Bayesian Analysis (2023) **18**, Number 2, pp. 415–435

Normal Approximation for Bayesian Mixed Effects Binomial Regression Models

Brandon Berman*, Wesley O. Johnson†, and Weining Shen‡

Abstract. Bayesian inference for generalized linear mixed models implemented with Markov chain Monte Carlo (MCMC) sampling methods have been widely used. In this paper, we propose to substitute a large sample normal approximation for the intractable full conditional distribution of the latent effects (of size k) in order to simplify the computation. In addition, we develop a second approximation involving what we term a sufficient reduction (SR). We show that the full conditional distributions for the model parameters only depend on a small, say $r \ll k$, dimensional function of the latent effects, and also that this reduction is asymptotically normal under mild conditions. Thus we substitute the sampling of an r dimensional multivariate normal for sampling the k dimensional full conditional for the latent effects. Applications to oncology physician data, to cow abortion data and simulation studies confirm the reasonable performance of the proposed approximation method in terms of estimation accuracy and computational speed.

Keywords: asymptotic approximation, binomial regression, generalized linear mixed models, Markov chain Monte Carlo, sufficient reduction.

1 Introduction

Generalized linear mixed models (GLMMs) have been widely used for modeling longitudinal and clustered data in many scientific applications (McCulloch, 1996; Diggle et al., 2002). Statistical inference for GLMMs has received a lot of interest over the past three decades. In frequentist literature, it is common to consider a likelihood-based approach (McCulloch, 1997; Breslow and Clayton, 1993), which often requires a painful high-dimensional numerical integration and the derivation of large-sample asymptotic results for inference. Consequently, Bayesian methods have become more attractive as they provide a useful alternative to numerical integration by adopting posterior sampling techniques such as Markov chain Monte Carolo (MCMC) methods. The MCMC methods for GLMMs are usually carried out using Gibbs sampling (Zeger and Karim, 1991) and Metropolis-Hastings methods (Gamerman, 1997); and those methods can now be conveniently implemented using standard software platforms such as BUGS (Spiegelhalter et al., 2003) and JAGS (Plummer, 2012).

Despite the popularity of MCMC methods for GLMMs, in some cases, posterior sampling can still be challenging due to computational issues such as the convergence

*Department of Statistics, University of California, Irvine
†Department of Statistics, University of California, Irvine
‡Department of Statistics, University of California, Irvine. Corresponding author, swn1989@gmail.com

 https://doi.org/10.1214/22-BA1312

of Markov chains and the lack of efficient sampling methods; particularly with a large number of latent effects (Hadfield, 2010). The main focus of this paper is to provide a simple-yet-useful alternative by considering a large sample approximation method that could have the potential to improve upon the computational efficiency of MCMC sampling for GLMMs. Large sample approximation is a classical approach that has been well explored in many areas of statistics, including Bayesian statistics and GLMMs. For example, asymptotic normality results were established for posterior distributions under different conditions (Walker, 1969; Chen, 1985). Laplace approximation (Tierney and Kadane, 1986) and the integrated nested Laplace approximation (INLA) (Rue et al., 2009) provided a computationally convenient first-order approximation to a posterior density function by a normal density centered at the posterior mode. For GLMMs, Breslow and Clayton (1993) proposed the penalized quasi-likelihood method by approximating the marginal quasi-likelihood function.

More recently, Yee et al. (2002) obtained an asymptotic joint normal approximation for the posterior distributions of two blocks of variables, i.e., one block of GLMM model parameters and another block of random effects. They first obtained the conditional asymptotic distributions for normalized block variables, and then, due to compatibility of the limiting distributions, they obtained the appropriate joint asymptotic distribution. This result was later generalized to a model with multiple blocks of variables by Su and Johnson (2006). Posterior normality was also obtained for stochastic processes by Weng and Tsai (2008). Fong et al. (2010) gave a comprehensive review of the implementation of INLA for GLMMs. Baghishani and Mohammadzadeh (2012) also obtained asymptotic normality results for the joint posterior distribution of the model parameters and random effects in GLMMs by using Stein's Identity.

Unlike the aforementioned methods that focus on establishing joint asymptotic normality results for general model forms, our interest here focuses on a large sample approximation for the block that is more difficult to sample, namely the one for latent effects. Using our normal approximation for the latents, the normal prior for regression coefficients is conditionally conjugate. The gamma prior for the precision is also conditionally conjugate.

We initially develop the concept for a particular GLMM, the mixed effects binomial regression model

$$Y_i \overset{\perp}{\sim} \text{Bin}(n_i, p_i), \quad \text{logit}(p_i) = u_i, \quad u_i \mid x_i, \beta \overset{\perp}{\sim} N(x_i\beta, 1/\tau),\ i = 1, \ldots, k, \tag{1}$$

where x_i is a $1 \times p$ vector with a one in the first slot and with $p-1$ predictor values for unit i in the remaining slots; β is a vector of regression coefficients. The main reason for considering this model is because the type of data we envision involves obtaining Bernoulli success/failure observations for units that are clustered. For example suppose we were to count the number of patients in each of k hospitals who acquired nosocomial infections while under treatment, during a specified amount of time. Since the number of treated patients at hospital i, n_i, would be known, the corresponding number of infections, y_i, might initially be regarded as a binomial count. But when considering multiple counts, we might realize that the individual Bernoullis within hospitals would

be correlated since the same environment, including hospital staff, physical space, management, etc is shared by all who are being treated. So extra binomial variation is to be expected and this model will account for it. A main goal for the ensuing analysis of such data would be to estimate regression effects, β, corresponding odds ratios, e^{β}, and probabilities of infection, expit $(x_i\beta) = \exp(x_i\beta)/\{1 + \exp(x_i\beta)\}$. We consider a more complex model later in the paper after exploring this one.

Denote the observed data as $y = \{y_1, \ldots, y_k\}$. The MCMC approach for making numerical approximations to the joint conditional distribution, $p(\beta, \tau, u \mid y)$ involves sampling from the full conditional distributions $p(\beta \mid \tau, u, y)$, $p(\tau \mid \beta, u, y)$, and $p(u \mid \beta, \tau, y)$. With conditionally conjugate priors for β and τ, the first two distributions are easy to sample. However sampling from the third generally involves adaptive rejection sampling (Gilks and Wild, 1992). We consider a large-sample normal approximation to replace the nonstandard conditional density kernels for the u_i's. This allows us to bypass the need of implementing the usual adaptive rejection step within the Gibbs sampler and hence renders more convenient computation. In particular, we explore two types of the normal approximation methods. One is to approximate the conditionals for each u_i separately with a normal distribution; and the other is to construct a *low-dimensional* statistic, T, in replacement of the potentially *high-dimensional* collection of all k u_i's and then approximate the conditional distribution of T with a normal distribution. In our current illustration we are able to sample a two-dimensional $T(u)$, instead of k dimensional u, in order to make full inferences about all regression parameters, probabilities of success and odds ratios,

The normal approximation and dimension reduction ideas that we develop in this paper are in general applicable to other mixed effect models although we focus on binomial regression for demonstration. Our goal is to introduce these ideas and methods, and to show their fittness, rather than to claim any superiority over existing computational methods. We will see clearly that JAGS is much faster than our methods due to its C++ based platform being superior to our R code. Moreover, their optimization of the (Gilks and Wild, 1992) adaptive rejection scheme is surely superior to ours. We also acknowledge that the Pólya-gamma data augmentation (Polson et al., 2013) scheme for binomial regression with logit link, which manages to perform Gibbs sampling exactly with an efficient augmentation scheme, is surely an attractive alternative to other possibilities including ours. Instead, the main goal of this paper is to present a neat idea and show its promise as a valuable alternative for analyzing Bayesian mixed effect models, especially when the number of latent effects is large. Normal approximation is a classical idea that has been widely studied in all areas of statistics. It is our hope that the work presented in this paper may pave the way for future investigation of more complicated GLMMs (e.g., with more levels of random effects and general link functions) and development of efficient computational toolkits.

We give a more detailed description of the proposed methodology in Section 2. We evaluate the empirical performance of the proposed method through an oncology physician data analysis and additional simulation studies in Sections 3 and 4. We further extend our method to a two-level logistic mixed effect model in Section 5, and we provide an elaborate sketch of theoretical details that supports its use in this more complex model. Several future working directions are discussed in Section 6.

