Secure Bayesian Model Averaging for Horizontally Partitioned Data

Joyee Ghosh[*]and Jerome P. Reiter[†]

Abstract

When multiple data owners possess records on different subjects with the same set of attributes—known as horizontally partitioned data—the data owners can improve analyses by concatenating their databases. However, concatenation of data may be infeasible because of confidentiality concerns. In such settings, the data owners can use secure computation techniques to obtain the results of certain analyses on the integrated database without sharing individual records. We present secure computation protocols for Bayesian model averaging and model selection for both linear regression and probit regression. Using simulations based on genuine data, we illustrate the approach for probit regression, and show that it can provide reasonable model selection outputs.

Key Words: Bayesian model averaging; Data confidentiality; Disclosure limitation; Markov chain Monte Carlo; Regression; Variable selection.

1 Introduction

In many contexts, data owners can improve statistical analysis by concatenating records from different databases. For example, genomics researchers at different universities or labs may seek to combine their small samples, so as to increase the precision of statistical models; local educational agencies might want to combine their students' data to have greater representation of the general student population; and, several companies may seek to pool data on their customers to improve marketing analyses. These settings are examples of horizontally partitioned data, i.e., multiple data owners possess records on different subjects with the same set of attributes. Often, however, the data owners are not willing or even legally permitted to combine their databases because of concerns about the confidentiality of data subjects' identities and sensitive attributes. These concerns can persist even after stripping records of unique identifiers like names, addresses, and tax identification numbers (Willenborg and de Waal, 2001).

[*] Joyee Ghosh is Assistant Professor, Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA 52242. Email `joyee-ghosh@uiowa.edu`

[†] Jerome P. Reiter is Associate Professor, Department of Statistical Science, Duke University, Durham, NC 27708. Email `jerry@stat.duke.edu`

Recently, researchers in computer science and statistical science have developed techniques that allow data owners to perform analyses on the combined data without actually sharing individual records. These techniques include secure linear regression analyses (Du *et al.*, 2004; Karr *et al.*, 2005), secure data mining with association rules (Kantarcioglu and Clifton, 2002; Vaidya and Clifton, 2002; Evfimievski *et al.*, 2004), secure model based clustering (Vaidya and Clifton, 2003; Lin *et al.*, 2005), secure logistic regression (Slavkovic *et al.*, 2007), and secure adaptive regression splines (Ghosh *et al.*, 2007). The literature on privacy-preserving data mining (Agrawal and Srikant, 2000; Lindell and Pinkas, 2000) contains related results.

In this article, we develop secure Bayesian model selection and model averaging for normal linear regression and probit regression with horizontally partitioned data. In Section 2, we review Bayesian model choice and secure computation protocols. In Section 3 and 4, we present the protocols for secure Bayesian model selection and model averaging for linear regression and probit regression, respectively. In Section 5, we apply the protocols for secure model selection for probit regression to data on the predictors of abnormal heart conditions. In Section 6, we conclude with a brief summary.

2 Background: Bayesian Model Choice and Secure Computation

2.1 Bayesian Model Choice

To fix ideas of Bayesian model averaging (BMA) and model selection, we present an overview of BMA for normal linear regression. Let $\mathbf{Y} = (Y_1, \ldots Y_n)'$ be a vector of univariate response variables, and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ be the set of corresponding p covariates. Let $\boldsymbol{\gamma} =$ $(\gamma_1, \ldots \gamma_p)^T \in \Gamma$ be a vector of indicator variables, such that $\gamma_j = 1$ when $\boldsymbol{x}_j$ is included as a column in the design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ and $\gamma_j = 0$ otherwise. Thus, all 2^p models can be represented by the possible configurations of $\boldsymbol{\gamma}$. For any model $\boldsymbol{\gamma}$, the linear regression of $\mathbf{Y}$ on $\mathbf{X}$ is

$$\mathbf{Y} \mid \mathbf{X}, \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \boldsymbol{\gamma} \sim \mathsf{N}(\mathbf{1}\alpha + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{I}_n/\phi), \tag{1}$$

where $\mathbf{1}$ is a vector of ones of length n, α is the intercept, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the vector of regression coefficients, ϕ is the precision parameter (reciprocal of the error variance), and $\mathbf{I}_n$ is an $n \times n$ identity matrix.

After specifying the choice of prior probabilities for models, $p(\boldsymbol{\gamma})$, and prior distributions for model-specific parameters, $p(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})$, the marginal likelihood of $\mathbf{Y}$ under model $\boldsymbol{\gamma}$ is obtained by integrating the likelihood of $\mathbf{Y}$ with respect to the prior distribution of model-specific parameters; that is,

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma}) = \int p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}_{\boldsymbol{\gamma}} \mid \mathbf{X}, \boldsymbol{\gamma}) d\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \tag{2}$$

where $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$. The posterior probability of any model $\boldsymbol{\gamma}$ is given by

$$p(\boldsymbol{\gamma} \mid \mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma} \in \Gamma} p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma})}. \tag{3}$$

For model selection, one can choose the $\boldsymbol{\gamma}$ with the highest posterior probability, the median probability model of Barbieri and Berger (2004) which includes all covariates $\boldsymbol{x}_j$ with $p(\gamma_j = 1 \mid \mathbf{X}, \mathbf{Y}) \geq 0.5$, or a set of high probability models.

An alternative to relying on a single model or a set of high probability models is to carry out inference and predictions based on all models using BMA. Under BMA, the posterior distribution of any quantity of interest is given by a weighted average of model-specific posterior distributions, with weights determined by the posterior probabilities of models. The importance of variable $\boldsymbol{x}_j$ is often summarized by its marginal posterior inclusion probability,

$$p(\gamma_j = 1 \mid \mathbf{X}, \mathbf{Y}) = \sum_{\boldsymbol{\gamma} \in \Gamma : \gamma_j = 1} p(\boldsymbol{\gamma} \mid \mathbf{X}, \mathbf{Y}). \tag{4}$$

For regression problems with 20 to 25 explanatory variables, exhaustive exploration of the model space is possible, if exact or approximate marginal likelihoods are available (Clyde *et al.*, 2011). When marginal likelihoods are intractable, one can use methods based on MCMC (Chib, 1995; Carlin and Chib, 1995) or those based on importance sampling, bridge sampling, and path sampling (Meng and Wong, 1996; Gelman and Meng, 1998). For higher dimensional model spaces, typically model search algorithms are used with the idea of exploring promising regions of the space, and BMA is based on the subset of visited models instead of the model space in its entirety.

For further reviews of BMA and Bayesian variable selection, see Hoeting *et al.* (1999); Dellaportas *et al.* (2002); Clyde and George (2004) and the references therein. Heaton and Scott (2010) describe recent developments in computational strategies for BMA in the context of linear models.

2.2 Secure Summation

Most secure summation protocols assume that the participating parties are "semi-honest:" each follows the agreed-on computational protocols properly, but may retain the results of intermediate computations. This assumption rules out scenarios where one or more parties abuse the protocol to learn information about another party's individual data values. We assume that the results of analyses of horizontally partitioned data are shared among all participating data owners and possibly disseminated to the broader public.

Following the presentations in Karr *et al.* (2005) and in Ghosh *et al.* (2007), consider $K > 2$ semi-honest data owners, such that owner a has a value v_a. The owners wish to compute $v = \sum_{a=1}^{K} v_a$ so that each owner a learns only the minimum possible about the other owners' values, namely the value of $v_{(-a)} = \sum_{\ell \neq a} v_\ell$. A variant of secure summation protocols (Benaloh, 1987) can be used to perform this computation.

One owner is designated the master owner and numbered 1. The remaining owners are numbered $2, \ldots, K$. To begin, owner 1 generates a random number R from a uniform distribution $(-m, m)$, where m is much larger than $|v_a|$ for all a and not revealed to other owners. Owner 1 adds R to its local value v_1 and sends $s_1 = (R + v_1)$ to owner 2. Since R is chosen randomly owner 2 learns essentially nothing about the actual value of v_1.

