Statistics and Probability Letters 81 (2011) 1920–1928

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Consistency of spike and slab regression

Hemant Ishwaran *, J. Sunil Rao

University of Miami, 1120 NW 14th Street, Miami FL, United States

ARTICLE INFO

Article history:
Received 1 February 2011
Received in revised form 8 August 2011
Accepted 9 August 2011
Available online 19 August 2011

Keywords:
Oracle property
Posterior mean
Rescaling
Shrinkage
Two-component prior

ABSTRACT

Spike and slab models are a popular and attractive variable selection approach in regression settings. Applications for these models have blossomed over the last decade and they are increasingly being used in challenging problems. At the same time, theory for spike and slab models has not kept pace with the applications. There are many gaps in what we know about their theoretical properties. An important property known to hold in these models is selective shrinkage: a unique property whereby the posterior mean is shrunk toward zero for non-informative variables only. This property has been shown to hold under orthogonality for continuous priors under the modified class of rescaled spike and slab models. In this paper, we extend this result to the general case and prove an oracle property for the posterior mean under a discrete two-component prior. An immediate consequence is that a strong selective shrinkage property holds. Interestingly, the conditions needed for our result to hold in the non-orthogonal setting are more stringent than in the orthogonal case and amount to a type of enforced sparsity condition that must be met by the prior.

1. Introduction

Spike and slab regression was introduced by Lempers (1971) and Mitchell and Beauchamp (1988) who adopted a Bayesian approach to subset selection in linear regression models. The expression "spike and slab", coined by Mitchell and Beauchamp (1988), referred to the prior for the regression coefficients used in their Bayesian hierarchy. This prior was chosen such that the regression parameters were mutually independent with a two-point mixture distribution made up of a uniform flat distribution (the slab) and a degenerate distribution at zero (the spike). Later, George and McCulloch (1993) suggested a different approach. Instead of directly modeling the regression parameters, they advocated using a normal hierarchy with a two-component mixing distribution for the variance. This yields conditional distributions for all parameters from elementary distributions, making it possible to efficiently sample the posterior by Gibbs sampling. This important computational feature made spike and slab regression feasible in large scale problems and greatly enhanced its popularity (Geweke and Meese, 1981; Chipman, 1996; Clyde et al., 1996; Kuo and Mallick, 1998; Brown et al., 1998; Seo et al., 2007; Park and Casella, 2008).

As discussed in Ishwaran and Rao (2005a), normal hierarchies subject to a normal-variance mixture prior for the regression parameter, such as those used in George and McCulloch (1993), constitute a wide class of models termed *spike and slab models*. Spike and slab models were modified to the class of rescaled spike and slab models (Ishwaran and Rao, 2005a). These new models rescale the response by the square root of the sample size divided by some suitable estimate of the population standard deviation. Rescaling was shown to induce a non-vanishing penalization effect, and ensures a selective shrinkage property in orthogonal models when used in tandem with a continuous bimodal prior (Ishwaran and Rao, 2005a). This property allows the posterior mean for the coefficients to shrink toward zero for truly zero coefficients, while for non-zero coefficients, posterior estimates are similar to the ordinary least squares (OLS) estimator. Ishwaran and

* Corresponding author.
E-mail address: hemant.ishwaran@gmail.com (H. Ishwaran).

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.08.005

Rao (2005b) used rescaled spike and slab models to analyze multigroup microarray data (an extension of previous work of Ishwaran and Rao, 2003) and showed that selective shrinkage was a sufficient condition for oracle like total misclassification.

These results, however, all presuppose an orthogonal design matrix. The main goal of this manuscript is to extend this work by showing that selective shrinkage can hold in both orthogonal and non-orthogonal settings for certain classes of rescaled spike and slab models. In fact, we establish a stronger result by showing under certain conditions that the so-called "oracle property" holds (Fan and Li, 2001). An estimator is said to have the Fan–Li oracle property if it is sparse and asymptotically normal and possesses the same limiting distribution as the OLS constrained to the true nonzero coefficients. In establishing these results, we focus on two-component priors. These discrete priors have only two components: an atom near zero which constitutes the spike and a large atom away from zero which constitutes the slab. Our reason for using these priors is that their simple construction make them an excellent theoretical tool for studying selective shrinkage. At the same time, our results are highly suggestive of the conditions needed for other types of priors, including continuous ones.

Our main result (Theorem 1 of Section 3) shows that for the oracle property to hold in non-orthogonal settings, the spike in the two-component prior must converge to zero at $O(n^{-1})$ rate and the slab must converge to ∞ at a $O(n)$ rate. However, in the orthogonal case, although the slab must continue to satisfy a rate condition, no such condition is needed for the spike. This latter finding is consistent with previous work. In particular, Ishwaran and Rao (2005a), who considered orthogonal models, showed that a prior with continuous right tail and a near-zero spike was sufficient for selective shrinkage to hold. A continuous right tail functions like a slab converging to ∞ by ensuring that the variance can be arbitrarily large with positive probability. The spike near zero ensures that the variance can take a small value with positive probability—however, no stringent conditions for the spike were required for selective shrinkage to hold. Thus, our results for the two-component prior match what has been observed for continuous priors in the orthogonal case. At the same time, it is natural to speculate that our findings for the non-orthogonal case for the two-component prior should also apply to continuous priors. If so, this would suggest that such priors must satisfy a type of enforced sparsity. Theorem 1 provides guidance for the regularization required.

