involve
a journal of mathematics

A new go-to sampler for Bayesian probit regression

Scott Simmons, Elizabeth J. McGuffey and Douglas VanDerwerken

msp

2020 vol. 13, no. 1

INVOLVE 13:1 (2020)
dx.doi.org/10.2140/involve.2020.13.77

A new go-to sampler for Bayesian probit regression

Scott Simmons, Elizabeth J. McGuffey and Douglas VanDerwerken

(Communicated by Jem Noelle Corcoran)

This paper introduces an independent-proposal Metropolis–Hastings sampler for Bayesian probit regression. The Gibbs sampler of Albert and Chib has been the default sampler since its introduction in 1993. We conduct a simulation study comparing the two methods under various combinations of predictor variables and sample sizes. The proposed sampler is shown to outperform that of Albert and Chib in terms of efficiency measured through autocorrelation, effective sample size, and computation time. We then demonstrate performance of the samplers on real data applications with analogous results.

1. Introduction

Modeling binary response data is important in statistics and related fields of biostatistics and econometrics. The classical approach is to fit a logistic or probit regression model via maximum likelihood and to conduct inference based on asymptotic theory. A thorough summary of this approach is given by [Amemiya 1981]. When the sample size n is small, however, reliance on asymptotic theory is dubious; for example, [Griffiths et al. 1987] shows that the maximum likelihood estimator (MLE) exhibits nonnegligible bias in finite samples.

One solution is to take a Bayesian approach. The Bayesian paradigm is based on Bayes' rule, which in our modeling context may be generally expressed as

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{\pi(\boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})} \propto \pi(\boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta}),$$

where $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the data likelihood; $\pi(\boldsymbol{\theta})$ is the prior distribution assigned to $\boldsymbol{\theta}$, representing one's a priori beliefs about $\boldsymbol{\theta}$; $p(\boldsymbol{y})$ is the marginal likelihood; and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the posterior distribution, describing one's updated beliefs about $\boldsymbol{\theta}$. Interest lies in conducting inference on $\boldsymbol{\theta}$ based on the posterior distribution.

For many data likelihoods, there exists a prior distribution that leads to a posterior distribution of the same family that may be expressed in closed form. There is no

MSC2010: 62C10, 62F15.
Keywords: probit regression, Markov chain Monte Carlo, Gibbs sampling, Metropolis–Hastings.

such conjugate prior for either the logistic or probit likelihoods, so computational techniques must be utilized to summarize and/or sample from the posterior distribution. For instance, [Zellner and Rossi 1984] recommends numerical integration when the number of predictors p is small and importance sampling with a multivariate t instrumental distribution otherwise.

The advent of Gibbs sampling [Geman and Geman 1984; Gelfand and Smith 1990] introduced a whole class of new methods for doing posterior inference. In its simplest form, Gibbs sampling generates random draws from a multivariate posterior distribution by iteratively generating univariate draws from the full conditional distributions $p(\theta_k \mid \boldsymbol{\theta}_{-k}, \boldsymbol{y})$, where $\boldsymbol{\theta}_{-k}$ is the parameter vector $\boldsymbol{\theta}$ excluding the k-th entry. Zeger and Karim [1991] were the first to apply Gibbs sampling to a logistic regression model. However, in this approach not all of the full conditional distributions are available in closed form.

In a now landmark paper, Albert and Chib [1993] introduced a novel Gibbs sampler for sampling from the posterior distribution of a probit regression model with a multivariate normal prior. Thanks to a clever data augmentation technique, the full conditional distributions in the Albert and Chib Gibbs sampler are all named distributions. Albert and Chib's method can be used to obtain *approximate* posterior samples from the logistic regression model, but it took more than a decade for researchers to devise an *exact* (i.e., asymptotically convergent) data-augmentation-based Gibbs sampler for the logistic regression model; several now exist [Holmes and Held 2006; Frühwirth-Schnatter and Frühwirth 2007; Polson et al. 2013]. Holmes and Held [2006], in addition to their contribution to Bayesian logistic regression, also introduced an ostensible improvement to Albert and Chib's probit Gibbs sampler in the form of a joint update of the regression and auxiliary parameters.

In this paper, we draw attention to a Markov chain Monte Carlo (MCMC) sampling algorithm related to the importance sampling approach of [Zellner and Rossi 1984] and demonstrate that it performs favorably in a wide variety of settings when compared to the Albert and Chib sampler. In particular, we recommend that samples from the probit posterior be obtained using an independence Metropolis–Hastings (MH) sampler with a multivariate t proposal distribution centered at the posterior mode and having variance equal to the inverse Hessian evaluated at the mode. Variations of this sampler have been mentioned in the literature, but to our knowledge no one has undertaken to systematically compare it to Albert and Chib's.

Although our algorithm is closely related to the importance sampling scheme of [Zellner and Rossi 1984], one minor difference is that we obtain actual samples from the posterior, allowing for natural comparison to the Albert and Chib sampler, whereas Zellner and Rossi approximated integrals with appropriately weighted draws from the multivariate t-distribution. Perhaps a more notable difference is that we use the inverse Hessian evaluated at the *mode* instead of the (moderately scaled) inverse Hessian evaluated at the MLE. This difference is nontrivial when the contribution from the prior is considerable.