For the remaining owners $a = 2, \ldots, K - 1$, the algorithm is as follows. Owner a receives $s_{a-1} = (R + \sum_{t=1}^{a-1} v_t)$ from which it learns essentially nothing about the actual values of $v_1, \ldots, v_{a-1}$. Owner a then computes and passes on to owner $a+1$ the quantity $s_a = (s_{a-1} + v_a) = (R + \sum_{t=1}^{a} v_t)$. Finally, owner K adds v_K to s_{K-1}, and sends the result s_K to owner 1. Owner 1, which knows R, then calculates v by subtraction, $v = (s_K - R)$ and shares this value with the other owners.

Many secure summation techniques assume that owners know $0 < v < m$ and use arithmetic mod m when passing sums. The arithmetic mod m reduces the amount of information leaked about R in the protocol and helps protect v_1 in particular. However, in the contexts we care about it is likely that owners will not know if some $v > 0$ or not, and arithmetic mod m does not determine the sign of v. Thus, this adapted protocol sacrifices some security for flexibility.

3 Secure BMA for Linear Regression

To illustrate secure BMA for linear regression, we use a discrete uniform prior distribution for the model space, which assigns equal probabilities to all models,

$$p(\boldsymbol{\gamma}) = \frac{1}{2^p}. \tag{5}$$

The protocol can be easily extended to more general beta-binomial prior distributions on the model space. We first show how to implement secure BMA with the popular Zellner's g-prior (Zellner, 1986; Liang *et al.*, 2008), and then extend the approach to the Zellner-Siow prior (Zellner and Siow, 1980). For these models, the secure BMA on horizontally partitioned data results in mathematically equivalent results as BMA based on the concatenated data.

Throughout the rest of the article, we assume that the data owners have unique records. If this is not reasonable, the data owners can use secure record linkage (Churches and Christen, 2004) to identify and delete duplicates before initiating the protocol. We also assume that owners are willing to share their sample sizes. If needed, the owners could use secure summation to compute the total sample size n.

3.1 Secure BMA with Zellner's g-prior

The Zellner's g-prior specification for model selection typically includes an improper prior distribution for the intercept α and precision ϕ, and a multivariate normal prior distribution for the regression coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in each model $\boldsymbol{\gamma}$, conditional on ϕ, so that

$$\begin{aligned} p(\alpha, \phi \mid \mathbf{X}, \boldsymbol{\gamma}) &\propto 1/\phi \\ \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{X}, \boldsymbol{\gamma}, \phi &\sim \mathrm{N}\left(\mathbf{0}, \frac{g}{\phi}(\boldsymbol{X}_{\boldsymbol{\gamma}}{}'\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}\right). \end{aligned} \tag{6}$$

Several choices for g have been proposed in the literature. We use the unit information prior with $g = n$. Without loss of generality, we assume all columns of $\mathbf{X}$ have mean zero.

The Zellner's g-prior is popular for model selection because it leads to a closed form expression for the marginal likelihood $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma})$ for any model $\boldsymbol{\gamma}$, namely

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma}) \propto (1+g)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}} \times \{1 + g(1 - \mathrm{R}_{\boldsymbol{\gamma}}{}^2)\}^{-\frac{(n-1)}{2}} \mathrm{TSS}^{-\frac{(n-1)}{2}}. \tag{7}$$

Here, $p_{\boldsymbol{\gamma}} = \sum_j \gamma_j$ is the dimension of model $\boldsymbol{\gamma}$, and $\mathrm{TSS} = \sum_{i=1}^n (Y_i - \bar{Y})^2$. For model $\boldsymbol{\gamma}$, $\mathrm{SSR}_{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}' \mathbf{X}_{\boldsymbol{\gamma}}{}' \mathbf{Y}$, where $\widehat{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} = (\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{Y}$ and $\hat{\alpha}$ are the ordinary least squares estimates of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and α, respectively. Finally, the coefficient of determination, $\mathrm{R}_{\boldsymbol{\gamma}}{}^2 = \mathrm{SSR}_{\boldsymbol{\gamma}}/\mathrm{TSS}$. Note that $\mathbf{X}_{\boldsymbol{\gamma}}$ does not include the column of ones corresponding to the intercept. Hence, to calculate $\mathrm{SSR}_{\boldsymbol{\gamma}}$, $n\bar{\mathbf{Y}}^2$ does not need to be subtracted from $\widehat{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}' \mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{Y}$.

For any new design point $\mathbf{X}^\star$, we can predict its corresponding response variable $\mathbf{Y}^\star$ using the expected value under the posterior predictive distribution under BMA. We have

$$E(\mathbf{Y}^\star \mid \mathbf{Y}, \mathbf{X}, \mathbf{X}^\star) \;=\; \sum_{\boldsymbol{\gamma}\in\Gamma} E(\mathbf{Y}^\star \mid \mathbf{Y}, \mathbf{X}, \mathbf{X}^\star, \boldsymbol{\gamma}) p(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}, \mathbf{X}^\star), \tag{8}$$

where $E(\mathbf{Y}^\star \mid \mathbf{Y}, \mathbf{X}, \mathbf{X}^\star, \boldsymbol{\gamma}) = E[\alpha | \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}]\mathbf{1} + \mathbf{X}^\star E[\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}]$, and $E[\alpha | \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}]$ and $E[\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}]$ denote the posterior means of the regression coefficients under model $\boldsymbol{\gamma}$. For the g-prior, the $E[\alpha | \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}] = \bar{\mathbf{Y}}$, and the $E[\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}] = \frac{g}{1+g}\widehat{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}$.

To implement secure BMA with the Zellner's g-prior, we need all the components of (7). Since n is known, the owners can compute the column means of the covariates and $\mathbf{Y}$ via $p+1$ secure summation steps. Given $\bar{Y}$, each owner can compute their share of TSS, and the components can be added via another secure summation. After subtracting off the column means to form their share of the centered design matrix $\mathbf{X}$, the owners need to compute $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$. Since each element of these matrices can be expressed as a sum, the owners can apply element-wise secure summation to obtain the required matrices.

For any model $\boldsymbol{\gamma}$, $\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{X}_{\boldsymbol{\gamma}}$ is the $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\gamma}}$ matrix formed by collecting the rows and columns of $\mathbf{X}'\mathbf{X}$ corresponding to unit entries in the vector $\boldsymbol{\gamma}$. One can similarly obtain $\mathbf{X}_{\boldsymbol{\gamma}}{}'\mathbf{Y}$ by taking entries of $\mathbf{X}'\mathbf{Y}$ corresponding to $\gamma_j = 1$. From these, each owner can compute $\mathrm{SSR}_{\boldsymbol{\gamma}} = \widehat{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}' \mathbf{X}_{\boldsymbol{\gamma}}{}' \mathbf{Y}$. The owners now have all the ingredients to calculate $\mathrm{R}_{\boldsymbol{\gamma}}{}^2$, and hence the marginal likelihood in (7). Given the model-specific marginal likelihoods and posterior means, the owners can perform model averaging, variable selection, or prediction.

For exhaustive model space explorations, the data owners can calculate the marginal likelihoods and predictions for all possible models to carry out exact secure BMA. For larger model spaces where enumeration of the marginal likelihoods of all models is computationally infeasible, BMA over the entire model space can be approximated by BMA over a subset of promising models visited by the model search algorithm under consideration. As the model search algorithm proceeds and a new model $\boldsymbol{\gamma}$ is sampled, the algorithm can calculate its corresponding marginal likelihood simply by taking the appropriate submatrices of $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{Y}$. The model search does not require extra rounds of secure summation.