The paper is organized as follows. Section 2 provides background on spike and slab models and the selective shrinkage property. The oracle result for the posterior mean of the regression coefficients is given in Theorem 1 of Section 3. This result follows from Lemma 1 which proves a key fact about the frequentist property of the ridge matrix associated with the posterior mean. Section 3 discusses implications of Theorem 1. Finally, we conclude in Section 4 with a summary and discussion of our results.

2. Spike and slab regression

We begin by introducing notation that will be used throughout the paper. Following this we provide background on spike and slab and rescaled spike and slab models.

Our results apply to linear regression models. Specifically, it will be assumed that the underlying true (data) model is of the form:

$$y_i = \beta_{1,0}x_{i,1} + \cdots + \beta_{p,0}x_{i,p} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $(\varepsilon_i)_{1\le i\le n}$ are independent random variables such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma_0^2 > 0$. Write X for the $n \times p$ design matrix corresponding to (1) and $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})^T$ for the true regression parameter. The variables $x_i = (x_{i,1}, \ldots, x_{i,p})^T$ and the response-vector $y = (y_1, \ldots, y_n)^T$ are assumed to be standardized such that

$$\sum_{i=1}^n x_{i,k} = 0, \qquad \sum_{i=1}^n x_{i,k}^2 = n, \qquad \sum_{i=1}^n y_i = 0. \tag{2}$$

Note that the last constraint is satisfied by centering $(y_i)_{1\le i\le n}$ by the mean. Because y is assumed centered, no intercept term is included in (1).

2.1. Spike and slab models

We now review some background material on spike and slab models. Ishwaran and Rao (2005a) characterized a spike and slab model as being any model with a Bayesian hierarchy specified as follows:

$$\begin{aligned}
&(y|X, \beta, \sigma^2) \sim \mathrm{N}(X\beta, \sigma^2 I_n)\\
&(\beta|\gamma) \sim \mathrm{N}(0, \Gamma)\\
&\gamma := (\gamma_1, \ldots, \gamma_p)^T \sim \pi(\cdot)\\
&\sigma^2 \sim \mu(\cdot).
\end{aligned} \tag{3}$$

Here $\beta = (\beta_1, \ldots, \beta_p)^T$ is the regression vector and $\Gamma = \mathrm{diag}\{\gamma_k\}_{1\le k\le p}$ is its $p \times p$ diagonal hypervariance matrix.

The prior π for the hypervariance γ plays a critical role in how effective (3) is for variable selection. A successful and popular choice for π are priors that make use of mixture distributions involving a spike near zero. As discussed earlier, the

prototype for such priors was discussed in George and McCulloch (1993). There, the prior for γ_k was assumed to have a two-component distribution of the form

$$(\gamma_k|\tau_k, c_k, \varpi_k) \overset{\text{ind}}{\sim} (1-\varpi_k)\delta_{\tau_k}(\cdot) + \varpi_k\delta_{c_k\tau_k}(\cdot), \quad k = 1, \ldots, p. \tag{4}$$

The value for $\tau_k > 0$ (the spike) is chosen to be some small value, where "small" is typically specified based on the data at hand, while $c_k > 0$, also data-specific, is chosen so that $c_k\tau_k$ (the slab) is sufficiently large. Selecting the two hyperparameters in this way allows γ_k to be either small or large, and this in turn enables the posterior of β_k to shrink toward zero or be some nonzero value. The values $(\varpi_k)_{k=1}^p$ are complexity parameters that influence the likelihood of a coefficient being shrunk toward zero. In principle, each variable can have a unique complexity value, but a common practice is to set $\varpi_k = 1/2$ for each k, in which case (4) is referred to as an indifference prior.

2.2. Rescaled spike and slab models

To improve variable selection properties of spike and slab models, Ishwaran and Rao (2005a) introduced a slightly different class of models referred to as rescaled spike and slab models. The results presented in this manuscript focus only on this class of models. Rescaled spike and slab models are models of the form

$$\begin{aligned}
&(y^*|X, \beta, \sigma^2) \sim \mathrm{N}(X\beta, n\sigma^2 I_n), \quad y^* = n^{1/2}y\\
&(\beta|\gamma) \sim \mathrm{N}(0, \Gamma)\\
&\gamma \sim \pi(\cdot)\\
&\sigma^2 \sim \mu(\cdot).
\end{aligned} \tag{5}$$

Observe that (5) differs slightly from (3) in that it uses a modified response y^* in place of y. Another difference is the rescaled variance appearing in the first level of (5) (scaled by a factor of n). Although these differences may appear cosmetic, the effects of rescaling are crucial to variable selection performance (Ishwaran and Rao, 2005a). Without the correct scaling (of size n), penalization vanishes asymptotically and the limiting distribution for the posterior mean of β equals that of the OLS. In contrast, rescaling ensures a different limit than the OLS, with rescaling acting as a penalization parameter. Variable selection improves because the posterior mean acquires a selective shrinkage property. For this paper, we adopt the following strong definition of selective shrinkage.

Definition 1. Let $\hat\beta^*$ be the posterior mean under a rescaled spike and slab model. Then, $\hat\beta^*$ is said to possess the strong selective shrinkage property if $n^{-1/2}\hat\beta^* = O_p(1)$ and $n^{-1/2}\hat\beta^*_k \xrightarrow{\text{p}} 0$ if and only if $\beta_{0,k} = 0$.