2 Method

2.1 Normal approximation

Following Gelfand et al. (1995), we consider a mixed effect binomial regression model that involves centering the random effects, u_i, on $x_i\beta$s. This parametrization also facilitates our normal approximation here. The augmented data likelihood is

$$\begin{aligned}
&\prod_{i=1}^{k} \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i} \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau(u_i - X_i\beta)^2}{2}\right) \\
&\propto \tau^{k/2} \exp\left\{\sum_{i=1}^{k} \left[y_i \operatorname{logit}(p_i) + n_i \log(1-p_i)\right] - \frac{\tau(u - X\beta)'(u - X\beta)}{2}\right\}, \qquad (2)
\end{aligned}$$

where $u = (u_1, u_2, \ldots, u_k)'$ and $X = (x_1', x_2', \ldots, x_k')'$.

We assign independent prior distributions on β and τ; $\beta \sim N_p(B_0, C)$ and $\tau \sim \mathrm{Ga}(a/2, b/2)$ where B_0, C, a, b are hyper-parameters determined using *a priori* knowledge. Consequently, the joint conditional for (β, u, τ) given the data is proportional to

$$\begin{aligned}
\tau^{(a+k)/2-1} \exp\left\{\left[\sum_{i=1}^{k} \left[y_i u_i - n_i \log(1+\exp(u_i))\right]\right] - \frac{\tau}{2}\left[(u - X\beta)'(u - X\beta)\right.\right. \\
\left.\left. +(\beta - B_0)'C^{-1}(\beta - B_0) + b\tau\right]\right\}. \qquad (3)
\end{aligned}$$

To approximate inferences using MCMC methods, we obtain the full conditional distributions:

$$\begin{aligned}
\beta \,|\, u, \tau, y, X &\sim N_p\left((\tau X'X + C^{-1})^{-1}(X'u\tau + C^{-1}B_0), (\tau X'X + C^{-1})^{-1}\right), \\
\tau \,|\, u, \beta, y, X &\sim \mathrm{Ga}\left(\frac{a+k}{2}, \frac{b + u'(I_k - X(X'X)^{-1}X')u + (\beta - \hat{\beta})'X'X(\beta - \hat{\beta})}{2}\right),
\end{aligned}$$

and

$$p(u \,|\, \beta, \tau, y, X) \propto \exp\left\{\sum_{i=1}^{k} n_i \left[\tilde{\pi}_i u_i - \log(1+\exp(u_i))\right]\right\} \exp\left\{-\frac{\tau}{2}(u - X\beta)'(u - X\beta)\right\}, \qquad (4)$$

where $\tilde{\pi}_i = y_i/n_i$ for all i and $\hat{\beta} = (X'X)^{-1}X'u$. Here we assume $\tilde{\pi}_i \in (0,1)$ since a Poisson approximation will work better than a normal approximation if $\tilde{\pi} = 0$ or 1. Since the u_i's are mutually independent, we only need to sample the full conditional distributions of the individual u_i's, which is usually accomplished with adaptive-rejection sampling or slice sampling.

We propose to approximate the conditional distribution for $u \mid \beta, \tau, y, X$ in (4) using a multivariate normal distribution with diagonal covariance structure due to the aforementioned conditional independence. This is accomplished by expanding the terms in the first exponent of (4) in a particular second order Taylor expansion and recognizing that the first term in the expansion does not involve u, the second term is zero, and the third term is a quadratic form in u. Moreover, since $(u - X\beta)'(u - X\beta)$ is itself a quadratic form in u, it is possible to complete the square to obtain a normal kernel for the conditional pdf for u. The precise result is given in Proposition 1 below and is proven in the Supplementary File (Berman et al., 2022), Section S1.

Let $\tilde{u}_i = \text{logit}(\tilde{\pi}_i)$ and $\tilde{u} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_k)'$, $\tilde{\pi} = (\tilde{\pi}_1, \ldots, \tilde{\pi}_k)'$, and $D_{n\tilde{\pi}(1-\tilde{\pi})}$ be a $k \times k$ diagonal matrix, $D_{n\tilde{\pi}(1-\tilde{\pi})} = \text{diag}\{n_i\tilde{\pi}_i(1-\tilde{\pi}_i) \colon i = 1, 2, \ldots, k\}$. The following proposition establishes a normal approximation for the conditional distribution of u.

Proposition 1. *The ratio of the kernel for the full conditional density function for $u \mid \beta, \tau, y, X$ defined in (4) and the kernel of the pdf for the following normal distribution*

$$N_k\left((D_{n\tilde{\pi}(1-\tilde{\pi})} + \tau I_k)^{-1}(D_{n\tilde{\pi}(1-\tilde{\pi})}\tilde{u} + \tau X\beta), (D_{n\tilde{\pi}(1-\tilde{\pi})} + \tau I_k)^{-1}\right). \tag{5}$$

converges in probability to one as $min_{i=1,\ldots,k}(n_i) \to \infty$, with k fixed.

2.2 Normal approximation with sufficient reduction

In most statistical applications, a large sample size is an advantageous feature as more information is usually helpful for understanding scientific phenomena. However, in our case, methods with a large k can be difficult since our normal approximation method relies on Gibbs sampling; and since we sample the full conditional for $u \mid \beta, \tau, y, X$ as a k-dimensional multivariate normal distribution, the computational costs increase as k increases. In this section we propose a solution that we call, "sufficient reduction", where the name comes from the concept of sufficient statistics, even though our method does not involve a sufficient statistic in the classical sense. It is a form of conditional sufficiency. Our proposal is motivated by examining the conditional distributions for $\beta \mid \tau, u, y, X$ and $\tau \mid \beta, u, y, X$.

Recall $\hat{\beta} = (X'X)^{-1}X'u$. If we let $T_1(u) = \hat{\beta}$, then the conditional distribution of β depends on u only through $T_1(u)$. The rate parameter of conditional distribution of τ depends only on u through $(\beta - T_1(u))'X'X(\beta - T_1(u)) + u'(I_k - X(X'X)^{-1}X')u$, which motivates us to select $T_2(u) = u'(I_k - X(X'X)^{-1}X')u$. To summarize, we consider the following "conditionally sufficient reduction" for u,

$$T(u) \equiv (T_1(u), T_2(u)),\ T_1(u) = (X'X)^{-1}X'u,\ T_2(u) = u'(I_k - X(X'X)^{-1}X)u. \tag{6}$$

Thus if we sample from the full conditional for T, we no longer need to sample the full conditional for u, or its approximation, in order to make inferences about (β, τ), or functions of it.

The conditional distributions for β and τ are now expressed as

$$\beta \mid T_1(u), \tau \sim N_p\left((\tau X'X + C^{-1})^{-1}(X'XT_1(u)\tau + C^{-1}B_0), (\tau X'X + C^{-1})^{-1}\right) \tag{7}$$

$$\tau \mid \beta, T_1(u), T_2(u) \sim \text{Ga}\left(\frac{a+k}{2}, \frac{b + T_2(u) + (\beta - T_1(u))' X' X (\beta - T_1(u))}{2}\right). \tag{8}$$

Since the conditional distribution of u is approximately multivariate normal, the conditional distribution for $T_1(u)$ is also approximately multivariate normal (exactly normal of course if u is exactly normal). The conditional distribution for $T_2(u)$ involves a quadratic function of u, and it can also be approximated by a normal distribution, for large k. More specifically, we consider a multivariate normal approximation for the joint conditional distribution of $(T_1(u), T_2(u))$. Before asserting formal asymptotic normality, it is straightforward to obtain the conditional moments for $T(u)$:

$$\begin{aligned}
&E\left[T_1(u) \mid \text{else}\right] = Au_0, \quad E\left[T_2(u) \mid \text{else}\right] = \text{tr}\left[BD_0\right] + u_0' B u_0, \\
&\text{var}[T_1(u) \mid \text{else}] = AD_0A', \quad \text{var}[T_2(u) \mid \text{else}] = 2\,\text{tr}\left[BD_0BD_0\right] + 4u_0'BD_0Bu_0, \\
&\text{cov}[T_1(u), T_2(u) \mid \text{else}] = 2AD_0Bu_0,
\end{aligned} \tag{9}$$

where $A = (X'X)^{-1}X' \in \mathbb{R}^{p\times k}$, $P_X = X(X'X)^{-1}X'$, $B = I_k - P_X$, $D_0 = (D_{n\tilde{\pi}(1-\tilde{\pi})} + \tau I_k)^{-1}$, $u_0 = D_0\gamma$, and $\gamma = D_{n\tilde{\pi}(1-\tilde{\pi})}\tilde{u} + X\beta\tau$. Based on (9), the posterior samples can be obtained by the usual Gibbs sampling of β, τ and (T_1, T_2).