3.2 Secure BMA with Zellner-Siow Prior

Zellner's g-prior is attractive for variable selection problems because it lends itself to exact marginal likelihood calculations. However, it suffers from what is commonly referred to as the "information paradox." The Bayes factor for comparing two models $\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}^*$ is defined as the ratio of the corresponding marginal likelihoods, $p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma})/p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma}^*)$. If $R^2_{\boldsymbol{\gamma}}$ for a model $\boldsymbol{\gamma}$ goes to 1, indicating that there is overwhelming evidence in the data in favor of the model, one would hope that the Bayes factor for the model $\boldsymbol{\gamma}$ with respect to the null model (which includes only the intercept) goes to infinity. However, it can be shown that the Bayes factor goes to a constant as $R^2_{\boldsymbol{\gamma}} \to 1$ (Zellner, 1986; Berger and Perichhi, 2001).

The Zellner-Siow prior (Zellner and Siow, 1980) is a modification of the g-prior that does not exhibit the information paradox (Liang *et al.*, 2008). It treats g as random, assigning an Inverse-Gamma prior distribution on g with shape parameter $1/2$ and scale parameter $n/2$. This corresponds to a multivariate Cauchy prior distribution on the regression coefficients. The marginal likelihood associated with this prior distribution is given by

$$p(\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\gamma}) \propto \mathrm{TSS}^{-\frac{(n-1)}{2}} \times \int_0^\infty (1+g)^{\frac{n-p_{\boldsymbol{\gamma}}-1}{2}} \{1 + g(1 - \mathrm{R}_{\boldsymbol{\gamma}}{}^2)\}^{-\frac{(n-1)}{2}} p(g) dg, \tag{9}$$

where $p(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} \frac{e^{-n/2g}}{g^{3/2}}$ corresponds to the prior on g. Unfortunately, this is not a closed form expression. However, Liang *et al.* (2008) derive a Laplace approximation (Tierney and Kadane, 1986) for the generic integral

$$\int_0^\infty (1+g)^{(n-p_\gamma-1+2b)/2} \{1 + g(1 - \mathrm{R}_{\boldsymbol{\gamma}}{}^2)\}^{-\frac{(n-1)}{2}} e^{-n/2g} g^{a-3/2} dg. \tag{10}$$

Setting $a = b = 0$ leads to an approximation for the integral in (9).

With the Zellner-Siow prior, we have $E[\alpha|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}] = \bar{\mathbf{Y}}$ and $E[\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}] = E[\frac{g}{g+1}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}]\widehat{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}$. Setting $a = 1, b = -1$ in (10) leads to an expression for $E[\frac{g}{1+g}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\gamma}]$, which in turn can be used for calculating the posterior mean of the regression coefficients and for predictions. The Laplace approximation involves solving a cubic equation in g with coefficients of g involving a, b, n, p_γ and $\mathrm{R}^2_{\boldsymbol{\gamma}}$. All of the component quantities can be obtained via secure summation as in Section 3.1.

4 Secure BMA for Probit Regression

We now describe secure BMA for probit regression. Let Y_i be a binary response variable for the ith data subject, and let $\boldsymbol{X}_i$ be a $p \times 1$ vector denoting the subject's covariates. Throughout this section, the first column of the design matrix is assumed to be a vector of ones corresponding to the intercept. The probit regression for model $\boldsymbol{\gamma}$ is given by

$$P(Y_i = 1 | \boldsymbol{X}_i, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) = \Phi(\boldsymbol{X}_{i\boldsymbol{\gamma}}' \boldsymbol{\beta}_{\boldsymbol{\gamma}}), \tag{11}$$

where Φ is the standard normal cumulative distribution function. We first discuss the formidable operational challenges to exact BMA for probit regression, and then propose an approximate BMA that is more feasible in practice.

4.1 Exact BMA With Data Augmentation

In (11), conjugate prior distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are not available. As a result, expressions for the marginal likelihoods needed for BMA generally are not in closed-form. However, the data augmentation approach of Albert and Chib (1993) enables the use of Gibbs sampling (Gelfand and Smith, 1990). In particular, let $\boldsymbol{Z} = (Z_1, \ldots, Z_n)'$ be a vector of latent variables such that, for all i,

$$Z_i | \boldsymbol{X}_i, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\gamma} \overset{ind}{\sim} \mathrm{N}(\boldsymbol{X}_{i\boldsymbol{\gamma}}' \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1)$$
$$Y_i = 1 \text{ if } Z_i > 0, \; Y_i = 0 \text{ if } Z_i \leq 0.$$

For the prior distribution for $\boldsymbol{\beta}$, we use

$$\beta_j | \gamma_j, \tau_j \quad \overset{ind}{\sim} \quad \mathrm{N}(0, \gamma_j \tau_j), \; j = 1, 2, \ldots p, \tag{12}$$

so that β_j has a degenerate prior distribution with all its mass at zero when $\gamma_j = 0$ and has a $\mathrm{N}(0, \tau_j)$ prior distribution otherwise.

We choose independent prior distributions on β_j primarily for computational convenience, as is often done in practice (Holmes and Held, 2006; Ghosh and Clyde, 2011). It is possible to use dependent priors for the regression coefficients, for example a g-prior with a non-orthogonal design matrix. Unlike for linear models, we are not aware of research indicating beneficial theoretical properties of g-priors for probit regression models. Regardless of the use of independent or multivariate normal priors, marginal likelihoods are not available in closed form in either case.

To illustrate the exact BMA with data augmentation, we use a discrete uniform prior distribution for $\boldsymbol{\gamma}$. Then, conditional distributions for the stochastic search Gibbs sampler (George and McCulloch, 1993) for drawing samples from the posterior distribution
$p(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{Z}|\boldsymbol{Y}, \mathbf{X})$ are as follows. For $i = 1, \ldots, n$, we have

$$Z_i|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{Y}, \mathbf{X} \sim \begin{cases} \mathrm{N}_+(\boldsymbol{X}_{i\boldsymbol{\gamma}}'\boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1) & \text{if } \mathrm{Y_i} = 1 \\ \mathrm{N}_-(\boldsymbol{X}_{i\boldsymbol{\gamma}}'\boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1) & \text{if } \mathrm{Y_i} = 0 \end{cases} \tag{13}$$

where N_+ and N_- denote normal densities truncated to be positive and negative, respectively. Let $\boldsymbol{\gamma}_{-j} = (\gamma_1, \ldots \gamma_{j-1}, \gamma_{j+1}, \ldots \gamma_p)'$, and $\boldsymbol{\beta}_{-j} =$
$(\beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots \beta_p)'$. For $j = 1, \ldots, p$, we have

$$\gamma_j|\boldsymbol{\gamma}_{-j}, \boldsymbol{\beta}_{-j}, \boldsymbol{Z}, \boldsymbol{Y}, \mathbf{X} \sim \mathrm{Ber}(\rho_j) \tag{14}$$

$$\beta_j|\boldsymbol{\gamma}, \boldsymbol{\beta}_{-j}, \boldsymbol{Z}, \boldsymbol{Y}, \mathbf{X} \sim \mathrm{N}(\gamma_j\widehat{\mu}_j, \gamma_j\widehat{\tau}_j), \tag{15}$$

where $\widehat{\tau}_j = (\tau_j^{-1} + \sum_{i=1}^n x_{ij}^2)^{-1}$, $\widehat{\mu}_j = \widehat{\tau}_j \sum_{i=1}^n x_{ij} Z_i^*$, and $Z_i^* = Z_i - \sum_{l \neq j} x_{il}\beta_l$. Here, $\mathrm{N}(u; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2 evaluated at u, $\rho_j = \mathrm{N}(0; 0, \tau_j)/\{\mathrm{N}(0; \widehat{\mu}_j, \widehat{\tau}_j) + \mathrm{N}(0; 0, \tau_j)\}$,

What steps are needed to implement the secure Gibbs sampler in horizontally partitioned data? Each data owner needs to sample the values of Z_i in its database from (13), which is trivial given $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. To sample $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$, the owners must securely compute all $\hat{\tau}_j$ and $\hat{\mu}_j$. Each $\hat{\tau}_j$ can be computed with a single round of secure summation. However, for each cycle of the Gibbs sampler, the owners must use a new round of secure summation to compute each $\hat{\mu}_j$. These depend on $\boldsymbol{Z}$ and $\boldsymbol{\beta}$, which change with each cycle. Furthermore, the data owners must use common draws of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ at any cycle. This can be done by designating one owner to sample and share values of these parameters at each cycle.

