_j\|^2 \\ & + \tau\left[\mu_j\mathbf{X}_j^T\mathbf{y} - \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}(\boldsymbol{\mu}_{-j}\mu_j + \mathbf{\Sigma}_{-j,j})\right] \end{aligned} \tag{7}$$

$$w_j = \operatorname{expit}(\eta_j) \tag{8}$$

where $1 \le j \le p$, $\operatorname{expit}(x) = \operatorname{logit}^{-1}(x) = \exp(x)/(1+\exp(x))$, $\mathbf{w} = (w_1 \ldots w_p)^T$, $\mathbf{W} = \operatorname{diag}(\mathbf{w})$, $\mathbf{\Omega} = \mathbf{w}\mathbf{w}^T + \mathbf{W}(\mathbf{I}-\mathbf{W})$, the symbol $\odot$ denotes the Hadamard product between two matrices and $\mathbf{D} = \tau(\mathbf{X}^T\mathbf{X})\odot\mathbf{W}\odot(\mathbf{I}-\mathbf{W}) + \sigma_\beta^{-2}\mathbf{I}$. Note that $\mathbf{D}$ is a diagonal matrix. Algorithm 1 below describes an iterative process for finding parameters satisfying this system of equations via a coordinate ascent scheme whose derivation can be found in Appendix A.

Note that we use the notation that for a general matrix $\mathbf{A}$, $\mathbf{A}_j$ is the jth column of $\mathbf{A}$, $\mathbf{A}_{-j}$ is $\mathbf{A}$ with the jth column removed. We write $A_{i,j}$ to be the value of the component corresponding to the ith row and jth column of $\mathbf{A}$, $\mathbf{A}_{i,-j}$ to be the vector corresponding to the ith row of $\mathbf{A}$ with the jth component removed and similarly $\mathbf{A}_{-i,j}$ to be the vector corresponding to the jth column of $\mathbf{A}$ with the ith component removed. The w_j's can be interpreted as an approximation to the posterior probability of $\gamma_j = 1$ given $\mathbf{y}$, that is, $p(\gamma_j = 1|\mathbf{y})$. Using this, our data based decision for including the jth covariate is w_j and if $w_j > 0.5$, say, we include the jth covariate in the model.

The VB approach gives rise to the lower bound

$$\log p(\mathbf{y};\rho) \quad \geq \sum_{\boldsymbol{\gamma}} \int q(\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma}) \log\left\{ \frac{p(\mathbf{y},\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma})}{q(\boldsymbol{\beta},\sigma^2,\boldsymbol{\gamma})} \right\} d\boldsymbol{\beta} d\sigma^2 \equiv \log \underline{p}(\mathbf{y};\rho)$$

where the summation is interpreted as a combinatorial sum over all possible binary configurations of $\boldsymbol{\gamma}$. At the bottom of Algorithm 1 the lower bound of $\log p(\mathbf{y};\rho)$ simplifies to

$$\begin{aligned} \log \underline{p}(\mathbf{y};\rho) \quad &= \tfrac{p}{2} - \tfrac{n}{2}\log(2\pi) - \tfrac{p}{2}\log(\sigma_\beta^2) + A\log(B) - \log\Gamma(A) \\ &\quad + \log\Gamma\left(A + \tfrac{n}{2}\right) - \left(A + \tfrac{n}{2}\right)\log(s) + \tfrac{1}{2}\log|\boldsymbol{\Sigma}| \\ &\quad - \tfrac{1}{2\sigma_\beta^2}\operatorname{tr}\left(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}\right) + \sum_{j=1}^{p}\left[w_j \log\left(\tfrac{\rho}{w_j}\right) + (1-w_j)\log\left(\tfrac{1-\rho}{1-w_j}\right)\right]. \end{aligned}$$

Let $\log \underline{p}^{(t)}(\mathbf{y};\rho)$ denote the value of the lower bound at iteration t. Algorithm 1 is terminated when the increase of the lower bound log-likelihood is negligible, that is,

$$|\log \underline{p}^{(t)}(\mathbf{y};\rho) - \log \underline{p}^{(t-1)}(\mathbf{y};\rho)| < \epsilon \qquad (9)$$

where ϵ is a small number. In our implementation we chose $\epsilon = 10^{-6}$. Note that Algorithm 1 is only guaranteed to converge to a local maximizer of this lower bound. For the $n < p$ case Algorithm 1 is efficiently implemented by calculating $\|\mathbf{y}\|^2$, $\mathbf{X}^T\mathbf{y}$ and $\mathbf{X}^T\mathbf{X}$ only once outside the main loop of the algorithm. Then each iteration of the algorithm can be implemented with cost $O(p^3)$ and storage $O(p^2)$.

To illustrate the effect of ρ on the sparsity of the VB method we consider the *prostate cancer* dataset originating from a study by [61]. The data consists of $n = 97$ samples with variables *lcavol*, *lweight* (log prostate weight), *age*, *lbph* (log of benign prostate hyperplasia amount), *svi* (seminal vesicle invasion), *lcp* (log of capsular penetration), *gleason* (Gleason score), *pgg45* (percent of Gleason scores 4 or 5), and *lpsa* (log of prostate specific antigen). [24] illustrate the effect of tuning parameter selection for ridge regression and Lasso for a linear response model using *lpsa* as the response variable and the remaining variables as predictors. We also consider the regularization paths produced by the SCAD

Algorithm 1 *Iterative scheme to obtain optimal $q^*(\boldsymbol{\theta})$ for our model.*

```
Input: $(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w})$
where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\sigma_\beta^2 > 0$, $A > 0$, $B > 0$, $\tau^{(0)} > 0$, $\rho \in (0,1)$ and $\mathbf{w}^{(1)} \in [0,1]^p$.
$t \leftarrow 1$ ; $\lambda \leftarrow \text{logit}(\rho)$
Cycle:
    $\mathbf{W}^{(t)} \leftarrow \text{diag}(\mathbf{w}^{(t)})$ ; $\boldsymbol{\Omega}^{(t)} \leftarrow \mathbf{w}^{(t)}{\mathbf{w}^{(t)}}^T + \mathbf{W}^{(t)}(\mathbf{I} - \mathbf{W}^{(t)})$
    $\boldsymbol{\Sigma}^{(t)} \leftarrow \left[\tau^{(t-1)}(\mathbf{X}^T\mathbf{X}) \odot \boldsymbol{\Omega}^{(t)} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1}$ ; $\boldsymbol{\mu}^{(t)} \leftarrow \tau^{(t-1)}\boldsymbol{\Sigma}^{(t)}\mathbf{W}^{(t)}\mathbf{X}^T\mathbf{y}$
    $s \leftarrow B + \frac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\mathbf{X}\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)} + \text{tr}\left\{(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Omega}^{(t)})(\boldsymbol{\mu}^{(t)}{\boldsymbol{\mu}^{(t)}}^T + \boldsymbol{\Sigma}^{(t)})\right\}\right]$
    $\tau^{(t)} \leftarrow (A + n/2)/s$
    $\mathbf{w}^* = [w_1^*, \ldots, w_p^*] \leftarrow \mathbf{w}^{(t)}$
     For $j = 1, \ldots, p$
        $\eta_j \leftarrow \lambda - \frac{1}{2}\tau^{(t)}\left[\left(\mu_j^{(t)}\right)^2 + \Sigma_{j,j}^{(t)}\right]\|\mathbf{X}_j\|^2$
            $+\tau^{(t)}\mathbf{X}_j^T\left[\mathbf{y}\mu_j^{(t)} - \mathbf{X}_{-j}\text{diag}(\mathbf{w}_{-j}^*)\left(\boldsymbol{\mu}_{-j}^{(t)}\mu_j^{(t)} + \boldsymbol{\Sigma}_{-j,j}^{(t)}\right)\right]$
        $w_j^* \leftarrow \text{expit}(\eta_j)$
    $\mathbf{w}^{(t+1)} \leftarrow \mathbf{w}^*$ ; $t \leftarrow t + 1$
until the increase of $\log \underline{p}(\mathbf{y}; \rho)$ is negligible.
```

penalty as implemented by the *R* package *ncvreg* [8]. These regularization paths as a function of λ are illustrated in Figure 1 where for our VB method the values of $\boldsymbol{\mu}$ (which serve as point estimates for $\boldsymbol{\beta}$) .

From Figure 1 we make several observations about the VB estimates:

(A) the estimated components of $\boldsymbol{\beta}$ appear to be stepwise functions of λ with components being either zero or constant for various ranges of λ; and

(B) large negative values of λ tend to give rise to simpler models and positive values tend to give rise to more complex models.

Note (A) holds only approximately but illustrates empirically the model selection properties of estimators obtained through Algorithm 1. This contrasts with the Lasso and other penalized regression methods where the analogous penalty parameter enforces shrinkage, and hence, bias for the estimates of non-zero coefficients. Observation (B) highlights that care is required for selecting ρ (or equivalently λ).

3. Theory

The properties of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, τ and $\{w_j\}_{1\le j\le p}$ when the system of equations (3)–(8) hold simultaneously are difficult to analyze. Instead we will analyze Algorithm 1 by examining the limiting properties of the estimators from one iteration to the next. In Appendix B we will show, under certain assumptions, the following two main results. The first result concerns the behavior of VB estimates when particular w_j's are small.

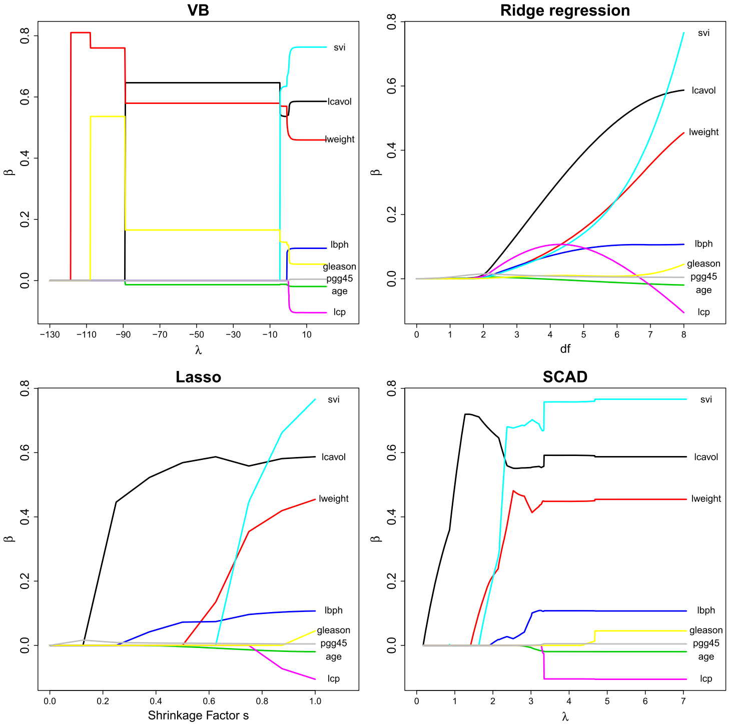

FIG 1. *Top left panel: An illustration the final values of the components of $\boldsymbol{\mu}$ for multiple runs of Algorithm 1 over a grid of $\lambda = logit(\rho)$ values where we have used $w_j^{(1)} = 1$, $j = 1, \ldots, p$, $\tau = 1000$ and hyperparameters selected as described in Section 4 on the prostate cancer dataset originating in [61]. Remaining panels: The regularization paths for Ridge, Lasso and SCAD penalized regression fits.*

Main Result 1 (Proof in Appendix B.1)**.** *Suppose that $w_j^{(t)} \ll 1$, $1 \leq j, k \leq p$. Then for observed* $\mathbf{y}$ *and* $\mathbf{X}$ *the updates in Algorithm 1 satisfy*

$$\tau^{(t)} = O(1), \qquad \mu_j^{(t)} = O(w_j^{(t)}), \qquad \Sigma_{j,k}^{(t)} = \begin{cases} \sigma_\beta^2 + O(w_j^{(t)}) & \text{if } j = k \\ O(w_j w_k) = O(w_j^{(t)}) & \text{if } j \neq k, \end{cases}$$

$$\text{and} \qquad w_j^{(t+1)} \leftarrow \text{expit}\left[\lambda - \tfrac{1}{2}\tau^{(t)}\|\mathbf{X}_j\|^2\sigma_\beta^2 + O(w_j^{(t)})\right].$$

Lemma 1 (Proof in Appendix B)**.** *Let a be a real positive number, then the quantities* $\text{expit}(-a) = \exp(-a) + O(\exp(-2a))$ *and* $\text{expit}(a) = 1 - \exp(-a) + O(\exp(-2a))$ *as $a \to \infty$.*

Remark: As a consequence of Main Result 1 and Lemma 1 we have that in Algorithm 1, if $w_j^{(t)}$ is small, updated value $w_j^{(t+1)}$ is approximately equal to $\exp(\lambda - \tau^{(t)}\|\mathbf{X}_j\|^2\sigma_\beta^2/2)$. Thus, when σ_β^2 is sufficiently large, when implemented on a computer, numerical underflow occurs and $w_j^{(t+1)}$ is represented on the computer as 0. This explains why Algorithm 1 provides sparse estimates of $\mathbf{w}$ and $\boldsymbol{\beta}$. Furthermore, all successive values of $w_j^{(T)}, T > t$ remain either small or numerically zero and may be removed safely from the algorithm, reducing the computational cost of the algorithm.

In order to establish various asymptotic properties in Main Result 2, we use the following assumptions [which are similar to those used in 73] and treat y_i and $\mathbf{x}_i$ as random quantities (only) in Main Result 2 and the proof of Main Result 2 in Appendix B.2:

(A1) for $1 \leq i \leq n$ the $y_i|\mathbf{x}_i = \mathbf{x}_i^T\boldsymbol{\beta}_0 + \varepsilon_i$ where ε_i and ε_j are independent if $i \neq j$, $\mathbb{E}(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma_0^2$ and $0 < \sigma_0^2 < \infty$, $\boldsymbol{\beta}_0$ are the true values of $\boldsymbol{\beta}$ and σ^2 with $\boldsymbol{\beta}_0$ being element-wise finite;
(A2) for $1 \leq i \leq n$ the random variables $\mathbf{x}_i \in \mathbb{R}^p$ are independent and identically distributed with p fixed;
(A3) the $p \times p$ matrix $\mathbf{S} \equiv \mathbb{E}(\mathbf{x}_i\mathbf{x}_i^T)$ is element-wise finite and $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p]$ where $\text{rank}(\mathbf{X}) = p$; and
(A4) for $1 \leq i \leq n$ the random variables $\mathbf{x}_i$ and ε_i are independent.

We view these as mild regularity conditions on the y_i's, ε_i's and the distribution of the covariates. Note that Assumption (A3) implicitly assumes that $n \geq p$. In addition to these we will assume:

(A5) for $1 \leq j, k \leq p$ the $\text{Var}(x_j x_k) < \infty$;
(A6) $\lambda \equiv \lambda_n$ varies with n, $\rho_n \equiv \text{expit}(\lambda_n)$ and satisfies $\lambda_n/n \to 0$ and $n\rho_n \to 0$ as $n \to \infty$.

Assumption (A5) will simplify later arguments, whereas Assumption (A6) is necessary for our method to identify the true model.

We now define some notation to simplify later proofs. For an indicator vector $\boldsymbol{\gamma}$ the square matrix $\mathbf{W}_{\boldsymbol{\gamma}}$ ($\mathbf{W}_{-\boldsymbol{\gamma}}$) is the principal submatrix of $\mathbf{W}$ by distinguishing (removing) rows and columns specified in $\boldsymbol{\gamma}$. The matrix $\mathbf{D}_{\boldsymbol{\gamma}}$ ($\mathbf{D}_{-\boldsymbol{\gamma}}$) is defined in the same manner. The matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ ($\mathbf{X}_{-\boldsymbol{\gamma}}$) is the submatrix of $\mathbf{X}$ by distinguishing (removing) columns specified in $\boldsymbol{\gamma}$. For example, suppose the matrix $\mathbf{X}$ has 4 columns, $\boldsymbol{\gamma} = (1, 0, 0, 1)^T$ then $\mathbf{X}_{\boldsymbol{\gamma}}$ is constructed using the first and forth columns of $\mathbf{X}$ and $\mathbf{W}_{\boldsymbol{\gamma}}$ is the submatrix of $\mathbf{W}$ consisting first and forth rows, and first and forth columns of $\mathbf{W}$. Similar notation, when indexing through a vector of indices $\mathbf{v}$, for example, if $\mathbf{v} = (1, 4)$, then $\mathbf{X}_{\mathbf{v}}$ is constructed using the first and the forth column of $\mathbf{X}$ and $\mathbf{W}_{\mathbf{v}}$ is the submatrix of $\mathbf{W}$ consisting of the first and forth rows, and the first and forth columns of $\mathbf{W}$. We rely on context to specify which notation is used. We denote $\mathbf{O}_p^v(\cdot)$ be a vector where each entry is $O_p(\cdot)$, $\mathbf{O}_p^m(\cdot)$ to be a matrix where each entry is $O_p(\cdot)$ and $\mathbf{O}_p^d(\cdot)$ be a diagonal matrix where diagonal elements are $O_p(\cdot)$. We use similar notation for $o_p(\cdot)$ matrices and vectors.

