Journal of Statistical Research *ISSN 0256 - 422 X*
2010, Vol. 44, No. 1, pp. 187-205
Bangladesh

VARIATIONAL BAYESIAN LOGISTIC REGRESSION MODEL SELECTION: AN IMPROVEMENT OVER LAPLACE?

Liang Zhang
Yahoo! Labs, Santa Clara, CA
Email: liangzha@yahoo-inc.com

David B. Dunson
Department of Statistical Science, Duke University, Durham, NC
Email: dunson@stat.duke.edu

SUMMARY

Increasingly, statisticians are faced with the problem of identifying interesting subsets of predictors from among a large number of candidates. Existing methods for variable selection, such as stochastic search algorithms, tend to explore the model space too slowly in large dimensions. Shotgun stochastic search (SSS) algorithms have been proposed as an efficient alternative. As current SSS algorithms rely on conjugacy, they are not appropriate for generalized linear models without use of approximation methods. This article compares the frequently used Laplace approximation with two alternatives based on Variational Bayes methods. The comparison is illustrated using several simulated data examples and an application to the problem of predicting conception using data on timing of intercourse in the menstrual cycle. This application also illustrates the problem of selection of interactions.

Keywords and phrases: Approximation; Large p, small n; Model uncertainty; Shotgun stochastic search; Subset selection; Variable selection.

1 Introduction

As the collection of massive amounts of information becomes more routine, there is a critical need for more efficient methods for identifying promising subsets of variables from among the very many candidates one is typically faced with. This problem occurs not only in genomics and bioinformatics studies, where it has received the most focus, but also in epidemiologic studies. For example, the application motivating this article focuses on using data on the timing of intercourse in the menstrual cycle to predict conception. The data consist of daily records of intercourse across the menstrual cycle and an indicator of conception status for women enrolled in a European study (Dunson *et al.*, 2002). Although the fertile interval of the menstrual cycle is only 5-6 days for most women (Dunson *et al.*, 1999), the timing of

the fertile days is highly uncertain (Wilcox *et al.*, 2000). Hence, there are many days in the menstrual cycle during which intercourse can potentially result in a pregnancy. In addition, potential interactions lead to a very high-dimensional set of candidate models.

One widely used approach for accommodating uncertainty in subset selection in linear regression models is the stochastic search variable selection (SSVS) algorithm originally proposed in George and McCulloch (1993), with numerous modifications later considered (George and McCulloch, 1997; Ishwaran and Rao, 2005; Casella and Moreno, 2006). SSVS algorithms rely on using a mixture prior for the regression coefficients, with one component concentration at 0, allowing a predictor to effectively drop out of the model. Gibbs sampling is then used to sample from the conditional posterior distributions of the coefficients, resulting in stochastic changes to the variables included in the model across the MCMC iterates. Such algorithms are very effective in modest sized models, but tend to explore the model space too slowly as the number of candidate variables increases.

Motivated by this problem, Hans (2005) and Hans *et al.* (2007) proposed a new regression model search algorithm named Shotgun Stochastic Search (SSS), with Jones *et al.* (2005) applying this approach to graphical models. Compared to the variety of alternative methods available (reviewed by Dellaportas *et al.* (2002)), Hans *et al.* (2007) argue that SSS is more efficient at rapidly identifying the models with the highest posterior probabilities in large model spaces. It is interesting to extend the SSS beyond linear regression to broader classes of regression models, such as generalized linear models (GLMs). Hans (2005) and Hans *et al.* (2007) proposed to use Laplace approximation for marginal likelihood approximations in GLM, and used SSS to search in the model space. Ntzoufras *et al.* (2003) proposed a reversible jump MCMC algorithm for posterior computation in GLMs when there is uncertainty in the predictors to be included. In recent work, Wang and George (2007) proposed an adaptive Bayesian criterion for variable selection in GLMs, relying on an integrated Laplace approximation to allow rapid computation.

Our initial goal is to consider applications of SSS algorithms to GLM variable selection in massive dimensions, with a particular emphasis on logistic regression models motivated by the fertility application. For subject i ($i = 1, \ldots, n$), let y_i denote the binary response variable and let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ denote a $p \times 1$ vector of candidate predictors, with γ_j a 0/1 indicator that the jth predictor is included in the model and $\mathbf{x}_{\gamma,i} = \{x_{ij} : \gamma_j = 1\}$ denoting the subset of included predictors. Then, the logistic regression model can be expressed as

$$\text{logit}\{\Pr(y_i = 1 \,|\, \mathbf{x}_i, \boldsymbol{\gamma})\} = \boldsymbol{x}'_{\gamma,i}\boldsymbol{\beta}_\gamma, \tag{1.1}$$

where $\text{logit}(x) = ln(\frac{x}{1-x})$, and $\boldsymbol{\beta}_\gamma$ is a vector of the unknown regression coefficients in the model indexed by $\boldsymbol{\gamma}$.

Letting $\pi(\boldsymbol{\gamma})$ denote the prior probability of model $\boldsymbol{\gamma}$, the posterior model probability is

$$p(\boldsymbol{\gamma} \,|\, \boldsymbol{Y}, \boldsymbol{X}) = \frac{p(\boldsymbol{\gamma})\, p(\boldsymbol{Y} \,|\, \boldsymbol{X}, \boldsymbol{\gamma})}{\sum_{\boldsymbol{\gamma}^* \in \Gamma} p(\boldsymbol{Y} \,|\, \boldsymbol{X}, \boldsymbol{\gamma}^*)} = \frac{p(\boldsymbol{\gamma})\, p(\boldsymbol{Y} \,|\, \boldsymbol{X}, \boldsymbol{\gamma})}{p(\boldsymbol{Y} \,|\, \boldsymbol{X})},$$

where Γ is the set of all 2^p possible subsets and the marginal likelihood of the data under

model $\boldsymbol{\gamma}$ is

$$p(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\gamma}) = \int p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}},\boldsymbol{\beta}_{\boldsymbol{\gamma}})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}})d\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tag{1.2}$$

which is not available analytically in most cases. In logistic regression, the marginal likelihood is expressed as the integral of the Bernoulli likelihood under the logistic model over the prior distribution for the coefficients, which does not have a closed form.

To complete a Bayesian specification of the model uncertainty problem, explicit choices are required for the prior probability of model $\boldsymbol{\gamma}$ and for the prior distribution on the coefficients within that model $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, for all $\boldsymbol{\gamma} \in \Gamma$. The standard choice of prior for $\boldsymbol{\gamma}$ corresponds to

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} \phi^{1(\gamma_j=1)}(1-\phi)^{1(\gamma_j=0)}, \tag{1.3}$$

where $1(\cdot)$ is a 0/1 indicator function, and ϕ is potentially given a beta hyperprior to allow the data to inform more strongly about model size.

