Bayesian Logistic Regression for
Small Areas with Numerous Households

Balgobin Nandram, Lu Chen, Shuting Fu and Binod Manandhar
Department of Mathematical Sciences, Worcester Polytechnic Institute,
Worcester, MA, USA

Abstract

We analyze binary data, available for a relatively large number (big data) of families (or households), which are within small areas, from a population-based survey. Inference is required for the finite population proportion of individuals with a specific character for each area. To accommodate the binary data and important features of all sampled individuals, we use a hierarchical Bayesian logistic regression model with each family (not area) having its own random effect. This modeling helps to correct for overshrinkage so common in small area estimation. Because there are numerous families, the computational time on the joint posterior density using standard Markov chain Monte Carlo (MCMC) methods is prohibitive. Therefore, the joint posterior density of the hyper-parameters is approximated using an integrated nested normal approximation (INNA) via the multiplication rule. This approach provides a sampling-based method that permits fast computation, thereby avoiding very time-consuming MCMC methods. Then, the random effects are obtained from the exact conditional posterior density using parallel computing. The unknown nonsample features and household sizes are obtained using a nested Bayesian bootstrap that can be done using parallel computing as well. For relatively small data sets (e.g., 5000 families), we compare our method with a MCMC method to show that our approach is reasonable. We discuss an example on health severity using the Nepal Living Standards Survey (NLSS).

Key words: Bayesian bootstrap, Big data, Integrated nested normal approximation, Overshrinkage, Sampling based method, Survey weights.

---

1. Introduction

In the second Nepal Living standards Survey (NLSS II), there are data from households. One question of interest is health status (good versus poor health), a binary variable, and there are several covariates that can explain the binary outcomes. Our interest is to provide smoothed estimates of the household proportions of members in good health for both sampled and non-sampled households. This is a general question. Direct estimation is not reliable for such a problem (not many members in each household), and there is a need to borrow strength from

---

Corresponding Author: Balgobin Nandram
E-mail: balnan@wpi.edu

the ensemble, as in small area estimation. This is generally done using Bayesian logistic regression. This problem is exasperated because there are numerous households (small areas). A hierarchical Bayesian logistic regression model with random effects is used to capture the variation of members within and across households. Markov chain Monte Carlo (MCMC) methods require extensive monitoring and the entire procedure to fit the model is time consuming mostly because there are numerous random effects. Therefore, the joint posterior density of the hyperparameters is approximated using an integrated nested normal approximation (INNA) via the multiplication rule of probability. This approach provides a sampling-based method that permits fast computation, thereby avoiding very time-consuming MCMC methods from the exact method. In this paper, our main contribution is to obtain an approximate joint posterior density for the hierarchical Bayesian logistic regression model and to get reasonable estimates and standard errors of small area parameters (e.g., household proportions).

The estimation of parameters of the binary logistic regression with random effects is not straight forward due to fact that the likelihood involves multiple integrals. In case of Bayesian analysis, a natural approach to inference in mixed models was proposed by Paulino, Silva and Achcar (2005). They estimated the random effects, which are treated as parameters in the presence of misclassified data. They also showed that if the posterior distribution is not possible to be obtained analytically, MCMC methods can be used to approximate them. Souza and Migon (2010) proposed that the inference problem can be solved in an easier way if the random effects of the mixed models are distributed as Student-t or finite mixture of normal distributions. Liu and Dey (2008) used prior distributions like skew-normal distribution. Santos, Loschi and Arellano-Valle (2013) provided different prior and posterior interpretations for the parameters in the logistic regression model with random intercepts when skew normal distribution are assumed to model random effects. They obtained the prior distributions for the different parameters and they discussed odds ratio and median odds ratio using skew-normal distributions for the random effects. They concluded that the misspecification of the random effects parameters can give poor estimate. Larsen et al. (2000) gave interpretations of the parameters in the logistic regression model with random effects. Finally, Chen, Ibrahim and Kim (2008) described important properties of logistic regression for a single population (no random effects), and showed how to implement Jeffreys's prior for binomial regression models.

Our work on logistic regression dates back to Nandram (1989) who discussed discrimination between the logit and the complementary log-log link functions. Nandram (2000) reviewed the paper of Ghosh et al. (1998) on generalized linear models, logistic regression and Poisson regression being two important special cases. Nandram and Erhardt (2005) showed how to analyze binary data with covariates to maintain conjugacy for both logistic and Poisson regression model. Nandram and Choi (2010) showed how to analyze binary data with covariates under nonignorable nonresponse. Different from most researchers, Nandram and Choi (2010) showed how to use logistic regression to obtain propensity scores, an interesting part of their paper, for small area estimation. Roberts, Rao and Kumar (1987) discussed logistic regression for sample survey data (not small area estimation). Nandram and Chen (1996) show how to accelerate the Gibbs sampler for a model with latent variables introduced earlier by Albert and Chib (1993) for Bayesian probit analysis.

Albert and Chib (1993) started an innovative stream of research on Bayesian probit analysis, not logistic regression; they agrued that logistic regression is approximately a special case of their probit analysis. However, we now know that this approximation is poor in the tails and

their algorithm is a poorly mixing Gibbs sampler (Holmes and Held 2006). For probit analysis, Holmes and Held (2006) showed how to solve this mixing problem by incorporating latent variables and using the block Gibbs sampler (i.e., some variables are drawn simultaneously). Holmes and Held (2006) extended their approach to logistic regression, albeit for a single sample, not for small area estimation as in the case of numerous small areas that we are studying here. Technically, even for a simple sample, their sampling algorithm is very complicated using rejection sampling, the Kolmogorov-Smirnov distribution, part of a representation of the standard logistic distribution, and a generalized inverse Gaussian distribution. However, once a user-friendly program is available, the complexity does not matter. Note that for simple logistic regression (i.e., a single sample), the Metropolis sampler or rejection sampling can be used in a straightforward manner, and this is faster and much simpler than the method of Holmes and Held (2006). However, it will be extremely difficult to apply their computational techniques in our case simply because the sum of two logistic random variables (we have two error terms, not one) is not another logistic random variable.

The other side of our application is that there are numerous small areas (households) and MCMC methods cannot handle them in real time. So our problem can be classified as a "big data" problem. Scott et al. (2013) defined "big data" as data that are too big to comfortably process on a single machine, either because of processor, memory, or disk bottlenecks. They considered consensus Monte Carlo methods which split the data to several machines. Communication between large numbers of machines is expensive (regardless of the amount of data being communicated), so there is a need for algorithms that perform distributed approximate Bayesian analyses with minimal communication. Consensus Monte Carlo operates by running a separate Monte Carlo algorithm on each machine, and then averaging individual Monte Carlo draws across machines. One of the examples they gave is on a hierarchical Poisson regression model (very close to logistic regression). But certainly how to split the data is problematic. Miroshnikov and Colon (2015) described parallel MCMC methods for non-Gaussian posterior distributions. Fortunately, in survey sampling the design generally uses stratification which is not artificial, and in this case, consensus Monte Carlo may not be needed; it will be a good idea for a large stratum.

The procedure we use to approximate the posterior density of the parameters of the hierarchical Bayesian logistic regression model, called the integrated nested normal approximation (INNA), has a closed resemblance to the integrated nested Laplace approximation (INLA); see Rue, Martino and Chopin (2009). INNA uses a sampling-based procedure, that is accommodated by the multiplication rule of probability; currently INLA is a fairly popular method for making approximations in complicated hierarchical Bayesian model. INLA is a promising alternative to MCMC for big data analysis. However, it requires posterior modes and, for numerous small areas, computation of modes becomes time-consuming and challenging for logistic regression model or any generalized linear mixed models. Yet INLA has found many useful applications. See, for example, Fong, Rue and Wakefield (2010) for an application on Poisson regression, and Illian, Sørbye and Rue (2012) for a realistic application on spatial point pattern data. We note that INLA can be problematic especially for logistic and Poisson hierarchical regression models, even if the modes can be computed. For example, Ferkingstad and Rue (2015), attempting to improve INLA, used a copula-based correction which adds complexity to INLA. For a comparison of INLA and MCMC, the paper by Held, Schrödle and Rue (2010) for cross-validatory predictive checks is interesting. Unfortunately, the computational cost of INLA is exponential in the dimension of the parameter space (or hyperparameter space

in the case of hierarchical models).

Of course, there are many other approximations in Bayesian statistics some of which can apply directly to logistic regression. Approximate Bayesian methods (ABC) were introduced in population genetics (e.g., Beaumont, Zhang and Balding 2002) to deal with intractable likelihood functions and it uses summary statistics. An important advance was made by Fearhead and Prangle (2012), who obtained a more principled approach to the construction of summary statistics. Jaakkola and Jordan (2000), Faes, Omerod and Wand (2011) and Omerod and Wand (2010) studied variational Bayes methods. Variation methods are very complicated, even for the simplest problem, logistic regression without random effects (Jaakkola and Jordan 2000), the analysis is not simple. Moreover, the approximate posteriors delivered by variational Bayes give good accuracy for individual marginal distributions, but not for the joint distribution as a whole.

We will not use any of these approximations. The nearest to our procedure is INLA that requires posterior modes and it is computationally costly to run the application we have in mind because of lack of conjugacy, numerical optimization is needed. Variational Bayes methods are mathematically too complicated and ABC is not accurate enough even though this is also an active area of research. Instead of finding the posterior modes, INNA finds approximate modes in closed form, facilitated by the empirical logistic transform (Cox ans Snell 1972). Here both the gradient vector (gradient vector is not zero though) and the gradient term is kept in the second order Taylor's series expansion of the posterior distribution of the regression coefficients. So our method does not need posterior modes as in INLA; this is an enormous saving in computing time, even more so for numerous households.

The plan of the rest of the paper is as follows. In Section 2, we describe our main contribution about our approximation to the joint posterior density. In particular, we describe the integrated nested normal approximation (INNA) and some theoretical results are provided. In Section 3, we present an illustrative example using the Nepal Living Standards Survey (NLSS II). We compare the approximate method with the exact method, which is presented in Appendix A. It is worth noting that the word "exact" refers to MCMC without further approximation. In Section 4, we have discussions and two extensions, both of them can be used to accommodate the NLSS II data better. Additional technical details are given in the appendices.

2.   Approximate Theory and Method

The method we developed here for many small areas can be applied to any generalized linear model in the same manner. Of course, the specific models will be different. For example, for the model for Poisson regression is different from the model for logistic regression. However, note that for logistic regression model, the unit level (binary data, not binomial counts) are used and for Poisson regression model the count data are collected at an aggregate level. Our application is on the Nepal Living Standards Survey (NLSS II) and we have binary data (good health versus poor health) for each individual within a household, and these households are within wards. Our theory applies to individuals within households or individuals within wards. We note that the number of members in the $i^{th}, i = 1, \ldots, L$, household in the population is N_i and all household members are sampled, but not all households in a ward are sampled. Our model will hold for all L households but it is convenient to present the model for the $\ell \leq L$ sampled households.

Our logistic regression model applies to all $L \geq \ell$ population households. We know how many households are in each ward (sampled or nonsampled) but we do not know how many members are in each nonsampled household. Let y_{ij} and $\underset{\sim}{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp-1})^T$, denote the responses and the p vector of covariates with an intercept ($x_{ij0} = 1$).

A standard hierarchical Bayesian logistic regression model is

$$y_{ij} | \underset{\sim}{\beta}, v_i \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{e^{x'_{ij}\underset{\sim}{\beta}+v_i}}{1 + e^{x'_{ij}\underset{\sim}{\beta}+v_i}} \right\}, j = 1, \ldots, n_i,$$

$$v_i | \delta^2 \overset{iid}{\sim} \text{Normal}(0, \delta^2), i = 1, \ldots, \ell,$$

$$\pi(\underset{\sim}{\beta}, \delta^2) \propto \frac{1}{(1+\delta^2)^2}, \delta^2 > 0.$$

Here, $v_i, i = 1, \ldots, \ell$, are the random effects and $\underset{\sim}{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$ are the regression coefficients with δ^2, the variance of the random effects.

For our approximation methods, we will use an equivalent model. It is convenient to separate $\underset{\sim}{\beta}$ into β_0 and $\underset{\sim}{\beta}_{(0)}$, where $\underset{\sim}{\beta}_{(0)} = (\beta_1, \beta_2, \ldots, \beta_p)'$. Omitting the intercept term from the covariate $\underset{\sim}{x}_{ij}$, we have

$$y_{ij} | \mu_i, \underset{\sim}{\beta}_{(0)} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}{1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}} \right\}, j = 1, \ldots, n_i,$$

$$\mu_i | \beta_0, \delta^2 \overset{iid}{\sim} \text{Normal}(\beta_0, \delta^2), i = 1, \ldots, \ell,$$

$$\pi(\underset{\sim}{\beta}, \delta^2) \propto \frac{1}{(1+\delta^2)^2}, \delta^2 > 0, \tag{1}$$

where essentially we have made the transformation $\mu_i = v_i + \beta_0, i = 1, \ldots, \ell$.

The parameters of interest are the household proportions,

$$P_i = \frac{1}{N_i} \sum_{j=1}^{N_i} I_{ij}, \quad I_{ij} \mid v_i, \underset{\sim}{\beta} \overset{ind}{\sim} \text{Bernoulli}\{ \frac{e^{x'_{ij}\underset{\sim}{\beta}+v_i}}{1 + e^{x'_{ij}\underset{\sim}{\beta}+v_i}} \}, i = 1, \ldots, \ell. \tag{2}$$

The P_i give smoothed estimates at the household level and actually predictions for the non-sample households. A similar formula can be written down for the wards. Because we are not linking the census to the NLSS, we do not have the covariates and the number of members in each nonsampled households, both being obtained using a Bayesian bootstrap (Rubin 1981) of the original samples.

To develop the approximate methodology, we will work with the no-intercept model. Then, using Bayes' theorem, the joint posterior density for the parameters is

$$\pi(\underset{\sim}{\mu}, \underset{\sim}{\beta}, \delta^2 | \underset{\sim}{y}) \propto \left\{ \prod_{i=1}^{\ell} [\prod_{j=1}^{n_i} \frac{e^{(x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i)y_{ij}}}{1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}][\frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(\mu_i-\beta_0)^2}{2\delta^2}}] \right\} \frac{1}{(1+\delta^2)^2}. \tag{3}$$

The posterior density in (3) is a non-standard density, and there are difficulties in fitting it using Markov chain Monte Carlo methods, more so when ℓ is large. This motivates our approximate methods.

We note that given $\underset{\sim}{\beta}, \delta^2, \underset{\sim}{y}$, the μ_i are independent with

$$\pi(\mu_i \mid \underset{\sim}{\beta}, \delta^2, \underset{\sim}{y}) \propto [\prod_{j=1}^{n_i} \frac{e^{(x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i)y_{ij}}}{1+e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}][\frac{1}{\sqrt{2\pi\delta^2}}e^{-\frac{(\mu_i-\beta_0)^2}{2\delta^2}}], i = 1,\dots,\ell. \tag{4}$$

For the nonsampled households, $\mu_i \mid \beta_0, \delta^2 \overset{iid}{\sim} \text{Normal}(\beta_0, \delta^2), i = 1,\dots,\ell$. More importantly with respect to posterior inference about P_i, the conditional posterior densities for the v_i can be written down easily.