Main Result 2 (Proof in Appendix B.2). *If* $\mathbf{w}^{(1)} = \mathbf{1}$ *and assumptions (A1)-*

(A5) hold then

$$\begin{gathered}\boldsymbol{\mu}^{(1)} = \boldsymbol{\beta}_0 + \mathbf{O}_p^v\big(n^{-1/2}\big), \quad \boldsymbol{\Sigma}^{(1)} = \frac{1}{n\tau^{(0)}}\Big[\mathbb{E}\big(\mathbf{x}_i\mathbf{x}_i^T\big)\Big]^{-1} + \mathbf{O}_p^m\big(n^{-3/2}\big), \\ \tau^{(1)} = \sigma_0^{-2} + O_p\big(n^{-1/2}\big)\end{gathered} \tag{10}$$

and for $1 \leq j \leq p$ *we have*

$$w_j^{(2)} = \text{expit}\left(\eta_j^{(2)}\right) = \begin{cases} \text{expit}\left[\lambda_n + \frac{n}{2\sigma_0^2}\mathbb{E}(x_j^2)\beta_{0j}^2 + O_p\big(n^{1/2}\big)\right] & j \in \boldsymbol{\gamma}_0, \\ \text{expit}\left[\lambda_n + O_p(1)\right] & j \notin \boldsymbol{\gamma}_0. \end{cases} \tag{11}$$

If, in addition to the aforementioned assumptions, Assumption (A6) holds, then for $t = 2$ *we have*

$$\begin{gathered}\boldsymbol{\mu}_{\boldsymbol{\gamma}_0}^{(2)} = \boldsymbol{\beta}_{0,\boldsymbol{\gamma}_0} + \mathbf{O}_p^v(n^{-1/2}), \quad \boldsymbol{\mu}_{-\boldsymbol{\gamma}_0}^{(2)} \leq \exp(\lambda_n\mathbf{1} + \mathbf{O}_p^v(\log n)), \\ \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0,\boldsymbol{\gamma}_0}^{(2)} = \tfrac{\sigma_0^2}{n}\Big[\mathbb{E}\big(\mathbf{x}_i\mathbf{x}_i^T\big)\Big]_{\boldsymbol{\gamma}_0,\boldsymbol{\gamma}_0}^{-1} + \mathbf{O}_p^m\big(n^{-3/2}\big), \quad \boldsymbol{\Sigma}_{-\boldsymbol{\gamma}_0,-\boldsymbol{\gamma}_0}^{(2)} = \sigma_\beta^2\mathbf{I} + \mathbf{E}^{(2)} \\ \textit{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0,-\boldsymbol{\gamma}_0}^{(2)} \leq \exp(\lambda_n\mathbf{1} + \mathbf{O}_p^m(1)),\end{gathered} \tag{12}$$

where $\mathbf{E}^{(2)} \leq \exp(\lambda_n\mathbf{1} + \mathbf{O}_p^m(\log n))$*. For* $1 \leq j \leq p$ *we have*

$$\begin{aligned} w_j^{(3)} &= \text{expit}\left(\eta_j^{(3)}\right) \\ &= \begin{cases} \text{expit}\left[\lambda_n + \frac{n}{2\sigma_0^2}\mathbb{E}(x_j^2)\beta_{0j}^2 + O_p\big(n^{1/2}\big)\right] & j \in \boldsymbol{\gamma}_0, \\ \text{expit}\left[\lambda_n - \frac{n}{2\sigma_0^2}\mathbb{E}(x_j^2)\sigma_\beta^2 + O_p\big(n^{1/2} + n^2\,\text{expit}(\lambda_n)\big)\right] & j \notin \boldsymbol{\gamma}_0. \end{cases}\end{aligned} \tag{13}$$

For $t > 2$ *we have*

$$\begin{gathered}\boldsymbol{\mu}_{\boldsymbol{\gamma}_0}^{(t)} = \boldsymbol{\beta}_{0,\boldsymbol{\gamma}_0} + \mathbf{O}_p^v(n^{-1/2}), \\ \boldsymbol{\mu}_{-\boldsymbol{\gamma}_0}^{(t)} \leq \exp(-\tfrac{n}{2}\sigma_0^{-2}s_{\mathit{min}}\sigma_\beta^2\mathbf{1} + \mathbf{O}_p^v(\lambda_n + n^{1/2})), \\ \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0,\boldsymbol{\gamma}_0}^{(t)} = \frac{\sigma_0^2}{n}\Big[\mathbb{E}\big(\mathbf{x}_i\mathbf{x}_i^T\big)\Big]_{\boldsymbol{\gamma}_0,\boldsymbol{\gamma}_0}^{-1} + \mathbf{O}_p^m\big(n^{-3/2}\big), \\ \boldsymbol{\Sigma}_{-\boldsymbol{\gamma}_0,-\boldsymbol{\gamma}_0}^{(t)} = \sigma_\beta^2\mathbf{I} + \mathbf{E}^{(t)} \\ \textit{and} \quad \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_0,-\boldsymbol{\gamma}_0}^{(t)} \leq \exp(-\tfrac{n}{2}\sigma_0^{-2}s_{\mathit{min}}\sigma_\beta^2\mathbf{1} + \mathbf{O}_p^m(\lambda_n + n^{1/2}))\end{gathered} \tag{14}$$

where $s_{\mathit{min}} = \min_{j\in-\boldsymbol{\gamma}_0}\mathbb{E}(x_j^2)$ *and* $\mathbf{E}^{(t)} \leq \exp(-\frac{n}{2}\sigma_0^{-2}s_{\mathit{min}}\sigma_\beta^2\mathbf{1} + \mathbf{O}_p^m(\lambda_n + n^{1/2}))$*. For* $1 \leq j \leq p$ *we have*

$$\begin{aligned} w_j^{(t+1)} &= \text{expit}\left(\eta_j^{(t+1)}\right) \\ &= \begin{cases} \text{expit}\left[\lambda_n + \frac{n}{2\sigma_0^2}\mathbb{E}(x_j^2)\beta_{0j}^2 + O_p\big(n^{1/2}\big)\right] & j \in \boldsymbol{\gamma}_0, \\ \text{expit}\left[\lambda_n - \frac{n}{2\sigma_0^2}\mathbb{E}(x_j^2)\sigma_\beta^2 + O_p\big(n^{1/2}\big)\right] & j \notin \boldsymbol{\gamma}_0. \end{cases}\end{aligned} \tag{15}$$

Remark: This result suggests, under assumptions (A1)–(A6) and in light of Lemma 1, that the vector $\mathbf{w}^{(t)}$ in Algorithm 1 approaches $\boldsymbol{\gamma}_0$ at an exponential rate in n. For example, if $j \notin \gamma_0$, then

$$w_j = \text{expit}\left[n\left\{-\frac{\beta_{0j}^2}{2\sigma_0^2}\mathbb{E}(x_j^2) + \frac{\lambda_n}{n} + O_p(n^{-1/2})\right\}\right]$$

Since $\lambda_n = o(n)$, the term inside the curly brackets is negative for sufficiently large n. An application of Lemma 1 shows that $w_j \to 0$ at an exponential rate.

Remark: To get some further intuition about the stepwise shape of the VB regularization path consider (4) and (7) when $\mathbf{w} = \mathbf{1}$. In this case we can rearrange (4) to obtain

$$\tau \mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}) = \sigma_\beta^{-2}\boldsymbol{\mu}.$$

Substituting this expression after rearranging (7) we obtain

$$\begin{aligned}\eta_j &= \lambda + \tfrac{1}{2}\tau\mu_j^2\|\mathbf{X}_j\|^2 + \mu_j\tau\mathbf{X}_j^T(\mathbf{y} - \mathbf{X}\boldsymbol{\mu}) - \tfrac{1}{2}\tau\boldsymbol{\Sigma}_{jj}\|\mathbf{X}_j\|^2 - \tau\mathbf{X}_j^T\mathbf{X}_{-j}\boldsymbol{\Sigma}_{-j,j}\\ &= \lambda + \tfrac{1}{2}\tau\mu_j^2\|\mathbf{X}_j\|^2 + \tau\mu_j^2/\sigma_\beta^2 - \tfrac{1}{2}\tau\Sigma_{j,j}\|\mathbf{X}_j\|^2 - \tau\mathbf{X}_j^T\mathbf{X}_{-j}\boldsymbol{\Sigma}_{-j,j}.\end{aligned}$$

For large σ_β^2 the third term is small and $\boldsymbol{\mu}$ will be approximately equal to the least squares estimate for $\boldsymbol{\beta}$. For large n, $\|\mathbf{X}_j\|^2$ and $\mathbf{X}_j^T\mathbf{X}_{-j}$ are $O_p(n)$, and $\boldsymbol{\Sigma} = O_p(n^{-1})$. In such circumstances η_j can be approximated by the dominant terms

$$\eta_j \approx \lambda + \tfrac{1}{2}\tau\mu_j^2\|\mathbf{X}_j\|^2.$$

When λ is a sufficiently large negative constant the updated w_j will be small. Main Result 1 and Lemma 1 shows that for large σ_β^2 all successive values of w_j will be extremely small. If λ is not sufficiently large then the term $\frac{1}{2}\tau\mu_j^2\|\mathbf{X}_j^2\|^2$ is $O_p(n)$ and so the updated value of the w_js will be close to 1.

4. Hyperparameter selection and initialization

We will now briefly discuss selecting prior hyperparameters. We use $A = B = 0.01$, $\sigma_\beta^2 = 10$ and initially set $\tau = 1000$. This leaves us to choose the parameter $\rho = \text{expit}(\lambda)$ and the initial values for $\mathbf{w}$. The theory in Section 3 and 4 suggests that if we choose $\mathbf{w} = \mathbf{1}$ and say $\lambda \propto -\sqrt{n}$ and provided with enough data then Algorithm 1 will select the correct model. However, in practice this is not an effective strategy in general since Algorithm 1 may converge to a local minimum (which means $\mathbf{w}$ should be carefully selected), all values of λ satisfy Assumption (A7) when n is fixed and we do not know how much data is sufficient for our asymptotic results to guide the choice of λ.

To avoid local maxima problems, [57] used a deterministic annealing variant of the EM algorithm proposed by [65] and it was proved to be successful in that context. We instead employ a simpler stepwise procedure which initially “adds” that variable j (by setting w_j to 1 for some j) which maximizes the lower bound $\log \underline{p}(\mathbf{y};\rho)$ with $\rho = \text{expit}(-0.5\sqrt{n})$. We then,

(I) For fixed $\mathbf{w}$ select the $\rho_j = \text{expit}(\lambda_j)$ which maximizes the lower bound $\log \underline{p}(\mathbf{y};\rho_j)$ where λ_j is an equally spaced grid between -15 and 5 of 50 points.

(II) Next, for each $1 \leq j \leq p$, calculate the lower bound $\log \underline{p}(\mathbf{y};\rho)$ when w_j is set to both 0 and 1. The value w_j is set to the value which maximizes $\log \underline{p}(\mathbf{y};\rho)$ if this value exceeds the current largest $\log \underline{p}(\mathbf{y};\rho)$.

(III) Return to (I).

Algorithm 2 *Iterative scheme to tune ρ and select initial $\mathbf{w}$ for Algorithm 1*

```
Input: $(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau)$ where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $A > 0$, $B > 0$, $\sigma_\beta^2 > 0$, $\mathbf{w}_{\text{curr}} = \mathbf{0}$ and $\tau > 0$
Set $M = 100$; $P = 50$; $\rho = \text{expit}(-0.5\sqrt{n})$; $\mathcal{L} = -\infty$
For $i = 1, \ldots, \max(p, P)$
    For $j = 1, \ldots, p$
        $\mathcal{L}_j \leftarrow \log \underline{p}(\mathbf{y}; \rho)$ from Algorithm 1 with input $\big(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}_j^{(1)}\big)$
    $k \leftarrow \text{argmax}_{1 \le j \le p}\{\mathcal{L}_j\}$;   If $\mathcal{L}_k > \mathcal{L}$ then set $\mathcal{L}$ to $\mathcal{L}_k$ and $\mathbf{w}$ to $\mathbf{w}_k^{(1)}$
For $i = 1, \ldots, M$
    For $j = 1, \ldots, J$
        $\mathcal{L}_j \leftarrow \log \underline{p}(\mathbf{y}; \rho_j)$ from Algorithm 1 with input $(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho_j, \mathbf{w})$
    $k \leftarrow \text{argmax}_{1 \le j \le p}\{\mathcal{L}_j\}$;   If $\mathcal{L}_k > \mathcal{L}$ then set $\mathcal{L}$ to $\mathcal{L}_k$ and $\rho$ to $\rho_k$
    For $j = 1, \ldots, p$
        $\mathcal{L}_0 \leftarrow \log \underline{p}(\mathbf{y}; \rho)$ from Algorithm 1 with input $\big(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}_j^{(0)}\big)$
        $\mathcal{L}_1 \leftarrow \log \underline{p}(\mathbf{y}; \rho)$ from Algorithm 1 with input $\big(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}_j^{(1)}\big)$
        $k \leftarrow \text{argmax}_{j \in \{0,1\}}\{\mathcal{L}_j\}$;   If $\mathcal{L}_k > \mathcal{L}$ then set $\mathcal{L}$ to $\mathcal{L}_k$ and $\mathbf{w}$ to $\mathbf{w}_j^{(k)}$
    If $\mathcal{L}$ does not improve return output of Algorithm 1 with input $\big(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}\big)$
```

This procedure is more specifically described in Algorithm 2. Note that in Algorithm 2 we use the notation $\mathbf{w}_j^{(k)}$ to denote the vector $\mathbf{w}$ with the jth element set to k.

5. Numerical examples

In the following numerical examples we only consider simulated, but hopefully sufficiently realistic, examples in order to reliably assess the empirical qualities of different methods where truth is known. We start with situations where $p = 41$ and $n = 80$ and with $p = 99$ and $n = 2118$. These examples have $n > p$, but where it is not compuationalyl feasible to enumerate all possible models. We then look at two $p > n$ examples with $n = 500$ and $p = 1000$ and with $n = 600$ and $p = 7381$. Our methods were implemented in *R* and all code was run on the first author's laptop computer (64 bit Windows 8 Intel i7-4930MX central processing unit at 3GHz with 32GB of random access memory).

We use the mean square error (MSE) to measure the quality of the prediction error, $\text{MSE} = \frac{1}{n}\sum_{i=1}^n (\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}})_i^2$. The F_1-score [see 66] is used to assess the quality of model selection defined to be the harmonic mean between precision and recall

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where

$$\text{precision} = \frac{TP}{TP + FP} \quad \text{and} \quad \text{recall} = \frac{TP}{TP + FN},$$

with TP, FP and FN being the number of true positives, false positives and false negatives respectively. Note that F_1 is a value between 0 and 1 and higher values are being preferred. We use this measure avoid preference of the two boundary models, that is selecting non or all of the variables. The performance of our VB approach is based on Algorithm 2. We compare the performance of our VB method against the Lasso, SCAD and MCP penalized regression methods as implemented by the *R* package *ncvreg* [8]. We make use of the extended BIC [15] to choose the tuning parameter λ that minimizes

$$\text{EBIC}(\lambda) = \log(\text{RSS}_\lambda/n) + \frac{d_\lambda}{n}\left[\log(n) + 2\log(p)\right],$$

where RSS_λ is the estimated residual sum of squares $\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda\|^2$, $\widehat{\boldsymbol{\beta}}_\lambda$ is the estimated value of $\boldsymbol{\beta}$ for a particular value of λ and d_λ is the number of non-zero elements of $\widehat{\boldsymbol{\beta}}_\lambda$. [69] showed that this criterion performs well in several contexts.

We also compared our method with the Expectation Maximization Variable Selection approach (EMVS) of [57]. We used the settings that the convergence parameter ϵ equals 10^{-4} and the initial value of σ^2 equals 1. The default initial values for the regression parameters failed to converge when p is large, in such a case we specified the initial values by screening down the non-zero coefficients using Lasso, SCAD and MCP solution paths.

Finally, we compared our method with the Bayesian Model Selection (BMS) method of [21, 22]. We used the settings that 10^6 samples were used for inference after discarding a burn-in of 10^3, the hyper-g prior distribution was used with the hyperparameter equal to 3 and same initial values for the regression parameters as in the EMVS were used.

Note that for all of the simulations we center the simulated values of the response and standardize the covariates for ease of comparison with EMVS and BMS.

5.1. Comparison with MCMC for model (1)

Comparisons between VB and MCMC are fraught with difficulty. In terms of computational cost per iteration VB has a similar cost to an MCMC scheme based on Gibbs sampling. The later method has a slightly higher cost from drawing samples from a set of full conditional distributions rather than calculating approximations of them. The full conditionals corresponding to the model (1) are given by

$$\begin{array}{rl}
\boldsymbol{\beta}|\text{rest} & \sim N\left[\left(\boldsymbol{\Gamma}\mathbf{X}^T\mathbf{X}\boldsymbol{\Gamma} + \sigma^2\sigma_b^{-2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2\left(\boldsymbol{\Gamma}\mathbf{X}^T\mathbf{X}\boldsymbol{\Gamma} + \sigma^2\sigma_b^{-2}\mathbf{I}\right)^{-1}\right] \\
\sigma^2|\text{rest} & \sim \text{Inverse-Gamma}\left[A + \frac{n}{2}, B + \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2\right] \\
\gamma_j|\text{rest} & \sim \text{Bernoulli}\left[\lambda - \frac{1}{2\sigma^2}\|\mathbf{X}_j\|^2\beta_j^2 + \sigma^{-2}\beta_j\mathbf{X}_j^T\left(\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\Gamma}_{-j}\boldsymbol{\beta}_{-j}\right)\right],
\end{array} \tag{16}$$

where $1 \leq j \leq p$ for $\gamma_j|\text{rest}$. Using these Gibbs sampling can be easily implemented.

Despite the similarity between Algorithm 1 and (16) a fair comparison of these methods is difficult since choices for when each of these methods are stopped and what statistics are used to compare the outputs of each of the methods can unduly favor one method or the other. This MCMC scheme is appropriate when determining the quality of the VB method for performing Bayesian inference for model (1). We do this to compare the quality of the VB via Algorithm 1 with its Gibbs sampling counterpart (16). However, to compare model selection performance we use BMS.

Firstly, comparison is hampered by the difficultly to determine whether a MCMC scheme has converged to its stationary distribution, or in the model selection context, whether the MCMC scheme has explored a sufficient portion of the model space. Furthermore, the number of samples required to make accurate inferences may depend on the data at hand and the choice of what inferences are to be made. For these reasons both an overly large number of burn-in and total samples drawn are commonly chosen. However, by making the number of burn-in samples sufficiently large MCMC methods can be made to be arbitrarily slower than VB.

Similarly, convergence tolerances for VB trade accuracy against speed. We have chosen ϵ in (9) to be 10^{-6}. Larger values of ϵ result in cruder approximations and smaller values of ϵ are usually wasteful. Since each completion of Algorithm 1 takes very little time we are able to tune the parameter ρ via Algorithm 2. In comparison, MCMC schemes can both be sensitive to the choice of hyperparameter values and prohibitively time consuming to tune in practice.

With the above in mind we consider using (16) with identical hyperparameters and ρ selected via Algorithm 2. For each of the examples we used 10^5 MCMC samples for inference after discarding a burn-in of 10^3. No thinning was applied. For the comparisons with MCMC in addition to F_1-score and MSE we also compare the posterior density accuracy, introduced in [16], defined by

$$\text{accuracy}(\theta_j) = 100 \times \left(1 - \frac{1}{2}\int |p(\theta_j|\mathbf{y}) - q(\theta_j)|d\theta_j\right)$$

where θ_j is an arbitrary parameter and is expressed as a percentage and the mean parameter bias for the regression coefficients

$$\text{BIAS} = \frac{1}{p}\sum_{j=1}^{p}(\beta_{0j} - \widehat{\beta}_j)^2.$$

In our tables and figures the observed MSE and BIAS are reported on negative log scale (where higher values are better) and bracketed values represent standard error estimates.

