Bayesian Variable Selection Using an Adaptive Powered Correlation Prior

Arun Krishna, Howard D. Bondell and Sujit K. Ghosh

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, U.S.A.

Correspondence Author: Howard D. Bondell

email: bondell@stat.ncsu.edu

Telephone: (919)515-1914; Fax: (919)515-1169

Abstract

The problem of selecting the correct subset of predictors within a linear model has received much attention in recent literature. Within the Bayesian framework, a popular choice of prior has been Zellner's g-prior which is based on the inverse of empirical covariance matrix of the predictors. An extension of the Zellner's prior is proposed in this article which allow for a power parameter on the empirical covariance of the predictors. The power parameter helps control the degree to which correlated predictors are smoothed towards or away from one another. In addition, the empirical covariance of the predictors is used to obtain suitable priors over model space. In this manner, the power parameter also helps to determine whether models containing highly collinear predictors are preferred or avoided. The proposed power parameter can be chosen via an empirical Bayes method which leads to a data adaptive choice of prior. Simulation studies and a real data example are presented to show how the power parameter is well determined from the degree of cross-correlation within predictors. The proposed modification compares favorably to the standard use of Zellner's prior and an intrinsic prior in these examples.

Keywords: Bayesian variable selection; Collinearity; Powered Correlation Prior; Zellner's g-prior

1　Introduction

Consider the linear regression model with n independent observations and let $\boldsymbol{y} = (y_1, \cdots, y_n)'$ be the vector of response variables. The canonical linear model can be written as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p)$ is an $n \times p$ matrix of explanatory variables with $\boldsymbol{x}_j = (x_{1j}, \cdots, x_{nj})'$ for $j = 1, \cdots, p$. Let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ be the corresponding vector of unknown regression parameters, and $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I})$. Throughout the paper, we assume $\boldsymbol{y}$ to be empirically centered to have mean zero, while the columns of $\boldsymbol{X}$ have been standardized to have mean zero and norm one, so $\boldsymbol{X}'\boldsymbol{X}$ will be the empirical correlation matrix.

Under the above regression model, it is assumed that only an unknown subset of the coefficients are nonzero, so that the variable selection problem is to identify this unknown subset. Bayesian approaches to the problem of selecting variables/predictors within a linear regression framework has received considerable attention over the years, for example see, Mitchell and Beauchamp (1988), Geweke (1996), George and McCulloch (1993, 1997), Brown, Vannucci and Fearn (1998), George (2000) and Chipman, George and McCulloch (2001) and Casella and Moreno (2006).

For the linear model, Zellner (1986) suggested a particular form of a conjugate Normal-Gamma family called the g-prior which can be expressed as

$$
\begin{aligned}
\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} &\sim N(0, \frac{\sigma^2}{g}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\
\sigma^2 &\sim IG(a_0, b_0),
\end{aligned}
\tag{1.2}
$$

where $g > 0$ is a known scaling factor and $a_0 > 0$, $b_0 > 0$ are known parameters of the Inverse Gamma distribution with mean $\frac{a_0}{b_0 - 1}$. The prior covariance matrix of $\boldsymbol{\beta}$ is the scalar multiple σ^2/g of the inverse Fisher information matrix, which concurrently depends on the observed data through the design matrix $\boldsymbol{X}$.

This particular prior has been widely adopted in the context of Bayesian variable selection due to its closed form calculations of all marginal likelihoods which is suitable for rapid

computations over a large number of submodels, and its simple interpretation that it can be derived from the idea of a likelihood for a pseudo- data set with the same design matrix $\boldsymbol{X}$ as the observed sample (see, Zellner (1986), George and Foster (2000), Smith and Kohn (1996), Fernandez, Ley and Steel (2001)).

In this paper, we point out a drawback of using Zellner's prior on $\boldsymbol{\beta}$ particularly when the predictors $(\boldsymbol{x}_j)$ are highly correlated. The conditional variance of $\boldsymbol{\beta}$ given σ^2 and $\boldsymbol{X}$ is based on the inverse of the empirical correlation of predictors and puts most of its prior mass in the direction that causes the regression coefficients of correlated predictors to be smoothed away from each other. So when coupled with model selection, Zellner's prior discourages highly collinear predictors to enter the models simultaneously by inducing a negative correlation between the coefficients.

We propose a modification of Zellner's g-prior by replacing $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ by $(\boldsymbol{X}'\boldsymbol{X})^{\lambda}$ where the power $\lambda \in \mathbb{R}$, controls the amount of smoothing of collinear predictors towards or away from each other accordingly as $\lambda > 0$ or $\lambda < 0$, respectively. For $\lambda > 0$, the new conditional prior variance of $\boldsymbol{\beta}$ puts more prior mass in the direction that corresponds to a strong prior smoothing of regression coefficients of highly collinear predictors towards each other. Therefore, by choosing $\lambda > 0$ our proposed modification in contrast, forces highly collinear predictors entering or exiting the model simultaneously (see Section 2). Hence, the use of the power hyperparameter λ to the empirical correlation matrix helps us to determine whether models with high collinear predictors are preferred or not.

The hyperparameter λ is further incorporated into the prior probabilities over model space with the same intentions of encouraging or discouraging the inclusion of groups of correlated predictors. The choice of hyperparameter is obtained via an empirical Bayes approach and the inference regarding model selection is then made based on the posterior probabilities. By allowing the power parameter λ to be chosen by the data, we let the data decide whether to include collinear predictors or not.