In conducting a high-dimensional model search, it is necessary to rapidly estimate or approximate the Bayes factors for massive numbers of pairs of models. The Bayes factor is the ratio of marginal likelihoods under competing models. A variety of approaches have been proposed in the literature for Bayes factor estimation, with some of the approaches relying on approximation of the marginal likelihoods for the individual models and other approaches bypassing the need to estimate the individual model marginal likelihoods in calculating the ratio. In this paper, we follow the approach of first obtaining marginal likelihood approximations and then using these to approximate the Bayes factor. There are many simulation-based methods that can be used for estimating the marginal likelihood for comparing a small number of competing models (Gelfand and Smith, 1990; Gelfand and Dey, 1994; Verdinelli and Wasserman, 1995; Chib, 1995; DiCiccio *et al.*, 1997; Gelman and Meng, 1998; Han and Carlin, 2001; Chib and Jeliazkov, 2001). However, most approaches require substantial numbers of samples within each model to produce an accurate estimate, so are impractical in conducting a model search.

Hence, it is necessary to focus on marginal likelihood approximations that can be calculated very quickly. The Laplace approximation provides a convenient and widely used approach, which often performs well (Tierney and Kadane, 1986). DiCiccio *et al.* (1997) provide the details of the Laplace approximation to the marginal likelihood. Raftery (1996) uses the Laplace approximation for Bayesian model selection in GLMs, while Hans (2005); Hans *et al.* (2007) combined this approach with SSS. There are also other alternatives for estimating the marginal likelihood in logistic regression. Jaakkola and Jordan (2000) proposed a variational Bayes (VB) approach to approximate the posterior of $\boldsymbol{\beta}$ by a variational transformation of the logistic function. Their paper suggests two different VB methods.

It is widely believed in the machine learning community that the VB approach provides an improvement over the Laplace approximation. Srebro and Jaakkola (2003) pointed out that for the Taylor expansion, the iterative improvement of the approximation is not always monotonic, resulting in no guarantee of convergence. They also claimed that the VB method

is more robust and the convergence is guaranteed. However, Wang and Titterington (2005); Consonni and Marin (2006) proved that "the covariance matrices from the variational Bayes approximations are usually "too small" compared with those for the maximum likelihood estimator". Our initial motivation was to apply the VB method in the high dimensional variable selection setting to improve upon Laplace-based methods, though we found that Laplace consistently outperformed VB in logistic regression.

We first introduce the Laplace and VB methods, and describe their use in applying the shotgun stochastic search algorithm to high-dimensional variable selection. This is followed by two simulation studies; the first compares accuracy of the marginal likelihood approximations, and the second assesses predictive performance using VB or Laplace model averaging. Finally, we implement the SSS algorithm with VB and Laplace for the fertility data set.

2 Marginal Likelihood Approximations

2.1 Laplace Method

In this subsection, we provide a brief review of the Laplace approximation to the marginal likelihood in logistic regression. Re-expressing marginal likelihood (1.2) in the form

$$\int \exp\Big[\log\big\{p(\boldsymbol{Y}\,|\,\mathbf{X}_{\boldsymbol{\gamma}},\boldsymbol{\beta}_{\boldsymbol{\gamma}})\,p(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\,\big\}\Big]d\boldsymbol{\beta}_{\boldsymbol{\gamma}},$$

expanding the logarithm function using Taylor series, and keeping the items up to the second order, we have the estimator

$$\hat{p}(\boldsymbol{Y}|\mathbf{X}_{\boldsymbol{\gamma}}) = (2\pi)^{(k_{\boldsymbol{\gamma}}+1)/2}|\hat{\Sigma}_{\boldsymbol{\gamma}}|^{1/2}\,p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})p(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}), \tag{2.1}$$

where $k_{\boldsymbol{\gamma}}$ is the size of model $\boldsymbol{\gamma}$, $\hat{\Sigma}_{\boldsymbol{\gamma}}^{-1}$ is the approximate posterior covariance matrix

$$\hat{\Sigma}^{-1} = -\frac{\partial^2}{\partial\hat{\beta}_i\partial\hat{\beta}_j}\Big[\log\big\{p(\boldsymbol{Y}|\mathbf{X}_{\boldsymbol{\gamma}},\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})p(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})\big\}\Big], \tag{2.2}$$

and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the posterior mode under model $\boldsymbol{\gamma}$

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = argmax_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\big\{p(\boldsymbol{Y}\,|\,\mathbf{X}_{\boldsymbol{\gamma}},\boldsymbol{\beta}_{\boldsymbol{\gamma}})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\big\}. \tag{2.3}$$

To find the posterior mode, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, for a logistic regression, a simple Newton-Raphson algorithm can be used, with Raftery (1996) suggesting a one-step approximation that allows use of maximum likelihood estimates. Here, we instead iterate the algorithm until convergence. If the prior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is $N(\mathbf{0},\tau I_{k_{\boldsymbol{\gamma}}+1})$, the algorithm is based on the following updating equation:

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t+1)} = \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)} - G(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)})^{-1}g(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)}), \tag{2.4}$$

where

$$G(\boldsymbol{\beta_\gamma}) = \frac{1}{\tau} I_{k+1} - \sum_{i=1}^{n} \boldsymbol{x}_{\boldsymbol{\gamma},i} \boldsymbol{x}'_{\boldsymbol{\gamma},i} \phi_{\boldsymbol{\gamma},i}(1-\phi_{\boldsymbol{\gamma},i}), \qquad g(\boldsymbol{\beta_\gamma}) = -\frac{\boldsymbol{\beta_\gamma}}{\tau} + \sum_{i=1}^{n} (y_i - \phi_{\boldsymbol{\gamma},i}) \boldsymbol{x}_{\boldsymbol{\gamma},i},$$

$$\phi_{\boldsymbol{\gamma},\boldsymbol{i}} = \Big(1 + \exp(-\boldsymbol{x}'_{\boldsymbol{\gamma},i} \boldsymbol{\beta_\gamma})\Big)^{-1} \qquad .$$

Iterative updating of $\boldsymbol{\beta_\gamma}$ tends to converge within a few iterations, so that the approximation to the marginal likelihood under model $\boldsymbol{\gamma}$, $\hat{p}(\boldsymbol{Y}|\mathbf{X}_{\boldsymbol{\gamma}})$ can be obtained very quickly.

2.2 Variational Bayes Approximations

In this section, we describe two VB approaches to approximate the logistic regression marginal likelihood. The first approach was suggested by Jaakkola and Jordan (2000), while the second approach is also briefly introduced in their paper, though they present it as less appealing than the first approach.