For the formal result, we treat the result in Proposition 1 as if it were exact. Thus we have $u \sim N_k(u_0, D_0)$ and therefore $T_1(u) = Au \sim N_p(Au_0, AD_0A')$. We then standardize T_1 and T_2 as $\tilde{T}_1 = (AD_0A')^{-1/2}(T_1 - Au_0), \tilde{T}_2 = (T_2 - E(T_2))/sd(T_2)$ and proceed to show that the joint distribution for $(\tilde{T}_1, \tilde{T}_2)$ converges to a $(p+1)$ dimensional standard multivariate normal. From a practical standpoint, when we do computations, we can just use the normal distribution with the moments given in Equations (9) when we sample the full conditional for T.

Technical assumptions (A1)-(A4) are given in the Supplementary File Section S1. Assumption (A1) is easily satisfied since $(I_k - sBD_0)$ should be very close to I_k as s is sufficiently small. Assumptions (A2) and (A3) require the existence of the asymptotic variance for $\tilde{T}_2$ and its covariance with $\tilde{T}_1$. Assumption (A4) is needed for handling the remainder terms in the Taylor expansion of the mgf in the proof. Now we state Proposition 2, which establishes the joint asymptotic distribution for $(\tilde{T}_1, \tilde{T}_2)$. The proof is in Supplementary File Section S1. There, we also give a simple illustration of the assumptions for a concrete example.

Proposition 2. *Suppose that Assumptions (A1)-(A4) hold,* $u \sim N_k(u_0, D_0)$ *and let* c *be a* $p \times 1$ *column vector defined in (A3), then*

$$\begin{pmatrix} \tilde{T}_1(u) \\ \tilde{T}_2(u) \end{pmatrix} \xrightarrow{L} N_{p+1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_p & c \\ c' & 1 \end{pmatrix}\right) \qquad (k \to \infty). \tag{10}$$

3 HDP data analysis

3.1 Data description

We demonstrate the use of our proposed approximation by analyzing a hospital, doctor, and patient (HDP) dataset simulated by UCLA's Institute of Digital Research and Education (IDRE), available at https://stats.idre.ucla.edu/r/codefragments/mesimulation/.

Despite the fact that the data were simulated, they provide a very nice conceptual situation with a moderately large k in order for us to illustrate our methods.

There are 308 oncology physicians in the dataset who have treated 6745 individuals for lung cancer, i.e, on average 21.9 patients treated by each physician. We are mainly interested in modeling the probability of a physician's ability to facilitate lung cancer remission using physician level covariates, including physician's years of experience (mean = 17.96, sd = 4.08), the number of malpractice lawsuits involving the physician (mean = 1.97, sd = 1.53), and whether or not the physician attended a top ranked medical school (22.4% attended top school). These covariates are believed to be helpful in quantifying a physician's ability to successfully facilitate cancerous tumor remission. For example, the more experience a physician has, ideally, the better the treatment. The number of malpractice lawsuits may have a negative effect on the probability of a physician's ability to successfully facilitate cancerous tumor remission, as it may be viewed as a sign of carelessness.

Let y_i be the number of patients that physician i has treated successfully (the cancer went into remission) and n_i be the total number of patients they treated. Also let $x_i = (1, \text{sExp}_i, \text{sLS}_i, \text{TMS}_i)$, where sExp and sLS are standardized experience and law suit scores (mean zero and variance 1) respectively, and where TMS is one if the physician went to a top medical school and zero otherwise. We consider a physician with about 18 years of experience, with about two law suits and who went to a non-top medical school to be a *baseline* physician. This concept will be useful when we specify priors.

Moreover, patients of the same physician, say i, would naturally share a common treatment environment and experience. Thus it is reasonable to assume that they would share the same random u_i whose distribution is centered on the $x_i\beta$ for that physician. Thus we regard their n_i patients as a cluster, and we propose the model in Equation (1) for these data. Accordingly, we don't expect all patients of the same physician to have exactly the same probability of successful treatment. In fact, the median success rate for physicians with covariates x_i is $\tilde{p}_i = \text{expit}(x_i\beta)$, and the corresponding 90th percentile of success probabilities is $\tilde{p}_{i,90} = \text{expit}(x_i\beta + 1.28\sigma)$, where $\sigma = 1/\sqrt{\tau}$.

We specify $\beta \sim N_4(0, I_4)$, and $\tau \sim Ga(1/2, 1/2)$ as our prior for the model parameters. On the logistic scale, a $N(0, 1)$ prior induces a prior on a probability that is not overly concentrated near zero or one. While priors with large variances have been used in attempts to be "noninformative", the induced prior on the probability scale attaches considerable mass to 0 and to 1, which can have a big effect on the posterior (Christensen et al., 2010, Chapter 8).

Part of the particular specification involves thinking about the success probability for baseline physicians, whose median probability of success is $\text{expit}(\beta_1)$. Our best guess for this probability 0.5 and we specify 95% certainty that it is between 0.1 and 0.9. Using this information, we find the $N(0, b)$ distribution for β_1 that induces this specification; $b = 1.1$ does the job. Additional considerations are made for the full specification, and are given in Supplementary File Section S2.

Parameter	Method	2.5%tile	Median	97.5%tile	Mean	SD
$\exp(\beta_2)$	Normal Appx	1.12	1.31	1.53	1.31	0.10
	JAGS	1.13	1.34	1.58	1.34	0.12
$\exp(\beta_3)$	Normal Appx	0.60	0.86	1.21	0.87	0.15
	JAGS	0.58	0.84	1.22	0.86	0.16
$\exp(\beta_4)$	Normal Appx	0.78	0.92	1.08	0.92	0.08
	JAGS	0.76	0.91	1.08	0.91	0.08
$\text{expit}(\beta_1)$	Normal Appx	0.30	0.34	0.38	0.34	0.02
	JAGS	0.28	0.32	0.36	0.32	0.02

Table 1: Posterior inferences for the HDP example.

3.2 Results

We implemented the proposed normal approximation within the Gibbs sampling and compared the results to the method used in JAGS. Both methods used the same prior as described in the previous section, the same number of iterations (5000) and the same burn-in (500). In Figure 1 we plot posterior pdfs for five quantities: $\text{expit}(\beta_1)$, $\text{expit}(\beta_1 + 1.28\sigma)$, $\exp(\beta_2)$, $\exp(\beta_3)$, and $\exp(\beta_4)$. The quantity $\text{expit}(\beta_1)$ represents the probability of a "baseline" physician (with average covariate values), i.e., approximately 18 years of experience, subject to approximately 2 lawsuits, and did not attend a top ranked medical school, to remit a patient's cancerous tumor. Then $\text{expit}(\beta_1 + 1.28\sigma)$ represents the 90^{th} percentile of cancer remission probabilities for baseline physicians. The other objects of interest, $\exp(\beta_2)$, $\exp(\beta_3)$, and $\exp(\beta_4)$, are odds ratios for the three predictors, experience, lawsuits, and top-ranked medical school, respectively.

We present numerical summaries for these quantities in Table 1. We find that our normal approximation method (we are not implementing the SR method here) works quite well given the minor discrepancy between the results from normal approximation and JAGS. For example, the posterior median for the odds ratio, $\exp(\beta_2)$, is 1.31 with a 95% PI of $(1.12, 1.53)$ under normal approximation; and it is 1.34 with 95% PI of $(1.13, 1.58)$ using JAGS. Thus if there are two physicians of equal education level and number of lawsuits, the type of physician with four more years of experience has odds of remission that are 31% higher than those for the less experienced physician. Similar interpretations can be obtained for other predictors.

In Figure 1, we see greater discrepancies for $\exp(\beta_4)$ and for $\text{expit}(\beta_1 + 1.28\sigma)$ compared with other parameters. Aside from the different graph scales, we believe this is also due to using the normal kernel for approximating the binomial likelihood when the expected number of successes np (or failures $n(1-p)$) is too small (Cochran, 1952). To further explore this phenomenon, we selected two physicians with the most extreme sample proportions of cancer remission, i.e., smallest (Physician # 4) and largest (Physician # 218). We also randomly selected two additional physicians (# 116 and 171) as benchmarks. We present inferences for their probabilities of cancer remission in Figure 2 and Table 2. It is clear that the proposed normal approximation works almost perfectly for the benchmark Physicians # 116 and 171, but is not as well for Physicians 4 and 218, especially in the tails. It is worth noting that Physician 4 only treated one patient out of 34 patients successfully while Physician 218 successfully treated 36 out of 40 patients.