While the idea of using an independence MH sampler with multivariate t proposal was not explicitly discussed in [Albert and Chib 1993], the authors do suggest that a difficult full conditional may be approximated by a normal or t-distribution centered at the mode and having variance determined by the curvature at the mode; however, normality of the posterior itself is dismissed for small n. In this paper we show that even in the small-n scenario the independence MH sampler is often more efficient.

Polson et al. [2013] discuss the independence MH sampler at length, but in the context of logistic regression, as opposed to probit. They acknowledge that MH was faster than the auxiliary-variable Gibbs sampler in five of eight data sets studied, but warn against its use when the model has a complex prior structure, as is the case for, e.g., mixed models, factor models, and models with spatiotemporal dependence. We realize that similar arguments could be made in favor of the probit Gibbs sampler, but this does not negate the importance of our contribution. We hope to show that the deference which many practicing Bayesians give the Albert and Chib algorithm in the context of "vanilla" probit regression might be better given to the independence MH sampler. To take an extreme example, [Greene 2012, p. 713] goes so far as to all but reject the Bayesian paradigm because he finds the Albert and Chib algorithm too slow—as if no other sampler were available! Even for the more complicated settings alluded to in [Polson et al. 2013], hybrid Metropolis–Hastings-within-Gibbs samplers may yet provide the "best of both worlds"; see, for example, [Gamerman 1997; Geweke and Tanizaki 2001]. Our work offers insight into the kinds of gains that may be attained by such hybrid approaches.

The paper will proceed as follows. In Section 2, we introduce the independence MH sampler we endorse herein and review Albert and Chib's auxiliary-variable approach. We also provide a brief and candid analysis of the computational efficiency of the adaptation by [Holmes and Held 2006]. Section 3 presents a simulation study that compares the performance of the proposed sampler to Albert and Chib's across a wide variety of simulated data sets, while Section 4 describes performance of the samplers across several probit regression examples taken from the literature. We close with a discussion of the results. The R code and data sets used for the simulation study and applications are included in the online supplement.

2. MCMC samplers for Bayesian probit regression

2.1. ***Albert and Chib Gibbs sampler.*** In the context of binary probit regression, assume that we have n independent binary random variables, $Y_i \overset{\text{ind}}{\sim} \text{Bern}(p_i)$. Each $p_i = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{x}_i'$ is the $1 \times (p+1)$ row vector of p covariates (with intercept) for observation i, $\boldsymbol{\beta}$ is a $(p+1) \times 1$ column vector of regression coefficients, and Φ

is the standard normal cdf. We assume a priori that $\boldsymbol{\beta} \sim \pi$, with π usually chosen to be $\mathcal{N}(b_0, B_0)$. (Here, we adopt the common practice of allowing π to represent both the pdf and cdf, as appropriate.) The likelihood is given as the product of Bernoulli pmfs. By Bayes' rule, the posterior distribution is

$$p(\boldsymbol{\beta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) p(\boldsymbol{y} \mid \boldsymbol{\beta}) = \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\beta}, \boldsymbol{x}_i)$$

$$= \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})^{y_i} (1 - \Phi(\boldsymbol{x}_i'\boldsymbol{\beta}))^{1-y_i}.$$

There is no conjugate prior for $\boldsymbol{\beta}$. To facilitate Gibbs sampling, Albert and Chib introduced a latent variable $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$, where $Z_i \overset{\text{ind}}{\sim} \mathcal{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1)$. Letting $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ otherwise ensures that each Y_i is independent and marginally Bern(p_i), as desired. The joint posterior of $\boldsymbol{Z}$ and $\boldsymbol{\beta}$ is

$$p(\boldsymbol{Z}, \boldsymbol{\beta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) p(\boldsymbol{Z} \mid \boldsymbol{\beta}) p(\boldsymbol{y} \mid \boldsymbol{Z}) = \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} p(Z_i \mid \boldsymbol{\beta}, \boldsymbol{x}_i) p(y_i \mid Z_i)$$

$$= \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \phi(Z_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, 1)[\mathbf{1}(y_i = 1)\mathbf{1}(Z_i > 0) + \mathbf{1}(y_1 = 0)\mathbf{1}(Z_i \le 0)],$$

where $\phi(Z_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ is the normal pdf with mean $\boldsymbol{x}_i'\boldsymbol{\beta}$ and variance 1, evaluated at Z_i. This joint posterior cannot be sampled from directly. However, the full conditional distributions are of nice form. In particular, the full conditional distribution of $\boldsymbol{\beta}$ given $\boldsymbol{Z}$ is

$$p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{y}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \phi(Z_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}, 1),$$

which is a multivariate normal distribution when π is normal. The full conditional distribution of $\boldsymbol{Z}$ given $\boldsymbol{\beta}$ is the product of the marginal full conditional distributions, the i-th of which is $p(Z_i \mid \boldsymbol{\beta}, \boldsymbol{y})$, where

$$(Z_i \mid \boldsymbol{\beta}, \boldsymbol{y}) \sim \begin{cases} \mathcal{TN}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1, 0, \infty) & \text{if } y_i = 1, \\ \mathcal{TN}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1, -\infty, 0) & \text{if } y_i = 0. \end{cases}$$