The remainder of the paper is structured as follows. In Section 2, we describe in detail the Powered Correlation Prior and provide a simple motivating example, when $p = 2$. Section 3,

describes the choice of new prior specifications for model selection. The Bayesian hierarchical model and the calculation of posterior probabilities are presented in Section 4. The superior performance of using the Powered Correlation Prior over Zellner's g-priors is illustrated with the help of simulation studies and real data examples in Section 5. Finally, in Section 6 we conclude with a discussion.

2   The Adaptive Powered Correlated Prior

Consider again a normal regression model as in (1.1), where $\boldsymbol{X}'\boldsymbol{X}$ represents the correlation matrix. Let $\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{\Gamma}\boldsymbol{D}\boldsymbol{\Gamma}'$ be the spectral decomposition, where the columns of $\boldsymbol{\Gamma}$ are the p orthonormal eigenvectors and $\boldsymbol{D}$ is the diagonal matrix with eigenvalues $d_1 \geq \cdots \geq d_p \geq 0$ as the diagonal entries. The powered correlation prior for $\boldsymbol{\beta}$ conditioned on σ^2 and $\boldsymbol{X}$ is defined as

$$\boldsymbol{\beta}|\sigma^2, \boldsymbol{X} \sim N(0, \frac{\sigma^2}{g}(\boldsymbol{X}'\boldsymbol{X})^\lambda), \tag{2.1}$$

where $(\boldsymbol{X}'\boldsymbol{X})^\lambda = \boldsymbol{\Gamma}\boldsymbol{D}^\lambda\boldsymbol{\Gamma}'$, with $g > 0$ and $\lambda \in \mathbb{R}$ controlling the strength and the shape, respectively, of the prior covariance matrix, for a given $\sigma^2 > 0$.

There are several priors which are special cases of the powered correlation prior. For instance, $\lambda = -1$ produces the Zellner's g-prior (1.2). By setting $\lambda = 0$ we have $(\boldsymbol{X}'\boldsymbol{X})^0 = I$ which gives us the ridge regression model of Hoerl and Kennard (1970), under this model β_j are given independent $N(0, \sigma^2/g)$ priors. Next we illustrate how λ controls the model's response to collinearity which is the main motivation for using the powered correlation prior.

Let $\boldsymbol{T} = \boldsymbol{X}\boldsymbol{\Gamma}$ and $\boldsymbol{\theta} = \boldsymbol{\Gamma}'\boldsymbol{\beta}$. The linear model can be written in terms of the principal components as:

$$\boldsymbol{y} \sim N(\boldsymbol{T}\boldsymbol{\theta}, \sigma^2) \ \text{ with } \ \boldsymbol{\theta} \sim N(0, \frac{\sigma^2}{g}\boldsymbol{D}^\lambda). \tag{2.2}$$

The columns of the new design matrix $\boldsymbol{T}$ are the principal components, and so the original prior on $\boldsymbol{\beta}$ can be viewed as independent mean zero normal priors on the principal component regression coefficients, with prior variance proportional to the power of the corresponding eigenvalues, $d_1^\lambda, \cdots, d_p^\lambda$. Principal components with d_i near zero indicate a presence of a near-

linear relationship between the predictors, and the direction determined by the corresponding eigenvectors are those which are uninformative from the data. A classical frequentist approach to handle collinearity is to use principal component regression, and eliminate those dimensions with very small eigenvalues. Then transform back to the original scale, so that no predictors are actually removed. Along the same lines, we shall illustrate on how changing the value of λ would affect the prior correlation and demonstrate the intuition behind our proposed modification. For a simple illustration consider the case with $p = 2$ with a positive correlation ρ between them so that

$$(\boldsymbol{X}'\boldsymbol{X})^\lambda = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}^\lambda. \tag{2.3}$$

It easily follows that in this case,

$$\boldsymbol{\Gamma} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \text{ and, } \boldsymbol{D}^\lambda = \begin{bmatrix} (1+\rho)^\lambda & 0 \\ 0 & (1-\rho)^\lambda \end{bmatrix}. \tag{2.4}$$

The first principal component of our new design matrix $\boldsymbol{T}$ can be written as the sum of the predictors and the second as the difference

$$\boldsymbol{T} = \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{x}_1 + \boldsymbol{x}_2 \\ \boldsymbol{x}_1 - \boldsymbol{x}_2 \end{bmatrix}' \text{ with, } \boldsymbol{\theta} \sim N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \frac{\sigma^2}{g} \begin{bmatrix} (1+\rho)^\lambda & 0 \\ 0 & (1-\rho)^\lambda \end{bmatrix} \right), \tag{2.5}$$

for $\lambda > 0$ the prior on the coefficient for the sum has mean zero and variance $(1 + \rho)^\lambda$, while the prior on the coefficient for the difference has mean zero and variance $(1 - \rho)^\lambda$.

As ρ in (2.5) increases, a smaller prior variance is given to the coefficient for the difference of the two predictors, and hence introduces more shrinkage to the principal component directions that are associated with small eigenvalues. So that larger λ forces the difference to be more likely closer to the prior mean (zero). On the original $\boldsymbol{\beta}$ scale this corresponds to strong prior smoothing of regression parameters corresponding to highly collinear predictors.

On the other hand $\lambda < 0$ places a large prior variance on the coefficient for the difference, and a smaller variance on the coefficient of the sum, thereby shrinking those directions which correspond to large eigenvalues. This has an effect of smoothing the regression parameters

corresponding to highly collinear predictors away from each other, forcing the two predictors to be negatively correlated.