2.1. Approximation to the Posterior Density

In this section we discuss the approximation to the joint posterior density in (3).

Let $f(\underset{\sim}{\tau}) = e^{h(\underset{\sim}{\tau})}$ denote the density of a vector of parameters $\underset{\sim}{\tau}$. Let $\underset{\sim}{g}$ denote the gradient vector and H the Hessian matrix at some point $\underset{\sim}{\tau}^*$.

Lemma 2.1. Let $h(\underset{\sim}{\tau})$ be a logconcave density function with the parameter $\underset{\sim}{\tau}$. Then, $\underset{\sim}{\tau}$ approximately has a multivariate normal distribution,

$$\underset{\sim}{\tau} \sim \text{Normal}(\underset{\sim}{\tau}^* - H^{-1}\underset{\sim}{g}, -H^{-1}).$$

Proof. Simply applying the second-order multivariate Taylor series of $h(\underset{\sim}{\tau})$ at $\underset{\sim}{\tau}^*$, we have

$$f(\underset{\sim}{\tau}) \approx f(\underset{\sim}{\tau}^*) + (\underset{\sim}{\tau} - \underset{\sim}{\tau}^*)'\underset{\sim}{g} + \frac{1}{2}(\underset{\sim}{\tau} - \underset{\sim}{\tau}^*)'H(\underset{\sim}{\tau} - \underset{\sim}{\tau}^*).$$

We remark that because of logconcavity, $-H$ is positive definite. Also because we are not required to find the mode of $h(\underset{\sim}{\tau})$, $\underset{\sim}{\tau}^*$ does not have to be the solution of the gradient vector set to the zero vector. So that the term involving $\underset{\sim}{g}$ is a correction to $\underset{\sim}{\tau}^*$.   □

Momentarily, we consider a flat prior $\underset{\sim}{\beta}^*_{(0)}$ and the μ_i (i.e., fiducial inference). That is,

$$y_{ij}|\mu_i, \underset{\sim}{\beta}_{(0)} \overset{ind}{\sim} \text{Bernoulli}\left\{\frac{e^{\underset{\sim}{x}'\underset{\sim}{\beta}_{(0)}+\mu_i}}{1+e^{\underset{\sim}{x}'\underset{\sim}{\beta}_{(0)}+\mu_i}}\right\}, j = 1,\dots,n_i, i = 1,\dots,\ell,$$

$$p(\underset{\sim}{\mu}, \underset{\sim}{\beta}_{(0)}) = 1.$$

The joint posterior density is

$$\pi(\underset{\sim}{\mu}, \underset{\sim}{\beta}|\underset{\sim}{y}) \propto \prod_{i=1}^{\ell}\left\{\prod_{j=1}^{n_i} \frac{e^{(x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i)y_{ij}}}{1+e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}\right\}. \tag{5}$$

The logarithm of the joint posterior density (or log-likelihood) is

$$\Delta = h(\underset{\sim}{\tau}) = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \{(\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)} + \mu_i)y_{ij} - \log(1 + e^{\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)} + \mu_i})\}.$$

Let $\underset{\sim}{\tau}' = (\underset{\sim}{\mu}', \underset{\sim}{\beta}'_{(0)})$. First, we find a convenient point to expand the log-likelihood in a multivariate Taylor's series expansion. In Appendix B, we show how to obtain quasi-modes for $\underset{\sim}{\beta}_{(0)}$ and $\mu_i, i = 1, \ldots, \ell$, of the log-likelihood function.

First, we use the empirical logistic transform z_i to get an estimate of μ_i, where

$$z_i = \log\left\{\frac{\bar{y}_i + \frac{1}{2n_i}}{1 - \bar{y}_i + \frac{1}{2n_i}}\right\}, i = 1, \ldots, \ell.$$

See Appendix C.

Second, obtain the first derivative of log-likelihood of $\underset{\sim}{\beta}_{(0)}$, use a first-order Taylor's expansion with μ_i replaced by z_i, and set to zero to get

$$\underset{\sim}{\beta}^*_{(0)} = [\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \underset{\sim}{x}_{ij}\underset{\sim}{x}'_{ij}]^{-1}[\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \underset{\sim}{x}_{ij}(y_{ij} - z_i)].$$

Third, we obtain quasi-modes for the μ_i, a refinement of the z_i. We use the log-likelihood of the μ_i with the regression coefficients replaced by $\underset{\sim}{\beta}^*_{(0)}$, and solve its first derivative for zeros using a first-order Taylor's series expansion to get

$$\mu_i^* = \log[\frac{\frac{1}{n_i}\sum_{j=1}^{n_i} e^{-\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}^*_{(0)}}}{(1 - \bar{y}_i + \frac{1}{2n_i})}], i = 1, \ldots, \ell.$$

Let $\underset{\sim}{\tau}^{*'} = (\underset{\sim}{\mu}^{*'}, \underset{\sim}{\beta}^{*'}_{(0)})$.

Next, we evaluate $\underset{\sim}{g}$ and H at the quasi-modes $\underset{\sim}{\tau} = \underset{\sim}{\tau}^*$,

$$\underset{\sim}{g} = \left(\begin{array}{cccc} \frac{\partial\Delta}{\partial\mu_1} & \cdots & \frac{\partial\Delta}{\partial\mu_\ell} & \frac{\partial\Delta}{\partial\underset{\sim}{\beta}_{(0)}} \end{array}\right)^T_{\underset{\sim}{\mu}=\underset{\sim}{\mu}^*, \underset{\sim}{\beta}_{(0)}=\underset{\sim}{\beta}^*_{(0)}},$$

$$H = \left(\begin{array}{cccc} \frac{\partial^2\Delta}{\partial\mu_1^2} & \cdots & 0 & \frac{\partial^2\Delta}{\partial\mu_1\partial\underset{\sim}{\beta}_{(0)}} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{\partial^2\Delta}{\partial\mu_\ell^2} & \frac{\partial^2\Delta}{\partial\mu_\ell\partial\underset{\sim}{\beta}_{(0)}} \\ \frac{\partial^2\Delta}{\partial\mu_1\partial\underset{\sim}{\beta}_{(0)}} & \cdots & \frac{\partial^2\Delta}{\partial\mu_\ell\partial\underset{\sim}{\beta}_{(0)}} & \frac{\partial^2\Delta}{\partial\underset{\sim}{\beta}^2_{(0)}} \end{array}\right)_{\underset{\sim}{\mu}=\underset{\sim}{\mu}^*, \underset{\sim}{\beta}_{(0)}=\underset{\sim}{\beta}^*_{(0)}}.$$

The partial derivatives can be expressed in terms of response y_{ij} and covariates $\underset{\sim}{x}_{ij}$ as

$$\frac{\partial \Delta}{\partial \underset{\sim}{\beta}_{(0)}} = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left(\underset{\sim}{x}_{ij} y_{ij} - \frac{\underset{\sim}{x}_{ij} e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}{1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}} \right)$$

$$\frac{\partial \Delta}{\partial \mu_i} = \sum_{j=1}^{n_i} \left(y_{ij} - \frac{e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}{1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}} \right), i = 1,\ldots,\ell,$$

$$\frac{\partial^2 \Delta}{\partial \underset{\sim}{\beta}_{(0)}^2} = -\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \frac{\underset{\sim}{x}_{ij}\underset{\sim}{x}'_{ij} e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}{(1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i})^2},$$

$$\frac{\partial^2 \Delta}{\partial \mu_i^2} = -\sum_{j=1}^{n_i} \frac{e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}{(1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i})^2}, i = 1,\ldots,\ell,$$

$$\frac{\partial^2 \Delta}{\partial \mu_i \partial v_{i'}} = 0, i \neq i' = 1,\ldots,\ell, \quad \frac{\partial^2 \Delta}{\partial \mu_i \partial \underset{\sim}{\beta}_{(0)}} = -\sum_{j=1}^{n_i} \frac{\underset{\sim}{x}_{ij} e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}{(1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i})^2}, i = 1,\ldots,\ell.$$

For the convenience of computation, denote $\underset{\sim}{g} = \begin{pmatrix} \underset{\sim}{g}_1 \\ g_2 \end{pmatrix}$ and $H = -\begin{pmatrix} D & C' \\ C & B \end{pmatrix}$, where

$$\underset{\sim}{g}_1 = \begin{pmatrix} \frac{\partial \Delta}{\partial \mu_1} & \cdots & \frac{\partial \Delta}{\partial \mu_\ell} \end{pmatrix}^T, g_2 = \frac{\partial \Delta}{\partial \underset{\sim}{\beta}_{(0)}},$$

$$B = -\frac{\partial^2 \Delta}{\partial \underset{\sim}{\beta}_{(0)}^2}, C = -\begin{pmatrix} \frac{\partial^2 \Delta}{\partial \mu_1 \partial \underset{\sim}{\beta}_{(0)}} & \cdots & \frac{\partial^2 \Delta}{\partial \mu_\ell \partial \underset{\sim}{\beta}_{(0)}} \end{pmatrix}, D = -\begin{pmatrix} \frac{\partial^2 \Delta}{\partial \mu_1^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{\partial^2 \Delta}{\partial \mu_\ell^2} \end{pmatrix}.$$

Note that D is a diagonal matrix.

$$\text{Let } -H^{-1} = \begin{pmatrix} D & C' \\ C & B \end{pmatrix}^{-1} = \begin{pmatrix} E & F' \\ F & G \end{pmatrix}, \text{ where}$$

$$E = D^{-1} + D^{-1}C'(B - CD^{-1}C')^{-1}CD^{-1}, F = -(B - CD^{-1}C')^{-1}CD^{-1}, G = (B - CD^{-1}C')^{-1}.$$

Lemma 2.2.   Assuming that the design matrix is full-rank and $0 < \sum_{j=1}^{n_i} y_{ij} < n_i, i = 1,\ldots,\ell$, the posterior density, $\underset{\sim}{\tau}|\underset{\sim}{y}$ in (5), is logconcave.

Proof. If $0 < \sum_{j=1}^{n_i} y_{ij} < n_i, i = 1,\ldots,\ell$, there are solutions to the gradient vector set to zero.

Let $p_{ij} = \frac{e^{x'\underset{\sim}{\beta}_{(0)}+\mu_i}}{1 + e^{x'\underset{\sim}{\beta}_{(0)}+\mu_i}}, j = 1,\ldots,n_i, i = 1,\ldots,\ell$. Then, A, B and C of the negative Hessian matrix can be written as,

$$B = \frac{\partial^2 \Delta}{\partial \underset{\sim}{\beta}_{(0)}^2} = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} p_{ij}(1 - p_{ij})\underset{\sim}{x}_{ij}\underset{\sim}{x}'_{ij},$$

$$D = \text{diagonal}(d_i, i = 1, \ldots, \ell), d_i = \frac{\partial^2 \Delta}{\partial \mu_i^2} = \sum_{j=1}^{n_i} p_{ij}(1 - p_{ij}), i = 1, \ldots, \ell$$

$$C = (\underset{\sim}{c}_i), \underset{\sim}{c}_i = \frac{\partial^2 \Delta}{\partial \mu_i \partial \underset{\sim}{\beta}_{(0)}} = \sum_{j=1}^{n_i} p_{ij}(1 - p_{ij})\underset{\sim}{x}_{ij}, i = 1, \ldots, \ell.$$

Clearly, D is positive definite. Thus, to show that $-H$ is positive definite, we show that its Schur complement of D, $S = B - CD^{-1}C'$, is positive definite (e.g., see Boyd and Vandenberghe 2004). Let $\omega_{ij} = p_{ij}(1 - p_{ij})/\sum_{j=1}^{n_i}\{p_{ij}(1 - p_{ij})\}, j = 1, \ldots, n_i, i = 1, \ldots, \ell$. Then, the Schur complement is

$$S = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} p_{ij}(1 - p_{ij}) \sum_{j=1}^{n_i} \omega_{ij}\underset{\sim}{x}_{ij}\underset{\sim}{x}'_{ij} - \sum_i^{\ell} \sum_j^{n_i} p_{ij}(1 - p_{ij}) \sum_{j=1}^{n_i} \omega_{ij}\underset{\sim}{x}_{ij} \sum_{j=1}^{n_i} \sum_{j=1}^{n_i} \omega_{ij}\underset{\sim}{x}'_{ij}.$$

It is now easy to show that

$$S = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \omega_{ij}(\underset{\sim}{x}_{ij} - \sum_{j=1}^{n_i} \omega_{ij}\underset{\sim}{x}_{ij})(\underset{\sim}{x}_{ij} - \sum_{j=1}^{n_i} \omega_{ij}\underset{\sim}{x}_{ij})'.$$

Therefore, $-H$ is positive definite, and $\underset{\sim}{\tau}|\underset{\sim}{y}$ is logconcave. $\qquad\square$

We next establish the first key result presented in Theorem 2.1.

Theorem 2.1. Assuming that the design matrix is full-rank and $0 < \sum_{j=1}^{n_i} y_{ij} < n_i$, the posterior density, $\underset{\sim}{\tau}|\underset{\sim}{y}$ in (5) is approximately a multivariate normal density, and

$$\underset{\sim}{\mu}|\underset{\sim}{\beta}_{(0)}, \underset{\sim}{y} \sim \text{Normal}\{\underset{\sim}{\mu}_\mu - D^{-1}C'(\underset{\sim}{\beta}_{(0)} - \underset{\sim}{\mu}_\beta), D^{-1}\} \text{ and } \underset{\sim}{\beta}_{(0)}|\underset{\sim}{y} \sim \text{Normal}\{\underset{\sim}{\mu}_\beta, G\},$$

where

$$\underset{\sim}{\mu}_\mu = \underset{\sim}{\mu}^* + Eg_1 + F'g_2 \text{ and } \underset{\sim}{\mu}_\beta = \underset{\sim}{\beta}^*_{(0)} + Fg_1 + Gg_2.$$

Proof. By Lemma 2.2 the posterior density is logconcave. By Lemma 2.1 the posterior density is approximately a multivariate normal density. We provide the approximate mean and variance to completely specify the multivariate normal density.

