A variational Bayes approach to variable selection

BY JOHN T. ORMEROD, CHONG YOU AND SAMUEL MÜLLER

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

18th of July 2014

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 estimates the model parameters with easily obtainable and appropriately sized standard error estimates; and, most strikingly, selects the true model at a 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: Mean field variational Bayes; spike and slab prior; Markov chain Monte Carlo.

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 (Akaike, 1973; Mallows, 1973; Schwarz, 1978), penalized regression (Tibshirani, 1996; Fan and Li, 2001; Fan and Peng, 2004) and Bayesian modeling approaches (Bottolo and Richardson, 2010; Hans et al., 2007; Li and Zhang, 2010; Stingo and Vannucci, 2011). 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 Johnstone and Titterington (2009); Fan and Lv (2010); Müller and Welsh (2010); Bülmann and van de Geer (2011); Johnson and Rossell (2012).

Bayesian model selection approaches have the advantage of being able to easily incorporate simultaneously many sources of variation, including prior knowledge. However, except for special cases such as linear models with very carefully chosen priors (see for example Liang et al., 2008; or Maruyama and George, 2011), Bayesian inference via Markov chain Monte Carlo (MCMC) for moderate to large scale problems is 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 efficiently (Nott and Kohn, 2005; Hans et al., 2007; O'Hara and Sillanpää, 2009; Bottolo and Richardson, 2010; Li and Zhang, 2010; Stingo and Vannucci, 2011). However, for sufficiently large scale problems even these approaches can be deemed to be too slow to be used in practice. Further drawbacks to these methods include sensitivity to prior choices, and there are no available diagnostics to determine whether the MCMC chain has either converged or explored a sufficient proportion of models in the model space.

Mean field variational Bayes (VB) is an efficient but approximate alternative to MCMC for Bayesian inference (Bishop, 2006; Ormerod and Wand, 2010). 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. Nevertheless, VB has shown to be an effective approach to several practical problems including document retrieval (Jordan, 2004), functional magnetic resonance imaging (Flandin and Penny, 2007; Nathoo et al., 2014), and cluster analysis for gene-expression data (Teschendorff et al., 2005). 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 posterior variances. Such criticism can be made on empirical, e.g., Wand et al. (2011); Carbonetto and Stephens (2011), or theoretical grounds, e.g., Wang and Titterington (2006); Rue et al. (2009). However, as previously shown in You et al. (2014) 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 (Hall et al., 2011a,b).

In this paper we consider a spike and slab prior on the regression coefficients (see Mitchell and Beauchamp, 1988; George and McCulloch, 1993) in order to encourage sparse estimators. 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) to automatically select variables to be included in the model. Our contributions are:

(i) We show how our VB method 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 appropriately sized standard error estimates;

(iii) Under these same assumptions 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 both promising and that extensions to more complicated settings should enjoy similar properties. Result (ii) is in keeping with consistency results of Bayesian inferential procedures (Casella et al., 2009). However, as VB methods are inexact these results are not applicable to VB-type approximations. Contribution (iii) is a remarkable result given the rate of convergence, but only holds for the case where $n > p$ and p is fixed (but still possibly very large). Similar rates of convergence were shown by Narisetty and He (2014) for an MCMC scheme. However, such results are impractical due to the inherent slowness of MCMC schemes.

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 may be viewed as a special case of VB), or VB include Huang et al. (2007), Rattray et al. (2009), Logsdon et al. (2010), Carbonetto and Stephens (2011), Wand and Ormerod (2011) and Ročková and George (2014). However, apart from Ročková and George (2014), these references did not attempt to understand how sparsity was achieved and the later reference did not consider proving the rates of convergence for the estimators they considered. Furthermore, each of these papers considered slightly different models and tuning parameter selection approaches to those here.

The spike and slab prior is a "sparsity inducing prior" since it has a discontinuous derivative at the origin (owing to the spike at zero). We could also employ other such priors on our regression coefficients. A deluge of such papers have recently considered such priors including Polson and Scott (2010), Carvalho et al. (2010), Griffin and Brown (2011), Armagan et al. (2013) and Neville et al. (2014). Such priors entail at least some shrinkage on the regression coefficients. We believe that some coefficient shrinkage is highly likely to be beneficial in practice. However, we have not pursued such priors here but focus on developing theory for what we believe is the simplest model deviating from the linear model which incorporates a sparsity inducing prior. Adapting the theory we develop for the spike and slab prior here to other sparsity inducing priors is beyond the scope of this paper.

Perhaps the most promising practical 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 Luts and Ormerod (2014). For example, it is not difficult to extend the methodology developed here to handle elaborate responses (Wand et al., 2011), missing data (Faes et al., 2011) or measurement error (Pham et al., 2013). This contrasts with criteria based procedures, penalized regression and some Bayesian procedures (for example Liang et al., 2008; Maruyama and George, 2011, where the models are chosen carefully so that an exact expression for marginal likelihood is obtainable). For these approaches there is usually no clear way of handling such complications.

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 6.

2 Bayesian linear model selection

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

$$\begin{aligned} \mathbf{y}|\boldsymbol{\beta}, \sigma^2 \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}), \quad \sigma^2 \sim \text{Inverse-Gamma}(A, B), \\ \beta_j|\gamma_j \sim \gamma_j N(0, \sigma_\beta^2) + (1-\gamma_j)\delta_0 \quad \text{and} \quad \gamma_j \sim \text{Bernoulli}(\rho), \quad j = 1, \ldots, p, \end{aligned} \tag{1}$$

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, Inverse-Gamma(A, B) is the inverse Gamma distribution with shape parameter A and scale parameter B, and δ_0 is the degenerate distribution with point mass at 0. 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 Ročková and George (2014) 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.

Replacing β_j with $\gamma_j \beta_j$ for $1 \leq j \leq p$ we can recast the model as

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

where $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_p)$ which is easier to deal with using VB. In the signal processing literature this is sometimes called the Bernoulli-Gaussian (Soussen et al., 2011) or binary mask model and is closely related to ℓ_0 regularization (see Murphy, 2012, Section 13.2.2). Wand and Ormerod (2011) 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.

The VB approach is summarized in many places including Bishop (2006); Rue et al. (2009); Ormerod and Wand (2010); Carbonetto and Stephens (2011); Faes et al. (2011); Wand et al. (2011); Murphy (2012); Pham et al. (2013); Luts and Ormerod (2014); Neville et al. (2014); You et al. (2014).

We refer the interested reader to these papers rather than summarize the approach again here. Using a variational Bayes approximation of $p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}|\mathbf{y})$ by

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

the optimal q-densities are of the form

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

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

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

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

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

$$\eta_j = \lambda - \tfrac{\tau}{2}(\mu_j^2 + \Sigma_{j,j})\|\mathbf{X}_j\|^2 + \tau\left[\mu_j\mathbf{X}_j^T\mathbf{y} - \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}(\boldsymbol{\mu}_{-j}\mu_j + \boldsymbol{\Sigma}_{-j,j})\right] \quad \text{and} \tag{6}$$

$$w_j = \text{expit}(\eta_j) \tag{7}$$

where $1 \leq j \leq p$, $\lambda = \text{logit}(\rho)$, $\mathbf{w} = (w_1 \ldots w_p)^T$, $\mathbf{W} = \text{diag}(\mathbf{w})$, $\boldsymbol{\Omega} = \mathbf{w}\mathbf{w}^T + \mathbf{W} \odot (\mathbf{I} - \mathbf{W})$ 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. 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. Later we will write $A_{i,j}$ to be the value the component corresponding to the ith row and jth column of $\mathbf{A}$ and $\mathbf{A}_{i,-j}$ is vector corresponding the ith row of $\mathbf{A}$ with the jth 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, $p(\gamma_j = 1|\mathbf{y})$, and can be used for model selection purposes, that is, the posterior probability 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

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

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

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

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,

$$|\log \underline{p}^{(t)}(\mathbf{y};\rho) - \log \underline{p}^{(t-1)}(\mathbf{y};\rho)| < \epsilon \tag{8}$$

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 $O(p^2)$.

Algorithm 1 *Iterative scheme to obtain optimal $q^*(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma})$ for our model.*

Input: $(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau, \rho, \mathbf{w})$ where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\sigma_\beta^2 > 0$, $A > 0$, $B > 0$, $\tau > 0$, $\rho \in (0,1)$ and $\mathbf{w} \in [0,1]^p$.

$\mathbf{W} \leftarrow \text{diag}(\mathbf{w}) \quad ; \quad \boldsymbol{\Omega} \leftarrow \mathbf{w}\mathbf{w}^T + \mathbf{W}(\mathbf{I} - \mathbf{W}) \quad ; \quad \lambda \leftarrow \text{logit}(\rho)$

Cycle:

$\boldsymbol{\Sigma} \leftarrow \left[\tau(\mathbf{X}^T\mathbf{X}) \odot \boldsymbol{\Omega} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1} \quad ; \quad \boldsymbol{\mu} \leftarrow \tau\boldsymbol{\Sigma}\mathbf{W}\mathbf{X}^T\mathbf{y}$

For $j = 1, \ldots, p$

$w_j \leftarrow \text{expit}\left[\lambda - \frac{1}{2}\tau(\mu_j^2 + \Sigma_{j,j})\|\mathbf{X}_j\|_2^2 + \tau\left\{\mu_j\mathbf{X}_j^T\mathbf{y} - \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}(\boldsymbol{\mu}_{-j}\mu_j + \boldsymbol{\Sigma}_{-j,j})\right\}\right]$

$\mathbf{w} \leftarrow [w_1, \ldots, w_p] \quad ; \quad \mathbf{W} \leftarrow \text{diag}(\mathbf{w})$

$\boldsymbol{\Omega} \leftarrow \mathbf{w}\mathbf{w}^T + \mathbf{W} \odot (\mathbf{I} - \mathbf{W})$

$s \leftarrow B + \frac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\mathbf{X}\mathbf{W}\boldsymbol{\mu} + \text{tr}\left((\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\right)\right] \quad ; \quad \tau \leftarrow (A + n/2)/s$

until the increase of $\log \underline{p}(\mathbf{y};\rho)$ is negligible.

To illustrate the effect of ρ on the sparsity of the VB method we consider the *prostate cancer* dataset originating from a study by Stamey et al. (1989). The data consists of $n = 97$ samples with variables `lcavol`, `lweight` (log prostate weight), `age`, `lbph` (log pf the amount of benign prostate hyperplasia), `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). Friedman et al. (2001) 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 penalty as implemented by the `R` package `ncvreg` (Breheny and Huang, 2011). These regularization paths are illustrated in Figure 1 where for our VB method the values of $\boldsymbol{\mu}$ (which serve as point estimates for $\boldsymbol{\beta}$) as a function of λ.