5.2. Example 1: Diets simulation

We use the following example modified from [25]. Let m_1 and n be parameters of this simulation which are chosen to be integers. For this example we suppose that there are two groups of diets with $n/2$ subjects in each group. We generate m_1+1 explanatory variables as follows. First, we generate a binary diet indicator z where, for each subject $i = 1, \ldots, n$, $z_i = I(i > n/2) - I(i \leq n/2)$. Next we

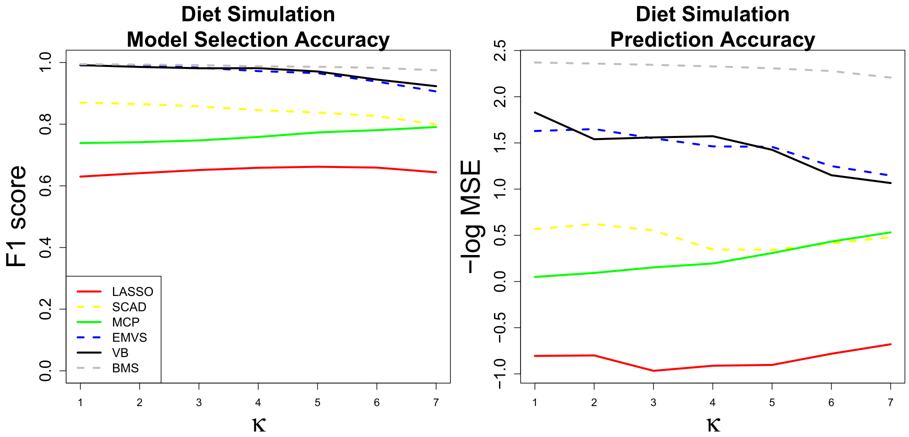

Fig 2. *Summaries of the model selection and prediction accuracies of VB, Lasso, SCAD and MCP methods for the Diet Simulation example.*

generate $\mathbf{x}_k = [x_{1,k}, \ldots, x_{n,k}]^T$, $k = 1, \ldots, m_1$, such that $x_{ik} = u_{ik} + z_i v_k$, where u_{ik} are independent uniform $(0, 1)$ random variables, $v_1, \ldots, v_{0.75m_1}$ are independent uniform $(0.25, 0.75)$ random variables, and $v_{0.75m_1+1}, \ldots, v_{m_1}$ are identically zero. Thus, we have m_1 variables, $x_1, \ldots, x_{m_1}$ where the first 75% of the x's depend on z. Finally, we generate the response vector as

$$\mathbf{y} = \beta_1 z + \beta_2 \mathbf{x}_1 + \beta_3 \mathbf{x}_2 + \beta_4 \mathbf{x}_3 + \sum_{k=5}^{m_1} \beta_k \mathbf{x}_{k-1} + \beta_{m_1+1} \mathbf{x}_{m_1} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is normally distributed with mean 0 and covariance $\sigma^2 \mathbf{I}$. For this simulation we set $m_1 = 40$, $n = 80$, $\sigma^2 = 1$, and $\boldsymbol{\beta} = (1 - (\kappa - 1)/12) \times (4.5, 3, -3, -3, \mathbf{0}^T, 3)$ where $\mathbf{0}^T$ is an $(m_1 - 4)$-dimensional vector of zeros and κ is a simulation parameter. The data $\mathbf{x}_1, \ldots, \mathbf{x}_{m_1}$ are generated according to four distinct categories whose interpretations are summarized in [25]. Correlations for the first $0.75m_1$ variables are around 0.8 in absolute magnitude. The remaining variables are independent from each other and the first $0.75m_1$ variables.

We generate 100 independent data sets for each value of κ in the set $\{1, 2, 3, 4, 5, 6, 7\}$ and apply each of the variable selection procedures we consider. Note that larger values of κ in the range $\kappa \in [1, 7]$ correspond to a smaller signal to noise ratio. [25] considered the case where $\kappa = 1$ and $n = 40$. The results are summarized in the two panels of Figures 2.

We can see that in the left panel of Figure 2, VB, EMVS and BMS work almostly equally well to select the correct model, BMS works slightly better for larger κ. In the right panel, VB and EMVS provide similar prediction errors which are not as well as BMS but much better than the rest of methods. Note that the mean times per simulation for our VB method, and the Lasso, SCAD, MCP, EMVS and BMS were 2.08, 0.06, 0.04, 0.03, 0.57 and 15.87 seconds respectively.

The results for the comparisons between VB and MCMC based on 100 simulations are summarized in Table 1. The posterior density accuracy is better for $\boldsymbol{\beta}$, but less so for σ^2 where accuracy decreases as the signal to noise ratio decreases. Note that the MCMC approach took an average of 17.55 seconds per simulation setting.

Table 1
Performance measure comparisons between VB and MCMC based on 100 simulations for the diet simulation example.

	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$	$\kappa = 5$	$\kappa = 6$	$\kappa = 6$
−log-MSE-VB	1.83 (0.42)	1.54 (0.52)	1.56 (0.46)	1.57 (0.42)	1.43 (0.40)	1.15 (0.40)	1.07 (0.37)
−log-MSE-MCMC	2.29 (0.05)	2.28 (0.05)	2.29 (0.05)	2.29 (0.05)	2.29 (0.05)	2.29 (0.06)	2.26 (0.06)
−log-BIAS-VB	3.55 (0.13)	3.13 (0.15)	3.16 (0.13)	3.23 (0.11)	3.01 (0.11)	2.26 (0.11)	2.26 (0.09)
−log-BIAS-MCMC	4.79 (0.01)	4.78 (0.01)	4.79 (0.01)	4.81 (0.01)	4.80 (0.01)	4.79 (0.01)	4.71 (0.01)
F_1-VB	0.99 (0.04)	0.99 (0.05)	0.98 (0.06)	0.98 (0.06)	0.97 (0.07)	0.94 (0.09)	0.92 (0.11)
F_1-MCMC	0.99 (0.03)	0.98 (0.03)	0.99 (0.03)	0.99 (0.03)	0.99 (0.03)	0.98 (0.03)	0.99 (0.03)
accuracy(β)	91.3 (9.11)	89.7 (12.5)	89.6 (12.7)	89.3 (12.9)	87.3 (15.9)	82.5 (19.0)	80.1 (18.9)
accuracy(σ^2)	86.9 (17.3)	84.1 (22.6)	83.4 (14.2)	82.4 (25.5)	78.3 (30.5)	69.0 (36.8)	64.0 (37.1)

5.3. *Example 2: Communities and crime data*

We use the *Communities and Crime* dataset obtained from the UCI Machine Learning Repository

http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

The data collected was part of a study by [56] combining socio-economic data from the 1990 United States Census, law enforcement data from the 1990 United States Law Enforcement Management and Administrative Statistics survey, and crime data from the 1995 Federal Bureau of Investigation's Uniform Crime Reports.

The raw data consists of 2215 samples of 147 variables the first 5 of which we regard as non-predictive, the next 124 are regarded as potential covariates while the last 18 variables are regarded as potential response variables. Roughly 15% of the data is missing. We proceed with a complete case analysis of the data. We first remove any potential covariates which contained missing values leaving 101 covariates. We also remove the variables *rentLowQ* and *medGrossRent* since these variables appeared to be nearly linear combinations of the remaining variables (the matrix $\mathbf{X}$ had two singular values approximately 10^{-9} when these variables were included). We use the *nonViolPerPop* variable as the response. We then remove any remaining samples where the response is missing. The remaining dataset consist of 2118 samples and 99 covariates. Finally, the response and covariates are standardized to have mean 0 and standard deviation 1. Empirical correlations between variables range from 3.3×10^{-5} to 0.999.

For this data we use the following procedure as the basis for simulations.

- Use the LARS algorithm to obtain the whole Lasso path and its solution vector $\boldsymbol{\beta}$:

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}$$

  for all positive values of λ. The solution for $\boldsymbol{\beta}$ is a piecewise function of λ with a finite number of pieces, say J, which can be represented by the set $\{\lambda^{(j)}, \boldsymbol{\beta}^{(j)}\}_{1\le j\le J}$.
- For the jth element in this path:
  - Let $\mathbf{X}^{(j)}$ be the columns of $\mathbf{X}$ corresponding to the non-zero elements of $\boldsymbol{\beta}^{(j)}$.
  - Find the least squares fit $(\widehat{\boldsymbol{\beta}}_{\text{LS}}^{(j)}, \widehat{\sigma}_j^2)$ of the data $(\mathbf{y}, \mathbf{X}^{(j)})$.
  - Simulate S datasets from the model $\mathbf{y} \sim N(\mathbf{X}^{(j)}\widehat{\boldsymbol{\beta}}_{\text{LS}}^{(j)}, \sigma^2\mathbf{I})$ for some value σ^2.

For this data we use $\sigma^2 = 1$, the first $J = 20$ elements of the LARS path and $S = 50$. Such datasets are simulated for each of these $J = 20$ elements. We use the *R* package *lars* [29] in the above procedure. Results for the comparisons between VB, Lasso, SCAD and MCP are summarized in Figure 3. We can see that our VB approach is competitive to other methods especially when model size is small. The Lasso performs well in model selection but gives larger prediction error while EMVS works less well in model selection but provides very stable prediction error in all simulation settings. The mean times per simulation for our VB method, and the Lasso, SCAD, MCP, EMVS and BMS were 6.92, 5.13, 3.49, 2.72, 0.73 and 40.69 seconds respectively.

The results for the comparisons between VB and MCMC based on 20 simulations are summarized in Table 2. In this table we see that parameter posterior density accuracies are nearly perfect for all parameters when model size is small and still reasonable when model size is large. Note that the MCMC approach took an average of 48.75 seconds per simulation setting.

5.4. Example 3: Simulated SNP data

For our first $p > n$ example we take a simulation setting from [11] that mimics some properties of single-nucleotide polymorphism (SNP) data. We used the *R* package *varbvs* [10] to generate the data. For all trials, we set $n = 500$, $p = 1000$, the number of non-zero coefficients to $m = 20$ and $\sigma = 3$. Note that for this example all covariates are uncorrelated. This process is repeated 50 times and the results are summarized in Figure 4. For this example all methods, except

TABLE 2
Performance measure comparisons between VB and MCMC based on 30 simulations for the communities and crime example.

Model size	1	2	3	4	5
−log-MSE-VB	7.96(0.00)	7.21(0.00)	6.26(0.00)	4.89(0.00)	4.71(0.00)
−log-MSE-MCMC	7.24(0.00)	6.77(0.00)	5.93(0.00)	5.38(0.00)	5.15(0.00)
−log-BIAS-VB	12.57(0.00)	11.82(0.00)	7.53(0.00)	6.97(0.00)	7.14(0.00)
−log-BIAS-MCMC	11.13(0.00)	8.58(0.00)	7.91(0.00)	8.04(0.00)	7.78(0.00)
F_1-VB	0.67(0.00)	0.80(0.00)	0.76(0.17)	0.61(0.15)	0.59(0.17)
F_1-MCMC	0.67(0.00)	0.80(0.00)	0.73(0.14)	0.69(0.11)	0.61(0.13)
accuracy(β)	98.67(0.53)	96.25(2.68)	95.26(2.92)	91.72(4.89)	91.91(5.30)
accuracy(σ^2)	97.90(2.40)	98.10(2.47)	97.75(2.95)	95.65(3.92)	95.95(3.10)
Model size	6	7	8	9	10
−log-MSE-VB	4.60(0.00)	4.42(0.01)	4.00(0.01)	3.86(0.01)	3.79(0.01)
−log-MSE-MCMC	4.96(0.00)	4.86(0.00)	4.70(0.01)	4.56(0.01)	4.44(0.00)
−log-BIAS-VB	7.39(0.00)	7.40(0.00)	7.35(0.00)	7.30(0.00)	7.41(0.00)
−log-BIAS-MCMC	7.67(0.00)	7.65(0.00)	7.66(0.00)	7.66(0.00)	7.50(0.00)
F_1-VB	0.59(0.10)	0.57(0.11)	0.56(0.11)	0.51(0.10)	0.48(0.07)
F_1-MCMC	0.55(0.11)	0.50(0.12)	0.55(0.11)	0.51(0.10)	0.47(0.08)
accuracy(β)	85.86(7.81)	84.83(9.25)	84.69(9.39)	84.33(9.26)	84.97(8.81)
accuracy(σ^2)	94.55(4.20)	93.70(4.94)	90.15(5.98)	88.30(7.11)	86.85(5.06)
Model size	11	12	13	14	15
−log-MSE-VB	3.76(0.01)	3.76(0.01)	3.67(0.01)	3.53(0.01)	3.46(0.01)
−log-MSE-MCMC	4.50(0.00)	4.50(0.00)	4.37(0.01)	4.17(0.01)	4.08(0.01)
−log-BIAS-VB	7.27(0.00)	7.27(0.00)	7.20(0.00)	6.92(0.00)	7.10(0.00)
−log-BIAS-MCMC	7.64(0.00)	7.64(0.00)	7.58(0.00)	7.22(0.00)	7.18(0.00)
F_1-VB	0.46(0.08)	0.46(0.08)	0.41(0.06)	0.40(0.08)	0.39(0.06)
F_1-MCMC	0.45(0.10)	0.45(0.10)	0.45(0.07)	0.41(0.11)	0.40(0.09)
accuracy(β)	86.04(10.17)	86.04(10.17)	84.72(9.20)	81.88(10.79)	82.56(9.75)
accuracy(σ^2)	87.90(6.85)	87.90(6.85)	87.95(6.78)	85.15(6.64)	82.55(5.99)
Model size	16	17	18	19	20
−log-MSE-VB	3.42(0.01)	3.38(0.01)	3.37(0.01)	3.29(0.01)	3.25(0.01)
−log-MSE-MCMC	4.05(0.01)	4.02(0.01)	3.98(0.01)	3.90(0.00)	3.90(0.00)
−log-BIAS-VB	6.98(0.00)	6.93(0.00)	6.92(0.00)	6.72(0.00)	6.68(0.00)
−log-BIAS-MCMC	7.23(0.00)	7.21(0.00)	7.12(0.00)	6.75(0.00)	6.74(0.00)
F_1-VB	0.36(0.07)	0.36(0.05)	0.35(0.06)	0.31(0.06)	0.28(0.05)
F_1-MCMC	0.38(0.10)	0.34(0.10)	0.35(0.09)	0.31(0.10)	0.30(0.09)
accuracy(β)	82.69(11.32)	76.87(15.91)	77.75(15.57)	74.15(17.34)	74.08(15.00)
accuracy(σ^2)	83.45(6.06)	82.45(8.29)	82.70(8.43)	80.95(8.80)	81.35(8.70)

for perhaps the Lasso had similar model selection accuracy. However, VB and BMS were superior when compared to the other selected methods in terms of perdiction accuracy and bias. The Lasso, SCAD, MCP, EMVS, VB and BMS methods took 0.1, 0.1, 0.1, 45, 197 and 299 seconds respectively.

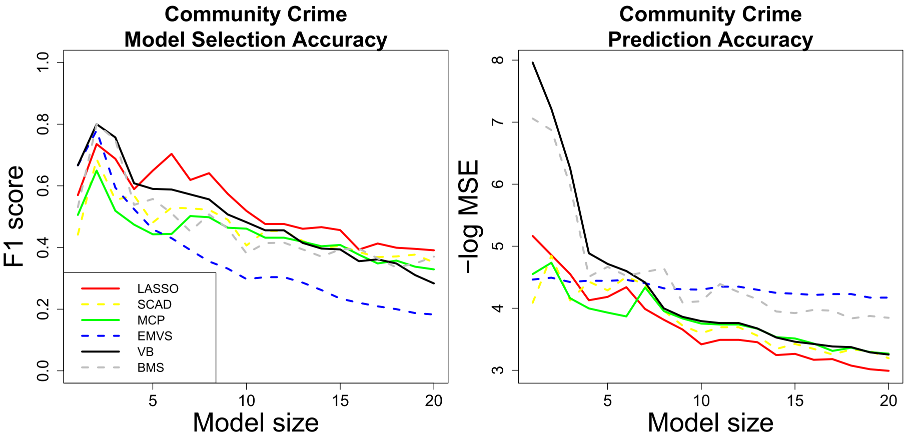

Fig 3. *Summaries of the model selection and prediction accuracies of VB, Lasso, SCAD and MCP methods for the Communities and Crime example.*

5.5. *Example 4: Simulated QTL data*

For our final $p > n$ simulation example we will use the design matrix based on an experiment on a backcross population of $n = 600$ individuals for a single large chromosome of 1800 cM. This giant chromosome was covered by 121 evenly spaced markers from [72]. Nine ofthe markers overlapped with QTL ofthe main effects and 13 out of the $\binom{121}{2} = 7260$ possible marker pairs had interaction effects. The $\mathbf{X}$ matrix combines the main effects and interaction effects to make a 600×7381 matrix. The values of the true coefficients are listed in Table 1 of [72] ranging from 0.77 to 4.77 in absolute magnitude and correlations range from 0 to 0.8 where most of the higher correlation occurs along the off-diagonal values of the correlation matrix of the covariates. Here we center the $\mathbf{X}$ matrix and simulate new data from $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ and the $\boldsymbol{\varepsilon}_i$ are independently drawn with $\boldsymbol{\varepsilon}_i \sim N(0, 20)$. Similar simulation studies were conducted in [72] and [37]. This process was repeated 50 times and the results are summarized in Figure 3. For this simulation setting VB has the best model selection accuracy, smallest MSEs and smallest parameter biases of all the methods compared. The Lasso, SCAD, MCP, EMVS, VB and BMS methods took 1.5, 1.5, 1.8, 1229, 2011, 5327 seconds respectively.

6. Extension to Bayesian robust linear regression

Here we will make the argument that variational Bayes allows relatively straightforward extensions to non-standard complications. We will further show that the above methodology can relatively easily be extended to model selection for robust fits of linear models. Here we do so by using the Laplace or double exponential distribution to model the response, i.e., we replace $\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim$

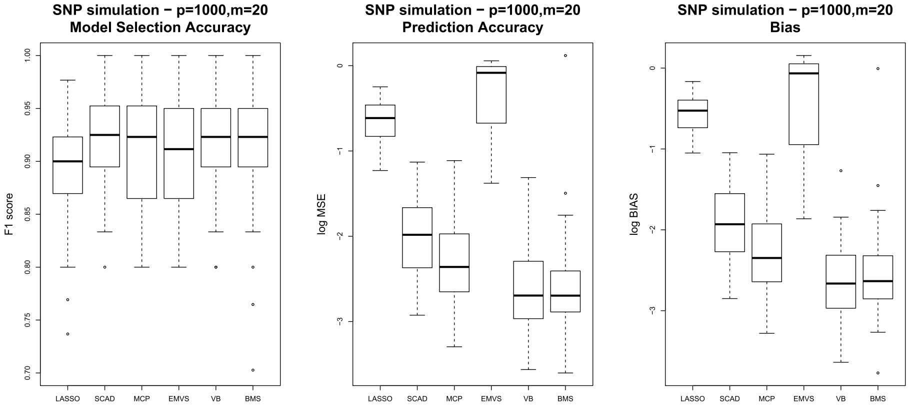

FIG 4. *Summaries of the model selection, prediction accuracies and coefficient biases of the VB, Lasso, SCAD, MCP, EMVS and BMS methods for theSimulated SNP data example.*

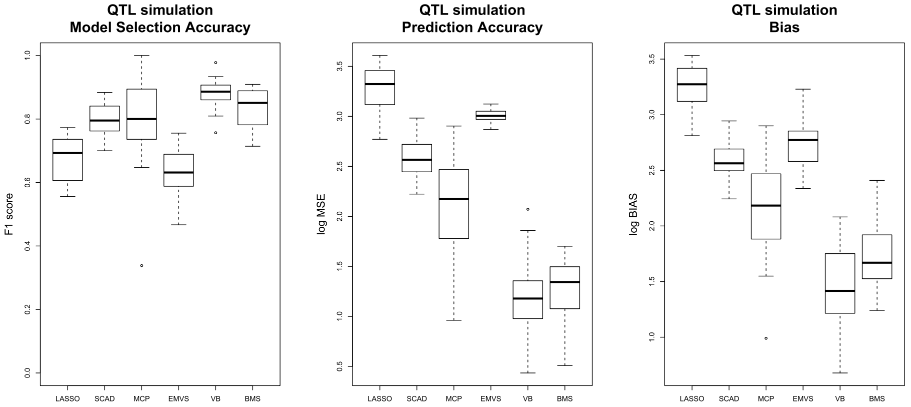

FIG 5. *Summaries of the model selection, prediction accuracies and coefficient biases of the VB, Lasso, SCAD, MCP, EMVS and BMS methods for the Simulated QTL data example.*