Since Gibbs samplers typically require thousands of cycles to ensure convergence, an exact secure BMA for probit regression requires a very large number of secure computation steps. This is computationally expensive and, arguably, impractical for many problems. Furthermore, the large number of intermediate results increases the risk of a breach in data confidentiality.

4.2 Secure BMA Using Normal Approximation

To get around the computational challenges of exact BMA for probit regression, we propose to use a normal approximation to the likelihood. For $i = 1, \ldots, n$, we use

$$Y_i \mid \boldsymbol{\gamma}, \mathbf{X}_i, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi \sim \mathrm{N}(\mathbf{X}_{i\boldsymbol{\gamma}}{}' \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1/\phi). \tag{16}$$

Using the Zellner-Siow prior distribution, the secure BMA can proceed as described in Section 3.2. The key benefit of the approximation is that only one round of secure computations is needed. However, unlike the protocol for linear regression, this protocol for probit regression is not guaranteed to conform mathematically to the BMA based on the concatenated data.

To investigate the performance of the approximate secure BMA empirically, we perform two simulation studies: one with an enumerable model space of 2^7 models and the other with a non-enumerable model space of dimension 2^{57}. Outputs for both exact and approximate secure methods do not depend on how the data are horizontally partitioned: the secure summation protocol guarantees the results will be exactly the same for the concatenated data. Thus, for each simulation study, we illustrate the results using a single data set.

4.2.1 Enumerable Model Space

In this section we compare the results of the approximate secure BMA protocol to the results of exact BMA on the Pima Indians Diabetes data (Ripley, 1996) provided in the library `MASS` in `R`.

For $i = 1, \ldots, 532$, let $Y_i = 1$ if the ith woman had diabetes and $Y_i = 0$ otherwise. There are seven covariates: *npreg* (number of times pregnant), *glu* (plasma glucose concentration in a glucose tolerance test), *bp* (diastolic blood pressure), *skin* (triceps skin fold thickness), *bmi* (body mass index), *ped* (diabetes pedigree function), and *age* (in years). We center and scale the design matrix so that the column means equal zero and the column standard deviations equal one. We compare the results on (i) the set of top ten models, (ii) marginal posterior inclusion probabilities, which are used to assess the importance of predictors, and (iii) predictive performance of BMA, based on training the methods on a training sample and calculating misclassification rates on a test sample.

To implement exact BMA via data augmentation, we assign the prior distribution in (12). We specify $\tau_j = 1$ for $j = 1, \ldots, p = 8$ (including the intercept). It is reasonable to use the same value of τ_j for all j because the predictors are standardized. This is commonly done in variable selection methods like ridge regression and lasso, so that all the coefficients can be interpreted on the same scale. We use $\tau_j = 1$ based on experience from variable selection with linear regression that indicates large prior variances have the undesirable effect of overwhelmingly favoring the null model (Liang *et al.*, 2008).

Posterior computation proceeds using the Gibbs sampler outlined in Section 4.1. We run the Gibbs sampler for 10,000 iterations, discarding the first 500 iterations as burn-in. We repeat the process 10 times to assess variability due to simulation.

For the approximate approach, the model space is small enough to enumerate fully; we do not need MCMC simulation. For each of the 2^7 models (the intercept is included in all models), we calculate the marginal likelihood using the Laplace approximation described in Section 3.2, and hence the marginal posterior inclusion probabilities. Computations are done with the `R` package `BAS` (Clyde, 2010).

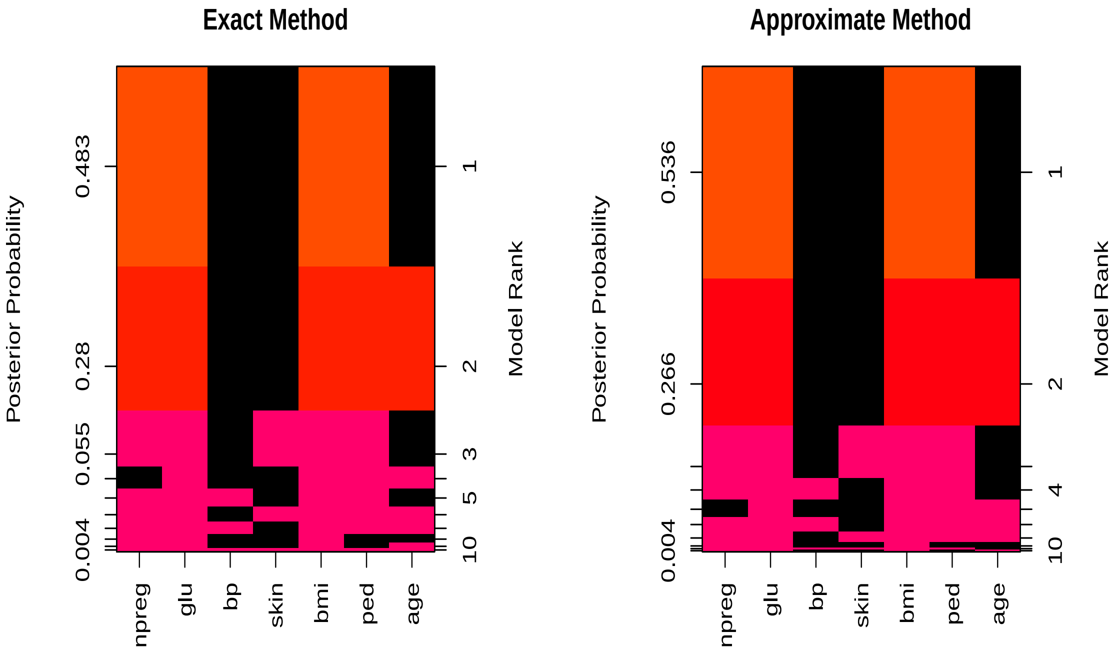

Figure 1: Top 10 models identified for the Pima Indians Diabetes data, using the exact secure BMA approach using a probit likelihood and the approximate approach using a normal likelihood; here models are represented by rows with the top model shown in the first row, predictors are represented by columns, and black regions denote predictors that were absent from a particular model.

Method	npreg	glu	bp	skin	bmi	ped	age
Exact	0.95	1.00	0.07	0.10	1.00	0.97	0.39
	(0.004)	(0.0)	(0.004)	(0.007)	(0.002)	(0.004)	(0.010)
Approx	0.96	1.00	0.08	0.08	1.00	0.99	0.36

Table 1: Estimates of marginal posterior inclusion probabilities for the Pima Indians Diabetes data, using the exact approach using a probit likelihood and the approximate approach using a normal likelihood; the mean and standard deviation (in parentheses) are reported based on 10 runs of the Gibbs sampler for the exact approach.

Figure 1 shows that the top models identified by the exact and approximate methods are very similar with respect to both the predictors that they contain and their posterior probabilities. Table 1 displays the corresponding estimates of posterior inclusion probabilities for each covariate, along with their counterparts from the normal approximation; these are in close agreement. The standard deviations across the ten runs are small, indicating that the number of MCMC iterations is adequate for the estimates to have stabilized.

To evaluate predictions, we split the data randomly into training and test samples. We choose the size of the training and test samples based on the `Pima.tr` and `Pima.te` in the library `MASS` in `R` (Ripley, 1996). The training sample consists of a randomly selected subset of 200 subjects and the test sample contains the remaining 332 subjects. We repeat this 10 times leading to 10 different random splits of the data to assess variability. The average misclassification rates using the probit and normal likelihoods are 23.89% and 22.95%, and the standard deviations (over the 10 random splits) are 1.47% and 1.10% respectively, so the two methods result in similar predictive performance.