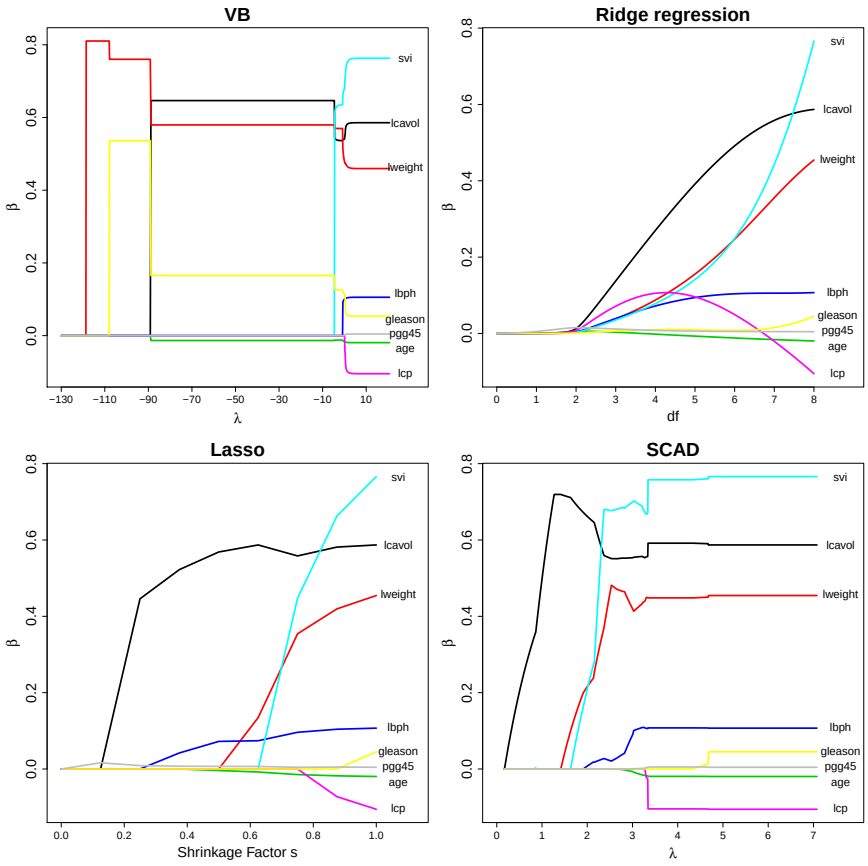

Figure 1: Top right panel: An illustration the final values of the components of $\boldsymbol{\mu}$ for multiple runs of Algorithm 1 over a grid of $\lambda = \text{logit}(\rho)$ values where we have used $w_j = 1$, $j = 1, \ldots, p$, $\tau = 1000$ and hyperparameters selected as described in Section 4 on the *prostate cancer* dataset originating in Stamey et al. (1989). Remaining panels: The regularization paths for Ridge, Lasso and SCAD penalized regression fits.

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, possibly bias for the estimates of the model coefficients. Observation (B) highlights that care is required for selecting ρ (or equivalently λ).

3 Theory

Instead of Algorithm 1, we analyze the properties of a slightly adjusted version, Algorithm 2 (consisting of Algorithm 2a and 2b), which applies the updates in a different order. Note that Algorithm 2 is less computationally efficient compared to Algorithm 1 due to the extra convergence step in Algorithm 2b. Also note that Algorithm 2 is technically not a VB algorithm. The reason is because $\eta_j^{(t+1)}$ in Algorithm 2a is updated with $\mathbf{W}_{-j}^{(t)} = \text{diag}(\mathbf{w}_{-j}^{(t)})$ instead of $\text{diag}(w_1^{(t+1)}, \ldots, w_{j-1}^{(t+1)}, w_{j+1}^{(t)}, \ldots, w_p^{(t)})$. Another important difference between Algorithm 1 and Algorithm 2 is that in Algorithm 2 we have also chosen $\mathbf{w}^{(1)} = \mathbf{1}$ to ensure that $\mathbf{w}^{(1)}$ is initialized to start from a correct model. By a correct model we mean that such a model includes all variables with non-zero coefficient values in the underlying true model. Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$. A correct model $\boldsymbol{\gamma}$ is the p-vector with elements such that $\gamma_j \in \{0, 1\}$ if $\beta_{0j} = 0$ and $\gamma_j = 1$ if $\beta_{0j} \neq 0$. Hence, the true model $\boldsymbol{\gamma}_0$ and the full model $\boldsymbol{\gamma} = \mathbf{1}$ are both correct models.

Although Algorithm 2 is less computationally efficient and not a VB algorithm, it simplifies analysis of the estimators. The properties of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$, τ and $\{w_j\}_{1 \leq j \leq p}$ when the system of equations (3)–(7) hold simultaneously are difficult to analyze. Algorithm 2 allows us to decouple the equations (3), (4) and (5) with equations (6) and (7) over the iterations of the algorithm. Based on Algorithm 2 our analysis of the theoretical properties assumes that (3), (4) and (5) hold exactly in each iteration in Algorithm 2a (namely at the completion of Algorithm 2b), and both (6) and (7) for $1 \leq j \leq p$ hold exactly at the completion of Algorithm 2a.

Algorithm 2a *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 $\mathbf{w}^{(1)} = \mathbf{1}$.

$t \leftarrow 1 \quad ; \quad \lambda \leftarrow \text{logit}(\rho)$

Cycle:

$(\boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)}, \tau^{(t)}) \leftarrow$ Output from Algorithm 2b with input $\left(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \mathbf{w}^{(t)}\right)$

$\mathbf{W}^{(t)} \leftarrow \text{diag}\left(\mathbf{w}^{(t)}\right)$

For $j = 1, \ldots, p$

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

$$w_j^{(t+1)} \leftarrow \text{expit}(\eta_j^{(t+1)})$$

$\mathbf{w}^{(t+1)} \leftarrow \left[w_1^{(t+1)}, \ldots, w_p^{(t+1)}\right]^T \quad ; \quad t \leftarrow t + 1$

until the increase of $\log \underline{p}(\mathbf{y}; \rho)$ is negligible.

In the next Appendix A 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.

Algorithm 2b *Iterative scheme to obtain optimal output* $(\boldsymbol{\mu}^{(t)}, \Sigma^{(t)}, \tau^{(t)})$ *in Algorithm 2a.*

Input: $(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau, \mathbf{w})$ where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $A > 0$, $B > 0$, $\sigma_\beta^2 > 0$, $\tau > 0$ and $\mathbf{w} \in [0,1]^p$.

$\mathbf{W} \leftarrow \text{diag}(\mathbf{w})$; $\boldsymbol{\Omega} \leftarrow \mathbf{w}\mathbf{w}^T + \mathbf{W}(\mathbf{I} - \mathbf{W})$

Cycle:

$$\boldsymbol{\Sigma} \leftarrow \left[\tau(\mathbf{X}^T\mathbf{X}) \odot \boldsymbol{\Omega} + \sigma_\beta^{-2}\mathbf{I}\right]^{-1} \quad ; \quad \boldsymbol{\mu} \leftarrow \tau\boldsymbol{\Sigma}\mathbf{W}\mathbf{X}^T\mathbf{y}$$

$$s \leftarrow B + \tfrac{1}{2}\left[\|\mathbf{y}\|^2 - 2\mathbf{y}^T\mathbf{X}\mathbf{W}\boldsymbol{\mu} + \text{tr}\left\{(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Omega})(\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma})\right\}\right] \quad ; \quad \tau \leftarrow (A + n/2)/s$$

until the increase of $\log \underline{p}(\mathbf{y}; q)$ is negligible

Output: $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \tau)$

Main Result 1 (Proof in Appendix A.1)**.** *Suppose that (3), (4) and (5) hold. Then for* $w_j \ll 1$, $1 \le j, k \le p$*, for observed* $\mathbf{y}$ *and* $\mathbf{X}$ *we have ,*

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

and the update for $w_j^{(t+1)}$ *in Algorithm 2a with small* $w_j^{(t)}$ *satisfies*

$$w_j^{(t+1)} \leftarrow \text{expit}\left[\lambda - \tfrac{1}{2}\tau^{(t)}\|\mathbf{X}_j\|^2\sigma_\beta^2 + O(w_j^{(t)})\right]. \tag{9}$$

Lemma 1 (Proof in Appendix A)**.** *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 2, if $w_j^{(t)}$ is small, the 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, $w_j^{(t+1)}$ is, for numerical purposes, identically 0. This explains why that Algorithms 1 and 2 provide sparse estimates of $\mathbf{w}$ and $\boldsymbol{\beta}$. Furthermore, all successive values of $w_j^{(t)}$ 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 You et al., 2014) and treat y_i and $\mathbf{x}_i$ as random quantities (only) in Main Result 2 and the proof of Main Result 2 in Section 6:

(A1) for $1 \le i \le n$ the $y_i|\mathbf{x}_i = \mathbf{x}_i^T\boldsymbol{\beta}_0 + \varepsilon_i$ where $\varepsilon_i \overset{\text{ind.}}{\sim} N(0, \sigma_0^2)$, $0 < \sigma_0^2 < \infty$, $\boldsymbol{\beta}_0$ are the true values of $\boldsymbol{\beta}$ and σ^2 with $\boldsymbol{\beta}_0$ being element-wise finite;

(A2) for $1 \le i \le n$ 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) the equations (3), (4) and (5) hold when $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and τ are replaced with $\boldsymbol{\mu}^{(t)}$, $\boldsymbol{\Sigma}^{(t)}$ and $\tau^{(t)}$ respectively; and

(A7) $\lambda \equiv \lambda_n$ 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) and (A6) will simplify later arguments, whereas Assumption (A7) is necessary for our method to identify the true model, which we will denote $\boldsymbol{\gamma}_0$.

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 $O_p(\cdot)$, $\mathbf{O}_p^m(\cdot)$ 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 A.2)**.** *If $w_j^{(1)} = 1$ for $1 \leq j \leq p$ and assumptions (A1)-(A6) hold then*

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

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 (A7) holds, then for $t = 2$ we have

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

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

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

For $t > 2$ *we have*

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

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

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

Remark: This result suggests, under assumptions (A1)–(A7) and in light of Lemma 1, that the vector $\mathbf{w}^{(t)}$ in Algorithm 2 approaches $\boldsymbol{\gamma}_0$ at an exponential rate in n.

4 Hyperparameter selection and initialization

We will now briefly discuss selecting prior hyperparameters. We have used $A = B = 0.01$, $\sigma_\beta^2 = 10$ 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.

Ročková and George (2014) used a deterministic annealing variant of the EM algorithm proposed by Ueda and Nakano (1998) to avoid local maxima problems and 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).

This procedure is more specifically described in Algorithm 3. Note that in Algorithm 3 we use the notation $\mathbf{w}_j^{(k)}$ to denote the vector $\mathbf{w}$ with the jth element set to k.

