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

Supplemental Materials

1 ODA-Probit Sampler

The ODA-Probit sampler for probit regression, introduced in Section 7, may be constructed as follows:

1. $\boldsymbol{Y_a} \mid \boldsymbol{\gamma}, \boldsymbol{Y_o}, \boldsymbol{Z_o} \sim \mathsf{N}\left(\boldsymbol{X_{a\gamma}}\widetilde{\boldsymbol{\beta}}_{\boldsymbol{o\gamma}}, (\mathbf{I}_{n_a} + \boldsymbol{X_{a\gamma}}(\boldsymbol{X}_{\boldsymbol{o\gamma}}^T\boldsymbol{X_{o\gamma}} + \Lambda_{\gamma})^{-1}\boldsymbol{X}_{\boldsymbol{a\gamma}}^T)\right)$, where

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

2. $\boldsymbol{\gamma}_j \mid \boldsymbol{Y_c}, \boldsymbol{Z_o} \overset{ind}{\sim} \mathsf{Ber}(\rho_j(\boldsymbol{Y_c}, 1, \lambda_j))$ for $j = 1, 2, \ldots, p$, where

$$\rho_j(\boldsymbol{Y_c}, 1, \lambda_j) \equiv O_j(\boldsymbol{Y_c}, 1, \lambda_j)/\left(1 + \mathbf{O}_j(\boldsymbol{Y_c}, 1, \lambda_j)\right)$$

$$O_j(\boldsymbol{Y_c}, 1, \lambda_j) \equiv \frac{\pi_j}{1 - \pi_j}\left(\frac{\lambda_j}{\delta_j + \lambda_j}\right)^{1/2} \exp\left\{\frac{1}{2}\,\frac{\delta_j}{\delta_j + \lambda_j}\,\widehat{\beta}_j^2\delta_j\right\} \text{ for } j = 1, \ldots, p$$

3. $\beta_j \mid \boldsymbol{Y_c}, \boldsymbol{\gamma}, \boldsymbol{Z_o} \overset{ind}{\sim} \mathsf{N}\left(\widehat{\beta}_j\gamma_j\frac{\delta_j}{\lambda_j + \delta_j}, \gamma_j(\lambda_j + \delta_j)^{-1}\right)$, where

$$\widehat{\beta}_j = \boldsymbol{x}_{\boldsymbol{c}j}^T\boldsymbol{Y_c}/(\boldsymbol{x}_{\boldsymbol{c}j}^T\boldsymbol{x}_{\boldsymbol{c}j}), \text{ for } j = 1, 2, \ldots p$$

4. $Y_{\boldsymbol{oi}} \mid \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{Y_a}, Z_{\boldsymbol{oi}} = \begin{cases} \mathsf{N}_+(\boldsymbol{x}_{\boldsymbol{\gamma}i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1) & \text{if} \quad Z_{\boldsymbol{oi}} = 1 \\ \mathsf{N}_-(\boldsymbol{x}_{\boldsymbol{\gamma}i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1) & \text{if} \quad Z_{\boldsymbol{oi}} = 0 \end{cases}$

   for $i = 1, 2, \ldots, n_o$, where N_+ and N_- denote truncated normal densities, truncated to be positive and negative, respectively.

2 Theoretical Validity of ODA-Cauchy and ODA-Probit

We now show theoretical support for validity of the ODA sampler for the scale mixture of normal distributions and probit regression.

Given a joint distribution $\pi(dw, dx, dy, dz)$ let $\{x^{(i)}, y^{(i)}, z^{(i)}\}$ be a Markov chain constructed as follows. A single iteration of the algorithm $\{(x^T, y^T, z^T) \to (x, y, z)\}$ consists of the 4 steps:

1. $x \sim \pi(x \mid y^T, z^T)$
2. $y \sim \pi(y \mid x, z^T)$
3. $w \sim \pi(w \mid x, y, z^T)$
4. $z \sim \pi(z \mid x, y, w)$

Proposition 1. *$\pi(dx, dy, dz) = \int_w \pi(dx, dy, dz, dw)$ is the stationary distribution of the above Markov chain.*