By Lemma 2.1, evaluating all appropriate quantities at $\underset{\sim}{\tau}^*$, the posterior mean is

$$\begin{pmatrix} \underset{\sim}{\mu}_\mu \\ \underset{\sim}{\mu}_\beta \end{pmatrix} = \underset{\sim}{\tau}^* - H^{-1}\underset{\sim}{g} = \begin{pmatrix} \underset{\sim}{\mu}^* \\ \underset{\sim}{\beta}^*_{(0)} \end{pmatrix} + \begin{pmatrix} E & F' \\ F & G \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 \end{pmatrix} = \begin{pmatrix} \underset{\sim}{\mu}^* + Eg_1 + F'g_2 \\ \underset{\sim}{\beta}^*_{(0)} + Fg_1 + Gg_2 \end{pmatrix}.$$

Also, by Lemma 1, the posterior variance is

$$-H^{-1} = \begin{pmatrix} D & C' \\ C & B \end{pmatrix}^{-1} = \begin{pmatrix} E & F' \\ F & G \end{pmatrix}.$$

That is, the approximate joint posterior density is

$$\begin{pmatrix} \mu \\ \beta_{(0)} \end{pmatrix} | y \sim \text{Normal} \left\{ \begin{pmatrix} \mu_\mu \\ \mu_\beta \end{pmatrix}, \begin{pmatrix} E & F' \\ F & G \end{pmatrix} \right\}.$$

Finally, using a standard property of the multivariate normal density, it follows that approximately,

$$\mu | \beta_{(0)}, y \sim \text{Normal}\{\mu_\mu - D^{-1}C'(\beta_{(0)} - \mu_\beta), D^{-1}\} \text{ and } \beta_{(0)} | y \sim \text{Normal}\{\mu_\beta, G\}.$$

$\square$

2.2. Integrated Nested Normal Approximation

In this section, we obtain the integrated nested normal approximation (INNA). INNA, which does not require posterior modes, is competitive to the integrated nested Laplace approximation (INLA) that requires posterior modes.

Next, using the normal priors for the μ_i and Theorem 1, we have the following approximate hierarchical Bayesian regression model,

$$\mu | \beta_{(0)}, y \sim \text{Normal}\{\mu_\mu - D^{-1}C'(\beta_{(0)} - \mu_\beta), D^{-1}\}$$

$$\beta_{(0)} | y \sim \text{Normal}\{\mu_\beta, G\}$$

$$\mu | \beta_0, \delta^2 \sim \text{Normal}\{\beta_0 j, \delta^2 I\}$$

$$\pi(\beta_0, \beta_{(0)}, \delta^2) \propto \frac{1}{(1+\delta^2)^2}, \delta^2 > 0, \tag{6}$$

where j is a vector of ones.

Then, using Bayes' theorem again, the approximate joint posterior density for the parameters μ, β and δ^2 is

$$\pi_a(\mu, \beta, \delta^2 | y) \propto e^{-\frac{1}{2}\left\{ \left[\mu - \left(\mu_\mu - D^{-1}C'(\beta_{(0)} - \mu_\beta)\right)\right]' D \left[\mu - \left(\mu_\mu - D^{-1}C'(\beta_{(0)} - \mu_\beta)\right)\right] + \left[\mu - \beta_0 j\right]'(\delta^2 I)^{-1}\left[\mu - \beta_0 j\right] \right\}}$$

$$\times \frac{|D|^{1/2}}{|\delta^2 I|^{1/2}|G|^{1/2}} \times \frac{1}{(1+\delta^2)^2} \times e^{-\frac{1}{2}\left[\beta_{(0)} - \mu_\beta\right]' G^{-1}\left[\beta_{(0)} - \mu_\beta\right]}. \tag{7}$$

Next, we state the main result of the paper in Theorem 2.2.

Theorem 2.2. Using the multiplication rule, the joint posterior density, $\pi(\mu, \beta, \delta^2 | y)$ in (7), can be approximated by

$$\pi_a(\mu, \beta, \delta^2 | y) = \pi_a(\mu | \beta, \delta^2, y)\pi_a(\beta | \delta^2, y)\pi_a(\delta^2 | y),$$

where the three densities on the right-hand side are to be determined.

Proof. First, it can be shown that

$$\underset{\sim}{\mu}|\underset{\sim}{\beta},\delta^2,y \sim \text{Normal}\left\{(D+\frac{1}{\delta^2}I)^{-1}(D\underset{\sim}{\mu}_\mu - C'(\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta)+\frac{1}{\delta^2}\beta_0\underset{\sim}{j}),(D+\frac{1}{\delta^2}I)^{-1}\right\}, \quad (8)$$

which is $\pi_a(\underset{\sim}{\mu} \mid \underset{\sim}{\beta},\delta^2,y)$. Because $(D+\frac{1}{\delta^2}I)$ is diagonal, given $\underset{\sim}{\beta},\delta^2,y$, the μ_i are independent. This is an important result because parallel computation can be done for μ_i, which accommodates time-consuming and massive storage challenges in big data analysis. This result holds for the exact conditional posterior density of the μ_i.

Second, because $\underset{\sim}{\mu}$ has a multivariate normal distribution, we can integrate out $\underset{\sim}{\mu}$ from the joint posterior density $\pi_a(\underset{\sim}{\mu},\underset{\sim}{\beta},\delta^2|y)$ to get the joint posterior density of $\underset{\sim}{\beta}$ and δ^2,

$$\pi_a(\underset{\sim}{\beta},\delta^2|\underset{\sim}{y}) \propto \frac{|D|^{1/2}}{|D+\frac{1}{\delta^2}I|^{1/2}|\delta^2I|^{1/2}|G|^{1/2}} \times \frac{1}{(1+\delta^2)^2} \times$$

$$e^{-\frac{1}{2}\left\{\left[\underset{\sim}{\mu}_\mu - D^{-1}C'(\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta)-\beta_0\underset{\sim}{j}\right]'(D^{-1}+\delta^2I)^{-1}\left[\underset{\sim}{\mu}_\mu - D^{-1}C'(\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta)-\beta_0\underset{\sim}{j}\right]+\left[\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta\right]'G^{-1}\left[\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta\right]\right\}} \quad (9)$$

and

$$\pi(\underset{\sim}{\beta}|\delta^2,\underset{\sim}{y}) \propto$$

$$e^{-\frac{1}{2}\left\{\left[\underset{\sim}{\mu}_\mu - D^{-1}C'(\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta)-\beta_0\underset{\sim}{j}\right]'(D^{-1}+\delta^2I)^{-1}\left[\underset{\sim}{\mu}_\mu - D^{-1}C'(\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta)-\beta_0\underset{\sim}{j}\right]+\left[\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta\right]'G^{-1}\left[\underset{\sim}{\beta}_{(0)}-\underset{\sim}{\mu}_\beta\right]\right\}}.$$
$$(10)$$

Let

$$\Delta_{(0)} = CD^{-1}(D^{-1}+\delta^2I)^{-1}D^{-1}C' + G^{-1},$$

$$\delta_0^2 = \underset{\sim}{j}'(D^{-1}+\delta^2I)^{-1}\underset{\sim}{j},$$

$$\underset{\sim}{\gamma} = CD^{-1}(D^{-1}+\delta^2I)^{-1}\underset{\sim}{j},$$

$$\begin{pmatrix} \omega_0 \\ \underset{\sim}{\omega}_{(0)} \end{pmatrix} = \begin{pmatrix} \delta_0^2 & \underset{\sim}{\gamma}' \\ \underset{\sim}{\gamma} & \Delta_{(0)} \end{pmatrix}^{-1} \begin{pmatrix} (\underset{\sim}{\mu}_\mu + D^{-1}C'\underset{\sim}{\mu}_\beta)'(D^{-1}+\delta^2I)^{-1}\underset{\sim}{j} \\ (\underset{\sim}{\mu}_\mu + D^{-1}C'\underset{\sim}{\mu}_\beta)'(D^{-1}+\delta^2I)D^{-1}C' + \underset{\sim}{\mu}'_\beta G^{-1} \end{pmatrix}.$$

After extensive algebraic manipulation, we can show that $\underset{\sim}{\beta}|\delta^2,y$ has an approximate multivariate normal density,

$$\begin{pmatrix} \beta_0 \\ \underset{\sim}{\beta}_{(0)} \end{pmatrix}|\delta^2,\underset{\sim}{y} \sim \text{Normal}\left\{\begin{pmatrix} \omega_0 \\ \underset{\sim}{\omega}_{(0)} \end{pmatrix}, \begin{pmatrix} \delta_0^2 & \underset{\sim}{\gamma}' \\ \underset{\sim}{\gamma} & \Delta_{(0)} \end{pmatrix}^{-1}\right\}, \quad (11)$$

which is denoted by $\pi_a(\underset{\sim}{\beta} \mid \delta^2,y)$.

Third, because the approximate conditional distribution of $\underset{\sim}{\beta}|\delta^2,y$ is a multivariate normal distribution, we can integrate out $\underset{\sim}{\beta}$ from the joint density of $\pi_a(\underset{\sim}{\beta},\delta^2 \mid y)$ in (9) to get the posterior density,

$$\pi_a(\delta^2|\underset{\sim}{y}) \propto \frac{1}{|\delta^2D+I|^{1/2}} \left|\begin{pmatrix} \delta_0^2 & \underset{\sim}{\gamma}' \\ \underset{\sim}{\gamma} & \Delta_{(0)} \end{pmatrix}\right|^{-\frac{1}{2}} \times \frac{1}{(1+\delta^2)^2}$$

$$\times e^{-\frac{1}{2}\left[(\mu_\mu+D^{-1}C'\mu_\beta)'(D^{-1}+\delta^2I)^{-1}(\mu_\mu+D^{-1}C'\mu_\beta)+\mu'_\beta G^{-1}\mu_\beta-w'\left(\left(\begin{array}{cc}\delta_0^2 & \gamma' \\ \gamma & \Delta_{(0)}\end{array}\right)\right)w\right]},$$

where $w' = (\omega_0, w'_{(0)})$.

$\square$

Our INNA method is sampling based and it uses random samples. Samples are obtained by first drawing from $\pi_a(\delta^2|y)$, then $\pi_a(\beta \mid \delta^2, y)$ and finally $\pi_a(\mu \mid \beta, \delta^2, y)$, the first two densities use standard draws, but the third is a nonstandard density.

Since $0 < \delta^2 < \infty$, we make a transformation to $\eta = \frac{1}{1+\delta^2}$ so that $0 < \eta < 1$. Then, posterior density, $\pi(\eta|y)$, is

$$\pi(\eta|y) \propto \left\{\frac{1}{|\delta^2D+I|^{1/2}}\left|\left(\begin{array}{cc}\delta_0^2 & \gamma' \\ \gamma & \Delta_{(0)}\end{array}\right)\right|^{-\frac{1}{2}}\right\}_{\delta^2=\frac{1-\eta}{\eta}} \times$$

$$\left\{e^{-\frac{1}{2}\left[(\mu_\mu+D^{-1}C'\mu_\beta)'(D^{-1}+\delta^2I)^{-1}(\mu_\mu+D^{-1}C'\mu_\beta)+\mu'_\beta G^{-1}\mu_\beta-w'\left(\left(\begin{array}{cc}\delta_0^2 & \gamma' \\ \gamma & \Delta_{(0)}\end{array}\right)\right)w\right]}\right\}_{\delta^2=\frac{1-\eta}{\eta}}. \quad (12)$$

The grid method is used to sample η.

INNA simply uses the multiplication rule to get samples from the approximate joint posterior density, which is

$$\pi_a(\mu, \beta, \delta^2 \mid y) = \pi_a(\mu, \mid \beta, \delta^2, y)\pi_a(\beta \mid \delta^2, y)\pi_a(\delta^2 \mid y), \quad (13)$$

where $\pi_a(\mu, \mid \beta, \delta^2, y)$ is given in (4), $\pi_a(\beta, \delta^2 \mid y)$ is given in (11) and $\pi_a(\delta^2 \mid y)$ is given in (12) after re-transformation. Equation (13) is the basis of our INNA approximation. This is simply the multiplication rule of probability; simply draw η from (12) and retransform to δ^2 to get $\pi_a(\delta^2 \mid y)$, draw β from $\pi_a(\beta \mid \delta^2, y)$ (multivariate normal and μ from $\pi_a(\mu, \mid \beta, \delta^2, y)$. Use a Metropolis algorithm with an approximate normal proposal density to draw samples of v_i independently given β, δ^2 and data. We run each Metropolis 100 times and picked the last one. If the Metropolis step fails (jumping rate is not in $(.25, .50)$, we use a grid method instead. Parallel computing can also be used in this latter step. This is performed in the same manner for the exact method. For our application with 3912 households, this latter step runs very fast. Of course, with much larger number of households, the computing time will be substantial, but now parallel computing is available.

2.3.   Comparison of the Two Methods

As a summary, we compare the approximate and exact methods. The exact method is given in Appendix A.

First, we note that the exact method actually uses the approximate method. We use a Metropolis step with $\pi_a(\beta, \delta^2|y)$, obtained from the approximate method. We fit a multivariate Student's t distribution with $\tilde{\eta}$ degrees of freedom to the iterates, $(\beta, \log(\delta^2))$ from the approximate method as the proposal density in the Metropolis step. It is standard to tune the Metropolis step by varying η.

We present two differences between the two methods. First, both methods are sampling based; the approximate method implements random samples and the exact method a Markov chain. Second, $\pi_a(\beta, \delta^2 \mid y)$ is used for the approximate method and a Metropolis step is used for $\pi(\beta, \delta^2 \mid y)$. It is this Metropolis step that is time-consuming (20 minutes versus 20 seconds); for a million households, we can prorate this time (80 hours versus 80 minutes), enormous savings.

We present three similarities between the two methods. First, $\pi(v_i \mid \beta, \delta^2, y), i = 1, \ldots, \ell$, are drawn the same way using a Metropolis step with proposal density $\pi_a(v_i \mid \beta, \delta^2, y), i = 1, \ldots, \ell$. Grids are used when the Metropolis step fails. Parallel computing can be done easily in both cases. Second, for $\pi(v_i \mid \beta, \delta^2, y), i = \ell+1, \ldots, L$, are normal densities ($\ell = 3912, L = 60,262$). Third, for prediction two Bayesian bootstraps are used to get the nonsampled household sizes and the nonsampled covariates ($\approx$ two million people). This is done within wards.

3.   Illustrative Example

In Section 3.1, we briefly describe the Nepal Living Standards Survey (NLSS II) and in Section 3.2, we use the health status data with five covariates to compare our approximate Bayesian logistic regression method with the exact one.

3.1.   Nepal Living Standards Survey

We use data from the Nepal Living Standards Survey (NLSS II, Central Bureau of Statistics, 2003-2004) to illustrate INNA with logistic regression. NLSS is a national household survey in Nepal, actually population based (i.e., interviews are done for individual household members). NLSS follows the World Bank's Living Standards Measurement Survey methodology with a two-stage stratified sampling scheme, which has been successfully applied in many parts of the world. It is an integrated survey which covers samples from the whole country and runs throughout the year. The main objective of the NLSS is to collect data from Nepalese households and provide information to monitor progress in national living standards. The NLSS gathers information on a variety of areas. It has collected data on demographics, housing, education, health, fertility, employment, income, agricultural activity, consumption, and various other areas. NLSS has records for 20,263 individuals from 3,912 households from a 326 wards (or psus) from a population of 60262 households and about two million Nepalese. We choose the binary variable, health status, from the health section of the questionnaire. There are hundreds of variables and we have selected five of the most important ones that can explain health status. We use health status as the binary variable with a selection of five pertinent covariates.