Algorithm 3 *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 $\tau > 0$
$M = 100$; $P = 50$; $\rho = \text{expit}(-0.5\sqrt{n})$; $\mathcal{L} = -\infty$

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 $\left(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}_j^{(1)}\right)$

  $k \leftarrow \text{argmax}_{1 \le j \le p}\{\mathcal{L}_j\}$; 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 $\left(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho_j, \mathbf{w}\right)$

  $k \leftarrow \text{argmax}_{1 \le j \le p}\{\mathcal{L}_j\}$; If $\mathcal{L}_k > \mathcal{L}$ then set $\mathcal{L}$ to $\mathcal{L}_k$ and ρ to ρ_k

  For $j = 1, \ldots, p$

    $\mathcal{L}_0 \leftarrow \log \underline{p}(\mathbf{y}; \rho)$ from Algorithm 1 with input $\left(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}_j^{(0)}\right)$
    $\mathcal{L}_1 \leftarrow \log \underline{p}(\mathbf{y}; \rho)$ from Algorithm 1 with input $\left(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}_j^{(1)}\right)$
    $k \leftarrow \text{argmax}_{j \in \{0,1\}}\{\mathcal{L}_j\}$; 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 $\left(\mathbf{y}, \mathbf{X}, \sigma_\beta^2, A, B, \tau_0, \rho, \mathbf{w}\right)$

---

5 Numerical examples

In the following three 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 considering a data set with few explanatory variables before looking at higher dimensional situations where $p = 41$ and $n = 80$ and in the third example where $p = 99$ and $n = 2118$.

We use the mean square error (MSE) to measure the quality of the prediction error defined by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}})_i^2$$

and the F_1-score to assess the quality of model selection defined to be the harmonic mean between precision and recall

$$F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad \text{where} \quad \text{precision} = \frac{TP}{TP+FP} \quad \text{and} \quad \text{recall} = \frac{TP}{TP+FN},$$

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 to prefer methods which do not select none or all of the variables. We compare the performance of our VB method against the Lasso, SCAD and MCP penalized regression methods using 10-fold cross-validation to choose the tuning parameter as implemented by the `R` package `ncvreg` (Breheny and Huang, 2011). Our methods were implemented in `R` and all code was run on the firrst author's laptop computer (64 bit Windows 8 Intel i7-4930MX central processing unit at 3GHz with 32GB of random access memory).

5.1 Comparison with MCMC

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 (2) are given by

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

from which 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 (2). However, if we wanted to compare model selection strategies, another MCMC scheme might be more appropriate to use for comparison. Here we focus on comparing the quality of the VB via Algorithm 1 with its Gibbs sampling counterpart (16).

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 (8) 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 3. 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 3. For each of the examples we used 10^5 MCMC samples for inference after a 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 Faes et al. (2011), 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 our results 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: Solid waste data

The following example we take from Müller and Welsh (2010) based on a simulation study using the solid waste data of Gunst and Mason (1980). The same settings were used in Shao (1993, 1997), Wisnowski et al. (2003), and Müller and Welsh (2005). We consider the model

$$y_i = \beta_0 x_{0i} + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i,$$

where $i = 1, \ldots, 40$, the errors ε_i are independent and identically distributed standard normal random variables, $\mathbf{x}_0$ is a column of ones, and the values for the solid waste data variables $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, and $\mathbf{x}_4$ are taken from (Shao, 1993, Table 1). Note that these explanatory variables are highly correlated. We consider $\boldsymbol{\beta}$ to be from the set

$$\boldsymbol{\beta}^T \in \big\{(2,0,0,0,0), (2,0,0,4,0), (2,0,0,4,8), (2,9,0,4,8), (2,9,6,4,8)\big\},$$

so that the number of true non-zero coefficients range from 1 to 5. Data for each of these 5 different values of β were simulated 1000 times and model selection accuracy (through F_1) and prediction accuracy (through MSE) results are summarized in Figure 2.

In Figure 2 we see in the first panel that VB selects the correct model for almost every simulation. We also see in the second panel that VB provides smaller prediction errors when compare to Lasso, SCAD and MCP penalty methods. All methods perform almost identically for the dense case, where the data generating model is the full model. The mean times per simulation for out VB method, and the Lasso, SCAD and MCP penalized regression methods were 0.73, 0.15, 0.17 and 0.18 seconds respectively.

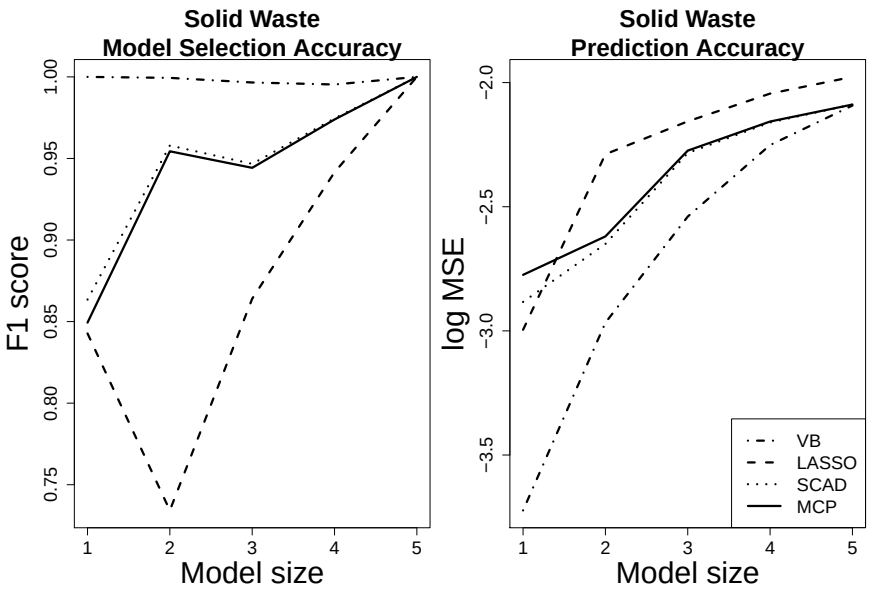

Figure 2: Summaries of the model selection and prediction accuracies of VB, Lasso, SCAD and MCP methods for the `Solid Waste` example.

The results for the comparisons between VB and MCMC based on 100 simulations are summarized in Table 1. Note that the MCMC approach took an average of 32.56 seconds per simulation setting. For each of the settings we see high agreement between VB and MCMC in terms of accuracy and model selection performance. For model sizes 3, 4 and 5 the prediction and bias measures are also very similar. However, for model sizes 1 and 2 the VB approach seems to have smaller prediction errors and less biased estimates of the regression coefficients.

5.3 Example 2: Diets simulation

We use the following example modified from Garcia et al. (2013). 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

Model Size	1	2	3	4	5
$-$log-MSE-VB	5.2 (0.24)	3.7 (0.15)	3.0 (0.10)	2.54 (0.08)	2.34 (0.07)
$-$log-MSE-MCMC	4.4 (0.13)	3.4 (0.11)	2.9 (0.09)	2.59 (0.08)	2.37 (0.07)
$-$log-BIAS-VB	6.8 (0.24)	5.3 (0.15)	3.6 (0.13)	2.41 (0.11)	2.05 (0.09)
$-$log-BIAS-MCMC	5.9 (0.11)	4.7 (0.11)	3.4 (0.11)	2.50 (0.10)	2.09 (0.09)
F_1-VB	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.99 (0.00)	1.00 (0.00)
F_1-MCMC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
accuracy(β)	97.2 (0.20)	93.1 (0.37)	94.3 (0.27)	93.7 (0.52)	96.1 (0.08)
accuracy(σ^2)	98.4 (0.27)	97.5 (0.18)	97.6 (0.12)	96.5 (0.22)	96.8 (0.04)

Table 1: Performance measure comparisons between VB and MCMC based on 100 simulations for the `solid waste` example.

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 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 $m_1 = 40$, $n = 80$, $\sigma^2 = 0.5$, and $\boldsymbol{\beta} = (1 - \kappa - 1/6) \times (4.5, 3, 3, 3, \mathbf{0}^T, 3)$ 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 Garcia et al. (2013).

We generate 100 independent data sets for each value of κ in the set $\{1, 2, 3, 4\}$ 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 larger signal to noise ratio. Garcia et al. (2013) considered the case where $\kappa = 1$ and $n = 40$. The results are summarized in the two panels of Figures 2.

In Figure 3 we see in the first panel that VB selects the correct model and in the second panel provides smaller prediction errors for almost every simulation with the exception of $\kappa = 4$ corresponding to the smallest signal to noise scenario. The mean times per simulation for out VB method, and the Lasso, SCAD and MCP penalized regression methods were 6.8, 0.4, 0.3 and 0.2 seconds respectively.

The results for the comparisons between VB and MCMC based on 100 simulations are summarized in Table 2. Here we see similar results to Table 1 except posterior density parameter accuracy is lower than for the `solid waste` example. In particular we note that MSE and parameter bi-

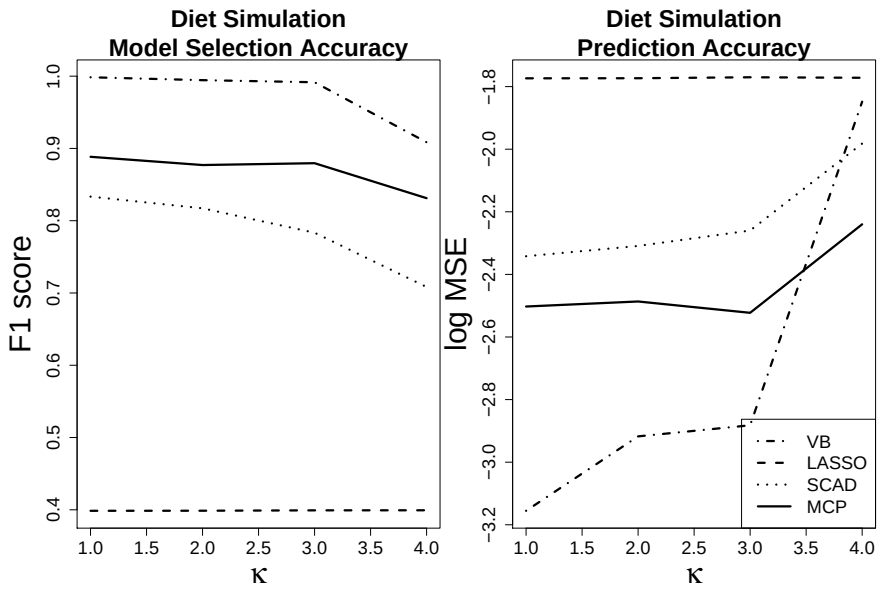

Figure 3: Summaries of the model selection and prediction accuracies of VB, Lasso, SCAD and MCP methods for the `Diet Simulation` example.

ases are either similar to MCMC or better for some settings. Note that the MCMC approach took an average of 131.4 seconds per simulation setting.