Hence in dimensions greater than two, in the presence of collinear predictors, λ has the flexibility to introduce more shrinkage in the directions that correspond to the small eigenvalues. This behavior motivates us to allow for the possibility of choosing alternative values for λ. In particular, we allow the data to determine the choice of λ using an empirical Bayes approach. We note that in the context of pricipal components regression, West (2003) allows for different prior variances on the principal component coefficients. However, our interest lies in the collinearity on the original scale.

3   Model Specification

The main focus of this paper is to use this powered correlation prior in a model selection problem. For the linear regression model in (1.1), it is typically the case that only an unknown subset of the coefficients β_j are non-zero, so in the context of variable selection we begin by indexing each candidate model with one binary vector $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_p)'$ where each element δ_j takes the value 1 or 0 depending on whether it is included or excluded from the model. More specifically, let

$$\delta_j = \begin{cases} 1 & \text{if } \boldsymbol{x}_j \text{ is included in the model,} \\ 0 & \text{if } \boldsymbol{x}_j \text{ is excluded from the model.} \end{cases} \tag{3.1}$$

We now rewrite the linear regression model, given $\boldsymbol{\delta}$ as

$$\boldsymbol{y} = \boldsymbol{X_\delta}\boldsymbol{\beta_\delta} + \boldsymbol{\epsilon}, \tag{3.2}$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 I)$, and $\boldsymbol{X_\delta}$ and $\boldsymbol{\beta_\delta}$ are the design matrix and the regression parameters of the model only including the predictors with $\delta_j = 1$. In the context of variable selection we can write the powered correlation prior as

$$\begin{aligned} \boldsymbol{\beta_\delta}|\boldsymbol{\delta}, \sigma^2, \boldsymbol{X} &\sim & N(0, \frac{\sigma^2}{g}(\boldsymbol{X'_\delta}\boldsymbol{X_\delta})^\lambda) \\ \text{with} \quad (\boldsymbol{X'_\delta}\boldsymbol{X_\delta})^\lambda &=& \boldsymbol{\Gamma_\delta} D_{\boldsymbol{\delta}}^\lambda \boldsymbol{\Gamma'_\delta}, \end{aligned}$$

$$\tag{3.3}$$

where $\boldsymbol{\Gamma}_{\boldsymbol{\delta}}$ is the matrix of eigenvectors and $\boldsymbol{D}_{\boldsymbol{\delta}}^{\lambda}$ is a diagonal matrix with diagonal entries as the eigenvalues of $(\boldsymbol{X}_{\boldsymbol{\delta}}'\boldsymbol{X}_{\boldsymbol{\delta}})^{\lambda}$.

Now that we have defined the prior for the coefficients given the model we now incorporate the same idea into the choice of prior for the inclusion indicators. With respect to Bayesian variable selection, a common prior for the inclusion indicators is, $p(\boldsymbol{\delta}) \propto \pi^{p_{\delta}}(1-\pi)^{p-p_{\delta}}$ (George and McCulloch, 1993,1997; George and Foster, 2000) where $p_{\boldsymbol{\delta}} = \sum_{j=1}^{p} \delta_j$ is the number of predictors in the model defined by $\boldsymbol{\delta}$, and π is the prior inclusion probability for each covariate. We can see this being equivalent to placing Bernoulli (π) priors on δ_j and thereby giving equal weight to any pair of equally-sized models. Setting $\pi = 1/2$ yields the popular uniform prior over model space formed by considering all subsets of predictors and, under this prior the posterior model probability is proportional to the marginal likelihood. A drawback of using this prior is that it puts most of its mass on models of size $\simeq p/2$ and it does not take into account the correlation between the predictors. Yuan and Lin (2005) proposed an alternative prior over model space.

$$P(\boldsymbol{\delta}|\pi) \propto \pi^{p_{\delta}}(1-\pi)^{p-p_{\delta}}|\boldsymbol{X}_{\boldsymbol{\delta}}'\boldsymbol{X}_{\boldsymbol{\delta}}|^{1/2}, \tag{3.4}$$

where $|.|$ denotes the determinant, and $|\boldsymbol{X}_{\boldsymbol{\delta}}'\boldsymbol{X}_{\boldsymbol{\delta}}| = 1$ if $p_{\boldsymbol{\delta}} = 0$. Since $|\boldsymbol{X}_{\boldsymbol{\delta}}'\boldsymbol{X}_{\boldsymbol{\delta}}|$ is small for models with highly collinear predictors, this prior discourages these models. We follow Yuan and Lin in that we use the information from the design matrix to build a prior for the model space. However, we do not necessarily want to penalize models with collinear predictors. We propose to incorporate the power parameter λ into a prior for $\boldsymbol{\delta}$ that could encourage or discourage inclusion of groups of correlated predictors. So we propose the following prior on model space:

$$P(\boldsymbol{\delta}|\lambda, \pi) \propto \pi^{p_{\delta}}(1-\pi)^{p-p_{\delta}}|\boldsymbol{X}_{\boldsymbol{\delta}}'\boldsymbol{X}_{\boldsymbol{\delta}}|^{-\lambda/2}. \tag{3.5}$$

So for large values of λ, the prior puts more of its mass on models with highly collinear predictors; while for $\lambda < 0$, penalizing models with collinear predictors. Hence coupled with the powered correlation prior, positive (negative) λ encourages (discourages) highly collinear predictors to enter the model simultaneously. Note that $\lambda = -1$ gives us Zellner's prior on

the coefficients coupled with the prior of Yuan and Lin on the models.