Here $\mathcal{TN}(\boldsymbol{x}_i'\boldsymbol{\beta}, 1, a, b)$ is the normal distribution truncated to (a, b). Once the full conditionals are derived, the Gibbs sampling algorithm is straightforward. First, initialize $\boldsymbol{\beta}^{(0)}$, typically at either the maximum likelihood estimate or posterior mode. Then draw $\boldsymbol{Z}^{(1)}$ from the full conditional distribution of $\boldsymbol{Z} \mid \boldsymbol{\beta}^{(0)}, \boldsymbol{y}$. Next, draw $\boldsymbol{\beta}^{(1)}$ from the full conditional distribution of $\boldsymbol{\beta} \mid \boldsymbol{Z}^{(1)}, \boldsymbol{y}$. This process is iterated until convergence [Geman and Geman 1984].

2.2. ***Holmes and Held Gibbs sampler.*** The Albert and Chib Gibbs sampler can be inefficient because of strong posterior correlation between $\boldsymbol{\beta}$ and $\boldsymbol{Z}$. Holmes and

Held [2006] adapted the Albert and Chib sampler by performing a joint update of the regression and auxiliary parameters. This is possible because of the factorization

$$p(\boldsymbol{\beta}, \boldsymbol{Z} \mid \boldsymbol{y}) = p(\boldsymbol{Z} \mid \boldsymbol{y}) p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{y}),$$

where $p(\boldsymbol{\beta} \mid \boldsymbol{Z}, \boldsymbol{y})$ is the full conditional given above and $p(\boldsymbol{Z} \mid \boldsymbol{y})$ is the marginal distribution of $\boldsymbol{Z}$ with $\boldsymbol{\beta}$ integrated out. Holmes and Held show that $p(\boldsymbol{Z} \mid \boldsymbol{y})$ is a multivariate truncated normal, which, while difficult to sample from directly, does allow for straightforward Gibbs sampling. In particular, $p(Z_i \mid \boldsymbol{Z}_{-i}, \boldsymbol{y})$ has a univariate truncated normal distribution. The authors claim that the additional computational burden of the joint sampling step is small compared with the efficiency gained from having less-correlated draws. In particular, they note that the innermost for-loop requires only minor adjustment to sample $p(Z_i \mid \boldsymbol{Z}_{-i}, \boldsymbol{y})$ instead of $p(Z_i \mid \boldsymbol{\beta}, \boldsymbol{Z}_{-i}, \boldsymbol{y})$ [Holmes and Held 2006, p. 160].

However, the authors do not mention the fact that under Albert and Chib's setup, sampling from the full conditional distributions $p(Z_i \mid \boldsymbol{\beta}, \boldsymbol{Z}_{-i}, \boldsymbol{y})$ can be vectorized, since the Z_i's are conditionally independent (rendering the conditioning on $\boldsymbol{Z}_{-i}$ unnecessary). Thus, there is no need for the extra for-loop that their joint sampling requires. In a compiled language, this may make little difference; but in an interpreted language like R — the language of choice for many if not most practicing statisticians—the difference is tremendous. In the R simulation presented by [Kapourani 2018], for example, the Albert and Chib Gibbs sampler acquires 10,000 samples in about 3 seconds on a standard laptop computer, whereas the Holmes and Held Gibbs sampler acquires 10,000 samples in about 84 seconds. The effective sample size for Holmes and Held is indeed greater (~ 1000 versus ~ 500), but a 2-fold increase in effective sample size is hardly worth a 28-fold increase in computation time.

There are, of course, compiled-language implementations of Bayesian probit regression. One example of such an implementation of Albert and Chib's Gibbs sampler is the `MCMCprobit` function within the MCMCpack R package [Martin et al. 2011], which interfaces with R but outsources the actual sampling to C++. We suspect that a similar implementation of the Holmes and Held Gibbs sampler would mitigate the computational discrepancies noted above but we are not aware of an R package with this functionality. Because we use R for implementing the proposed method, we compare the proposed method only to the Albert and Chib sampler in terms of efficiency. Both samplers are coded using appropriate vectorization in R only.

2.3. ***Independent-proposal MH.*** Holmes and Held [2006] have noted that the Albert and Chib Gibbs sampler can be slow to converge when $\boldsymbol{Z}$ and $\boldsymbol{\beta}$ are highly correlated. The proposed sampler is intended to overcome this inefficiency. It samples from $p(\boldsymbol{\beta} \mid \boldsymbol{y})$ using the Metropolis–Hastings algorithm with independent

$t_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ proposals,[1] where $\boldsymbol{\mu}$ is the posterior mode and $\boldsymbol{\Sigma}$ is the inverse Hessian evaluated at $\boldsymbol{\mu}$. The mode can be found with the Newton–Raphson algorithm, which requires the gradient and Hessian of the log posterior. Importantly, these are both available in closed form.