	$\kappa = 1$	$\kappa = 2$	$\kappa = 3$	$\kappa = 4$
$-$log-MSE-VB	3.33 (0.06)	3.30 (0.07)	3.27 (0.08)	2.63 (0.13)
$-$log-MSE-MCMC	2.79 (0.06)	2.76 (0.06)	2.77 (0.05)	2.69 (0.06)
$-$log-BIAS-VB	4.63 (0.08)	4.59 (0.09)	4.55 (0.10)	3.74 (0.16)
$-$log-BIAS-MCMC	4.15 (0.07)	4.15 (0.06)	4.15 (0.07)	4.03 (0.07)
F_1-VB	1.00 (0.00)	0.99 (0.00)	0.99 (0.01)	0.91 (0.02)
F_1-MCMC	0.98 (0.00)	0.97 (0.00)	0.98 (0.00)	0.97 (0.01)
accuracy(β)	93.5 (0.16)	93.4 (0.20)	93.3 (0.20)	91.1 (0.41)
accuracy(σ^2)	86.4 (0.98)	85.5 (1.31)	84.5 (1.45)	69.3 (2.99)

Table 2: Performance measure comparisons between VB and MCMC based on 100 simulations for the `diet simulation` example.

5.4 Example 3: 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 Redmond and Baveja (2002) 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.

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 $\{\lambda^{(j)}, \boldsymbol{\beta}^{(j)}\}_{1 \leq j \leq J}$.

- 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 $\sigma^2 = 0.1$, 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` (Hastie and Efron, 2013) in the above procedure. Results for the comparisons between VB, Lasso, SCAD and MCP are summarized in Figure 4 where we see again that VB has, except for a few model settings, the best model selection performance and smallest prediction error. The mean times per simulation for out VB method, and the Lasso, SCAD and MCP penalized regression methods were 45, 12, 8 and 6 seconds respectively.

The results for the comparisons between VB and MCMC based on 30 simulations are summarized in Table 3. In this table we see that parameter posterior density accuracies are nearly perfect for all parameters for nearly every simulation setting. Similarly to the previous two examples we again see that MSEs and parameter biases are either similar or better for VB in comparison to MCMC. Note that the MCMC approach took an average of 339 seconds per simulation setting.

Model size	1	2	3	4	5
$-$log-MSE-VB	11.2 (0.34)	10.6 (0.19)	10.5 (0.15)	10.3 (0.13)	9.8 (0.14)
$-$log-MSE-MCMC	2.2 (0.08)	2.5 (0.06)	2.4 (0.09)	2.5 (0.08)	2.6 (0.08)
$-$log-BIAS-VB	15.8 (0.33)	15.2 (0.19)	14.7 (0.19)	14.3 (0.17)	13.24 (0.31)
$-$log-BIAS-MCMC	6.9 (0.03)	6.9 (0.02)	6.9 (0.03)	6.9 (0.03)	6.97 (0.03)
F_1-VB	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)
F_1-MCMC	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)
accuracy(β)	99.5 (0.04)	99.5 (0.04)	99.4 (0.05)	99.4 (0.04)	99.0 (0.14)
accuracy(σ^2)	98.9 (0.19)	99.1 (0.17)	99.1 (0.24)	99.3 (0.11)	98.7 (0.37)
Model size	6	7	8	9	10
$-$log-MSE-VB	9.7 (0.18)	9.9 (0.08)	8.5 (0.03)	8.9 (0.08)	9.7 (0.14)
$-$log-MSE-MCMC	2.6 (0.07)	2.4 (0.08)	2.5 (0.09)	2.5 (0.08)	2.6 (0.09)
$-$log-BIAS-VB	12.8 (0.31)	13.3 (0.11)	11.5 (0.04)	12.0 (0.07)	12.2 (0.33)
$-$log-BIAS-MCMC	7.0 (0.03)	7.0 (0.03)	7.0 (0.03)	7.0 (0.02)	7.0 (0.02)
F_1-VB	0.96 (0.01)	0.92 (0.00)	0.86 (0.01)	0.89 (0.01)	0.93 (0.01)
F_1-MCMC	0.96 (0.01)	0.92 (0.00)	0.88 (0.01)	0.90 (0.01)	0.96 (0.01)
accuracy(β)	98.0 (0.17)	99.2 (0.10)	97.6 (0.23)	97.7 (0.24)	97.0 (0.23)
accuracy(σ^2)	97.9 (0.51)	99.2 (0.15)	93.6 (1.05)	95.2 (1.05)	97.8 (0.43)
Model size	11	12	13	14	15
$-$log-MSE-VB	9.0 (0.08)	9.0 (0.08)	9.8 (0.12)	9.0 (0.10)	9.0 (0.09)
$-$log-MSE-MCMC	2.4 (0.08)	2.4 (0.08)	2.7 (0.06)	2.5 (0.09)	2.6 (0.07)
$-$log-BIAS-VB	10.0 (0.10)	10.0 (0.10)	12.0 (0.41)	10.1 (0.17)	10.1 (0.07)
$-$log-BIAS-MCMC	7.0 (0.03)	7.0 (0.03)	7.0 (0.03)	7.0 (0.04)	7.0 (0.03)
F_1-VB	0.80 (0.01)	0.80 (0.01)	0.85 (0.01)	0.78 (0.01)	0.87 (0.01)
F_1-MCMC	0.85 (0.01)	0.85 (0.01)	0.87 (0.01)	0.85 (0.01)	0.94 (0.01)
accuracy(β)	95.6 (0.38)	95.6 (0.38)	97.3 (0.29)	94.6 (0.42)	95.4 (0.34)
accuracy(σ^2)	93.9 (0.94)	93.9 (0.94)	96.2 (0.93)	92.3 (0.99)	94.5 (0.75)
Model size	16	17	18	19	20
$-$log-MSE-VB	9.4 (0.12)	9.4 (0.11)	9.6 (0.11)	9.9 (0.07)	9.8 (0.06)
$-$log-MSE-MCMC	2.5 (0.09)	2.6 (0.08)	2.5 (0.06)	2.6 (0.08)	2.5 (0.09)
$-$log-BIAS-VB	11.6 (0.22)	10.4 (0.07)	10.5 (0.07)	10.3 (0.16)	11.2 (0.26)
$-$log-BIAS-MCMC	7.1 (0.03)	7.9 (0.03)	7.1 (0.03)	7.0 (0.03)	7.1 (0.03)
F_1-VB	0.93 (0.01)	0.90 (0.01)	0.91 (0.00)	0.91 (0.00)	0.94 (0.01)
F_1-MCMC	0.94 (0.01)	0.96 (0.01)	0.96 (0.01)	0.96 (0.01)	0.96 (0.00)
accuracy(β)	95.7 (0.26)	94.9 (0.32)	95.4 (0.27)	96.1 (0.32)	95.5 (0.56)
accuracy(σ^2)	96.7 (0.46)	95.9 (0.53)	96.2 (0.54)	97.7 (0.46)	97.6 (0.36)

Table 3: Performance measure comparisons between VB and MCMC based on 30 simulations for the `communities and crime` example.

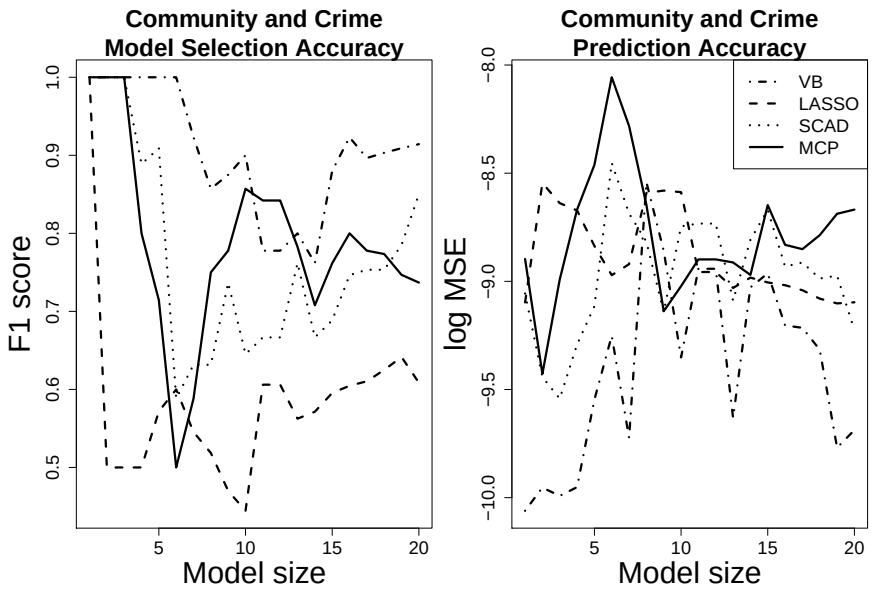

Figure 4: Summaries of the model selection and prediction accuracies of VB, Lasso, SCAD and MCP methods for the `Communities and Crime` example.

6 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.

A drawback of our theory is that we only consider the case where $p < n$. This might be mitigated to a certain extent but the use of screening procedures such as sure independence screening (Fan and Lv, 2008) which can be seen as searching for a correct model with high probability. 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. A second important extension would be to analyze the effect of more elaborate shrinkage 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 for example Neville et al., 2014). 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.

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).

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: In Proceedings of the 2nd International Symposium on Information Theory. Akademiai Kiad6, Budapest, pp. 267–281.

Armagan, A., Dunson, D. B., Lee, J., 2013. Generalized double Pareto shrinkage. Statistica Sinica 23, 119–143.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer, New York.

Bishop, Y. M. M., Fienberg, S. E., Holland, P. W., 2007. Discrete multivariate analysis: Theory and Practice. Springer.

Bottolo, L., Richardson, S., 2010. Evolutionary stochastic search for Bayesian model exploration. Bayesian Analysis 5, 583–618.

Breheny, P., Huang, J., 2011. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. The Annals of Applied Statistics 5, 232–253.

Bülmann, P., van de Geer, S., 2011. Statistics for High Dimensional Data. Springer.

Carbonetto, P., Stephens, M., 2011. Scalable variational inference for bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Analysis 6, 1–42.

Carvalho, C. M., Polson, N. G., Scott, J. G., 2010. The horseshoe estimator for sparse signals. Biometrika 97, 465–480.

Casella, G., Girón, F. J., Martńez, M. L., Moreno, E., 2009. Consistency of Bayesian procedures for variable selection. The Annals of Statistics 37, 1207–1228.

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

Fan, J., Li, R., 2001. Variable selection via onconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96, 1348–1360.

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

Fan, J., Lv, J., 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20, 101–148.

Fan, J., Peng, H., 2004. Nonconcave penalized likelihood with a diverging number of parameters. The Annals of Statistics 32, 928–961.

Flandin, G., Penny, W. D., 2007. Bayesian fMRI data analysis with sparse spatial basis function priors. NeuroImage 34, 1108–1125.

Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning. Springer.

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

George, E. I., McCulloch, B. E., 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association 88, 1023–1032.

Griffin, J., Brown, P., 2011. Bayesian adaptive lassos with non-convex penalization. Australian New Zealand Journal of Statistics 53, 423–442.

Gunst, R. F., Mason, R. L., 1980. Regression analysis and its application: a data-oriented approach. New York: Marcel Dekker.