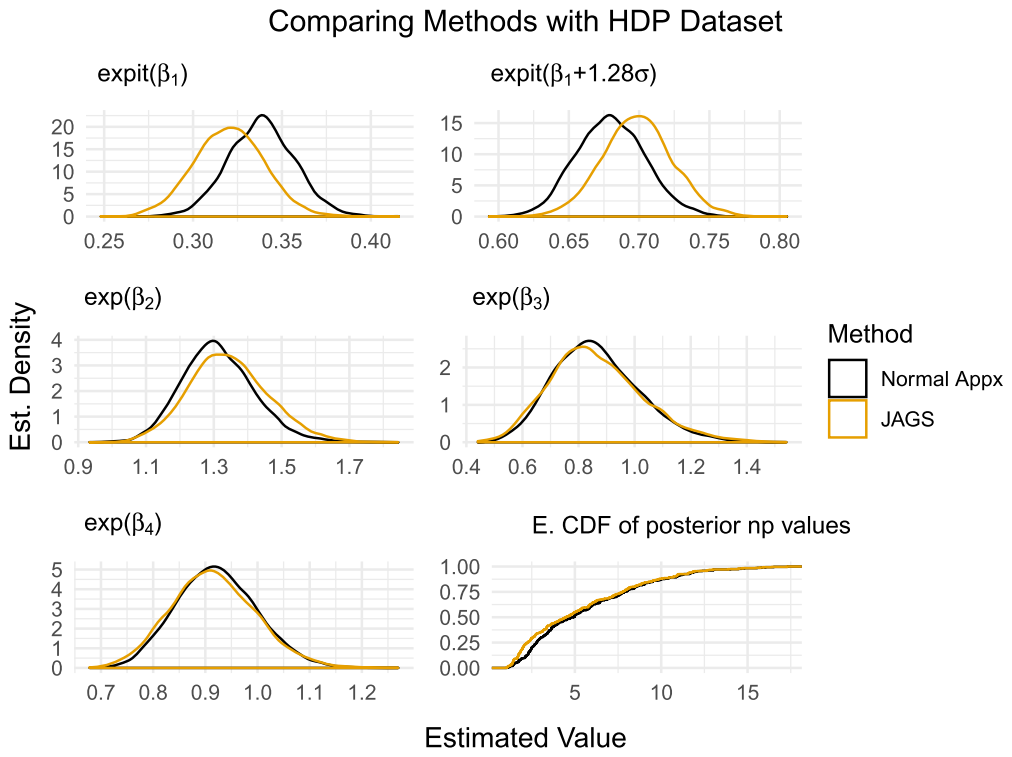

Figure 1: Posterior density plots for HDP data with JAGS and Gibbs sampling with normal approximation.

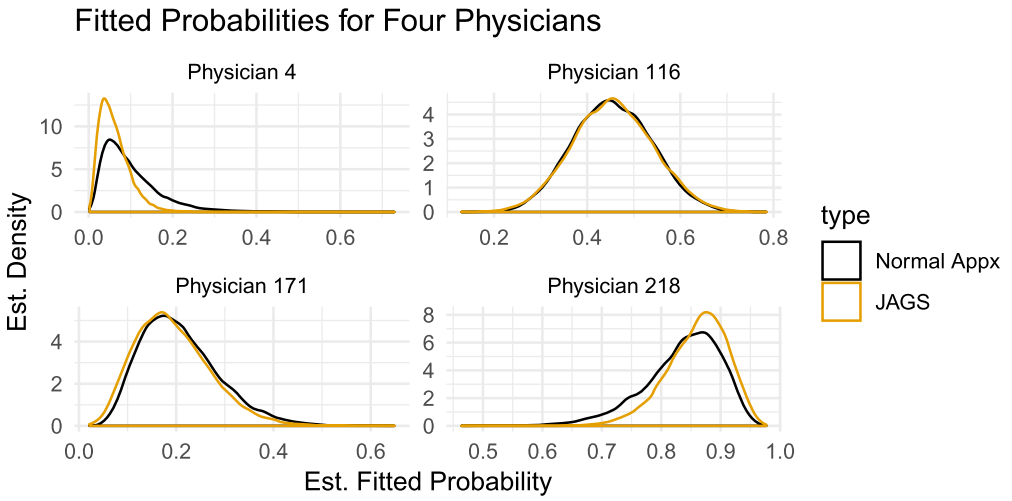

Figure 2: Fitted probability densities for four select physicians in HDP example.

Physician	Method	2.5%tile	Median	97.5%tile	Mean	SD
116	Normal Appx	0.29	0.45	0.62	0.45	0.08
	JAGS	0.29	0.45	0.62	0.45	0.09
171	Normal Appx	0.09	0.20	0.40	0.21	0.08
	JAGS	0.07	0.18	0.36	0.19	0.08
4	Normal Appx	0.02	0.08	0.29	0.10	0.07
	JAGS	0.01	0.05	0.15	0.06	0.04
218	Normal Appx	0.68	0.85	0.93	0.84	0.06
	JAGS	0.74	0.87	0.94	0.86	0.05

Table 2: Summary of fitted probabilities for four selected physicians in HDP example.

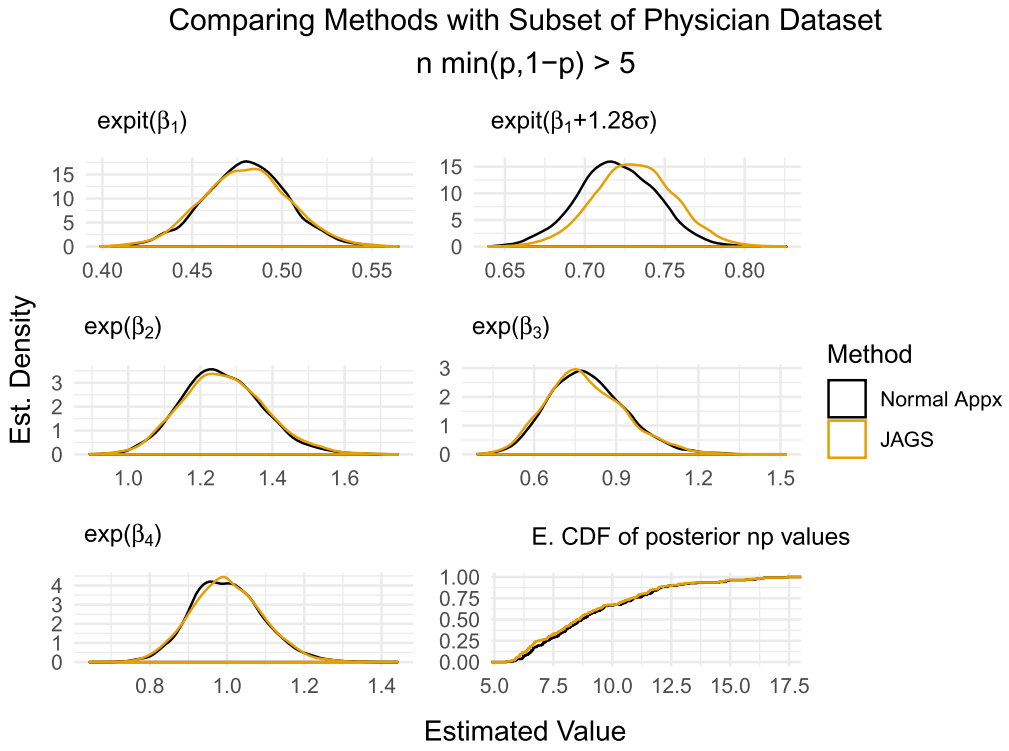

Figure 3: Estimated conditional densities based on data with expected counts > 5.

In Figure 1, we also plot the empirical CDF of the posterior medians for $n_i \min(p_i, 1-p_i) : i = 1, \ldots, 308$ for physicians in the lower right picture. It is clear that about half (150 out of 308) of the physicians have an estimated np value less than 5, which explains why the normal approximation does not work perfectly when p is close to the boundary of $(0, 1)$. We also present results after filtering out the physicians with $n \min(p, 1-p) > 5$ in Figure 3. Results from normal approximation method and JAGS align quite well.

