Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://amstat.tandfonline.com/loi/uasa20

Rao-Blackwellization for Bayesian Variable Selection and Model Averaging in Linear and Binary Regression: A Novel Data Augmentation Approach

Joyee Ghosh[a] & Merlise A. Clyde[a]

[a] Joyee Ghosh is Assistant Professor, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242-1409. Merlise A. Clyde is Professor, Department of Statistical Science, Duke University, Durham, NC 27705. This research was supported by National Institutes of Health grants NIH/NIEHS 5T32ES007018, NIH/NIEHS P30 ES10126, 1-RC1HL099863-01, and NIH R01-HL090559-01 and National Science Foundation grants DMS-0406115 and DMS-0422400. The authors thank Steve MacEachern and Scott Schmidler for fruitful discussions. The authors thank the editor, the associate editor, and two referees for helpful comments.
Published online: 24 Jan 2012.

To cite this article: Joyee Ghosh & Merlise A. Clyde (2011) Rao-Blackwellization for Bayesian Variable Selection and Model Averaging in Linear and Binary Regression: A Novel Data Augmentation Approach, Journal of the American Statistical Association, 106:495, 1041-1052, DOI: 10.1198/jasa.2011.tm10518

To link to this article: http://dx.doi.org/10.1198/jasa.2011.tm10518

Supplementary materials for this article are available online. Please click the JASA link at http://pubs.amstat.org.

Rao–Blackwellization for Bayesian Variable Selection and Model Averaging in Linear and Binary Regression: A Novel Data Augmentation Approach

Joyee GHOSH and Merlise A. CLYDE

Choosing the subset of covariates to use in regression or generalized linear models is a ubiquitous problem. The Bayesian paradigm addresses the problem of model uncertainty by considering models corresponding to all possible subsets of the covariates, where the posterior distribution over models is used to select models or combine them via Bayesian model averaging (BMA). Although conceptually straightforward, BMA is often difficult to implement in practice, since either the number of covariates is too large for enumeration of all subsets, calculations cannot be done analytically, or both. For orthogonal designs with the appropriate choice of prior, the posterior probability of any model can be calculated without having to enumerate the entire model space and scales linearly with the number of predictors, p. In this article we extend this idea to a much broader class of nonorthogonal design matrices. We propose a novel method which augments the observed nonorthogonal design by at most p new rows to obtain a design matrix with orthogonal columns and generate the "missing" response variables in a data augmentation algorithm. We show that our data augmentation approach keeps the original posterior distribution of interest unaltered, and develop methods to construct Rao–Blackwellized estimates of several quantities of interest, including posterior model probabilities of any model, which may not be available from an ordinary Gibbs sampler. Our method can be used for BMA in linear regression and binary regression with nonorthogonal design matrices in conjunction with independent "spike and slab" priors with a continuous prior component that is a Cauchy or other heavy tailed distribution that may be represented as a scale mixture of normals. We provide simulated and real examples to illustrate the methodology. Supplemental materials for the manuscript are available online.

KEY WORDS: Gibbs sampling; MCMC; Missing data; Model probability; Model uncertainty; Orthogonal design; Posterior probability.

1. INTRODUCTION

Linear models remain among the most popular methods for modeling relationships between a Gaussian response variable and a set of explanatory variables and provide a foundation for many nonparametric regression methods with the appropriate selection of basis vectors, such as splines, wavelets, and kernel regression (Clyde and George 2004). The generalization of the Gaussian linear model to other exponential families leads to the familiar generalized linear models, for example, probit or logistic regression for binary outcomes. For many applications, it is routine to collect data on many variables of interest and entertain a wide variety of possible models. Computational advances have facilitated the Bayesian treatment of such problems, where the Bayesian approach to model uncertainty proceeds by treating the model as an unknown parameter, and assigns prior probabilities to all possible models under consideration. Given the observed data, these probabilities are updated via Bayes theorem to obtain posterior probabilities of models, which may be used in the selection of a model via decision theoretic approaches taking into consideration other costs (Fouskakis, Ntzoufras, and Draper 2009) or in Bayesian model averaging (BMA) using the full joint posterior distribution (see Draper 1995, Hoeting et al. 1999, or Clyde and George 2004 for reviews and additional references).

While it is straightforward to formulate the model uncertainty problem under the Bayesian paradigm, its implementation often becomes nontrivial in large problems, particularly with highly correlated predictors. Heaton and Scott (2010) compare several recent stochastic search algorithms and note that traditional Markov chain Monte Carlo (MCMC) methods for sampling from the posterior distribution of models often fail to reach regions of the model space with high posterior probability. For identifying high probability models, they argue in favor of using alternative stochastic search algorithms (Berger and Molina 2005; Scott and Carvalho 2008; Clyde, Ghosh, and Littman 2011) that rely on having analytic expressions for marginal likelihoods, such as with Zellner's g-prior (Zellner 1986) or mixtures of g-priors (Zellner and Siow 1980; Liang et al. 2008); the renormalized likelihoods are used to estimate posterior model probabilities and marginal inclusion probabilities in lieu of (noisy) Monte Carlo frequencies. While these methods are better at identifying high probability models, both Heaton and Scott (2010) and Clyde, Ghosh, and Littman (2011) note substantial disagreement in estimates such as posterior inclusion probabilities between MCMC methods and these alternative stochastic search methods in high-dimensional problems. Clyde and Ghosh (2010) prove that estimators based on renormalized likelihoods lead to biased estimates of inclusion probabilities and other quantities under BMA, which in turn may result in higher mean squared errors. Alternative choices are independent proper "spike and slab" variable selection priors (Ishwaran and Rao 2005) which lead to generalized ridge regression estimates. Despite progress over the last decade, there is clearly still a need for improved algorithms and estimators in the variable selection/model averaging problem.

Joyee Ghosh is Assistant Professor, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242-1409 (E-mail: *joyee-ghosh@uiowa.edu*). Merlise A. Clyde is Professor, Department of Statistical Science, Duke University, Durham, NC 27705 (E-mail: *clyde@stat.duke.edu*).
This research was supported by National Institutes of Health grants NIH/NIEHS 5T32ES007018, NIH/NIEHS P30 ES10126, 1-RC1HL099863-01, and NIH R01-HL090559-01 and National Science Foundation grants DMS-0406115 and DMS-0422400. The authors thank Steve MacEachern and Scott Schmidler for fruitful discussions. The authors thank the editor, the associate editor, and two referees for helpful comments.

Journal of the American Statistical Association
September 2011, Vol. 106, No. 495, Theory and Methods
DOI: 10.1198/jasa.2011.tm10518

In the case of linear regression with designs matrices with orthogonal columns and independent normal priors on regression coefficients, one can obtain closed form expressions for posterior probabilities of models and marginal posterior inclusion probabilities (conditional on σ^2), without explicit enumeration of the entire model space of 2^p models. Such "orthogonal designs" arise naturally in the context of designed experiments or wavelet regression, where $p = n$. The computational advantages from orthogonality have been exploited by Chipman, Kolaczyk, and McCulloch (1997), Clyde, Parmigiani, and Vidakovic (1998), Clyde and George (1999, 2000), Johnstone and Silverman (2005), among others, to provide computationally efficient estimators in nonenumerable model spaces that provide adaptive shrinkage with outstanding frequentist properties. For general design matrices, Clyde, DeSimone, and Parmigiani (1996), Clyde and Parmigiani (1996) transform the nonorthogonal design to an orthogonal design and developed efficient algorithms for Bayesian model averaging (BMA) based on importance sampling. Their approach led to better performance in terms of prediction in many cases, but does not readily lend itself for variable selection as models are defined in terms of combinations of the original predictor variables. In this article, we propose a novel method that augments a nonorthogonal design with additional rows, such that the columns of the resulting "complete" design matrix are mutually orthogonal. Unlike Clyde, DeSimone, and Parmigiani (1996) the Orthogonal Data Augmentation algorithm presented in Section 2 permits both model averaging and model selection in terms of the original variables. The response variables corresponding to the newly introduced rows are treated as missing data and are sampled using MCMC algorithms. Exploiting properties of orthogonal designs, in Section 2.2 we construct "Rao–Blackwellized" (RB) estimates of posterior model probabilities by marginalizing over the missing responses and prove that the RB estimates have smaller variances than the ergodic averages. A key feature of the ODA formulation is the use of Rao–Blackwellization to provide a simple method for estimating the mass of unsampled models. In Section 3 we discuss how to select the augmented design and its properties. We compare the ODA algorithm and RB estimator to other methods in Section 4 using the simulation design of Nott and Kohn (2005), for which $p = 15$, so that enumeration is feasible and estimates can be compared to the truth. We show that our estimates of model probabilities can outperform Monte Carlo estimates and provide accurate estimates of the unsampled mass. The ODA formulation allows extensions to heavier tailed prior distributions on regression coefficients constructed as scale mixtures of normals. In Section 5, we show how to implement ODA with independent Cauchy priors and compare the results to the multivariate Zellner–Siow Cauchy prior (Zellner and Siow 1980). In Section 6 we apply our methods to the ozone data analyzed by Friedman and Silverman (1989) and more recently by Liang et al. (2008), with a model space of dimension 2^{44} that prohibits enumeration. Empirically, we show that ODA does as well or better than BMA with the Zellner g prior, Zellner–Siow Cauchy prior or shrinkage methods using the lasso (Tibshirani 1996) or horseshoe prior (Carvalho, Polson, and Scott 2010). Finally, using the latent variable formulations, we may extend the class of models to binary regression. We illustrate the ODA method for probit regression in Section 7 using the well-known Pima Indian diabetes dataset. In Section 8 we conclude with possible extensions of our method.

2. ORTHOGONAL DATA AUGMENTATION

The basis of the popular Expectation–Maximization (EM) algorithm of Dempster, Laird, and Rubin (1977) or the Data Augmentation (DA) algorithm of Tanner and Wong (1987) (and their many extensions) is to augment the observed data with missing data, so that the resulting complete data model is much simpler. Letting $\mathbf{Y_o} \equiv (Y_1, Y_2, \ldots, Y_{n_o})^T$ denote the observed response vector of length n_o, we may augment it with a vector of length n_a of unknown missing response variables $\mathbf{Y_a}$, so that the complete response data is denoted as $\mathbf{Y_c}^T = (\mathbf{Y_o}^T, \mathbf{Y_a}^T)$, with length $n_c = n_o + n_a$. In the regression context, we must also specify a design matrix for the missing $\mathbf{Y_a}$. For fractional factorial experiments, there is a natural choice for this design matrix based on the remaining fraction from the full factorial experiment. We propose a method that can be used to construct a design matrix for the more general observational setting.