Hall, P., Ormerod, J. T., Wand, M. P., 2011a. Theory of Gaussian variational approximation for a Poisson mixed model. Statistica Sinica 21, 369–389.

Hall, P., Pham, T., Wand, M. P., Wang, S. S. J., 2011b. Asymptotic normality and valid inference for Gaussian variational approximation. The Annals of Statistics 39, 2502–2532.

Hans, C., Dobra, A., West, M., 2007. Shotgun stochastic search for "large p" regression. Journal of the American Statistical Association 102, 507–516.

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

Horn, R. A., Johnson, C. R., 2012. Matrix Analysis. Cambridge University Press.

Huang, J. C., Morris, Q. D., Frey, B. J., 2007. Bayesian inference of microRNA targets from sequence and expression data. Journal of Computational Biology 14, 550–563.

Johnson, V. E., Rossell, D., 2012. Bayesian model selection in high-dimensional settings. Journal of the American Statistical Association 107, 649–660.

Johnstone, I. M., 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.

Jordan, M. I., 2004. Graphical models. Statistical Science 19, 140–155.

Li, F., 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.

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

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

Luts, J., Ormerod, J. T., 2014. Mean field variational Bayesian inference for support vector machine classification. Computational Statistics and Data Analysis 73, 163–176.

Mallows, C. L., 1973. Some comments on Cp. Technometrics 15, 661–675.

Maruyama, Y., George, E. I., 2011. Fully Bayes factors with a generalized g-prior. The Annals of Statistics 39, 2740–2765.

Mitchell, T. J., Beauchamp, J. J., 1988. Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023–1032.

Müller, S., Welsh, A. H., 2005. Outlier robust model selection in linear regression. Journal of the American Statistical Association 100, 1297–1310.

Müller, S., Welsh, A. H., 2010. On model selection curves. International Statistical Review 78, 240–256.

Murphy, K. P., 2012. Machine Learning: A Probabilistic Perspective. The MIT Press, London.

Narisetty, N. N., He, X., 2014. Bayesian Variable Selection with Shrinking and Diffusing Priors. The Annals of Statistics 42, 789–817.

Nathoo, F. S., Babul, A., Moiseev, A., Virji-Babul, N., Beg, M. F., 2014. A variational Bayes spatiotemporal model for electromagnetic brain mapping. Biometrics 70, 132–143.

Neville, S., Ormerod, J. T., Wand, M. P., 2014. Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. Electronic Journal of Statistics.

Nott, D. J., Kohn, R., 2005. Adaptive sampling for Bayesian variable selection. Biometrika 92, 747–763.

O'Hara, R. B., Sillanpää, M. J., 2009. A review of Bayesian variable selection methods: what, how and which. Bayesian analysis 4, 85–117.

Ormerod, J. T., Wand, M. P., 2010. Explaining variational approximations. The American Statistician 64, 140–153.

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

Polson, N. G., Scott, J. G., 2010. Shrink globally, act locally: sparse Bayesian regularization and prediction. In: Bernardo, J., Bayarri, M., J.O. Berger, A. D., Heckerman, D., Smith, A., M.West (Eds.), Bayesian Statistics 9. Oxford University Press, Oxford.

Rattray, M., Stegle, O., Sharp, K., Winn, J., 2009. Inference algorithms and learning theory for Bayesian sparse factor analysis. In: Journal of Physics: Conference Series. Vol. 197. p. 012002.

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

Ročková, V., George, E. I., 2014. EMVS: The EM approach to Bayesian variable selection. Journal of the American Statistical Association 109, 828–846.

Rue, H., Martino, S., 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.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.

Shao, J., 1993. Linear model selection by cross-validation. Journal of the American statistical Association 88, 486–494.

Shao, J., 1997. An asymptotic theory for linear model selection. Statistica Sinica 7, 221–242.

Soussen, C., Idier, J., Brie, D., Duan, J., 2011. From Bernoulli–Gaussian deconvolution to sparse signal restoration. Signal Processing, IEEE Transactions on 59, 4572–4584.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., 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.

Stingo, F. C., Vannucci, M., 2011. Variable selection for discriminant analysis with Markov random field priors for the analysis of microarray data. Bioinformatics 27, 495–501.

Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D., Caldas, C., 2005. A variational bayesian mixture modelling framework for cluster analysis of gene-expression data. Bioinformatics 21, 3025–3033.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statatistical Society, Series B 58, 267–288.

Ueda, N., Nakano, R., 1998. Deterministic annealing EM algorithm. Neural Networks 11, 271–282.

Wand, M. P., Ormerod, J. T., 2011. Penalized wavelets: Embedding wavelets into semiparametric regression. Electronic Journal of Statistics 5, 1654–1717.

Wand, M. P., Ormerod, J. T., Padoan, S. A., Frühwirth, R., 2011. Mean Field Variational Bayes for Elaborate Distributions. Bayesian Analysis 6, 847–900.

Wang, B., 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.

Wisnowski, J. W., Simpson, J. R., Montgomery, D. C., Runger, G. C., 2003. Resampling methods for variable selection in robust regression. Computational Statistics & Data Analysis 43, 341–355.

You, C., Ormerod, J. T., 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.

A variational Bayes approach to variable selection
Appendices

BY JOHN T. ORMEROD, CHONG YOU AND SAMUEL MÜLLER

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

26th of July 2014

Appendix A: Proofs

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

□

Result 1. *If $w_j^{(t)} > 0$, $1 \leq j \leq p$, 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 (Horn and Johnson, 2012, Section 7.1). By definition $\mathbf{a}^T\boldsymbol{\Omega}\mathbf{a} = \mathbf{a}^T\left[\mathbf{w}\mathbf{w}^T + \mathbf{W}(\mathbf{I}-\mathbf{W})\right]\mathbf{a} = (\sum_{j=1}^p a_j w_j)^2 + \sum_{j=1}^p a_j^2 w_j(1-w_j)$. As $0 < w_j^{(t)} \leq 1$, $1 \leq j \leq p$, we have $w_j(1-w_j) \geq 0$ and hence $\sum_{j=1}^p a_j^2 w_j(1-w_j) \geq 0$. Again, as $w_j^{(t)} > 0$, $1 \leq j \leq p$, we have $(\sum_{j=1}^p a_j w_j)^2 > 0$ for any non-zero vector $\mathbf{a}$. Hence, the result is as stated.

□

Let

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

and

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

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 positive definite and $w_j \in (0,1]$, $1 \leq j \leq p$ and $\alpha \geq 0$ then the function dof$(\alpha, \mathbf{w})$ is monotonically decreasing in α and satisfies $0 <$ dof$(\alpha, \mathbf{w}) \leq p$.*

Proof of Result 2 : Let the eigenvalue decomposition (17) hold. Since $w_j \in (0,1]$, $1 \leq j \leq p$ is positive by Result 1 the matrix $\boldsymbol{\Omega}$ is positive definite. By the Schur product theorem (Horn and Johnson, 2012, Theorem 7.5.2), the matrix $(\mathbf{X}^T\mathbf{X}) \odot \boldsymbol{\Omega}$ is also positive definite. Hence, $\nu_i > 0$, $i = 1, \ldots, p$. Then, using properties of the orthonormal matrix $\mathbf{U}$, we have

$$\operatorname{dof}(\alpha, \mathbf{w}) = \operatorname{tr}\left[\mathbf{U}\operatorname{diag}(\boldsymbol{\nu})\mathbf{U}^T\left(\mathbf{U}\operatorname{diag}(\boldsymbol{\nu})\mathbf{U}^T + \alpha\mathbf{I}\right)^{-1}\right] = \sum_{j=1}^p \frac{\nu_j}{\nu_j + \alpha}.$$

Clearly, dof$(\alpha, \mathbf{w})$ is monotonically decreasing in α, dof$(0, \mathbf{w}) = p$ and dof$(\alpha, \mathbf{w})$ only approaches zero as $\alpha \to \infty$.

□

The next lemma follows from Horn and Johnson (2012, Section 0.7.3):

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

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

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

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,* $\mathbf{a} = [a_1, \ldots, a_p]^T$ *be a real vector, and let the elements of the vector* $\mathbf{b} = [b_1, \ldots, b_p]^T$ *be positive. Then the quantity* $\mathbf{a}^T \left[\mathbf{M} + diag(\mathbf{b})\right]^{-1} \mathbf{a}$ *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 $\mathbf{m}_{12} = [M_{12}, \ldots, M_{1p}]^T$, $\mathbf{B}_2 = \text{diag}(b_2, \ldots, b_p)$ 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,

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

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 (Horn and Johnson, 2012, 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 $b_j, 2 \leq j \leq p$ after a relabeling argument.

□

Result 3. *Suppose that (3) and (5) hold. Then*

$$\tau = \frac{2A + n - dof(\tau^{-1}\sigma_\beta^{-2}, \mathbf{w})}{2B + \|\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu}\|^2 + \boldsymbol{\mu}^T[(\mathbf{X}^T\mathbf{X}) \odot \mathbf{W} \odot (\mathbf{I} - \mathbf{W})]\boldsymbol{\mu}}. \tag{21}$$

Proof of Result 3: Substituting (3) into (5) we may rewrite τ as

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

The result then follows after rearranging.

□

Remark: Note from Result 3 that the numerator in (21) may become negative when $2A + n < p$ and σ_β^2 is sufficiently large. The practical implication of this is that Algorithm 1 can fail to converge in practice in these situations. For these reasons, and to simplify our results, we only consider the case where $n > p$.

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 4. *Suppose that (3), (4) and (5) hold. Then $\tau_L \leq \tau \leq \tau_U$ where*

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

Proof of Result 4: The upper bound for τ follows from Result 3 and the inequalities (a) $\text{dof}(\tau^{-1}\sigma_\beta^{-2}, \mathbf{w}) > 0$ (from Result 2); (b) $\|\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu}\|^2 \geq \|\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\|^2$ (from least squares results); and (c) $\boldsymbol{\mu}^T[(\mathbf{X}^T\mathbf{X}) \odot \mathbf{W} \odot (\mathbf{I} - \mathbf{W})]\boldsymbol{\mu} \geq 0$ (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 τ first note $\|\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu}\|^2 \leq \|\mathbf{y}\|^2 + \|\mathbf{X}\mathbf{W}\boldsymbol{\mu}\|^2$ via the triangle inequality. Hence, using Result 2 we have

$$\tau \geq \frac{2A + n - p}{2B + \|\mathbf{y}\|^2 + \boldsymbol{\mu}^T[\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W} + (\mathbf{X}^T\mathbf{X}) \odot \mathbf{W} \odot (\mathbf{I} - \mathbf{W})]\boldsymbol{\mu}} = \frac{2A + n - p}{2B + \|\mathbf{y}\|^2 + \boldsymbol{\mu}^T[(\mathbf{X}^T\mathbf{X}) \odot \boldsymbol{\Omega}]\boldsymbol{\mu}}.$$

Let the eigenvalue decomposition (17) hold. Then

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

where the last line follows from Lemma 3. Combining this inequality, using the fact from Result 2 that $\text{dof}(\tau^{-1}\sigma_\beta^{-2}, \mathbf{w}) \leq p$ and Result 3 obtains the lower bound on τ.