$N(\mathbf{X\Gamma}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ in (1) by

$$\mathbf{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim \text{Laplace}(\mathbf{X\Gamma}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

Following [2] we can represent the Laplace distribution by the normal scale-mixture

$$y_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, a_i \sim \text{N}(\mathbf{x}_i^T\mathbf{\Gamma}\boldsymbol{\beta}, a_i^{-1}\sigma^2), \quad \text{with} \quad a_i \sim \text{Inverse-Gamma}(1, 1/2),$$

where the remaining elements of the model specification are identical to (1). Note that the above representation follows from the fact that

$$\int_0^\infty p(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, a_i)p(a_i)\mathrm{d}a_i = p(y_i|\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2),$$

Algorithm 3 *Iterative scheme to obtain parameters in optimal q-densities for our model.*

```
Input: $(\mathbf{y}, \mathbf{X}, \tau_\beta, A, B, \tau, \rho, \mathbf{w}, \widetilde{\mathbf{A}})$
where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\sigma_\beta^2 > 0$, $A > 0$, $B > 0$, $\tau > 0$, $\rho \in (0,1)$, $\text{diag}(\widetilde{\mathbf{A}}) = \mathbf{1}$ and $\mathbf{w} \in [0,1]^p$.
$\mathbf{W} \leftarrow \text{diag}(\mathbf{w})$ ; $\boldsymbol{\Omega} \leftarrow \mathbf{w}\mathbf{w}^T + \mathbf{W}(\mathbf{I} - \mathbf{W})$ ; $\lambda \leftarrow \text{logit}(\rho)$
Cycle:
    $\boldsymbol{\Sigma} \leftarrow \left[\tau(\mathbf{X}^T\widetilde{\mathbf{A}}\mathbf{X}) \odot \boldsymbol{\Omega} + \tau_\beta \mathbf{I}\right]^{-1}$ ; $\boldsymbol{\mu} \leftarrow \tau \boldsymbol{\Sigma}\mathbf{W}\mathbf{X}^T\widetilde{\mathbf{A}}\mathbf{y}$
    For $j = 1, \ldots, p$
        $w_j \leftarrow \text{expit}\left[\lambda - \frac{1}{2}\tau \mathbf{X}_j^T \widetilde{\mathbf{A}}\mathbf{X}_j(\mu_j^2 + \Sigma_{jj})\right.$
            $+\tau \mathbf{X}_j^T \widetilde{\mathbf{A}}\left[\mathbf{y}\mu_j - \mathbf{X}_{-j}\mathbf{W}_{-j}(\boldsymbol{\mu}_{-j}\mu_j + \boldsymbol{\Sigma}_{-j,j})\right]$
        $\mathbf{w} \leftarrow [w_1, \ldots, w_p]^T$ ; $\mathbf{W} \leftarrow \text{diag}(\mathbf{w})$
    $\boldsymbol{\Omega} \leftarrow \mathbf{w}\mathbf{w}^T + \mathbf{W}(\mathbf{I} - \mathbf{W})$
    $s \leftarrow B + \frac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\widetilde{\mathbf{A}}\mathbf{X}\mathbf{W}\boldsymbol{\mu} + \text{tr}\left((\mathbf{X}^T\widetilde{\mathbf{A}}\mathbf{X} \odot \boldsymbol{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\right)\right]$
    $\tau \leftarrow (A + n/2)/s$
     For $i = 1, \ldots, n$
        $\widetilde{A}_i \leftarrow \tau^{-1/2}\left[y_i^2 - 2y_i\mathbf{x}_i^T\mathbf{W}\boldsymbol{\mu} + \text{tr}\left((\mathbf{x}_i\mathbf{x}_i^T \odot \boldsymbol{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\right)\right]^{-1/2}$
     $\widetilde{\mathbf{A}} \leftarrow \text{diag}(\widetilde{A}_1, \ldots, \widetilde{A}_n)$
Until the increase of $\log \underline{p}_{\text{Laplace}}(\mathbf{y}; \rho)$ is negligible.
```

which can be shown using properties of the inverse Gaussian distribution. Using a variational Bayes approximation of $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{a}|\mathbf{y})$ by

$$q(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{a}) = q(\boldsymbol{\beta})q(\sigma^2)\left[\prod_{j=1}^{p} q(\gamma_j)\right]\left[\prod_{i=1}^{n} q(a_i)\right]$$

the optimal q-densities are of the form

$$\begin{aligned} q^*(\boldsymbol{\beta}) \quad &\text{is a} \quad N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{density}, \\ q^*(\sigma^2) \quad &\text{is a} \quad \text{Inverse-Gamma}(A + n/2, s) \quad \text{density}, \\ q^*(\gamma_j) \quad &\text{is a} \quad \text{Bernoulli}(w_j) \quad \text{density for } j = 1, \ldots, p, \\ \text{and} \quad q^*(a_j) \quad &\text{is a Inverse-Gaussian}(\widetilde{A}_j, 1) \quad \text{density for } i = 1, \ldots, n, \end{aligned}$$

where the optimal values for the parameters are obtained via Algorithm 3 which is derived in Appendix C. If x has an inverse Gaussian distribution, denoted $x \sim \text{Inverse-Gaussian}(\mu, \lambda)$ with mean μ and variance μ^3/λ, then it has density

$$p(x) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2x\mu^2}\right\}, \quad x, \mu, \lambda > 0.$$

At the bottom of Algorithm 3 the lower bound on $\log p(\mathbf{y}; \rho)$ simplifies to

$$\log \underline{p}_{\text{Laplace}}(\mathbf{y}; \rho) = \log \underline{p}(\mathbf{y}; \rho) + \frac{n}{2}\log(2\pi) - n\log(2) - \sum_{i=1}^{n} \frac{1}{2\widetilde{A}_i},$$

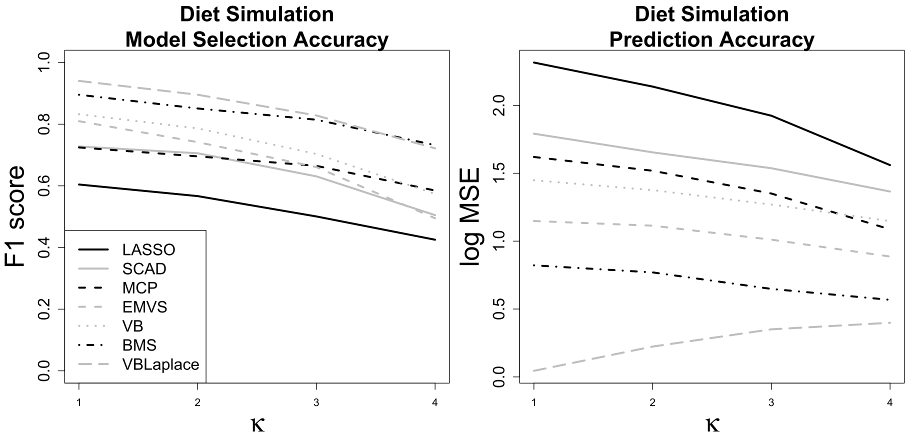

Fig 6. *Summaries of the model selection and prediction accuracies of VB, Lasso, SCAD and MCP methods for the Communities and Crime example with Laplace distribute errors.*

where $\log \underline{p}(\mathbf{y};\rho)$ is the expression for the lower bound defined in Section 2. Note that the $\widetilde{A}_i$ can be rewritten as

$$\tau^{-1/2}\left[(y_i - \mathbf{x}_i^T\mathbf{W}\boldsymbol{\mu})^2 + \mathbf{x}_i^T\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{x}_i + w_i(1-w_i)((\mathbf{x}_i^T\boldsymbol{\mu})^2 + \mathbf{x}_i^T\boldsymbol{\Sigma}\mathbf{x}_i)\right]^{-1/2}.$$

Note that the term $(y_i - \mathbf{x}_i^T\mathbf{W}\boldsymbol{\mu})^2$ can be interpreted as an estimate of the error in prediction of the ith sample whereas $\mathbf{x}_i^T\mathbf{W}\boldsymbol{\Sigma}\mathbf{W}\mathbf{x}_i$ can be interpreted as a measure of the influence for the ith sample. Thus, the procedure will downweight both outliers and high influence points simultaneously.

6.1. *Example 5: Diets simulation with Laplace errors*

We now revisit Example 1 with Laplace distributed errors, i.e., where $\boldsymbol{\varepsilon}$ follows the Laplace distribution with scale parameter $\sigma^2 = 0.5$. The results are summarized in Figure 6, which suggests that the performance of the other methods compared is impaired by the non-Gaussian distributed errors. The VB method based on the Laplace distribution has best model selection performance and prediction accuracy over whole range of κ, except for the case where $\kappa = 4$ where BMS has comparable accuracy.

7. Conclusion

In this paper we have provided theory for a new approach which induces sparsity on the estimates of the regression coefficients for a Bayesian linear model. We have shown that these estimates are consistent, can be used to obtain valid standard errors, and that the true model can be found at an exponential rate in n. Our method performs well empirically compared to the penalized regression approaches on the numerical examples we considered and is both much faster and highly accurate when comparing to MCMC.

Our theory is limited to assuming that $p < n$. This might be mitigated to a certain extent by the use of screening procedures such as SIS [18] or the High-dimensional Ordinary Least-squares Projection (HOLP) method of [70] which can be seen as searching for a correct model with high probability whereas extending the theory to $p > n$ is future research.

Theoretical extensions include considering the case where both p and n diverge. Such theory would be important for understanding how the errors of our estimators behave as p grows relative to n. Expansions along this line of research would require combining the theory developed here with a detailed understanding of how ridge regression estimators behave as n and p grow such as developed by [31] or [70].

A second important theoretical extension would be to analyze the effect of more elaborate sparisity inducing priors on the regression coefficients, e.g., where the normal "slab" in the spike and slab is replaced by the Laplace, horseshoe, negative-exponential-gamma and generalized double Pareto distributions [see 67]. Another theoretical extension includes adapting the theory presented here to non-Gaussian response. However, such methodological (as opposed to theoretical) extensions would be relatively straightforward, as would extensions which handle missing data or measurement error highlighting the strength and flexibility of our approach.

Appendix A: Derivation of Algorithm 1

The q-densities corresponding to Algorithm 1 are derived below. The density $q(\boldsymbol{\beta})$ is given by:

$$\begin{aligned} q(\boldsymbol{\beta}) &\propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\left\{-\tfrac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2 - \tfrac{||\boldsymbol{\beta}||^2}{2\sigma_\beta^2}\right\}\right] \\ &\propto \exp\left[-\tfrac{1}{2}\boldsymbol{\beta}^T\left(\tau(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}+\sigma_\beta^{-2}\mathbf{I}\right)\boldsymbol{\beta}+\boldsymbol{\beta}^T\mathbf{W}\mathbf{X}^T\mathbf{y}\tau\right] \\ &= \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma}), \end{aligned}$$

where $\boldsymbol{\Sigma} = (\tau(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}+\sigma_\beta^{-2}\mathbf{I})^{-1}$, $\boldsymbol{\mu} = \tau\boldsymbol{\Sigma}\mathbf{W}\mathbf{X}^T\mathbf{y}$, $\mathbf{w} = \mathbb{E}_q\boldsymbol{\gamma}$, $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, $\boldsymbol{\Omega} = \mathbf{w}\mathbf{w}^T + \mathbf{W}\odot(\mathbf{I}-\mathbf{W})$ and $\tau = \mathbb{E}_q(1/\sigma^2)$. Note that in the above derivation we have used

$$\begin{aligned} \mathbb{E}_q(\boldsymbol{\Gamma}\mathbf{X}^T\mathbf{X}\boldsymbol{\Gamma}) &= \mathbb{E}_q((\mathbf{X}^T\mathbf{X})\odot(\boldsymbol{\gamma}\boldsymbol{\gamma}^T)) \\ &= (\mathbf{X}^T\mathbf{X})\odot\mathbb{E}_q(\boldsymbol{\gamma}\boldsymbol{\gamma}^T) \\ &= (\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}. \end{aligned}$$

The density $q(\sigma^2)$ is given by:

$$\begin{aligned} q(\sigma^2) \quad \propto \quad \exp\Big[\mathbb{E}_{-q(\sigma^2)}\Big\{ &-\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}\|^2 \\ &-(A+1)\log(\sigma^2) - \frac{B}{\sigma^2}\Big\}\Big]. \end{aligned}$$

Hence, $q(\sigma^2) = \text{Inverse-Gamma}(A + \frac{n}{2}, s)$, where

$$s = B + \frac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\mathbf{X}\mathbf{W}\boldsymbol{\mu} + \text{tr}\left((\mathbf{X}^T\mathbf{X}\odot\boldsymbol{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\right)\right].$$

Next noting that $\gamma_j = \gamma_j^2$ as $\gamma_j \in \{0,1\}$, the optimal $q(\gamma_j)$, $1 \le j \le p$, take the form

$$q(\gamma_j) \quad \propto \quad \exp\left[\gamma_j\,\mathbb{E}_{-q_{\gamma_j}}\{\text{logit}(\rho) + \frac{\beta_j}{\sigma^2}\mathbf{X}_j^T\big(\mathbf{y} - \mathbf{X}_{-j}\mathbf{W}_{-j}\boldsymbol{\beta}_{-j}\big) - \frac{\beta_j^2}{2\sigma^2}\mathbf{X}_j^T\mathbf{X}_j\}\right].$$

Hence, $q(\gamma_j) = \text{Bern}(w_j)$, where $w_i = \text{expit}(\eta_j)$ and

$$\eta_j = \lambda - \tfrac{1}{2}\tau\mathbf{X}_j^T\mathbf{X}_j(\mu_j^2 + \Sigma_{jj}) + \tau\mathbf{X}_j^T\left[\mathbf{y}\mu_j - \mathbf{X}_{-j}\mathbf{W}_{-j}(\boldsymbol{\mu}_{-j}\mu_j + \boldsymbol{\Sigma}_{-j,j})\right].$$

Appendix B: Proofs

Proof Lemma 1: Note that $|\,\text{expit}(-a) - \exp(-a)| = \exp(-2a)/(1+\exp(-a)) < \exp(-2a)$, and also note that $\text{expit}(a) = 1 - \text{expit}(-a)$. Hence the result is as stated. □

Result 1. *If $w_j^{(t)} > 0$, $1 \le j \le p$, then $\boldsymbol{\Omega}$ is positive definite.*

Proof of Result 1: A matrix is positive definite if and only if for all non-zero real vector $\mathbf{a} = [a_1, \ldots, a_p]^T$ the scalar $\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a}$ is strictly positive [30, Section 7.1]. By definition $\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a} = \mathbf{a}^T\big[\mathbf{w}\mathbf{w}^T + \mathbf{W}(\mathbf{I} - \mathbf{W})\big]\mathbf{a} = (\sum_{j=1}^p a_jw_j)^2 + \sum_{j=1}^p a_j^2w_j(1-w_j)$. As $0 < w_j^{(t)} \le 1$, $1 \le j \le p$, we have $w_j(1-w_j) \ge 0$ and hence $\sum_{j=1}^p a_j^2w_j(1-w_j) \ge 0$. Again, as $w_j^{(t)} > 0$, $1 \le j \le p$, we have $(\sum_{j=1}^p a_jw_j)^2 > 0$ for any non-zero vector $\mathbf{a}$. Hence, the result is as stated. □

Let

$$\text{dof}(\alpha, \mathbf{w}) = \text{tr}\left[(\mathbf{X}^T\mathbf{X}\odot\boldsymbol{\Omega})\left\{(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega} + \alpha\mathbf{I}\right\}^{-1}\right]$$

and

$$\mathbf{U}\text{diag}(\boldsymbol{\nu})\mathbf{U}^T \quad \text{be the eigenvalue decomposition of} \quad (\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}, \tag{17}$$

where $\mathbf{U}$ is an orthonormal matrix and $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_p]^T$ is a vector of eigenvalues of $(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}$.

Result 2. *Suppose $\mathbf{X}^T\mathbf{X}$ is semi-positive definite and $w_j \in (0,1]$, $1 \le j \le p$ and $\alpha \ge 0$ then the function dof$(\alpha, \mathbf{w})$ is monotonically decreasing in α and satisfies $0 <$ dof$(\alpha, \mathbf{w}) \le$ rank$(\mathbf{X}^T\mathbf{X}\odot\boldsymbol{\Omega}) \le p$.*

Proof of Result 2 : Let the eigenvalue decomposition (17) hold. Since $w_j \in (0,1]$, $1 \le j \le p$ is positive, by Result 1 the matrix $\boldsymbol{\Omega}$ is positive definite. By the Schur product theorem [30, Theorem 7.5.2], the matrix $(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}$ is also semi-positive definite. Hence, ν_i, $i = 1,\dots,p$ defined in Equation (17) is non-negative and number of non-zero $\nu_i, i = 1,\dots,p$ equals $\text{rank}(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Omega})$. Then, using properties of the orthonormal matrix $\mathbf{U}$, we have

$$\begin{aligned}
\text{dof}(\alpha,\mathbf{w}) &= \text{tr}\left[\mathbf{U}\text{diag}(\boldsymbol{\nu})\mathbf{U}^T\left(\mathbf{U}\text{diag}(\boldsymbol{\nu})\mathbf{U}^T + \alpha\mathbf{I}\right)^{-1}\right] \\
&= \sum_{j=1}^{p} \frac{\nu_j}{\nu_j + \alpha} \\
&= \sum_{\nu_j \neq 0} \frac{\nu_j}{\nu_j + \alpha} \leq \text{rank}(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Omega}).
\end{aligned}$$

Note that $\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Omega}$ is a $p \times p$ matrix, hence $\text{rank}(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Omega}) \leq p$. Clearly, $\text{dof}(\alpha,\mathbf{w})$ is monotonically decreasing in α and $\text{dof}(\alpha,\mathbf{w})$ only approaches zero as $\alpha \to \infty$. □

The next lemma follows from [30, Section 0.7.3]:

Lemma 2. *The inverse of a real symmetric matrix can be written as*

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ -\mathbf{C}^{-1}\mathbf{B}^T & \mathbf{I} \end{bmatrix}\begin{bmatrix} \widetilde{\mathbf{A}} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{I} & -\mathbf{B}\mathbf{C}^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \quad (18)$$

$$= \begin{bmatrix} \widetilde{\mathbf{A}} & -\widetilde{\mathbf{A}}\mathbf{B}\mathbf{C}^{-1} \\ -\mathbf{C}^{-1}\mathbf{B}^T\widetilde{\mathbf{A}} & \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{B}^T\widetilde{\mathbf{A}}\mathbf{B}\mathbf{C}^{-1} \end{bmatrix} \quad (19)$$

where $\widetilde{\mathbf{A}} = \left(\mathbf{A} - \mathbf{B}\mathbf{C}^{-1}\mathbf{B}^T\right)^{-1}$ *provided all inverses in (18) and (19) exist.*

Lemma 3. *Let* $\mathbf{M}$ *be a real positive definite symmetric* $p \times p$ *matrix,* $\mathbf{a} = [a_1,\dots,a_p]^T$ *be a real vector, and let the elements of the vector* $\mathbf{b} = [b_1,\dots,b_p]^T$ *be positive. Then the quantity* $\mathbf{a}^T\left[\mathbf{M} + \text{diag}(\mathbf{b})\right]^{-1}\mathbf{a}$ *is a strictly decreasing function of any element of* $\mathbf{b}$.