2.2.1 Approach I

By Bayes' Theorem,

$$p(\boldsymbol{Y}|\mathbf{X}_{\boldsymbol{\gamma}}) = \frac{p(\boldsymbol{Y}|\mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}^*_{\boldsymbol{\gamma}}) p(\boldsymbol{\beta}^*_{\boldsymbol{\gamma}})}{p(\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}|\boldsymbol{Y}, \mathbf{X}_{\boldsymbol{\gamma}})}, \tag{2.5}$$

where $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$ is any vector in $\Re^{k_{\boldsymbol{\gamma}}}$. Expression (2.5) is commonly used in estimating marginal likelihoods via Monte Carlo sampling (e.g. Chen, 2005; Chib and Jeliazkov, 2001). Following Jaakkola and Jordan (2000), for any single observation y_i, we have

$$p(y_i|\mathbf{x}_{\boldsymbol{\gamma},i}, \boldsymbol{\beta_\gamma}) \geq p(y_i|\mathbf{x}_{\boldsymbol{\gamma},i}, \boldsymbol{\beta_\gamma}, \xi) = g(\xi) \exp\{(X_i - \xi)/2 + \lambda(\xi)(X_i^2 - \xi^2)\}, \tag{2.6}$$

where $g(\xi) = (1 + e^{-\xi})^{-1}$, $\lambda(\xi) = [1/2 - g(\xi)]/(2\xi)$, and $X_i = (2y_i - 1)\boldsymbol{x}'_{\boldsymbol{\gamma},i}\boldsymbol{\beta_\gamma}$.

The inequality in (2.6) holds for any ξ. Because $\log(p(y_i|\mathbf{x}_{\boldsymbol{\gamma},i}, \boldsymbol{\beta_\gamma}, \xi))$ has a quadratic form, when the prior of $\boldsymbol{\beta_\gamma}$ is Gaussian, the posterior $p(\boldsymbol{\beta_\gamma}|\boldsymbol{x}_{\boldsymbol{\gamma},i}, y_i, \mathbf{x}_{\boldsymbol{\gamma},i}, \xi)$ is also Gaussian. Potentially, one can choose ξ so that $p(\boldsymbol{\beta_\gamma}|\boldsymbol{x}_{\boldsymbol{\gamma},i}, y_i, \mathbf{x}_{\boldsymbol{\gamma},i}, \xi)$ provides a good approximation to $p(\boldsymbol{\beta_\gamma}|\boldsymbol{x}_{\boldsymbol{\gamma},i}, y_i, \mathbf{x}_{\boldsymbol{\gamma},i})$. From (2.6),

$$\int p(y_i|\mathbf{x}_{\boldsymbol{\gamma},i}, \boldsymbol{\beta_\gamma}, \xi) p(\boldsymbol{\beta_\gamma}) d\boldsymbol{\beta_\gamma} \geq \int p(y_i|\mathbf{x}_{\boldsymbol{\gamma},i}, \boldsymbol{\beta_\gamma}, \xi) p(\boldsymbol{\beta_\gamma}) d\boldsymbol{\beta_\gamma}. \tag{2.7}$$

Hence, the best possible approximation to the posterior utilizing the bound in (2.6) is achieved by choosing the value of ξ that maximizes the right hand side of this inequality. This maximization can proceed via the Jaakkola and Jordan (2000) EM algorithm. Letting $N(\boldsymbol{\mu}_{\boldsymbol{\gamma},0}, \boldsymbol{\Sigma}_{\gamma,\mathbf{0}})$ denote the prior for $\beta_{\boldsymbol{\gamma}}$, initializing $i = 0$, and choosing an arbitrary positive starting point for ξ, the algorithm iterates through the following steps for $i = 1, \ldots, n$:

1. Apply the following updating equations:

$$\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\gamma},i} = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\gamma},i-1} + 2|\lambda(\xi)|\boldsymbol{x}_{\boldsymbol{\gamma},i}\boldsymbol{x}'_{\boldsymbol{\gamma},i} \qquad \boldsymbol{\mu}_{\boldsymbol{\gamma},i} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma},i}\big\{\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\gamma},i-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + (y_i - 1/2)\boldsymbol{x}_{\boldsymbol{\gamma},i}\big\}$$

2. Update ξ by:
$$\xi^2 = \boldsymbol{x}'_{\boldsymbol{\gamma},i}\boldsymbol{\Sigma}_{\boldsymbol{\gamma},i}\boldsymbol{x}_{\boldsymbol{\gamma},i} + (\boldsymbol{x}'_{\boldsymbol{\gamma},i}\boldsymbol{\mu}_{\boldsymbol{\gamma},i})^2 \tag{2.8}$$
Go back to Step 1 and repeat until convergence (Jaakkola and Jordan (2000) claim 6-7 iterations is sufficient).

3. Let $i = i + 1$ and go to step 1 until all subjects are added.

The estimated posterior $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{Y}, \boldsymbol{X}_{\boldsymbol{\gamma}}) \overset{d}{=} N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \boldsymbol{\mu}_{\boldsymbol{\gamma},n}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma},n})$. In using this VB approximation to the posterior to obtain an approximation to the marginal likelihood via (2.5), we find substantial sensitivity to the value of $\boldsymbol{\beta}^*_{\boldsymbol{\gamma}}$. Such sensitivity has been noted in previous use of (2.5) in approximating marginal likelihoods, and we follow common practice in using the posterior mean of $\boldsymbol{\beta}$. This is convenient, as the posterior mean conveniently corresponds to $\boldsymbol{\mu}_{\boldsymbol{\gamma},n}$, the value obtained at the final iteration of the above EM algorithm. We refer to the resulting estimator of the marginal likelihood as the VB1 estimator.

2.2.2 Approach II

By the variational transformation (2.6), we can also obtain

$$p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}) \geq p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\xi}) = \int \prod_{i=1}^{n} p(y_i|\mathbf{x}_{\boldsymbol{\gamma},i}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \xi_i) p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}. \tag{2.9}$$

The right hand side of the inequality (2.9) provides an alternative estimator of the marginal likelihood, with the vector $\boldsymbol{\xi}$ chosen to maximize the lower bound. As in Section 3.1, the EM algorithm can be used to estimate the optimal value of $\boldsymbol{\xi}$. Initializing $\boldsymbol{\xi}$, the algorithm iterates as follows until convergence:

1. Update the estimated posterior covariance and mean as
$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma},0}^{-1} + 2\sum_{i=1}^{n} |\lambda(\xi_i)| \boldsymbol{x}_{\boldsymbol{\gamma},i}\boldsymbol{x}'_{\boldsymbol{\gamma},i} \qquad \boldsymbol{\mu}_{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\big\{\boldsymbol{\Sigma}_{\boldsymbol{\gamma},0}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma},0} + \sum_{i=1}^{n}(y_i - 1/2)\boldsymbol{x}_{\boldsymbol{\gamma},i}\big\}$$

2. Update the variational parameters $\boldsymbol{\xi}$ by:
$$\xi_i^2 = \boldsymbol{x}_{\boldsymbol{\gamma},i}^T\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\boldsymbol{x}_{\boldsymbol{\gamma},i} + (\boldsymbol{x}'_{\boldsymbol{\gamma},i}\boldsymbol{\mu}_{\boldsymbol{\gamma}})^2, \quad i = 1, \ldots, n, \tag{2.10}$$