□

A.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 $\mathbf{P}_j$ by

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

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

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

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 5. *If (3) holds then*

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

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

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

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

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

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

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

If (3) and (4) hold then

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

and

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

and if (3), (4) and (5) hold then

$$\eta_j \quad = \lambda + \left(\tfrac{1}{2} \tau \|\mathbf{X}_j\|^2 + w_j^{-1} \sigma_\beta^{-2}\right) \mu_j^2 - \left(\tfrac{1}{2} \tau \|\mathbf{X}_j\|^2 - w_j \tau \mathbf{X}_j^T \mathbf{P}_j \mathbf{X}_j\right) \Sigma_{j,j}, \qquad 1 \leq j \leq p. \tag{31}$$

Proof of Result 5: For a given indicator vector $\boldsymbol{\gamma}$ we can rewrite (3) as

$$\begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} & \boldsymbol{\Sigma}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}} \\ \boldsymbol{\Sigma}_{-\boldsymbol{\gamma},\boldsymbol{\gamma}} & \boldsymbol{\Sigma}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \tau \mathbf{W}_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}} \mathbf{W}_{\boldsymbol{\gamma}} + \mathbf{D}_{\boldsymbol{\gamma}} & \tau \mathbf{W}_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{-\boldsymbol{\gamma}} \mathbf{W}_{-\boldsymbol{\gamma}} \\ \mathbf{W}_{-\boldsymbol{\gamma}} \mathbf{X}_{-\boldsymbol{\gamma}}^T \mathbf{X}_1 \mathbf{W}_{\boldsymbol{\gamma}} \tau & \tau \mathbf{W}_{-\boldsymbol{\gamma}} \mathbf{X}_{-\boldsymbol{\gamma}}^T \mathbf{X}_{-\boldsymbol{\gamma}} \mathbf{W}_{-\boldsymbol{\gamma}} + \mathbf{D}_{-\boldsymbol{\gamma}} \end{bmatrix}^{-1}.$$

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,

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

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,

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

and (30) follows by letting $\boldsymbol{\gamma} = \mathbf{e}_j$. Lastly, rearranging (4) we find

$$\begin{aligned}
\left[\tau \mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W} + \mathbf{D}\right]\boldsymbol{\mu} &= \tau \mathbf{W}\mathbf{X}^T\mathbf{y} \\
\Rightarrow \tau \mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W}\boldsymbol{\mu} + \mathbf{D}\boldsymbol{\mu} &= \tau \mathbf{W}\mathbf{X}^T\mathbf{y} \\
\Rightarrow \mathbf{D}\boldsymbol{\mu} &= \tau \mathbf{W}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu}) \\
\Rightarrow w_j^{-1} d_j \mu_j &= \tau \mathbf{X}_j^T(\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu}), \quad \text{for } 1 \leq j \leq p.
\end{aligned}$$

Hence, via a simple algebraic manipulation, η_j may be written as

$$\begin{aligned}
\eta_j &= \lambda - \tfrac{1}{2}\tau(\mu_j^2 + \Sigma_{j,j})\|\mathbf{X}_j\|^2 - \tau \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}\boldsymbol{\Sigma}_{-j,j} \\
&\quad + \mu_j \tau \mathbf{X}_j^T(\mathbf{y} - \mathbf{X}_{-j}\mathbf{W}_{-j}\boldsymbol{\mu}_{-j} - \mathbf{X}_j w_j \mu_j) + \mu_j^2 \tau w_j \|\mathbf{X}_j\|^2 \\
&= \lambda + \left(w_j - \tfrac{1}{2}\right)\tau\|\mathbf{X}_j\|_2^2\mu_j^2 - \tfrac{1}{2}\tau\|\mathbf{X}_j\|^2\Sigma_{j,j} + \tau \mathbf{X}_j^T(\mathbf{y} - \mathbf{X}\mathbf{W}\boldsymbol{\mu})\mu_j - \tau \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}\boldsymbol{\Sigma}_{-j,j} \\
&= \lambda + \left[\left(w_j - \tfrac{1}{2}\right)\tau\|\mathbf{X}_j\|_2^2 + w_j^{-1}D_j\right]\mu_j^2 - \tfrac{1}{2}\tau\|\mathbf{X}_j\|^2\Sigma_{j,j} - \tau \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}\boldsymbol{\Sigma}_{-j,j}.
\end{aligned} \tag{32}$$

Substituting (26), (27) and (30) into (32) and simplifying gives (31).

□

From Result 4 we have that τ is bounded and that $(y_i, \mathbf{x}_i)$ are observed so that all quantities are deterministic. From Equation (30) we see that μ_j is clearly $O(w_j)$ as $\mathbf{P}_j$ does not depend on w_j. Noting that $\lim_{w_j \to 0} \Sigma_{j,j} = \sigma_\beta^2$ follows from Equation (26) and the result for $\Sigma_{j,j}$ follows after a Taylor series argument. The result $\Sigma_{j,k} = O(w_j w_k)$, $j \neq k$ follows from Equation (28). We can see that the update for $w_j^{(t)}$ in Algorithm 2a is as stated by combining Equation (32) with the fact that $\mu_j^{(t)} = O(w_j^{(t)})$ and $\Sigma_{j,j} = \sigma_\beta^2 + O(w_j^{(t)})$. This completes the proof of Main Result 1.

A.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 1–5 are still valid, when assuming random quantities $\mathbf{y}$ and $\mathbf{X}$. Define the following stochastic sequences:

$$\mathbf{A}_n = n^{-1}\mathbf{X}^T\mathbf{X}, \quad \mathbf{b}_n = n^{-1}\mathbf{X}^T\mathbf{y}, \quad c_n = \text{dof}(\tau^{(t)}\sigma_\beta^{-2}, \mathbf{1}) \quad \text{and} \quad \boldsymbol{\beta}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{33}$$

Assuming (A1)–(A4) You et al. (2014) 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 (Bishop et al., 2007, 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

$$\begin{aligned}
\mathbf{A}_n &= \mathbf{S} + \mathbf{O}_p^m\big(n^{-1/2}\big), & \mathbf{A}_n^{-1} &= \mathbf{S}^{-1} + \mathbf{O}_p^m\big(n^{-1/2}\big), \\
\|\mathbf{X}_j\|^2 &= n\mathbb{E}(x_j^2) + O_p\big(n^{1/2}\big), & \|\boldsymbol{\epsilon}\|^2 &= n\sigma_0^2 + O_p\big(n^{1/2}\big), \\
n^{-1}\mathbf{X}\boldsymbol{\epsilon} &= \mathbf{O}_p^v\big(n^{-1/2}\big) \quad \text{and} & \mathbf{b}_n &= n^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}) = \mathbf{S}\boldsymbol{\beta}_0 + \mathbf{O}_p^v\big(n^{-1/2}\big).
\end{aligned} \tag{34}$$

Before we improve upon the results of You et al. (2014) 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 6. *Assume (A1)–(A6), then for $t > 0$ we have $\tau^{(t)} = O_p(1)$ and $1/\tau^{(t)} = O_p(1)$.*

Proof of Result 6: Using (A6) equations (3), (4) and (5) hold and we can use Result 4 to obtain $\tau_U^{-1} < \tau^{-1} < \tau_L^{-1}$ where

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

By (A1)–(A4) and the strong law of large numbers

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

Similarly we have, $\frac{1}{n}\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{b}_n^T\mathbf{A}_n^{-1}\mathbf{b}_n \overset{\text{a.s.}}{\to} \boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0$ and hence, $\tau_L^{-1} \overset{\text{a.s.}}{\to} \sigma_0^2 + 2\boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0$.

In a similar manner to τ_L we have

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

and $\frac{1}{n}\|\mathbf{y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\|^2 = \frac{1}{n}\|\mathbf{y}\|^2 - \frac{1}{n}\mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \overset{\text{a.s.}}{\to} \sigma_0^2$. Hence, by the continuous mapping theorem

$$\tau_L \overset{\text{a.s.}}{\to} [2\boldsymbol{\beta}_0^T\mathbf{S}\boldsymbol{\beta}_0 + \sigma_0^2]^{-1} \quad \text{and} \quad \tau_U \overset{\text{a.s.}}{\to} \sigma_0^{-2}$$

so that τ is bounded almost surely between two constants. Finally, since almost sure convergence implies convergence in probability the result holds.

□

We will next derive some properties for correct models. Note $\boldsymbol{\beta}_{0,-\boldsymbol{\gamma}} = \mathbf{0}$ by definition. In the following, we denote $j \in \boldsymbol{\gamma}$ if $\gamma_j = 1$ and $j \notin \boldsymbol{\gamma}$ if $\gamma_j = 0$. In the main result we assume that $\mathbf{w}^{(t)}$ is "close" to a correct model in probability (defined in Result 7). 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 2a, where $T > t$. In the next few results we use the following quantities:

$$\begin{aligned}\mathbf{T}_1 = \mathbf{T}_2 - \mathbf{T}_3\mathbf{T}_4, \qquad \mathbf{T}_2 = (n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}}) \odot (\boldsymbol{\Omega}_{\boldsymbol{\gamma}}^{(t)} - \mathbf{1}\mathbf{1}^T) + (n\tau^{(t)}\sigma_\beta^2)^{-1}\mathbf{I}, \\ \mathbf{T}_3 = (n\tau^{(t)}\sigma_\beta^2)\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}})\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}\left[\mathbf{I} + \mathbf{T}_5\right]^{-1}, \qquad \mathbf{T}_4 = \mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})\mathbf{W}_{\boldsymbol{\gamma}}^{(t)}, \\ \mathbf{T}_5 = (n\tau^{(t)}\sigma_\beta^2)(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{X}_{-\boldsymbol{\gamma}}) \odot \boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)} \quad \text{and} \quad \mathbf{t}_1 = (\mathbf{W}_{\boldsymbol{\gamma}}^{(t)} - \mathbf{I})(n^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{y}) - \mathbf{T}_3\mathbf{W}_{-\boldsymbol{\gamma}}^{(t)}(n^{-1}\mathbf{X}_{-\boldsymbol{\gamma}}^T\mathbf{y}).\end{aligned} \tag{35}$$

Result 7. *Assume (A1)–(A6) hold. Let $\boldsymbol{\gamma}$ be a correct model. Suppose that $\mathbf{w}^{(t)}$ is close to $\boldsymbol{\gamma}$ in probability in the following sense*