Finally, we investigated the sensitivity of variable selection with respect to the choice of hyperparameters by repeating analyses using $\tau_j = 4$ and $\tau_j = 10$. While the inclusion probabilities changed somewhat, conclusions regarding the importance of predictors (defining important predictors as those with marginal inclusion probabilities > 0.5) were not affected.

4.2.2 Non-enumerable Model Space

In this section we carry out a simulation study with $p = 57$ predictors resulting in a non-enumerable model space. We create the data set by adding 50 noise variables generated as N(0,1) random variables to the Pima Indians data. For large model spaces like this, typically the size of the MCMC sample is a small fraction of the model space. The estimates of model probabilities based on Monte Carlo frequencies are no longer reliable because repeated visits to models are rare. Thus, we compare the methods based on (i) marginal posterior inclusion probabilities and (ii) predictive performance of BMA. We use the same prior specification as in the enumerable case and run the Gibbs sampler for 100,000 iterations, discarding the first 500 as burn-in. For the approximate approach we also cannot enumerate the model space, so that we run a Gibbs sampler for 100,000 iterations with a 500 burn-in. We run both Gibbs samplers 10 times to assess the variability across runs.

Method	npreg	glu	bp	skin	bmi	ped	age	noise-13	noise-36
Exact	0.97 (0.002)	1.00 (0.000)	0.08 (0.001)	0.10 (0.001)	1.00 (0.000)	0.97 (0.001)	0.39 (0.006)	0.97 (0.001)	0.93 (0.001)
Approx	0.98 (0.000)	1.00 (0.000)	0.12 (0.001)	0.13 (0.001)	1.00 (0.000)	0.99 (0.000)	0.47 (0.002)	0.98 (0.000)	0.97 (0.001)

Table 2: Estimates of marginal posterior inclusion probabilities for the data set with 57 predictors (created by adding 50 noise predictors to the Pima Indians Diabetes data), using the exact approach using a probit likelihood and the approximate approach using a normal likelihood; the mean and standard deviation (in parentheses) are reported based on 10 runs of the Gibbs sampler for both approaches.

Table 2 displays the estimates of posterior inclusion probabilities for the seven covariates in the original Pima Indians data and for covariates with marginal inclusion probabilities greater than 0.5. The estimates from the Gibbs samplers based on the probit model and the normal approximation are in close agreement for most predictors, and the standard deviations across the ten runs are reasonably small. Both methods wrongly identified two noise variables as important predictors. For out of sample prediction, we split the data into ten random splits as before. The average misclassification rates using the exact and approximate approaches are 23.13% and 22.74%, and the corresponding standard deviations are 1.81% and 1.95% respectively.

We examined the sensitivity of the variable selection to the choice of τ_j as in Section 4.2.1. Once again, the four important predictors in the original Pima data continued to be identified as important, although the inclusion probabilities changed somewhat. However, when setting $\tau_j = 10$, the marginal probability of one of the wrongly identified noise predictors decreased appreciably to 0.54.

4.3 Secure BMA Using Importance Sampling

In variable selection and model averaging problems for generalized linear models, the Bayesian Information Criterion, abbreviated BIC (Schwarz, 1978), and Laplace approximations to the marginal likelihood are often used in lieu of exact marginal likelihoods (Tierney and Kadane, 1986; Raftery, 1996; Clyde, 2002). In this section, we develop secure implementations of BMA for probit regression using both the BIC and a Laplace approximation.

For both we use a similar two-step process. First, because the posterior distribution of γ with these approximations is expensive to sample from securely, we sample models using the normal approximation of Section 4.2. Second, we re-weight the corresponding samples from the normal approximation by using importance sampling. This procedure requires multiple rounds of secure summation operations, but potentially much fewer rounds compared to the secure data augmentation algorithm. Thus, it strikes a balance in computation and accuracy between the methods in Section 4.1 and 4.2.

4.3.1 Secure Importance Sampling for BIC Approximation

The BIC approximation does not depend on the specified prior distributions under each model, so that it offers some robustness to the choice of the prior distributions. This can be particularly convenient with horizontally partitioned data, as it eliminates the need for the data owners to agree on a set of common values for prior hyperparameters.

For enumerable model spaces the posterior probabilities $p(\boldsymbol{\gamma}|\mathbf{X},\mathbf{Y})$ are readily available for the normal proposal distribution, as well as the probit target distribution, based on the BIC approximation. We have

$$p(\boldsymbol{\gamma} \mid \mathbf{X},\mathbf{Y}) = \frac{\exp\left(-\mathrm{BIC}(\boldsymbol{\gamma})/2\right)p(\boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}\in\Gamma}\exp\left(-\mathrm{BIC}(\boldsymbol{\gamma})/2\right)p(\boldsymbol{\gamma})}, \tag{17}$$

where $\mathrm{BIC}(\boldsymbol{\gamma}) = -2l_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) + \log(n)p_{\boldsymbol{\gamma}}$, $l_{\boldsymbol{\gamma}}()$ is the log-likelihood, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the maximum likelihood estimate of $\boldsymbol{\beta}_{\gamma}$, and $p_{\boldsymbol{\gamma}}$ is the number of parameters under model $\boldsymbol{\gamma}$. For non-enumerable model spaces the posterior probabilities are available up to a normalizing constant. Note that even if the entire model space is enumerated with the approximations in (17), the resulting posterior probabilities would differ from their counterparts from the data augmentation Gibbs sampler that uses the specified normal prior for the regression coefficients.

The importance sampling algorithm proceeds by drawing say T models from the proposal distribution, i.e., the posterior distribution over models using the linear model and the Zellner-Siow prior. For enumerable model spaces one can use independent Monte Carlo sampling; for larger model spaces one can sample models via MCMC. Both of these require only one round of secure matrix computations, as described in Section 4.2. We next calculate the importance sampling weights $w'_t = w_t / \sum_{t=1}^{T} w_t$, where w_t is the ratio of the marginal likelihoods for model $\boldsymbol{\gamma}^{(t)}$ under the target and proposal distributions respectively (the ratio of prior model probabilities under the target and proposal distributions is 1). To calculate $E[h(\boldsymbol{\gamma})]$ under the target distribution, one would use the usual importance sampling estimator $\sum_{t=1}^{T} h(\boldsymbol{\gamma}^{(t)}) w'_t$. For small to moderate p, many of the T models will overlap, and weights need to be calculated only for the unique set of sampled models. This reduces unnecessary rounds of secure matrix computation. For each of the unique sampled models, the maximum likelihood estimates (MLE) of $\boldsymbol{\beta}_\gamma$ under the probit model, the log-likelihood evaluated at the MLE, and hence the BIC, can be computed securely using a secure IWLS (iteratively weighted least squares) algorithm (Slavkovic *et al.*, 2007).

To illustrate, we first apply the importance sampling algorithm with BIC to the Pima Indians Diabetes data. Table 3 displays the estimated inclusion probabilities for various values of T. As expected, with an increase in T, the accuracy of the estimates also increases. For all T, the secure importance sampling algorithm produces estimates that are reasonably close to the BIC approximation based on the concatenated data. For example, the inclusion probability of the covariate *age* based on BIC is 0.27, and the estimate based on importance sampling is 0.26. In contrast, the estimate based on the the normal approximation of Section 4.2 alone is 0.36, which is much higher than 0.27. The maximum number of unique sampled models (over 10 replicates) for $T = 100,\ 1,000$ and $10,000$ were as small as 11, 20 and 29 respectively. Thus, the algorithm requires far fewer rounds of secure summations than the exact BMA approach of Section 4.1.

Figure 2 compares the estimates of model probabilities obtained from enumerating the model space with the BIC approximation to the importance sampling estimates. There is close agreement between the two estimates, even for a relatively small sample size of $T = 100$.