Health status is covered in Section 8 of the questionnaire. This section collected information on chronic and acute illnesses, uses of medical facilities, expenditures on them and health

status. Health status questionnaire is asked for every individual that was covered in the survey. The health status questionnaire has four options. For our purpose we make it a binary variable, good health or poor health.

In the NLSS II, Nepal is divided into wards (psu's) and within each ward there are a number of households. The sampling design of the NLSS II used two-stage stratified sampling. A sample of psu's was selected using PPS sampling and then twelve households were systematically selected from each ward. Thus, households have equal probability of selection. But while individuals in a household has equal probability of selection, these probabilities will vary with the size of the households. That is, over the entire Nepal, each individual has unequal probability of selection.

We choose five relevant covariates which can influence health status from the same NLSS survey for the integrated nested normal approximation (INNA) logistic regression method. They are age, nativity, sex, area and religion. We created binary variables nativity (Indigenous = 1, Non-indigenous = 0), religion ((Hindu = 1, Non-Hindu = 0), sex (Male = 1, Female = 0) and area (Urban = 1, Rural = 0). We standardize age. Older age and child age are more vulnerable than younger age. Indigenous people can have different health status from migrated people. Similarly, health status of urban and rural citizens could be different. Interest is on the smoothed proportion of household members in good health (binary). The NLSS 2003-2004 sample consists 3,912 households (roughly 20 thousand people) from a population from 326 wards with 60,221 households. From Census 2001, the population of Nepal consists of 35,000 wards from 4 million households with roughly 20 million people. We will predict the finite population proportions of household members in good health based for the 60,221 households, not the entire Nepal. This can be done using the same methodology because Nepal consists of six strata.

We use a Bayesian hierarchical logistic regression model to predict all nonsampled households in the sampled wards. To obtain smoothed estimates for the sampled households, we also predict these. We use Bayesian bootstraps (Rubin 1981) for unknown household sizes and non-sampled covariates; the bootstrapping is done within wards. The 2001 Census can potentially provide these two pieces of information, but there is a mis-match between the households in the census and the NLSS (a record linkage can be performed). We note, however, that there is linkage between the wards, but this information is not useful to household estimates.

3.2.   Numerical Comparisons

We predict the household proportions of members in good health for 60,221 households. This analysis is based on 3912 sample households from 326 wards (PSUs). Our primary purpose is to compare the approximate method with the exact method when there are random effects at the household level. We want to show that one can safely use the approximate method to save computational time. Our secondary purpose is to compare the exact methods when there are random effects at the household level only and random effects at the ward level. We note that there are only 326 wards, considerably less than the number of households. In this case, a data analyst might try to save computational time by only using the random effects at the ward level, but this procedure is not sensible.

First, in Figures 1, 2 and 3, we compare respectively the posterior means (PMs), posterior standard deviations (PSDs) and posterior coefficient of variations (CVs) as our primary purpose. The PMs are very closed, followed by the PSDs and then the CVs. For the PMs, the points lie very closely on the 45° straight line through the origin. This is similar for the PSDs and CVs. However, the plot of the PSD is a bit thicker and the plot for the CVs has larger CVs more spread out. But overall, these approximations are quite acceptable to data analysts, scientists and engineers. We also show the posterior densities (PDs) of the hyperparameters. In Figure 7 we plot the PDs of δ^2 and β_0 and in Figure 8 we plot the remaining five βs. We can see that they are all unimodal, with the modes about the same, but, as expected, the spread is a bit larger for the exact method. But these differences are not alarming. However, these differences are much smaller when inference is made about the household proportions.

We also compare the approximate method versus the exact method when random effects are used only at the ward level. Again, the PMs are very good; see Figure 4. The PSDs are more spread out (Figure 5) and the plot of the CVs (Figure 6) is a similar to the household random effects. Again, the approximate method and the exact method are reasonably closed, of course, not as closed as for the household level random effects.

Finally, we compare the exact method at the ward level to the exact method at the household level. Figures 9, 10 and 11 show that the comparisons are rather poor; there are no clear indications of 45° straight line through the origin. We have examined this further in Tables 1, 2 and 3, where we cross-classified the PMs in $0 - .2, .2 - .4, .4 - .8, .8 - 1$, the PSDs in $.0 - .1$, $.1 - .2, .2 - .3, .3 - .4, .4 - .5$, and the CVs in $.02 - .05, .05 - .10, .10 - .25, .25 - .50, > .50$. For all three tables a majority of the points fall along the north-west to south-east diagonal, but there are many points off the diagonals. Thus, one must have random effects at the household level.

We conclude that the approximation at the household level is reasonable. The approximation is desirable because one can perform the computations in real time. One should not use random effects only at the ward level to cut corners. This can be misleading.

4. Concluding Remarks

We make three statistical comments. First, the approximate method is necessary when there are a large number (millions) of households (clusters or areas). Second, it is difficult to use the census data effectively but it is desirable (matching problem). Third, it is possible to obtain similar approximations for spatial priors and Dirichlet process priors (under investigation)

We make four computational comments. First, parallel computing is needed for big data (numerous small areas). Second, consensus Monte Carlo method, although problematic, is needed for large data (storage problems). For the NLSS survey, stratification already exists; Nepal has six strata and our INNA procedure can be applied in each stratum in parallel. Third, MCMC (not INLA) methods are useful for small datasets; good approximations are needed for large datasets (big data). Finally, INNA is potentially useful because modes are not required (INLA needs modes).

We mention two possible extensions. The first extension is about survey weights. The sec-

ond extension is to a sub-area model.

PPS sampling is used in the first-stage of the survey design. Thus, there are survey weights (design, not adjusted weights). All households (each member) in a psu has the same weight. So we can proceed in one of the two ways in our analysis. First, we can use an adjusted logistic likelihood incorporating the survey weights. We discuss a single area, then we show how to extend it to small area problems. Let $\omega_i, i = 1, \ldots, n$, denote the survey weights for sampling from a single area. Let

$$n \geq n_e = (\sum_{i=1}^{n} \omega_i)^2 / (\sum_{i=1}^{n} \omega_i^2), \quad \tilde{\omega}_i = n_e \frac{\omega_i}{\sum_{i=1}^{n} \omega_i}, i = 1, \ldots, n;$$

see Potthof, Woodbury and Manton (1992) for pioneering work on equivalent sample sizes. For $(y_i, \tilde{\omega}_i, \underset{\sim}{x}_i), i = 1, \ldots, n$, we have

$$p(y_i \mid \underset{\sim}{\beta}) \propto \left\{ \frac{e^{y_i \underset{\sim}{x}_i' \underset{\sim}{\beta}}}{1 + e^{\underset{\sim}{x}_i' \underset{\sim}{\beta}}} \right\}^{\tilde{\omega}_i}, y_i = 0, 1, i = 1, \ldots, n,$$

$$p(y_i \mid \underset{\sim}{\beta}) = \frac{e^{y_i \tilde{\underset{\sim}{x}}_i' \underset{\sim}{\beta}}}{1 + e^{\tilde{\underset{\sim}{x}}_i' \underset{\sim}{\beta}}}, y_i = 0, 1,$$

$$\tilde{\underset{\sim}{x}}_i = \tilde{\omega}_i \underset{\sim}{x}_i, i = 1, \ldots, n.$$

For small areas, with $(y_{ij}, \tilde{\omega}_{ij}, \underset{\sim}{x}_{ij}), \ j = 1, \ldots, n_i, \ i = 1, \ldots, \ell$, we have

$$p(y_{ij} \mid \underset{\sim}{\beta}, v_i) = \frac{e^{y_{ij}(\tilde{\underset{\sim}{x}}_{ij}' \underset{\sim}{\beta} + v_i)}}{1 + e^{\tilde{\underset{\sim}{x}}_{ij}' \underset{\sim}{\beta} + v_i}}, \ y_{ij} = 0, 1,$$

$$\tilde{\underset{\sim}{x}}_{ij} = \tilde{\omega}_{ij} \underset{\sim}{x}_{ij}, \ j = 1, \ldots, n_i, \ i = 1, \ldots, \ell.$$

Our second extension is to sub-area model. One example already discussed in the literature is that of Torabi and Rao (2014), who extended the Fay-Herriot model (Fay and Herriot 1979), not for logistic regression. For our problem, the areas are the wards and sub-areas are the households. Let $i = 1, \ldots, \ell$, denote the areas and $j = 1, \ldots, n_i$, denote the sub-areas. We assume that

$$y_{ijk} \mid \underset{\sim}{\beta}, v_i, \mu_{ij} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{e^{\underset{\sim}{x}_{ijk}' \underset{\sim}{\beta} + v_i + \mu_{ij}}}{1 + e^{\underset{\sim}{x}_{ijk}' \underset{\sim}{\beta} + v_i + \mu_{ij}}} \right\}, k = 1, \ldots, m_{ij},$$

$$\mu_{ij} \mid \sigma^2 \overset{ind}{\sim} \text{Normal}(0, \sigma^2), j = 1, \ldots, n_i,$$

$$v_i \mid \delta^2 \overset{ind}{\sim} \text{Normal}(0, \delta^2), i = 1, \ldots, \ell,$$

$$\pi(\underset{\sim}{\beta}, \delta^2, \sigma^2) \propto \frac{1}{(1 + \sigma^2)^2} \frac{1}{(1 + \delta^2)^2}, \sigma^2 > 0, \delta^2 > 0.$$

For logistic regression, this research is currently in progress.

Appendix A   Exact Method for Logistic Regression

Recall the Bayesian logistic model with covariates that we worked on with INNA method

$$y_{ij}|\mu_i,\underset{\sim}{\beta}_{(0)} \overset{ind}{\sim} \text{Bernoulli} \left\{ \frac{e^{\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}}{1+e^{\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}} \right\},$$

$$\mu_i|\beta_0,\delta^2 \overset{iid}{\sim} \text{Normal}(\beta_0,\delta^2),$$

$$\pi(\underset{\sim}{\beta},\delta^2) \propto \frac{1}{(1+\delta^2)^2}, \delta^2 > 0, i=1,...,\ell, j=1,...,n_i. \tag{A.1}$$

According to Bayes' theorem, the joint posterior density of the parameters $(\underset{\sim}{\mu},\underset{\sim}{\beta},\delta^2|\underset{\sim}{y})$ is

$$\pi(\underset{\sim}{\mu},\underset{\sim}{\beta},\delta^2|\underset{\sim}{y}) \propto \pi(\underset{\sim}{y}|\underset{\sim}{\mu},\underset{\sim}{\beta}_{(0)}) \times \pi(\underset{\sim}{\mu}|\beta_0,\delta^2) \times \pi(\underset{\sim}{\beta},\delta^2)$$

$$\propto \prod_{i=1}^{\ell} \left\{ \left[\prod_{j=1}^{n_i} \frac{e^{(\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i)y_{ij}}}{1+e^{\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}} \right] \left[\frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(\mu_i-\beta_0)^2}{2\delta^2}} \right] \right\} \frac{1}{(1+\delta^2)^2}.$$

Theorem A.1. The joint posterior density, $\pi(\underset{\sim}{\mu},\underset{\sim}{\beta},\delta^2|\underset{\sim}{y})$, is proper provided that the design matrix is full rank and $0 < \sum_{j=1}^{n_i} y_{ij} < n_i, i=1,\ldots,\ell$.

Proof. With a flat prior on the μ_i and $\underset{\sim}{\beta}$, the same argument as in Lemma 2.2 gives logconcavity of the joint posterior density. Putting a logconcave prior on the μ_i does not change the logconavity of $\pi(\underset{\sim}{\beta},\underset{\sim}{\mu} \mid \delta^2,y)$ because the product of two logconcave densities is another logconcave density. In addition, logconcave densities have sub-exponential tails and their moment generating functions exist (see Dharmadhikari and Joag-Dev 1988). That is, $\int \pi(\underset{\sim}{\mu},\underset{\sim}{\beta} \mid \delta^2,y)d\underset{\sim}{\beta}d\underset{\sim}{\mu} = a(\delta^2)$ finite for all δ^2. Therefore, $\int a(\delta^2)\pi(\delta^2)d\delta^2 < \infty$ as long as $\pi(\tilde{\delta^2})$ is proper as for $\pi(\delta^2) = 1/(1+\delta^2)^2$. So that, $\pi(\underset{\sim}{\mu},\underset{\sim}{\beta},\delta^2|\underset{\sim}{y})$ is proper. $\square$

The standard MCMC logistic regression exact method is complicated to work with and it takes longer time to get posterior samples. We apply Metropolis Hastings sampler to draw samples for parameters $\underset{\sim}{\beta}$, δ^2 and $\underset{\sim}{\mu}$.

The idea of exact method is to get full conditional posterior distributions for all of the parameters in the model, and then get a large number of independent samples of each parameter with its full conditional posterior density.