3. Strong selective shrinkage and the oracle property for the posterior mean

Under orthogonality, a sufficient condition for selective shrinkage is that π should have: (i) a right tail that is continuous; and (ii) a spike near zero (see Ishwaran and Rao, 2005a, Theorem 6). The continuity of the right tail allows nonzero coefficients to be minimally penalized, and thus little or no shrinkage of the posterior mean occurs for such coefficients, whereas the spike near zero induces sparsity by shrinking the posterior mean of zero coefficients.

Here we prove that selective shrinkage (of the strong type) holds in general. Interestingly, we find that the conditions needed for consistency in the non-orthogonal case involve more stringent constraints on the behavior of the prior at zero than in the orthogonal case. To establish this result, we shall prove the stronger assertion that the posterior mean possesses the oracle property. Because the oracle property is difficult to prove in general, we will adopt the following simplification. Our proof will be based on the slightly modified rescaled spike and slab model:

$$\begin{aligned}
&(\tilde y|X, \beta) \sim \mathrm{N}(X\beta, nI_n), \quad \tilde y = \hat\sigma^{-1}n^{1/2}y\\
&(\beta|\gamma) \sim \mathrm{N}(0, \Gamma).
\end{aligned} \tag{6}$$

This model differs from (5) because we have replaced y^* by $\tilde y$, where $\tilde y$ is scaled by the additional factor $\hat\sigma^{-1}$ (here $\hat\sigma^2$ is an estimator for σ_0^2 satisfying certain properties; we say more about this shortly). Rescaling by $\hat\sigma^{-1}$ allows us to remove the prior for σ^2, which simplifies our arguments.

Our results will be based on a two-component prior for γ. We assume γ_k takes either a small value $w > 0$ (the spike) or a large value $W > 0$ (the slab) specified according to

$$\begin{aligned}
&(\gamma_k|\varpi) \overset{\text{ind}}{\sim} (1-\varpi)\delta_w(\cdot) + \varpi\delta_W(\cdot), \quad k = 1, \ldots, p\\
&\varpi \sim \text{Uniform}[0, 1].
\end{aligned} \tag{7}$$

Notice that (7) is similar to (4), except that we have extended the hierarchy to include a prior for the complexity parameter ϖ. Here we assume ϖ has a uniform density on $[0, 1]$. Theorem 1 will show why this is a reasonable choice.

3.1. Conditions

Here we state the conditions needed for our proof. We will need the following additional notation to describe these. Map each γ to the model comprising those coefficients for which $\gamma_k = W$ (i.e., these are the promising variables not shrunk toward zero). Thus, each γ is uniquely identified with a model comprising those regression coefficients with indices in some set $\alpha \subseteq \{1, \ldots, p\}$. We think of α and the model corresponding to γ as being interchangeable.

Write α_0 for the true model and let $p_0 = \#\alpha_0$ be its cardinality. For our results, we make use of the following regularity conditions:

$$\hat{\sigma}^2 \xrightarrow{\text{p}} s_0^2, \quad \text{where } 0 < s_0^2 < \infty. \tag{8}$$

$$\alpha_0 \neq \emptyset \quad \text{and } \alpha_0 \neq \{1, \ldots, p\}. \tag{9}$$

$$C_n = n^{-1} X^T X > 0 \quad \text{and} \lim_{n\to\infty} C_n = C > 0. \tag{10}$$

Condition (8) is mild and requires only that $\hat{\sigma}^2$ converges to a positive, finite value; a consistent estimator for σ_0^2 is not necessary (see Remarks 1 and 2 for some practical examples of how to select $\hat{\sigma}^2$). Condition (9) is purely a technical requirement: selective shrinkage is not well defined if all coefficients are zero, or if the true model is the full model. However, our results continue to hold even if (9) does not. Condition (10) assumes positive definiteness of C_n and is a standard assumption for the design matrix.

Remark 1. One simple way to select $\hat{\sigma}^2$ is to use the unbiased estimator for σ_0^2 obtained from least-squares. Let $\hat{y}_{\text{OLS}}$ denote the predictor for y based on the OLS estimator. Define

$$\hat{\sigma}^2 = \frac{1}{n-p} \|y - \hat{y}_{\text{OLS}}\|^2.$$

If we strengthen our assumption for $(\varepsilon_i)_{i=1}^n$ to include a bounded fourth moment, so that $(\varepsilon_i)_{i=1}^n$ are independent such that $\mathbb{E}(\varepsilon_i) = 0$ and $\mathbb{E}(\varepsilon_i^2) = \sigma_0^2$ and $\mathbb{E}(\varepsilon_i^4) \leq M < \infty$, then $\hat{\sigma}^2$ is consistent. To see why, first note that $\hat{\sigma}^2 = (\varepsilon^T\varepsilon - \varepsilon^T H\varepsilon)/(n-p)$, where $H = X(X^TX)^{-1}X$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Observe that H is well defined by (10). The first term $\varepsilon^T\varepsilon/(n-p)$ has a finite mean $n\sigma_0^2/(n-p)$ converging to σ_0^2 and a variance of order $O(n^{-1})$. Therefore, $\varepsilon^T\varepsilon/(n-p) \xrightarrow{\text{p}} \sigma_0^2$. For the second term, using Markov's inequality, for each $\delta > 0$,

$$\mathbb{P}\left\{\varepsilon^T H\varepsilon \geq (n-p)\delta\right\} \leq \frac{\mathbb{E}(\varepsilon^T H\varepsilon)}{(n-p)\delta} = \frac{\text{trace}(H\mathbb{E}(\varepsilon\varepsilon^T))}{(n-p)\delta} = \frac{p\sigma_0^2}{(n-p)\delta} \to 0.$$

Deduce that $\hat{\sigma}^2 \xrightarrow{\text{p}} \sigma_0^2$.