To begin, let $\mathbf{X_o} = [\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_p]$ denote the $n_o \times (p+1)$ observed design where $\mathbf{x}_0$ is an $n \times 1$ vector of ones representing an intercept. Without loss of generality we assume that the remaining p predictor variables $\mathbf{x}_j$ for $j > 0$ have been centered so that they are orthogonal to the intercept ($\mathbf{x}_j^T \mathbf{x}_0 = 0$) and have been rescaled by dividing by their standard deviation (using n_o, rather than $n_o - 1$ in the denominator). This standardization leads to the diagonal elements of $\mathbf{X_o}^T\mathbf{X_o}$ being the sample size n_o and the off-diagonal elements of $\mathbf{X_o}^T\mathbf{X_o}$ being n_o times the correlation of the columns of $\mathbf{X_o}$. Such standardizations are commonly used in ridge regression or other shrinkage methods such as the lasso so that coefficients may be interpreted on the same scale. The normal linear model using the full design matrix may be written as

$$\mathbf{Y_o} = \mathbf{X_o}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a vector of regression coefficients, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_{n_o})$, $\mathbf{I}_{n_o}$ is the identity matrix of dimension n_o and σ^2 is the variance.

For the variable selection problem, models may be represented by a binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T \in \{0,1\}^p \equiv \boldsymbol{\Gamma}$, where $\gamma_j = 0$ implies that variable $\mathbf{x}_j$ has been removed from the model or equivalently that $\gamma_j = 0 \Leftrightarrow \beta_j = 0$, so that under model $\boldsymbol{\gamma}$

$$\mathbf{Y_o}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma} \sim \mathrm{N}\big(\mathbf{X_{o\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2\mathbf{I}_{n_o}\big), \tag{2}$$

where $\mathbf{X_{o\gamma}}$ is the $n \times (p_{\boldsymbol{\gamma}} + 1)$ design matrix and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the $p_{\boldsymbol{\gamma}} + 1$ dimensional vector of nonzero model specific regression coefficients; by default, the intercept is always assumed to be included in $\mathbf{X_{o\gamma}}$. To transform the problem of model search with a general design matrix to one with an orthogonal design, we augment the rows of the observed design matrix $\mathbf{X_o}$ with a $n_a \times (p+1)$ design matrix $\mathbf{X_a}$, such that the resulting "complete" design matrix

$$\mathbf{X_c} = \begin{bmatrix} \mathbf{X_o} \\ \mathbf{X_a} \end{bmatrix} \tag{3}$$

has orthogonal columns, $\mathbf{X_c}^T\mathbf{X_c} = \mathbf{X_o}^T\mathbf{X_o} + \mathbf{X_a}^T\mathbf{X_a} = \mathbf{D}$, where $\mathbf{D} = \mathrm{diag}(\delta_0, \delta_1, \ldots, \delta_p)$ is a diagonal matrix with $\delta_j > 0$. For now, we assume that such a matrix will exist and defer discussion of how to select $\mathbf{X_a}$ until Section 3. Under model $\boldsymbol{\gamma}$, given model specific parameters $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2)$ and the design $\mathbf{X_a}$, $\mathbf{Y_a}$ has

the same distribution as the observed data $\mathbf{Y_o}$ and is independent of $\mathbf{Y_o}$, leading to the complete data model

$$\mathbf{Y_c}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}, \sim \mathrm{N}\big(\mathbf{X_{c\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2\mathbf{I}_{n_c}\big). \tag{4}$$

We begin with limiting conjugate prior distributions for β_0 and σ^2 and conjugate prior distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$

$$p(\beta_0) \propto 1, \tag{5a}$$

$$\beta_j|\sigma^2, \lambda_j, \gamma_j \overset{\text{ind}}{\sim} \mathrm{N}\left(0, \sigma^2\frac{\gamma_j}{\lambda_j}\right) \quad \text{for } j = 1, \ldots, p, \tag{5b}$$

$$p(\sigma^2) \propto 1/\sigma^2, \tag{5c}$$

where β_j is degenerate at 0 if $\gamma_j = 0$. As both β_0 and σ^2 are assumed to be included in all models, this and arguments based on orthogonal parametrizations and invariance to scale and location transformation have been used to justify this objective choice in many variable selection applications; see Liang et al. (2008), Berger, Pericchi, and Varshavsky (1998) for more details.

Using independent Bernoulli prior distributions on the inclusion indicators γ_j

$$p(\boldsymbol{\gamma}|\boldsymbol{\pi}) = \prod_{j=1}^{p} \pi_j^{\gamma_j}(1-\pi_j)^{1-\gamma_j} \tag{6}$$

the posterior distribution of $\boldsymbol{\gamma}$ given σ^2 and the complete data $\mathbf{Y_c}$ also has an independent product Bernoulli distribution

$$p(\boldsymbol{\gamma}|\mathbf{Y_c}, \sigma^2) = \prod_{j=1}^{p}\big\{p(\gamma_j = 1|\mathbf{Y_c}, \sigma^2)^{\gamma_j} \times (1 - p(\gamma_j = 1|\mathbf{Y_c}, \sigma^2))^{1-\gamma_j}\big\} \tag{7a}$$

with posterior inclusion probability $\rho_j(\mathbf{Y_c}, \sigma^2, \lambda_j) \equiv p(\gamma_j = 1|\mathbf{Y_c}, \sigma^2)$ expressed via posterior odds as

$$\rho_j(\mathbf{Y_c}, \sigma^2, \lambda_j) \equiv O_j(\mathbf{Y_c}, \sigma^2, \lambda_j)/(1 + O_j(\mathbf{Y_c}, \sigma^2, \lambda_j)),$$

$$O_j(\mathbf{Y_c}, \sigma^2, \lambda_j) \equiv \frac{\pi_j}{1-\pi_j}\left(\frac{\lambda_j}{\delta_j + \lambda_j}\right)^{1/2} \times \exp\left\{\frac{1}{2}\frac{\delta_j}{\delta_j + \lambda_j}\frac{\widehat{\beta}_j^2}{\sigma^2}\delta_j\right\} \quad \text{for } j = 1, \ldots, p. \tag{7b}$$

The estimate $\widehat{\beta}_j$ is the jth element of $\widehat{\boldsymbol{\beta}} = (\mathbf{X_c}^T\mathbf{X_c})^{-1}\mathbf{X_c}^T\mathbf{Y_c}$, the vector of OLS regression coefficients of $\boldsymbol{\beta}$ using the complete data $\mathbf{Y_c}, \mathbf{X_c}$. Because $\mathbf{X_c}^T\mathbf{X_c}$ is diagonal, the inversion is trivial. The quantity, $\widehat{\beta}_j^2\delta_j/\sigma^2$ in the exponential term of the posterior odds is the square of the Z-score for testing the hypothesis $\beta_j = 0$, thus the posterior probabilities provide a way of converting frequentist test statistics into the posterior probabilities of hypotheses via Bayes theorem.

Predictions under orthogonal designs with BMA are greatly simplified. To predict a future $\mathbf{Y}^\star$ observed at $\mathbf{X}^\star$, the mean of the posterior predictive distribution

$$E(\mathbf{Y}^\star|\mathbf{Y_c}, \sigma^2) = \mathbf{x}_0^\star\hat{\beta}_0 + \sum_{j=1}^{p}\mathbf{x}_j^\star p(\gamma_j = 1|\mathbf{Y_c}, \sigma^2)\frac{\delta_j}{\delta_j + \lambda_j}\widehat{\beta}_j \tag{8}$$

takes the form of a multiple shrinkage estimator with linear shrinkage of the OLS estimate induced from the normal prior distribution given that β_j is nonzero, and nonlinear data-dependent shrinkage from the posterior inclusion probability $p(\gamma_j = 1|\mathbf{Y_c}, \sigma^2)$ that arises from the uncertainty of whether β_j is nonzero. Note that in the orthogonal case, the summation is over p terms as opposed to 2^p terms in BMA with nonorthogonal designs. Of course, while $\mathbf{Y_a}$ is not necessary under orthogonal designs, σ^2 is generally unknown. In the context of wavelets with no missing data ($\mathbf{Y_c} = \mathbf{Y_o}$), Clyde, Parmigiani, and Vidakovic (1998) use robust estimates of σ^2 from the data. As using a "plug-in" estimate for σ^2 may underestimate uncertainty, they suggest using a Gibbs sampler and calculate RB estimates of inclusion probabilities and fitted values. In the next section, we introduce a two-component Gibbs sampler based on the idea of orthogonal data augmentation that allows us to incorporate uncertainty in $\mathbf{Y_a}$ and σ^2.

2.1 ODA Algorithm

Given the model for the complete data (4) and the prior specification in (5a)–(5c) and (6), we may construct a two block Gibbs sampler with $[\sigma^2, \mathbf{Y_a}]$ in one block and the vector $[\boldsymbol{\gamma}]$ in the other using the following sequence of distributions:

$$1/\sigma^2|\boldsymbol{\gamma}, \mathbf{Y_o} \sim \mathrm{G}\left(\frac{n_o - 1}{2}, \frac{(\mathbf{Y_o}^T\mathbf{Y_o} - \widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}}^T(\mathbf{X}_{\mathbf{o}\boldsymbol{\gamma}}^T\mathbf{X_{o\gamma}} + \Lambda_{\boldsymbol{\gamma}})\widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}})}{2}\right), \tag{9a}$$

$$\mathbf{Y_a}|\sigma^2, \boldsymbol{\gamma}, \mathbf{Y_o} \sim \mathrm{N}\big(\mathbf{X_{a\gamma}}\widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}}, \sigma^2\big(\mathbf{I}_{n_a} + \mathbf{X_{a\gamma}}(\mathbf{X}_{\mathbf{o}\boldsymbol{\gamma}}^T\mathbf{X_{o\gamma}} + \Lambda_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\mathbf{a}\boldsymbol{\gamma}}^T\big)\big), \tag{9b}$$

$$\gamma_j|\mathbf{Y_a}, \sigma^2, \mathbf{Y_o} \overset{\text{ind}}{\sim} \mathrm{Ber}(\rho_j(\mathbf{Y_c}, \sigma^2, \lambda_j)) \quad \text{for } j = 1, 2, \ldots, p, \tag{9c}$$

where

$$\widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}} = (\mathbf{X}_{\mathbf{o}\boldsymbol{\gamma}}^T\mathbf{X_{o\gamma}} + \Lambda_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\mathbf{o}\boldsymbol{\gamma}}^T\mathbf{Y_o} \tag{10}$$

is the posterior mean for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under model $\boldsymbol{\gamma}$ and observed data $\mathbf{Y_o}$, $\rho_j(\mathbf{Y_c}, \sigma^2, \lambda_j)$ is given by Equation (7b), and $\Lambda_{\boldsymbol{\gamma}}$ is a $(p_{\boldsymbol{\gamma}} + 1) \times (p_{\boldsymbol{\gamma}} + 1)$ diagonal matrix with first entry $\lambda_0 = 0$ and the remaining $p_{\boldsymbol{\gamma}}$ entries are given by the subset of λ_j where $\gamma_j = 1$. The block $[\mathbf{Y_a}, \sigma^2]$ is sampled by drawing σ^2 from $p(\sigma^2|\boldsymbol{\gamma}, \mathbf{Y_o})$, and then drawing $\mathbf{Y_a}$ from $p(\mathbf{Y_a}|\boldsymbol{\gamma}, \sigma^2, \mathbf{Y_o})$. Alternatively the joint distribution could be decomposed by drawing first $\mathbf{Y_a}|\boldsymbol{\gamma}, \mathbf{Y_o}$ and then drawing $\sigma^2|\mathbf{Y_c}, \boldsymbol{\gamma}$. In practice we noted no difference, but prefer the former order as the draws for σ^2 are independent of $\mathbf{Y_a}$ given $\boldsymbol{\gamma}$ and are from the full conditional in the collapsed sampler where $\mathbf{Y_a}$ has been integrated out. Because of the independence of $\boldsymbol{\gamma}$ given $[\mathbf{Y_c}, \sigma^2]$, the entire block may be drawn in parallel.