3.1 Choice of g

The parameter g defines the strength of the powered correlation prior. The choice of g is complicated in that large values of g will result in the prior dominating the likelihood, and small values of g would favor the null model (George and Foster, 2000). Various choices of g have been proposed over the years. For example, Smith and Kohn (1996) performed variable selection involving splines with a fixed value of $g = .01$. However the choice of g may also depend on the sample size n, or the number of predictors p. George and Foster (2000) propose an empirical Bayes method for estimating g from its marginal likelihood. Foster and George (1994) recommended using $g = 1/p^2$ based on a Risk Inflation Criterion (RIC). Kass and Wasserman (1995) suggests the unit information prior, where the amount of information in the prior corresponds to the amount of information in one observation, leading to $g = 1/n$. This leads to the Bayes factor as an approximation of the BIC. Fernandez *et. al.* (2001) suggest $g = 1/max(n, p^2)$ called the Benchmark Prior (BRIC), which is a combination of RIC and BIC. More recently Liang *et. al.*(2008) suggest a mixture of g-priors as an alternative to the default g-priors.

Since the scale of $(X'_{\delta}X_{\delta})^{\lambda}$ will depend on λ, we first standardize so that we may separate out the scale of g from that of $(X'_{\delta}X_{\delta})^{\lambda}$. To do so we modify (3.3) as

$$
\begin{aligned}
\beta_{\delta}|\delta, \sigma^2, X &\sim N(0, \frac{\sigma^2}{g}k(X'_{\delta}X_{\delta})^{\lambda}) \\
\text{with } k \equiv k(\lambda, \delta, X) &= \frac{Tr[(X'_{\delta}X_{\delta})^{-1}]}{Tr[(X'_{\delta}X_{\delta})^{\lambda}]}.
\end{aligned}
$$

(3.6)

This has an effect of setting the trace of $(X'_{\delta}X_{\delta})^{\lambda}$ to be equal to that of using $\lambda = -1$ regardless of the choice of λ. We then choose $g = 1/n$, as in the unit information prior (Kass and Wasserman, 1995).

For $(X'_{\delta}X_{\delta})^{\lambda} = \Gamma_{\delta}D_{\delta}^{\lambda}\Gamma_{\delta}$, k can be considered as the ratio of the average eigenvalues with those of $\lambda = -1$, $\frac{\sum_j D_{\delta_j}^{-1}}{n}/\frac{\sum_j D_{\delta_j}^{\lambda}}{n}$. Instead of the trace one could have opted to choose

8

the determinant, i.e. the ratio of the product of the eigenvalues. An advantage of using the average of the eigenvalues is that it provides more stability and in turn helps prevent the prior from dominating the likelihood. We note that other choices of standardization and choice of g are possible and are left for future investigations.

4   Model Selection using Posterior Probabilities

In the Bayesian framework, a set of prior distributions is specified on the parameters $\boldsymbol{\theta_\delta} = (\boldsymbol{\beta_\delta}, \sigma^2)$ for each model, along with a meaningful set of prior model probabilities $P(\boldsymbol{\delta}|\lambda, \pi)$ over the class of all models. Model selection is then done based on the posterior probabilities. Using the set of priors defined in the previous section, we can now construct a hierarchical Bayesian model to perform variable selection

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{X} &\sim N(\boldsymbol{X_\delta\beta_\delta}, \sigma^2 I) \\
\boldsymbol{\beta_\delta}|\boldsymbol{\delta}, \sigma^2, \boldsymbol{X} &\sim N(\boldsymbol{\beta_0}, \frac{\sigma^2}{g}k(\boldsymbol{X'_\delta X_\delta})^\lambda), \\
\sigma^2 &\sim IG(\frac{\gamma_o}{2}, \frac{\gamma_o}{2}), \\
P(\boldsymbol{\delta}|\lambda, \pi) &\propto \pi^{p_\delta}(1-\pi)^{p-p_\delta}|\boldsymbol{X'_\delta X_\delta}|^{-\lambda/2},
\end{aligned}
\tag{4.1}
$$

where k is as defined in (3.6) The key idea in computing the posterior model probabilities is to obtain the marginal likelihood of the data under model $\boldsymbol{\delta}$ by integrating out the model parameters

$$
P(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{X}) = \int P(\boldsymbol{y}|\boldsymbol{\theta_\delta}, \boldsymbol{\delta}, \boldsymbol{X})P(\boldsymbol{\theta_\delta}|\boldsymbol{\delta}, \boldsymbol{X})d\boldsymbol{\theta_\delta}.
\tag{4.2}
$$

The choice of conjugate priors allows us to analytically compute the above integral. Using the hierarchical model and integrating out $\boldsymbol{\theta_\delta}$ we obtain the conditional distribution of $\boldsymbol{y}$ given $\boldsymbol{\delta}$ and $\boldsymbol{X}$,

$$
\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{X} \sim t_{(\gamma_o+n)}\left\{\boldsymbol{X_\delta\beta_0}, (\boldsymbol{I_n} + \frac{k}{g}\boldsymbol{X_\delta}(\boldsymbol{X'_\delta X_\delta})^\lambda \boldsymbol{X'_\delta})\right\}.
\tag{4.3}
$$

Then model comparison is done via the posterior probabilities,

$$
P(\boldsymbol{\delta}|\boldsymbol{y}, \boldsymbol{X}) \propto P(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{X})P(\boldsymbol{\delta}|\lambda, \pi)
\tag{4.4}
$$

9

In order to fully specify our prior distribution we need to specify g, γ_o, π and λ. We choose $g = 1/n$ the unit information prior proposed by Kass and Wasserman (1995). For γ_o, after trying various choices, we saw that the model selected was not sensitive to the value of γ_o chosen, and since there is little or no information about this hyperparameter we decided to set γ_o to a constant, which has led to reasonable results as pointed out by George and McCulloch (1997). Following these lines, we set $\gamma_o = 0.01$ for the rest of the article, which corresponds to placing a non-informative prior on σ^2.