Remark 2. Here we describe a more general $\hat{\sigma}^2$ estimator. This method can be used with both parametric and nonparametric methods but works best with procedures known to yield low test-set mean-squared-error (MSE). For example, we recommend using a machine learning method such as Random Forests (Breiman, 2001). Although we do not provide a rigorous proof, we outline why our proposed estimator should satisfy (8). Let $\hat{f}$ denote the predictor for y. Let y_{new} be an N-dimensional vector of new test-set observations and let $\hat{f}_{\text{new}}$ be the predictor of y_{new}. Let $f_{\text{new}} = \mathbb{E}(y_{\text{new}})$. Taking expectations over ε_{new} (independent error terms associated with the test data):

$$\begin{aligned} N^{-1}\mathbb{E}\|y_{\text{new}} - \hat{f}_{\text{new}}\|^2 &= N^{-1}\mathbb{E}\|f_{\text{new}} + \varepsilon_{\text{new}} - \hat{f}_{\text{new}}\|^2 \\ &= N^{-1}\mathbb{E}\|\hat{f}_{\text{new}} - f_{\text{new}}\|^2 + N^{-1}\mathbb{E}\|\varepsilon_{\text{new}}\|^2 \\ &= N^{-1}\mathbb{E}\|\hat{f}_{\text{new}} - f_{\text{new}}\|^2 + \sigma_0^2. \end{aligned}$$

The first term on the right is the test-set MSE. If this value is small (which we would expect with an accurate predictor), then $N^{-1}\mathbb{E}\|y_{\text{new}} - \hat{f}_{\text{new}}\|^2$ will nearly equal σ_0^2. Thus, not only will this estimator satisfy (8), but it also provides a means for estimating σ_0^2. In practice, to calculate $N^{-1}\mathbb{E}\|y_{\text{new}} - \hat{f}_{\text{new}}\|^2$, one could average $N^{-1}\|y_{\text{new}} - \hat{f}_{\text{new}}\|^2$ over different test-sets (in our experience we have found using out-of-bag MSE estimates from Random Forests to be quite good). Finally, note that even if $\hat{f}_{\text{new}}$ is not perfectly accurate, we always obtain an upper bound to σ_0^2:

$$N^{-1}\mathbb{E}\|y_{\text{new}} - \hat{f}_{\text{new}}\|^2 \geq \sigma_0^2.$$

Thus it remains bounded away from zero.

3.2. The oracle result

To prove the oracle result, we first state a key lemma. Let $\hat{\beta} = \hat{\sigma} n^{-1/2} \hat{\beta}^*$, where $\hat{\beta}^*$ is the posterior mean from (6)–(7). Then

$$\hat{\beta} = n^{-1} \mathbb{E}(\Sigma | \tilde{y}) X^T y, \tag{11}$$

where $\Sigma = (C_n + \Gamma^{-1})^{-1}$. This shows that the effects of shrinkage are captured exclusively by the posterior mean of the ridge matrix Σ. The following result describes frequentist properties of Σ under rate conditions on w and W and will be fundamental to proving the oracle property.

Lemma 1. *Assume (1) is the true model, the data has been standardized as in (2), and that regularity conditions (8)–(10) hold. Under the rescaled spike and slab model (6) with prior (7), if $W \to \infty$ and $w \to 0$ such that $W = O(n)$ and $w = O(n^{-1})$, then*

$$\mathbb{E}(\Sigma | \tilde{y}) \xrightarrow{\text{p}} C_{[\alpha_0]}^{-1},$$

where $C_{[\alpha_0]}^{-1}$ is the $p \times p$ symmetric matrix equal to zero everywhere except along the coordinates corresponding to α_0 where it equals $(C_{(\alpha_0)})^{-1}$, where the subscript (α) indicates a term including only those indices in α.

To understand how Lemma 1 implies the oracle property, it is instructive to consider the orthogonal setting, $C_n = I_p$. Define $\mathbb{J}_k = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ to be the p-dimensional vector with the value of 1 in the kth coordinate and 0's elsewhere. Under the asserted conditions, it follows that

$$\mathbb{E}(\Sigma | \tilde{y}) \xrightarrow{\text{p}} \sum_{k \in \alpha_0} \mathbb{J}_k \mathbb{J}_k^T.$$

Moreover, because $n^{-1/2} X^T y = n^{1/2} \beta_0 + O_p(1)$, this implies from (11) that

$$n^{1/2} \left(\hat{\beta} - (1 + o_p(1)) \sum_{k \in \alpha_0} \beta_{0,k} \mathbb{J}_k \right) \overset{\text{d}}{\rightsquigarrow} \sigma_0 \left(\sum_{k \in \alpha_0} \mathbb{J}_k \mathbb{J}_k^T \right) Z,$$

for some random vector $Z \in \mathbb{R}^p$. Notice that the limiting covariance for $\hat{\beta}$ is zero if $\beta_{0,k} = 0$. Furthermore, by slightly tightening our assumptions regarding $(\varepsilon_i)_{i=1}^n$, we can assert asymptotic normality for Z, from which the oracle property follows.

Lemma 1 allows us to extend this argument to arbitrary X-designs. This yields our main result, Theorem 1, stated next. Some remarks helpful for interpreting the theorem are given following this.