The sampler proceeds as follows. We initialize $\boldsymbol{\beta}^{(0)}$ at $\boldsymbol{\mu}$. Then, we draw a proposal, $\boldsymbol{\beta}^*$ from $t_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (with pdf denoted by g), and set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^*$, with probability α, and $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$ with probability $1-\alpha$, where

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\beta}^* \mid \boldsymbol{y}) g(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^*)}{p(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{y}) g(\boldsymbol{\beta}^* \mid \boldsymbol{\beta}^{(t)})}\right\} = \min\left\{1, \frac{p(\boldsymbol{\beta}^* \mid \boldsymbol{y}) g(\boldsymbol{\beta}^{(t)})}{p(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{y}) g(\boldsymbol{\beta}^*)}\right\}.$$

Note that the second equality is justified by the fact that the proposals are independent of the current state; hence the conditioning is superfluous. Since the proposed jump, $\boldsymbol{\beta}^*$, is independent of the current state, $\boldsymbol{\beta}^{(t)}$, the average jump size is relatively large. This makes for lower autocorrelation between draws and requires fewer draws to attain a desired effective sample size. It is well known that when the sample size n is large compared to the number of predictors p, the posterior is well approximated by a normal distribution, so an independent-proposal MH algorithm with multivariate t proposals will do well. This paper investigates how well the approximation works when n is medium or small. If, for certain combinations of n and p, the $t_5(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a very poor approximation, performance of the proposed sampler will suffer and Albert and Chib's method may be preferred. The next two sections provide evidence that the proposed sampler is often better, however.

3. Simulation study

3.1. ***Settings.*** We compare the Albert and Chib and independent-proposal MH samplers on a variety of simulated data sets which were created by varying the number of predictors (p) and sample size (n). For a given data set, the prior standard deviation (σ) was also varied. Comparisons were made for all $4^3 = 64$ combinations of $p \in \{2, 5, 10, 20\}$, $n \in \{5p, 10p, 25p, 50p\}$, and $\sigma \in \{0.1, 1, 10, 100\}$. Predictors were taken from a single 1000×20 matrix $\boldsymbol{X}$ with independent rows. Column vectors were sampled and assigned as follows:

$$\begin{array}{llll}
X_1 \sim \text{Unif}(-1,1), & X_6 = X_5^3, & X_{11} \sim \text{Unif}(0,2), & X_{16} \sim \text{Unif}(0,6), \\
X_2 = X_1^2, & X_7 \sim \text{Unif}(-5,1), & X_{12} \sim \text{Unif}\left(-1, \tfrac{1}{2}\right), & X_{17} \sim \text{Exp}\left(\tfrac{1}{2}\right), \\
X_3 \sim \text{Exp}(5), & X_8 = X_7^2, & X_{13} = X_{11}^2, & X_{18} \sim \text{Exp}(2), \\
X_4 = X_3^2, & X_9 \sim \text{Exp}\left(\tfrac{1}{3}\right), & X_{14} = X_{12}^2, & X_{19} \sim \text{Unif}(-1,1), \\
X_5 \sim \text{Exp}\left(\tfrac{1}{5}\right), & X_{10} = X_9^2, & X_{15} \sim \mathcal{N}(3,1), & X_{20} = X_{19}^3;
\end{array}$$

[1] We chose the t-distribution with five degrees of freedom because it approximates the asymptotic normal posterior distribution but has relatively heavy tails, enabling approximation in the small-n case as well.

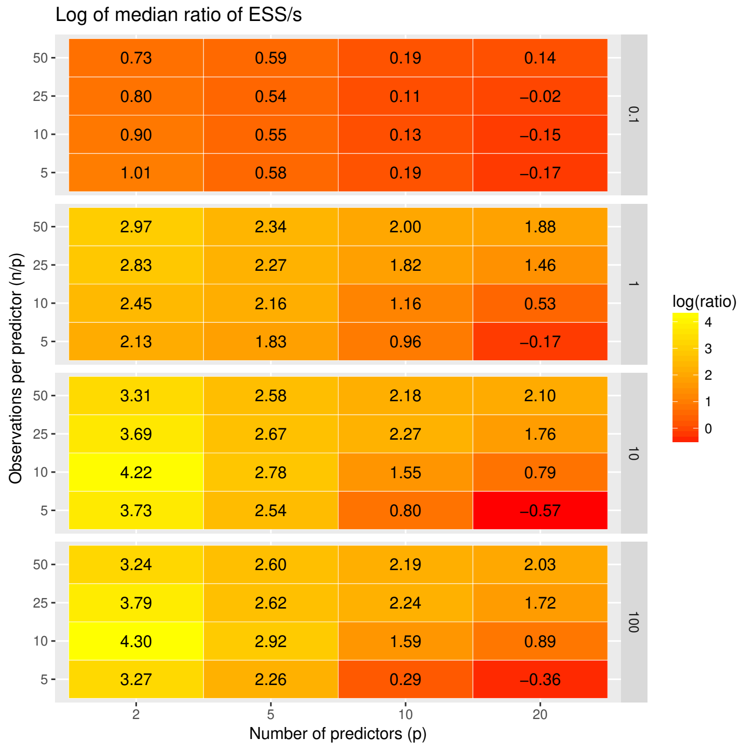

Figure 1. The effective sample size per second of computation time is almost always better for the proposed method than for Albert and Chib Gibbs sampling. The few exceptions correspond to situations with either unrealistically narrow priors or unrealistically few observations per predictor.