We next apply both the importance sampling algorithm with BIC and a Gibbs sampler based on the BIC approximation to the simulated data with 57 predictors, 7 of which are the predictors in the Pima Indians Diabetes data and the remaining 50 are noise. Table 4 displays the estimated inclusion probabilities for various values of T. The inclusion probabilities for the BIC based Gibbs sampler does not seem to converge until running 100,000 iterations. The importance sampling

Method	npreg	glu	bp	skin	bmi	ped	age
Enumeration-BIC	0.94	1.00	0.05	0.05	1.00	0.95	0.27
IS-BIC: $T = 100$	0.92 (0.061)	1.00 (0.000)	0.04 (0.014)	0.06 (0.017)	1.00 (0.002)	0.94 (0.054)	0.26 (0.068)
IS-BIC: $T = 1,000$	0.94 (0.013)	1.00 (0.000)	0.05 (0.005)	0.05 (0.005)	1.00 (0.002)	0.95 (0.017)	0.27 (0.018)
IS-BIC: $T = 10,000$	0.94 (0.003)	1.00 (0.000)	0.05 (0.001)	0.05 (0.001)	1.00 (0.000)	0.95 (0.005)	0.27 (0.003)

Table 3: Estimates of marginal posterior inclusion probabilities for the Pima Indians Diabetes data, under enumeration using BIC and under importance sampling with BIC (IS-BIC) to the marginal likelihood, with sample sizes $T = 100$, $T = 1,000$, and $T = 10,000$; the mean and standard deviation (in parentheses) are reported based on 10 runs of the importance sampler for each value of T.

estimates for 100,000 iterations are in close agreement with the Gibbs sampler. Note that here the number of unique sampled models is almost as large as the MCMC sample size because models are rarely sampled twice in large model spaces like this. However, it is still beneficial to use the importance sampling algorithm compared to the BIC based Gibbs sampler. This is because the Gibbs sampler needs to cycle through $p = 57$ inclusion indicators and compute “p” BICs to sample a single model, whereas the importance sampler needs to calculate the BIC only for the final sampled model. In short, the BIC based Gibbs sampler has p times more rounds of secure summation. The importance sampling is also preferable to the data augmentation Gibbs sampler in terms of having fewer rounds of secure computation. Each iteration of the data augmentation Gibbs sampler consists of cycling through p inclusion indicators, whereas one iteration of the importance sampler needs one BIC calculation that usually involved fewer than 10 iterations of the IWLS algorithm. So, each iteration of the data augmentation Gibbs sampler will typically involve $(p - 10)$ more rounds of secure computation.

4.3.2 Secure Importance Sampling for Laplace Approximation

In the previous section, the use of BIC was motivated primarily by its robustness to the choice of the prior distribution rather than its consistency properties in estimating Bayes factors. For small dimensional model spaces, one can improve upon the BIC approximation by using Laplace approximations instead. Laplace approximations account for prior distributions, offering consistent estimates of the Bayes factors under the specified prior distribution.

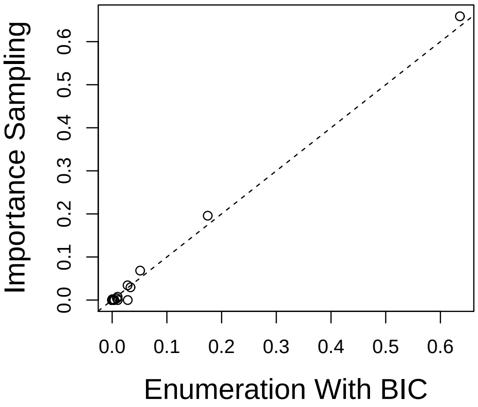

Figure 2: Comparison of posterior probabilities of all 2^7 models for the Pima Indians Diabetes data, using i) BIC approximation to the marginal likelihood for a probit link function (based on enumerating all 2^7 models) and ii) Importance sampling estimates for a single run with $T = 100$ (unsampled models are automatically assigned an estimate zero).

We use the Laplace approximation developed by Raftery (1996), which requires the MLEs and the observed Fisher information matrix. Using the prior specification in (12), this Laplace approximation is

$$\log(p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\gamma})) \simeq l_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) + \lambda_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) +$$

$$\boldsymbol{\lambda}_{\boldsymbol{\gamma}}{}'(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})^T(\boldsymbol{F}_{\boldsymbol{\gamma}}+\boldsymbol{G}_{\boldsymbol{\gamma}})^{-1}\{\boldsymbol{I}_{p_{\boldsymbol{\gamma}}} - \frac{1}{2}\boldsymbol{F}_{\boldsymbol{\gamma}}(\boldsymbol{F}_{\boldsymbol{\gamma}}+\boldsymbol{G}_{\boldsymbol{\gamma}})^{-1}\}\boldsymbol{\lambda}_{\boldsymbol{\gamma}}{}'(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) - \frac{1}{2}\log|\boldsymbol{F}_{\boldsymbol{\gamma}}+\boldsymbol{G}_{\boldsymbol{\gamma}}| + \frac{p_{\boldsymbol{\gamma}}}{2}\log(2\pi), \quad (18)$$

where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the MLE of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, $\boldsymbol{F}_{\boldsymbol{\gamma}}$ is the observed Fisher Information matrix, $\boldsymbol{G}_{\boldsymbol{\gamma}}$ is the inverse of the prior covariance matrix of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, and $p_{\boldsymbol{\gamma}}$ is the number of parameters, all under model $\boldsymbol{\gamma}$. For the chosen independent normal prior with $\tau_j = \tau$ for all j, we have

$$\lambda_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = -\frac{p_{\boldsymbol{\gamma}}}{2}\log(2\pi) - \frac{p_{\boldsymbol{\gamma}}}{2}\log(\tau) - \frac{1}{2\tau}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^T\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}},$$

$$\boldsymbol{\lambda}'_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}) = -\frac{1}{\tau}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}.$$

Both $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ and $l_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})$ can be obtained via a secure IWLS algorithm. The observed Fisher

Method	npreg	glu	bp	skin	bmi	ped	age	noise-13	noise-36
GS-BIC: $T = 1,000$	1.00	1.00	0.05	0.00	1.00	0.90	0.20	1.00	0.85
	(0.000)	(0.000)	(0.158)	(0.000)	(0.000)	(0.211)	(0.258)	(0.000)	(0.242)
GS-BIC: $T = 10,000$	1.00	1.00	0.10	0.00	1.00	0.80	0.30	1.00	0.90
	(0.000)	(0.000)	(0.211)	(0.000)	(0.000)	(0.350)	(0.422)	(0.000)	(0.211)
GS-BIC: $T = 100,000$	0.97	1.00	0.05	0.05	1.00	0.96	0.23	0.96	0.88
	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)	(0.002)	(0.001)	(0.001)
IS-BIC: $T = 1,000$	0.96	1.00	0.04	0.07	0.99	0.98	0.23	0.96	0.90
	(0.024)	(0.000)	(0.018)	(0.033)	(0.013)	(0.027)	(0.049)	(0.028)	(0.075)
IS-BIC: $T = 10,000$	0.97	1.00	0.04	0.06	1.00	0.96	0.24	0.96	0.90
	(0.014)	(0.000)	(0.007)	(0.008)	(0.002)	(0.017)	(0.024)	(0.011)	(0.026)
IS-BIC: $T = 100,000$	0.97	1.00	0.04	0.05	1.00	0.96	0.23	0.95	0.88
	(0.005)	(0.000)	(0.004)	(0.003)	(0.001)	(0.009)	(0.006)	(0.003)	(0.011)

Table 4: Estimates of marginal posterior inclusion probabilities for the data set with 57 predictors (created by adding 50 noise variables to the Pima Indians Diabetes data), under Gibbs sampling (GS-BIC) and importance sampling (IS-BIC) with BIC approximation to the marginal likelihood, for sample sizes $T = 1,000$, $T = 10,000$, and $T = 100,000$; the mean and standard deviation (in parentheses) are reported based on 10 runs of the samplers for each value of T.

information matrix $\boldsymbol{F}_{\boldsymbol{\gamma}}$ is obtained as part of the secure IWLS algorithm. All other quantities in (18) are functions of the prior hyperparameters, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$, and $\boldsymbol{F}_{\boldsymbol{\gamma}}$. Hence, the calculation does not involve any more secure summations than the BIC approximation of Section 4.3.1.