Theorem 1. *Assume that the conditions of Lemma 1 hold and that $(\varepsilon_i)_{i=1}^n$ are i.i.d. and $n^{-1} \max_{i=1}^n \|x_i\|_2^2 \to 0$, where x_i is the ith row of X. Then:*

(i) $\hat{\beta}_{(\alpha_0^c)} = o_p(1)$.

(ii) $n^{1/2} \left(\hat{\beta}_{(\alpha_0)} - (1 + o_p(1)) \beta_{0,(\alpha_0)} \right) \overset{\text{d}}{\rightsquigarrow} N(0, \sigma_0^2 (C_{(\alpha_0)})^{-1})$.

Thus, the rescaled posterior mean, $\hat{\beta}$, has the Fan–Li oracle property.

Remark 3. Clearly (i) and (ii) imply that $n^{-1/2} \hat{\beta}^*$ shrinks to zero in probability for the truly zero coefficients and is stochastically bounded otherwise. Thus, Theorem 1 immediately implies that the posterior mean, $\hat{\beta}^*$, has the strong selective shrinkage property.

Remark 4. The rate conditions for the prior can be considerably weakened in the orthogonal case. One can show that for the slab, any sequence satisfying $n - \log W \to \infty$ is permitted, whereas for the spike there is no imposed constraint on w at all. Interestingly, this shows that the growth rate for W and the shrinkage rate for w are considerably more stringent in correlated settings.

Remark 5. In the orthogonal case, the posterior has correct asymptotic complexity recovery. Under a Beta(a, a) prior for ϖ,

$$\mathbb{E}(\varpi | \tilde{y}) \xrightarrow{\text{p}} (p_0 + a)/(p + 2a).$$

If p is large, this will be close to the true complexity p_0/p, for any reasonably selected a. As one example, the value $a = 1$, corresponding to a uniform prior, would be appropriate.

4. Discussion

In this article we established an oracle property for rescaled spike and slab models under general X-design scenarios. One immediate consequence is that the posterior mean for our class of models has the strong selective shrinkage property. In order to prove our result, we made use of a simplified two-component prior. The simple construction of this prior made it an excellent tool for studying selective shrinkage. In practical settings, however, discrete priors are difficult to use and instead continuous priors are often advocated. Nevertheless, we believe our results are applicable to more general priors and moreover should prove useful for developing regularization strategies for them. We believe this to be the case for several reasons. First, we found that our results matched up with what has been found for continuous priors in orthogonal settings: namely, that the right tail behavior of the prior is crucial for selective shrinkage to hold. Secondly, we suspect that having a small spike at zero, as we do for our two-component prior, is an essential ingredient to successful variable selection in non-orthogonal settings, regardless of whether the prior is discrete or not. We conjecture that for strong selective shrinkage and oracle-like properties to hold, the prior must satisfy an enforced sparsity property. Theorem 1 suggests that the prior must have a spike near zero on order $O(n^{-1})$.

Acknowledgments

This work was partially funded by the National Science Foundation grant DMS-0705037. We thank the Associate Editor and two referees for helpful comments.

Appendix. Proofs

Proof of Lemma 1. Throughout the proof we use a subscript of α to indicate dependence upon γ whenever this distinction is necessary. For example, Σ_α refers to $(C_n + \Gamma_\alpha^{-1})^{-1}$, where Γ_α is the diagonal matrix comprised of hypervariances defined by α. Subscripts of the form (α) are used to indicate a term containing only those coordinates corresponding to α. Thus $X_{(\alpha)}$ is the X matrix constrained to those columns containing α. Also, the notation $O(\cdot)$ and $o(\cdot)$, and their stochastic counterparts $O_p(\cdot)$ and $o_p(\cdot)$, are used not only for random variables but also vectors and matrices. In the latter two cases, this is taken to mean convergence under the metric $\left(\sum_k A_k^2\right)^{1/2}$ for vectors and $\left(\sum_{j,k} A_{j,k}^2\right)^{1/2}$ for matrices, which corresponds to uniform pointwise convergence as the dimension of all vectors and matrices are finite and because the cardinality of α is finite and fixed. Finally, throughout the proof we implicitly assume that condition (9) holds, but the proof can be suitably modified if the condition does not hold.

The posterior for (γ, ϖ) equals

$$\pi(\gamma, \varpi|\tilde{y}) = \int_{\mathbb{R}^p} \pi(\beta, \gamma, \varpi|\tilde{y})\, \mathrm{d}\beta.$$

To be able to work this out, we first need to derive $\pi(\beta, \gamma, \varpi|\tilde{y})$. Define $Z_n = n^{-1/2}X^T y$ and recall that $\Sigma = (C_n + \Gamma^{-1})^{-1}$ where $C_n = n^{-1}X^T X$. By (6) and (7), deduce that

$$\begin{aligned}
\pi(\beta, \gamma, \varpi|\tilde{y}) &\propto \pi(\gamma|\varpi)\pi(\beta|\gamma) \exp\left(-\frac{1}{2n}\|\tilde{y} - X\beta\|^2\right) \\
&\propto \pi(\gamma|\varpi)|2\pi\Gamma|^{-1/2} \exp\left(-\frac{1}{2}\beta^T\Gamma^{-1}\beta + \hat{\sigma}^{-1}Z_n^T\beta - \frac{1}{2}\beta^T C_n\beta\right) \\
&\propto \pi(\gamma|\varpi)|\Gamma\Sigma^{-1}|^{-1/2} \exp\left(\frac{1}{2\hat{\sigma}^2}Z_n^T\Sigma Z_n\right) \\
&\quad \times \left[|2\pi\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(\beta - \hat{\sigma}^{-1}\Sigma Z_n)^T\Sigma^{-1}(\beta - \hat{\sigma}^{-1}\Sigma Z_n)\right)\right].
\end{aligned}$$