We also did an analysis of these data restricted to the physicians with $n \min(p, 1 -$

$p) \geq 10$, and the analogous figure to Figure 3 shows perfect alignment of estimated pdfs based on JAGS and the normal approximation.

Finally we conclude the analysis by briefly actually analyzing the data. We use the normal approximation to make inferences, but inferences would be virtually/practically identical if we used the JAGS output. The estimated probability of patient remission for a baseline physician is estimated to be 0.34 (0.30, 0.38). Observe from Figure 1 that the corresponding 90th percentile of remission probabilities for baseline doctors would be estimated to be considerably larger than the median (50th percentile). An additional four years experience is estimated to improve the odds of remission by about 30%, with a 95% probability interval for that improvement of 12 to 53 percent. This effect is statistically important and plausibly practically important as well. The estimated effect of having two additional law suits against a physician is a reduction in estimated odds by about 14%. However this is not statistically important since the possibility of both reductions and increases in these odds are well inside of the corresponding 95% interval. Finally, the effect of attending a top medical school does not appear to have any practical or statistical importance. It is clear from Figure 2 that there is a considerable range of estimated probabilities of success across the types of physicians in the data. We remind the reader that the data were simulated.

4 Simulations

In this section we use simulations to further evaluate the empirical performance of the proposed normal approximation method. We generate the data from a realistic setting that is similar to what we observed from the HDP data example. More specifically, we consider three covariates, including the intercept, a continuous covariate that follows a standard normal distribution and a binary covariate following a Bernoulli distribution with success probability of 22% (the same percentage with the medical school covariate from the HDP example). The true coefficient values are set to be the same with the posterior medians obtained from the HDP data analysis using JAGS with the prior $\beta \sim N_3(0, I)$ and $\tau \sim \text{Gamma}(1/2, 1/2)$. We vary the number of observational units, k, i.e., the number of physicians in the HDP data example, in the set $\{100, 300, 500\}$; and generate n_i (e.g,. number of patients being treated by physician i) from a Poisson distribution with parameter λ, where λ takes values in $\{25, 50, 100\}$.

We compare the performance of four methods including (i) the method implemented by JAGS, (ii) the normal approximation method proposed in Section 2.1, (iii) the sufficient reduction method in 2.2, and (iv) an adaptive-rejection sampling (ARS) method that we programmed. All four methods generate MCMC samples using Gibbs sampling and they all update the blocks for β and τ by sampling the exact full conditional distributions. The exact full conditional distribution for u is sampled using adaptive-rejection Gilks and Wild (1992) in methods (i) and (iv), using different code of course. And finally, the full conditional for u is sampled approximately using our normal approximation in method (ii), and the full conditional for T is sampled approximately with a normal approximation in our method (iii). The main reason that we include the ARS method (iv) is to allow a fair comparison of computational time because methods (ii-iv) (except

		expit(β_1)				expit($\beta_1 + 1.28\sigma$)			
		JAGS	ARS	Normal Approx.	Sufficient Reduc.	JAGS	ARS	Normal Approx.	Sufficient Reduc.
$k = 100$	$\lambda = 25$	0.040 (0.029)	0.045 (0.028)	0.039 (0.041)	0.045 (0.041)	0.041 (−0.007)	0.046 (−0.007)	0.041 (−0.031)	0.045 (−0.031)
	$\lambda = 50$	0.033 (0.016)	0.034 (0.017)	0.036 (0.024)	0.036 (0.024)	0.040 (0.006)	0.040 (0.006)	0.040 (−0.007)	0.040 (−0.007)
	$\lambda = 100$	0.031 (0.004)	0.031 (0.005)	0.031 (0.009)	0.031 (0.009)	0.042 (−0.003)	0.042 (−0.003)	0.042 (−0.010)	0.042 (−0.010)
$k = 300$	$\lambda = 25$	0.027 (0.024)	0.028 (0.025)	0.038 (0.037)	0.038 (0.037)	0.026 (−0.011)	0.026 (−0.011)	0.037 (−0.033)	0.037 (−0.034)
	$\lambda = 50$	0.019 (0.009)	0.019 (0.010)	0.023 (0.018)	0.023 (0.018)	0.024 (−0.004)	0.024 (−0.004)	0.027 (−0.017)	0.027 (−0.017)
	$\lambda = 100$	0.018 (0.001)	0.018 (0.002)	0.018 (0.006)	0.018 (0.006)	0.023 (−0.004)	0.023 (−0.004)	0.024 (−0.011)	0.024 (−0.011)
$k = 500$	$\lambda = 25$	0.025 (0.023)	0.025 (0.024)	0.036 (0.036)	0.036 (0.036)	0.023 (−0.013)	0.023 (−0.013)	0.024 (−0.036)	0.024 (−0.036)
	$\lambda = 50$	0.016 (0.009)	0.016 (0.010)	0.020 (0.018)	0.020 (0.018)	0.019 (−0.005)	0.019 (−0.005)	0.023 (−0.018)	0.023 (−0.018)
	$\lambda = 100$	0.014 (0.005)	0.014 (0.006)	0.015 (0.010)	0.015 (0.010)	0.017 (2e-04)	0.017 (3e-04)	0.018 (−0.007)	0.018 (−0.007)

Table 3: Simulation results: The Mean Absolute Deviation and Bias (in parenthesis) for parameters expit(β_1) and expit($\beta_1 + 1.28\sigma$) based on 100 Monte-Carlo replications.

JAGS) are all implemented in R (MCMC iterations is 5000, with the first 500 as burn-in) while the core function in JAGS has been optimized for mass use and is written in C++, which is much faster than R.

We summarize the mean absolute deviation (MAD) and the bias of the posterior medians for each parameter of interest based upon 100 Monte-Carlo replications in Tables 3 and 4. We find that both the MAD and bias decrease as either k or λ increases in most cases. The estimation accuracy of all four methods are comparable in most cases. Both the normal approximation and sufficient reduction seem to work quite well, especially for larger values of k and λ. When λ is small, e.g., first row of Table 3, normal approximation works better than ARS and SR, and this advantage becomes less noticeable as the λ increases. This is expected because λ is the mean parameter for n_i, which controls the accuracy of the normal approximation to a binomial likelihood with n_i trials. In other words, a smaller λ value leads to a worse normal approximation and SR will perform even worse because of the additional loss of information. Similar findings have been observed in Tan (2021) and Goplerud (2021) when normal approximation is used in variational inference.

We also use wall time, the elapsed time the computer used to execute the program, to quantify the computational efficiency for three methods. The mean wall time for each of the three methods is summarized in Table 5. We find that the sufficient reduction method is faster than the normal approximation and both of those are faster than the adaptive-rejection method we implemented. Wall time for the SR method improves over the normal approximation as k grows, as expected.

Next we relate the average running time in Table 5 to the computational complexity of normal approximation and SR methods. We assume p is fixed. For the normal approx-

		$\exp(\beta_2)$				$\exp(\beta_3)$			
		JAGS	ARS	Normal Approx.	Sufficient Reduc.	JAGS	ARS	Normal Approx.	Sufficient Reduc.
$k = 100$	$\lambda = 25$	0.171 (−0.043)	0.170 (−0.044)	0.162 (−0.068)	0.161 (−0.068)	0.247 (0.051)	0.246 (0.051)	0.232 (0.056)	0.232 (0.057)
	$\lambda = 50$	0.167 (−0.020)	0.166 (−0.020)	0.158 (−0.037)	0.158 (−0.037)	0.265 (0.046)	0.264 (0.045)	0.252 (0.047)	0.253 (0.047)
	$\lambda = 100$	0.163 (0.005)	0.162 (0.006)	0.157 (−0.004)	0.156 (−0.005)	0.266 (0.013)	0.266 (0.013)	0.259 (0.015)	0.258 (0.015)
$k = 300$	$\lambda = 25$	0.097 (−0.047)	0.097 (−0.047)	0.100 (−0.069)	0.100 (−0.068)	0.143 (0.009)	0.142 (0.009)	0.133 (0.016)	0.132 (0.016)
	$\lambda = 50$	0.097 (−0.002)	0.098 (−0.002)	0.093 (−0.019)	0.093 (−0.020)	0.161 (0.018)	0.161 (0.018)	0.154 (0.022)	0.153 (0.022)
	$\lambda = 100$	0.095 (−0.006)	0.095 (−0.006)	0.092 (−0.017)	0.092 (−0.017)	0.145 (0.013)	0.145 (0.013)	0.142 (0.016)	0.141 (0.016)
$k = 500$	$\lambda = 25$	0.080 (−0.046)	0.080 (−0.046)	0.086 (−0.069)	0.086 (−0.069)	0.127 (0.021)	0.125 (0.021)	0.117 (0.028)	0.117 (0.028)
	$\lambda = 50$	0.078 (−0.009)	0.078 (−0.009)	0.076 (−0.026)	0.076 (−0.026)	0.113 (0.009)	0.113 (0.009)	0.107 (0.014)	0.107 (0.014)
	$\lambda = 100$	0.072 (−0.008)	0.072 (−0.008)	0.0711 (−0.019)	0.071 (−0.019)	0.114 (−0.025)	0.113 (−0.024)	0.110 (−0.022)	0.110 (−0.021)