For enumerable model spaces, once $\log(p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\gamma}))$ are calculated via (18), one can use (3) to calculate approximate posterior probabilities. For small dimensional model spaces (with respect to the sample size), these estimates should be in accord with those obtained from the exact data augmentation Gibbs sampler in Section 4.1. However, because enumeration of model spaces is generally expensive from the perspective of secure computation, we instead use an importance sampling algorithm akin to the one in Section 4.3.1, with importance weights now based on the Laplace approximation instead of the BIC approximation.

Method	npreg	glu	bp	skin	bmi	ped	age
Enumeration-LA	0.95	1.00	0.07	0.10	1.00	0.97	0.39
IS-LA: $T = 100$	0.93	1.00	0.07	0.11	1.00	0.96	0.38
	(0.047)	(0.000)	(0.024)	(0.030)	(0.003)	(0.033)	(0.063)
IS-LA: $T = 1,000$	0.94	1.00	0.08	0.10	1.00	0.97	0.39
	(0.010)	(0.000)	(0.007)	(0.009)	(0.002)	(0.012)	(0.018)
IS-LA: $T = 10,000$	0.95	1.00	0.08	0.10	1.00	0.97	0.40
	(0.002)	(0.000)	(0.002)	(0.002)	(0.000)	(0.003)	(0.004)

Table 5: Estimates of marginal posterior inclusion probabilities for the Pima Indians Diabetes data, under enumeration using a Laplace approximation (LA), and under importance sampling with Laplace approximation (IS-LA) to the marginal likelihood, with sample sizes $T = 100$, $T = 1,000$, and $T = 10,000$; the mean and standard deviation (in parentheses) are reported based on 10 runs of the importance sampler for each value of T.

We apply this importance sampling algorithm with Laplace approximation to the Pima Indians Diabetes data. Table 5 displays the corresponding estimated inclusion probabilities. For all T, the secure importance sampling algorithm produces estimates that are reasonably close to the Laplace approximation based on the concatenated data. Moreover, the inclusion probability estimates for all covariates are very close to the exact estimates from the secure data augmentation Gibbs sampler, reported in Table 1. This is expected, because the Laplace approximation approximates the Bayes factors under the correct posterior distribution corresponding to the specified independent normal prior. The unique number of sampled models for all values of T was less than 29, justifying the use of a sampling algorithm instead of enumerating the model space, which involves many more secure summations.

For high dimensional model spaces, when the number of predictors grows with the sample size, the theoretical justification of the Laplace approximation or a good numerical study of its performance is not yet available; see Berger *et al.* (2003). Thus, we do not illustrate the Laplace approximations for the high-dimensional model space example of dimension 2^{57}.

5 Application to SPECT Heart Data

In this section, we apply the methods for secure BMA to genuine data with a moderate number of covariates and more challenging variable selection task. We use the SPECT heart data (Kurgan *et al.*, 2001), available from the University of California Irvine Machine Learning Repository. The data comprise information about 267 patients and their cardiac Single Proton Emission Computed Tomography (SPECT) images. Based on their SPECT images, a cardiologist classifies the patients into two categories: normal and abnormal heart conditions. The explanatory variables include 20 binary features extracted from the images using a machine learning algorithm. The goal is to use these image features to predict the cardiologist's diagnosis.

We note that the original data include 22 predictors. For purposes of illustration, we removed two predictors that generate perfect predictions—for either of these predictors, whenever $x = 1$ we always have $y = 1$—because including them makes MLEs numerically unstable. We discuss issues for secure computation related to such separability further in Section 6.

Using the full model space of size 2^{20}, we implemented i) the exact data augmentation Gibbs sampler, ii) the approximate Gibbs sampler for normal approximation, iii) the BIC based Gibbs sampler, and iv) the BIC based importance sampler. We use a standardized design matrix and the same choices of prior distributions as in Section 4. We run each algorithm for 100,000 iterations, discarding the first 500 samples as burn-in.

Figure 3 displays the inclusion probabilities from the exact BMA and the three approximate methods. In general there is close agreement on the inclusion probabilities. Defining important (unimportant) predictors as those with inclusion probabilities greater (less) than 0.5, we can summarize the results from Figure 3 as follows. There are no points in the upper left block of Figure 3, so that the approximate methods did not wrongly select any unimportant predictors. There are two points in the lower right block showing that the approximate methods missed these two important predictors. Both these predictors have relatively low inclusion probabilities (0.56 and 0.64). These results illustrate that predictors without appreciably large posterior inclusion probabilities under the probit model may not be identified as important with the approximate methods.

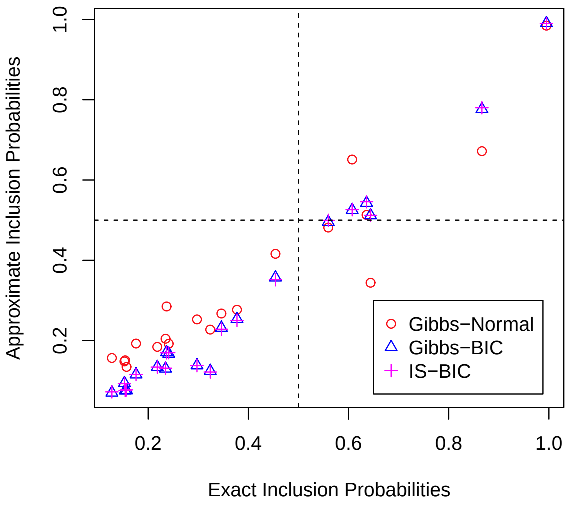

Figure 3: Comparison of posterior marginal inclusion probabilities for the heart data, using i) the exact data augmentation Gibbs sampler, ii) the normal approximation based Gibbs sampler (Gibbs-Normal), iii) the BIC based Gibbs sampler (Gibbs-BIC), and iv) the BIC based importance sampler (IS-BIC).

We also investigated out of sample predictions for the exact approach and the normal approximation. We split the data randomly into two parts: training data comprising 80% of the samples and test data comprising the remainder. The misclassification rate in the test data was 26.41% for the exact approach and 24.53% for the approximate approach. A different random split of the data resulted in rates of 15.09% and 13.21%, respectively.

Finally, we performed sensitivity analysis of variable selection to the choice of prior hyperparameters. Unlike the Pima data (with 7 or 57 predictors), here the results are sensitive to the specification of τ_j. With $\tau_j = 4$, three of the predictors that were deemed as important when $\tau_j = 1$ no longer had inclusion probabilities greater than 0.5. These initially had inclusion probabilities in the range 0.56–0.64. For $\tau_j = 10$, only one predictor was flagged as important; this had an initial inclusion probability nearly equal to 1. These results are in accord with the discussion in Liang *et al.* (2008): the larger the prior variance on the regression parameters, the more the null model is favored.