After convergence, the approximation to the marginal likelihood is then

$$p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\xi}) = \prod_{i=1}^{n} g(\xi_i) \exp\left\{ -\sum_{i=1}^{n}\left(\frac{\xi_i}{2} + \lambda(\xi_i)\xi_i^2\right)\right\} \frac{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{1/2}}{|\boldsymbol{\Sigma}_{\boldsymbol{\gamma},0}|^{1/2}} \exp\left\{ \frac{\boldsymbol{\mu}'_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}} - \boldsymbol{\mu}'_{\boldsymbol{\gamma},0}\boldsymbol{\Sigma}_{\boldsymbol{\gamma},0}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma},0}}{2}\right\}. \tag{2.11}$$

We refer to this estimator as VB2. Jaakkola and Jordan (2000) proposed this approach as an alternative to VB1, but believed that VB1 is cleaner because optimizing the variational

parameters sequentially instead of jointly is usually better. However, we find that VB2 has the advantage of producing a much more stable estimator of the marginal likelihood than VB1, which has a disturbing degree of sensitivity to $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^*$. VB2 does take longer to compute, particularly in cases involving large number of subjects, which implies a high-dimensional $\boldsymbol{\xi}$.

3 Shotgun Stochastic Search

Hans (2005) and Hans *et al.* (2007) proposed the Shotgun Stochastic Search (SSS) algorithm as an alternative to SSVS for searching for high posterior probability models in cases in which the model space is defined by all possible subsets of a high-dimensional vector of predictors. Let $\boldsymbol{\gamma}' \in \eta(\boldsymbol{\gamma})$ denote the subset of Γ corresponding to those models in a neighborhood of $\boldsymbol{\gamma}$, defined to consist of all those models obtained by adding or deleting a single predictor or substituting the predictor for another from among the candidates.

The SSS algorithm searches the model space by iterating through the following steps a large number of times after choosing an initial model $\boldsymbol{\gamma}$:

1. Proceeding in parallel, calculate scores for all models in $\eta(\boldsymbol{\gamma})$, the neighborhood of the current model $\boldsymbol{\gamma}$, with these scores defined to be proportional to $p^*(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}_{\boldsymbol{\gamma}}) = p(\boldsymbol{\gamma})\hat{p}(\mathbf{Y} \mid \mathbf{X}_{\boldsymbol{\gamma}})$, where $\hat{p}(\mathbf{Y} \mid \mathbf{X}_{\boldsymbol{\gamma}})$ is an estimate of the marginal likelihood under model $\boldsymbol{\gamma}$. Note that one cannot calculate the posterior model probability $p(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}_{\boldsymbol{\gamma}})$ as the normalizing constant involves summing over all possible subsets.

2. Randomly select one new model $\boldsymbol{\gamma}'$ from $\eta(\boldsymbol{\gamma})$ by sampling with probabilities proportional to $p^*(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{X}_{\boldsymbol{\gamma}})^\alpha$, where $\alpha \in [0, 1]$ is an annealing parameter.

By using annealing, with the annealing parameter tuned based on the problem, one limits the tendency of SSVS algorithms to remain for long intervals in local regions of the model space. There is no approach currently available for optimally choosing the annealing parameter, and instead one can try different values and select the one which finds models with highest posterior probabilities in training runs. For example, Jones *et al.* (2005) use the SSS method for graphical model selection in a gene expression data set with 150 genes, and find an annealing parameter value of 50 to yield good performance relative to other values.

Empirically, the SSS algorithm has proven very efficient at finding the top models compared to MCMC methods. Hans *et al.* (2007) compared the SSS algorithm with the MCMC model composition (MC^3) algorithm (Madigan *et al.*, 1995; Raftery, 1996) and the Gibbs sampling algorithm (George and McCulloch, 1997) on a standard dataset and show that SSS is much better than the other two in terms of the accumulated posterior mass of the models visited given the same number of model evaluations. Based on a simulation study they also show empirically that the SSS algorithm improves over MCMC approaches in terms of the average number of steps needed before the true model is found.

4 Simulation Analysis

Unfortunately, it is very difficult to compare the three methods described in Section 2 (Laplace, VB1, VB2) theoretically. If the marginal likelihoods were analytically tractable, then one would not need these approximation methods, and it has proven difficult to theoretically justify the tightness of the VB lower bounds, as this is entirely problem dependent. Hence, we rely on simulations to assess relative performance. In assessing accuracy of the marginal likelihood approximations, one challenge is that we lack knowledge of the exact marginal likelihoods even for simulated data. To address this problem, we follow the approach of implementing importance sampling (IS) for a very large number of samples, with the resulting estimated marginal likelihood used as the gold standard. Up to about 20-30 dimensions, importance sampling is well known to produce very accurate results if sufficient numbers of samples are collected. However, such an approach is certainly not a useful method in practice in model search problems, as it requires a substantial computational burden to estimate the marginal likelihood for a single model, and when there are many models under consideration, it is necessary practically to rapidly approximate the marginal likelihood so one can spend more time searching for good models. We also tried a number of recently proposed Monte Carlo methods for estimating the marginal likelihood, but found that alternative approaches often did not converge to the same estimate even when using a 100,000s of samples. In contrast, one could easily judge convergence of importance sampling, and collect sufficient numbers of samples to produce a highly-accurate estimate.

In Section 4.1, we assess the relative performance of Laplace, VB1 and VB2 in estimating marginal likelihoods for logistic regression models in different simulated data sets. In Section 4.2, we then compare predictive performance of Bayesian model averaging using posterior model probabilities estimated under the three different approaches.

4.1 Accuracy of Marginal Likelihood Estimation

Let $g(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ correspond to the multivariate t distribution, with low degrees of freedom ($\nu = 3$) and with the mean and variance chosen as the VB2 approximated posterior mean and variance. Then, rewrite $p(\boldsymbol{Y}_{\boldsymbol{\gamma}} \mid \boldsymbol{X}_{\boldsymbol{\gamma}})$ in (1.2) as

$$p(\boldsymbol{Y} \mid \boldsymbol{X}_{\boldsymbol{\gamma}}) = \int \frac{p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{g(\boldsymbol{\beta}_{\boldsymbol{\gamma}})} g(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}. \tag{4.1}$$

By simulating samples of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ from $g(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$, the marginal likelihood can be estimated by

$$p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}) \approx \sum_{i=1}^{N} \frac{p(\boldsymbol{Y}|\boldsymbol{X}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(i)})p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(i)})}{g(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(i)})}, \tag{4.2}$$

where $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(i)}$ is the ith sample from $g(\boldsymbol{\beta})$. We use $N = 100,000$ samples, since we find this empirically to give a highly accurate estimate of the marginal likelihood.