Table 4: Simulation results: The Mean Absolute Deviation and Bias (in parenthesis) for parameters $\exp(\beta_2)$ and $\exp(\beta_3)$ based on 100 Monte-Carlo replications.

k	λ	Normal	SR	ARS
100	25	152	121	2064
	50	294	233	4149
	100	270	220	4163
300	25	453	214	6256
	50	488	209	6300
	100	1293	441	12077
500	25	1008	431	9965
	50	1053	440	10496
	100	1111	472	10770

Table 5: Simulation results: Mean Wall Time (in seconds) for proposed normal approximation method (Normal), sufficient reduction (SR) and ARS.

imation method in Section 2.1, its computational complexity in theory is $O(k^3 + \lambda k)$, where the k^3 term comes form the matrix inversion of k by k matrices in Proposition 1, and the λk term comes from the matrix multiplication for n_i times where n_i is generated following a Poisson distribution with mean parameter of λ in this simulation. For the SR method, its computational complexity is $O(k^2 + \lambda k)$, where the k^2 term comes from the matrix multiplication in Equation (9). It is hence clear that the computational gain of the proposed SR method is mainly driven by avoiding inverting $k \times k$ matrices, which becomes substantial as k becomes large. These results are reflected in Table 5, e.g., the time saving of SR over normal approximation improves as k becomes larger because both k^2 and k^3 dominate the λk term. When k is small (e.g., 100), the computational time nearly doubles as λ changes from 25 to 50 as expected. On the other hand, when k

is large (e.g., 500), increasing λ from 25 to 100 does not quite change the computational time. For other entries in the table (e.g., $k = 300$), the results are less interpretable probably because the terms associated with k^3 and λk are at comparable orders.

5 Extension to a two-level logistic mixed regression model

5.1 Model setup

In this section, we further demonstrate the proposed normal approximation and sufficient reduction ideas with a two-level logistic regression model as follows,

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad \text{logit}(p_{ij}) = X_{ij}\beta + u_i, \quad u_i \overset{\perp}{\sim} N(S_i\gamma, \tau^{-1}), \tag{11}$$

where Y_{ij} is a binary response variable for j^{th} subject within cluster i with $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$, X_{ij} is a $1 \times p$ vector of subject-level covariates, and u_i is a normal random effect for the i^{th} cluster with a precision parameter τ. We center u_i on $S_i\gamma$ where S_i is an $1 \times q$ vector of unique covariate information for cluster i, γ is the associated coefficient parameter, and β is the $1\times p$ vector of regression coefficients at the observational unit level.

We use independent normal priors for the regression parameters β and γ and consider a gamma prior for τ as follows,

$$\beta \sim N_p(B, C)\ \gamma \sim N_q(B_0, C_0),\ \tau \sim \text{Gamma}(a/2, b/2).$$

Then the joint posterior is proportional to

$$\begin{aligned}
&\exp\left\{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}(X_{ij}\beta + u_i) - \log(1 + \exp(X_{ij}\beta + u_i))\right\} \\
&\quad \times \exp\left\{-\sum_{i=1}^{k}\frac{\tau(u_i - S_i\gamma)^2}{2}\right\}\tau^{k/2} \\
&\quad \times \exp\left\{-\frac{(\beta - B)'C^{-1}(\beta - B)}{2} - \frac{(\gamma - B_0)'C_0^{-1}(\gamma - B_0)}{2}\right\}\tau^{a/2-1}\exp\left\{-\frac{b\tau}{2}\right\}.
\end{aligned} \tag{12}$$

To further simplify the algebra, we introduce some matrix notation. Let $u = (u_1, \ldots, u_k)'$ such that $u \sim N_k(S\gamma, I_k\tau^{-1})$, where $S = (S_1, \ldots, S_k)$. Let $N = \sum_{i=1}^{k} n_i$ be the total number of observations, $X = (X_{11}', X_{12}', \ldots, X_{kn_k}')'$ be an $N\times p$ matrix, and $y = (y_{11}, y_{12}, \ldots, y_{kn_k})'$ be an $N\times 1$ vector of responses of y_{ij}s. Finally, we define Z_{ij} as a $1 \times k$ vector of zeros except the i^{th} entry is one for $i = 1, \ldots, k$, and let $Z = (Z_{11}', Z_{12}', \ldots, Z_{kn_k}')'$ be an $N\times k$ matrix. Then the joint posterior kernel in (12) can be rewritten as

$$\exp\left\{y'(X\beta + Zu) - \frac{\tau(u - S\gamma)'(u - S\gamma)}{2} - \frac{(\beta - B)'C^{-1}(\beta - B)}{2}\right.$$

$$- \frac{(\gamma - B_0)' C_0^{-1} (\gamma - B_0)}{2} \Bigg\}$$
$$\times \exp\left\{ - \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log(1 + \exp(X_{ij}\beta + Z_{ij}u)) \right\} \tau^{a/2+k/2-1} \exp\left\{ -\frac{b\tau}{2} \right\}. \quad (13)$$

Based on this joint posterior kernel, we can easily obtain the conditional distribution of τ and γ given other parameters as follows,

$$\tau \,|\, \text{else} \sim \text{Gamma}\left(\frac{a+k}{2}, \frac{b + (u - S\gamma)'(u - S\gamma)}{2} \right),$$
$$\gamma \,|\, \text{else} \sim N_q\left((S'S + C_0^{-1})^{-1} (S'S\hat{\gamma} + C_0^{-1} B_0), (S'S + C_0^{-1})^{-1} \right). \quad (14)$$

The conditional distribution for (β, u) is proportional to

$$\exp\left\{ y'(X\beta + Zu) - \frac{\tau(u - S\gamma)'(u - S\gamma)}{2} - \frac{(\beta - B)' C^{-1} (\beta - B)}{2} \right\}$$
$$\times \exp\left\{ - \sum_{i=1}^{k} \sum_{j=1}^{n_i} \log(1 + \exp(X_{ij}\beta + Z_{ij}u)) \right\}, \quad (15)$$

which is not recognizable. In the literature, it is common to use a Metropolis sampling or adaptive-rejection sampling approach to sample from this type of distribution; it is our goal in the next section to explore the possibility of a normal approximation.

5.2 Normal approximation and sufficient reduction

We first consider the normal approximation for (15). By some calculation (details provided in Supplementary File Section S3), we have the following normal approximation results for the conditional distributions of β and u,

$$\beta \,|\, \text{else} \sim N_p\left(\mu_\beta, (X' D_{\tilde{\pi}(1-\tilde{\pi})} X + C^{-1})^{-1} \right), \quad (16)$$
$$u \,|\, \text{else} \sim N_k\left(\mu_u, (Z' D_{\tilde{\pi}(1-\tilde{\pi})} Z + \tau I_k)^{-1} \right), \quad (17)$$

where

$$\mu_\beta = (X' D_{\tilde{\pi}(1-\tilde{\pi})} X + C^{-1})^{-1} \left(X' D_{\tilde{\pi}(1-\tilde{\pi})} X\tilde{\beta} + C^{-1} B + X' D_{\tilde{\pi}(1-\tilde{\pi})} Z(\tilde{u} - u) \right)$$
$$\mu_u = (Z' D_{\tilde{\pi}(1-\tilde{\pi})} Z + \tau I_k)^{-1} \left(Z' D_{\tilde{\pi}(1-\tilde{\pi})} Z\tilde{u} + \tau S\gamma + Z' D_{\tilde{\pi}(1-\tilde{\pi})} X(\tilde{\beta} - \beta) \right),$$

and $D_{\tilde{\pi}(1-\tilde{\pi})} = \text{diag}\{\text{expit}(X_{ij}\tilde{\beta}_{ij} + Z_{ij}\tilde{u}_{ij}) \colon j = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, k\}$, where the definitions of $\tilde{\beta}$ and $\tilde{u}$ are given in Supplementary File Section S3. Based on this result, we can easily design a block Gibbs sampling algorithm that allows us to directly sample from β, u together with γ and τ using (14).