Proof of Lemma 3: Let the matrix $\mathbf{M} + \text{diag}(\mathbf{b})$ be partitioned as

$$\mathbf{M} + \text{diag}(\mathbf{b}) = \begin{bmatrix} M_{11} + b_1 & \mathbf{m}_{12}^T \\ \mathbf{m}_{12} & \mathbf{M}_{22} + \mathbf{B}_2 \end{bmatrix}$$

where $\mathbf{m}_{12} = [M_{12},\dots,M_{1p}]^T$, $\mathbf{B}_2 = \text{diag}(b_2,\dots,b_p)$ and

$$\mathbf{M}_{22} = \begin{bmatrix} M_{22} & \cdots & M_{2p} \\ \vdots & \ddots & \vdots \\ M_{p2} & \cdots & M_{pp} \end{bmatrix}.$$

Then, by Equation (18) in Lemma 2,

$$\mathbf{a}^T\left[\mathbf{M} + \text{diag}(\mathbf{b})\right]^{-1}\mathbf{a} = \frac{c_1^2}{b_1 + M_{11} - \mathbf{m}_{12}^T(\mathbf{M}_{22} + \mathbf{B}_2)^{-1}\mathbf{m}_{12}} + \mathbf{c}_2^T(\mathbf{M}_{22} + \mathbf{B}_2)^{-1}\mathbf{c}_2 \quad (20)$$

where

$$\mathbf{c} = \begin{bmatrix} 1 & -\mathbf{m}_{12}^T(\mathbf{M}_{22}+\mathbf{B}_2)^{-1} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \mathbf{a} \qquad \text{and} \qquad \mathbf{c}_2 = [c_2, \ldots, c_p]^T.$$

Note any principal submatrix of a positive definite matrix is positive definite [30, Chapter 7.1.2]. Hence, the matrix $(\mathbf{M}+\text{diag}(\mathbf{b}))^{-1}$ is positive definite and $(b_1 + M_{11} - \mathbf{m}_{12}^T(\mathbf{M}_{22}+\mathbf{B}_2)^{-1}\mathbf{m}_{12})^{-1}$ is a positive scalar. Clearly, (20) is strictly decreasing as b_1 increases. The result follows for $b_j, 2 \le j \le p$ after a relabeling argument. □

The following result bounds the values that τ can take and is useful because these bounds do not depend on $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ or $\mathbf{w}$.

Result 3. *Suppose the updates in Algorithm 1 hold, then $\tau_L \le \tau^{(t)} \le \tau_U$ for all t where*

$$\tau_L = \frac{2A+n-p}{2B + 2\|\mathbf{y}\|^2 + 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + p\frac{2A+n-p}{(2A+n)\tau^{(0)}}}$$

and

$$\tau_U = \frac{2A+n}{2B + \|\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\|^2}.$$

Proof of Result 3: From Algorithm 1 we can express $\tau^{(t)}$ as

$$\begin{aligned} \tau^{(t)} &= (2A+n)\Big[2B + \|\mathbf{y} - \mathbf{X}\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)}\|^2 \\ &\quad + \boldsymbol{\mu}^{(t)^T}[(\mathbf{X}^T\mathbf{X}) \odot \mathbf{W}^{(t)} \odot (\mathbf{I} - \mathbf{W}^{(t)})]\boldsymbol{\mu}^{(t)} \\ &\quad + \tau^{(t-1)^{-1}}\text{dof}(\tau^{(t-1)^{-1}}\sigma_\beta^{-2}, \mathbf{w}^{(t)})\Big]^{-1}. \end{aligned}$$

The upper bound for $\tau^{(t)}$ where $t > 0$ follows from the above equation and following these inequalities:

(a) $\text{dof}(\tau^{(t-1)^{-1}}\sigma_\beta^{-2}, \mathbf{w}) > 0$ for any $\tau^{(t-1)} > 0$;
(b) $\|\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu}\|^2 \ge \|\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\|^2$ (from least squares results); and
(c) $\boldsymbol{\mu}^T[(\mathbf{X}^T\mathbf{X}) \odot \mathbf{W} \odot (\mathbf{I} - \mathbf{W})]\boldsymbol{\mu} \ge 0$ (as $(\mathbf{X}^T\mathbf{X}) \odot \mathbf{W} \odot (\mathbf{I} - \mathbf{W})$ is clearly at least positive semidefinite).

To obtain a lower bound for $\tau^{(t)}$ first note that for any vector $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$, $\|\mathbf{u}+\mathbf{v}\|^2 \le 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$, as $\|\mathbf{u}+\mathbf{v}\|^2 + \|\mathbf{u}-\mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$ and $\|\mathbf{u}-\mathbf{v}\|^2$ is non-negative. Hence $\|\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu}\|^2 \le 2\|\mathbf{y}\|^2 + 2\|\mathbf{X}\mathbf{W}\boldsymbol{\mu}\|^2$. Using Result 2 and the fact $\boldsymbol{\mu}^T[(\mathbf{X}^T\mathbf{X}) \odot \mathbf{W} \odot (\mathbf{I} - \mathbf{W})]\boldsymbol{\mu} \ge 0$ we have

$$\begin{aligned} \tau^{(t)} &\ge (2A+n)\Big[2B + 2\|\mathbf{y}\|^2 + p/\tau^{(t-1)} \\ &\quad + \boldsymbol{\mu}^{(t)^T}[2\mathbf{W}^{(t)}\mathbf{X}^T\mathbf{X}\mathbf{W}^{(t)} + (\mathbf{X}^T\mathbf{X}) \odot \mathbf{W}^{(t)} \odot (\mathbf{I} - \mathbf{W}^{(t)})]\boldsymbol{\mu}^{(t)}\Big]^{-1} \\ &\ge (2A+n)\Big[2B + 2\|\mathbf{y}\|^2 + p/\tau^{(t-1)} \\ &\quad + \boldsymbol{\mu}^{(t)^T}[2\mathbf{W}^{(t)}\mathbf{X}^T\mathbf{X}\mathbf{W}^{(t)} + 2(\mathbf{X}^T\mathbf{X}) \odot \mathbf{W}^{(t)} \odot (\mathbf{I} - \mathbf{W}^{(t)})]\boldsymbol{\mu}^{(t)}\Big]^{-1} \\ &= \frac{2A+n}{2B + 2\|\mathbf{y}\|^2 + 2\boldsymbol{\mu}^{(t)^T}[(\mathbf{X}^T\mathbf{X}) \odot \boldsymbol{\Omega}^{(t)}]\boldsymbol{\mu}^{(t)} + p/\tau^{(t-1)}}. \end{aligned}$$

Let the eigenvalue decomposition (17) hold. Then

$$\begin{aligned}
&\boldsymbol{\mu}^T[(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}]\boldsymbol{\mu}\\
&\quad= \mathbf{y}^T\mathbf{X}\mathbf{W}\big[\mathbf{U}\text{diag}(\boldsymbol{\nu})\mathbf{U}^T + \tau^{-1}\sigma_\beta^{-2}\mathbf{I}\big]^{-1}\mathbf{U}\text{diag}(\boldsymbol{\nu})\mathbf{U}^T\\
&\quad\quad\times\big[\mathbf{U}\text{diag}(\boldsymbol{\nu})\mathbf{U}^T + \tau^{-1}\sigma_\beta^{-2}\mathbf{I}\big]^{-1}\mathbf{W}\mathbf{X}^T\mathbf{y}\\
&\quad= \sum_{j=1}^{p}\frac{\nu_j(\mathbf{U}^T\mathbf{W}\mathbf{X}^T\mathbf{y})_j^2}{(\nu_j+\tau^{-1}\sigma_\beta^{-2})^2}\\
&\quad\leq \mathbf{y}^T\mathbf{X}^T\mathbf{W}\big[(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega}\big]^{-1}\mathbf{W}\mathbf{X}^T\mathbf{y}\\
&\quad= \mathbf{y}^T\mathbf{X}\big[\mathbf{X}^T\mathbf{X} + \mathbf{W}^{-1}\big\{(\mathbf{X}^T\mathbf{X})\odot\mathbf{W}\odot(\mathbf{I}-\mathbf{W})\big\}\mathbf{W}^{-1}\big]^{-1}\mathbf{X}^T\mathbf{y}\\
&\quad\leq \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y},
\end{aligned}$$

where the last line follows from Lemma 3. Combining this inequality the lower bound for $\tau^{(t)}, t > 0$ can be expressed as

$$\tau_L^{(t)} = \frac{2A+n}{2B + 2\|\mathbf{y}\|^2 + 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + p/\tau^{(t-1)}}. \tag{21}$$

Note that $\tau^{(t-1)} \geq \tau_L^{(t-1)}$ and expand the recursive inequality, we obtain

$$\begin{aligned}
\tau^{(t)} \quad &\geq \quad \frac{2A+n}{2B + 2\|\mathbf{y}\|^2 + 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + p/\tau_L^{(t-1)}}\\
&\geq \quad \frac{2A+n}{\frac{p^t}{(2A+n)^{t-1}\tau^{(0)}} + (2B + 2\|\mathbf{y}\|^2 + 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})\sum_{k=0}^{t-1}\frac{p^k}{(2A+n)^k}}.
\end{aligned}$$

Note that as $p < n$,

$$\sum_{k=0}^{t-1}\frac{p^k}{(2A+n)^k} \leq \frac{2A+n}{2A+n-p} \quad \text{and} \quad \frac{p^t}{(2A+n)^{t-1}} < p.$$

Hence,

$$\tau_L^{(t)} \geq \frac{2A+n-p}{2B + 2\|\mathbf{y}\|^2 + 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + p\frac{2A+n-p}{(2A+n)\tau^{(0)}}},$$

which is independent to t. Hence the lower bound on $\tau^{(t)}$ is as stated. □

B.1. Proof of Main Result 1

It is clear from the numerical example in Section 2 that sparsity in the vector $\boldsymbol{\mu}$ is achieved (at least approximately). In order to understand how sparsity is achieved we need to understand how the quantities $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\eta}$ behave when elements of the vector $\mathbf{w}$ are small. Define the $n \times n$ matrix $\mathbf{P}_j$ for $1 \leq j \leq p$ by

$$\mathbf{P}_j \equiv \mathbf{X}_{-j}\mathbf{W}_{-j}\left(\mathbf{W}_{-j}\mathbf{X}_{-j}^T\mathbf{X}_{-j}\mathbf{W}_{-j} + \tau^{-1}\mathbf{D}_{-j}\right)^{-1}\mathbf{W}_{-j}\mathbf{X}_{-j}^T; \tag{22}$$

for $j \neq k, 1 \leq j, k \leq p$ we define

$$\begin{aligned}\mathbf{P}_{(j,k)} \quad &\equiv \mathbf{X}_{-(j,k)}\mathbf{W}_{-(j,k)}\left(\mathbf{W}_{-(j,k)}\mathbf{X}_{-(j,k)}^T\mathbf{X}_{-(j,k)}\mathbf{W}_{-(j,k)} + \tau^{-1}\mathbf{D}_{-(j,k)}\right)^{-1}\\ &\quad\times\mathbf{W}_{-(j,k)}\mathbf{X}_{-(j,k)}^T,\end{aligned}$$

and for a indicator vector $\boldsymbol{\gamma}$ we define

$$\mathbf{P}_{\boldsymbol{\gamma}} \equiv \mathbf{X}_{-\boldsymbol{\gamma}}\mathbf{W}_{-\boldsymbol{\gamma}}\left(\mathbf{W}_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}\mathbf{W}_{-\boldsymbol{\gamma}} + \tau^{-1}\mathbf{D}_{-\boldsymbol{\gamma}}\right)^{-1}\mathbf{W}_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}^T. \tag{23}$$

Result 4. *If (3) holds then*

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} = \left(\tau\mathbf{W}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}} + \mathbf{D}_{\boldsymbol{\gamma}} - \tau\mathbf{W}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}}\right)^{-1} \tag{24}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}} = -\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}\mathbf{W}_{-\boldsymbol{\gamma}}\left(\mathbf{W}_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}\mathbf{W}_{-\boldsymbol{\gamma}} + \tau^{-1}\mathbf{D}_{-\boldsymbol{\gamma}}\right)^{-1}; \tag{25}$$

for $1 \leq j \leq p$ we have

$$\Sigma_{j,j} \quad = \left(\sigma_\beta^{-2} + \tau w_j\|\mathbf{X}_j\|^2 - \tau w_j^2\mathbf{X}_j^T\mathbf{P}_j\mathbf{X}_j\right)^{-1}, \tag{26}$$

and

$$\boldsymbol{\Sigma}_{-j,j} \quad = -\left(\tau\mathbf{W}_{-j}\mathbf{X}_{-j}^T\mathbf{X}_{-j}\mathbf{W}_{-j} + \mathbf{D}_{-j}\right)^{-1}\mathbf{W}_{-j}\mathbf{X}_{-j}^T\mathbf{X}_j(\tau w_j\Sigma_{j,j}); \tag{27}$$

and for $j \neq k$, $1 \leq j, k \leq p$ we have

$$\begin{aligned}\Sigma_{j,k} \quad &= -\tau w_j w_k\mathbf{X}_j^T(\mathbf{I} - \mathbf{P}_{(j,k)})\mathbf{X}_k\\ &\quad\times\Big[\left(\sigma_\beta^{-2} + \tau w_j\|\mathbf{X}_j\|^2 - \tau w_j^2\mathbf{X}_j^T\mathbf{P}_{(j,k)}\mathbf{X}_j\right)\\ &\quad\times\left(\sigma_\beta^{-2} + \tau w_k\|\mathbf{X}_k\|^2 - \tau w_k^2\mathbf{X}_k^T\mathbf{P}_{(j,k)}\mathbf{X}_k\right)\\ &\quad-\{\tau w_j w_k\mathbf{X}_j^T(\mathbf{I} - \mathbf{P}_{(j,k)})\mathbf{X}_k\}^2\Big]^{-1}.\end{aligned} \tag{28}$$

If (3) and (4) hold then

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} \quad = \tau\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^T(\mathbf{I} - \mathbf{P}_{\boldsymbol{\gamma}})\mathbf{y}; \tag{29}$$

and

$$\mu_j = \frac{\tau w_j\mathbf{X}_j^T(\mathbf{I} - \mathbf{P}_j)\mathbf{y}}{\sigma_\beta^{-2} + \tau w_j\|\mathbf{X}_j\|^2 - \tau w_j^2\mathbf{X}_j^T\mathbf{P}_j\mathbf{X}_j}, \qquad 1 \leq j \leq p. \tag{30}$$

Proof of Result 4: For a given indicator vector $\boldsymbol{\gamma}$ we can rewrite (3) as

$$\begin{aligned}&\begin{bmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} & \boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}}\\ \boldsymbol{\Sigma}_{-\boldsymbol{\gamma},\boldsymbol{\gamma}} & \boldsymbol{\Sigma}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}\end{bmatrix}\\ &\quad= \begin{bmatrix}\tau\mathbf{W}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}} + \mathbf{D}_{\boldsymbol{\gamma}} & \tau\mathbf{W}_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}\mathbf{W}_{-\boldsymbol{\gamma}}\\ \tau\mathbf{W}_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}} & \tau\mathbf{W}_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}\mathbf{W}_{-\boldsymbol{\gamma}} + \mathbf{D}_{-\boldsymbol{\gamma}}\end{bmatrix}^{-1}.\end{aligned}$$

Equations (24) and (25) can be obtained by applying Equation (19) in Lemma 2 and equations (26) and (27) can be obtained by letting $\boldsymbol{\gamma} = \mathbf{e}_j$ (where $\mathbf{e}_j$ is the zero vector except for the value 1 in the jth entry). Similarly,

$$\begin{aligned}
\Sigma_{1,2} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \Big[\mathbf{D}_{(1,2)} + \tau \mathbf{W}_{(1,2)} \mathbf{X}_{(1,2)}^T (\mathbf{I} - \mathbf{P}_{(1,2)}) \mathbf{X}_{(1,2)} \mathbf{W}_{(1,2)}\Big]^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\
&= -\tau w_1 w_2 \mathbf{X}_1^T \big(\mathbf{I} - \mathbf{P}_{(1,2)}\big)\mathbf{X}_2 \Big/ \Big[\big(\tau w_1^2 \mathbf{X}_1^T \big(\mathbf{I} - \mathbf{P}_{(1,2)}\big)\mathbf{X}_1 + D_1\big) \\
&\qquad \times \big(\tau w_2^2 \mathbf{X}_2^T \big(\mathbf{I} - \mathbf{P}_{(1,2)}\big)\mathbf{X}_2 + D_2\big) - \big(\tau w_1 w_2 \mathbf{X}_1^T \big(\mathbf{I} - \mathbf{P}_{(1,2)}\big)\mathbf{X}_2\big)^2 \Big] \\
&= -\tau w_1 w_2 \mathbf{X}_1^T \big(\mathbf{I} - \mathbf{P}_{(1,2)}\big)\mathbf{X}_2 \Big/ \Big[\Big(\sigma_\beta^{-2} + \tau w_1 \|\mathbf{X}_1\|^2 - \tau w_1^2 \mathbf{X}_1^T \mathbf{P}_{(1,2)} \mathbf{X}_1\Big) \\
&\qquad \times \Big(\sigma_\beta^{-2} + \tau w_2 \|\mathbf{X}_2\|^2 - \tau w_2^2 \mathbf{X}_2^T \mathbf{P}_{(1,2)} \mathbf{X}_2\Big) \\
&\qquad - \big\{\tau w_1 w_2 \mathbf{X}_1^T (\mathbf{I} - \mathbf{P}_{(1,2)}) \mathbf{X}_2\big\}^2 \Big]
\end{aligned}$$

and (28) follows after a relabeling argument. Equation (29) follows by substituting $\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}}$ into,

$$\boldsymbol{\mu}_{\boldsymbol{\gamma}} = \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} & \boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}} \end{bmatrix} \begin{bmatrix} \tau \mathbf{W}_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{y} \\ \tau \mathbf{W}_{-\boldsymbol{\gamma}} \mathbf{X}_{-\boldsymbol{\gamma}}^T \mathbf{y} \end{bmatrix}$$

and (30) follows by letting $\boldsymbol{\gamma} = \mathbf{e}_j$. □

From Result 3 we have that $\tau^{(t)}$ is bounded for all t as $(y_i, \mathbf{x}_i)$ are observed so that all quantities are deterministic. From Equation (30) we see that $\mu_j^{(t)}$ is clearly $O(w_j^{(t)})$ as $\mathbf{P}_j$ does not depend on w_j. Noting that $\lim_{w_j \to 0} \Sigma_{j,j}^{(t)} = \sigma_\beta^2$ follows from Equation (26) and the result for $\Sigma_{j,j}^{(t)}$ follows after a Taylor series argument. The result $\Sigma_{j,k}^{(t)} = O(w_j^{(t)} w_k^{(t)})$, $j \neq k$ follows from Equation (28). We can see that the update for $w_j^{(t+1)}$ in Algorithm 1 is as stated by combining Equation (7) with the fact that $\mu_j^{(t)} = O(w_j^{(t)})$, $\Sigma_{j,j}^{(t)} = \sigma_\beta^2 + O(w_j^{(t)})$, $\Sigma_{j,k}^{(t)} = O(w_j^{(t)} w_k^{(t)})$ and from Result 3 we have $\tau^{(t)} = O(1)$. This completes the proof of Main Result 1.