To motivate our choice of $g(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$, first note that it is standard practice to use multivariate Gaussian approximations to the posterior, which are justified by Bernstein von Mises

theorems that ensure asymptotic normality of the posterior. However, in finite samples, the normal approximation may have insufficiently heavy tails to be valid as a proposal for use in importance sampling. For this reason, we follow the approach of inflating the heaviness of the tails to satisfying the condition that the tails of the proposal are at least as heavy as the tails of the target. The use of 3 degrees of freedom will lead to very heavy tails, so is a conservative choice.

In assessing accuracy of Laplace, VB1 and VB2 approximations, we simulated data sets under three different cases in which there were 6 candidate predictors, with the size of the true model equal to 1, 3, or 5 predictors. For each case, we simulated 100 data sets having $n = 50$ samples per data set, with the coefficients for the predictors that were included sampled independently from $N(0, 4^2)$. For one simulated data set for each model size, we used importance sampling (IS), Laplace, VB1 and VB2 to estimate the marginal likelihoods under each of the $2^6 = 64$ possible models, and we then sorted the models by the marginal likelihood estimates based on IS.

Figure 1 plots the IS, Laplace, VB1 and VB2 estimated log marginal likelihoods under the 64 sorted models in the model size 1 simulations, while Figures 2 and 3 present the corresponding results for the size 3 and size 5 simulations, respectively. Because the IS estimate was considered indistinguishable from the true marginal likelihood, all three plots indicate that the Laplace approximation is very close to the true marginal likelihood. Interestingly, the VB methods tend to be highly accurate for models with relatively low marginal likelihoods, but tend to substantially underestimate the marginal likelihood for good models. For example, when there is only one predictor, the 32 models that do not include the true predictor have a low marginal likelihood, and all the three approximation methods have very similar estimates to the importance sampling results. On the other hand, for the other 32 models that include the true predictor, the VB1 and VB2 estimates are both poor.

To avoid randomness from one data set, we have also tested the marginal likelihood estimation performance by different methods on all 100 data sets. The results are very similar to what we have shown in Figure 1, 2, and 3. As an example, Figure 4 shows the difference between the highly accurate IS estimate and the three fast approximate estimates averaging over the 100 simulated data sets in the model size 3 case. For each data set, we sort all the possible models by their estimated marginal likelihood by importance sampling, and compute the difference between importance sampling and the other three methods. Then for each rank of the model, we take the average of the differences. We can see from the plot that although Laplace approximation is underestimating a little bit the marginal likelihood, its estimation line is almost parallel to the estimation line made by importance sampling, which makes the model selection by Laplace approximation more robust. For the two variational methods, as we have addressed before, their estimation precision both become worse as the model becomes better, while VB2 is better than VB1.

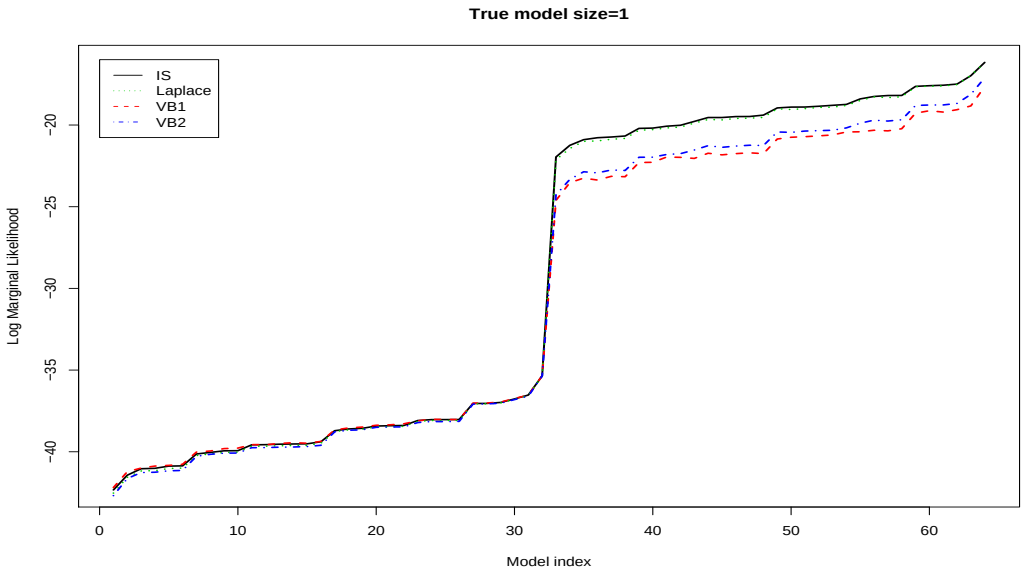

Figure 1: The estimated log-marginal likelihoods under importance sampling (IS), Laplace and the two variational Bayes methods (VB1, VB2) in a simulation case in which the true model size is 1. The models are ordered to be increasing in the IS estimated marginal likelihood.

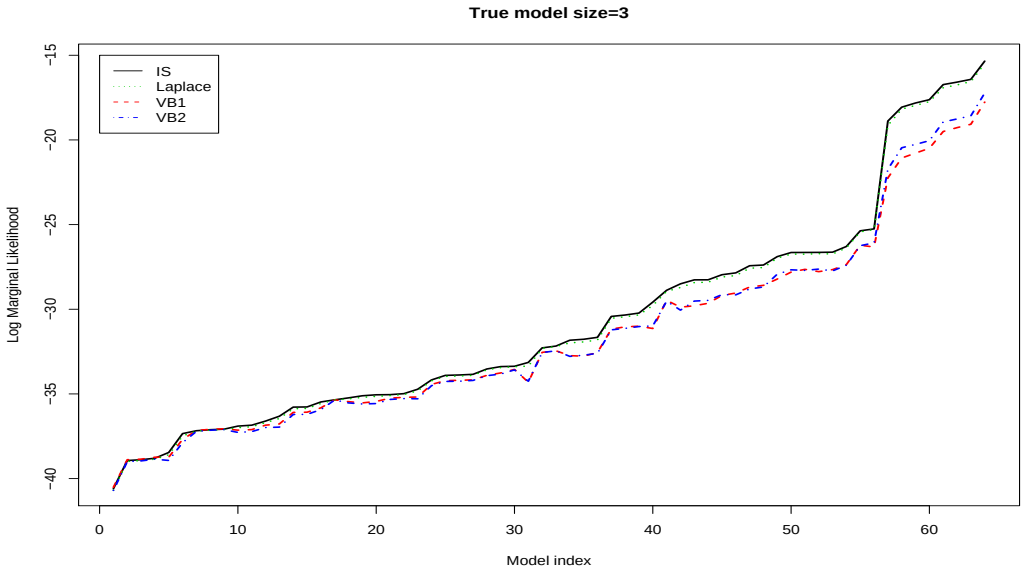

Figure 2: The estimated log-marginal likelihoods under IS, Laplace, VB1 and VB2 when the true model size is 3.