Proof. Let $K(y^T, z^T; dx)$, $K(x, z^T; dy)$, $K(x, y, z^T; dw)$, $K(x, y, w; dz)$ denote the corresponding transition kernels.

$$\begin{aligned}
& \int_{x^T}\int_{y^T}\int_{z^T}\int_{w} \pi(dx^T, dy^T, dz^T)K(y^T, z^T; dx)K(x, z^T; dy)K(x, y, z^T; dw)K(x, y, w; dz) \\
= & \int_{y^T}\int_{z^T}\int_{w} \pi(dx, dy^T, dz^T)K(x, z^T; dy)K(x, y, z^T; dw)K(x, y, w; dz) \\
= & \int_{z^T}\int_{w} \pi(dx, dy, dz^T)K(x, y, z^T; dw)K(x, y, w; dz) \\
= & \int_{w} \pi(dx, dy, dw)K(x, y, w; dz) \\
= & \pi(dx, dy, dz)
\end{aligned}$$

□

ODA for Scale Mixtures of Normals

In this case the parameters of interest $\{x, y, z\}$ are $\{(\sigma^2, \boldsymbol{Y_a}), \boldsymbol{\gamma}, \boldsymbol{\lambda}\}$; the extra parameter $w = \boldsymbol{\beta}$ is introduced because one cannot draw $\boldsymbol{\lambda}$ directly from the distribution $p(\boldsymbol{\lambda} \mid \sigma^2, \boldsymbol{Y_a}, \boldsymbol{\gamma})$. All distributions are considered to be conditional on the observed data $\boldsymbol{Y_o}$. Thus it immediately follows that the ODA algorithm constructed for the scale mixture of normal priors converges to the correct posterior distribution.