The parameters (λ, π) are very influential and informative with respect to the model selected and it is of utmost importance that we choose them carefully. Thus, we propose an empirical Bayes approach to select $\pi \in (0, 1)$ and $\lambda \in \mathbb{R}$ by marginalizing over $\boldsymbol{\delta}$ and maximizing the marginal likelihood function given by

$$m(\boldsymbol{y}|\boldsymbol{X}, \pi, \lambda) = \sum_{\boldsymbol{\delta}} P(\boldsymbol{y}|\boldsymbol{\delta}, \boldsymbol{X}) P(\boldsymbol{\delta}|\lambda, \pi). \tag{4.5}$$

When the number of predictors, p is of moderate size (e.g., $p \leq 20$) the above sum can be computed by evaluating (4.2) for each model via complete enumeration for a given (π, λ). Numerical optimization is then used to maximize $m(\boldsymbol{y}|\boldsymbol{X}, \pi, \lambda)$ defined in (4.5) to obtain the pair $(\hat{\pi}, \hat{\lambda})$. Specifically, we fix λ on a fine grid and for each λ, we maximize over $\pi \in (0, 1)$, and obtain $\hat{\pi}(\lambda)$ and obtain $\hat{\lambda} = \text{argmax } m(\boldsymbol{y}|\boldsymbol{X}, \hat{\pi}(\lambda), \lambda)$.

5   Simulations and Real Data

We shall now compare our proposed method to the standard use of Zellner's prior with a uniform prior over model space, i.e. the common approach

$$\begin{aligned}
\boldsymbol{\beta_\delta}|\sigma^2, \boldsymbol{\delta}, \boldsymbol{X} &\sim N(0, \sigma^2(\boldsymbol{X}'_{\boldsymbol{\delta}}\boldsymbol{X_\delta})^{-1}/g) \quad \text{where,} \quad g = 1/n, \\
\sigma^2 &\sim IG(\frac{\gamma_o}{2}, \frac{\gamma_o}{2}), \quad \gamma_o = .01, \\
P(\boldsymbol{\delta}) &= (1/2)^p.
\end{aligned}$$

$$\tag{5.1}$$

We also compared our method to the fully automatic Bayesian variable selection procedure proposed by Casella and Moreno (2006) where posterior probabilities are computed using intrinsic priors (Berger and Pericchi, 1996) which eliminates the need for tuning parameters. However, we note that this procedure was not specifically designed to handle correlated predictors.

In this section we evaluate the performance of using our proposed method in selecting the correct subset of predictors as compared to the two above mentioned methods, based on a simulated data involving highly collinear predictors. Comparisons are also presented for one real dataset.

5.1 Simulation Study

For the simulated example, we consider the true model

$$y = x_1 + x_2 + \epsilon \qquad \text{where } \epsilon \sim N(0, 1). \tag{5.2}$$

We generate p predictors from a multivariate normal with $\text{Cov}(\boldsymbol{x}_j, \boldsymbol{x}_k) = \rho^{|j-k|}$, for $\rho = 0.9$. For Case one, we fixed $p = 4$, while for Case two we used $p = 12$, so that in the first case there were 2 unimportant predictors, while in case two, there were 10. For both cases, we generated 1000 datasets each with $n = 30$ observations.

5.1.1 Case 1: p=4

Using the empirical Bayes approach mentioned earlier we compute the optimal pair $\hat{\lambda} = 1.6$ and $\hat{\pi} = .15$ which maximizes the marginal likelihood obtained under complete enumeration of all possible $2^4 - 1$ models. The estimates $(\hat{\lambda}, \hat{\pi})$ are the values obtained after averaging over the 1000 replications (Figure 1).

Figure (1) goes here.

From Table 1, the performance of our proposed method appears quite good compared to the other two methods. We see that Zellner's as well as the intrinsic prior's chooses single variable models with over half of its posterior probability. In contrast, using the powered

11

correlation prior smoothes the regression parameters of the correlated variables towards each other, by giving more prior information in the direction that are less determined by the data, and selects the correct model, $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with an overwhelming 0.622 posterior probability.

Table 1 also lists the number of times (in %) each model was selected as the model with highest posterior probability out of 1000 replications by the three methods. We see that the Powered Correlation Prior based method picks the correct model, $(\boldsymbol{x}_1, \boldsymbol{x}_2)$, 68.7 % of the time.

Table (1) goes here.

5.1.2   Case 2: p=12

Similar to the previous case, the optimal values $(\hat{\lambda} = 1.7, \hat{\pi} = 0.12)$ were obtained by averaging over 1000 replications (Figure 2) which maximizes the marginal likelihood function. The performance of the powered correlation priors in terms of selecting correlated predictors is also similar to the previous case.

Figure (2) goes here.

From Table 2 it is clear that Zellner's prior penalizes models with high collinearity, thereby putting more posterior mass on single variable models. In contrast, the powered correlation prior method favors the true model $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with maximum average posterior probability of 0.145 and the correct model was selected 46.1 % of the time. For the intrinsic prior, the maximum posterior model is the model including only $\boldsymbol{x}_1$ and the correct model is selected only 9.8% of the time.

Table (2) goes here.