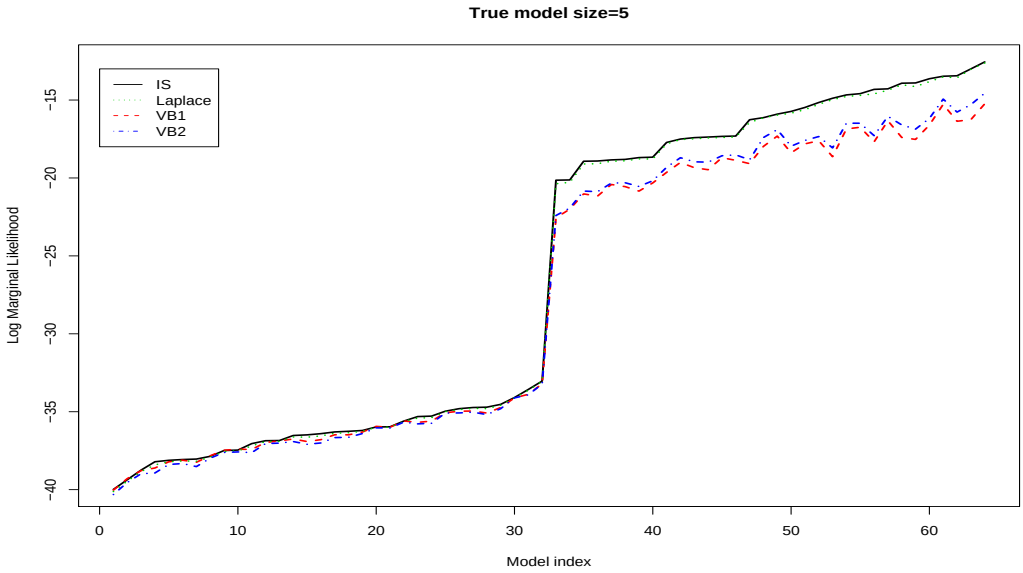

Figure 3: The estimated log-marginal likelihoods under IS, Laplace, VB1 and VB2 when the true model size is 5.

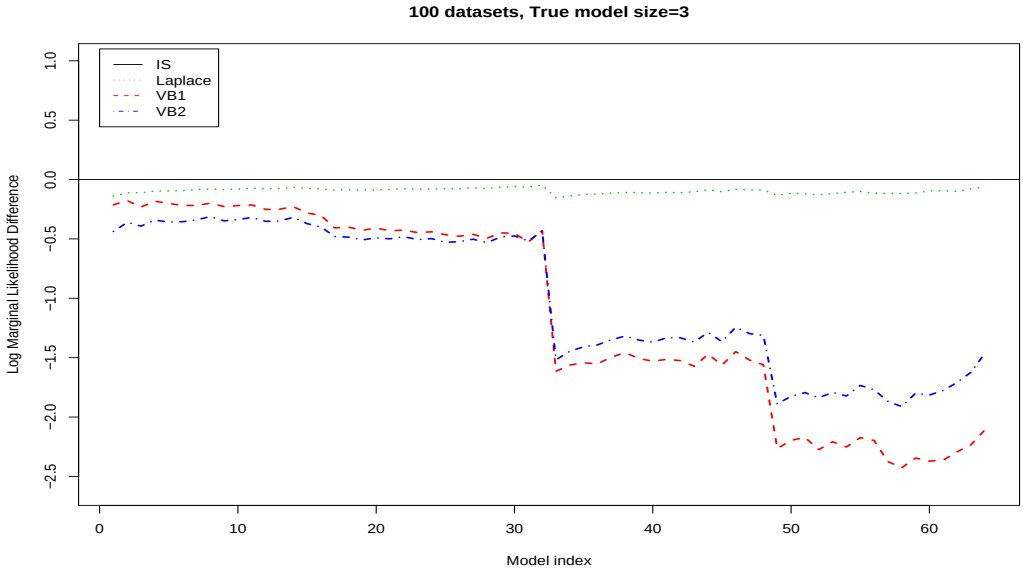

Figure 4: The average difference between the log-marginal likelihood estimated by IS and the estimates under Laplace, VB1 and VB2 for the 100 simulated data sets when the true model size is 3.

4.2 Prediction Performance by Bayesian Model Averaging

When the model space is large, there are typically many models with similar posterior probabilities, so that it is not ideal to base prediction on a single selected model, and model averaging is recommended. In particular, given predictors $\mathbf{X}^*$ for a new set of subject, the model-averaged predictive distribution of $\mathbf{y}^*$ is

$$p(\mathbf{y}^*|\boldsymbol{X}^*, \boldsymbol{Y}, \boldsymbol{X}) = \sum_{\boldsymbol{\gamma}\in\Gamma} p(\mathbf{y}^*|\boldsymbol{X}^*_{\boldsymbol{\gamma}})p(\boldsymbol{\gamma}|\boldsymbol{Y}, \boldsymbol{X}). \tag{4.3}$$

In binary data, when $p(y_{n+1}|\boldsymbol{x}_{n+1}, \boldsymbol{Y}, \boldsymbol{X}) > 0.5$, subject $n+1$ is predicted to have $y_{n+1} = 1$ and otherwise the subject is predicted to have $y_{n+1} = 0$.

To test the predictive performance of the three methods, we simulate a data set with 200 potential predictors and 500 samples, with the size of the true model to be 9. After using SSS to search for the top 50 models for each marginal likelihood estimation method, we implement model averaging to do out-of-sample predictions for an additional 2000 samples. The results show that the Laplace approximation has a misclassification rate of 307/2000, while both VB1 and VB2 have a misclassification rate of 305/2000. In contrast, conducting prediction based on maximization of the true logistic regression model resulted in a rate of 308/2000. Hence, all three methods did as well as a frequentist analysis under the true model.

The average of the estimated posterior probabilities of $y_i = 1$ ($y_i = 0$) for subjects in the test sample with $y_i = 1$ ($y_i = 0$) was 0.728 (0.803) under the Laplace approximation and 0.709 (0.781) for both VB1 and VB2. Furthermore, the root mean square errors (square root of the sum of the square of the difference between the true probabilities of $y_i = 1$ and the estimated probabilities) of Laplace approximation, VB1, and VB2 are 2.615, 3.348, and 3.351, respectively. Hence, there is some gain in predictive performance for the Laplace approximation relative to the VB approaches in a high dimensional logistic regression setting.

5 Daily Fecundability Data Analysis

5.1 Description of Data and Scientific Problem

We now apply SSS with the three different approximation methods to build a predictive model for the probability of conception in a menstrual cycle based on daily records of intercourse timing. Data were drawn from the European Study of Daily Fecundability (ESDF), which followed women using the sympto-thermal method of natural family planning, collecting daily data on intercourse, basal body temperature and characteristics of cervical mucus secretiions. We focus on intercourse data in a 19 day window indexed relative to the last day of hypothermia, which is a commonly used marker of the ovulation day that can be obtained based on the basal body temperature charts. The 19 window started 12 days prior to the marker of ovulation and ended 6 days after, which means that ovulation corresponds to day 13. Intercourse data consisted of a 0/1 indicator of intercourse for each day in each menstrual cycle under study.