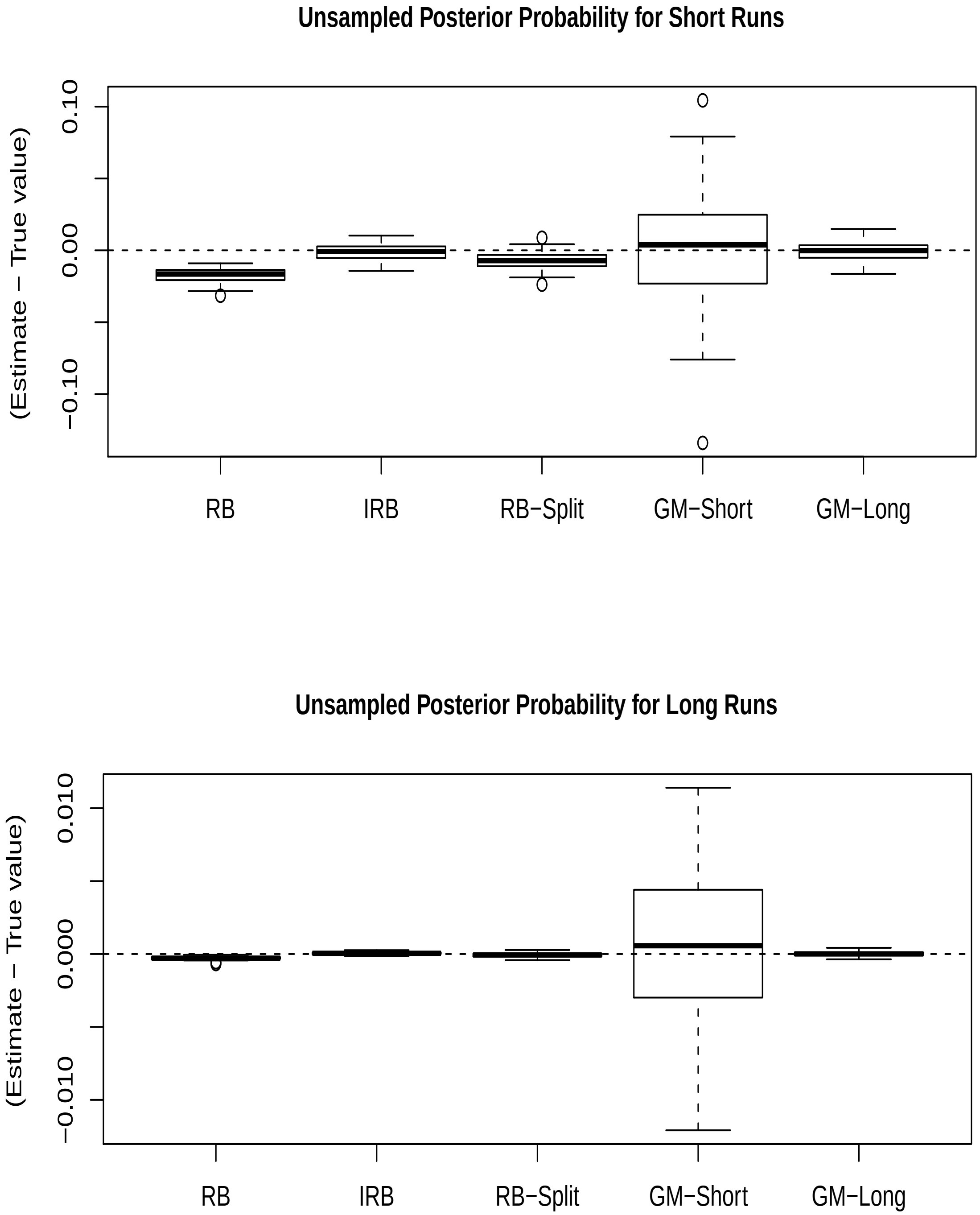
ODA for Probit Regression

The above proof can be also used to show the validity of the ODA sampler for probit regression. Here the parameters of interest $\{x, y, z\}$ are $\{\boldsymbol{Y_a}, \boldsymbol{\gamma}, \boldsymbol{Y_o}\}$; the extra parameter $w = \boldsymbol{\beta}$ is introduced because it is difficult to draw $\boldsymbol{Y_o}$ directly from the distribution $p(\boldsymbol{Y_o} \mid \boldsymbol{Y_a}, \boldsymbol{\gamma})$. All distributions are considered to be conditional on the observed binary data $\boldsymbol{Z_o}$.

3 Estimation of Unsampled Posterior Mass

We compute the Rao-Blackwell (RB) estimate (equation (15) in the manuscript) to estimate the remaining posterior mass in each of the 100 runs of ODA for the Nott-Kohn simulations and compare this to the exact value under enumeration. Figure 1 shows boxplots of (estimate - true value) for each estimator for short and long runs of ODA. From these simulations and other comparisons (not reported) we have found that the usual RB estimate of posterior probability of unsampled models for ODA, shows a slight tendency towards underestimation. This is not surprising because the sample of $(\boldsymbol{Y_a}, \sigma^2)$ used to calculate the estimate has been generated from a chain that visited these models at least once, and thus favor them over the unsampled models. This leads to an overestimation of the probabilities of sampled models and hence an underestimation of unsampled ones. Although the bias is negligible (even in the short run the mass is estimated within ± 0.018), we show two possible alternatives to correct it. One option is to run an independent Markov chain, and calculate the estimate (in equation (15) of the manuscript) based on $(\boldsymbol{Y_a}, \sigma^2)$ generated from this new chain. We shall 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 $(\boldsymbol{Y_a}, \sigma^2)$ 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. George and McCulloch (1997) constructed an estimate of the normalizing constant C for the posterior distribution of $\boldsymbol{\gamma}$, $p(\boldsymbol{\gamma} \mid \boldsymbol{Y_o}) = Cp(\boldsymbol{Y_o} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma} \mid \boldsymbol{\pi})$. They run a preliminary independent MCMC, to obtain an initial set A of $\boldsymbol{\gamma}$ and show that $\hat{C} = [\sum_{i=1}^{K} \mathbf{1}(\boldsymbol{\gamma}^{(k)} \in A)]/[K \sum_{\boldsymbol{\gamma} \in \mathbf{A}} p(\boldsymbol{Y_o} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma} \mid \boldsymbol{\pi})]$ is a consistent estimate of C. An estimate of the unsampled mass is then $1 - \hat{C} \sum_{\boldsymbol{\gamma} \in \boldsymbol{\Gamma}_K} p(\boldsymbol{Y_o} \mid \boldsymbol{\gamma})p(\boldsymbol{\gamma} \mid \boldsymbol{\pi})$. They suggest a short preliminary run of 100 iterations, which is labeled GM Small in the figure. They suggest that the accuracy of the estimator will increase when $p(A \mid \boldsymbol{Y_o})$ is large; we consider a larger set A based on an independent run of the same length as the original Markov chain, which is reported under GM Large. From the boxplots in Figure 1 it is evident that the bias of all estimators is negligible, with GM Small exhibiting the most variability in both short and long runs. IRB and GM Large are 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. Note that the GM method can be implemented only when the marginal likelihood is available in closed form, as in the case of the normal errors with conjugate normal priors on the regression coefficients. Thus unlike ODA, GM cannot be used to estimate the unsampled probability for more general scenarios such as for heavy-tailed Student-t priors for the regression coefficients in linear regression, or for generalized linear models like probit regression. For enumerable problems where marginal likelihoods are not available in closed form, ODA clearly has an advantage over GM in estimating not only the unsampled probability but also model probabilities for all models, as we illustrate with probit regression in Section 7 of the manuscript.