In this simulation study the true model contains two highly correlated predictors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Given that we know the true data-generating mechanism, an objective comparison criterion is the ability of the methods to correctly identify the underlying true model. We see from Table 1 and 2 that the posterior probability for the true model using our proposed method is much larger than for any other model. Hence our method is able to correctly

identify the true set of predictors even when they are highly correlated. Alternatively, the two other approaches choose a single predictor model and give very little posterior probability to the true model. Model performance could be further evaluated based on prediction accuracy, but we have not explored that aspect in this simulation study, as the main goal was to examine whether the methods identify the entire correct set of variables that contribute to the explanation of the response y.

5.2 Real Data Example

We consider a real dataset to demonstrate the performance of our method. For our real data example we use the data on NCAA graduation rates (Mangold, Bean and Adams, 2003) where there are 97 observations and 19 predictors. The response variable is the average graduation rates for each of the 97 colleges (see Appendix for a description of the dataset). Mangold, Bean and Adams used this dataset with the goal of showing that successful sports programs raise graduation rates. This dataset is of specific interest to us, due to the presence of high correlation among the variables. We fit a main effects only model with the 19 possible predictors.

For this dataset we obtain the optimal values of $\hat{\lambda} = 1.9$ and $\hat{\pi} = 0.19$ (Figure 3). Posterior probabilities are computed using these optimal values by complete enumeration of all $2^{19} - 1$ possible models.

Figure (3) goes here.

In Table 3 we compare the posterior model probabilities by using our proposed method to those obtained using the standard Zellners g-prior and the fully automatic intrinsic priors. The highest posterior model selected using the powered correlation prior to predict the average graduation rates is a 6 variable model, as compared to Zellners which selects a 5 variable model by dropping x_{17} (Acceptance Rate) from the model chosen by our approach. This could be attributed to the high correlation between x_2 and x_{17} ($\rho = .81$). The intrinsic prior approach picks out a simpler (fewer predictors) model as its highest posterior probability model.

Table (3) goes here.

13

Model comparison and validation are now made based on the average predicted error, where the parameter estimates are obtained by computing the posterior mean of $\boldsymbol{\beta_\delta}$ for each given configuration of $\boldsymbol{\delta}$ and $\boldsymbol{y}$ for each of the three methods. Table 3 reports the average mean square predictive error along with their standard errors obtained using 5-fold crossvalidation (CV), whose estimates are first averaged across 10 cross-validation splits to reduce variability, and then replicated 1000 times. We see that the top two models picked out by the powered correlation prior's posterior probabilities has a significantly lower prediction error than that of the models selected using the two other methods. Hence, both the simulation and the real data example show strong support for the use of our proposed powered correlation prior.

6 Discussion

In this paper we have demonstrated that within a linear model framework the powered correlation prior helps to resolve the problem of selecting subsets using a suitable modification of Zellners g-prior when the predictors are highly correlated. By using simulated and the real data examples we have illustrated that the powered correlation prior tends to perform better in terms of choosing the correct model than the standard Zellners prior and the intrinsic prior for correlated predictors. The choice of hyperparameter λ obtained using a empirical Bayes method controls the degree of smoothing of correlated predictors towards or away from each other.

For a large number of predictors (e.g. $p > 30$), a attractive feature of this prior is that all the parameters can be integrated out analytically to obtain a closed form for the unnormalized posterior model probabilities. Hence a simple Gibbs sampler over model space (George and McCulloch, 1997) can be implemented to approximate the marginal for each pair (λ, π). This can be implemented on a two dimensional grid, and although it may take significant computation time, it remains feasible.

Model averaging for linear regression models has received considerable attention (Raftery, Madigan and Hoeting, 1997). This method accounts for model uncertainty by averaging over all possible models. It is possible to extend the use of our proposed prior to perform model

averaging via the use of posterior probabilities.

There has also been considerable interest in Bayesian variable selection for generalized linear models. The selection criteria are based on extensions of Bayesian methods used in linear regression framework. While beyond the scope of this paper, one can extend the powered correlation prior used here to generalized linear models.

ACKNOWLEDGEMENTS

Appendix

Brief Description of NCAA Data

Data from Mangold, Bean, Adams (2003), Journal Of Higher Education, p. 540-562, "The Impact of Intercollegiate Athletics on Graduation Rates Among Major NCAA Division I Universities." The data were taken from the 1996-99 editions of the US News "Best Colleges in America" and from the US Department of Education data and includes 97 NCAA Division 1A schools. The authors hoped to show that successful sports programs raise graduation rates. Here is a list describing briefly the response variable and 19 predictors.

Y	Average 6 yr graduation rate for 1996, 1997, 1998
$x1$	% Students in top 10 Percent HS
$x2$	ACT COMPOSITE 25TH
$x3$	% On living campus
$x4$	% First-time undergraduates
$x5$	Total Enrollment/1000
$x6$	% Courses taught by TAs
$x7$	Composite of basketball ranking
$x8$	In-state tuition/1000
$x9$	Room and board/1000
$x10$	Avg BB home attendance
$x11$	Full Professor Salary
$x12$	Student to faculty ratio
$x13$	% White
$x14$	Assistant professor salary
$x15$	Population of city where located
$x16$	% Faculty with PHD
$x17$	Acceptance rate
$x18$	% Receiving loans
$x19$	% Out of state

References

Berger, J.O. and Pericchi, L. R. (1996) The Intrinsic Bayes Factor for Model Selection and Prediction. *J. Amer. Statist. Assoc.*, **91**, 109-122.

Brown, P. J., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *J. Roy. Statist. Soc. Ser. B* **60**, 627-642.

Casella, G. and Moreno, E. (2006) Objective Bayesian variable selection. *J. Amer. Statist. Assoc.*, **101**, 157-167.