First, we integrate $\underset{\sim}{\mu}$ from the posterior density to get the joint posterior density of $\underset{\sim}{\beta},\delta^2|\underset{\sim}{y}$ as

$$\pi(\underset{\sim}{\beta},\delta^2|\underset{\sim}{y}) \propto \int_{\Omega} \prod_{i=1}^{\ell} \left\{ \prod_{j=1}^{n_i} \frac{e^{(\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i)y_{ij}}}{1+e^{\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}} \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(\mu_i-\beta_0)^2}{2\delta^2}} \right\} \frac{1}{(1+\delta^2)^2} d\underset{\sim}{\mu}$$

$$= \frac{1}{(1+\delta^2)^2} \prod_{i=1}^{\ell} \left\{ \int_{-\infty}^{\infty} \frac{e^{\sum_{j=1}^{n_i}(\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i)y_{ij}}}{\prod_{j=1}^{n_i}\left[1+e^{\underset{\sim}{x}'_{ij}\underset{\sim}{\beta}_{(0)}+\mu_i}\right]} \frac{1}{\sqrt{2\pi\delta^2}} e^{-\frac{(\mu_i-\beta_0)^2}{2\delta^2}} d\mu_i \right\}.$$

Notice that this is not a simple distribution function for the integration, so we apply numerical integration. Divide the integration domain to m equal intervals $[t_{k-1}, t_k], k = 1, ..., m$. Let $z_i = \frac{\mu_i - \beta_0}{\delta}$ with standard normal distribution. We get an approximate density (very accurate though),

$$\pi(\underset{\sim}{\beta}, \delta^2 | \underset{\sim}{y}) \propto \frac{1}{(1+\delta^2)^2} \left(\frac{1}{\sqrt{\delta^2}}\right)^\ell \prod_{i=1}^\ell \left\{ \sum_{k=1}^m \int_{t_{k-1}}^{t_k} \frac{e^{\sum_{j=1}^{n_i} (x'_{ij}\underset{\sim}{\beta}_{(0)} + \mu_i)y_{ij}}}{\prod_{j=1}^{n_i} \left[1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)} + \mu_i}\right]} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu_i - \beta_0)^2}{2\delta^2}} d\mu_i \right\}$$

$$= \frac{1}{(1+\delta^2)^2} \prod_{i=1}^\ell \left\{ \sum_{k=1}^m \int_{t_{k-1}}^{t_k} \frac{e^{\sum_{j=1}^{n_i} (x'_{ij}\underset{\sim}{\beta}_{(0)} + \beta_0 + z_i\delta)y_{ij}}}{\prod_{j=1}^{n_i} \left[1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)} + \beta_0 + z_i\delta}\right]} \frac{1}{\sqrt{2\pi}} e^{-\frac{z_i^2}{2}} dz_i \right\}.$$

Take the middle point of each interval $[t_{k-1}, t_k]$ to estimate the cumulative density function, and denote $\hat{z}_k = \frac{t_k + t_{k-1}}{2}$. We have the following deduction

$$\pi(\underset{\sim}{\beta}, \delta^2 | \underset{\sim}{y}) \approx \frac{1}{(1+\delta^2)^2} \prod_{i=1}^\ell \left\{ \sum_{k=1}^m \frac{e^{\sum_{j=1}^{n_i} (x'_{ij}\underset{\sim}{\beta}_{(0)} + \beta_0 + \hat{z}_k\delta)y_{ij}}}{\prod_{j=1}^{n_i} \left[1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)} + \beta_0 + \hat{z}_k\delta}\right]} \int_{t_{k-1}}^{t_k} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \right\}.$$

The integration is now over a standard normal distribution. We consider the interval (-3, 3) for numerical integration, since this domain (standard normal) covers 99.74% of the distribution that we are dealing with. We take m=100 grid points. Then the joint posterior density for $\underset{\sim}{\beta}$ and δ^2 can be expressed as

$$\pi(\underset{\sim}{\beta}, \delta^2 | \underset{\sim}{y}) \approx \frac{1}{(1+\delta^2)^2} \prod_{i=1}^\ell \left\{ \sum_{k=1}^m \frac{e^{\sum_{j=1}^{n_i} (x'_{ij}\underset{\sim}{\beta}_{(0)} + \beta_0 + \hat{z}_k\delta)y_{ij}}}{\prod_{j=1}^{n_i} \left[1 + e^{x'_{ij}\underset{\sim}{\beta}_{(0)} + \beta_0 + \hat{z}_k\delta}\right]} (\Phi(t_k) - \Phi(t_{k-1})) \right\}. \qquad \text{(A.2)}$$

We use the Metropolis sampler to draw samples from the joint posterior density of $\pi(\underset{\sim}{\beta}, \delta^2 | \underset{\sim}{y})$ drawing $\underset{\sim}{\beta}$ and δ^2 simultaneously. We use the transformation $\log(\delta^2) = \beta_{p+1}$. We obtain a proposal density for $\underset{\sim}{\beta}$ using a multivariate Student's t density as follows. First, when we run the approximate method, we obtain the approximate mean, $\hat{\beta}$ and covariance matrix $\hat{\Sigma}$ of $\underset{\sim}{\beta}$. So we take $\underset{\sim}{\beta} \sim \text{Normal}(\hat{\underset{\sim}{\beta}}, \sigma^2\hat{\Sigma}), \eta/\sigma^2 \sim \text{Gamma}(\eta/2, 1/2)$. Tuning of the Metropolis sampler is obtained by varying η (e.g., $\eta = 8$ corresponds to approximately a logistic random vector). We run the Metropolis sampler in a nonstandard manner, as the Markov chain runs, we reserve those iterates when the algorithm moves. In this way while tuning, is required to get a reasonable jumping rate, no other diagnostics (autocorrelation, effective sample size, test of stationarity, thinning) are needed.

The number of nonsampled households in the sampled PSUs are known, but the number of members of a nonsampled household and their covariates are unknown. For each sampled PSU, we obtain the empirical distribution of the number of members and their covariates. Then, we

use two Bayesian bootstraps to sample the number of members and their covariates from the pool of the PSU. For both bootstraps we use 1000 samples.

Appendix B   Quasi-Modes for Logistic Regression

Now we have to specify $\beta^*_{\sim(0)}$, $\mu^*_{\sim}$, g and H. Consider the log likelihood function

$$f(\underset{\sim}{\tau}) = \log h(\underset{\sim}{\tau}) = \log \left\{ \prod_{i=1}^{\ell} \prod_{j=1}^{n_i} \frac{e^{(x'_{ij}\beta_{\sim(0)}+\mu_i)y_{ij}}}{1+e^{x'_{ij}\beta_{\sim(0)}+\mu_i}} \right\}$$

$$= \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ (x'_{ij}\underset{\sim}{\beta}_{(0)} + \mu_i)y_{ij} - \log(1+e^{x'_{ij}\beta_{\sim(0)}+\mu_i}) \right\}. \tag{B.1}$$

As an estimate of μ_i, we use the empirical logistic transform z_i,

$$\hat{\mu}^*_i = z_i = \log \left\{ \frac{\bar{y}_i + \frac{1}{2n_i}}{1 - \bar{y}_i + \frac{1}{2n_i}} \right\}.$$

See Appendix C. Plug $\hat{\mu}^*_i$ in the log likelihood function (B.1) and consider it as a function of $\underset{\sim}{\beta}_{(0)}$ only,

$$g(\underset{\sim}{\beta}_{(0)}) = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ (x'_{ij}\underset{\sim}{\beta}_{(0)} + \hat{\mu}^*_i)y_{ij} - \log(1+e^{x'_{ij}\beta_{\sim(0)}+\hat{\mu}^*_i}) \right\}.$$

The first derivative function of $g(\underset{\sim}{\beta}_{(0)})$ over $\underset{\sim}{\beta}_{(0)}$ is

$$g'(\underset{\sim}{\beta}_{(0)}) = \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ \underset{\sim}{x}_{ij}y_{ij} - \frac{\underset{\sim}{x}_{ij}e^{(x'_{ij}\beta_{\sim(0)}+\hat{\mu}^*_i)}}{1+e^{x'_{ij}\beta_{\sim(0)}+\hat{\mu}^*_i}} \right\}$$

$$= \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ \underset{\sim}{x}_{ij}y_{ij} - \underset{\sim}{x}_{ij}(1+e^{-(x'\beta_{\sim(0)}+\hat{\mu}^*_i)})^{-1} \right\}. \tag{B.2}$$

Typically, we can solve the equation $g'(\underset{\sim}{\beta}_{(0)}) = 0$ for the mode as the maximum likelihood estimator (MLE), but here it is not easy to solve the equation because $g'(\underset{\sim}{\beta}_{(0)})$ is complex. We use first order Taylor series approximation to simplify the above function. Since the first order Taylor expansion of $(1+e^{-(x'_{ij}\beta_{\sim(0)}+\hat{\mu}^*_i)})^{-1}$ equals $(1-e^{-(x'_{ij}\beta_{\sim(0)}+\hat{\mu}^*_i)})$, (B.2) equals to

$$\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ \underset{\sim}{x}_{ij}y_{ij} - \underset{\sim}{x}_{ij}(1-e^{-(x'_{ij}\beta_{\sim(0)}+\hat{\mu}^*_i)}) \right\}. \tag{B.3}$$

This is still complex. We apply Taylor series again and get expansion of the term $e^{-(x'\beta_{\sim(0)}+\hat{\mu}^*_i)}$ to the first order as $(1-(x'\underset{\sim}{\beta}_{(0)}+\hat{\mu}^*_i))$. Thus (B.3) approximately equals

$$\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ \underset{\sim}{x}_{ij}y_{ij} - \underset{\sim}{x}_{ij}(1-(1-(x'_{ij}\underset{\sim}{\beta}_{(0)}+\hat{\mu}^*_i))) \right\}$$

$$= \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \left\{ \underset{\sim}{x}_{ij}(y_{ij} - \hat{\mu}_i^*) - \underset{\sim}{x}_{ij}(x'_{ij}\underset{\sim}{\beta}_{(0)}) \right\}. \tag{B.4}$$

(B.4) is easy to solve. Solve for $g'(\underset{\sim}{\beta}_{(0)}) = 0$, and we can get the approximate posterior mode of $\underset{\sim}{\beta}_{(0)}$

$$\underset{\sim}{\beta}^*_{(0)} = [\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \underset{\sim}{x}_{ij}\underset{\sim}{x}'_{ij}]^{-1}[\sum_{i=1}^{\ell} \sum_{j=1}^{n_i} \underset{\sim}{x}_{ij}(y_{ij} - \hat{\mu}_i^*)]. \tag{B.5}$$

Plug $\underset{\sim}{\beta}^*_{(0)}$ in the likelihood function (B.1) and consider it as a function of $\underset{\sim}{\mu}$ only,

$$q(\mu_i) = \log \prod_{j=1}^{n_i} \frac{e^{(x'_{ij}\beta^*_{(0)} + \mu_i)y_{ij}}}{1 + e^{x'_{ij}\beta^*_{(0)} + \mu_i}} = \sum_{j=1}^{n_i} \left\{ (x'_{ij}\underset{\sim}{\beta}^*_{(0)} + \mu_i)y_{ij} - \log(1 + e^{x'_{ij}\beta^*_{(0)} + \mu_i}) \right\}.$$

The first derivative function of $q(\mu_i)$ over μ_i is

$$q'(\mu_i) = \sum_{j=1}^{n_i} \left\{ y_{ij} - \frac{e^{(x'_{ij}\beta^*_{(0)} + \mu_i)}}{1 + e^{x'_{ij}\beta^*_{(0)} + \mu_i}} \right\} = \sum_{j=1}^{n_i} \left\{ y_{ij} - (1 + e^{-(x'_{ij}\beta^*_{(0)} + \mu_i)})^{-1} \right\}. \tag{B.6}$$

Similar to above, we apply Taylor series approximation

$$(1 + e^{-(x'_{ij}\beta^*_{(0)} + \mu_i)})^{-1} \approx (1 - e^{-\mu_i}e^{-x'_{ij}\beta^*_{(0)}}).$$

So (B.6) equals

$$\sum_{j=1}^{n_i} \left\{ y_{ij} - (1 - e^{-\mu_i}e^{-x'_{ij}\beta^*_{(0)}}) \right\}.$$

Solve for $q'(\mu_i) = 0$, then the approximate posterior mode (quasi-mode) of μ_i can be obtained as

$$\mu_i^* = \log[\frac{\sum_{j=1}^{n_i} e^{-x'_{ij}\beta^*_{(0)}}}{n_i(1 - \bar{y}_i)}].$$

Notice that the term $(1 - \bar{y}_i)$ in denominator may cause trouble for this posterior mode, because the binary response variable could lead to $\bar{y}_i = 1$ for some i, so that $(1 - \bar{y}_i) = 0$. We borrow the idea from the empirical logistic transform (ELT) and make a little adjustment to avoid 0's in denominator

$$\mu_i^* \approx \log[\frac{\frac{1}{n_i}\sum_{j=1}^{n_i} e^{-x'_{ij}\beta^*_{(0)}}}{1 - \bar{y}_i + \frac{1}{2n_i}}]. \tag{B.7}$$

Appendix C   Empirical Logistic Transform (ELT)

We consider the empirical logistic transform (ELT) without covariates for binary data. See Cox and Snell (1972) for the empirical logistic transform (ELT) that accommodates binary data. Letting y denote a binomial random variable with success probability p, the empirical logistic transform, Z, is

$$Z = \log(\frac{Y + \frac{1}{2}}{n - Y + \frac{1}{2}}),$$

and the corresponding variance V is

$$V = \frac{(n+1)(n+2)}{n(Y+1)(n-Y+1)}.$$

Then, using the approximation of Cox and Snell (1972), Z has a normal distribution with mean θ and variance V, where θ is unknown according to Cox and Snell (1972). Suppose y is the variable of length ℓ. Each of the binary response $y_i(i = 1, ..., \ell)$ follows a binomial distribution with corresponding number of observations n_i and probability p_i. The goal is to estimate the Bernoulli probability parameter p_i. Here we assume that

$$y_i \stackrel{ind}{\sim} \text{Binomial}\{n_i, p_i\}$$

and for logistic transform we define $z_i = \log(\frac{y_i + \frac{1}{2}}{n_i - y_i + \frac{1}{2}})$ as the the empirical logistic transforms, and $V_i = \frac{(n_i+1)(n_i+2)}{n_i(y_i+1)(n_i-y_i+1)}$ as the associated variances. Then,

$$z_i \mid \mu_i \stackrel{ind}{\sim} \text{Normal}(\mu_i, V_i).$$

We can actually start with this approximation based on the empirical logistic transform. However, this approximation will not work for binary data with covariates at the unit level, but we will make use it in our approximation for logistic regression with binary data in a less important manner.

Acknowledgments

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669-679.

Beaumont, M. A., Zhang, W. and Balding, D. J. (2002). Approximate Bayesian Computation in Population Genetics. *Genetics*, **162**, 2025-2035.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, Cambridge, United Kingdom.

Central Bureau of Statistics Thapathali. Nepal living Standards Survey 2003/04. *Statistical Report Volume 1*, Kathmandu, Nepal.

Chen, M-H, Ibrahim, J. G. and Kim, S. (2008). Properties and Implementation of Jeffreys's Prior in Binomial Regression Models. *Journal of the American Statistical Association*, **103**, 1659-1664.

Cox, D. R. and Snell, E. J. (1972). *Analysis of Binary Data, Second Edition*. Chapman Hall. **2**, 31-32.

Dharmadhikari, S. W. and Joag-Dev, K. (1988). *Unimodality, Convexity and Applications*. Academic Press, New York.

Fearnhead, P. and Prangle, D. (2012). Constructing Summary Statistics for Approximate Bayesian computation: Semi-automatic ABC. *Journal of the Royal Statistical Society*, **B(74)**, 419-474.

Faes, C., Ormerod, J. T. and Wand, M. P. (2011). Variational Bayesian Inference for Parametric and Nonparametric Regression With Missing Data. *Journal of the American Statistical Association*, **106**, 959-971.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: an Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, **74**, 269-277.

Ferkingstad, E. and Rue, H. (2015). Improving the INLA approach for approximate Bayesian inference for latent Gaussian models. *Electronic Journal of Statistics*, **9**, 2706-2731.

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

Ghosh, M., Natarajan, K., Stroud, T. W. F. and Bradley P. Carlin, B P. (1998). Generalized Linear Models for Small-Area Estimation. *Journal of the American Statistical Association*, **93, 441**, 273-282

Held, L., Schrödle, B. and Rue, H. (2010). Posterior and Cross-Validatory Predictive Checks: A Comparison of MCMC and INLA. *Statistical Modeling and Regression Structures, Physica-Verig HD*, **2010**, 91-110.

Holmes, C. C. and Held, L. (2006). Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. *Bayesian Analysis*, **1(1)**, 145-168.

Illian, J. B., Sørbye, S. H. and Rue, H. (2012). A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA). *The Annals of Applied Statistics*, **6(4)**, 1499-1530.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian Parameter Estimation via Variation Methods. *Statistics and Computing*, **10**, 25-37.