The Markov chain may be used to "marginalize" over the distribution of $\mathbf{Y_a}$, so that we account for uncertainty in $\mathbf{Y_a}$ appropriately, however, $\mathbf{X_a}$ has been held fixed throughout. This in fact does not alter the posterior distribution $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}|\mathbf{X_o}, \mathbf{Y_o})$ as

$$p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}|\mathbf{X_o}, \mathbf{X_a}, \mathbf{Y_o}) = \int p(\mathbf{Y_a}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}|\mathbf{X_o}, \mathbf{X_a}, \mathbf{Y_o})\, d\mathbf{Y_a}$$

$$
\begin{aligned}
&= \int p(\mathbf{Y_o}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}, \mathbf{X_o}) p(\mathbf{Y_a}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}, \mathbf{X_a}) \\
&\quad \times p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2|\boldsymbol{\gamma}) p(\boldsymbol{\gamma})\, d\mathbf{Y_a} \\
&\quad \times \Bigg(\sum_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} \int\int\int p(\mathbf{Y_o}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}, \mathbf{X_o}) p(\mathbf{Y_a}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}, \mathbf{X_a}) \\
&\quad \times p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2|\boldsymbol{\gamma}) p(\boldsymbol{\gamma})\, d\mathbf{Y_a}\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\sigma^2 \Bigg)^{-1} \\
&= p(\mathbf{Y_o}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}, \mathbf{X_o}) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2|\boldsymbol{\gamma}) p(\boldsymbol{\gamma}) \\
&\quad \times \Bigg(\sum_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}} \int\int p(\mathbf{Y_o}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}, \mathbf{X_o}) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2|\boldsymbol{\gamma}) \\
&\quad \times p(\boldsymbol{\gamma})\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\sigma^2 \Bigg)^{-1} \\
&= p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{\gamma}|\mathbf{X_o}, \mathbf{Y_o}) \qquad (11)
\end{aligned}
$$

is the distribution of interest. We now show how Rao–Blackwellization may be used with the ODA sampler to reduce Monte Carlo variation in estimates.

2.2 Rao–Blackwellized Estimates of Quantities of Interest

Ergodic averages given the output of MCMC may be used to estimate functions of interest, such as model probabilities, inclusion probabilities and predictions. Let $\mathbf{1}(\boldsymbol{\gamma} = \boldsymbol{\gamma}^*)$ denote the indicator function for the event $\boldsymbol{\gamma} = \boldsymbol{\gamma}^\star$, then the ergodic average or Monte Carlo (MC) frequency estimator of the probability of model $\boldsymbol{\gamma}^\star$ is $\widehat{p}^{\,\mathrm{MC}}(\boldsymbol{\gamma}^\star|\mathbf{Y_o}) = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}(\boldsymbol{\gamma}^{(k)} = \boldsymbol{\gamma}^\star)$, where K is the number of draws from the Markov chain. Similarly posterior inclusion probabilities may be estimated by taking an ergodic average of $\mathbf{1}(\gamma_j{}^{(k)} = 1)$ and in general for a function $t(\boldsymbol{\gamma})$,

$$
\widehat{t}^{\,\mathrm{MC}}(\boldsymbol{\gamma}) = \frac{1}{K}\sum_{k=1}^{K} t\big(\boldsymbol{\gamma}^{(k)}\big). \qquad (12)
$$

"Rao–Blackwellization" of estimates (Gelfand and Smith 1990) has been suggested as a way to construct improved estimates by replacing $t(\boldsymbol{\gamma}^{(k)})$ with its expectation given other components in the sampler, and then applying the ergodic average to marginalize over the components. Applying this to the ODA sampler, leads to RB estimates of $t(\boldsymbol{\gamma})$

$$
\begin{aligned}
\widehat{t}^{\,\mathrm{RB}}(\boldsymbol{\gamma}) &= \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}\big[t\big(\boldsymbol{\gamma}^{(k)}\big)|\mathbf{Y_o}, \mathbf{Y_a}^{(k)}, \sigma^{2(k)}\big] \\
&= \frac{1}{K}\sum_{k=1}^{K} \mathrm{E}\big[t\big(\boldsymbol{\gamma}^{(k)}\big)|\mathbf{Y_c}^{(k)}, \sigma^{2(k)}\big]. \qquad (13)
\end{aligned}
$$

Because of the product Bernoulli posterior for $\boldsymbol{\gamma}$ conditional on $\mathbf{Y_a}, \sigma^2$, analytic expressions for the conditional expectation of the indicators of models, individual inclusion indicators or predictions is trivial. For ODA, the RB estimators of model probabilities and inclusion probabilities take the form

$$
\widehat{p}^{\,\mathrm{RB}}(\boldsymbol{\gamma}|\mathbf{Y_o}) = \frac{1}{K}\sum_{k=1}^{K}\prod_{j=1}^{p} \rho_j\big(\mathbf{Y_c}^{(k)}, \sigma^{2(k)}, \lambda_j\big)^{\gamma_j} \times \big(1 - \rho_j\big(\mathbf{Y_c}^{(k)}, \sigma^{2(k)}, \lambda_j\big)\big)^{1-\gamma_j}, \qquad (14a)
$$

$$
\widehat{p}^{\,\mathrm{RB}}(\gamma_j = 1|\mathbf{Y_o}) = \frac{1}{K}\sum_{k=1}^{K} \rho_j\big(\mathbf{Y_c}^{(k)}, \sigma^{2(k)}, \lambda_j\big), \qquad (14b)
$$

where $\rho_j(\mathbf{Y_c}, \sigma^2, \lambda_j)$ is the marginal inclusion probability given $\mathbf{Y_c}, \sigma^2, \lambda_j$ (7b). For a matrix $\mathbf{X}$ (at observed or new data points), RB estimates for model averaging of $\mathbf{X}\boldsymbol{\beta}$ are given by

$$
\widehat{\mathrm{E}}^{\mathrm{RB}}(\beta_j|\mathbf{Y_o}) = \frac{1}{K}\sum_{k=1}^{K} \rho_j\big(\mathbf{Y_c}^{(k)}, \sigma^{2(k)}, \lambda_j\big)\frac{\delta_j}{\delta_j + \lambda_j}\widehat{\beta}_j^{(k)}, \qquad (14c)
$$

$$
\widehat{\mathrm{E}}^{\mathrm{RB}}(\mathbf{X}\boldsymbol{\beta}|\mathbf{Y_o}) = \mathbf{x}_0\frac{1}{K}\sum_{k=1}^{K}\widehat{\beta}_0^{(k)} + \sum_{j=1}^{p}\mathbf{x}_j\widehat{\mathrm{E}}^{\mathrm{RB}}(\beta_j|\mathbf{Y_o}), \qquad (14d)
$$

where $\widehat{\beta}_j^{(k)}$ is the least squares estimate of β_j from the complete data $\mathbf{Y_c}^{(k)}, \mathbf{X_c}$ for the kth iteration.

The RB estimate in (14a) may be used to estimate the posterior model probability for any model, including those that are not sampled in the MCMC (whose ergodic average is 0). This gives us a simple way of estimating the remaining unsampled posterior mass. Let $\boldsymbol{\Gamma}_K$ denote the set of unique models in the K draws from the ODA sampler. Because the estimates of model probabilities given by Equation (14a) sum to 1 for the entire model space $\boldsymbol{\Gamma}$, we can estimate the total posterior probability of unsampled models by

$$
\begin{aligned}
\widehat{p}^{\,\mathrm{RB}}(\boldsymbol{\Gamma} - \boldsymbol{\Gamma}_K|\mathbf{Y_o}) &= \sum_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma} - \boldsymbol{\Gamma}_K} \widehat{p}^{\,\mathrm{RB}}(\boldsymbol{\gamma}|\mathbf{Y_o}) \\
&= 1 - \sum_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_K} \widehat{p}^{\,\mathrm{RB}}(\boldsymbol{\gamma}|\mathbf{Y_o}). \qquad (15)
\end{aligned}
$$

RB estimates are commonly believed to reduce the variance over ergodic averages, however, Liu, Wong, and Kong (1994) and Geyer (1995) show that this is not always true when the dependence in the Markov chain is taken into account. For the two block ODA algorithm, however, Rao–Blackwellization is guaranteed to result in variance reduction:

Theorem 1. Under the assumption that the Markov chain induced by the ODA sampler has reached its stationary distribution

$$
\begin{aligned}
\mathrm{var}[\widehat{p}^{\,\mathrm{RB}}(\boldsymbol{\gamma}|\mathbf{Y_o})] &\le \mathrm{var}[\widehat{p}^{\,\mathrm{MC}}(\boldsymbol{\gamma}|\mathbf{Y_o})], \\
\mathrm{var}[\widehat{p}^{\,\mathrm{RB}}(\gamma_j = 1|\mathbf{Y_o})] &\le \mathrm{var}[\widehat{p}^{\,\mathrm{MC}}(\gamma_j = 1|\mathbf{Y_o})].
\end{aligned}
$$

Proof. The ODA algorithm is a two-component Gibbs sampler with components $[\boldsymbol{\gamma}]$ and $[\mathbf{Y_a}, \sigma^2]$, which is also referred to as "data augmentation." The proof follows immediately from theorem 4.1 of Liu, Wong, and Kong (1994), who show that for such a Gibbs sampler, the RB estimator leads to a reduction in variance compared to the Monte Carlo estimate in estimating $\mathrm{E}[t(\boldsymbol{\gamma})|\mathbf{Y_o}]$ for any scalar function $t(\boldsymbol{\gamma})$.