As described by Bigelow *et al.* (2004) and Scarpa *et al.* (2006), increasing mucus score tends to be highly predictive of an increased day-specific probability of conception. Scarpa and Dunson (2007) used Bayesian variable selection combined with a decision-theoretic analysis to identify optimal rules for timing intercourse to achieve conception. Their analysis focused on simple rules based on the timing in the cycle, allowing for an interaction with timing and the effect of the mucus score on the probability of conception. However, given the use of typical SSVS methods in their analysis, it was not computationally possible to consider models that allow for interactions. In particular, although it is typically assumed that sperm introduced on different days commingle and then compete independently in attempting to fertilize the ovum, it is quite possible biologically that the independent competing risks assumption is not fully accurate biologically. However, in allowing interactions between the effects of intercourse on different days, one obtains an enormous number of possible regression models.

The European data base is particular suited to exploring different models, because it is very large compared to typical prospective studies of fecundability and collected quite detailed records. Data were available for 2832 cycles from 660 different couples. Previous analyses of day-specific conception probabilities have focused on simple, biologically-based competing risk models, which require conception probabilities to be zero if no intercourse is reported within the potentially fertile interval of the cycle (refer, for example, to Dunson and Stanford, 2005). Here, we instead focus on a logistic regression model and avoid the assumption that no reported intercourse implies zero conception probabilities, motivated by the fact that intercourse will be unreported some of the time. For example, most studies of this type have at least a few "immaculate conceptions" in which conception occurs in cycles with no reported intercourse.

5.2 High-dimensional Logistic Regression

We build a standard logistic regression model to study the relationship between intercourse and women's conception time. The response variable Y is set as the binary conception variable, and $\boldsymbol{X}$ consists of the 19-day intercourse variable, and all the second-order interactions between them. As a result, in total there are 190 potential predictors of Y in our logistic regression model. The prior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is $N(\mathbf{0}, \boldsymbol{I}_{\boldsymbol{\gamma}})$, which corresponded to a ridge regression shrinkage prior that expressed our view that the coefficients for the included predictors should have a low probability of falling outside of the interval within ± 2 of 0. To set the sparsity prior for model selection, we assume apriori that on average there are 10 predictors in the model ($\phi = 10/190$), and the annealing parameter for SSS is 1. For all the 2832 observations, we run SSS in parallel for 10,000 iterations using 50 CPU cores, and record the top models by the three marginal likelihood approximation methods. The top 5 models found by each method are listed in Table 1. We can also notice in the table that Laplace approximation is the fastest approximation method in the three methods.

Dunson *et al.* (1999) reported that the probability of conception is near zero unless intercourse occurs in a five day fertile interval ending on the day of ovulation. This result

Method	Top 5 Models	Score	Running Time (s)
Laplace	9 10 11 12 (9 10) (10 11) (11 12)	-1147.41	2140
	9 10 11 12 (6 8) (10 11) (11 12)	-1147.59	
	9 10 11 12 (8 13) (9 10) (10 11) (11 12)	-1147.65	
	9 10 11 12 (3 14) (9 10) (10 11) (11 12)	-1147.75	
	9 10 11 12 (9 10) (9 12) (10 11) (11 12)	-1147.97	
VB1	10 11 12 (3 9) (10,11) (11,12)	-1153.32	9304
	9 10 11 12 (9 10) (10 11) (11 12)	-1153.55	
	10 11 12 (3 9) (8 13) (10 11) (11 12)	-1153.79	
	9 10 11 12 (9 10) (11 12)	-1153.82	
	9 10 11 12 (9 10) (9 12) (10 11) (11 12)	-1153.86	
VB2	9 10 11 12 (9 10) (10 11) (11 12)	-1148.84	72020
	9 10 11 12 (6 8) (9 10) (10 11) (11 12)	-1149.17	
	9 10 11 12 (9 12) (10 11) (11 12)	-1149.19	
	10 11 12 (3 9) (10 11) (11 12)	-1149.20	
	9 10 11 12 (6 8) (9 10) (11 12)	-1149.28	

Table 1: The top 5 models of the daily fecundability data obtained by SSS and the three marginal likelihood approximation methods. The scores are proportional to the posterior probabilities of the top models which are obtained by the Laplace, VB1 and VB2.

is consistent with days 8-13 being included in the model for the conception probability. To our knowledge, all previous analyses of day-specific conception probabilities have assumed independent competing risks, therefore have not allowed interactions between intercourse acts occurring on different days. Interestingly, our results suggest that a narrower four day fertile interval ending one day prior to the estimate day of ovulation is appropriate (days 9, 10, 11 and 12), but with interactions of intercourse occurring between days (9,10), (10,11), and (11,12). (The sparsity prior plays a role here. If the sparsity prior is less sparse, then the model will be richer, and days 8, 13 will be in the model with a lot of other non-appealing days or interactions). These interactions all have high marginal inclusion probabilities. It can also be suggested that intercourse happening in two continuous days may increase the probability of conception. This is exactly shown by the coefficients estimated in the top models containing the interactions of days (10,11) and (11,12), while for the interaction (9,10), the probability of conception with intercourse in both days is less than the intercourse in day 10 only, but more than that in day 9 only, because the influence of day 9 is relatively low (not sure about the reason). On the other hand, several top models also select some other interactions, such as (6,8) and (8,13), where the days themselves are not included in the model. Those interactions usually have a positive coefficient, which means they may

Method (True Value)	50%	60%	70%	80%	90%
Laplace (Y=1)	0.1987	0.2047	0.2091	0.2108	0.2025
VB1 (Y=1)	0.2023	0.2101	0.2141	0.2168	0.2105
VB2 (Y=1)	0.2021	0.2089	0.2141	0.2171	0.2100
Laplace (Y=0)	0.1347	0.1348	0.1399	0.1415	0.1451
VB1 (Y=0)	0.1465	0.1459	0.1520	0.1541	0.1577
VB2 (Y=0)	0.1466	0.1460	0.1519	0.1541	0.1576

Table 2: The mean predictive probabilities of conception for true values of response that are equal to 1 or 0 in the test data respectively. Different training sample sizes are tried (50% - 90%).

also contribute to the conception a little.

To assess the model selection performance of the three methods, we use cross-validation to compare the three methods. First, 10% of the data, which mean 283 observations, are randomly picked out to form the test data. For the other 2549 observations, we try different sample size of the training data, i.e., randomly select 1416, 1699, 1982, 2266 and 2549 observations (50% - 90% of the total 2832 cycles) to form 5 individual training data sets. To compare different approximation methods for different sizes of the training data sets, we first use SSS to do the model selection, and then use Bayesian model averaging on the top 50 models to obtain predictive probabilities of conception for the test data. The other settings such as priors are the same as the analysis for the full data.