and these were then standardized to have mean 0 and variance 1. For a given (p, n) pair, the first p columns and the first n rows of $\boldsymbol{X}$ constituted the predictor matrix, the true parameter vector was the first p elements of the 20-element vector $\boldsymbol{\beta} = (+1, -1, +1, -1, \ldots, -1)$, and n binary responses were generated as Bernoulli realizations having expected values $\Phi(\boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1})$. MCMC samples were then obtained for each $\sigma \in \{0.1, 1, 10, 100\}$ using the same data with the different priors. Data were generated anew for each of 50 iterations, for a total of $50 \times 4^3 = 3200$ $(n, p, \sigma, \text{iter})$ combinations.

For each $(n, p, \sigma, \text{iter})$ combination, both samplers were initialized at the posterior mode and run for 10,999 iterations, the first 999 of which were discarded as burn-in. The worst-dimension autocorrelation,[2] the effective sample size (ESS), and

[2]The worst-dimension autocorrelation is the maximum autocorrelation observed for any univariate β after burn-in.

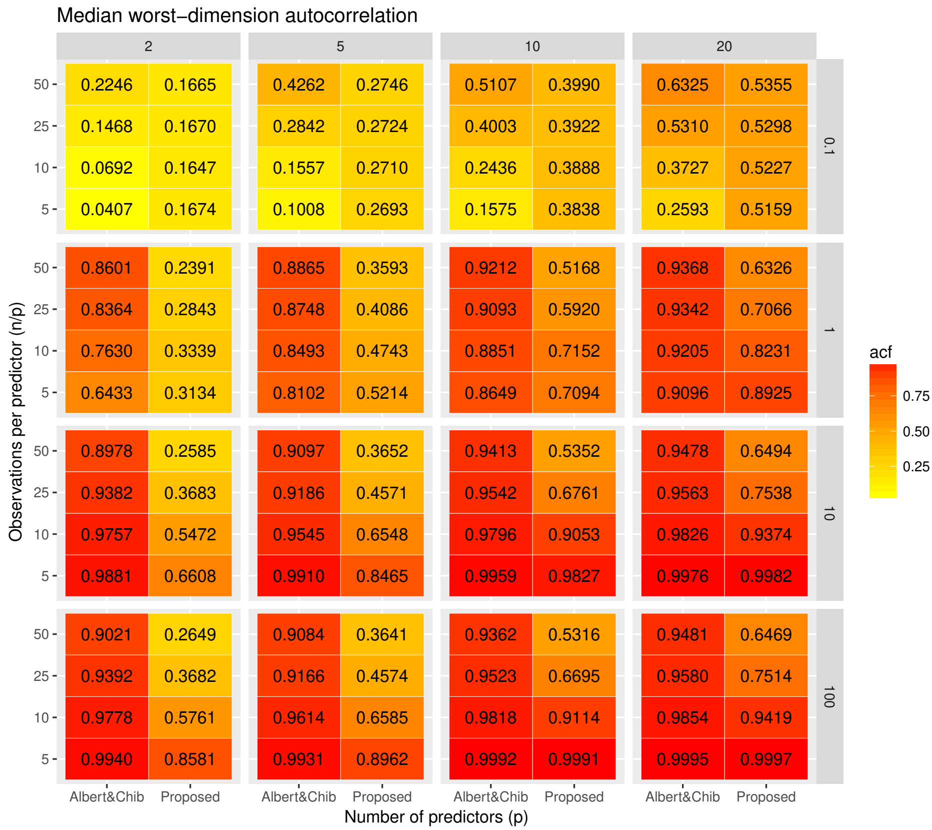

Figure 2. Median worst-dimension lag-1 autocorrelations by sampler.

computation time (in seconds) were recorded for each run. A computation-time-adjusted effective sample size ratio was constructed by dividing the ESS per second of computation time for the independent-proposal MH by the ESS per second of computation time for the Albert and Chib Gibbs sampler. For each (p, n, σ) triplet, the median of these metrics across the 50 iterations was recorded.

3.2. ***Results.*** Figure 1 depicts the log of the median value, across 50 iterations, of the computation-time-adjusted ESS ratio. Positive values indicate that the proposed sampler is more efficient, whereas negative values indicate that Albert and Chib is more efficient; values exceeding 2.30 indicate an order-of-magnitude improvement (since $e^{2.30} = 10.0$).

