# Selection of prior distributions for multivariate small area models with application to small area health insurance estimates*

Ryan Janicki
Center for Statistical Research and Methodology, U. S. Census Bureau

**Abstract**
In sample surveys, it is often the case that there is insufficient sample size to obtain reliable direct estimates for a parameter of interest for certain domains. Precision can be increased, at the cost of reliance on possibly misspecified models, by introducing small area models which "borrow strength" by connecting different areas and incorporating auxiliary covariate information. This article considers multivariate generalized linear models for analyzing survey data, with special attention given to the mixed effect multinomial logistic regression model. Because it is possible that the predictors used may be measured with error, the small area models considered include error terms which attempt to account for possible measurement errors, in addition to the usual area-specific random effects. A comparison of the model where area-specific random effects are correlated is made with the model where area-specific random effects are assumed to have independent components. A general theorem is presented which gives necessary and sufficient conditions for the propriety of the posterior. An example is given where a simulated data set is analyzed using the model to estimate the proportion in different income levels for different demographic groups. This example compares the results of the Gibbs sampler when the posterior is proper, to the Markov chain from a Gibbs sampler when the posterior distribution is improper. The results of this example indicate that we can have improved estimates over estimates based on a small area model for single components when we use area-specific random effects with correlated components.

## 1. Introduction

In sample surveys, it is often the case that there is insufficient sample size to obtain reliable direct estimates for a parameter of interest for certain small areas. By small area (or small domain), we mean a subgroup of a population, such as a geographic region, or a cross-classification of demographic factors, such as age, race, or gender. There has been great demand for reliable estimates at progressively smaller domains. For example, the Small Area Health Insurance Estimates (SAHIE) program provides estimates of the number of people without health insurance by demographic groups and income categories at the state and county level. The Small Area Income and Poverty Estimates (SAIPE) program produces estimates of income and poverty levels by demographic groups at the state, county, and school district level. It is important that the estimates produced by these programs are reliable, as the estimates help determine the allocation of federal funds and the administration of federal programs.

Most surveys provide little information on one or more small areas, since they are often designed to produce accurate estimates at a higher level of aggregation, and obtaining sufficient sample sizes at the small area level can be prohibitively expensive. Due to the small sample size, the direct estimate (the estimate based only on the small area specific data) can have inadequate precision to be considered reliable. For this reason, small area modelling approaches are often used to improve upon direct estimates by using an appropriate model

*This report is released to inform interested parties of (ongoing) research and to encourage discussion (of work in progress.) The views expressed are those of the author and not necessarily those of the U.S. Census Bureau. The author would like to thank Jerry Maples, Sam Szelepka and Eric Slud for their careful review of this paper, as well as Donald Malec for many discussions about this problem.

which allow estimates to "borrow strength" by relating similar small areas and making use of relevant covariate information from other sources, such as administrative records.

Bayesian methods are widely used for small area problems, as a hierarchical approach can be particularly effective in connecting local areas. The full hierarchical Bayesian method can have computational advantages, as there are many algorithms available, such as the Gibbs sampler, and the adaptive rejection algorithm, which make sampling from the posterior distribution straightforward, and there are software packages, such as BUGS and JAGS, which can be used to implement these algorithms. Also, using a full hierarchical Bayes model and specifying prior distributions eliminates the need to integrate over the unobserved random effects.

However, there are several difficulties in the Hierarchical Bayes approach. First, the posterior distribution is often intractable, necessitating the use of Markov Chain Monte Carlo techniques to sample from the posterior distribution, and it can be difficult to choose an appropriate proposal distribution for the Metropolis-Hastings algorithm to achieve a suitable acceptance rate. Second, there is difficulty selecting an appropriate prior distribution. One approach is to choose conjugate priors, which simplify computations required to sample from the posterior distribution. However, without strong prior information, it can be difficult to select the values of the hyperparameters. Also, if the hyperparameters are chosen so that a "vague," or minimally informative proper prior distribution is used, the rate of convergence of the Gibbs sampler for the full set of parameters may be reduced due to the widely dispersed mass of the resulting posterior (Natarajan and Kass, 2000).

Since there is typically is little information about the hyperparameters, it is desirable to use a flat "noninformative" prior. Noninformative priors are designed to reflect the lack of information about the hyperparameter. Motivated by this desire to reflect lack of information, and a desire for invariance properties, noninformative priors are often improper, in the sense that the integral of the prior distribution over the parameter space is infinite. The danger in using an improper prior is that the resulting posterior distribution could also be improper.

Much has been written concerning the propriety of the posterior distribution for univariate hierarchical Bayesian models. Hobert and Casella (1996) considered the linear model, and gave conditions for the propriety of the posterior distribution when a uniform distribution is used on the regression parameters and improper inverse gamma distributions are used on the variance components. Ghosh et al. (1998) considered univariate generalized linear models for small area estimation problems, and gave conditions for the propriety of the posterior distribution when a uniform distribution is used on the regression parameters, and vague, but proper inverse gamma distributions are used on the area-specific random effects components. Sun et al. (2001) extended the work of Ghosh et al. (1998) by allowing improper inverse gamma priors for the variance components.