Unsampled Posterior Probability for Short Runs

Unsampled Posterior Probability for Long Runs

Figure 1: Short and long runs of Nott-Kohn example: Comparison of RB (Rao-Blackwell), IRB (Independent Rao-Blackwell), RB-Split (Rao-Blackwell Split), and GM (George and McCulloch) estimates of unsampled posterior probability, for normal prior distributions on regression coefficients; results are based on 100 replicates of ODA

4 Sensitivity Analysis for Choice of $\boldsymbol{X_a}$

We have used the Cholesky decomposition to generate the augmented design matrix, $\boldsymbol{X_a}$, for the ODA algorithms considered in the manuscript. We have used Cholesky because it is much faster than generating design matrices based on singular value decomposition (SVD). Currently the difference in CPU times is negligible, as both are using C/FORTRAN routines called from R. We plan to re-write the R code for ODA in C/FORTRAN to speed up calculations and we think in that case using Cholesky will be more advantageous (in terms of speed), especially if we consider algorithms where we need to generate $\boldsymbol{X_a}$ in every iteration. We now perform a sensitivity analysis for the results obtained for different choices of $\boldsymbol{X_a}$. We compare the results based on Cholesky reported in the manuscript to two other augmented designs based on SVD:
i) A symmetric $\boldsymbol{X_a}$ given by $\boldsymbol{X_a} = \boldsymbol{U}\boldsymbol{\alpha}^{1/2}\boldsymbol{U}^T$ as described in Theorem 2 of the paper
ii) A non-symmetric $\boldsymbol{X_a}$ given by $\boldsymbol{X_a} = \boldsymbol{\alpha}^{1/2}\boldsymbol{U}^T$, where $\boldsymbol{U}$ and $\boldsymbol{\alpha}$ are defined as before. We will refer to these three choices as Cholesky, SVD-S (symmetric SVD) and SVD-NS (non-symmetric SVD).

We run ODA for the Nott-Kohn simulation example with the two choices SVD-S and SVD-NS, for 6,400 draws, and repeat it 100 times, so that our results are comparable to those from the short runs of ODA (Cholesky) reported in Table 1 of the manuscript. It is evident from Figures 2 and 3 that the results for the three choices of design matrices are very similar, implying that the ODA algorithm is not sensitive to the choice of $\boldsymbol{X_a}$.