B.2. Proof of Main Result 2

For the remainder of this section we will assume that $\mathbf{y}$ and $\mathbf{X}$ (and consequently $\boldsymbol{\mu}^{(t)}$,$\boldsymbol{\Sigma}^{(t)}$,$\tau^{(t)}$ and $\mathbf{w}^{(t)}$ for $t = 0, 1, \ldots$) are random quantities. Note that Results 1–4 are still valid, when assuming random quantities $\mathbf{y}$ and $\mathbf{X}$. Define the following stochastic sequences:

$$\begin{gathered}
\mathbf{A}_n = n^{-1}\mathbf{X}^T\mathbf{X}, \quad \mathbf{b}_n = n^{-1}\mathbf{X}^T\mathbf{y}, \\
c_n = \text{dof}(\tau^{(t)}\sigma_\beta^{-2}, \mathbf{1}) \quad \text{and} \quad \boldsymbol{\beta}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.
\end{gathered} \tag{31}$$

Assuming (A1)–(A4) [73] proved consistency results for Bayesian linear models. We will need stronger results to prove consistency of the estimates corresponding to Algorithm 2. Lemma 4 will aid in obtaining these results.

Lemma 4 ([6], Theorem 14.4.1)**.** *If $\{X_n\}$ is a stochastic sequence with $\mu_n = \mathbb{E}(X_n)$ and $\sigma_n^2 = Var(X_n) < \infty$, then $X_n - \mu_n = O_p(\sigma_n)$.*

Hence, from Lemma 4 and assumptions (A1)–(A5) we have

$$\begin{aligned}
\mathbf{A}_n &= \mathbf{S} + \mathbf{O}_p^m\big(n^{-1/2}\big),\\
\mathbf{A}_n^{-1} &= \mathbf{S}^{-1} + \mathbf{O}_p^m\big(n^{-1/2}\big),\\
\|\mathbf{X}_j\|^2 &= n\mathbb{E}(x_j^2) + O_p\big(n^{1/2}\big),\\
\|\boldsymbol{\epsilon}\|^2 &= n\sigma_0^2 + O_p\big(n^{1/2}\big),\\
n^{-1}\mathbf{X}\boldsymbol{\epsilon} &= \mathbf{O}_p^v\big(n^{-1/2}\big) \quad \text{and}\\
\mathbf{b}_n &= n^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}) = \mathbf{S}\boldsymbol{\beta}_0 + \mathbf{O}_p^v\big(n^{-1/2}\big).
\end{aligned} \tag{32}$$

Before we improve upon the results of [73] we need to show that $\tau^{(t)}$ is bounded for all t. In fact $\tau^{(t)}$ is bounded in probability for all t as the following result shows.

Result 5. *Assume (A1)–(A5), then for $t > 0$ we have $\tau^{(t)} = O_p(1)$ and $1/\tau^{(t)} = O_p(1)$.*

Proof of Result 5: Using Result 3, we obtain $\tau_L < \tau^{(t)} < \tau_U$ and $\tau_U^{-1} < 1/\tau^{(t)} < \tau_L^{-1}$ for $t > 1$ where

$$\begin{aligned}
\tau_U^{-1} &= \frac{2B + \|\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\|^2}{2A+n}\\
&= \left(\frac{n}{2A+n}\right)\frac{2B + \|\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\|^2}{n}
\end{aligned}$$

and $\frac{1}{n}\|\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\|^2 = \frac{1}{n}\|\mathbf{y}\|^2 - \frac{1}{n}\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. By (A1)–(A4) and the strong law of large numbers

$$\begin{aligned}
\tfrac{1}{n}\|\mathbf{y}\|^2 &= \tfrac{1}{n}\|\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}\|^2 = \tfrac{1}{n}\boldsymbol{\beta}_0^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}_0 + \tfrac{1}{n}2\boldsymbol{\varepsilon}^T\mathbf{X}\boldsymbol{\beta}_0 + \tfrac{1}{n}\|\boldsymbol{\varepsilon}\|^2\\
&\overset{\text{a.s.}}{\to} \boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0 + 2\mathbb{E}(\varepsilon_i)\mathbb{E}(\mathbf{x}_i^T)\boldsymbol{\beta}_0 + \mathbb{E}(\varepsilon_i^2) = \boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0 + \sigma_0^2.
\end{aligned}$$

Using (32), we have $\frac{1}{n}\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{b}_n^T\mathbf{A}_n^{-1}\mathbf{b}_n \overset{\text{P}}{\to} \boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0$ and hence $\tau_U^{-1} \overset{\text{P}}{\to} \sigma_0^2$. By the continuous mapping theorem $\tau_U \overset{\text{P}}{\to} \sigma_0^{-2}$.

In a similar manner to τ_U we have

$$\begin{aligned}
\tau_L^{-1} &= \frac{2B + 2\|\mathbf{y}\|^2 + 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + p\frac{2A+n-p}{(2A+n)\tau^{(0)}}}{2A+n-p}\\
&= \left(\frac{n}{2A+n-p}\right)\frac{2B + 2\|\mathbf{y}\|^2 + 2\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} + p\frac{2A+n-p}{(2A+n)\tau^{(0)}}}{n}\\
&\overset{\text{P}}{\to} 2\sigma_0^2 + 4\boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0.
\end{aligned}$$

By the continuous mapping theorem $\tau_L \overset{\text{P}}{\to} [4\boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0 + 2\sigma_0^2]^{-1}$. Hence $\tau^{(t)}, t > 1$ is bounded in probability between two constants. Similarly, $\tau^{(1)}$ is bounded as $\tau^{(0)} = 1$ is $O_p(1)$. □

We will now define what we call "correct models" and the "true model". Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$.

Definition: A *correct model* $\boldsymbol{\gamma}$ is a p-vector with elements such that $\gamma_j \in \{0,1\}$ if $\beta_{0j} = 0$ and $\gamma_j = 1$ if $\beta_{0j} \neq 0$.

Definition: The *true model* $\boldsymbol{\gamma}_0$ is the p-vector with elements such that $\gamma_j = 0$ if $\beta_{0j} = 0$ and $\gamma_j = 1$ if $\beta_{0j} \neq 0$.

Hence, for example, the true model $\boldsymbol{\gamma}_0$ and the full model $\boldsymbol{\gamma} = \mathbf{1}$ are both correct models. We will next derive some properties for correct models. Note that, by definition, $\boldsymbol{\beta}_{0,-\boldsymbol{\gamma}} = \mathbf{0}$ and we denote $j \in \boldsymbol{\gamma}$ if $\gamma_j = 1$ and $j \notin \boldsymbol{\gamma}$ if $\gamma_j = 0$.

Definition: Let $\boldsymbol{\gamma}$ be a correct model. Then we say that $\mathbf{w}^{(t)}$ is "close" to $\boldsymbol{\gamma}$ in probability if

$$w_j^{(t)} = \begin{cases} 1 - d_{nj} & j \in \boldsymbol{\gamma} \\ d_{nj} & j \notin \boldsymbol{\gamma} \end{cases}, 1 \leq j \leq p,$$

where d_{nj}, $1 \leq j \leq p$, is a sequences of positive random variables such that nd_{nj} converges in probability to zero.

In the main results we assume that $\mathbf{w}^{(t)}$ is close to a correct model. Under this assumption we prove, in the following order, that:

- $\boldsymbol{\mu}^{(t)}$ is a consistent estimator of $\boldsymbol{\beta}$;
- $\tau^{(t)}$ is a consistent estimator of σ_0^{-2};
- $\boldsymbol{\Sigma}^{(t)} = \text{cov}(\boldsymbol{\beta}_{\text{LS}}) + \mathbf{O}_p^m(n^{-3/2})$; and
- $\mathbf{w}^{(t+1)}$ is also "close" to the true model in probability.

We can then use these results recursively to obtain similar results for the Tth iteration of the Algorithm 1, where $T > t$. In the next few results we use the following quantities:

$$\begin{aligned}
\mathbf{T}_1 &= \mathbf{T}_2 - \mathbf{T}_3\mathbf{T}_4, \\
\mathbf{T}_2 &= (n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}) \odot (\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{(t)} - \mathbf{1}\mathbf{1}^T) + (n\tau^{(t-1)}\sigma_\beta^2)^{-1}\mathbf{I}, \\
\mathbf{T}_3 &= (n\tau^{(t-1)}\sigma_\beta^2)\mathbf{T}_4^T\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}\left[\mathbf{I} + \mathbf{T}_5\right]^{-1}, \\
\mathbf{T}_4 &= \mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}, \\
\mathbf{T}_5 &= (n\tau^{(t-1)}\sigma_\beta^2)(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}) \odot \boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)} \quad \text{and} \\
\mathbf{t}_1 &= (\mathbf{W}_{\boldsymbol{\gamma}}^{(t)} - \mathbf{I})(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}) - \mathbf{T}_3\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{y}).
\end{aligned} \tag{33}$$

Result 6. *Assume (A1)–(A5) hold. Let* $\boldsymbol{\gamma}$ *be a correct model. Suppose that* $\mathbf{w}^{(t)}$

is close to $\boldsymbol{\gamma}$ *then*

$$\begin{aligned}
\mathbf{T}_1 &= \mathbf{O}_p^m(n^{-1}),\\
\mathbf{T}_2 &= \mathbf{O}_p^m(n^{-1}),\\
\mathbf{T}_3 &= \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty),\\
\mathbf{T}_4 &= \mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty),\\
\mathbf{T}_5 &= \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \quad \textit{and}\\
\mathbf{t}_1 &= \mathbf{O}_p^v(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

Proof of Result 6: Firstly,

$$\begin{aligned}
\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{(t)} - \mathbf{1}\mathbf{1}^T &= (\mathbf{w}_{\boldsymbol{\gamma}}^{(t)})(\mathbf{w}_{\boldsymbol{\gamma}}^{(t)})^T + \mathbf{W}_{\boldsymbol{\gamma}}^{(t)}(\mathbf{I} - \mathbf{W}_{\boldsymbol{\gamma}}^{(t)}) - \mathbf{1}\mathbf{1}^T\\
&= (\mathbf{1} - \mathbf{d}_{n,\boldsymbol{\gamma}})(\mathbf{1} - \mathbf{d}_{n,\boldsymbol{\gamma}})^T + (\mathbf{I} - \mathbf{D}_{n,\boldsymbol{\gamma}})\mathbf{D}_{n,\boldsymbol{\gamma}} - \mathbf{1}\mathbf{1}^T\\
&= \mathbf{d}_{n,\boldsymbol{\gamma}}\mathbf{d}_{n,\boldsymbol{\gamma}}^T - \mathbf{1}\mathbf{d}_{n,\boldsymbol{\gamma}}^T - \mathbf{d}_{n,\boldsymbol{\gamma}}\mathbf{1}^T + (\mathbf{I} - \mathbf{D}_{n,\boldsymbol{\gamma}})\mathbf{D}_{n,\boldsymbol{\gamma}}\\
&= \mathbf{O}_p^m(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty)
\end{aligned}$$

where $\mathbf{D}_{n,\boldsymbol{\gamma}} = \mathrm{diag}(\mathbf{d}_{n,\boldsymbol{\gamma}})$. Similarly, $\boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)} = \mathbf{O}_p^d(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Again, using (32) and Result 5 we have $\mathbf{T}_2 = [\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}_p^m(n^{-1/2})] \odot \mathbf{O}_p^m(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty) + \mathbf{O}_p^d(n^{-1}) = \mathbf{O}_p^m(n^{-1} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty)$. Next, using (32) and Result 5 we have

$$\begin{aligned}
\mathbf{T}_5 &= (n\tau^{(t-1)}\sigma_\beta^2)(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}) \odot \boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)}\\
&= O_p(n)[\mathbf{S}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}} + \mathbf{O}_p(n^{-1/2})] \odot \mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\\
&= \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

Expanding and simplifying the above equation obtains the result for $\mathbf{T}_5$. Now since, using the assumption of $\mathbf{w}^{(t)}$ being close to $\boldsymbol{\gamma}$ we have $n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty = o_p(1)$ and so $\mathbf{T}_5 = \mathbf{o}_p^m(1)$. By the continuous mapping theorem, we have $(\mathbf{I}+\mathbf{T}_5)^{-1} = \mathbf{I} + \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Next, $\mathbf{T}_4 = \mathbf{D}_{n,-\boldsymbol{\gamma}}[\mathbf{S}_{-\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}_p^m(n^{-1/2})](\mathbf{I} - \mathbf{D}_{n,\boldsymbol{\gamma}})$.

Expanding and simplifying the above expression obtains the result for $\mathbf{T}_4$. Furthermore,

$$\mathbf{T}_3 = n\tau^{(t-1)}\sigma_\beta^2\mathbf{T}_4^T(\mathbf{I} + \mathbf{T}_5)^{-1} = O_p(n)\mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)[\mathbf{I} + \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)].$$

Expanding and simplifying the above expression obtains the result for $\mathbf{T}_3$. Substituting the order expressions for $\mathbf{T}_2$, $\mathbf{T}_3$ and $\mathbf{T}_4$ in the expression for $\mathbf{T}_1$. Then expanding and simplifying obtains the result for $\mathbf{T}_1$. Finally, using (32) we have

$$\begin{aligned}
\mathbf{t}_1 &= (\mathbf{W}_{\boldsymbol{\gamma}}^{(t)} - \mathbf{I})(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}) - \mathbf{T}_3\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{y})\\
&= \mathbf{O}_p^m(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty)[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\boldsymbol{\beta}_{0,\boldsymbol{\gamma}} + \mathbf{O}_p^v(n^{-1/2})]\\
&\qquad -\mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\mathbf{O}_p^m(\|\mathbf{d}_{n-\boldsymbol{\gamma}}\|_\infty)[\mathbf{S}_{-\boldsymbol{\gamma},\boldsymbol{\gamma}}\boldsymbol{\beta}_{0,\boldsymbol{\gamma}} + \mathbf{O}_p^v(n^{-1/2})]\\
&= \mathbf{O}_p^v(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty + n\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)
\end{aligned}$$

which simplifies to the result for $\mathbf{t}_1$ under the assumption that $d_{nj} = o_p(n^{-1})$.□

Result 7. *Assume (A1)–(A5) hold. If* $\mathbf{w}^{(t)}$ *is close to a correct model* $\boldsymbol{\gamma}$ *in probability then*

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{(t)} &= \boldsymbol{\beta}_{0,\boldsymbol{\gamma}} + \mathbf{O}_p^v(n^{-1/2}),\\
\boldsymbol{\mu}_{-\boldsymbol{\gamma}}^{(t)} &= \mathbf{O}_p^v(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty),\\
\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{(t)} &= (n\tau^{(t-1)})^{-1}\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{-1} + \mathbf{O}_p^m(n^{-3/2})\\
\boldsymbol{\Sigma}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)} &= \sigma_\beta^2\mathbf{I} + \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \qquad \textit{and}\\
\boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)} &= \mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

Proof of Result 7: Firstly, note that

$$\tau(\mathbf{X}^T\mathbf{X})\odot\boldsymbol{\Omega} + \sigma_\beta^{-2}\mathbf{I} = \tau\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W} + \mathbf{D} \tag{34}$$

by definition. Using equations (23), (24), (34) and Result 5 we have

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{(t)} \quad = \quad & \left(n\tau^{(t-1)}\right)^{-1}\Big[(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})\odot\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{(t)} + (n\tau^{(t-1)}\sigma_\beta^2)^{-1}\mathbf{I}\\
& -\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}})\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}\\
& \times\left\{(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}})\odot\boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)} + (n\tau^{(t-1)}\sigma_\beta^2)^{-1}\mathbf{I}\right\}^{-1}\\
& \times\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}\Big]^{-1}\\
= \quad & \left(n\tau^{(t-1)}\right)^{-1}\Big[(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_1\Big]^{-1}\\
= \quad & \left(n\tau^{(t-1)}\right)^{-1}\Big[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}_p^m(n^{-1/2}) + \mathbf{O}_p^m(n^{-1})\Big]^{-1}\\
= \quad & (n\tau^{(t-1)})^{-1}\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{-1} + \mathbf{O}_p^m(n^{-3/2}) = \mathbf{O}_p^m(n^{-1}).
\end{aligned}$$

Using equations (23), (29), (32) and (34), Result 6, and the continuous mapping theorem we have

$$\begin{aligned}
\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{(t)} &= \tau^{(t-1)}\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{(t)}\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}\mathbf{X}_{\boldsymbol{\gamma}}^T(\mathbf{I}-\mathbf{P}_{\boldsymbol{\gamma}}^{(t)})\mathbf{y}\\
&= \Big[(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_1\Big]^{-1}\Big[\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}) - \mathbf{W}_{\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}})\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}\\
&\quad \times\left\{(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}})\odot\boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)} + (n\tau^{(t-1)}\sigma_\beta^2)^{-1}\mathbf{I}\right\}^{-1}\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{y})\Big]\\
&= \Big[(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_1\Big]^{-1}\Big[(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}) + \mathbf{t}_1\Big]\\
&= \Big[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{-1} + \mathbf{O}_p^m(n^{-1/2} + \|\mathbf{T}_1\|_\infty)\Big]\Big[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\boldsymbol{\beta}_{0,\boldsymbol{\gamma}} + \mathbf{O}_p^v(n^{-1/2} + \|\mathbf{t}_1\|_\infty)\Big]\\
&= \boldsymbol{\beta}_{0,\boldsymbol{\gamma}} + \mathbf{O}_p^v(n^{-1/2} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

Since by assumption $\|\mathbf{d}_n\|_\infty = o_p(n^{-1})$ we have $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{(t)}$ as stated. Using equations (23), (24) and (34) we have

$$\begin{aligned}
\boldsymbol{\Sigma}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)} \quad &= \Big[\sigma_\beta^{-2}\mathbf{I} + \tau^{(t-1)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}})\odot(n\boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)})\\
&\qquad -n\mathbf{T}_4\left\{(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_2\right\}^{-1}\mathbf{T}_4^T\Big]^{-1}.
\end{aligned}$$

From Equation (32) and Result 6, we can show that $(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})+\mathbf{T}_2=\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}+\mathbf{O}_p^m(n^{-1/2})$ and $\mathbf{T}_4=\mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Using the continuous mapping theorem we find that

$$\begin{aligned}
&\tau^{(t-1)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}})\odot(n\boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)})-n\mathbf{T}_4\left[(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})+\mathbf{T}_2\right]^{-1}\mathbf{T}_4^T\\
&\quad=O_p(1)[\mathbf{S}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}+\mathbf{O}_p^m(n^{-1/2})]\odot\mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\\
&\qquad\quad-n\mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\left[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}+\mathbf{O}_p^m(n^{-1/2})\right]\mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\\
&\quad=\mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

Noting that by assumption $d_{nj}=o_p(n^{-1})$ and applying the continuous mapping theorem, we obtain the result for $\boldsymbol{\Sigma}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)}$. Next, from equations (23), (29), (32) and Result 6 we obtain

$$\begin{aligned}
\boldsymbol{\mu}_{-\boldsymbol{\gamma}}^{(t)}\quad&=\tau^{(t-1)}\boldsymbol{\Sigma}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)}\Big[(n\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)})(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{y})\\
&\qquad-n\mathbf{T}_4\left\{(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})+\mathbf{T}_2\right\}^{-1}\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y})\Big]\\
&=O_p(1)[\sigma_\beta^2\mathbf{I}+\mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)]\\
&\qquad\times\Big[[\mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)][\mathbf{S}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}\boldsymbol{\beta}_{0,-\boldsymbol{\gamma}}+\mathbf{O}_p^v(n^{-1/2})]\\
&\qquad-n[\mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)][\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}+\mathbf{O}_p^m(n^{-1/2})]\\
&\qquad\times[\mathbf{I}-\mathbf{O}_p^d(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty)][\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\boldsymbol{\beta}_{0,\boldsymbol{\gamma}}+\mathbf{O}_p^v(n^{-1/2})]\Big]\\
&=\mathbf{O}_p^v(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

Lastly, using equations (23), (25), (32), Result 6 and by the assumption that $d_{nj}=o_p(n^{-1})$ we obtain

$$\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)}\quad&=-\boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{(t)}\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}\left[\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}\odot\boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)}+(\tau^{(t-1)}\sigma_\beta^2)^{-1}\mathbf{I}\right]^{-1}\\
&=-\left[(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})+\mathbf{T}_1\right]^{-1}n^{-1}\tau^{(t-1)^{-1}}\mathbf{T}_3\\
&=\left[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}+\mathbf{O}_p^m(n^{-1/2})\right]^{-1}\times\mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

After expanding the above expression and dropping appropriate lower order terms the result is proved. □

Result 8. *Assume (A1)–(A5) hold. If* $\mathbf{w}^{(t)}$ *is close to a correct model* $\boldsymbol{\gamma}$ *then* $\tau^{(t)}=\sigma_0^{-2}+O_p(n^{-1/2})$.