Liu, Wong, and Kong (1994) and Liu (1994) also contrast collapsing and grouping as ways of reducing Monte Carlo variation and recommend collapsing in general. A collapsed Gibbs sampler results from integrating out one or more components of a Gibbs sampler analytically. Grouping or blocking refers to drawing two or more components together from their joint

conditional distribution, rather than drawing them individually from their respective conditional distributions. Because of the conjugate priors for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2$, the marginal posterior distribution for $\boldsymbol{\gamma}$ is available as

$$p(\boldsymbol{\gamma}|\mathbf{Y_o}) = \frac{p(\mathbf{Y_o}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}|\pi)}{\sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}} p(\mathbf{Y_o}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}|\pi)}, \tag{16}$$

where

$$p(\mathbf{Y_o}|\boldsymbol{\gamma}) \propto |\Lambda_{\boldsymbol{\gamma}}|^{+1/2}|\mathbf{X}_{\mathbf{o}\boldsymbol{\gamma}}^T\mathbf{X}_{\mathbf{o}\boldsymbol{\gamma}} + \Lambda_{\boldsymbol{\gamma}}|^{-1/2} \times (\|\mathbf{Y_o} - \mathbf{X}_{\mathbf{o}\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}}\|^2 + \widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}}^T\Lambda_{\boldsymbol{\gamma}}\widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}})^{-(n-1)/2}, \tag{17}$$

$|\Lambda_{\boldsymbol{\gamma}}|^+$ is the determinant of the lower positive definite $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\gamma}}$ block of $\Lambda_{\boldsymbol{\gamma}}$ obtained by excluding the first row and column, and $\widetilde{\boldsymbol{\beta}}_{\mathbf{o}\boldsymbol{\gamma}}$ is the posterior mean of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under the observed data only given by Equation (10). As it is not possible to sum over the models in $\boldsymbol{\Gamma}$ for large problems, a collapsed Gibbs sampler may be used to draw the components of $\boldsymbol{\gamma}$ from the Bernoulli full conditional distributions $p(\gamma_j|\boldsymbol{\gamma}_{(j)}, \mathbf{Y_o})$, where $\boldsymbol{\gamma}_{(j)}$ is $\boldsymbol{\gamma}$ with the jth component removed. This Gibbs sampler is equivalent to the popular integrated Stochastic Search Variable Selection (SSVS) algorithm introduced by George and McCulloch (1997) and may also be viewed as a collapsed Gibbs sampler induced by integrating out $\mathbf{Y_a}$ and σ^2 in the ODA algorithm. Liu (1994) shows that collapsing reduces the norm of the forward operator of the Gibbs sampler over that of a grouped Gibbs sampler, although this does not imply faster convergence or better mixing for nonreversible chains (Liu 1994; Liu, Wong, and Kong 1994). In general, one must balance computational ease, speed and mixing. By introducing the missing $\mathbf{Y_a}, \sigma^2$, we are able to generate $\boldsymbol{\gamma}$ in one block, while in the collapsed case, we must propose each γ_j conditional on the remaining components. One iteration of the collapsed Gibbs sampler requires p marginal likelihood evaluations and the solution of Equation (10) to obtain a new model; thus standard implementations[1] of SSVS are computationally more expensive than a single iteration of ODA, even with the generation of the latent $\mathbf{Y_a}$ and σ^2.

3. EXISTENCE AND CHOICE OF AUGMENTED DESIGNS

Our goal is to find a matrix $\mathbf{X_a}$ given the observed matrix $\mathbf{X_o}$ such that $\mathbf{X_c}^T\mathbf{X_c} = \mathbf{X_o}^T\mathbf{X_o} + \mathbf{X_a}^T\mathbf{X_a} = \mathbf{D}$ is diagonal. Furthermore, $\mathbf{X_a}$ must be real and $\mathbf{X_a}^T\mathbf{X_a}$ should be positive semidefinite (psd), which implies that it has real, nonnegative eigenvalues. There is a long, rich history describing conditions on the possible eigenvalues of the sums of two symmetric (Hermitian) matrices (see Fulton 2000 for a summary and key results). The earliest significant results are due to Weyl (1912) and may be restated for our purposes in the following lemma:

Lemma 1. Let $\mathbf{A}$, $\mathbf{O}$, and $\mathbf{D}$ be three real symmetric matrices of dimension d with ordered eigenvalues $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_d$, $o_1 \geq o_2 \cdots \geq o_d$, and $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_d$, respectively. Then for $\mathbf{D} = \mathbf{A} + \mathbf{O}$, $\delta_{i+j-1} \leq \alpha_i + o_j$ whenever $i + j - 1 \leq d$.

Given $\mathbf{D}$ and $\mathbf{X_o}^T\mathbf{X_o}$, the theorem below suggests how we may now construct $\mathbf{X_a}$.

[1] George and McCulloch (1997) discuss efficient updating under g-priors.

Theorem 2. Let $\mathbf{O} \equiv \mathbf{X_o}^T\mathbf{X_o}$ denote a psd symmetric matrix with ordered eigenvalues $o_1 \geq o_2 \geq \cdots \geq o_{p+1} \geq 0$ and $\mathbf{D}$ be a diagonal matrix with elements $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_{p+1}$. If $\delta_i \geq o_1$ for $1 \leq i \leq p+1$, then a real augmented design matrix $\mathbf{X_a}$ of dimension $(p+1) \times (p+1)$ always exists.

Proof. As both $\mathbf{O}$ and $\mathbf{D}$ are symmetric with real eigenvalues, $\mathbf{A} \equiv \mathbf{D} - \mathbf{X_o}^T\mathbf{X_o}$ is a real symmetric matrix with spectral decomposition $\mathbf{U}\boldsymbol{\alpha}\mathbf{U}^T$ for $\mathbf{U}$ a $(p+1) \times (p+1)$ orthogonal matrix and diagonal matrix $\boldsymbol{\alpha}$ of ordered eigenvalues $\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{p+1}$. From Lemma 1, $\delta_{i+j-1} - o_j \leq \alpha_i$ whenever $i + j - 1 \leq p + 1$, thus in order for $\mathbf{A}$ to be positive semidefinite, $\delta_{i+j-1} \geq o_j$ for $i + j - 1 \leq p + 1$, and in particular for $j = 1$, $\delta_i \geq o_1$ for $1 \leq i \leq p + 1$. Setting $\mathbf{X_a} = \mathbf{U}\boldsymbol{\alpha}^{1/2}\mathbf{U}^T$, the symmetric matrix square root of $\mathbf{A}$, ensures that $\mathbf{X_a}^T\mathbf{X_a} \equiv \mathbf{A}$ is psd and that $\mathbf{X_a}$ is real.

3.1 Choice of $\mathbf{D}$

If $\delta_i = o_1$ for $i = 1, \ldots, p+1$, then the eigenvalues of $\mathbf{X_a}^T\mathbf{X_a}$ will be $o_1 - o_i$ and there will be $r = p + 1 - m$ nonzero eigenvalues, where m is the number of eigenvalues of $\mathbf{O}$ that are equal to o_1. The construction of $\mathbf{X_a}^T\mathbf{X_a}$ above reduces to a $(p+1) \times (p+1)$ rank r matrix, $\mathbf{X_a}^T\mathbf{X_a} = \sum_{i=m+1}^{p+1}(o_1 - o_i)\mathbf{u}_i\mathbf{u}_i^T$ where $\mathbf{u}_i$ are the eigenvectors of $\mathbf{X_a}^T\mathbf{X_a}$. If all the eigenvalues of $\mathbf{X_o}^T\mathbf{X_o}$ are distinct then this implies $\mathbf{X_a}^T\mathbf{X_a}$ is of rank p.

Since the columns of $\mathbf{X_o}$ have been standardized, the choice of equal diagonal elements of $\mathbf{D}$, $\delta = \delta_1 = \cdots = \delta_{p+1}$ is not unreasonable as the columns of $\mathbf{X_o}$ are all on the same scale. Furthermore, this choice leads to the same sample precision for all $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under the complete data. In the special case that the original matrix has orthogonal columns, the eigenvalues $o_i = n_o$, thus the augmented design under the choice $\delta = o_1$ leads to a rank 0 matrix, hence augmentation is unnecessary.

Large δ increases the column sum of squares for $\mathbf{X_c}$ and hence will lead to less Monte Carlo variation for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ under the augmented data posterior. However, this also inflates the column sum of squares of $\mathbf{X_a}$ as $\mathbf{X_o}^T\mathbf{X_o}$ is fixed. As a consequence, the augmented design points may be large in magnitude in contrast to the observed points, which increases the leverage of the augmented points.

Theorem 3. Let $\mathbf{X_a} = \mathbf{U}\boldsymbol{\alpha}^{1/2}\mathbf{U}^T$ where $\mathbf{U}\boldsymbol{\alpha}\mathbf{U}^T = \delta\mathbf{I} - \mathbf{X_o}^T\mathbf{X_o}$ for $\delta = o_1 + \epsilon$ and o_1 the maximum eigenvalue of $\mathbf{X_o}^T\mathbf{X_o}$. Then $\epsilon = 0$ minimizes the leverage of the design points in $\mathbf{X_a}$.

Proof. The leverages under the complete design are the diagonal elements of

$$\mathbf{X_c}(\mathbf{X_c}^T\mathbf{X_c})^{-1}\mathbf{X_c}^T = \delta^{-1}\begin{bmatrix} \mathbf{X_o}\mathbf{X_o}^T & \mathbf{X_o}\mathbf{X_a}^T \\ \mathbf{X_a}\mathbf{X_o}^T & \mathbf{X_a}\mathbf{X_a}^T \end{bmatrix}$$

with lower block, $\delta^{-1}\mathbf{X_a}\mathbf{X_a}^T = \delta^{-1}\mathbf{U}\boldsymbol{\alpha}^{1/2}\mathbf{U}^T\mathbf{U}\boldsymbol{\alpha}^{1/2}\mathbf{U}^T = \delta^{-1} \times \mathbf{U}\boldsymbol{\alpha}\mathbf{U}^T$. The leverage for row i simplifies to $\sum_j(1 - o_j/\delta)u_{ij}^2$. Since $\delta \geq o_1$, the leverages are minimized for $\delta = o_1$.

While ideally we set $\delta = o_1$, in practice this choice sometimes leads to computationally unstable solutions with a small negative value for the smallest eigenvalue of $\mathbf{X_a}^T\mathbf{X_a}$. To ensure that $\mathbf{X_a}^T\mathbf{X_a}$ is positive definite we use $\epsilon = 0.001$. This definition of $\mathbf{D}$ will be assumed throughout the rest of the article.

3.2 Invariance

In Theorem 2, we took $\mathbf{X_a}$ to be the symmetric square root of $\mathbf{D} - \mathbf{X_0}^T\mathbf{X_0}$, however, any matrix square root of $\mathbf{D} - \mathbf{X_0}^T\mathbf{X_0}$ may be used to create an augmented design. We characterize the possible solutions and show that the distribution of the augmented $\mathbf{Y_a}$ is invariant under the choice of square root.

Lemma 2. Let $\mathcal{S}$ denote the set of $(p+1) \times (p+1)$ matrix square roots of $\mathbf{A} = \mathbf{D} - \mathbf{X_0}^T\mathbf{X_0}$ for $\mathbf{A} > 0$ such that for $\mathbf{S} \in \mathcal{S}$, $\mathbf{S}^T\mathbf{S} = \mathbf{A}$. Let $\mathcal{O}_{p+1}$ denote the group of $(p+1) \times (p+1)$ orthogonal matrices with group action $(\cdot)$ matrix multiplication. Then any $\mathbf{S}^* \in \mathcal{S}$ may be written as $\mathbf{S}^* = \mathbf{OS}$ for some $\mathbf{O} \in \mathcal{O}_{p+1}$ and $\mathbf{S}$ in $\mathcal{S}$.