Chipman, H., George, E. I. and McCulloch, R. E. (2001) The practical implementation of Bayesian model selection, vol. 38 of *IMS Lecture Notes - Monograph Series*

Fernandez, C., Ley, E. and Steel, M. F. (2001) Benchmark priors for Bayesian model averaging. *J. Econometrics*, **100**, 381-427.

Foster, D. P. and George, E. I. (1994) The risk inflation criterion for multiple regression. *Ann. Statist.*, **22**, 1947-1975.

Hoerl, A. E. and Kennard, R. (1970) Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

George, E. I(2000) The variable selection problem. *J. Amer. Statist. Assoc.*, **95**, 1304-1308.

George, E. I. and Foster, D. P. (2000) Calibration and empirical Bayes variable selection. *Biometrika*, **87**, 731-747.

George, E, I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc.*, **7**, 881-889.

George, E, I. and McCulloch, R. E. (1997) Approaches for Bayesian variable selection. *Statist. Sinica*, **7**, 339-374.

Geweke, J. (1996) Variable selection and model comparison in regression. In *Bayesian Statistics 5: Proceedings of the Fifth Valencia International Meeting*, (eds. J. M. Bernardo, J. O. Berger, A. P. David and A. F. M. Smith), pp. 609-620. Oxford Univ. Press.

Kass, R. E. and Wasserman, L. (1995) A reference Bayesian test for nested hypothesis and its relationship to the schwarz criterion. *J. Amer. Statist. Assoc.*, **90**, 928-934.

Liang, F., Paulo, R., Molina, G., Clyde, M., A. and Berger, J., O. (2008) Mixtures of g-priors

for Bayesian Variable Selection. *J. Amer. Statist. Assoc.*, **103**, 410-423.

Mangold, W. D., Bean, L. and Adams, D. (2003) The Impact of Intercollegiate Athletics on Graduation Rates among Major NCAA Division I Universities: Implications for College Persistence Theory and Practice, *Journal Of Higher Education*, pp. 540-562.

Mitchell, T. J. and Beauchamp, J. J. (1988) Bayesian variable selection in linear regression (with discussion). *J. Amer. Statist. Assoc.*, **83**, 1023-1032.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) Bayesian model averaging for linear regression models. *J. Amer. Statist. Assoc.*, **92**, 179-191.

Smith, M. and Kohn, R. (1996) Nonparametric regression using Bayesian variable selection. *J. Econometrics*, **75**, 314-343.

West, M. (2003) Bayesian factor regression models in the large p, small n paradigm. In *Bayesian Statistics 7* (eds: Bernardo et al.). Oxford University Press.

Yuan, M. and Lin, Y. (2005) Efficient empirical Bayes variable selection and estimation in linear models, *J. Amer. Statist. Assoc.*, **100**, 1215-1224.

Zellner, A. (1986) On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, (eds. P. K. Goel and A. Zellner), pp. 233-243. North-Holland/Elsevier.
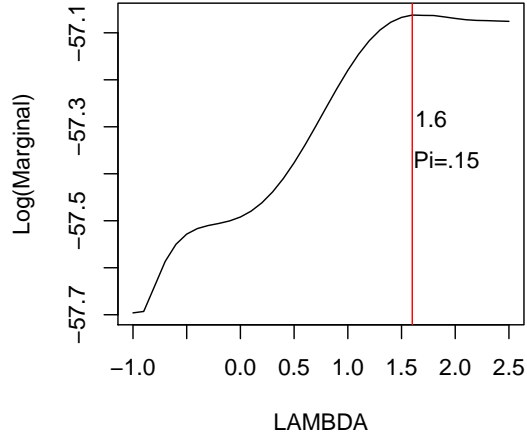
Figure 1: Plot of λ vs. Log$[m(\boldsymbol{y}|\boldsymbol{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to case 1: $p = 4$. Averaged over 1,000 simulations. The vertical line represents the location of the global maximum.

Figure 2: Plot of λ vs. Log $[m(\boldsymbol{y}|\boldsymbol{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to case 2: $p = 12$, averaged over 1,000 simulations. The vertical line represents the location of the global maximum.
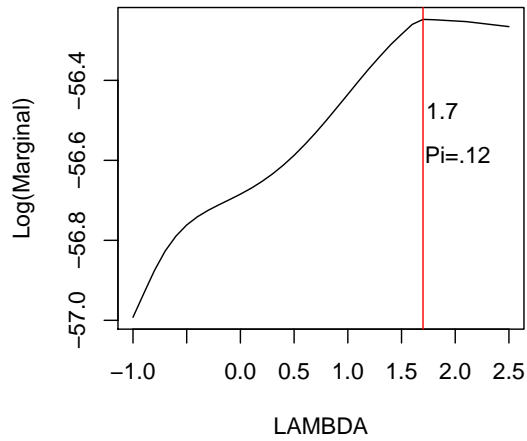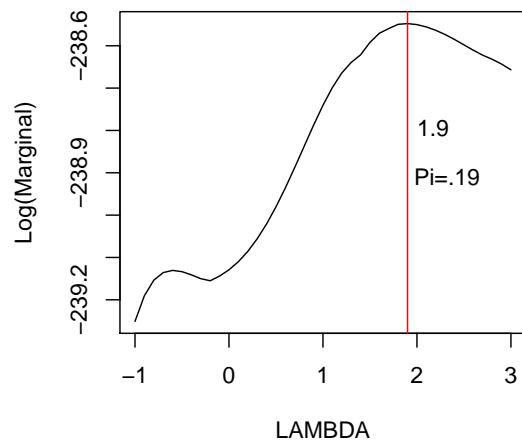
19

Figure 3: Plot of λ vs. $\text{Log}[m(\boldsymbol{y}|\boldsymbol{X}, \pi, \lambda)]$, maximized over $\pi \in (0, 1)$, corresponding to the NCAA Dataset. The vertical line represents the location of the global maximum.