Liu, J. and Dey, D. K. (2008). Skew random effects in multilevel binomial models: an alternative to nonparametric approach. *Statistical Modelling*, **8(3)**, 221-241

Larsen, K., Petersen, J. H., Budtz-Jørgensen, E. and Endahl, L. (2000). Interpreting Parameters in the Logistic Regression Model with Random Effects. *Biometrics*, **56(3)**, 909-914.

Miroshnikov, A. and Conlon, E. M. (2015). Parallel Markov Chain Monte Carlo for Non-Gaussian Posterior Distributions. *Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts*, pp. 1-32.

Nandram, B. (2000). Bayesian Generalized Linear Models for Inference about Small Areas. *In Generalized linear models: A Bayesian Perspective*, Eds. D. K. Dey, S.K. Ghosh and B.K. Mallick, Marcel Dekker, Chapter 6, 91-114.

Nandram, B. (1989). Discrimination between Complementary Log-log and Logistic Model for Ordinal Data. *Communications in Statistics, Theory and Methods*, **18**, 2155-2164.

Nandram, B. and Chen, M-H. (1996). Reparameterizing the Generalized Linear Model to Accelerate Gibbs Sampler Convergence. *Journal of Statistical Computation and simulation*, **54**, 129-144.

Nandram, B. and Choi, J. W. (2010). A Bayesian Analysis of Body Mass Index Data from Small Domains Under Nonignorable Nonresponse and Selection. *Journal of the American Statistical Association*, **105**, 120-135.

Nandram, B. and Erhardt, E. (2005). Fitting Bayesian Two-Stage Generalized Linear Models Using Random Samples via the SIR Algorithm. *Sankhya*, **66**, 733-755.

Ormerod, J. T., and Wand, M. P. (2010). Explaining Variational Approximations. *The American Statistician*, **64**, 140-153.

Paulino, C.D., Silva, G. and Achcar, J.A. (2005). Bayesian Analysis of Correlated Misclassified Binary Data. *Computational Statistics & Data Analysis*, **49(4)**, 1120-1131.

Potthoff, R.F., Woodbury, M.A. and Manton, K.G. (1992). Equivalent Sample Size and Equiv-

alent Degrees of Freedom Refinements for Inference Using Survey Weights under Super-population Models. *Journal of the American Statistical Association*, **87**, 383-396.

Roberts, G., Rao, J. N. K. and Kumar, S. (1987). Logistic Regression Analysis of Sample Survey Data. *Biometrika*, **74(1)**, 1-12.

Rubin, D. B. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, **9(1)**, 130-134.

Rue, H., Martino, S. and Chopin, IN. (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society* Series B, **71(2)**, 319-392.

Santos, C. C., Loschi, R. H. and Arellano-Valle, R. B. (2013). Parameter Interpretation in Skewed Logistic Regression with Random Intercept. *Bayesian Analysis*, **8(2)**, 381-410.

Souza, A.D. and Migon, H.S. (2010). Bayesian outlier analysis in binary regression. *Journal of Applied Statistics*, **37(8)**, 1355-1368.

Scott, S. L., Blocker, A. W., Bonassi1, F. V., Chipman, H. A., George, E. I. and McCulloch, R. E. (2013). Bayes and Big Data: The Consensus Monte Carlo Algorithm. *Technical Report*, Google, Inc., 1-22.

Torabi, M. and Rao, J. N. K. (2014). On Small Area Estimation under a Sub-Area Level Model. *Journal of Multivariate Analysis*,**127**, 36-55.

Table 1: Categorical tables for 60,221 households by posterior mean of model (2) with random effects at the ward level projected to the households and the model (1) at household level

			Model 2			
Model 1	.0 − .2	.2 − .4	.4 − .6	.6 − .8	.8 − 1	Total
.0 − .2	1	19	586	228	29	863
.2 − .4	2	187	1845	139	44	2217
.4 − .6	251	1730	29861	9638	664	42144
.6 − .8	87	548	9115	3491	253	13494
.8 − 1	9	24	1022	384	64	1503

Table 2: Categorical tables for 60,221 households by posterior standard deviation of model (2) with random effects at the ward level projected to the households and the model (1) at household level

			Model 2			
Model 1	.0 − .1	.1 − .2	.2 − .3	.3 − .4	.4 − .5	Total
.0 − .1	5	66	626	41	1	739
.1 − .2	13	233	2411	105	2	2764
.2 − .3	348	2826	46500	3391	37	53102
.3 − .4	5	180	2859	553	3	3600
.4 − .5	0	0	13	2	1	16

Table 3: Categorical tables for 60,221 households by posterior coefficient of variation of model (2) with random effects at the ward level projected to the households and the model (1) at household level

			Model 2			
Model 1	.02 − .05	.05 − .10	.10 − .25	.25 − .50	> .50	Total
.02 − .05	0	0	4	20	9	33
.05 − .10	0	3	20	188	66	277
.10 − .25	1	4	80	856	368	1309
.25 − .50	28	128	791	23694	10535	35176
> .50	22	62	399	9996	12947	23426

Figure 1: Comparison of the posterior means (PM) of the household proportions by the approximate method and the MCMC method
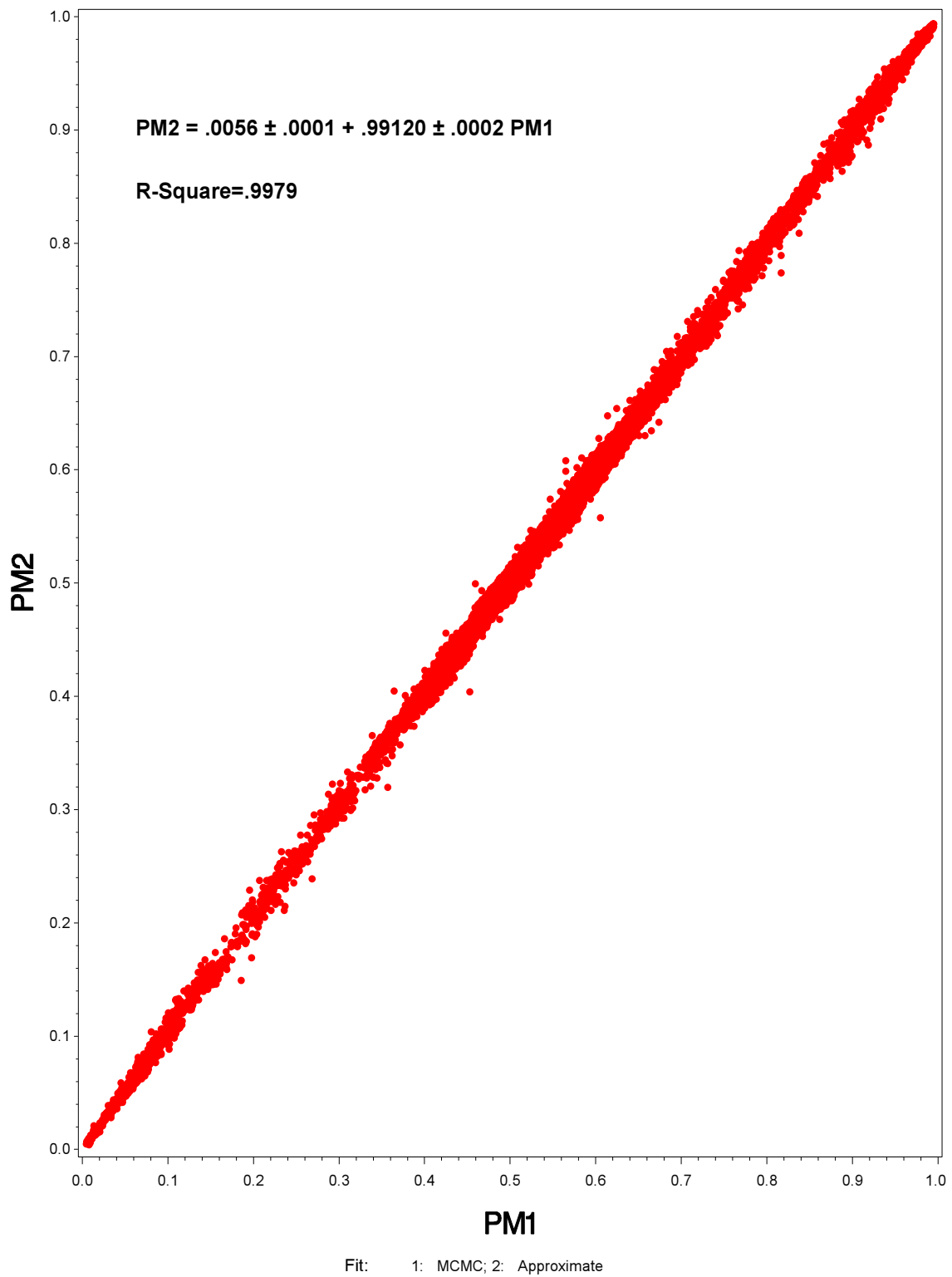

PM2 = .0056 ± .0001 + .99120 ± .0002 PM1

R-Square=.9979

Fit: 1: MCMC; 2: Approximate

Figure 2: Comparison of the posterior standard deviations (PSD) of the household proportions by the approximate method and the MCMC method
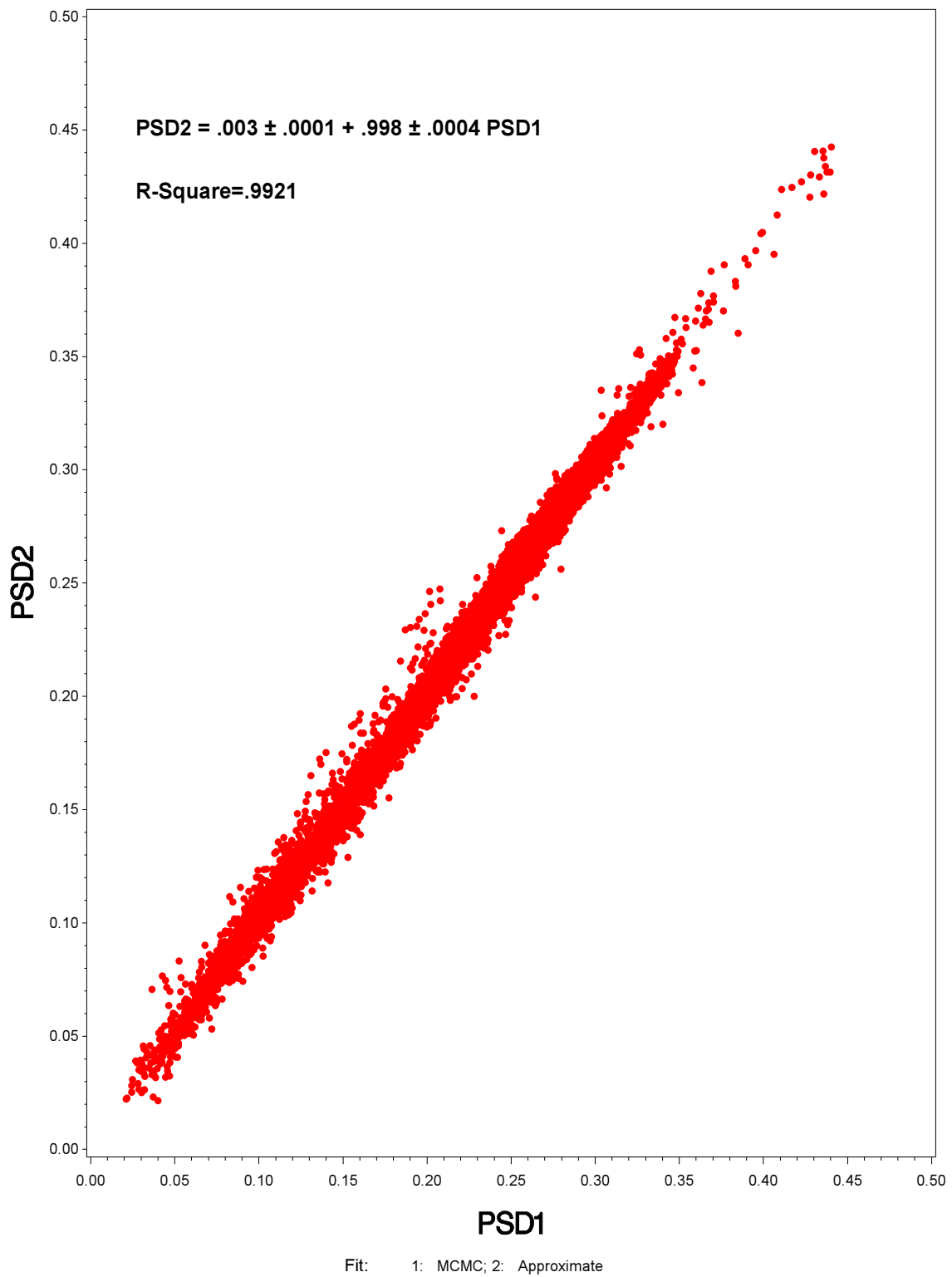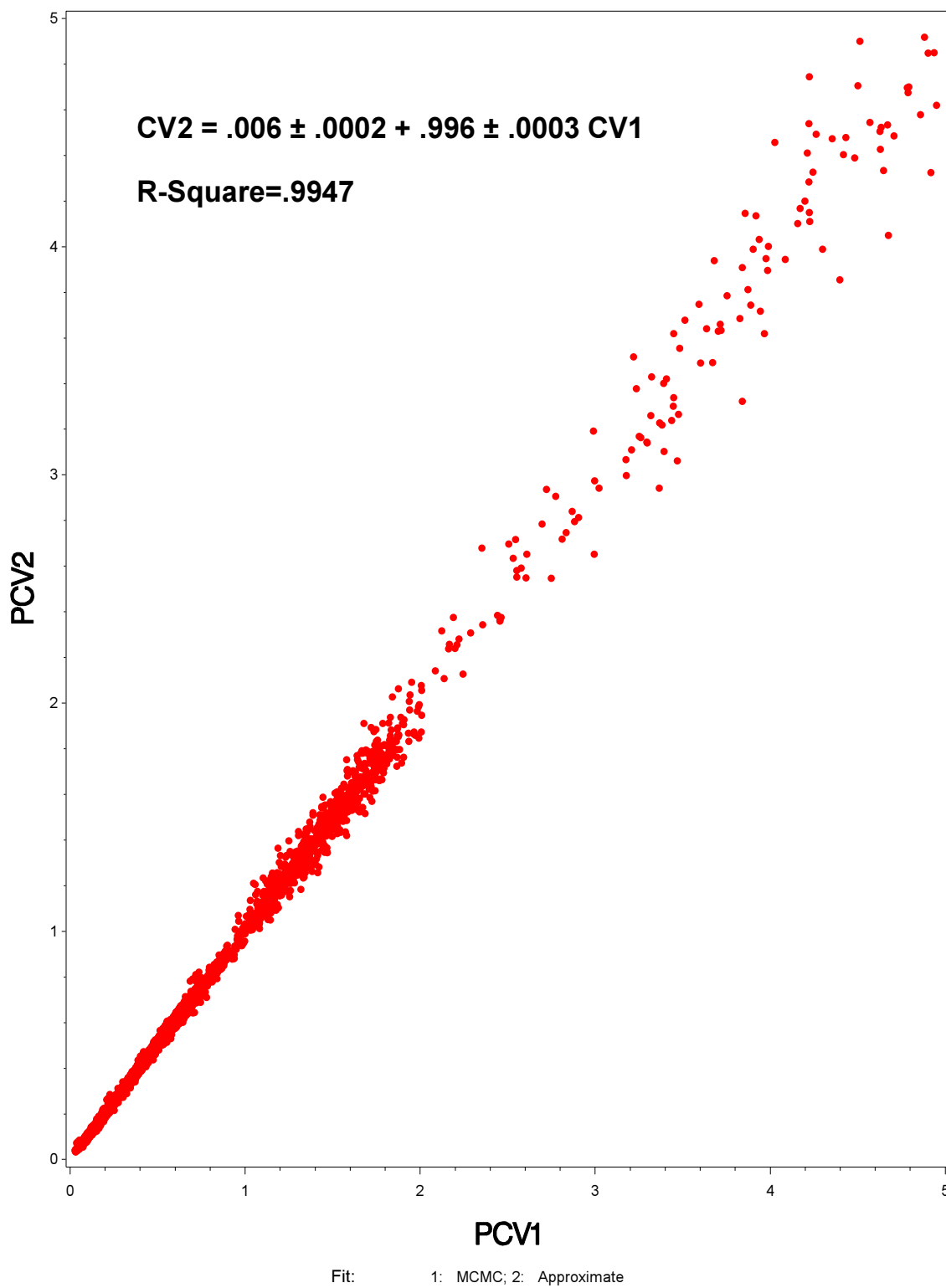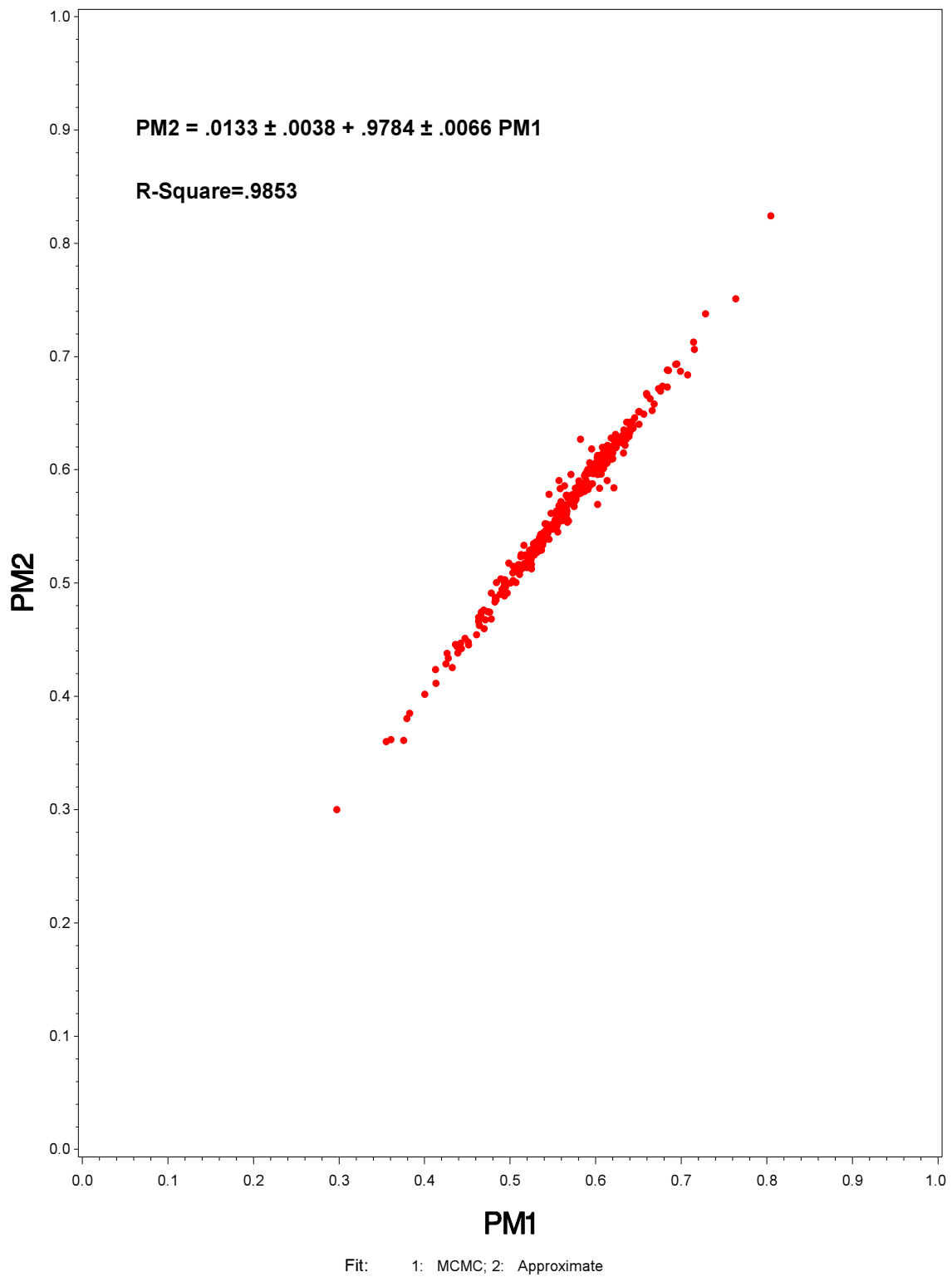