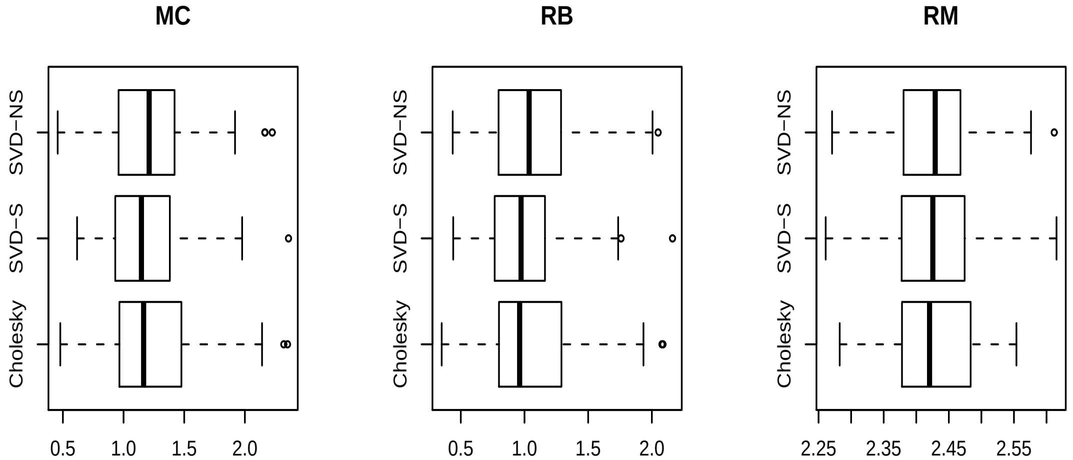


Figure 2: Nott-Kohn simulation example: Comparison of square root of the mean squared error ×100 for estimation of the 15 posterior inclusion probabilities using Monte Carlo frequencies (MC), Rao-Blackwellized (RB) and renormalized marginal likelihoods (RM) under short runs (1 minute) of ODA with the augmented design matrix $\boldsymbol{X_a}$ obtained by i)Cholesky ii)symmetric SVD (SVD-S) and iii)non-symmetric SVD (SVD-NS). The boxplots are based on 100 replicates for each choice of $\boldsymbol{X_a}$.

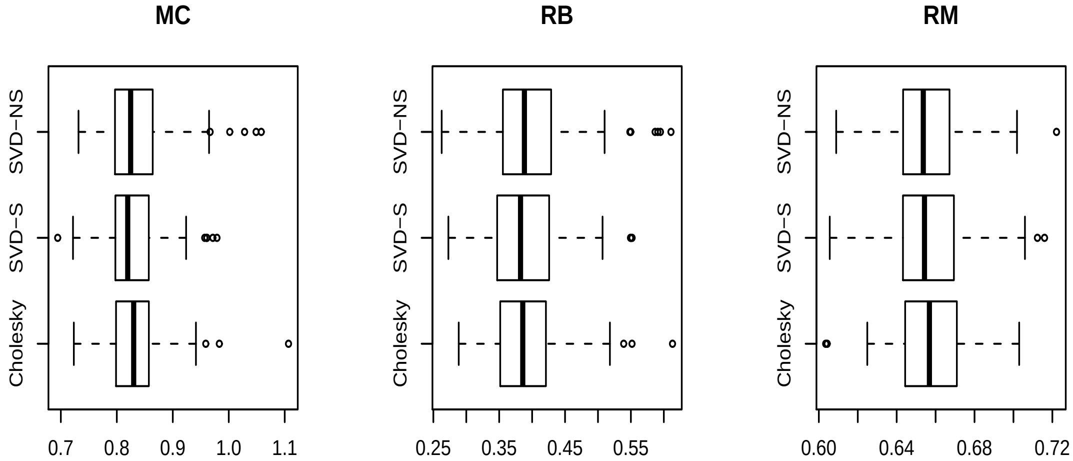


Figure 3: Nott-Kohn simulation example: Comparison of square root of the mean squared error $\times 10^4$ for estimation of the 2^{15} posterior model probabilities using Monte Carlo frequencies (MC), Rao-Blackwellized (RB) and renormalized marginal likelihoods (RM) under short runs (1 minute) of ODA with the augmented design matrix $\boldsymbol{X_a}$ obtained by i)Cholesky ii)symmetric SVD (SVD-S) and iii)non-symmetric SVD (SVD-NS). The boxplots are based on 100 replicates for each choice of $\boldsymbol{X_a}$.

5 Variable Key for Pima Indians Diabetes Data

npreg	-	number of pregnancies
glu	-	plasma glucose concentration in an oral glucose tolerance test
bp	-	diastolic blood pressure (mm Hg)
skin	-	triceps skin fold thickness (mm)
bmi	-	body mass index (weight in kg/(height in m)^2)
ped	-	diabetes pedigree function
age	-	age in years
type	-	"Yes" or "No" for diabetic according to WHO criteria

References

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