The term in square brackets on the right-hand side is a multivariate normal distribution in β. Integrating over β, it follows that

$$\pi(\gamma, \varpi|\tilde{y}) \propto \pi(\gamma|\varpi)|\Gamma\Sigma^{-1}|^{-1/2} \exp\left(\frac{1}{2\hat{\sigma}^2}Z_n^T\Sigma Z_n\right).$$

The prior probability for α, denoted by $\Pr(\alpha)$, is uniquely determined by γ. By the definition of the two-component prior (7), one can easily see that

$$\begin{aligned}
\Pr(\alpha) &= \int_0^1 \prod_{\gamma_k = W} \varpi \prod_{\gamma_k = w} (1 - \varpi)\, \mathrm{d}\varpi \\
&= \int_0^1 \varpi^{p_\alpha}(1 - \varpi)^{p - p_\alpha}\, \mathrm{d}\varpi,
\end{aligned}$$

where p_α equals the number of coordinates of γ equal to W; i.e., p_α is the cardinality of α. Mapping each γ to its model α and integrating over ϖ, deduce that

$$\mathbb{E}(\Sigma|\tilde{y}) = \frac{\sum_\alpha \Pr(\alpha)\,\Sigma_\alpha\,|\Gamma_\alpha\,\Sigma_\alpha^{-1}|^{-1/2}\exp\left(\frac{1}{2\hat{\sigma}^2}Z_n^T\,\Sigma_\alpha Z_n\right)}{\sum_\alpha \Pr(\alpha)\,|\Gamma_\alpha\,\Sigma_\alpha^{-1}|^{-1/2}\exp\left(\frac{1}{2\hat{\sigma}^2}Z_n^T\,\Sigma_\alpha Z_n\right)}. \tag{12}$$

Let P_α be a $p\times p$ orthogonal matrix that rotates the coordinate axes so that the first p_α coordinates correspond to α. Partition $P_\alpha\,\Sigma_\alpha P_\alpha^T$ as follows:

$$P_\alpha\,\Sigma_\alpha P_\alpha^T = \left(P_\alpha C_n P_\alpha^T + P_\alpha\,\Gamma_\alpha^{-1}P_\alpha^T\right)^{-1} = \begin{pmatrix} A_{1,1} & A_{1,2} \\ A_{1,2}^T & w^{-1}A_{2,2}\end{pmatrix}^{-1},$$

where

$$\begin{aligned} A_{1,1} &= n^{-1}X_{(\alpha)}^T X_{(\alpha)} + W^{-1}I_{(\alpha)} \\ A_{1,2} &= n^{-1}X_{(\alpha)}^T X_{(\alpha^c)} \\ A_{2,2} &= n^{-1}wX_{(\alpha^c)}^T X_{(\alpha^c)} + I_{(\alpha^c)}. \end{aligned}$$

By standard matrix algebra,

$$P_\alpha\,\Sigma_\alpha P_\alpha^T = \begin{pmatrix} B_{1,1} & B_{1,2} \\ B_{1,2}^T & B_{2,2}\end{pmatrix}$$

where

$$\begin{aligned} B_{1,1} &= \left(A_{1,1} - wA_{1,2}A_{2,2}^{-1}A_{1,2}^T\right)^{-1} \\ B_{1,2} &= -A_{1,1}^{-1}A_{1,2}B_{2,2} \\ B_{2,2} &= w\left(A_{2,2} - wA_{1,2}^T A_{1,1}^{-1}A_{1,2}\right)^{-1}. \end{aligned}$$

By assumption (10), $A_{1,1}$, $A_{1,1}^{-1}$, $A_{2,2}$, $A_{2,2}^{-1}$, and $A_{1,2}$ are all well defined, and all have well defined limits. By the rates imposed on W and w, one can show that

$$\begin{aligned} \Sigma_\alpha &= P_\alpha^T\begin{pmatrix}(C_{n,(\alpha)})^{-1} & 0 \\ 0 & 0\end{pmatrix}P_\alpha + O(n^{-1}) \\ &= C_{n,[\alpha]}^{-1} + O(n^{-1}). \end{aligned} \tag{13}$$

Now observe that $\Gamma_\alpha\,\Sigma_\alpha^{-1} = \Gamma_\alpha C_n + I$. Therefore,

$$P_\alpha\,\Gamma_\alpha\,\Sigma_\alpha^{-1}P_\alpha^T = \left(P_\alpha\,\Gamma_\alpha P_\alpha^T\right)\left(P_\alpha C_n P_\alpha^T\right) + I = \begin{pmatrix} WA_{1,1} & WA_{1,2} \\ wA_{1,2}^T & A_{2,2}\end{pmatrix}.$$

The determinant of the left-hand side equals $|\Gamma_\alpha\,\Sigma_\alpha^{-1}|$. Hence, using standard properties for determinants, we have

$$|\Gamma_\alpha\,\Sigma_\alpha^{-1}| = W^{p_\alpha}|A_{1,1}|\,|A_{2,2} - wA_{1,2}^T A_{1,1}^{-1}A_{1,2}|.$$