6 Summary

In this article, we have shown that it is possible to perform secure Bayesian model averaging and Bayesian variable selection for linear and probit regression in the case of horizontally partitioned data. For normally distributed errors and conjugate priors like Zellner's g-prior, all quantities needed for BMA are available as sufficient statistics that can be computed using a secure summation protocol. Our approach also extends to the popular Zellner-Siow multivariate Cauchy prior for the regression coefficients. Credible intervals corresponding to the posterior distributions of ϕ, $\boldsymbol{\beta}$, or $\mathbf{Y}^*$ can be constructed via sampling from their corresponding posterior distributions, conditional on the set of sampled models, without any extra rounds of secure summation. For probit regression, we illustrated the use of a Gaussian approximation to the likelihood as a tool for BMA, which requires far fewer rounds of secure summation than the exact data augmentation approach does. While our focus in this paper has been on probit regression and approximations for it, technically the normal approximation can be viewed as an approximation to other likelihoods for binary response variables as well, such as logistic regression, as pointed out by one reviewer. Finally, we illustrated how the normal approximation can serve as a proposal distribution in importance sampling when using BIC or Laplace approximations to marginal likelihoods.

It was pointed out by one reviewer that the normal linear regression model assumes homoscedastic variance, which does not hold for the probit model, so that the approximation may not work well for data with small sample size or high variability in the variances of the response variables. Clyde (1999) suggests variance-stabilizing transformations for model selection in Bayesian generalized linear models, like square-root transformations for Poisson regression, or arcsine square-root transformation for binomial regression models with each observation having $n_i > 1$ trials. For binary regressions in which each observation is based on only one trial, such transformations are not available. We did further simulations to evaluate the performance of the approximate BMA of Section 4.2 in small samples with high variability in the variances. The approximate method tended to identify important predictors reasonably well, except for those with inclusion probability close to 0.5. In such data, it may be preferable to use the secure BMA approaches with importance sampling, which entail fewer secure computations than the exact data augmentation Gibbs sampler.

For binary regression models with many predictors, the approximate approaches to secure model selection can be compromised when the data exhibit separability. In particular, the maximum likelihood estimates may not converge or can exhibit numerical instability, in which case approaches based on the BIC cannot be trusted; in such cases, the normal approximation is the only practical secure BMA approach. For example, when we used the SPECT heart data that includes the two predictors causing separability, the frequentist standard errors for these two coefficients in a probit regression model were on the order of 300 times larger than the standard errors for other predictors' coefficients, revealing numerical instability of the MLE. We do not see quick practical fixes to the problems caused by separability in secure BMA for probit regression. However, agencies can take steps to identify potential separability and adjust analyses accordingly. For example, each agency can search for separability within their own data and share its findings with other agencies. If necessary, the agencies collectively can share this information without revealing sources via a secure data integration protocol like the one of Karr *et al.* (2005). We note that similar steps could be taken to identify possibly collinear predictors; these also would be evident from the covariances in $\mathbf{X}'\mathbf{X}$ in the normal approximation.

Once the offending variables causing separability have been identified, the agencies can choose to remove these variables from analysis if it is deemed scientifically relevant to do so. Alternatively, when the goal of the probit modeling is classification, the agencies could proceed in steps. First, the agencies remove all cases that are perfectly predicted by the offending variables, e.g., remove the cases with $x = 1$ for the two offending predictors in the SPECT heart data. Second, the agencies remove the offending variables from the model space. Third, the agencies implement secure model selection on the remaining cases and variables. For classification of new cases, we predict $y = 1$ when $x = 1$ for the offending variables, and we use the model from secure BMA when $x = 0$ for the offending variables. The accuracy of this two-step approach is a subject for future research.

Acknowledgements

The authors thank the editor, associate editor and two referees for valuable comments and suggestions. The first author was supported by the Old Gold Summer Fellowship, University of Iowa. The second author was supported by National Science Foundation grant CNS 1012141.

References

Agrawal, R. and Srikant, R. (2000). Privacy-preserving data mining. In *Proceedings of the 2000 ACM SIGMOD on Management of Data*, 439–450.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 422, 669–679.

Barbieri, M. and Berger, J. (2004). Optimal predictive model selection. *Annals of Statistics* **32**, 3, 870–897.

Benaloh, J. (1987). Secret sharing homomorphisms: keeping shares of a secret secret. In A. Odlyzko, ed., *Advances in Cryptography: CRYPTO86*, vol. 263, 251–260. Springer, New York.

Berger, J. and Perichhi, L. (2001). Objective bayesian methods for model selection: introduction and comparison [with discussion]. In P.Lahiri, ed., *Institute of Mathematical Statistics Lecture Notes, Monograph Series volume 38, Beachwood Ohio*, 135–207.

Berger, J. O., Ghosh, J. K., and Mukhopadhyay, N. (2003). An overview of robust Bayesian analysis. *Journal of Statistical Planning and Inference* **112**, 241–258.

Carlin, B. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B* **57**, 473–384.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90**, 1313–1321.

Churches, T. and Christen, P. (2004). Some methods for blindfolded record linkage. *BMC Medical Informatics and Decision Making* **4**, 9.

Clyde, M. (1999). Bayesian model averaging and model search strategies (with discussion). In *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, 157–185.

Clyde, M. (2002). Model averaging. In S. J. Press, ed., *Subjective and objective Bayesian statistics: principles, models and applications*. John wiley and Sons, Inc.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science* **19**, 1, 81–94.

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

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). Bayesian adaptive sampling for variable selection and model averaging. *Journal of Computational and Graphical Statistics* **20**, 1, 80–101.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing* **12**, 1, 27–36.

Du, W., Han, Y., and Chen, S. (2004). Privacy-preserving multivariate statistical analysis: linear regression and classification. In *Proceedings of the 4th SIAM International Conference on Data Mining*, 222–233.

Evfimievski, A., Srikant, R., Agrawal, R., and Gehrke, J. (2004). Privacy preserving mining of association rules (invited journal version). *Information Systems* **29**, 4, 343–364.

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.

Gelman, A. and Meng, X.-L. (1998). Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science* **13**, 163–185.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

Ghosh, J. and Clyde, M. A. (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.

Ghosh, J., Reiter, J. P., and Karr, A. F. (2007). Secure computation with horizontally partitioned data using adaptive regression splines. *Computational Statistics and Data Analysis* **51**, 5813–5820.

Heaton, M. and Scott, J. (2010). Bayesian computation and the linear model. In M.-H. Chen, D. K. Dey, P. Mueller, D. Sun, and K. Ye, eds., *Frontiers of Statistical Decision Making and Bayesian Analysis*.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with discussion). *Statistical Science* **14**, 4, 382–401.

Holmes, C. C. and Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* **1**, 145–168.

Kantarcioglu, M. and Clifton, C. (2002). Privacy-preserving distributed mining of association rules on horizontally partitioned data. In *The ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD 2002), Madison, Wisconsin*, 24–31.

Karr, A., Lin, X., Sanil, A., and Reiter, J. (2005). Secure regressions on distributed databases. *Journal of Computational and Graphical Statistics* **14**, 263–279.

Kurgan, L., Cios, K., Tadeusiewicz, R., Ogiela, M., and Goodenday, L. (2001). Knowledge discovery approach to automated cardiac spect diagnosis. *Artificial Intelligence in Medicine* **23:2**, 149–169.

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.

Lin, X., Clifton, C., and Zhu, Y. (2005). Privacy preserving clustering with distributed em mixture modeling. *International Journal of Knowledge and Information Systems* **8**, 1, 68–81.

Lindell, Y. and Pinkas, B. (2000). Privacy-preserving data mining. In *Advances in Cryptology: CRYPTO2000*, 36–54. Springer, New York.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica* **6**, 831–860.

Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 251–266.

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

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

Slavkovic, A. B., Nardi, Y., and Tibbits, M. M. (2007). Secure logistic regression of horizontally and vertically partitioned distributed databases. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, 723 –728.

Tierney, L. and Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82–86.

Vaidya, J. and Clifton, C. (2002). Privacy preserving association rule mining in vertically partitioned data. In *The 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, Canada*, 639–644.

Vaidya, J. and Clifton, C. (2003). Privacy-preserving k-means clustering over vertically partitioned data. In *The 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC*, 206–215.

Willenborg, L. and de Waal, T. (2001). *Elements of Statistical Disclosure Control*. Springer Verlag, New York.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. North-Holland/Elsevier.

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)*, 585–603.