PSD2 = .003 ± .0001 + .998 ± .0004 PSD1

R-Square=.9921

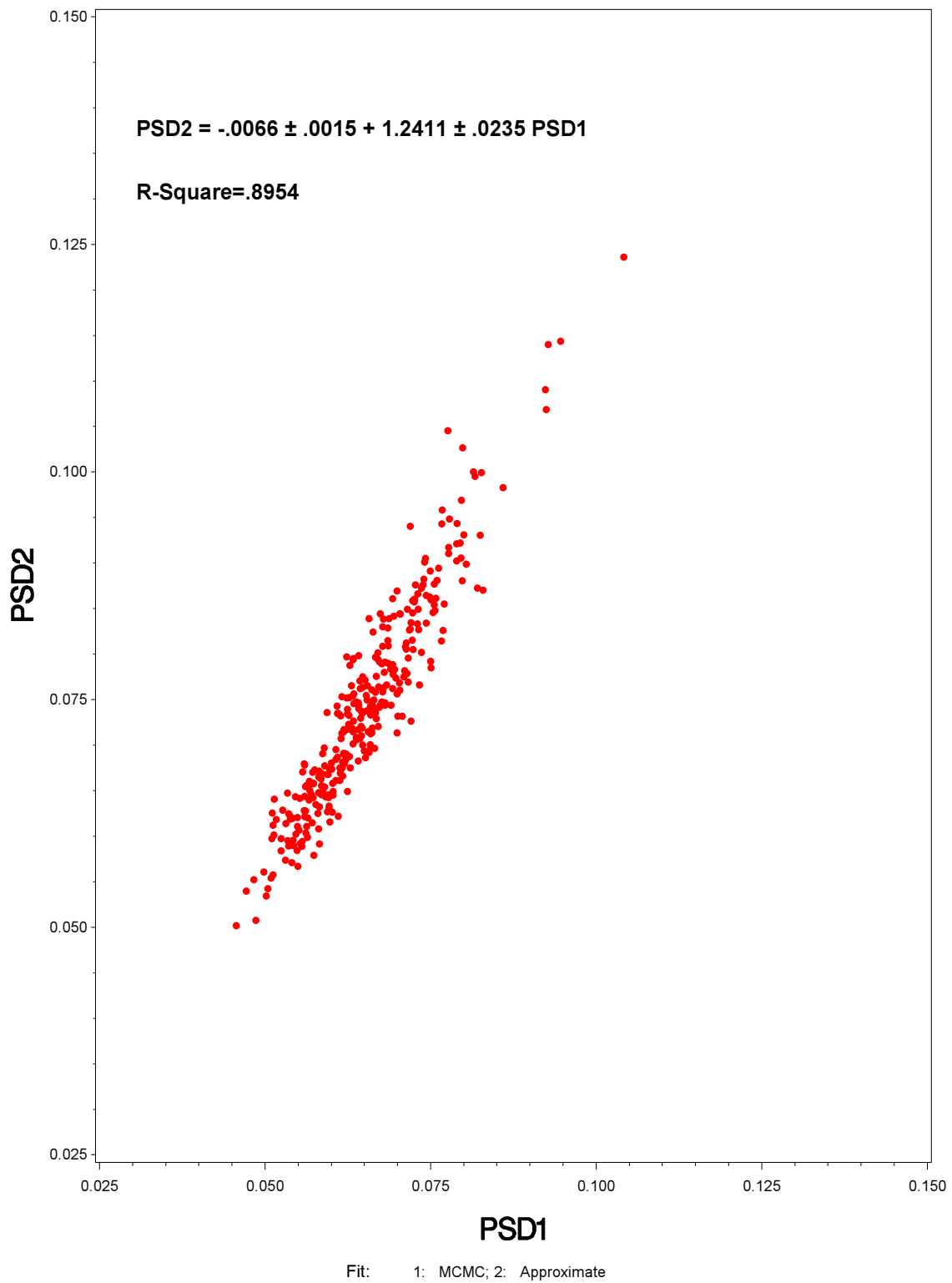Fit: 1: MCMC; 2: Approximate

Figure 3: Comparison of the posterior coefficient of variations (CV) of the household proportions by the approximate method and the MCMC method
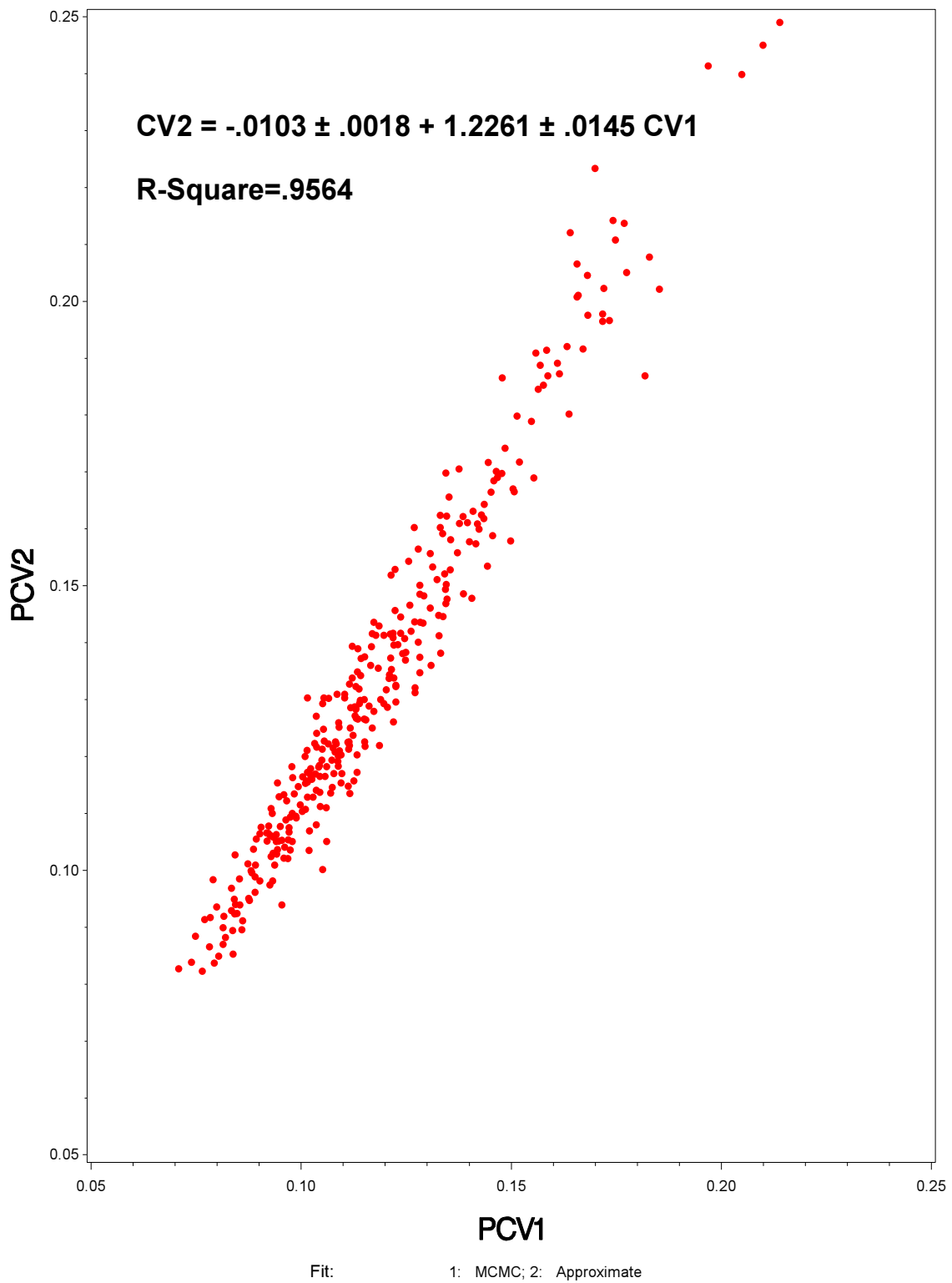

Fit: 1: MCMC; 2: Approximate

Figure 4: Comparison of the posterior means (PM) of the ward proportions by the approximate method and the MCMC method

Fit: 1: MCMC; 2: Approximate

Figure 5: Comparison of the posterior standard deviations (PSD) of the ward proportions by the approximate method and the MCMC method

Fit: 1: MCMC; 2: Approximate

Figure 6: Comparison of the posterior coefficient of variations (CV) of the ward proportions by the approximate method and the MCMC method