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

where d_{nj}, $1 \leq j \leq p$, is a sequences of positive random variables such that nd_{nj} converges in probability to zero. Then we say such a $w^{(t)}$ is close to the correct model $\boldsymbol{\gamma}$ in probability and

$$\mathbf{T}_1 = \mathbf{O}_p^m(n^{-1} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty + n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty^2), \quad \mathbf{T}_2 = \mathbf{O}_p^m(n^{-1} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty), \quad \mathbf{T}_3 = \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty),$$
$$\mathbf{T}_4 = \mathbf{O}_p^m(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty), \quad \mathbf{T}_5 = \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \quad \textit{and} \quad \mathbf{t}_1 = \mathbf{O}_p^v(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty).$$

Proof of Result 7: 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 $\mathbf{D}_{n,\boldsymbol{\gamma}} = \text{diag}(\mathbf{d}_{n,\boldsymbol{\gamma}})$. Similarly, $\boldsymbol{\Omega}_{-\boldsymbol{\gamma}}^{(t)} = \mathbf{O}_p^d(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Again, using (34) and Result 6 we have $\mathbf{T}_2 = [\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}_p^m(n^{-1/2})] \odot \mathbf{O}_p^d(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty) + \mathbf{O}_p^d(n^{-1}) = \mathbf{O}_p^m(n^{-1} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty)$. Next, using (34) and Result 6 we have

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

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 $\boldsymbol{\gamma}$ (in the above sense) 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,

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

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 (34) we have

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

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

□

Result 8. *Assume (A1)–(A6) hold. If $\mathbf{w}^{(t)}$ is close to a correct model $\boldsymbol{\gamma}$ in probability in the sense of Result 7, then*

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

Proof of Result 8: Firstly, note that

$$\tau(\mathbf{X}^T\mathbf{X}) \odot \mathbf{\Omega} + \sigma_\beta^{-2}\mathbf{I} = \tau\mathbf{W}\mathbf{X}^T\mathbf{X}\mathbf{W} + \mathbf{D} \tag{36}$$

by definition. Using equations (23), (24) and (36) we have

$$\begin{aligned}
\mathbf{\Sigma}^{(t)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} &= \left(n\tau^{(t)}\right)^{-1}\Big[(n^{-1}\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}) \odot \mathbf{\Omega}^{(t)}_{\boldsymbol{\gamma}} + (n\tau^{(t)}\sigma^2_\beta)^{-1}\mathbf{I} - \mathbf{W}^{(t)}_{\boldsymbol{\gamma}}(n^{-1}\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}})\mathbf{W}^{(t)}_{-\boldsymbol{\gamma}} \\
&\quad \times \left\{(n^{-1}\mathbf{X}^T_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}) \odot \mathbf{\Omega}^{(t)}_{-\boldsymbol{\gamma}} + (n\tau^{(t)}\sigma^2_\beta)^{-1}\mathbf{I}\right\}^{-1}\mathbf{W}^{(t)}_{-\boldsymbol{\gamma}}(n^{-1}\mathbf{X}^T_{-\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}})\mathbf{W}^{(t)}_{\boldsymbol{\gamma}}\Big]^{-1} \\
&= \left(n\tau^{(t)}\right)^{-1}\Big[(n^{-1}\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_1\Big]^{-1}.
\end{aligned} \tag{37}$$

Using equations (23), (29), (34) and (36), Result 7, and the continuous mapping theorem we have

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

Since by assumption $\|\mathbf{d}_n\|_\infty = o_p(n^{-1})$ we have $\boldsymbol{\mu}^{(t)}_{\boldsymbol{\gamma}}$ as stated. Using equations (23), (24) and (36) again we have

$$\mathbf{\Sigma}^{(t)}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}} = \Big[\sigma^{-2}_\beta\mathbf{I} + \tau^{(t)}(n^{-1}\mathbf{X}^T_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}) \odot (n\mathbf{\Omega}^{(t)}_{-\boldsymbol{\gamma}}) - n\mathbf{T}_4\left\{(n^{-1}\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_2\right\}^{-1}\mathbf{T}^T_4\Big]^{-1}.$$

From Equation (34) and Result 7, we can show that $(n^{-1}\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_2 = \mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}^m_p(n^{-1/2} + \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$ and $\mathbf{T}_4 = \mathbf{O}^m_p(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Using the continuous mapping theorem we find that

$$\begin{aligned}
&\tau^{(t)}(n^{-1}\mathbf{X}^T_{-\boldsymbol{\gamma}}\mathbf{X}_{-\boldsymbol{\gamma}}) \odot (n\mathbf{\Omega}^{(t)}_{-\boldsymbol{\gamma}}) - n\mathbf{T}_4\left[(n^{-1}\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_2\right]^{-1}\mathbf{T}^T_4 \\
&\quad = O_p(1)[\mathbf{S}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}} + \mathbf{O}^m_p(n^{-1/2})] \odot \mathbf{O}^m_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \\
&\qquad -n\mathbf{O}^m_p(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\left[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}^m_p(n^{-1/2} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty)\right]\mathbf{O}^m_p(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \\
&\quad = \mathbf{O}^m_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty).
\end{aligned}$$

Noting that by assumption $d_{nj} = o_p(n^{-1})$ and applying the continuous mapping theorem, we obtain the result for $\mathbf{\Sigma}^{(t)}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}}$. Next, from equations (23), (29), (34) and Result 7 we obtain

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

Lastly, using equations (23), (25), (34), Result 7 and by the assumption that $d_{nj} = o_p(n^{-1})$ we obtain

$$\begin{aligned}
\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\gamma},-\boldsymbol{\gamma}} &= -\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} \mathbf{W}^{(t)}_{\boldsymbol{\gamma}} \mathbf{X}^T_{\boldsymbol{\gamma}} \mathbf{X}_{-\boldsymbol{\gamma}} \mathbf{W}^{(t)}_{-\boldsymbol{\gamma}} \left[\mathbf{X}^T_{-\boldsymbol{\gamma}} \mathbf{X}_{-\boldsymbol{\gamma}} \odot \boldsymbol{\Omega}^{(t)}_{-\boldsymbol{\gamma}} + (\tau^{(t)}\sigma^2_\beta)^{-1} \mathbf{I} \right]^{-1} \\
&= -\left[(n^{-1} \mathbf{X}^T_{\boldsymbol{\gamma}} \mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_1 \right]^{-1} \sigma^2_\beta \mathbf{T}^T_4 \left(\mathbf{I} + \mathbf{T}_5 \right)^{-1} \\
&= \left[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}^m_p(n^{-1/2}) + \mathbf{O}^m_p(n^{-1} + \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty + n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|^2_\infty) \right]^{-1} \mathbf{O}^m_p(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \\
&\quad \times \left[\mathbf{I} + \mathbf{O}^m_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty) \right].
\end{aligned}$$

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

Result 9. *Assume (A1)–(A6) hold. If* $\mathbf{w}^{(t)}$ *is close to a correct model* $\boldsymbol{\gamma}$ *in probability in the sense of Result 7, then*

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

Proof of Result 9: Using Result 3 the value $\tau^{(t)}$ satisfies

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

where

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

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

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

Using Equation (34) we have $n^{-1}\|\boldsymbol{\varepsilon}\|^2 = \sigma_0^2 + O_p(n^{-1/2})$ and $n^{-1}\boldsymbol{\varepsilon}^T\mathbf{X} = \mathbf{O}^v_p(n^{-1/2})$ and $n^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{S} + \mathbf{O}^m_p(n^{-1/2})$. Note that from Result 8 we have $\boldsymbol{\mu}^{(t)}_{\boldsymbol{\gamma}} = \boldsymbol{\beta}_{0\boldsymbol{\gamma}} + \mathbf{O}^v_p(n^{-1/2})$ and $\boldsymbol{\mu}^{(t)}_{-\boldsymbol{\gamma}} = \mathbf{O}^v_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Then $\boldsymbol{\mu}^{(t)} = \boldsymbol{\beta}_0 + \mathbf{e}_n$ where $\mathbf{e}_{n,\boldsymbol{\gamma}} = \mathbf{O}^v_p(n^{-1/2})$ and $\mathbf{e}_{n,-\boldsymbol{\gamma}} = \mathbf{O}^v_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Lastly, by assumption

$$\begin{aligned}
\mathbf{W}^{(t)}\boldsymbol{\mu}^{(t)} - \boldsymbol{\beta}_0 &= \begin{bmatrix} \mathbf{e}_{n,\boldsymbol{\gamma}} - \mathbf{d}_{n,\boldsymbol{\gamma}} \odot \boldsymbol{\beta}_{0,\boldsymbol{\gamma}} - \mathbf{d}_{n,\boldsymbol{\gamma}} \odot \mathbf{e}_{n,\boldsymbol{\gamma}} \\ \mathbf{d}_{n,-\boldsymbol{\gamma}} \odot \mathbf{e}_{n,-\boldsymbol{\gamma}} \end{bmatrix} = \begin{bmatrix} \mathbf{O}^v_p(\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty + \|\mathbf{e}_{n,\boldsymbol{\gamma}}\|_\infty) \\ \mathbf{O}^v_p(\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\mathbf{e}_{n,-\boldsymbol{\gamma}}\|_\infty) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{O}^v_p(n^{-1/2} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty) \\ \mathbf{O}^v_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|^2_\infty) \end{bmatrix}.
\end{aligned}$$

Hence, $T_2 = \sigma_0^2 + O_p(n^{-1/2} + \|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty + n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|^2_\infty)$. By assumption $\|\mathbf{d}_{n,\boldsymbol{\gamma}}\|_\infty$ and $n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|^2_\infty$ 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 (34) 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 8 we have $(\mu_j^{(t)})^2 = \beta_{0j}^2 + O_p(e_{nj})$. Hence, $T_3 = O_p(\|\mathbf{d}_n\|_\infty)$ and so

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

□

Result 10. *Assume (A1)–(A6) hold. If* $\mathbf{w}^{(t)}$ *is close to a correct model* $\boldsymbol{\gamma}$ *in probability in the sense of Result 7, then*

$$\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} = \frac{\sigma_0^2}{n}\mathbf{S}^{-1}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}^m_p(n^{-3/2}).$$

Proof of Result 10: Using Equation (37), Results 7 and 9, the matrix $\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}$ may be written as

$$\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} \quad = (n\tau^{(t)})^{-1}\Big[(n^{-1}\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{X}_{\boldsymbol{\gamma}}) + \mathbf{T}_1\Big]^{-1}$$

where $\mathbf{T}_1$, $\mathbf{T}_4$, $\mathbf{T}_5$ are as defined by (35). Using similar arguments to Result 8 we have

$$\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} \quad = n^{-1}[\sigma_0^{-2} + O_p(n^{-1/2})]\Big[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} + \mathbf{O}^m_p(n^{-1/2}) + \mathbf{O}^m_p(n^{-1} + \|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty + n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|^2_\infty)\Big]^{-1}.$$

The result is proved after application of the continuous mapping theorem, expanding and dropping appropriate lower order terms to the above expression.