Observe that $A_{1,1}\to C_{(\alpha)}$, $A_{2,2}\to I_{(\alpha^c)}$, and $wA_{1,2}^T A_{1,1}^{-1}A_{1,2}\to 0$. Deduce that

$$W^{p_\alpha/2}|\Gamma_\alpha\,\Sigma_\alpha^{-1}|^{-1/2}\to|C_{(\alpha)}|^{-1/2},\quad \text{as } n\to\infty. \tag{14}$$

Now multiply the numerator and denominator of (12) by the value

$$W^{p_0/2}\exp\left(-\frac{1}{2\hat{\sigma}^2}Z_n^T\,\Sigma_{\alpha_0}Z_n\right).$$

Recalling our assumption (2), $n^{-1/2}X^T\varepsilon = O_p(1)$ because

$$n^{-1}\mathbb{E}\left(\sum_{i=1}^n x_{i,k}\varepsilon_i\right)^2 = \sigma_0^2 n^{-1}\sum_{i=1}^n x_{i,k}^2 = \sigma_0^2,\quad 1\le k\le p.$$

Hence, by (13),

$$Z_n^T\,\Sigma_\alpha Z_n - Z_n^T\,\Sigma_{\alpha_0}Z_n = Z_n^T\left(C_{n,[\alpha]}^{-1} - C_{n,[\alpha_0]}^{-1}\right)Z_n + O_p(1). \tag{15}$$

The $O_p(1)$ term on the right-hand side of (15) holds because $Z_n^T Z_n = O_p(n)$ due to

$$Z_n = n^{-1/2} X^T X \beta_0 + n^{-1/2} X^T \varepsilon = n^{1/2} C_n \beta_0 + O_p(1).$$

By (14) and (15), the dominating term asymptotically for each term in the sum of the numerator (or the denominator) of (12) is

$$W^{(p_0 - p_\alpha)/2} \exp\left(\frac{1}{2\hat{\sigma}^2} \left[Z_n^T \left(C_{n,[\alpha]}^{-1} - C_{n,[\alpha_0]}^{-1} \right) Z_n + O_p(1) \right] \right). \tag{16}$$

Consider the term inside the exponent of (16). By (8), $\hat{\sigma}^2$ remains bounded away from zero and is finite, thus the key term is

$$Z_n^T \left(C_{n,[\alpha]}^{-1} - C_{n,[\alpha_0]}^{-1} \right) Z_n.$$

Let $\hat{\beta}_{\mathrm{OLS},(\alpha)} = (X_{(\alpha)}^T X_{(\alpha)})^{-1} X_{(\alpha)}^T y$ denote the constrained OLS estimator for α. With rearrangement one can show

$$\begin{aligned} Z_n^T C_{n,[\alpha]}^{-1} Z_n &= n^{-1} y^T X_{(\alpha)} (C_{n,(\alpha)})^{-1} X_{(\alpha)}^T y \\ &= \left(X_{(\alpha)} \hat{\beta}_{\mathrm{OLS},(\alpha)} \right)^T \left(X_{(\alpha)} \hat{\beta}_{\mathrm{OLS},(\alpha)} \right). \end{aligned}$$

The right-hand side is the squared ℓ_2-length of the projection of y onto X_α.

For each α, let $\nu_1, \ldots, \nu_n$ be an orthonormal basis for $\mathbb{R}^n$ such that the first p_α vectors span the column space of $X_{(\alpha)}$ and the first p_{α_0} vectors span the column space of $X_{(\alpha_0)}$. It is clear that such an orthonormal basis can always be constructed for α if either $\alpha \subseteq \alpha_0$ or $\alpha_0 \subset \alpha$. We assume that this is the case for now. Let $\theta_k = y^T \nu_k$. Then by the definition of the OLS, $X_{(\alpha)} \hat{\beta}_{\mathrm{OLS},(\alpha)} = \sum_{k \le p_\alpha} \theta_k \nu_k$ and

$$Z_n^T C_{n,[\alpha]}^{-1} Z_n = \sum_{k=1}^{p_\alpha} \theta_k^2.$$

Consequently if $\alpha \subseteq \alpha_0$,

$$Z_n^T \left(C_{n,[\alpha]}^{-1} - C_{n,[\alpha_0]}^{-1} \right) Z_n = - \sum_{k=p_\alpha+1}^{p_{\alpha_0}} \theta_k^2.$$

Because $\theta_k^2 = O_p(n)$ if $k \le p_{\alpha_0}$, the exponent in (16) becomes the dominant term and converges to zero in probability (the only exception being when $\alpha = \alpha_0$; then the exponent is exactly zero). On the other hand if $\alpha_0 \subset \alpha$, then $p_\alpha > p_0$ and $W^{(p_0 - p_\alpha)/2} \to 0$. Furthermore, $\theta_k^2 = O_p(1)$ for $k > p_{\alpha_0}$ (this follows from $y^T \nu_k = \varepsilon^T \nu_k$ and $\mathbb{E}(\varepsilon^T \nu_k)^2 = \sigma_0^2$, because $\|\nu_k\|_2^2 = 1$). Thus when $p_\alpha > p_0$, the polynomial $W^{(p_0 - p_\alpha)/2}$ becomes the dominating term and once again (16) converges to zero. Consequently, if α is a model such that either $\alpha \subseteq \alpha_0$ or $\alpha_0 \subset \alpha$, then the only α with a nonzero limiting contribution to either the numerator or denominator of (12) is α_0.