Proof. The orthogonal group "acts" on the left of $\mathcal{S}$ (see Eaton 1983, p. 186. definition 6.1), that is, for all $\mathbf{S} \in \mathcal{S}$ and $\mathbf{O}_1, \mathbf{O}_2 \in \mathcal{O}_{p+1}$, $(\mathbf{O}_1\mathbf{O}_2) \cdot \mathbf{S} = \mathbf{O}_1 \cdot (\mathbf{O}_2 \cdot \mathbf{S})$ with identity group element $\mathbf{I}_{p+1}$, $\mathbf{I}_{p+1} \cdot \mathbf{S} = \mathbf{S}$. Then proposition 6.1 of Eaton (1983) may be used to show that $\mathbf{OS}$ is one-to-one and onto from $\mathcal{S}$ to $\mathcal{S}$.

Theorem 4. Let $\mathbf{Y_a} \sim \mathrm{N}(\mathbf{S}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ where $\mathbf{S} \in \mathcal{S}$, the set of matrix square roots of $\delta\mathbf{I} - \mathbf{X_0}^T\mathbf{X_0}$. Then the distribution of $\mathbf{Y_a}$ is invariant to the choice of $\mathbf{S}$.

Proof. For $\mathbf{O} \in \mathcal{O}_{p+1}$, $\mathbf{Y_a^*} \equiv \mathbf{OY_a} \stackrel{d}{=} \mathbf{OS}\boldsymbol{\beta} + \mathbf{O}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Since $\mathbf{S}^* = \mathbf{OS} \in \mathcal{S}$, $\mathbf{Y_a^*} \sim \mathrm{N}(\mathbf{S}^*\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

The above result says that if $\mathbf{Y_a}$ is an augmentation such that $\mathbf{Y_a} \sim \mathrm{N}(\mathbf{S}\boldsymbol{\beta}, \sigma^2\mathbf{I})$, then for any other matrix square root of $\delta\mathbf{I} - \mathbf{X_0}^T\mathbf{X_0}$, say $\mathbf{S}^*$ (where $\mathbf{S}^* = \mathbf{OS}$ for some $\mathbf{O} \in \mathcal{O}_{p+1}$), the corresponding augmentation $\mathbf{Y_a^*}$ (obtained under left-multiplication by the orthogonal matrix $\mathbf{O}$) is also normally distributed as $\mathrm{N}(\mathbf{S}^*\boldsymbol{\beta}, \sigma^2\mathbf{I})$. Because it is fast to compute, we take $\mathbf{X_a}$ to be the Cholesky decomposition of $\mathbf{D} - \mathbf{X_0}^T\mathbf{X_0}$ for the remainder of the article. Results for sensitivity analysis using two other choices of $\mathbf{X_a}$ based on singular value decompositions are reported in the Supplemental Materials. These results suggest that the ODA algorithm is not sensitive to the choice of $\mathbf{X_a}$.

4. SIMULATION STUDY

We compare ODA and other MCMC samplers for data generated using the simulation design in Nott and Kohn (2005) (similar to Raftery, Madigan, and Hoeting 1997; Fernández, Ley, and Steel 2001) with a sample size $n = 50$ and 15 predictors. The first column of the design matrix $\mathbf{X}$ is a column of ones and the next 10 columns, $\mathbf{x}_1, \ldots, \mathbf{x}_{10}$, are generated as independent $\mathrm{N}(0, 1)$ variables. The last five columns, $\mathbf{x}_{11}, \ldots, \mathbf{x}_{15}$ depend on the first five predictors $[\mathbf{x}_{11}, \ldots, \mathbf{x}_{15}] = [\mathbf{x}_1, \ldots, \mathbf{x}_5](0.3, 0.5, 0.7, 0.9, 1.1)^T(1, 1, 1, 1, 1) + \mathbf{E}$ where $\mathbf{E}$ is a 50×5 matrix of independent $\mathrm{N}(0, 1)$ random variables. This induces strong correlations among the last five variables and moderate correlations between them and the first five predictors. Given $\mathbf{X}$, the response variable is generated as $\mathbf{Y_o} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}^* = (4, 2, 0, 0, 0, -1, 0, 1.5, 0, 0, 0, 1, 0, 0.5, 0, 0)^T$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(0, 2.5^2\mathbf{I})$.

With 15 predictors and conjugate prior distributions, the model space of $2^{15} = 32,768$ possible models may be enumerated easily to provide the true posterior distribution under model averaging. For all methods we use a design matrix $\mathbf{X_o}$ obtained by standardizing $\mathbf{X}$ as described in Section 2 and set $\lambda_j = 1$ in the Gaussian prior distribution for the coefficients in (5b). If the predictors were uncorrelated, this choice would be equivalent to a unit-information prior (Kass and Raftery 1995). To complete the prior specification, we take $\pi_j = 1/2$ corresponding to an uniform prior distribution on the model space.

We compare ODA with two alternative MCMC algorithms. The first is SSVS, which is a collapsed version of ODA, described previously. Because the SSVS algorithm and the related MC^3 algorithm of Raftery, Madigan, and Hoeting (1997) use one component at a time updates, high correlation among predictors may make it difficult to escape local modes, leading to poor mixing in practice. To improve mixing, we use a Metropolis–Hastings algorithm based on a mixture kernel that randomly selects a γ_j and switches γ_j to $1 - \gamma_j$ thus adding or deleting a variable as in the MC^3 proposal (Raftery, Madigan, and Hoeting 1997), combined with a random swap proposal that randomly exchanges a predictor in the current model with one that is not. This simple random swap (RS) algorithm greatly aids in escaping local modes (Denison, Mallick, and Smith 1998; Nott and Green 2004; Clyde, Ghosh, and Littman 2011) and often performs as well as the more complicated Swendsen–Wang method of Nott and Green (2004). All MCMC methods use the same likelihood for the observed data and prior distributions given above for ODA, so that all methods should provide samples from the same target distribution.

For SSVS and RS we compute estimates of posterior model probabilities and inclusion probabilities based on (i) Monte Carlo frequencies of visits to models (MC) and (ii) marginal likelihoods of the unique models in the sample renormalized (RM) over the set of unique sampled models ($\boldsymbol{\Gamma}_K$),

$$\begin{aligned} \widehat{p}^{\mathrm{RM}}(\boldsymbol{\gamma}|\mathbf{Y_o}) &= \frac{p(\mathbf{Y_o}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_K} p(\mathbf{Y_o}|\boldsymbol{\gamma})p(\boldsymbol{\gamma})} I(\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_K), \\ \widehat{\mathbf{p}}^{\mathrm{RM}}(\gamma_j = 1) &= \sum_{\boldsymbol{\gamma}\in\boldsymbol{\Gamma}_K} \gamma_j \widehat{p}^{\mathrm{RM}}(\boldsymbol{\gamma}|\mathbf{Y_o}). \end{aligned} \tag{18}$$

4.1 Comparisons of Algorithms and Estimators

We compared the three algorithms based on a fixed running time as a single "iteration" of each of the samplers is not equivalent: ODA updates $\boldsymbol{\gamma}$ as one block with a much simpler full conditional, but requires the generation of the $p+1$ dimensional vector $\mathbf{Y_a}$ and the scalar σ^2; one iteration of SSVS cycles through the p components of $\boldsymbol{\gamma}$ with p marginal likelihood evaluations for the full conditionals; RS changes one or two components of $\boldsymbol{\gamma}$ and requires a single marginal likelihood evaluation in the Metropolis–Hastings ratio. Running each algorithm for one hour[2] results in 450,000, 46,667, and 700,000 iterations for ODA, SSVS, and RS, respectively; all simulations were done in R on a Dell workstation with 8 3.2 GHz Intel Xeon CPUs. Ideally one should run MCMC algorithms long enough to ensure convergence of model probabilities and marginal inclusion probabilities, however, this would require the number of MCMC iterations to be much more than the size of the model space. As this is computationally prohibitive for large model spaces, the typical number of draws is only a small fraction of the dimension of the model space. To reproduce such

[2] All algorithms were programmed in native R.

Table 1. Nott–Kohn simulation example: Comparison of square root of the average mean squared error for estimation of the 15 posterior inclusion probabilities (Incl. prob.) and 2^{15} posterior model probabilities (Model prob.) using Monte Carlo frequencies (MC), Rao–Blackwellized (RB) and renormalized marginal likelihoods (RM) under short runs (1 minute) and long runs (1 hour) of each of the ODA, SSVS, and RS algorithms

Square root average MSE	Short runs Incl. prob. (×100)	Long runs Incl. prob. (×100)	Short runs Model prob. ($\times 10^4$)	Long runs Model prob. ($\times 10^4$)
ODA MC	1.28	0.15	0.83	0.09
ODA RB	1.11	0.13	0.39	0.05
ODA RM	2.43	0.12	0.66	0.02
SSVS MC	1.98	0.23	2.32	0.27
SSVS RM	6.48	0.73	3.22	0.14
RS MC	2.03	0.24	1.42	0.17
RS RM	3.02	0.14	0.92	0.02

NOTE: The calculations are based on 100 replicates of each algorithm.

a scenario but where we can still evaluate the exact posterior distribution, we also run the algorithms for one minute, resulting in 6400, 667, and 10,000 draws (19.5%, 2%, 30.5% of the model space) for ODA, SSVS, and RS, respectively. For comparison, enumeration of the model space took 2.7 minutes. To compare how well the different estimators and algorithms estimate model probabilities and inclusion probabilities, we use $\text{MSE} = \|\hat{\mathbf{p}} - \mathbf{p}\|^2/d$ where $\hat{\mathbf{p}}$ is an estimate of the vector of probabilities, $\mathbf{p}$ is the value obtained under enumeration of the model space, and d is the length of the vector, $d = 2^p$ for the model probabilities and $d = p$ for inclusion probabilities. For each method, we ran the MCMC algorithm 100 times with random starting values, and computed the average MSE, taking the mean over the 100 runs (Table 1).

The results in Table 1 indicate that the Monte Carlo estimates from ODA are better than MC estimates from the other algorithms for the same running time. Likewise using the renormalized marginal likelihoods (RM) is best under the ODA chain. The root MSE results show that even within the short run we may estimate inclusion probabilities within plus or minus 0.02 under ODA RB. The Rao–Blackwellization leads to a reduction in MSE over the MC estimates for both short and long runs, as the theory suggests, however, the renormalized estimates are better only in the long runs. As the RM estimates are Fisher consistent (they equal the true value when the population is enumerated), the MSE goes to zero as the number of unique models converges to $|\mathbf{\Gamma}|$. The long runs of ODA, SSVS, and RS sample 8484, 4018, and 8274 unique models (on average), corresponding to 99.5, 96.27, and 99.4% of the posterior mass (respectively). However, for the short runs ODA, SSVS, and RS sample 84.9, 52.1, and 80.2% of the mass, respectively, and the superior performance of the RM estimates of inclusion probabilities is diminished over the MC estimates for all algorithms. These results confirm those in Clyde and Ghosh (2010), who show that renormalization may result in larger MSE over MC estimates when the number of iterations is small relative to the model space, because of the inherent bias in RM estimators. As ODA RB has better performance compared to MC and RM in the short runs, this should provide improved estimates for nonenumerable model spaces where a smaller fraction of the posterior mass is sampled. We now turn to estimation of the unsampled mass, which may provide some guidance when ODA RB is preferable to ODA RM and whether one should continue running the MCMC.