Proof of Result 8: In Algorithm 1 the value $\tau^{(t)}$ satisfies

$$\begin{aligned}
\tau^{(t)}\quad&=(2A+n)\Big[2B+\|\mathbf{y}-\mathbf{X}\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)}\|^2\\
&\qquad+(\boldsymbol{\mu}^{(t)})^T[(\mathbf{X}^T\mathbf{X})\odot\mathbf{W}^{(t)}\odot(\mathbf{I}-\mathbf{W}^{(t)})]\boldsymbol{\mu}^{(t)}\\
&\qquad+\mathrm{dof}((\tau^{(t-1)})^{-1}\sigma_\beta^{-2},\mathbf{w}^{(t)})/\tau^{(t-1)}\Big]^{-1}\\
&=\frac{1+2A/n}{2B/n+T_1+T_2+T_3},
\end{aligned}$$

where

$$T_1 = n^{-1}\tau^{(t-1)^{-1}}\text{dof}((\tau^{(t-1)})^{-1}\sigma_\beta^{-2}, \mathbf{w}^{(t)}), \quad T_2 = n^{-1}\|\mathbf{y} - \mathbf{X}\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)}\|^2$$
$$\text{and} \quad T_3 = (\boldsymbol{\mu}^{(t)})^T[\mathbf{A}_n \odot \mathbf{W}^{(t)} \odot (\mathbf{I} - \mathbf{W}^{(t)})]\boldsymbol{\mu}^{(t)}.$$

Firstly, $T_1 = O_p(n^{-1})$ follows from results 2 and 5. Secondly, using $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$ we have

$$\begin{aligned} T_2 &= n^{-1}\|\boldsymbol{\varepsilon} + \mathbf{X}(\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)} - \boldsymbol{\beta}_0)\|^2 \\ &= n^{-1}\|\boldsymbol{\varepsilon}\|^2 + 2(n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X})(\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)} - \boldsymbol{\beta}_0) \\ &\qquad + (\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)} - \boldsymbol{\beta}_0)^T(n^{-1}\mathbf{X}^T\mathbf{X})(\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)} - \boldsymbol{\beta}_0). \end{aligned}$$

Using Equation (32) we have $n^{-1}\|\boldsymbol{\varepsilon}\|^2 = \sigma_0^2 + O_p(n^{-1/2})$ and $n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X} = \mathbf{O}_p^v(n^{-1/2})$ and $n^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{S} + \mathbf{O}_p^m(n^{-1/2})$. Note that from Result 7 we have $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{(t)} = \boldsymbol{\beta}_{0\boldsymbol{\gamma}} + \mathbf{O}_p^v(n^{-1/2})$ and $\boldsymbol{\mu}_{-\boldsymbol{\gamma}}^{(t)} = \mathbf{O}_p^v(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Then $\boldsymbol{\mu}^{(t)} = \boldsymbol{\beta}_0 + \mathbf{e}_n$ where $\mathbf{e}_{n,\boldsymbol{\gamma}} = \mathbf{O}_p^v(n^{-1/2})$ and $\mathbf{e}_{n,-\boldsymbol{\gamma}} = \mathbf{O}_p^v(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Lastly, by assumption

$$\begin{aligned} \mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)} - \boldsymbol{\beta}_0 &= \left[\begin{array}{c} \mathbf{e}_{n,\boldsymbol{\gamma}} - \mathbf{d}_{n,\boldsymbol{\gamma}} \odot \boldsymbol{\beta}_{0,\boldsymbol{\gamma}} - \mathbf{d}_{n,\boldsymbol{\gamma}} \odot \mathbf{e}_{n,\boldsymbol{\gamma}} \\ \mathbf{d}_{n,-\boldsymbol{\gamma}} \odot \mathbf{e}_{n,-\boldsymbol{\gamma}} \end{array}\right] \\ &= \left[\begin{array}{c} \mathbf{O}_p^v(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty + \|\mathbf{e}_{n,\boldsymbol{\gamma}}\|_\infty) \\ \mathbf{O}_p^v(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\mathbf{e}_{n,-\boldsymbol{\gamma}}\|_\infty) \end{array}\right] \\ &= \left[\begin{array}{c} \mathbf{O}_p^v(n^{-1/2} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty) \\ \mathbf{O}_p^v(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty^2) \end{array}\right]. \end{aligned}$$

Hence, $T_2 = \sigma_0^2 + O_p(n^{-1/2} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty + n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty^2)$. By assumption $\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty$ and $n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty^2$ are of smaller order than $n^{-1/2}$ so $T_2 = \sigma_0^2 + O_p(n^{-1/2})$. Next,

$$T_3 = \sum_{j=1}^p (n^{-1}\|\mathbf{x}_j\|^2)(w_j^{(t)}(1 - w_j^{(t)}))(\mu_j^{(t)})^2.$$

Using (32) we have $n^{-1}\|\mathbf{x}_j\|^2 = \mathbb{E}(x_j^2) + O_p(n^{-1/2})$. Using the assumption for $w_j^{(t)}$ we have $w_j^{(t)}(1 - w_j^{(t)}) = d_{nj}(1 - d_{nj}) = O_p(d_{nj})$ and from Result 7 we have $(\mu_j^{(t)})^2 = \beta_{0j}^2 + O_p(e_{nj})$. Hence, $T_3 = O_p(\|\mathbf{d}_n\|_\infty)$ and so

$$\tau^{(t)} = \frac{1 + O_p(n^{-1})}{\sigma_0^2 + O_p(n^{-1/2})} = \sigma_0^{-2} + O_p(n^{-1/2}).$$

□

We will now examine the updates for η_j and w_j, $1 \le j \le p$ defined at Line 8–9 in Algorithm 1. Note that $\mathbf{w}^*$ is updated one component, w_j^*, at a time iterating through $j = 1, \ldots, p$ and in each iteration w_j^* is updated using w_j. Also note that $\mathbf{w}^{(t+1)}$ is updated after $\mathbf{w}^*$ is completely updated. Here we will employ the notation $\boldsymbol{\gamma}^* \subseteq \boldsymbol{\gamma}$ to mean that $\gamma_j = \gamma_j^*$ for all $j \in \boldsymbol{\gamma}^*$. We will now show that, provided the $\mathbf{w}^{(t)}$ is close to a correct model $\boldsymbol{\gamma}$ and $\mathbf{w}^*$ is close to a correct model $\boldsymbol{\gamma}^*$ such that $\gamma^* \subseteq \boldsymbol{\gamma}$ then the next updated $\mathbf{w}^*$ is close to a correct model $\widetilde{\boldsymbol{\gamma}}$ such that $\widetilde{\boldsymbol{\gamma}} \subseteq \boldsymbol{\gamma}$.

Result 9. *Assume (A1)-(A6). If* $\mathbf{w}^{(t)}$ *and* $\mathbf{w}^*$ *are close to two correct models* $\boldsymbol{\gamma}$ *and* $\boldsymbol{\gamma}^*$ *respectively, such that* $\boldsymbol{\gamma}^* \subseteq \boldsymbol{\gamma}$*, i.e.,*

$$w_k^{(t)} = \begin{cases} 1 - d_{nk} & k \in \boldsymbol{\gamma} \\ d_{nk} & k \notin \boldsymbol{\gamma} \end{cases}$$

and

$$w_k^* = \begin{cases} 1 - d_{nk}^* & k \in \boldsymbol{\gamma}^* \\ d_{nk}^* & k \notin \boldsymbol{\gamma}^* \end{cases} \qquad 1 \le k \le p,$$

where $\{d_{nk}\}_{1\le k\le p}$ *and* $\{d_{nk}^*\}_{1\le k\le p}$ *by Definition 1 are two sequences of positive random variables such that* nd_{nk} *and* nd_{nk}^* *are converging in probability to zero. Then if the next* $\mathbf{w}^*$ *is to update* w_j^**, we have*

$$\eta_j \quad = \begin{cases} \lambda_n + \frac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\beta_{0j}^2 + O_p(n^{1/2}) & j \in \boldsymbol{\gamma}_0 \\ \lambda_n + O_p(1) & j \in \boldsymbol{\gamma} \text{ and } j \notin \boldsymbol{\gamma}_0 \\ \lambda_n - \frac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\sigma_\beta^2 + O_p(n^{1/2} + n^2\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) & j \notin \boldsymbol{\gamma} \end{cases}$$

and after updating $\mathbf{w}^*$ *with* $w_j^* = w_j = \text{expit}(\eta_j)$ *in Algorithm 1 the* $\mathbf{w}^*$ *is close to a correct model* $\widetilde{\boldsymbol{\gamma}}$ *where for* $1 \le k \le p$ *we have*

$$w_k^* = \begin{cases} 1 - \widetilde{d}_{nk} & j = k \text{ and } k \in \boldsymbol{\gamma}_0 \\ \widetilde{d}_{nk} & j = k \text{ and } k \notin \boldsymbol{\gamma}_0 \\ 1 - d_{nk}^* & j \neq k \text{ and } k \in \boldsymbol{\gamma}^* \\ d_{nk}^* & j \neq k \text{ and } k \notin \boldsymbol{\gamma}^* \end{cases}$$

and

$$\widetilde{\gamma}_k = \begin{cases} 1 & j = k \text{ and } k \in \boldsymbol{\gamma}_0 \\ 0 & j = k \text{ and } k \notin \boldsymbol{\gamma}_0 \\ \gamma_k^* & \text{otherwise} \end{cases}$$

where

- $\widetilde{d}_{nj} = \exp\left[-\frac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\beta_{0j}^2 - \lambda_n + O_p(n^{1/2})\right]$ *if* $j \in \boldsymbol{\gamma}_0$*;*
- $\widetilde{d}_{nj} = O_p(\exp(\lambda_n))$ *if* $j \in \boldsymbol{\gamma}$ *and* $j \notin \boldsymbol{\gamma}_0$*; and*
- $\widetilde{d}_{nj} = \exp\left[\lambda_n - \frac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\sigma_\beta^2 + O_p(n^{1/2} + n^2\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\right]$ *if* $j \notin \boldsymbol{\gamma}$.

Proof of Result 9: Consider the update η_j at Line 11 in Algorithm 1. From results 7 and 8 we have that if $\mathbf{w}^{(t)}$ is close to a correct model $\boldsymbol{\gamma}$ then

$$\begin{aligned} \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{(t)} &= \boldsymbol{\beta}_{0,\boldsymbol{\gamma}} + \mathbf{O}_p^v(n^{-1/2}), \\ \boldsymbol{\mu}_{-\boldsymbol{\gamma}}^{(t)} &= \mathbf{O}_p^v(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty), \\ \boldsymbol{\Sigma}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)} &= \sigma_\beta^2\mathbf{I} + \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \\ \boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}}^{(t)} &= \mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty), \\ \tau^{(t)} &= \sigma_0^{-2} + O_p(n^{-1/2}) \quad \text{and} \\ \boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}^{(t)} &= \mathbf{O}_p^m(n^{-1}). \end{aligned}$$

We also have the fact that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}$. Hence,

$$\begin{aligned} \eta_j &= \lambda - \tfrac{1}{2}\tau^{(t)}\Big((\mu_j^{(t)})^2 + \Sigma_{j,j}^{(t)}\Big)\|\mathbf{X}_j\|^2 \\ &\quad + \tau^{(t)}\mathbf{X}_j^T\Big[(\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon})\mu_j^{(t)} - \mathbf{X}_{-j}\mathrm{diag}(\mathbf{w}_{-j}^*)\big(\boldsymbol{\mu}_{-j}^{(t)}\mu_j^{(t)} + \boldsymbol{\Sigma}_{-j,j}^{(t)}\big)\Big] \\ &= \lambda + T_6 - T_7 + T_8 - T_9 + T_{10} \end{aligned}$$

where

$$\begin{aligned} T_6 &= \tau^{(t)}\|\mathbf{X}_j\|^2\left(\beta_{0,j}\mu_j^{(t)} - \tfrac{1}{2}(\mu_j^{(t)})^2\right), \\ T_7 &= \tfrac{1}{2}\tau^{(t)}\|\mathbf{X}_j\|^2\Sigma_{j,j}^{(t)} \\ T_8 &= \tau^{(t)}\mu_j^{(t)}\sum_{k\neq j}\mathbf{X}_j^T\mathbf{X}_k\left(\beta_{0,k} - w_k^*\mu_k^{(t)}\right), \\ T_9 &= \tau^{(t)}\sum_{k\neq j}\mathbf{X}_j^T\mathbf{X}_k w_k^*\Sigma_{k,j}^{(t)} \\ \text{and}\quad T_{10} &= \tau^{(t)}\mu_j^{(t)}\mathbf{X}_j^T\boldsymbol{\epsilon}. \end{aligned}$$

Firstly,

- If $j \in \boldsymbol{\gamma}_0$ then

$$\begin{aligned} T_6 &= \tfrac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]\left[\beta_{0j}^2 + O_p(n^{-1/2})\right] \\ &= \tfrac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\beta_{0j}^2 + O_p(n^{1/2}). \end{aligned}$$

- If $j \in \boldsymbol{\gamma}$ and $j \notin \boldsymbol{\gamma}_0$ then

$$T_6 = \tfrac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]O_p(n^{-1}) = |O_p(1)|.$$

- If $j \notin \boldsymbol{\gamma}$ then

$$\begin{aligned} T_6 &= \tfrac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]O_p(n^2\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty^2) \\ &= |O_p(n^3\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty^2)|. \end{aligned}$$

Secondly,

- If $j \in \boldsymbol{\gamma}$ then

$$T_7 = \tfrac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]O_p(n^{-1}) = |O_p(1)|.$$

- If $j \notin \boldsymbol{\gamma}$ then

$$\begin{aligned} T_7 &= \tfrac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]\left[\sigma_\beta^2 + O_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\right] \\ &= \tfrac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\sigma_\beta^2 + O_p(n^{1/2} + n^2\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \end{aligned}$$

Next note that since $\boldsymbol{\gamma}^* \subseteq \boldsymbol{\gamma}$ we have $\gamma_k = \gamma_k^* = 1$ for all $k \in \boldsymbol{\gamma}^*$ so that if $k \in \boldsymbol{\gamma}^*$ then $k \in \boldsymbol{\gamma}$ and the consequently the combination $k \in \boldsymbol{\gamma}^*$ then $k \notin \boldsymbol{\gamma}$ is not possible.

Next consider $\beta_{0,k} - w_k^*\mu_k^{(t)}$.

- If $k \in \boldsymbol{\gamma}$ and $k \in \boldsymbol{\gamma}^*$ then

$$\beta_{0,k} - w_k^* \mu_k^{(t)} = \beta_{0,k} - (1 - d_{nk}^*)(\beta_{0,k} + O_p(n^{-1/2})) = O_p(n^{-1/2}).$$

- If $k \in \boldsymbol{\gamma}$ and $k \notin \boldsymbol{\gamma}^*$ then

$$\beta_{0,k} - w_k^* \mu_k^{(t)} = \beta_{0,k} - d_{nk}^*(\beta_{0,k} + O_p(n^{-1/2})) = O_p(d_{nk}^* n^{-1/2}).$$

- If $k \notin \boldsymbol{\gamma}$ and $k \notin \boldsymbol{\gamma}^*$ then

$$\beta_{0,k} - w_k^* \mu_k^{(t)} = \beta_{0,k} - d_{nk}^* O_p(n \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) = O_p(n d_{nk}^* \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).$$

noting that if $k \notin \boldsymbol{\gamma}$ or if $k \notin \boldsymbol{\gamma}^*$ then $\beta_{0,k} = 0$ (since both $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ are correct models). Hence,

$$\begin{aligned} &\sum_{k \neq j} \mathbf{X}_j^T \mathbf{X}_k \left(\beta_{0,k} - w_k^* \mu_k^{(t)}\right) \\ &\quad = n \left[\mathbb{E}(x_j x_k) + O_p\big(n^{-1/2}\big)\right] O_p(n^{-1/2}) \\ &\quad = O_p(n^{1/2}). \end{aligned}$$

The second line follows from the assumption that $d_{nj} = o_p(n^{-1})$ and $d_{nj}^* = o_p(n^{-1})$ for all j so that $O_p(n^2 d_{nk}^* \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) = o_p(1)$. Hence,

$$T_8 \quad = \begin{cases} O_p(n^{1/2}) & j \in \boldsymbol{\gamma}_0 \\ O_p(1) & j \in \boldsymbol{\gamma}, j \notin \boldsymbol{\gamma}_0 \\ O_p(n^{3/2 \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty}) & j \notin \boldsymbol{\gamma}. \end{cases}$$

Next consider $w_k^* \Sigma_{k,j}$.