Fit: 1: MCMC; 2: Approximate

Figure 7: Plots of the empirical posterior densities of δ^2 and β_0 for the approximate method and the MCMC method
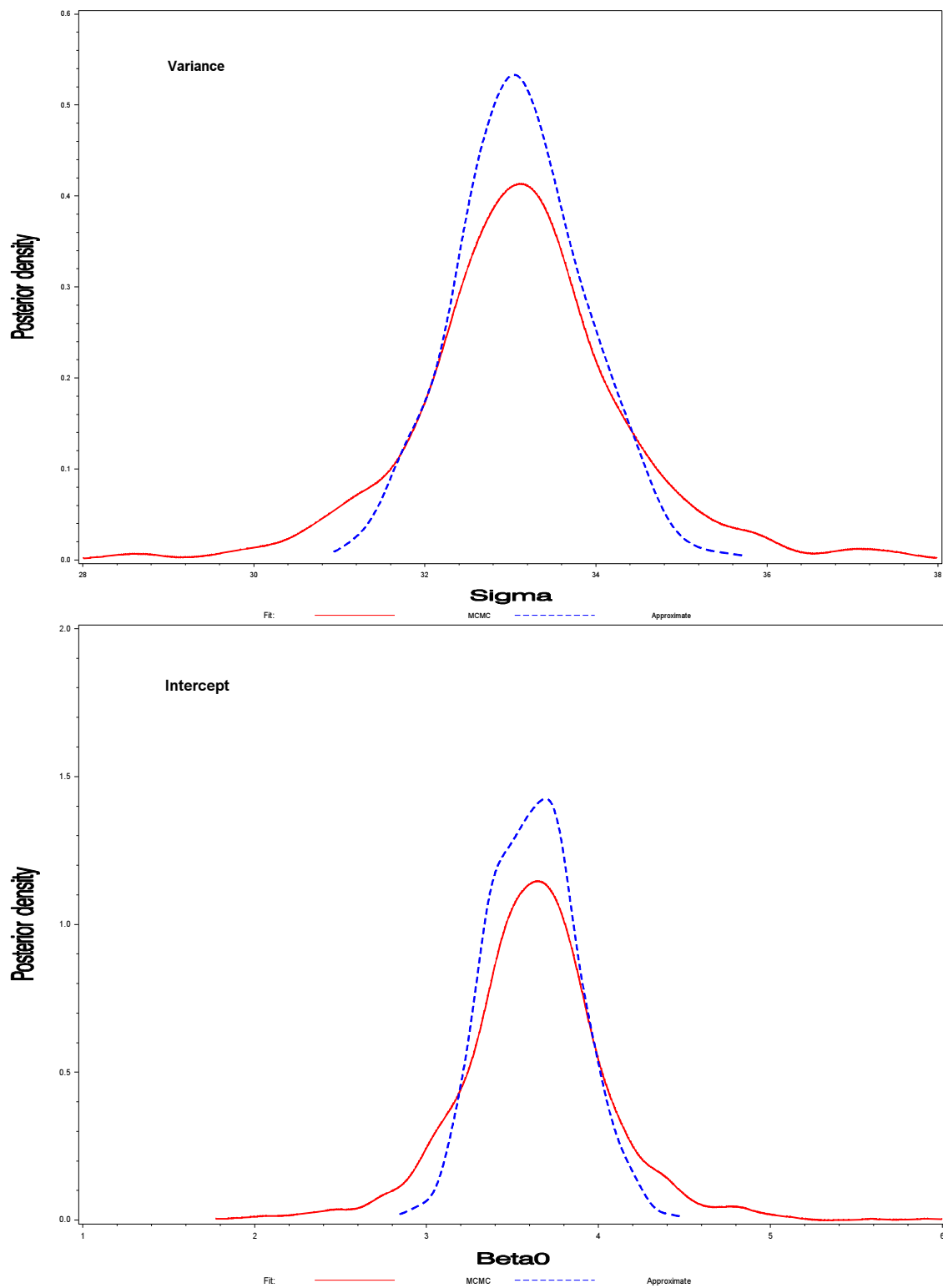
Figure 8: Plots of the empirical posterior densities of regression coefficients for the approximate method and the MCMC method
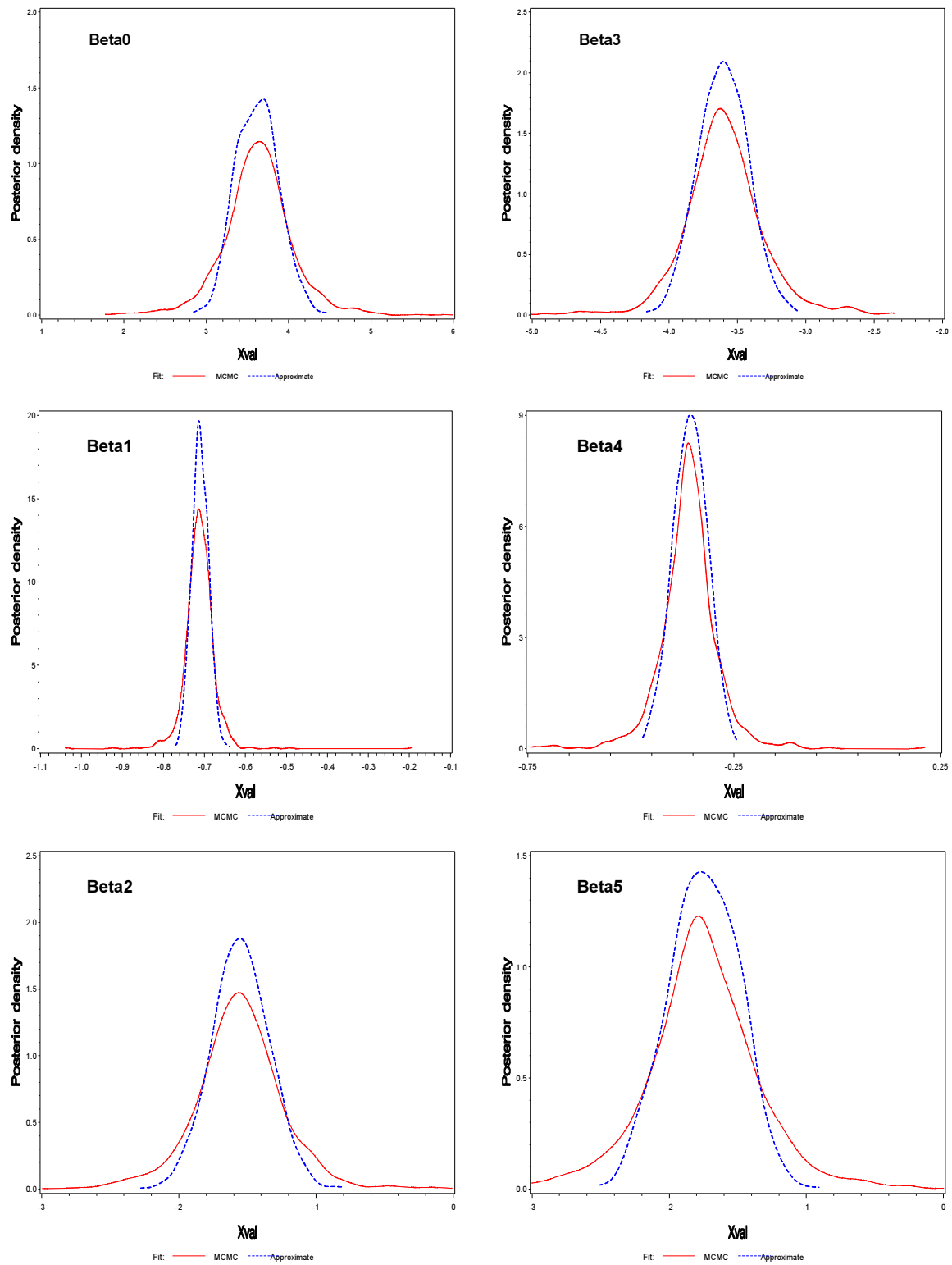
Figure 9: Comparison of the posterior means (PM) of the household proportions by the two MCMC methods
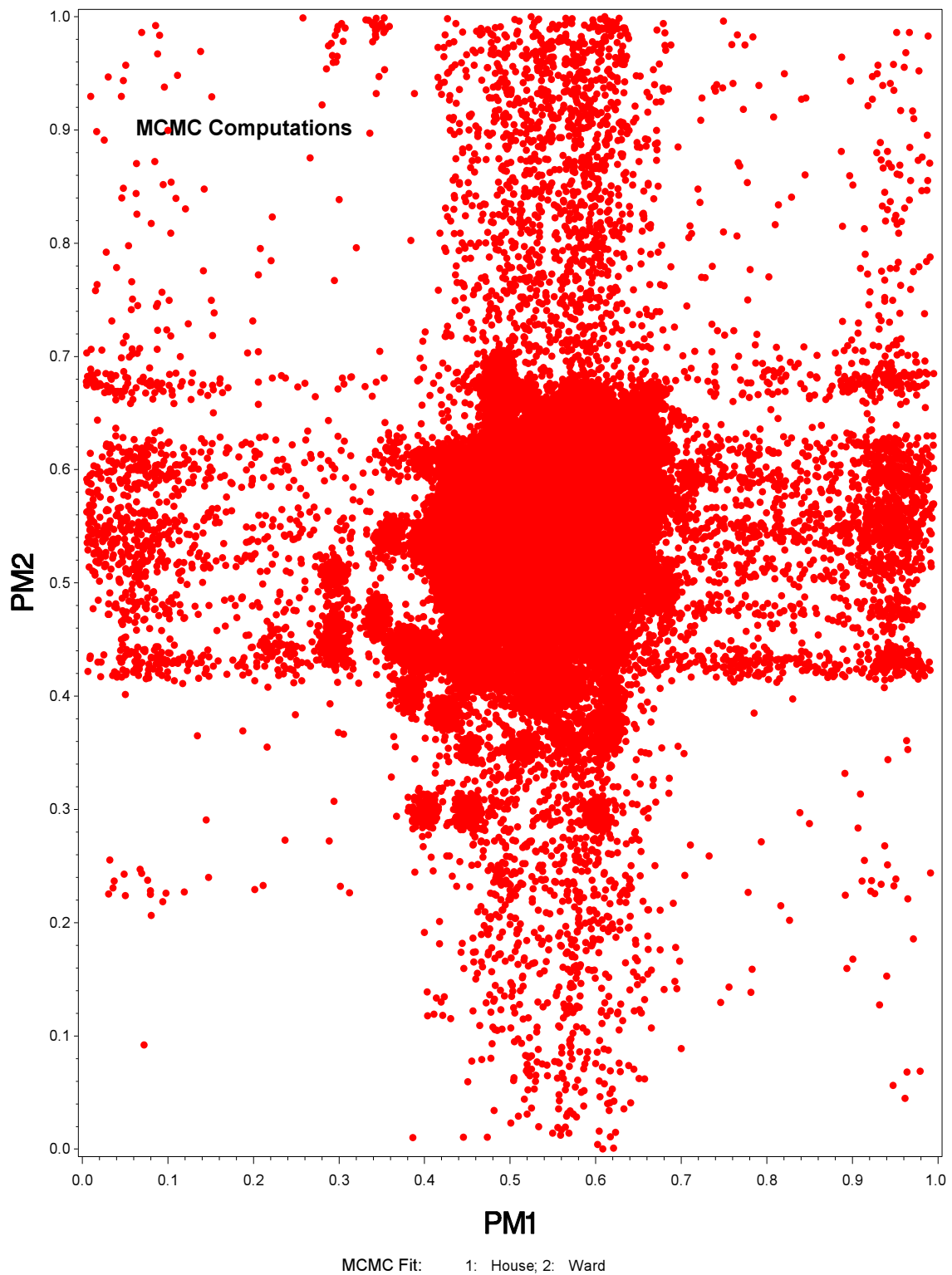

MCMC Fit: 1: House; 2: Ward

Figure 10: Comparison of the posterior standard deviations (PSD) of the household proportions by the two MCMC methods
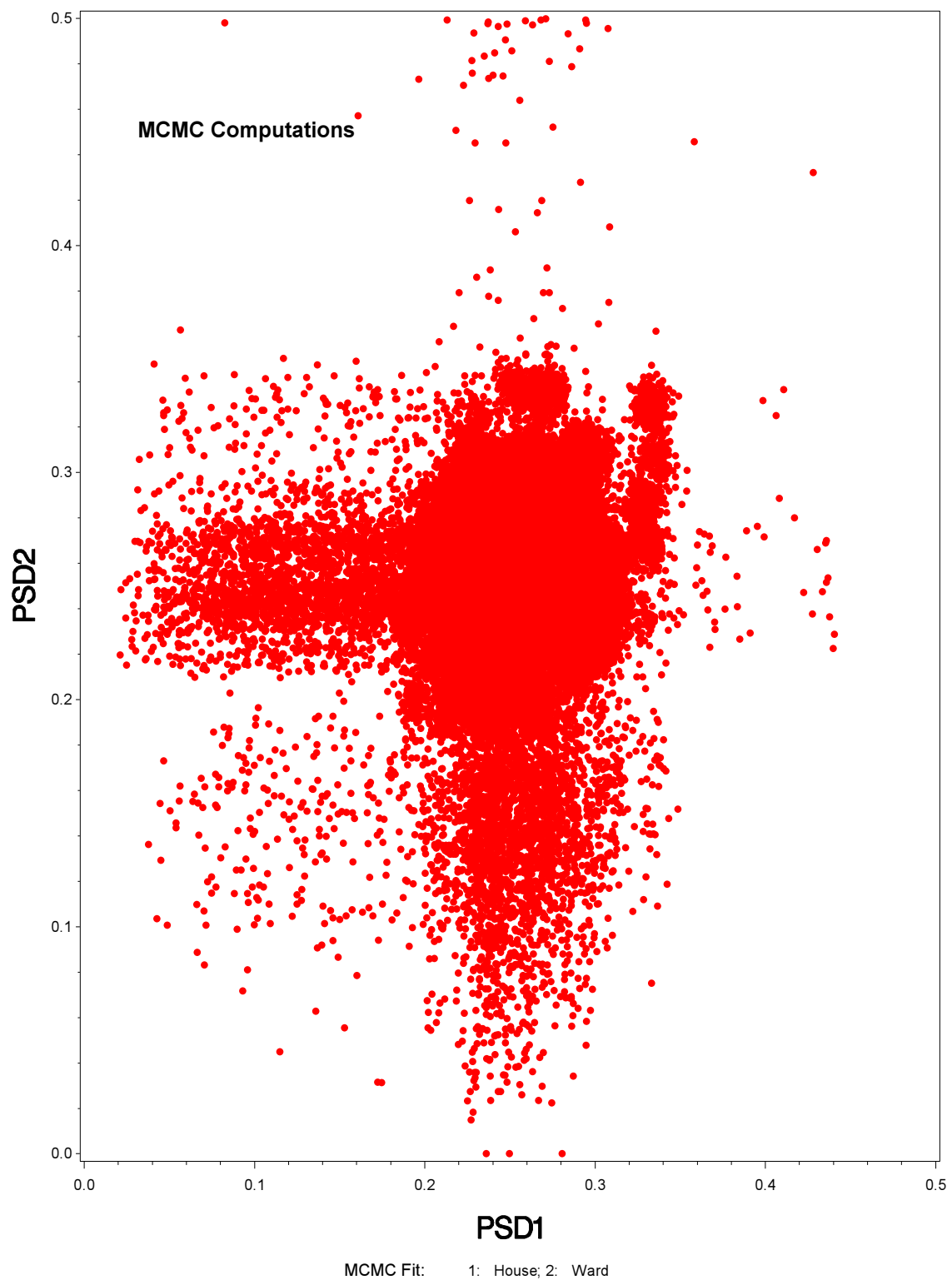

MCMC Fit: 1: House; 2: Ward

Figure 11: Comparison of the posterior coefficient of variations (CV) of the household proportions by the two MCMC methods

MCMC Fit: 1: House; 2: Ward