4.2 Estimation of Unsampled Posterior Mass

We can use the RB estimate in (15) to estimate the remaining posterior mass for the Nott–Kohn simulations and compare them to the exact value under enumeration. Using a short preliminary sample from the MCMC, George and McCulloch (1997) constructed an estimate of the normalizing constant C, for the posterior distribution of $\boldsymbol{\gamma}$ given in (16), $p(\boldsymbol{\gamma}|\mathbf{Y_o}) = Cp(\mathbf{Y_o}|\boldsymbol{\gamma})p(\boldsymbol{\gamma}|\boldsymbol{\pi})$, which in turn leads to an estimate of the unsampled mass. We found that the average bias was negligible for both estimators, although the George and McCulloch estimator exhibits much greater run-to-run variability (see Figure 1 in Supplemental Materials).

This ODA estimate of unsampled mass exhibits a slight tendency towards underestimation; a possible explanation is included in the Supplemental Materials. Although the bias is negligible, we provide two alternatives to correct it. One option is to run an independent Markov chain, and calculate the estimate in Equation (15), based on $(\mathbf{Y_a}, \sigma^2)$ generated from this new chain. We refer to this unbiased estimate as the Independent Rao–Blackwellized (IRB) estimate. Another approach that does not require additional simulation is to split randomly the MCMC samples from the original chain into two halves. The collection of $(\mathbf{Y_a}, \sigma^2)$ from the first half can be used to estimate the probabilities of models from the second half and vice-versa; adding the two estimates provides an estimate of the sampled mass from the entire chain, which is then subtracted from one to estimate the remaining mass. We call this the Rao–Blackwell Split (RB–Split) estimate. The Supplemental Materials contain an expanded version of this section with a detailed comparison of all estimators. Among the methods, IRB is the best, but at the expense of running a second independent chain. As a compromise that reduces bias, but does not increase computational complexity, we recommend the RB–Split method.

A key feature of the ODA method for estimating probabilities is that it does not require the marginal likelihoods to be available in closed form unlike the method of George and McCulloch (1997) to provide estimates of the unsampled mass. For enumerable problems where marginal likelihoods are not available in closed form, ODA clearly has an advantage in estimating not only the unsampled probability but also model probabilities for *all* models, as we illustrate in Section 7. We first show how to extend the method to allow alternatives to Gaussian prior distributions, another important example where marginal likelihoods are not available.

5. ODA WITH SCALE MIXTURES OF NORMALS

Normal prior distributions on the nonzero coefficients as in (5b) have been widely used for Bayesian variable selection, as the conjugate prior permits tractable marginal likelihoods used by collapsed MCMC algorithms that sample only from the model space. While normal priors are appealing from a computational perspective, Jeffreys (1961) rejected the normal prior

for hypothesis testing of a mean from a single normal population and recommended Cauchy priors instead. There are two potential problems with Gaussian prior distributions. First, posterior probabilities are sensitive to the prior variance, where large variances may lead to false rejection of the hypothesis that the coefficient is nonzero, the well-known "Lindley's" or "Bartlett's" paradox. Second, the normal prior does not have bounded influence, thus even as the t statistic increases to infinity, the prior mean of zero exerts influence on the marginal likelihood so that Bayes factors in favor of the alternative do not go to infinity but are instead bounded. The lack of bounded influence also affects the posterior mean, as the linear shrinkage of the sample mean is independent of the data. Liang et al. (2008) show that these problems arise in the regression context using the normal g-prior (Zellner 1986) and prove that scale mixtures of Zellner's g-prior, obtained by placing a prior distribution on g resolve these problems. These mixtures of g-priors are multivariate priors with the same prior correlation structure as in the observed data, which may not be desirable in the presence of multicollinearity. In the orthogonal wavelet setting, Clyde and George (2000) use scale mixtures of independent normal distributions for both error distributions as well as wavelet coefficients, which provides both robustness to using a prior mean of zero and robustness to outliers in the data. Johnstone and Silverman (2004, 2005) also demonstrate the superior performance of mixtures of a point mass and heavy-tailed prior distributions compared to mixtures of a point mass and Gaussian prior distributions. Scale mixtures of normal distributions have also seen a resurgence of interest for shrinkage estimators (priors with no explicit probability at zero) (Carvalho, Polson, and Scott 2010; Polson and Scott 2010) where MAP estimates are easy to obtain. These scale mixtures may be easily incorporated in the ODA algorithm by placing a prior distribution on the λ_j in (5b), leading to heavy-tailed prior distributions like the Cauchy, Student-t, double exponential, or horseshoe for the nonzero regression coefficients. We illustrate the algorithm for Cauchy prior distributions below.

5.1 ODA for Cauchy Prior Distributions

We add another layer of hierarchy to the normal prior in (5b):

$$\lambda_j \overset{\text{iid}}{\sim} \text{G}(\alpha/2, \alpha/2), \qquad j = 1, \ldots, p, \tag{19}$$

which leads to a marginal distribution for β_j given γ_j and σ^2 that is Student-t with location and scale parameters 0 and σ^2, and α degrees of freedom. In particular, the choice $\alpha = 1$ corresponds to $\beta_j|\gamma_j, \sigma^2 \overset{\text{iid}}{\sim} \text{C}(0, \sigma^2\gamma_j)$. Collapsing over γ_j, the prior distribution for β_j is a mixture of a point mass at zero and a Cauchy distribution, a robust version of the "spike-and-slab" prior.

With Cauchy prior distributions, marginal likelihoods may no longer be expressed in closed form and hence explicit expressions for posterior probabilities even where enumeration is feasible do not exist. We may add a step to sample the λ_j's, leading to the following ODA–Cauchy algorithm. For the complete data model in (4) and prior distributions in (5a)–(5c), (6), and (19) generate σ^2, $\mathbf{Y_a}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\lambda}$ from the following sequence of distributions:

1. $(\sigma^2, \mathbf{Y_a})|\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{Y_o}$ from Equations (9a)–(9b)
2. $\gamma_j|\mathbf{Y_c}, \sigma^2, \boldsymbol{\lambda}$ for $j = 1, \ldots, p$ from Equation (9c)
3. $\beta_j|\mathbf{Y_c}, \sigma^2, \boldsymbol{\gamma}, \boldsymbol{\lambda} \overset{\text{ind}}{\sim} \text{N}(\widehat{\beta}_j\gamma_j\frac{\delta_j}{\lambda_j+\delta_j}, \sigma^2\gamma_j(\lambda_j + \delta_j)^{-1})$ where $\widehat{\beta}_j = \mathbf{x_c}_j^T\mathbf{Y_c}/\mathbf{x_c}_j^T\mathbf{x_c}_j$ for $j = 1, 2, \ldots, p$
4. $\lambda_j|\mathbf{Y_c}, \sigma^2, \boldsymbol{\gamma}, \beta_j \overset{\text{ind}}{\sim} \text{G}(\frac{\alpha+\gamma_j}{2}, \frac{\alpha+\gamma_j\beta_j^2/\sigma^2}{2})$ for $j = 1, 2, \ldots, p$.

Note, the full conditional distribution of the scale parameter vector, $\boldsymbol{\lambda}$, is not available in closed form when $\boldsymbol{\beta}$ is integrated out and hence we draw β_j in step 3 to facilitate the drawing of $\boldsymbol{\lambda}$, but integrate out $\boldsymbol{\beta}$ from steps 1 and 2 to exploit the collapsed structure of the original ODA algorithm. We give a proof in the supplemental materials that the stationary distribution for this ODA sampler is the desired posterior distribution of $(\sigma^2, \mathbf{Y_a}, \boldsymbol{\gamma}, \boldsymbol{\lambda})$. The RB estimates for ODA–Cauchy are obtained by averaging over the K draws of $\mathbf{Y_a}$, σ^2 and λ.

5.2 Comparison of Different Prior Distributions

In this section we use the Nott–Kohn simulations to compare the performance of model averaging with ODA–Cauchy (ODA with independent Cauchy priors) and ODA–Normal (ODA with normal priors) for parameter estimation under squared error loss. In addition, we compare the independent priors in ODA to BMA using the multivariate g-prior (Zellner 1984) with $g = n_o$ (a unit-information prior), the multivariate Zellner–Siow (Zellner and Siow 1980) Cauchy prior distribution, and BMA using BIC to approximate model probabilities. The latter three estimates were computed under enumeration of the model space using the R package BAS on CRAN (Clyde 2010).

We generate 100 simulated datasets for the Nott–Kohn example and compare the different prior distributions based on their sum of squared errors (SSE) in estimating $\boldsymbol{\beta}$, $\text{SSE} = \sum_{j=1}^{p}(\beta_j - \widetilde{\beta}_j)^2$, where β_j is specified in Section 4 and $\widetilde{\beta}_j$ is the BMA estimate of β_j. BAS is a sampling without replacement algorithm that is guaranteed to enumerate the model space for number of predictors less than 20–25 (Clyde, Ghosh, and Littman 2011). As the number of predictors is 15, we ran BAS for 2^{15} iterations for enumeration, providing exact posterior means under BMA for the g-prior. BAS uses a one-dimensional Laplace approximation to calculate the marginal likelihood for Zellner–Siow prior (Liang et al. 2008). ODA–Cauchy and ODA–Normal are run for 2^{15} iterations after discarding a burn-in of 5000. CPU times for running ODA and BAS are not directly comparable currently as ODA is written currently in interpreted R code, while the sampling algorithm in the BAS package is written in C/FORTRAN and makes extensive use of the BLAS library. We plan to make an R package available and will recode in C/Fortran to speed up the calculations and expect that "an iteration" of ODA to be close to that of the BAS package as both algorithms require calculation of the OLS estimates, and use independent Bernoulli distributions to generate models. Currently for each of the 100 simulated datasets, ODA takes around 5 minutes, whereas the other prior distributions take around 0.2–0.3 seconds using the package BAS.

For each simulated dataset, we identified the algorithm that had the smallest SSE, and computed the relative efficiency of each algorithm to the one with the minimum SSE, $\text{EFF}(\text{algorithm}_k) = \text{SSE}(\text{algorithm}_k)/\min(\text{SSE})$. Boxplots of the relative efficiencies are shown in Figure 1. ODA–Normal and ODA–Cauchy have boxplots that are most concentrated around one, indicating that they are often producing the smallest SSE, and at other times performing almost as well as the

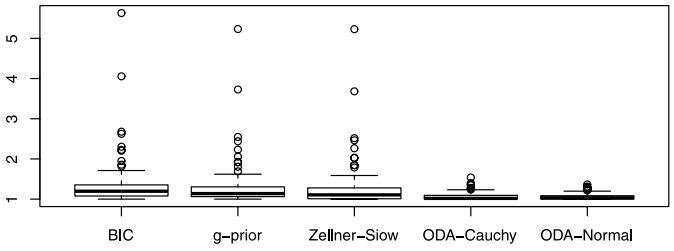

Figure 1. Boxplots of relative efficiency (SSE(algorithm$_k$)/min(SSE) for each simulated dataset) for the Nott–Kohn simulation study for BIC, g-prior, Zellner–Siow multivariate Cauchy prior, independent Cauchy, independent normal.

prior with the smallest SSE. For this simulation study, there is little difference among Zellner's $g-$prior and the Zellner–Siow Cauchy prior. Similarly the results for the independent normal and Cauchy priors based on ODA are very close, thus most of the gain is from the generalized ridge estimates based on the independent prior used in ODA. The mean SSE calculated over 100 replicates for ODA–Normal, ODA–Cauchy, Zellner–Siow, g-prior, and BIC are 2.08, 2.09, 2.34, 2.40, and 2.48 respectively.