Next we consider the sufficient reduction by letting $n_i \to \infty$ for each $i = 1, \ldots, k$ and having $k \to \infty$. Under this setting, we would expect the dimension for u is large, such

that a direct sampling from its distribution is expensive. Similarly with the definition of (T_1, T_2) in Section 2.2, we define a new set of statistics as follows,

$$T_1(u) = (S'S)^{-1}S'u, \quad T_2(u) = u'(I_k - S(S'S)^{-1}S')u,$$
$$T_3(u) = (X'D_{\tilde{\pi}(1-\tilde{\pi})}X + C^{-1})^{-1}(X'D_{\tilde{\pi}(1-\tilde{\pi})}Z)u.$$

Here $T = (T_1, T_2, T_3)$ can be viewed as "sufficient reduction" for u. Then the conditional distributions for τ, γ, and β are

$$\begin{aligned}
\tau \,|\, \text{else} &\sim \text{Gamma}\left(\frac{a+k}{2}, \frac{b + T_2(u) + (T_1(u)-\gamma)'S'S(T_1(u)-\gamma)}{2}\right), \\
\gamma \,|\, \text{else} &\sim N_q\left((\tau S'S + C_0^{-1}(\tau S'ST_1(u) + C_0^{-1}B_0), (\tau S'S + C_0^{-1})^{-1}\right), \\
\beta \,|\, \text{else} &\sim N_p\Big((X'D_{\tilde{\pi}(1-\tilde{\pi})}X + C^{-1})^{-1}(X'D_{\tilde{\pi}(1-\tilde{\pi})}(X\tilde{\beta} + Z\tilde{u}) + C^{-1}B) - T_3(u), \\
&\qquad (X'D_{\tilde{\pi}(1-\tilde{\pi})}X + C^{-1})^{-1}\Big). \qquad (18)
\end{aligned}$$

We will also need a conditional distribution of $T \,|\, \text{else}$ to form a Gibbs sampler for our SR method. Similarly with Proposition 2, we consider a normal approximation for their joint conditional distribution. This approximation intuitively makes sense because T_1 and T_3 are affine transformations of a multivariate normal distribution, and T_2 is a weighted sum of squares of normally distributed random variables. We present the means and covariances for (T_1, T_2, T_3) in Supplementary File Section S3.2.

In terms of computational complexity, normal approximation is at $O(N^2 + k^3)$ and SR is at $O(N^2 + k^2) = O(N^2)$ since $N = \sum_{i=1}^{k} n_i$. Therefore we expect a substantial computational saving when k is large and n_i's are fixed.

5.3 Cow abortion data analysis

We use the developed methodology to analyze cow abortion data. This data set consist of 13145 cows across 9 herds, where the herds vary in size from 116 to 2711 cows, with a mean (median) herd size of 1460 (1490) cows, and a standard deviation of approximately 719 cows. The data were previously analyzed by Hanson et al. (2003), and the main interest here is to model the probability of a spontaneous abortion in dairy cows given two covariates, the gravidity (GR) and days open (DO). Here gravidity refers to the number of times the cow was pregnant before the current pregnancy (mean is 3.2 pregnancies, median is 3, and the standard deviation is 1.5) and days open is the number of days between the most recent birth and conception (mean is 95.6 days, median is 79, and the standard deviation is 48.4). We consider the following two-level logistic model,

$$Y_{ij} \sim \text{Bernoulli}(p_{ij}), \quad \text{logit}(p_{ij}) = X_{\text{DO},ij}\beta_1 + X_{\text{GR},ij}\beta_2 + u_i, \quad u_i \overset{\text{i.i.d.}}{\sim} N(\gamma, \tau^{-1}),$$

where Y_{ij} is a binary response with 1 indicating a recorded abortion for j^{th} cow in the i^{th} herd for all j cows ($j = 1, 2, \ldots, n_i$) in herd i ($i = 1, 2, \ldots, 9$). Since both covariates that we consider are quantitative, we standardize them prior to analysis. We then consider

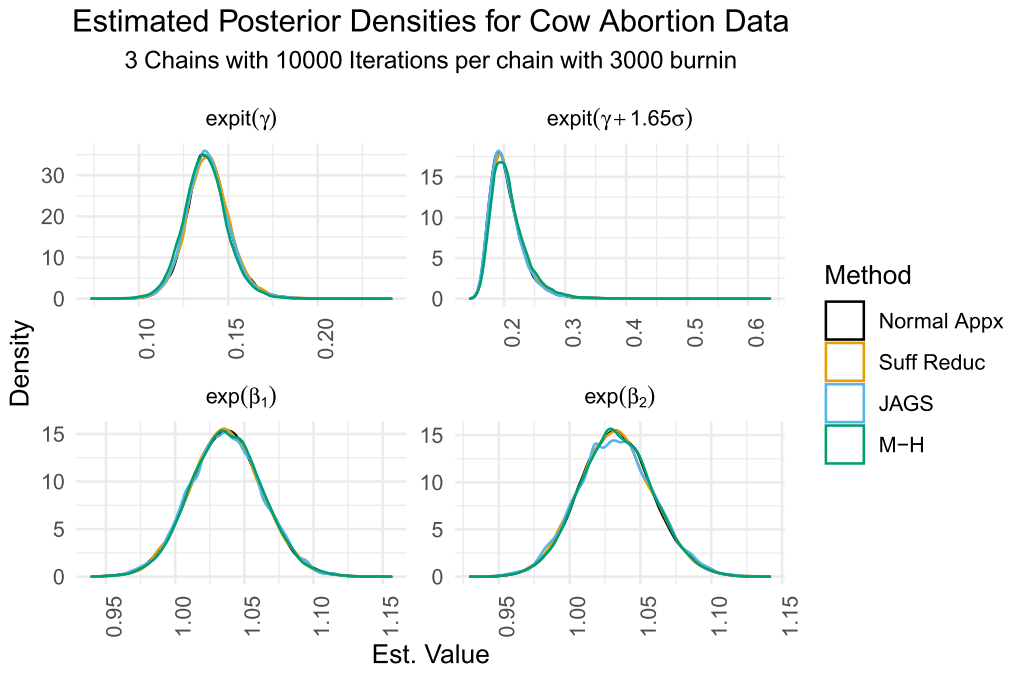

Figure 4: Posterior Densities of Quantities of Interest in the Cow Abortion Example.

the following independent priors for the parameters. A detailed discussion of the choice of prior is provided in Supplementary File Section S3.3.

$$\beta \sim N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}10 & 0\\0 & 10\end{pmatrix}\right), \quad \gamma \sim N(-2, 10), \quad \tau \sim \text{gamma}(1, 0.05).$$

We implemented four methods to analyze these data: (i) the adaptive rejection sampling used in JAGS, which will serve as the benchmark for comparison; (ii) the proposed normal approximation within the Gibbs sampling; (iii) sufficient reduction method; and (iv) a Metropolis Hasting algorithm by using the normal approximation as a proposal distribution. Method (iv) is new to the presentation. For each of those methods, we ran three Markov Chains, each with 10,000 iterations and the first 3,000 were discarded as burn-in.

We focus on four quantities of interest, (a) expit(γ), which is the probability of an average cow (taking average values for cow's days open and gravidity) abortion; (b) expit($\gamma + 1.65\sigma$), which is the 95^{th} percentile for the probability of abortion for the average cow, (c) $\exp(\beta_1)$, the odds ratio for days open; and (d) $\exp(\beta_2)$, the odds ratio for gravidity. We summarize the posterior densities for those four quantities in Figure 4, it is clear that the posterior densities obtained from all four methods are almost identical, which confirms the excellent approximating accuracy for our proposed three methods. The good normal approximation performance is probably due to the large values of n_i (number of cows for each herd), e.g., the smallest herd has 116 cows.