□

Result 11. *Assume (A1)–(A6) hold. If* $\mathbf{w}^{(t)}$ *is close to a correct model* $\boldsymbol{\gamma}$ *in probability in the sense of Result 7, then*

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

Proof of Result 11: If equations (3), (4) and (5) hold then by Equation (31) we have

$$\eta_j^{(t+1)} \quad = \lambda_n + T_6 + T_7 + T_8 + T_9,$$

where

$$T_6 = \tfrac{1}{2}\tau^{(t)}\|\mathbf{X}_j\|^2(\mu_j^{(t)})^2, \qquad T_7 = (w_j^{(t)})^{-1}\sigma_\beta^{-2}(\mu_j^{(t)})^2,$$
$$T_8 = -\tfrac{1}{2}\tau^{(t)}\|\mathbf{X}_j\|^2\Sigma^{(t)}_{j,j} \qquad \text{and} \qquad T_9 = w_j^{(t)}\tau^{(t)}\mathbf{X}_j^T\mathbf{P}_j^{(t)}\mathbf{X}_j\Sigma^{(t)}_{j,j}.$$

Note $T_6 \geq 0$ and $T_7 \geq 0$. Next, using Result 8 we have that if $j \in \boldsymbol{\gamma}$ and $j \in \boldsymbol{\gamma}_0$ then $(\mu_j^{(t)})^2 = (\beta_{0j} + O_p(n^{-1/2}))^2 = \beta_{0j}^2 + O_p(n^{-1/2})$. If $j \in \boldsymbol{\gamma}$ and $j \notin \boldsymbol{\gamma}_0$ then $(\mu_j^{(t)})^2 = (0 + O_p(n^{-1/2}))^2 = O_p(n^{-1})$. If $j \notin \boldsymbol{\gamma}$ and then $(\mu_j^{(t)})^2 = O_p(n^2\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|^2_\infty)$. Hence, using Result 9 and Equation (34) we have

$$\begin{aligned} T_6 \quad &= \begin{cases} \frac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]\left[\beta_{0j}^2 + O_p(n^{-1/2})\right] & j \in \boldsymbol{\gamma} \text{ and } j \in \boldsymbol{\gamma}_0, \\ \frac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]O_p(n^{-1}), & j \in \boldsymbol{\gamma} \text{ and } j \notin \boldsymbol{\gamma}_0, \\ \frac{n}{2}\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]O_p(n^2\|\mathbf{d}_n, -\boldsymbol{\gamma}\|^2_\infty), & j \notin \boldsymbol{\gamma} \end{cases} \\ &= \begin{cases} \frac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\beta_{0j}^2 + O_p(n^{1/2}) & j \in \boldsymbol{\gamma} \text{ and } j \in \boldsymbol{\gamma}_0, \\ |O_p(1)| & j \in \boldsymbol{\gamma} \text{ and } j \notin \boldsymbol{\gamma}_0, \\ |O_p(n^3\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|^2_\infty)| & j \notin \boldsymbol{\gamma}. \end{cases} \end{aligned}$$

For T_7 we need to consider the behavior of $(w_j^{(t)})^{-1}(\mu_j^{(t)})^2$. Note when $j \in \boldsymbol{\gamma}$ we have $(w_j^{(t)})^{-1} = 1 + O_p(|d_{nj}|)$ and this along with Equation (30) yields

$$T_7 \quad = \begin{cases} [1 - |O_p(d_{nj})|][\beta_{0j}^2 + O_p(n^{-1/2})]\sigma_\beta^{-2} = O_p(1) & j \in \boldsymbol{\gamma} \text{ and } j \in \boldsymbol{\gamma}_0, \\ [1 - |O_p(d_{nj})|]O_p(n^{-1})\sigma_\beta^{-2} = O_p(n^{-1}) & j \in \boldsymbol{\gamma} \text{ and } j \notin \boldsymbol{\gamma}_0, \\ |O_p(n^2\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)| & j \notin \boldsymbol{\gamma}. \end{cases}$$

Next, note that $T_8 \leq 0$ and from Result 10 we have $\mathbf{\Sigma}^{(t)}_{\boldsymbol{\gamma},\boldsymbol{\gamma}} = \frac{\sigma_0^2}{n}[\mathbf{S}_{\boldsymbol{\gamma},\boldsymbol{\gamma}}]^{-1} + \mathbf{O}_p^m(n^{-3/2}) = \mathbf{O}_p^m(n^{-1})$ and $\mathbf{\Sigma}^{(t)}_{-\boldsymbol{\gamma},-\boldsymbol{\gamma}} = \sigma_\beta^2\mathbf{I} + \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)$. Hence, with Result 9 and we have

$$
\begin{aligned}
T_8 &= \begin{cases} -\frac{n}{2}\left[\sigma_0^{-2} + O_p\big(n^{-1/2}\big)\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right] O_p\big(n^{-1}\big) & j \in \boldsymbol{\gamma}, \\ -\frac{n}{2}\left[\sigma_0^{-2} + O_p\big(n^{-1/2}\big)\right]\left[\mathbb{E}(x_j^2) + O_p\big(n^{-1/2}\big)\right]\left[\sigma_\beta^2 + O_p(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)\right] & j \notin \boldsymbol{\gamma}, \end{cases} \\
&= \begin{cases} -|O_p(1)| & j \in \boldsymbol{\gamma}, \\ -\frac{n}{2}\sigma_0^{-2}\mathbb{E}(x_j^2)\sigma_\beta^2 + O_p\big(n^{1/2} + n^2\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty\big) & j \notin \boldsymbol{\gamma}. \end{cases}
\end{aligned}
$$

Lastly, using Equation (22) we have

$$
\begin{aligned}
\mathbf{X}_j^T\mathbf{P}_j^{(t)}\mathbf{X}_j &= \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}\left(\mathbf{W}_{-j}\mathbf{X}_{-j}^T\mathbf{X}_{-j}\mathbf{W}_{-j} + \tau^{-1}\mathbf{D}_{-j}\right)^{-1}\mathbf{W}_{-j}\mathbf{X}_{-j}^T\mathbf{X}_j \\
&\leq \mathbf{X}_j^T\mathbf{X}_{-j}\mathbf{W}_{-j}\left(\mathbf{W}_{-j}\mathbf{X}_{-j}^T\mathbf{X}_{-j}\mathbf{W}_{-j}\right)^{-1}\mathbf{W}_{-j}\mathbf{X}_{-j}^T\mathbf{X}_j \\
&= \mathbf{X}_j^T\mathbf{X}_{-j}\left(\mathbf{X}_{-j}^T\mathbf{X}_{-j}\right)^{-1}\mathbf{X}_{-j}^T\mathbf{X}_j \\
&= n(n^{-1}\mathbf{X}_j^T\mathbf{X}_{-j})\left(n^{-1}\mathbf{X}_{-j}^T\mathbf{X}_{-j}\right)^{-1}(n^{-1}\mathbf{X}_{-j}^T\mathbf{X}_j) \\
&= n[\mathbf{S}_{j,-j} + \mathbf{O}_p^v(n^{-1/2})]\left[\mathbf{S}_{-j,-j} + \mathbf{O}_p^v(n^{-1/2})\right]^{-1}[\mathbf{S}_{-j,j} + \mathbf{O}_p^v(n^{-1/2})] \\
&= O_p(n).
\end{aligned}
$$

The second line follows from Lemma 3 and the other lines follow from simplifying and using Equation (34). Hence,

$$
T_9 = \begin{cases} (1+d_{nj})\left[\sigma_0^{-2} + O_p(n^{-1/2})\right]O_p(n)O_p(n^{-1}) = O_p(1) & j \in \boldsymbol{\gamma}, \\ d_{nj}[\sigma_0^{-2} + O_p(n^{-1/2})][O_p(n)][\sigma_\beta^2\mathbf{I} + \mathbf{O}_p^m(n\|\mathbf{d}_{n,-\boldsymbol{\gamma}}\|_\infty)] = O_p(nd_{nj}) & j \notin \boldsymbol{\gamma}. \end{cases}
$$

Combining the expressions for T_6, T_7, T_8 and T_9 and using the assumption that $d_{nj} = o_p(n^{-1})$ obtains the result.

□

Remark: At this stage we can see how Assumption (A7) comes into play. When $j \in \boldsymbol{\gamma}$ and $j \in \boldsymbol{\gamma}_0$ we do not want λ_n to dominate $n\mathbb{E}(x_j^2)\beta_{0j}^2/2\sigma_0^2$ and for $j \notin \boldsymbol{\gamma}$ and $j \in \boldsymbol{\gamma}_0$ we want λ_n to be of larger order than any $O_p(1)$ term. Hence, presuming that (A1)–(A7) hold, updates of the form $w_j^{(t+1)} \leftarrow \text{expit}(\eta_j^{(t+1)})$ will take us from a correct model to another correct model, with the jth variable possibly set to a value close to zero and effectively "removed."

Proof of Main Result 2: If $w_j^{(1)} = 1$ for $1 \leq j \leq p$ and assumptions (A1)–(A7) hold then $\mathbf{w}^{(1)} = \mathbf{1}$ corresponds to a correct model $\boldsymbol{\gamma} = \mathbf{1}$ and Results 8–11 hold with the sequence $d_{nj} = 0$ for $1 \leq j \leq p$ where the convergence rate of nd_{nj} being obviously satisfied. Hence, equations (10) and (11) are proved. In order to apply these theorems again for $j \in \boldsymbol{\gamma}_0$ we need

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

and for $j \notin \boldsymbol{\gamma}_0$ we need $w_j^{(2)} = \text{expit}\left[\lambda_n + O_p(1)\right] = d_{nj}$ for some sequence of random variables d_{nj}, $1 \leq j \leq p$ satisfying nd_{nj} converging in probability to zero. Hence,

- If $\lambda_n > 0$ and $\lambda_n \to \infty$ as $n \to \infty$ then $w_j^{(2)}$ will converge in probability to 1 for $1 \leq j \leq p$.

- 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 $w_j^{(2)}$ will be $O_p(1)$.
- If $\lambda_n < 0$ and $\lambda_n/n \to \kappa$ for some constant κ then $w_j^{(2)}$ 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 $w_j^{(2)}$ will converge in probability to 0 for $1 \leq j \leq p$.
- If $\lambda_n \to -\infty$ and $\lambda_n/n \to 0$ then d_{nj} will converge to 0 for $1 \leq j \leq p$, but for $j \notin \boldsymbol{\gamma}_0$ the sequence nd_{nj} may not converge in probability to zero.

Thus, we require $\lambda_n \to -\infty$, $\lambda_n/n \to 0$ and $n\,\mathrm{expit}(\lambda_n) = n\rho_n \to 0$. These are the conditions specified by Assumption (A7). Hence, under the assumptions (A1)–(A7) then we can apply Results 8–11 with $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0$ (the true model) to prove equations (12) and (13). We now note that the term $n^2\,\mathrm{expit}(\lambda_n)$ is $o_p(n)$ or smaller by Assumption (A7). However, by Assumption (A7) 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 8–11 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 8–11 recursively to prove the Main Result 2 for all t.

□