- If $j \in \boldsymbol{\gamma}, k \in \boldsymbol{\gamma}$ and $k \in \boldsymbol{\gamma}^*$ then

$$w_k^* \Sigma_{k,j} = (1 - d_{nk}^*) O_p(n^{-1}) = O_p(n^{-1}).$$

- If $j \in \boldsymbol{\gamma}, k \in \boldsymbol{\gamma}$ and $k \notin \boldsymbol{\gamma}^*$ then

$$w_k^* \Sigma_{k,j} = O_p(d_{nk}^* n^{-1}) = o_p(n^{-2}).$$

- If $j \in \boldsymbol{\gamma}, k \notin \boldsymbol{\gamma}$ and $k \notin \boldsymbol{\gamma}^*$ then

$$w_k^* \Sigma_{k,j} = O_p(d_{nk}^* \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) = o_p(n^{-2}).$$

- If $j \notin \boldsymbol{\gamma}, k \in \boldsymbol{\gamma}$ and $k \in \boldsymbol{\gamma}^*$ then

$$w_k^* \Sigma_{k,j} = (1 - d_{nk}^*) O_p(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) = O_p(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).$$

- If $j \notin \boldsymbol{\gamma}, k \in \boldsymbol{\gamma}$ and $k \notin \boldsymbol{\gamma}^*$ then

$$w_k^* \Sigma_{k,j} = O_p(d_{nk}^* \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).$$

- If $j \notin \boldsymbol{\gamma}, k \notin \boldsymbol{\gamma}$ and $k \notin \boldsymbol{\gamma}^*$ then

$$w_k^* \Sigma_{k,j} = d_{nk}^*(\sigma_\beta^2 + O_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)) = O_p(d_{nk}^*).$$

Hence,

$$T_9 = \begin{cases} O_p(1) & j \in \boldsymbol{\gamma} \\ o_p(1) & j \notin \boldsymbol{\gamma}. \end{cases}$$

Finally by noting that $\mathbf{X}_j$ is independent to $\boldsymbol{\epsilon}$, we have

$$T_{10} = \begin{cases} (\beta_{0,k} + O_p(n^{-1/2}))O_p(n^{1/2}) = O_p(n^{1/2}) & j \in \boldsymbol{\gamma} \text{ and } j \in \boldsymbol{\gamma}_0, \\ O_p(n^{-1/2})O_p(n^{1/2}) = O_p(1) & j \in \boldsymbol{\gamma} \text{ and } j \notin \boldsymbol{\gamma}_0, \\ O_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)O_p(n^{1/2}) = O_p(n^{3/2}\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) & j \notin \boldsymbol{\gamma}, \end{cases}$$

From T_6, T_7, T_8, T_9 and T_{10} we obtain the expression for η_j^* in the result. From Lemma 1 we have

$$w_k^* = \begin{cases} 1 - \widetilde{d}_{nk} & k \in \boldsymbol{\gamma}_0 \\ \widetilde{d}_{nk} & k \notin \boldsymbol{\gamma}_0, \end{cases}$$

where all possible $\widetilde{d}_{nj}$ are specified in the result.

Now note that

- If $\lambda_n > 0$ and $\lambda_n \to \infty$ as $n \to \infty$ then w_j^* defined in Algorithm 1 will converge in probability to 1.
- If λ_n is $O_p(1)$ then d_{nj} will converge to zero at a faster rate than required for $j \in \boldsymbol{\gamma}_0$, for $j \notin \boldsymbol{\gamma}_0$ the value w_j^* will be $O_p(1)$.
- If $\lambda_n < 0$ and $\lambda_n/n \to \kappa$ for some constant κ then w_j^* may not converge in probability to 1 depending on the size of κ.
- If $\lambda_n < 0$ and λ_n grows at a faster rate than $O_p(n)$ then w_j^* will converge in probability to 0.
- If $\lambda_n \to -\infty$ and $\lambda_n/n \to 0$ then d_{nj} will converge to 0, but for $j \notin \boldsymbol{\gamma}_0$ the sequence $n\widetilde{d}_{nj}$ may not converge in probability to zero.

Thus, we require $\lambda_n \to -\infty$, $\lambda_n/n \to 0$ and $n\,\text{expit}(\lambda_n) = n\rho_n \to 0$. These are the conditions specified by Assumption (A6). Thus, under Assumption (A6) the vector $\mathbf{w}^*$ will be close to the model $\widetilde{\boldsymbol{\gamma}}$. □

Proof of Main Result 2: If $w_j^{(1)} = 1$ for $1 \le j \le p$ and assumptions (A1)–(A6) hold then $\mathbf{w}^{(1)} = \mathbf{1}$ corresponds to a correct model $\boldsymbol{\gamma} = \mathbf{1}$ and results 7–8 hold with $d_{nj} = 0$ for $1 \le j \le p$. Applying Result 9 repeatedly over indexes $1 \le j \le p$ obtains the result for $\mathbf{w}^{(2)}$ with the sequence $d_{nj} = 0$ for $1 \le j \le p$ where the convergence rate of nd_{nj} being satisfied. Hence, equations (10) and (11) are proved. We can then apply results 7–8 to prove Equation (12). We now note that the term $n^2\,\text{expit}(\lambda_n)$ is $o_p(n)$ or smaller by Assumption (A6). However, by Assumption (A6) this term and λ_n in $w_j^{(3)}$ with $j \notin \boldsymbol{\gamma}_0$ are dominated by $-n\mathbb{E}(x_j^2)\sigma_\beta^2/2\sigma_0^2$. Thus, we have $w_j^{(3)} = 1 - d_{nj}$ for $j \in \boldsymbol{\gamma}_0$ and $w_j^{(3)} = d_{nj}$ for

$j \notin \boldsymbol{\gamma}_0$ where d_{nj} are sequences of random variables with $n^2 d_{nj}$ converging in probability to zero. Thus, after applying Results 6–8 repeatedly over indexes $1 \leq j \leq p$ the equations (14) and (15) are proved for $t = 3$. However, these results give rise to the same conditions for $t = 4$ as those required for $t = 3$. Thus, we can continue applying Results 6–8 recursively to prove the Main Result 2 for all t. □

Appendix C: Deriviation of Algorithm 3

The q-densities corresponding to Algorithm 3 are:

$$\begin{aligned} q(\boldsymbol{\beta}) \quad & \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\left\{\textstyle\sum_{i=1}^n -\frac{a_i}{2\sigma^2}||y_i - \mathbf{x}_i^T\boldsymbol{\Gamma}\boldsymbol{\beta}||^2 - \frac{||\boldsymbol{\beta}||^2}{2\sigma_\beta^2}\right\}\right] \\ & \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\left\{-\tfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta})^T\mathrm{diag}(\mathbf{a})(\mathbf{y} - \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}) - \tfrac{||\boldsymbol{\beta}||^2}{2\sigma_\beta^2}\right\}\right] \\ & = \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \end{aligned}$$

where $\boldsymbol{\Sigma} = (\tau(\mathbf{X}^T\widetilde{\mathbf{A}}\mathbf{X}) \odot \boldsymbol{\Omega} + \tau_\beta\mathbf{I})^{-1}$, $\boldsymbol{\mu} = \tau\boldsymbol{\Sigma}\mathbf{W}\mathbf{X}^T\widetilde{\mathbf{A}}\mathbf{y}$, $\mathbf{w} = \mathbb{E}_q\boldsymbol{\gamma}$, $\mathbf{W} = \mathrm{diag}(\mathbf{w})$, $\boldsymbol{\Omega} = \mathbf{w}\mathbf{w}^T + \mathbf{W} \odot (\mathbf{I} - \mathbf{W})$, $\tau = \mathbb{E}_q(1/\sigma^2)$, $\tau_\beta = \sigma_\beta^{-2}$ and $\widetilde{\mathbf{A}} = \mathbb{E}_q\mathrm{diag}(\mathbf{a})$. Similarly, we have

$$\begin{aligned} q(\sigma^2) \quad \propto \exp\Big[\mathbb{E}_{-q(\sigma^2)}\Big\{ & -\tfrac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta})^T\mathrm{diag}(\mathbf{a})(\mathbf{y} - \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}) \\ & -\frac{n}{2}\log\sigma^2 - (A+1)\log\sigma^2 - \frac{B}{\sigma^2}\Big\}\Big] \end{aligned}$$

Hence, $q(\sigma^2) = \text{Inverse-Gamma}(A + \frac{n}{2}, s)$, where

$$s \quad = B + \frac{1}{2}\left[\mathbf{y}^T\widetilde{\mathbf{A}}\mathbf{y} - 2\mathbf{y}^T\widetilde{\mathbf{A}}\mathbf{X}\mathbf{W}\boldsymbol{\mu} + \mathrm{tr}\left((\mathbf{X}^T\widetilde{\mathbf{A}}\mathbf{X} \odot \boldsymbol{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\right)\right].$$

Next noting that $\gamma_j = \gamma_j^2$ as $\gamma_j \in \{0, 1\}$, the optimal $q(\gamma_j)$, $1 \leq j \leq p$, takes the form

$$q(\gamma_j) \quad \propto \exp\left[\gamma_j\,\mathbb{E}_{-q_{3j}}\left\{\lambda + \tfrac{\beta_j}{\sigma^2}\mathbf{X}_j^T\widetilde{\mathbf{A}}\big(\mathbf{y} - \mathbf{X}_{-j}\mathbf{W}_{-j}\boldsymbol{\beta}_{-j}\big) - \tfrac{\beta_j^2}{2\sigma^2}\mathbf{X}_j^T\widetilde{\mathbf{A}}\mathbf{X}_j\right\}\right].$$

Hence, $q(\gamma_j) = \mathrm{Bern}(w_j)$, where $w_i = \mathrm{expit}(\eta_j)$ and

$$\eta_j = \lambda - \frac{1}{2}\tau\mathbf{X}_j^T\widetilde{\mathbf{A}}\mathbf{X}_j(\mu_j^2 + \Sigma_{jj}) + \tau\mathbf{X}_j^T\widetilde{\mathbf{A}}\left[\mathbf{y}\mu_j - \mathbf{X}_{-j}\mathbf{W}_{-j}(\boldsymbol{\mu}_{-j}\mu_j + \boldsymbol{\Sigma}_{-j,j})\right].$$

Next, we have

$$\begin{aligned} q(a_j) \quad & \propto \quad \exp\left[\mathbb{E}_{-q(a_j)}\left\{\frac{1}{2}\log a_j - \frac{a_j}{2\sigma^2}||y_j - \mathbf{x}_j^T\boldsymbol{\Gamma}\boldsymbol{\beta}||^2 - 2\log a_j - \frac{1}{2a_j}\right\}\right] \\ & \propto \quad \exp\left[-\frac{3}{2}\log a_j - \frac{a_j}{2}\mathbb{E}_q\frac{1}{\sigma^2}||y_j - \mathbf{x}_j^T\boldsymbol{\Gamma}\boldsymbol{\beta}||^2 - \frac{1}{2a_j}\right] \\ & = \quad \text{Inverse-Gaussian}\left(\widetilde{A}_j, 1\right), \end{aligned}$$

where

$$\begin{aligned}\widetilde{A}_j &= \left(\mathbb{E}_q \frac{1}{\sigma^2}||y_j - \mathbf{x}_j^T \boldsymbol{\Gamma}\boldsymbol{\beta}||^2\right)^{-1/2} \\ &= \tau^{-1/2}\left[y_j^2 - 2y_j\mathbf{x}_j^T\mathbf{W}\boldsymbol{\mu} + \text{tr}\left((\mathbf{x}_j\mathbf{x}_j^T \odot \boldsymbol{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\right)\right]^{-1/2}.\end{aligned}$$

Acknowledgments

This research was partially supported by an Australian Research Council Early Career Award DE130101670 (Ormerod) an Australian Research Council Discovery Project DP130100488 (Müller) and an Australian Postgraduate Award (You). We would like to thank Veronika Ročková for providing code corresponding to the EVMS method.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *In Proceedings of the 2nd International Symposium on Information Theory* 267–281. Akademiai Kiad6, Budapest. MR0483125

[2] Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* **36** 99–102. MR0359122

[3] Ariascastro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electronic Journal of Statistics* **8** 328–354. MR3195119

[4] Bartlett, M. (1957). A Comment on D. V. Lindley's statistical paradox. *Biometrika* **44** 533–534. MR0086727

[5] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York. MR2247587

[6] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (2007). *Discrete multivariate analysis: Theory and Practice*. Springer. MR2344876

[7] Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis* **5** 583–618. MR2719668

[8] Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics* **5** 232–253. MR2810396

[9] Bülmann, P. and van de Geer, S. (2011). *Statistics for High Dimensional Data*. Springer. MR2807761

[10] Carbonetto, P. (2012). varbvs 1.10. Variational inference for Bayesian variable selection. R package. http://cran.r-project.org.

[11] Carbonetto, P. and Stephens, M. (2011). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis* **6** 1–42. MR2896713

[12] Casella, G., Girón, F. J., Martńez, M. L. and Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics* **37** 1207–1228. MR2509072

[13] Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. W. (2014). Bayesian linear regression with sparse priors. *Annals of Statistics* **43** 1986–2018. MR3375874

[14] Castillo, I. and van der Vaart, A. W. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics* **40** 2069–2101. MR3059077

[15] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

[16] Faes, C., Ormerod, J. T. and Wand, M. P. (2011). Variational Bayesian inference for parametric and nonparametric regression with missing data. *Journal of the American Statistical Association* **106** 959–971. MR2894756

[17] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360. MR1946581

[18] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B* **70** 849–911. MR2530322

[19] Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20** 101-148. MR2640659

[20] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32** 928–961. MR2065194

[21] Feldkircher, M. and Zeugner, S. (2009). Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in Bayesian model averaging. *IMF Working Paper* **09/202**.

[22] Feldkircher, M. and Zeugner, S. (2013). BMS 03.3. Bayesian Model Averaging Library. R package. http://cran.r-project.org.

[23] Flandin, G. and Penny, W. D. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage* **34** 1108-1125.

[24] Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The Elements of Statistical Learning*. Springer. MR1851606

[25] Garcia, T. P., Müller, S., Carroll, R. J., Dunn, T. N., Thomas, A. P., Adams, S. H., Pillai, S. D. and Walzem, R. L. (2013). Structured variable selection with q-values. *Biostatistics* **14** 695–707.

[26] Hall, P., Ormerod, J. T. and Wand, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica* **21** 369–389. MR2796867

[27] Hall, P., Pham, T., Wand, M. P. and Wang, S. S. J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *The Annals of Statistics* **39** 2502–2532. MR2906876

[28] Hans, C., Dobra, A. and West, M. (2007). Shotgun stochastic search

for "large p" regression. *Journal of the American Statistical Association* **102** 507–516. MR2370849

[29] Hastie, T. and Efron, B. (2013). lars 1.2. Least angle regression, lasso and forward stagewise regression. R package. http://cran.r-project.org.

[30] Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press. MR2978290

[31] Hsu, D., Kakade, S. and Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics* **14** 569-600. MR3201956

[32] Huang, J. C., Morris, Q. D. and Frey, B. J. (2007). Bayesian inference of MicroRNA targets from sequence and expression data. *Journal of Computational Biology* **14** 550–563. MR2344257

[33] Johnson, V. E. and Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association* **107** 649-660. MR2980074

[34] Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **367** 4237-4253. MR2546386

[35] Jordan, M. I. (2004). Graphical models. *Statistical Science* **19** 140-155. MR2082153

[36] Lai, R. C. S., Hannig, J. and Lee, T. C. M. (2015). Generalized fiducial inference for ultrahigh dimensional regression. *Journal of the American Statistical Association* **110** 760–772. MR3367262

[37] Li, S. M. J. Z. (2012). Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. *Genetics* **190** 231–249.

[38] Li, F. and Zhang, N. R. (2010). Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. *Journal of the American Statistical Association* **105** 1202–1214. MR2752615

[39] Liang, F., Paulo, R., Molina, G., Clyde, M. A. and Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association* **103** 410–423. MR2420243

[40] Logsdon, B. A., Hoffman, G. E. and Mezey, J. G. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics* **11** 1–13.

[41] Luenberger, D. G. and Ye, Y. (2008). *Linear and Nonlinear Programming*, 3rd edition ed. Springer, New York. MR2423726

[42] Luts, J. and Ormerod, J. T. (2014). Mean field variational Bayesian inference for support vector machine classification. *Computational Statistics and Data Analysis* **73** 163–176. MR3147981

[43] Mallows, C. L. (1973). Some comments on Cp. *Technometrics* **15** 661–675.

[44] Martin, R., Mess, R. and Walker, S. G. Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23**. MR3624879

[45] Martin, R. and Walker, S. G. (2014). Asymptotically minimax empir-

ical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics* **8** 2188–2206. MR3273623

[46] Maruyama, Y. and George, E. I. (2011). Fully Bayes factors with a generalized g-prior. *The Annals of Statistics* **39** 2740–2765. MR2906885

[47] Müller, S. and Welsh, A. H. (2010). On model selection curves. *International Statistical Review* **78** 240–256.

[48] Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press, London.

[49] Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42** 789–817. MR3210987

[50] Nathoo, F. S., Babul, A., Moiseev, A., Virji-Babul, N. and Beg, M. F. (2014). A variational Bayes spatiotemporal model for electromagnetic brain mapping. *Biometrics* **70** 132–143. MR3251674

[51] Nott, D. J. and Kohn, R. (2005). Adaptive sampling for Bayesian variable selection. *Biometrika* **92** 747–763. MR2234183

[52] O'Hara, R. B. and Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Analysis* **4** 85–117. MR2486240

[53] Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician* **64** 140–153. MR2757005

[54] Pham, T. H., Ormerod, J. T. and Wand, M. P. (2013). Mean field variational Bayesian inference for nonparametric regression with measurement error. *Computational Statistics and Data Analysis* **68** 375–387. MR3103783

[55] Rattray, M., Stegle, O., Sharp, K. and Winn, J. (2009). Inference algorithms and learning theory for Bayesian sparse factor analysis. In *Journal of Physics: Conference Series* **197** 012002.

[56] Redmond, M. and Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* **141** 660–678.

[57] Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association* **109** 828-846. MR3223753

[58] Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* **71** 319–392. MR2649602 MR2649602

[59] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

[60] Soussen, C., Idier, J., Brie, D. and Duan, J. (2011). From Bernoulli–Gaussian deconvolution to sparse signal restoration. *Signal Processing, IEEE Transactions on* **59** 4572–4584. MR2882966

[61] Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A. and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate:

II. radical prostatectomy treated patients. *Journal of Urology* **141** 1076–1083.

[62] Stingo, F. C. and Vannucci, M. (2011). Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. *Bioinformatics* **27** 495–501.

[63] Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D. and Caldas, C. (2005). A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics* **21** 3025-3033.

[64] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statatistical Society, Series B* **58** 267–288. MR1379242

[65] Ueda, N. and Nakano, R. (1998). Deterministic annealing EM algorithm. *Neural Networks* **11** 271–282.

[66] Van Rijsbergen, C. J. (1979). *Information Retrieval (2nd ed.).* Butterworth.

[67] Wand, M. P. and Ormerod, J. T. (2011). Penalized wavelets: Embedding wavelets into semiparametric regression. *Electronic Journal of Statistics* **5** 1654–1717. MR2870147

[68] Wand, M. P., Ormerod, J. T., Padoan, S. A. and Frühwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* **6** 847–900. MR2869967

[69] Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association* **104** 1512–1524. MR2750576

[70] Wang, X. and Chen, L. (2016). High dimensional ordinary least squares projection for screening variables. *Journal of The Royal Statistical Society Series B* **78** 589–611. MR3506794

[71] Wang, B. and Titterington, D. M. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis* **1** 625–650. MR2221291

[72] Xu, S. (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63** 513–521. MR2370810

[73] You, C., Ormerod, J. T. and Müller, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Australian and New Zealand Journal of Statistics* **56** 83–87. MR3200293

[74] Zellner, A. (1986). On Assessing Prior Distributions and Bayesian Regression Analysis With g-Prior Distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.) 233–243. North-Holland/Elsevier. MR0881437