6. EXAMPLE: OZONE DATA

We illustrate the performance of ODA using the ozone data (Friedman and Silverman 1989). There are 330 observations and eight meteorological variables, where the response variable is ground level ozone in Los Angeles. Considering the same second order interactions and square terms, as in Liang et al. (2008), there are 44 predictors in all, leading to a model space of 2^{44} models.

For ODA–Normal we may calculate the exact marginal likelihoods of the sampled models, which provides a diagnostic to check convergence for posterior model probabilities. We run ODA–Normal for one million iterations and discard the first 200,000 as burn-in. For model spaces as large as this one, most of the MC estimates of model probabilities are around $1/K$, where K is the number of MCMC iterations. Figure 2 highlights this problem and contrasts the true marginal likelihoods with their RB and MC estimates. Ideally the points should be tightly clustered around the 45° line. While RB offers an improvement over the MC estimates, RB also exhibits considerable Monte Carlo variation. This example illustrates the difficulty of estimating model probabilities in higher dimensions, which is still an open problem. In problems like this, we do not advocate using the RB estimates of model probabilities for selecting the highest probability model or a set of top models. The ODA algorithm is still useful for implementing BMA with heavy-tailed prior distributions in large model spaces as this one, and provides reliable estimates of marginal inclusion probabilities for variable selection or predictions under BMA.

We explore the predictive performance of the ODA method with normal and Cauchy prior distributions. Following Liang et al. (2008), we leave out a randomly chosen half of the data as training data and predict the ozone concentration for the remaining half ($n^* = n_o/2$) using BMA estimates for ODA–Normal and ODA–Cauchy. For comparison we also include BMA predictions under the g-prior (with $g = n_o/2$) and the Zellner–Siow prior using the R package BAS, the lasso (Tibshirani 1996; Efron et al. 2004) using the R package lars (Hastie and Efron 2007) and the Bayesian Horseshoe (Carvalho, Polson, and Scott 2010) using the R package monomvn (Gramacy 2010). The penalty in the lasso may be viewed as the log of a double exponential prior, which is a scale mixture of normals, while the horseshoe prior is yet another scale mixture of normals. While both result in nonlinear shrinkage rules, neither include a mixture component that puts positive probability on zero. We compare methods on the basis of the predictive root mean square error (RMSE) defined as $\sqrt{\sum_{i=1}^{n^*}(Y_i - \widehat{Y}_i)^2/(n^*)}$ where $Y_1, \ldots, Y_{n^*}$ are the response variables in the randomly selected test sample and $\widehat{Y}_i$ is the BMA estimate of Y_i.

We run the ODA algorithms for 2^{15} iterations after discarding a 5000 burn-in and run BAS for $2^{15} + 5000$ iterations. Empirically, predictions under model averaging appear to converge faster than model probabilities in these large spaces, thus we use a shorter MCMC run. The entire procedure is repeated 10 times to have 10 different sets of training and test samples. The resulting RMSE from different prior distributions (and algorithms)

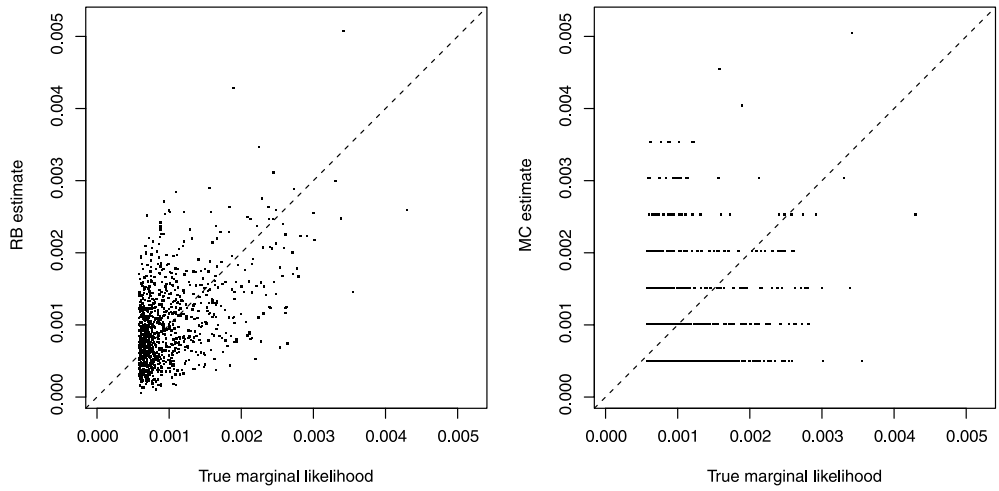

Figure 2. Comparison of Rao–Blackwellized (RB) and Monte Carlo (MC) estimates of (renormalized) marginal likelihoods of top 1000 models for the ozone data for ODA–Normal with one million iterations.

are quite similar, with the between sample variation dominating. As before the best methods are ODA, with root average MSE (RAMSE) of 4.098 for ODA–Cauchy, 4.102 for ODA–Normal, 4.131 for horseshoe, 4.153 for lasso, and 4.168 for both the Zellner g-prior and Zellner–Siow priors. Overall, the ODA methods are competitive with some of the best available methods.

7. PROBIT REGRESSION

We now show how to extend the ODA framework to binary regression models. Letting $\mathbf{Z_o}$ denote the binary response variable of length n_o, a probit regression model may be expressed as

$$P(Z_{\mathbf{o}i} = 1 | \mathbf{x}_{\boldsymbol{\gamma}i}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) = \Phi(\mathbf{x}_{\boldsymbol{\gamma}i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}), \qquad i = 1, 2, \ldots, n_o, \tag{20}$$

where Φ is the normal cumulative distribution function (cdf). Using the method of Albert and Chib (1993), one can augment the observed data $\{\mathbf{Z_o}, \mathbf{X_o}\}$ by latent data $\mathbf{Y_o}$ such that

$$Y_{\mathbf{o}i} | \boldsymbol{\beta}_{\boldsymbol{\gamma}} \overset{\text{ind}}{\sim} \mathrm{N}(\mathbf{x}_{\boldsymbol{\gamma}i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1), \tag{21}$$

where $Z_{\mathbf{o}i} = \mathbf{1}_{(Y_{\mathbf{o}i} > 0)}$. Using the prior specifications given in (5b) with $\sigma^2 = 1$ and $\lambda_j = 1$ for $j = 1, \ldots, p$ and (6), we can construct the ODA–Probit sampler, described in the Supplemental Materials.

To illustrate the application of the ODA–Probit algorithm, we use the Pima Indians Diabetes dataset available in the MASS library in R. The response variable is an indicator of whether the women had diabetes according to World Health Organization criteria. There are seven explanatory variables: npreg, glu, bp, skin, bmi, ped, and age. We provide a variable key in the Supplemental Materials.

We run 10 ODA–Probit chains for 300,000 iterations after a burn-in of 5000, with random starting values. On average, 39 (± 1) unique models were sampled, with an estimate of the remaining mass of 1.2×10^{-5} (range 2.2×10^{-6}–1.9×10^{-5} over the 10 chains), using the RB-split estimate. Inclusion probabilities were stable over the 10 runs, with estimates of marginal inclusion probabilities $p(\gamma_{\text{npreg}} = 1) = 0.947$ (sd $= 0.002$), $p(\gamma_{\text{glu}} = 1) = 1.000$ (sd $= 1.254 \times 10^{-29}$), $p(\gamma_{\text{bp}} = 1) = 0.075$ (sd $= 0.0002$), $p(\gamma_{\text{skin}} = 1) = 0.099$ (sd $= 0.001$), $p(\gamma_{\text{bmi}} = 1) = 0.997$ (sd $= 0.0004$), $p(\gamma_{\text{ped}} = 1) = 0.969$ (sd $= 0.001$), and $p(\gamma_{\text{age}} = 1) = 0.389$ (sd $= 0.003$). For generalized linear models, analytic expressions for marginal likelihoods are not available in closed form; thus even for an enumerable model space of dimension 2^7, there is no way of obtaining closed form expressions for true model probabilities or marginal inclusion probabilities. Here we can enumerate all 2^7 models and estimate the model probabilities using the ODA RB estimates for *any* model. We observe that npreg, glu, bmi, and ped are identified as important covariates, which were also flagged as important in an article by Holmes and Held (2006). Although their prior distribution for $\boldsymbol{\beta}$ is slightly different from ours, Holmes and Held (2006) reported the inclusion probability for age was 0.129, suggesting doubt about the importance of age, while Ripley (1996) found that the best AIC model included age, dropping only bp and skin. Looking more closely at the distribution of $\rho_{\text{age}}^{\text{RB}}$ from one of the chains, we found that it is highly bimodal, with support near zero and one. Images of the model space show that age is actually included in two out of the top four models (Figure 3), with the highest probability model under ODA–Probit corresponding to the second best model in terms of AIC and the best AIC model corresponding to the second highest posterior probability model. The Bayes factor for comparing the top two models is 1.11, suggesting that these two models receive similar support from the data. In this case, model averaging may be preferable to model selection. Splitting the data as in Ripley (1996), and refitting the ODA–Probit model with the 200 training observations to predict for the remaining 332, we find that the misclassification rate under ODA–Probit with model averaging is 18.7%, which is better than the highest probability model under ODA–Probit (19.4%), the best AIC model (19.9%), or the best BIC model (20.2%).

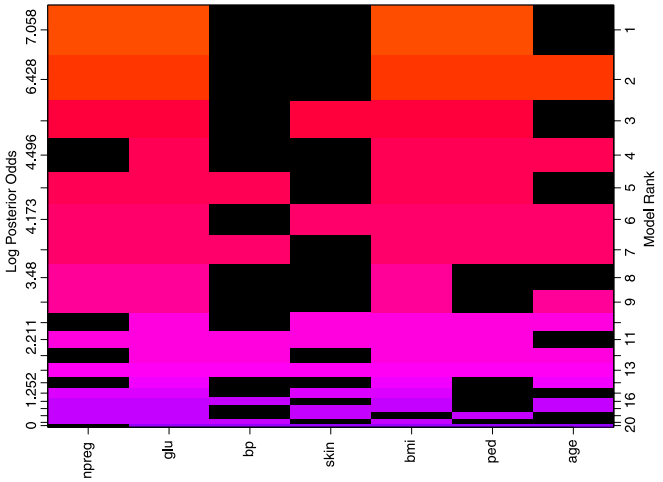

Figure 3. Top 20 models for the Pima Indians Diabetes data. Each row corresponds to a model, with the highest probability model at the top, and each column corresponds to a variable, with black regions indicating variables that are excluded from the model. The y-axis is the log Bayes factor for comparing each model to the 20th best model out of all (2^7) models. Log Bayes factors for comparing any other two models may be found by subtraction.