Parameter	Method	2.5%tile	median	97.5%tile	mean	sd
expit(γ)	Normal Appx	0.12	0.14	0.16	0.14	0.01
	Suf. Reduc.	0.12	0.14	0.17	0.14	0.01
	JAGS	0.12	0.14	0.16	0.14	0.01
	M-H	0.11	0.14	0.16	0.14	0.01
expit($\gamma + 1.65\sigma$)	Normal Appx	0.17	0.20	0.27	0.20	0.03
	Suf. Reduc.	0.17	0.20	0.28	0.21	0.03
	JAGS	0.17	0.20	0.27	0.20	0.03
	M-H	0.17	0.20	0.28	0.21	0.03
exp(β_1)	Normal Appx	0.99	1.04	1.09	1.04	0.03
	Suf. Reduc.	0.99	1.04	1.09	1.04	0.03
	JAGS	0.99	1.04	1.09	1.04	0.03
	M-H	0.99	1.04	1.09	1.04	0.03
exp(β_2)	Normal Appx	0.98	1.03	1.08	1.03	0.03
	Suf. Reduc.	0.98	1.03	1.08	1.03	0.03
	JAGS	0.98	1.03	1.09	1.03	0.03
	M-H	0.98	1.03	1.08	1.03	0.03

Table 6: Posterior Estimates and Quantities using Cow Abortion Data.

We also present the summary statistics for the quantities of interest in Table 6. Since we have standardized the covariates, the interpretation of the odds ratios is based on a one standard deviation increase in the covariate. For example, exp(β_1), represents the change in the odds of spontaneous abortion between two cows with identical gravidity values, but one cow has a days open value that is 48.4 days longer than the other. Similarly, exp(β_2), is the change in odds of spontaneous abortion between two cows with identical days open, but one standard deviation more in gravidity value, approximately 1.5 more previous pregnancies.

The computational time (in seconds) is 725 for normal approximation, 873 for sufficient reduction, and 948 for the Metropolis Hasting algorithm using the normal approximation as a proposal. Sufficient reduction is slower than normal approximation in this example since $k = 9$ is quite small. The M-H algorithm takes the longest time because it requires calculating the normal approximation as a middle step.

6 Discussion

In this paper, we explored the idea of large sample approximation for enhanced MCMC sampling in two GLMMs. This was implemented by replacing the usual Metropolis–Hastings and adaptive-rejection sampler with a direct sampling from a normal distribution within the Gibbs sampler. In the future, it will be of interest to extend the idea of normal approximation and sufficient reduction for more complex models, e.g., general forms of GLMM and additional hierarchies in the model. Computationally, as complexity grows, it will be of interest to implement the proposed method in a more efficient software platform such as C++ and to develop faster algorithms for precision matrix

calculation. Missing from our presentation so far is an investigation of the effects of "big data" values for k on various methods, including those not considered here. We plan to investigate potential improvements for data sets with values of k that are orders of magnitude larger than those considered here.

Supplementary Material

Web-based Supplementary File for "Normal approximation for Bayesian mixed effects binomial regression model" (DOI: 10.1214/00-BA1312SUPP; .pdf). The supplementary material contains technical details and proofs that the full conditionals for u and for $T(u)$ are asymptotically normal, as well as details for the more complex model.

References

Baghishani, H. and Mohammadzadeh, M. (2012). "Asymptotic normality of posterior distributions for generalized linear mixed models." *Journal of Multivariate Analysis*, 111: 66–77. MR2944406. doi: https://doi.org/10.1016/j.jmva.2012.05.003. 416

Berman, B., Johnson, W. O., and Shen, W. (2022) "Supplementary Material for "Normal approximation for Bayesian mixed effects binomial regression models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1312SUPP. 419

Breslow, N. E. and Clayton, D. G. (1993). "Approximate Inference in Generalized Linear Mixed Models." *Journal of the American Statistical Association*, 88(421): 9–25. MR1394064. doi: https://doi.org/10.2307/2291379. 415, 416

Chen, C. F. (1985). "On asymptotic normality of limiting density functions with Bayesian implications." *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(3): 540–546. MR0844485. 416

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. CRC Press. MR2682928. 421

Cochran, W. G. (1952). "The χ^2 Test of Goodness of Fit." *The Annals of Mathematical Statistics*, 23(3): 315–345. MR0049531. doi: https://doi.org/10.1214/aoms/1177729380. 422

Diggle, P., Diggle, P. J., Heagerty, P., Liang, K. Y., Heagerty, P. J., and Zeger, S. (2002). *Analysis of Longitudinal Data*. Oxford: Oxford University Press. MR2049007. 415

Fong, Y., Rue, H., and Wakefield, J. (2010). "Bayesian inference for generalized linear mixed models.,." *Biostatistics*, 11(3): 397–412. 416

Gamerman, D. (1997). "Sampling from the posterior distribution in generalized linear mixed models." *Statistics and Computing*, 7(1): 57–68. 415

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). "Efficient parametrisations for

normal linear mixed models." *Biometrika*, 82(3): 479–488. MR1366275. doi: https://doi.org/10.1093/biomet/82.3.479. 418

Gilks, W. R. and Wild, P. (1992). "Adaptive Rejection Sampling for Gibbs Sampling." *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2): 337–348. 417, 425

Goplerud, M. (2021). "Fast and Accurate Estimation of Non-Nested Binomial Hierarchical Models Using Variational Inference." *Bayesian Analysis*, 1(1): 1–28. 426

Hadfield, J. D. (2010). "MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package." *Journal of statistical software*, 33(2): 1–22. 416

Hanson, T., Bedrick, E. J., Johnson, W. O., and Thurmond, M. C. (2003). "A mixture model for bovine abortion and foetal survival." *Statistics in Medicine*, 22(10): 1725–1739. 430

McCulloch, C. E. (1996). "Conference on Applied Statistics in Agriculture." In *An Introduction to Generalized Linear Mixed Models*, volume 8. MR1993816. 415

McCulloch, C. E. (1997(1997). "Maximum likelihood algorithms for generalized linear mixed models." *Journal of the American Statistical Association*, 92(437): 162–170. MR1436105. doi: https://doi.org/10.2307/2291460. 415

Plummer, M. (2012). "JAGS Version 3.3. 0 user manual." Technical report, International Agency for Research on Cancer, Lyon, Franc. 415

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 417

Rue, H., Martino, S., and Chopin, N. (2009). "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2): 319–392. MR2649602. doi: https://doi.org/10.1111/j.1467-9868.2008.00700.x. 416

Spiegelhalter, D., Thomas, A., Best, N., and Lunn, D. (2003). "WinBUGS user manual." Technical report. 415

Su, C. L. and Johnson, W. O. (2006). "Large-sample joint posterior approximations when full conditionals are approximately normal: application to generalized linear mixed models." *Journal of the American Statistical Association*, 101(474): 795–811. MR2256190. doi: https://doi.org/10.1198/016214505000001311. 416

Tan, L. S. (2021). "Use of model reparametrization to improve variational Bayes." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(1): 30–57. MR4220983. doi: https://doi.org/10.1111/rssb.12399. 426

Tierney, L. and Kadane, J. (1986). "Accurate approximations for posterior moments and marginal densities." *Journal of the American Statistical Association*, 81(393): 82–86. MR0830567. 416

Walker, A. M. (1969). "On the asymptotic behaviour of posterior distributions." *Journal of the Royal Statistical Society: Series B (Methodological)*, 31(1): 80–88. MR0269000. 416

Weng, R. C. and Tsai, W. C. (2008). "Asymptotic posterior normality for multiparameter problems." *Journal of statistical planning and inference*, 138(12): 4068–4080. MR2455988. doi: https://doi.org/10.1016/j.jspi.2008.03.034. 416

Yee, J. L., Johnson, W. O., and Samaniego, F. J. (2002). "Asymptotic approximations to posterior distributions via conditional moment equations." *Biometrika*, 89(4): 755–767. 416

Zeger, S. L. and Karim, M. R. (1991). "Generalized linear models with random effects: a Gibbs sampling approach." *Journal of the American Statistical Association*, 86: 79–86. MR1137101. 415

Acknowledgments

The authors thank the editor Michele Guindani, the associate editor, and anonymous reviewers for their valuable comments and suggestions that greatly improved the paper.