Now we consider the scenario when $\alpha \not\subseteq \alpha_0$ and $\alpha_0 \not\subseteq \alpha$. Now construct an orthonormal basis so that the first $P' = p_{\alpha'}$ vectors span the column space of $X_{(\alpha')}$, the first $P'' = p_{\alpha'} + p_{\alpha''}$ vectors span the column space of $(X_{(\alpha')}, X_{(\alpha'')})$, and the first $P''' = p_{\alpha'} + p_{\alpha''} + p_{\alpha'''}$ vectors span the column space of $(X_{(\alpha')}, X_{(\alpha'')}, X_{(\alpha''')})$, where $\alpha' = \alpha \cap \alpha_0^c$, $\alpha'' = \alpha \cap \alpha_0$, and $\alpha''' = \alpha^c \cap \alpha_0$. Because the squared-length of the projection of y onto $\{\nu_j\}_{j=P'+1}^{P'''}$ is less than the squared-length of the projection of y onto $X_{(\alpha_0)}$, we have

$$Z_n^T C_{n,[\alpha_0]}^{-1} Z_n \ge \sum_{k=P'+1}^{P'''} \theta_k^2,$$

and therefore

$$Z_n^T \left(C_{n,[\alpha]}^{-1} - C_{n,[\alpha_0]}^{-1} \right) Z_n \le \sum_{k=1}^{P''} \theta_k^2 - \sum_{k=P'+1}^{P'''} \theta_k^2 = \sum_{k=1}^{P'} \theta_k^2 - \sum_{k=P''+1}^{P'''} \theta_k^2.$$

The first sum on the right is $O_p(1)$ while the second sum is $O_p(n)$. Thus, the right-hand side converges to $-\infty$ in probability. Because of this, even though p_α may be larger than p_{α_0}, the dominating term in (16) is the exponent and hence α has a vanishing contribution in (12).

Combining these results with (13) deduce that $\mathbb{E}(\Sigma|\tilde{y}) \xrightarrow{\mathrm{p}} C_{[\alpha_0]}^{-1}$. □

Proof of Theorem 1. Using Lemma 1, deduce that

$$\begin{aligned} n^{1/2}\hat{\beta} &= n^{-1/2}\mathbb{E}(\Sigma|\tilde{y})X^T y \\ &= n^{-1/2}\mathbb{E}(\Sigma|\tilde{y})X^T(X\beta_0+\varepsilon) \\ &= n^{1/2}\mathbb{E}(\Sigma|\tilde{y})C_n\beta_0 + \left(C_{[\alpha_0]}^{-1}+o_p(1)\right)n^{-1/2}X^T\varepsilon. \end{aligned}$$

Invoking a standard triangular central limit theorem (Srivastava, 1967, Corollary B1) for $n^{-1/2}X^T\varepsilon$, deduce that

$$\left(C_{[\alpha_0]}^{-1}+o_p(1)\right)n^{-1/2}X^T\varepsilon \stackrel{\mathrm{d}}{\rightsquigarrow} C_{[\alpha_0]}^{-1}N(0,\sigma_0^2 C).$$

Therefore,

$$n^{1/2}\left(\hat{\beta}-\left[C_{[\alpha_0]}^{-1}+o_p(1)\right]C_n\beta_0\right) \stackrel{\mathrm{d}}{\rightsquigarrow} N(0,\sigma_0^2 C_{[\alpha_0]}^{-1}),$$

and (i) and (ii) now follow. □

References

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Brown, P.J., Vannucci, M., Fearn, T., 1998. Multivariate Bayesian variable selection and prediction. J. Roy. Statist. Soc. Ser. B 60, 627–641.
Chipman, H., 1996. Bayesian variable selection with related predictors. Canad. J. Statist. 24, 17–36.
Clyde, M., DeSimoned, H., Parmigiani, G., 1996. Prediction via orthogonalized model mixing. J. Amer. Statist. Assoc. 91, 1197–1208.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88, 881–889.
Geweke, J., Meese, R., 1981. Estimating regression models of finite but unknown order. Internat. Econom. Rev. 22, 55–70.
Ishwaran, H., Rao, J.S., 2003. Detecting differentially expressed genes in microarrays using Bayesian model selection. J. Amer. Statist. Assoc. 98 (462), 438–455.
Ishwaran, H., Rao, J.S., 2005a. Spike and slab variable selection: frequentist and Bayesian strategies. Ann. Statist. 33, 730–773.
Ishwaran, H., Rao, J.S., 2005b. Spike and slab gene selection for multigroup microarray data. J. Amer. Statist. Assoc. 100 (471), 764–780.
Kuo, L., Mallick, B.K., 1998. Variable selection for regression models. Sankhyā Ser. B 60, 65–81.
Lempers, F.B., 1971. Posterior Probabilities of Alternative Linear Models. Rotterdam University Press, Rotterdam.
Mitchell, T.J., Beauchamp, J.J., 1988. Bayesian variable selection in linear regression. J. Amer. Statist. Assoc. 83, 1023–1036.
Park, T., Casella, G., 2008. The Bayesian Lasso. J. Amer. Statist. Assoc. 103, 681–686.
Seo, D.M., Goldschmidt-Clermont, P.J., West, M., 2007. Of mice and men: sparse statistical modeling in cardiovascular genomics. Ann. Appl. Stat. 1, 152–178.
Srivastava, M.S., 1967. On fixed-width confidence bounds for regression parameters and mean vector. J. Roy. Statist. Soc. Ser. B 29, 132–140.