8. DISCUSSION

In the majority of problems outside Gaussian regression with conjugate prior distributions, the marginal likelihood of models and hence model probabilities are not analytically tractable, even for relatively small model spaces. The orthogonal data augmentation algorithm provides Rao–Blackwellized estimates of model probabilities, inclusion probabilities and predictions which provide improvements over current algorithms with normal priors, but more importantly allow one to consider more robust prior specifications, such as independent Cauchy priors. For linear regression ODA–Cauchy provides results that are equivalent or better than the "gold-standard" Zellner–Siow Cauchy prior, lasso or horseshoe. The ODA algorithm for probit regression provides Rao–Blackwellized estimates of model probabilities and inclusion probabilities so that one does not have to rely on asymptotic approximations, such as BIC, for model probabilities.

There are several extensions of the current method that are possible. For illustration purposes, we have used a fixed uniform prior on the model space throughout; this assumption may

be relaxed by placing a prior distribution, for example a Beta distribution, on π_j and updating π_j from its full conditional. The restriction to independent priors on the β_j may also be relaxed. For normal and mixture of normal prior distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ that are obtained from the distribution of coefficients in the full model and conditioning on a subset of β being equal to zero (conditionally compatible priors such as the g-prior or Zellner–Siow Cauchy prior), we may incorporate the prior precision Φ in the solution of of the augmented design $\mathbf{X}_\mathbf{0}^T\mathbf{X}_\mathbf{0} + \mathbf{X}_\mathbf{a}^T\mathbf{X}_\mathbf{a} + \Phi = \mathbf{D}$. For mixtures of g-priors, $\Phi = g^{-1}\mathbf{X}_\mathbf{0}^T\mathbf{X}_\mathbf{0}$, so this would lead to updating $\mathbf{X}_\mathbf{a}^T\mathbf{X}_\mathbf{a}$ after sampling g. Similarly for the $p > n$ case, the incorporating the prior precision into the solution for $\mathbf{X}_\mathbf{a}$ will lead to a full rank solution.

Data augmentation with parameter expansion has been shown to improve convergence in many cases (Liu and Wu 1999; Meng and van Dyk 1999; Hobert and Marchev 2008). We experimented with several parameter expansion schemes based on scale changes for the latent data (not reported here), but did not see any improvement. Liu and Wu (1999) suggest that if the set of transformation forms a locally compact group and the prior on the expansion parameter corresponds to Haar measure, that there is a well-defined parameter expansion algorithm that is optimal in terms of convergence. For ODA a natural choice that preserves the invariance of the augmented data is to consider the group of orthogonal rotations $\mathcal{O}_{p+1}$ as the expansion parameter. Generating random orthogonal matrices, however, will increase computational complexity of the algorithm; standard algorithms for generating orthogonal matrices are of order $O(p^3)$ although Genz (1998) describes methods using butterfly matrices that are of order $O(\log(p)p^2)$. It remains to be seen if improvements in convergence offset computational requirements, however, by allowing the augmented design matrix to be random, we may be able to further reduce the leverage of the augmented cases.

SUPPLEMENTARY MATERIALS

Algorithm, proof, variable key: Detailed algorithm for ODA–Probit, proof of proposition for ODA–Cauchy and ODA–Probit, and variable key for Pima Indians Diabetes dataset. (supp-rev.pdf)

R Code: R code for the ODA algorithm. (code.zip)

[Received August 2010. Revised December 2010.]

REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [1050]

Berger, J. O., and Molina, G. (2005), "Posterior Model Probabilities via Path-Based Pairwise Priors," *Statistica Neerlandica*, 59, 3–15. [1041]

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998), "Bayes Factors and Marginal Distributions in Invariant Situations," *Sankhya, Ser. A*, 60, 307–321. [1043]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [1042,1048,1049]

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997), "Adaptive Bayesian Wavelet Shrinkage," *Journal of the American Statistical Association*, 92, 1413–1421. [1042]

Clyde, M., and George, E. I. (1999), "Empirical Bayes Estimation in Wavelet Nonparametric Regression," in *Bayesian Inference in Wavelet Based Models*, New York: Springer-Verlag, pp. 309–322. [1042]

——— (2000), "Flexible Empirical Bayes Estimation for Wavelets," *Journal of the Royal Statistical Society, Ser. B*, 62, 681–698. [1042,1048]

——— (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94. [1041]

Clyde, M., and Parmigiani, G. (1996), "Orthogonalizations and Prior Distributions for Orthogonalized Model Mixing," in *Modelling and Prediction Honoring Seymour Geisser*, New York: Springer-Verlag. [1042]

Clyde, M., DeSimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208. [1042]

Clyde, M., Ghosh, J., and Littman, M. (2011), "Bayesian Adaptive Sampling for Variable Selection and Model Averaging," *Journal of Computational and Graphical Statistics*, 20, 80–101. [1041,1046,1048]

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998), "Multiple Shrinkage and Subset Selection in Wavelets," *Biometrika*, 85, 391–401. [1042,1043]

Clyde, M. A. (2010), "BAS: Bayesian Adaptive Sampling for Bayesian Model Averaging," R package version 0.90. [1048]

Clyde, M. A., and Ghosh, J. (2010), "A Note on the Bias in Estimating Posterior Probabilities in Variable Selection," Technical Report 2010-11, Duke University. [1041,1047]

Dempster, A., Laird, N. M., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–22. [1042]

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society, Ser. B*, 60, 333–350. [1046]

Draper, D. (1995), "Assessment and Propagation of Model Uncertainty" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 57, 45–70. [1041]

Eaton, M. L. (1983), *Multivariate Statistics: A Vector Space Approach*, New York: Wiley. [1046]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [1049]

Fernández, C., Ley, E., and Steel, M. F. (2001), "Benchmark Priors for Bayesian Model Averaging," *Journal of Econometrics*, 100, 381–427. [1046]

Fouskakis, D., Ntzoufras, I., and Draper, D. (2009), "Bayesian Variable Selection Using Cost-Adjusted BIC, With Application to Cost-Effective Measurement of Quality of Health Care," 3, 663–690. [1041]

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling," *Technometrics*, 31, 3–39. [1042,1049]

Fulton, W. (2000), "Eigenvalues, Invariant Forms, Highest Weights, and Schubert Calculus," *Bulletin of the American Mathematical Society*, 37, 209–249. [1045]

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [1044]

Genz, A. (1998), "Methods for Generating Random Orthogonal Matrices," in *Monte Carlo and Quasi-Monte Carlo Methods*, eds. H. Niederreiter and J. Spanier, Berlin: Springer-Verlag, pp. 199–213. [1051]

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–374. [1045,1047]

Geyer, C. J. (1995), "Conditioning in Markov Chain Monte Carlo," *Journal of Computational and Graphical Statistics*, 4, 148–154. [1044]

Gramacy, R. B. (2010), "monomvn: Estimation for Multivariate Normal and Student-t Data With Monotone Missingness," R package version 1.8-3. [1049]

Hastie, T., and Efron, B. (2007), "lars: Least Angle Regression, Lasso and Forward Stagewise," R package version 0.9-7. [1049]

Heaton, M., and Scott, J. (2010), "Bayesian Computation and the Linear Model," in *Frontiers of Statistical Decision Making and Bayesian Analysis*, eds. M.-H. Chen, D. K. Dey, P. Mueller, D. Sun, and K. Ye, New York: Springer. [1041]

Hobert, J. P., and Marchev, D. (2008), "A Theoretical Comparison of the Data-Augmentation, Marginal Augmentation and PX–DA Algorithms," *The Annals of Statistics*, 36, 532–554. [1051]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial" (with discussion), *Statistical Science*, 14, 382–401; corrected version available at *http://www.stat.washington.edu/www/research/online/hoeting1999.pdf*. [1041]

Holmes, C. C., and Held, L. (2006), "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression," *Bayesian Analysis*, 1, 145–168. [1050]

Ishwaran, H., and Rao, J. (2005), "Spike and Slab Variable Selection: Frequentist and Bayesian Strategies," *The Annals of Statistics*, 33, 730–773. [1041]

Jeffreys, H. (1961), *Theory of Probability*, Oxford: Oxford University Press. [1047]

Johnstone, I., and Silverman, B. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *The Annals of Statistics*, 32, 1594–1649. [1048]

——— (2005), "Empirical Bayes Selection of Wavelet Thresholds," *The Annals of Statistics*, 33, 1700–1752. [1042,1048]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795. [1046]

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008), "Mixtures of *g*-Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. [1041-1043,1048,1049]

Liu, J. S. (1994), "The Collapsed Gibbs Sampler in Bayesian Computations With Application to a Gene Regulation Problem," *Journal of the American Statistical Association*, 89, 958–966. [1044,1045]

Liu, J. S., and Wu, Y. N. (1999), "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274. [1051]

Liu, J. S., Wong, W. H., and Kong, A. (1994), "Covariance Structure of the Gibbs Sampler With Applications to the Comparisons of Estimators and Augmentation Schemes," *Biometrika*, 81, 27–40. [1044,1045]

Meng, X. L., and van Dyk, D. A. (1999), "Seeking Efficient Data Augmentation Schemes via Conditional and Marginal Augmentation," *Biometrika*, 86, 301–320. [1051]

Nott, D. J., and Green, P. J. (2004), "Bayesian Variable Selection and the Swendsen–Wang Algorithm," *Journal of Computational and Graphical Statistics*, 13, 141–157. [1046]

Nott, D. J., and Kohn, R. (2005), "Adaptive Sampling for Bayesian Variable Selection," *Biometrika*, 92, 747–763. [1042,1046]

Polson, N. G., and Scott, J. G. (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," in *Bayesian Statistics 9*, eds. J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, Oxford: Oxford University Press. [1048]

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191. [1046]

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge: Cambridge University Press. [1050]

Scott, J. G., and Carvalho, C. M. (2008), "Feature-Inclusion Stochastic Search for Gaussian Graphical Models," *Journal of Computational and Graphical Statistics*, 17, 790–808. [1041]

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82, 528–540. [1042]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288. [1042,1049]

Weyl, H. (1912), "Das asymptotische Verteilungsgesetz der Eigenwerte lineare partieller Differentialgleichungen," *Mathematische Annalen*, 71, 441–479. [1045]

Zellner, A. (1984), "Posterior Odds Ratios for Regression Hypotheses: General Considerations and Some Specific Results," in *Basic Issues in Econometrics*, ed. A. Zellner, Chicago: University of Chicago Press, pp. 275–305. [1048]

——— (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*, North-Holland/Elsevier, pp. 233–243. [1041,1048]

Zellner, A., and Siow, A. (1980), "Posterior Odds Ratios for Selected Regression Hypotheses," in *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia (Spain)*, Valencia, Spain: University of Valencia Press, pp. 585–603. [1041,1042,1048]