In this cross-validation analysis, it is not suitable to use 0.5 threshold to make prediction on Y based on the predictive probabilities, because the percentage of 1's in the population is so low (about 0.153), otherwise it may give a prediction set with all 0's. A nice way to look at is to compare the mean predictive probabilities of conception given the true value of Y in the test data for different sizes of the training data, and different approximation methods (Table 2).

From Table 2, the VB approximations does not have noticeably worse performance than Laplace in terms of prediction. This is because VB1 and VB2 order the models almost correctly, but just under-estimate the marginal likelihoods for the better models, flattening out the posterior probabilities across the better models. This may have a modest impact on predictive performance that may not show up compellingly in the cross validation exercise, but clearly in the previous simulation analysis. Also, it is quite intriguing that Laplace usually gives less mean predictive probabilities of conception than VB1 and VB2, and the predictive probabilities using training data with sizes from 50% to 80% keep increasing in each row, while we think what happens to the 90% training data is just because of randomness.

6 Conclusion

This article was originally motivated by the goal of using variational Bayes approximations to improve methods for high-dimensional model selection and averaging. The VB approaches have been increasingly widely used in machine learning applications, and have conceptual appeal in resulting from maximization of a formal lower bound on the marginal likelihood. Although the tightness of the lower bound is in general quite difficult to assess theoretically, the good performance of VB procedures in various predictive settings has been reassuring. However, to our knowledge, the performance of VB relative to traditional Laplace methods of estimating the marginal likelihood with the goal of model selection has not been assessed.

In the setting of logistic regression model selection, this article uses simulations to compare the accuracy of Laplace and two types of VB approximations (VB1, VB2). We find that Laplace is highly accurate, while VB1 and VB2 have a disturbing tendency to underestimate the marginal likelihood for high posterior probability models. When the goal is model selection or accurate estimate of posterior model probabilities, this type of underestimation is particularly troubling, since one will under-estimate the probabilities for good models and over-estimate the probabilities for bad models. On the positive side, the VB approaches do tend to rank the models appropriately; it is only the scores that are misestimated. Perhaps for this reason, we have observed only a slight decrease in predictive performance for the VB approaches relative to Laplace in settings with few important predictors and a high-dimensional set of candidates.

The performance of the VB approach is critically dependent on the accuracy of the product factorization of the joint posterior. This gives us a clue as to the reason for our results. In particular, it is our expectation in the variable selection setting that the product factorization provides a good approximation for bad models, since bad models correspond to exclusion of important predictors and inclusion of predictors that actually have no impact. However, we expect that the approximation breaks down when several important predictors are included and the coefficients for these predictors are correlated *a posteriori*, as one would typically expect in regression models. Because the implementation of VB methods is so tied to the product factorization, it is difficult to entirely eliminate this problem. However, one strategy is to attempt a factorization under a parameterization chosen to reduce posterior dependence.

Acknowledgments

The authors are very grateful to the referee for helpful comments which improved this paper.

References

Bigelow, J., Dunson, D., Stanford, J., Ecochard, R., Gnoth, C., and Colombo, B. (2004). Mucus observations in the fertile window: a better predictor of conception than timing

of intercourse. *Human Reproduction* **19**, 889–892.

Casella, G. and Moreno, E. (2006). Objective bayesian variable selection. *Journal of the American Statistical Association* **101**, 157–167.

Chen, M. (2005). Computing marginal likelihoods from a single mcmc output. *Statistica Neerlandica* **59**, 16–29.

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

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the metropolis-hastings output. *Journal of the American Statistical Association* **96**, 270–281.

Consonni, G. and Marin, J. (2006). Mean field variational bayesian inference for latent variable models. *Computational Statistics and Data Analysis* To appear.

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

DiCiccio, T. J., Kass, R. E., and Wasserman, L. (1997). Computing bayes factors by combining simulation and asymptitoc approximations. *Journal of the American Statistical Association* **92**, 903–915.

Dunson, D., Baird, D., Wilcox, A., and Weinberg, C. (1999). Day-specific probabilities of clinical pregnancy based on two studies with imperfect measures of ovulation. *Human Reproduction* **14**, 1835–1839.

Dunson, D., Colombo, B., and Baird, D. (2002). Changes with age in the level and duration of fertility in the menstrual cycle. *Human Reproduction* **17**, 1399–1403.

Dunson, D. and Stanford, J. (2005). Bayesian inferences on predictors of conception probabilities. *Biometrics* **61**, 126–133.

Gelfand, A. and Dey, D. (1994). Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society. Series B (Methodological)* **56**, 3, 501–514.

Gelfand, A. and Smith, A. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85**, 410, 398–409.

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

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

George, E. and McCulloch, R. (1997). Approaches for bayesian variable selection. *Statistica Sinica* **7**, 339–373.

Han, C. and Carlin, B. (2001). Markov chain monte carlo methods for computing bayes factors: A comparative review. *Journal of the American Statistical Association* **96**, 1122–1132.

Hans, C. (2005). *Regression model search and uncertainty with many predictors*. PhD. Thesis, Duke University.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* **102**, 507–516.

Ishwaran, H. and Rao, J. (2005). Spike and slab variable selection: Frequentist and bayesian strategies. *Annals of Statistics* **33**, 730–773.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing* **10**, 25–37.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science* **20**, 388–400.

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review/Revue Internationale de Statistique* 215–232.

Ntzoufras, I., Dellaportas, P., and Forster, J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference* **111**, 165–180.

Raftery, A. (1996). Approximate bayes factors and accounting for model uncertainty in generalised linear models. *Biometrika* **83**, 2, 251.

Scarpa, B. and Dunson, D. (2007). Bayesian methods for searching for optimal rules for timing of intercourse to achieve pregnancy. *Statistics in Medicine* **26**, 1920–1936.

Scarpa, B., Dunson, D., and Colombo, B. (2006). Cervical mucus secretions on the day of intercourse: An accurate marker of highly fertile days. *European Journal of Obstetrics Gynecology and Reproductive Biology* **125**, 72–78.

Srebro, N. and Jaakkola, T. (2003). Weighted low-rank approximations. *Proceedings of the Twentieth International Conference on Machine Learning (ICML)* .

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

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes Factors Using a Generalization of the Savage-Dickey Density Ratio. *Journal of the American Statistical Association* **90**, 430.

Wang, B. and Titterington, D. M. (2005). Inadequacy of interval estimates corresponding to variational bayesian approximations. *10th International workshop on Artifical Intelligence and Statistics* .

Wang, X. and George, E. (2007). Adaptive bayesian criteria in variable selection for generalized linear models. *Statistica Sinica* **17**, 667–690.

Wilcox, A., Dunson, D., and Baird, D. (2000). The timing of the "fertile window" in the menstrual cycle: day specific estimates from a prospective study. *British Medical Journal* **321**, 1259–1262.