Zellner's			PoCor			Intrinsic Prior		
Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected
x_2	.321	37.6	x_1, x_2	.622	68.7	x_1	.468	51.9
x_1	.221	20.4	x_2	.116	10.4	x_2	.410	44.2
x_1, x_2	.105	11.2	x_1, x_2, x_3	.089	3.4	x_1, x_2	.039	.62
x_1, x_2, x_4	.074	5.9	x_1, x_2, x_4	.040	3.1	x_1, x_3	.020	.36
x_1, x_2, x_3	.055	4.4	x_1, x_2, x_3, x_4	.039	2.5	x_3	.020	.24
x_2, x_4	.052	2.8	x_1	.030	2.3	x_1, x_4	.017	1.07
x_2, x_3	.045	2.6	x_1, x_3	.009	1.4	x_2, x_3	.010	.26
x_1, x_3	.035	2.1	x_2, x_3	.003	1.0	x_2, x_4	.008	.20
x_1, x_3, x_4	.024	1.4	x_2, x_3, x_4	.002	.9	x_4	.002	.12
x_1, x_2, x_3, x_4	.019	1.1	x_1, x_3, x_4	.002	.6	x_1, x_2, x_4	.003	.08

Table 1: For Case 1, Comparing Average Posterior Probabilities, corresponding to the case 1: $p = 4$, averaged across 1,000 simulations. % Selected is the number of times (in %) each model was selected as the highest posterior model out of 1000 replications. Zellners represents use of Zellners prior as in (5.1). PoCor represents our proposed modification as in (4.1). Intrinsic Prior represents the fully automatic procedure proposed by Casella and Moreno (2006)

Zellner's			PoCor			Intrinsic Prior		
Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected	Subset	Avg Post Prob	%Selected
x_1	.068	28.9	x_1, x_2	.145	46.1	x_1	.039	33.2
x_1, x_2	.051	10.6	x_1, x_2, x_3	.109	12.4	x_2	.016	14.5
x_1, x_3	.025	10	x_1, x_3	.090	7.8	x_1, x_2	.013	9.8
x_1, x_4	.016	5.9	x_2, x_3	.078	5.5	x_1, x_2, x_3	.011	5.4
x_1, x_5	.0133	5.4	x_1	.056	5.1	x_1, x_3	.009	2.3
x_1, x_2, x_5	.013	4.4	x_1, x_2, x_4	.020	4.7	x_1, x_2, x_4	.008	.89
x_1, x_2, x_6	.012	3.5	x_1, x_2, x_3, x_4	.016	3.6	x_1, x_3, x_4	.008	.76
x_2, x_2, x_8	.011	3.5	x_1, x_2, x_5	.010	3.4	x_1, x_2, x_5	.007	.54

Table 2: Case 2, Comparing Average Posterior Probabilities, corresponding to the case 2: $p = 12$, averaged over 1,000 Simulations. % Selected is the number of times (in %) each model was selected as the highest posterior model out of 1000 replications. Zellners represents use of Zellners prior as in (5.1). PoCor represents our proposed modification as in (4.1). Intrinsic Prior represents the fully automatic procedure proposed by Casella and Moreno (2006)

Zellner's			PoCor			Intrinsic Prior		
Subset	Post Prob	C.V. Pred Err	Subset	Post Prob	C.V. Pred Err	Subset	Post Prob	C.V. Pred Err
x_2, x_3, x_4, x_5, x_7	.042	54.38 (0.615)	$x_2, x_3, x_4, x_5, x_7, x_{17}$	.036	51.53 (0.561)	x_2, x_4, x_7	.066	53.97 (0.530)
x_2, x_3, x_4, x_7	.041	55.57 (0.646)	$x_2, x_3, x_4, x_5, x_7, x_{17}, x_{18}$	.030	52.38 (0.619)	x_2, x_3, x_4, x_5, x_7	.040	52.94 (0.541)
x_2, x_3, x_4, x_5	.028	56.74 (0.599)	$x_1, x_2, x_3, x_4, x_5, x_7$	.028	53.46 (0.608)	x_2, x_3, x_4, x_5	.028	54.09 (0.602)
$x_2, x_3, x_4, x_5, x_7, x_{18}$	.017	55.17 (0.609)	$x_2, x_3, x_4, x_5, x_7, x_{18}$	.021	54.08 (0.572)	x_2, x_4, x_5, x_7	.022	54.82 (0.576)
$x_2, x_3, x_4, x_5, x_7, x_{17}$	.015	54.89 (0.623)	x_2, x_3, x_4, x_5, x_7	.018	54.51 (0.568)	x_2, x_3, x_4	.016	58.71 (0.617)
$x_1, x_2, x_3, x_4, x_5, x_7$	.013	55.64 (0.611)	x_1, x_2, x_3, x_4, x_5	.015	56.13 (0.601)	x_2, x_4, x_7, x_{11}	.011	58.67 (0.594)
$x_2, x_3, x_4, x_7, x_8, x_9$	.011	56.54 (0.628)	$x_2, x_3, x_4, x_5, x_7, x_{10}$	.009	54.75 (0.554)	x_2, x_4, x_{11}	.010	60.08 (0.581)

Table 3: Comparing Posterior Probabilities and average prediction errors for the models of the NCAA Data. The entries in parenthesis are the standard errors obtained by 1000 replications.