The results reveal that the proposed method outperforms Albert and Chib in most conditions. In particular, in 58 of 64 conditions, the proposed method was more efficient, often by an order of magnitude. In four of the six conditions under which Albert and Chib was better, the improvement was by less than 20%; and all of these involved small ($\sigma = 1$) or very small ($\sigma = 0.1$) prior standard deviations. Such highly informative priors are in our experience quite rare, as such strong prior knowledge would essentially negate the need to collect data or make posterior inference.

The remaining two conditions, represented by the red cells in the bottom right corners of the $\sigma = 10$ and $\sigma = 100$ subgrids, respectively, reflect situations where sampling was very poor for both methods. In neither of these scenarios was Albert and Chib particularly successful in an absolute sense. Its median worst-dimension autocorrelations were 0.9976 and 0.9995 for $\sigma = 10$ and $\sigma = 100$, respectively (see Figure 2). In short, neither sampler is doing well here, though Albert and Chib is nominally better.

For these two conditions, even if the posterior were well sampled (using a dependent-proposal Metropolis–Hastings algorithm, for example), it is hard to envision it being of much use. These cells correspond to small-n, large-p scenarios where prior knowledge is poor; neither the likelihood nor the prior contains much information, so parameters cannot be estimated reliably. This is why others have cautioned against using probit regression when $n/p < 10$ [Long 1997].

In short, for a wide variety of settings, the proposed method is more efficient than the Albert and Chib Gibbs sampler at producing independent samples from the posterior. The few exceptions correspond to situations with exceptionally narrow priors or exceptionally broad likelihoods, none of which are particularly likely to be encountered in practice.

4. Applications

While the simulation study demonstrated success across a wide range of conditions, the real test is to apply the sampler to actual data. The four applications we consider in this section use previously studied data sets seen as canonical in the probit regression literature. For each application, we compare sampler performance on five outcomes: posterior means of the coefficients, autocorrelation, effective sample size, average jump size (in Euclidean distance), and computation time (in seconds). We also report the computation-time-adjusted ESS ratio, calculated as in the simulation study.

4.1. ***Prostate cancer.*** The data for our first application comes from a prostate cancer study on 53 patients and was reported in [Miller et al. 1980]. Chib [1995] used probit regression to predict whether cancer has spread to surrounding lymph nodes using five predictor variables: the patient's age, the level of serum acid phosphate, positive or negative X-ray results, the size of the tumor, and the pathological grade of the tumor. Results from Albert and Chib's sampler and the proposed MH sampler are presented in Table 1. The posterior means of the coefficients are almost identical, while the proposed sampler has lower computation time and more than double the effective sample size of the Albert and Chib sampler.

	prostate cancer		Finney data		grade prediction		German health care	
	proposed	Albert and Chib	proposed	Albert and Chib	proposed	Albert and Chib	proposed	Albert and Chib
β_0	1.52	1.53	−6.28	−6.28	−8.41	−8.39	−0.13	−0.13
β_1	−0.04	−0.04	3.10	3.10	1.82	1.82	0.01	0.01
β_2	1.67	1.67	2.58	2.58	0.06	0.06	−0.01	−0.01
β_3	1.29	1.28	—	—	1.58	1.58	−0.01	−0.01
β_4	1.02	1.02	—	—	—	—	−0.15	−0.15
β_5	0.50	0.50	—	—	—	—	0.07	0.07
β_6	—	—	—	—	—	—	0.36	0.36
ACF	0.34	0.60	0.20	0.55	0.36	0.75	0.30	0.41
ESS	45,828	20,113	64,923	28,475	43,798	11,801	2,600	2,013
avg. jump	1.89	1.86	2.18	1.83	2.27	1.75	0.06	0.07
comp. time	2.32	3.86	1.19	2.86	1.17	2.74	49.46	50.13
adj. ESS ratio	3.79		5.48		8.69		1.31	

Table 1. Application results. For each sampler, we report posterior means of $\boldsymbol{\beta}$, autocorrelation (ACF), effective sample size (ESS), average jump size, and computation time (in seconds). We compare the two samplers with a computation-time-adjusted ESS ratio; values larger than 1 favor the proposed sampler.

4.2. ***Finney data.*** In [Albert and Chib 1993], they use data from [Finney 1947] to illustrate the accuracy of their method as the number of simulation iterations increases. As shown in Table 1, the β estimates of the two samplers are identical to the hundredths place. The proposed sampler has nearly double the effective sample size, nearly a third of the autocorrelation, and less than half the computation time.

4.3. ***Grade prediction.*** The data set for our third application is from an econometric analysis textbook [Greene 2012], an earlier version of which was cited in [Albert and Chib 1993]. The data comes from a study by [Spector and Mazzeo 1980], whose motivation was to measure the effect of a personalized system of instruction on students' grades. As part of their analysis, they modeled grade improvement as predicted by three variables: the student's grade point average, his/her score on an economic literacy test, and whether or not he/she participated in a personalized system of instruction. In Table 1, note the close similarity in the posterior means of the coefficients, despite the small sample size of 32. The efficiency comparison in this application tells a similar story to the previous two, with the proposed sampler outperforming the Albert and Chib sampler in every category.

4.4. ***German health care.*** The data for our final application is from the same econometric analysis textbook [Greene 2012] as the grade-prediction data. The predictors of age, education, monthly income, marriage status, whether or not children under 16 live in the home, and gender are used to model the binary response of whether or not the subject has visited the doctor in the last three months. The data set contains 27,326 observations on 7,293 German families. Greene uses this model as an example of how slow and inefficient the Albert and Chib method can be. In Table 1 we see that the proposed sampler is only slightly faster; however, the ACF and ESS show that the draws from the proposed sampler are less correlated. Although it took both samplers approximately 50 seconds to run 5,000 iterations, in terms of effective samples, the proposed sampler collected roughly 30% more and thus obtained a similar amount of independent information from the posterior in much less time than Albert and Chib's sampler.

5. Conclusion

Albert and Chib's auxiliary-variable Gibbs sampler [1993] revolutionized Bayesian probit regression and continues to be utilized with wide success in many applications today. As suggested by [Polson et al. 2013], Albert and Chib's sampler is likely the better choice to accommodate more sophisticated modeling frameworks, such as those that involve, for example, dependence among observations. However, many probit regression models used in practice do not require such complexities, and we propose that for this basic modeling context, the independence MH sampler described in this paper should be the new go-to sampler.

We have demonstrated through a simulation study that, in terms of computational efficiency, our proposed sampler outperforms that of Albert and Chib in all but two scenarios, both of which correspond to settings unlikely to be encountered in practice: extremely strong prior knowledge or very diffuse likelihoods. In all other cases, our sampler provides efficiency improvement over the Albert and Chib sampler, and in many cases the scale of improvement is over an order of magnitude.

Beyond simulated settings, we have also demonstrated a striking efficiency advantage provided by our proposed sampler when applied to four real data sets frequently analyzed in probit regression literature. Upon comparing the proposed sampler to Albert and Chib's, we obtain nearly identical coefficient estimates, and our sampler consistently produces higher effective sample sizes in less computation time.

References

[Albert and Chib 1993] J. H. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data", *J. Amer. Statist. Assoc.* **88**:422 (1993), 669–679. MR Zbl

[Amemiya 1981] T. Amemiya, "Qualitative response models: a survey", *J. Econ. Lit.* **19**:4 (1981), 1483–1536.

[Chib 1995] S. Chib, "Marginal likelihood from the Gibbs output", *J. Amer. Statist. Assoc.* **90**:432 (1995), 1313–1321. MR Zbl

[Finney 1947] D. J. Finney, "The estimation from individual records of the relationship between dose and quantal response", *Biometrika* **34**:3-4 (1947), 320–334. Zbl

[Frühwirth-Schnatter and Frühwirth 2007] S. Frühwirth-Schnatter and R. Frühwirth, "Auxiliary mixture sampling with applications to logistic models", *Comput. Statist. Data Anal.* **51**:7 (2007), 3509–3528. MR Zbl

[Gamerman 1997] D. Gamerman, "Sampling from the posterior distribution in generalized linear mixed models", *Stat. Comput.* **7**:1 (1997), 57–68.

[Gelfand and Smith 1990] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities", *J. Amer. Statist. Assoc.* **85**:410 (1990), 398–409. MR

[Geman and Geman 1984] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images", *IEEE Trans. Pattern Anal. Mach. Intell.* **6**:6 (1984), 721–741. Zbl

[Geweke and Tanizaki 2001] J. Geweke and H. Tanizaki, "Bayesian estimation of state-space models using the Metropolis–Hastings algorithm within Gibbs sampling", *Comput. Statist. Data Anal.* **37**:2 (2001), 151–170. MR Zbl

[Greene 2012] W. H. Greene, *Econometric analysis*, 7th ed., Pearson, Upper Saddle River, NJ, 2012.

[Griffiths et al. 1987] W. E. Griffiths, R. C. Hill, and P. J. Pope, "Small sample properties of probit model estimators", *J. Amer. Statist. Assoc.* **82**:399 (1987), 929–937. MR

[Holmes and Held 2006] C. C. Holmes and L. Held, "Bayesian auxiliary variable models for binary and multinomial regression", *Bayesian Anal.* **1**:1 (2006), 145–168. MR Zbl

[Kapourani 2018] A. C. Kapourani, "Gibbs sampling for Bayesian binary probit", online tutorial, 2018, available at https://rpubs.com/cakapourani/bayesian-binary-probit-model.

[Long 1997] J. S. Long, *Regression models for categorial and limited dependent variables*, Advanced quantitative techniques in the social sciences **7**, SAGE, Thousand Oaks, CA, 1997.

[Martin et al. 2011] A. D. Martin, K. M. Quinna, and J. H. Park, "MCMCpack: Markov chain Monte Carlo in R", *J. Stat. Softw.* **42** (2011), art. id. 9.

[Miller et al. 1980] R. G. Miller, B. Efron, B. W. Brown, and L. E. Moses, "Prediction analysis for binary data", pp. 3–18 in *Biostatistics casebook*, Wiley, New York, 1980.

[Polson et al. 2013] N. G. Polson, J. G. Scott, and J. Windle, "Bayesian inference for logistic models using Pólya-Gamma latent variables", *J. Amer. Statist. Assoc.* **108**:504 (2013), 1339–1349. MR Zbl

[Spector and Mazzeo 1980] L. C. Spector and M. Mazzeo, "Probit analysis and economic education", *J. Econ. Ed.* **11**:2 (1980), 37–44.

[Zeger and Karim 1991] S. L. Zeger and M. R. Karim, "Generalized linear models with random effects; a Gibbs sampling approach", *J. Amer. Statist. Assoc.* **86**:413 (1991), 79–86. MR

[Zellner and Rossi 1984] A. Zellner and P. E. Rossi, "Bayesian analysis of dichotomous quantal response models", *J. Econometrics* **25**:3 (1984), 365–393. MR Zbl

Received: 2019-01-28 Accepted: 2019-11-14

scott.thomas.simmons@gmail.com *US Naval Academy, Annapolis, MD, United States*

emcguffe@usna.edu *US Naval Academy, Annapolis, MD, United States*

vanderwe@usna.edu *US Naval Academy, Annapolis, MD, United States*

involve

msp.org/involve

INVOLVE YOUR STUDENTS IN RESEARCH

Involve showcases and encourages high-quality mathematical research involving students from all academic levels. The editorial board consists of mathematical scientists committed to nurturing student participation in research. Bridging the gap between the extremes of purely undergraduate research journals and mainstream research journals, *Involve* provides a venue to mathematicians wishing to encourage the creative involvement of students.

MANAGING EDITOR

Kenneth S. Berenhaut — Wake Forest University, USA

BOARD OF EDITORS

Colin Adams	Williams College, USA
Arthur T. Benjamin	Harvey Mudd College, USA
Martin Bohner	Missouri U of Science and Technology, USA
Amarjit S. Budhiraja	U of N Carolina, Chapel Hill, USA
Pietro Cerone	La Trobe University, Australia
Scott Chapman	Sam Houston State University, USA
Joshua N. Cooper	University of South Carolina, USA
Jem N. Corcoran	University of Colorado, USA
Toka Diagana	University of Alabama in Huntsville, USA
Michael Dorff	Brigham Young University, USA
Sever S. Dragomir	Victoria University, Australia
Joel Foisy	SUNY Potsdam, USA
Errin W. Fulp	Wake Forest University, USA
Joseph Gallian	University of Minnesota Duluth, USA
Stephan R. Garcia	Pomona College, USA
Anant Godbole	East Tennessee State University, USA
Ron Gould	Emory University, USA
Sat Gupta	U of North Carolina, Greensboro, USA
Jim Haglund	University of Pennsylvania, USA
Johnny Henderson	Baylor University, USA
Glenn H. Hurlbert	Virginia Commonwealth University, USA
Charles R. Johnson	College of William and Mary, USA
K. B. Kulasekera	Clemson University, USA
Gerry Ladas	University of Rhode Island, USA
David Larson	Texas A&M University, USA
Suzanne Lenhart	University of Tennessee, USA
Chi-Kwong Li	College of William and Mary, USA
Robert B. Lund	Clemson University, USA
Gaven J. Martin	Massey University, New Zealand
Mary Meyer	Colorado State University, USA
Frank Morgan	Williams College, USA
Mohammad Sal Moslehian	Ferdowsi University of Mashhad, Iran
Zuhair Nashed	University of Central Florida, USA
Ken Ono	Univ. of Virginia, Charlottesville
Yuval Peres	Microsoft Research, USA
Y.-F. S. Pétermann	Université de Genève, Switzerland
Jonathon Peterson	Purdue University, USA
Robert J. Plemmons	Wake Forest University, USA
Carl B. Pomerance	Dartmouth College, USA
Vadim Ponomarenko	San Diego State University, USA
Bjorn Poonen	UC Berkeley, USA
Józeph H. Przytycki	George Washington University, USA
Richard Rebarber	University of Nebraska, USA
Robert W. Robinson	University of Georgia, USA
Javier Rojo	Oregon State University, USA
Filip Saidak	U of North Carolina, Greensboro, USA
Hari Mohan Srivastava	University of Victoria, Canada
Andrew J. Sterge	Honorary Editor
Ann Trenk	Wellesley College, USA
Ravi Vakil	Stanford University, USA
Antonia Vecchio	Consiglio Nazionale delle Ricerche, Italy
John C. Wierman	Johns Hopkins University, USA
Michael E. Zieve	University of Michigan, USA

PRODUCTION

Silvio Levy, Scientific Editor

Cover: Alex Scorpan

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2020 is US $205/year for the electronic version, and $275/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues and changes of subscriber address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

© 2020 Mathematical Sciences Publishers

involve

2020 vol. 13 no. 1