In this paper we consider multivariate extensions of univariate small area models and investigate prior distributions that result in proper posterior distributions. One of the most commonly used multivariate small area models is a multivariate extension of the Fay-Herriot model (Rao, 2003, p. 81). Suppose we have a $d \times 1$ vector of survey estimators $\widehat{\theta}_i = \left(\widehat{\theta}_{i1}, \ldots, \widehat{\theta}_{id}\right)^T$ and

$$\widehat{\theta}_i = \theta_i + \mathbf{e}_i,\ i = 1, \ldots, m \tag{1}$$

where $\theta_i = (\theta_{i1}, \ldots, \theta_{id})^T$ with $\theta_{ij} = g_j\left(\overline{Y}_{ij}\right), j = 1, \ldots, d$, and the sampling errors $\mathbf{e}_i = (e_{i1}, \ldots, e_{id})^T$ are independent $d-$variate normal, $N_d(\mathbf{0}, \Psi_i)$, with known covariance matrices $\Psi_i$, conditional on $\theta_i$. Here, $\overline{Y}_{ij}$ is the $i$th small area mean for the $j$th characteristic. The means $\theta_i$ are related to area-specific auxiliary data $\mathbf{X}_{ij}$ through a linear model

$$\theta_i = \mathbf{X}_i\beta + \mathbf{v}_i,\ i = 1, \ldots, m \tag{2}$$

where the area-specific random effects $\mathbf{v}_i$ are independent $N_d(\mathbf{0},\Sigma)$, $\mathbf{X}_i$ is a $d \times p$ matrix with rows $\mathbf{X}_{ij}$, and $\beta$ is a $p$–dimensional vector of unknown coefficients. Combining equations (1) and (2) gives

$$\widehat{\theta}_i = \mathbf{X}_i\beta + \mathbf{v}_i + \mathbf{e}_i.$$

It was argued by Fay (1987) that the multivariate Fay-Herriot model can lead to more efficient estimators of the small area means $\overline{Y}_{ij}$ than a univariate approach, because it takes advantage of the correlations between the components of $\widehat{\theta}_i$.

Inference can be made by applying the hierarchical Bayes approach after specifying a prior distribution on the model parameters $(\beta,\Sigma,\Psi)$. Datta et al. (1998) considered the selection of prior distributions for the multivariate Fay-Herriot model, with unknown variance components, and used the hierarchical Bayes approach to obtain model-based estimates. They found necessary and sufficient conditions for the propriety of the resulting posterior distribution corresponding to a certain class of improper prior distributions on the components of variance matrices.

The multivariate Fay-Herriot model assumes normality, which may not always be an appropriate assumption if, for example, the responses are categorical. In this paper we consider multivariate generalized linear models which can be used for small area problems. This approach can be used to model survey data that may not be continuous, and allows "borrowing strength" by relating similar small areas and incorporating area-specific covariate information. It also incorporates "errors-in-variables" modeling, which allows for the possibility that model predictors, such as administrative records, may be measured with error. Section 2 introduces the bivariate generalized linear model and gives conditions for the propriety of the posterior distribution. Section 3 extends the bivariate model to a multivariate model, when the components of the area-specific random effects are assumed to be independent. Finally, a simulated data set that is similar to the data set used by the SAHIE program is analyzed in Section 4.

## 2. A bivariate hierarchical small area model

Let $\{\mathbf{Y}_{ij}, i = 1,\ldots,m, j = 1,\ldots,n_i\}$ be conditionally independent, 2-dimensional random vectors, given a parameter $\theta_{ij}$, with sampling distribution belonging to a natural bivariate exponential family, so that

$$f(\mathbf{y}_{ij} \mid \theta_{ij}) = \rho(\mathbf{y}_{ij}) \exp\left\{\theta_{ij}^T \mathbf{y}_{ij} - \psi(\theta_{ij})\right\},$$

where $\theta_{ij}^T = (\theta_{ij1}, \theta_{ij2})$ is the natural parameter. For $k = 1,2$, let $g_k$ be known, monotone differentiable functions such that

$$g_k\left(\theta_{ijk}\right) = \mathbf{X}_{ijk}\beta + \gamma_{ik} + \varepsilon_{ijk},$$

where $\mathbf{X}_{ijk}$ is a vector of covariates, $\beta \in \mathbb{R}^p$ is a vector of unknown coefficients, and $\gamma_i$ is a vector of area-specific effects. We assume that the $(2\sum_{i=1}^m n_i \times p)$ matrix $\mathbf{X}$, with rows $\mathbf{X}_{ijk}$, is of full rank $p \leq 2\sum_{i=1}^m n_i$.

The reason for the inclusion of $\varepsilon_{ijk}$ is that the usual assumption of regression models, that predictors are measured without error, is not always appropriate. Standard regression models assume that the regressors are observed without error and that the model is used to account for the errors in the response variables. However, it can often be the case that some regressors are measured with error. In particular, if administrative records are used as predictors, as they are in the SAHIE model, there could be nonnegligible sampling error. The inclusion of $\varepsilon_{ijk}$ is an attempt to model this measurement error. This approach is known as errors-in-variables modeling (Fisher, 2003).

To complete the model specification, we use the prior distributions

$$\begin{aligned} \gamma_i = (\gamma_{i1}, \gamma_{i2})^T \mid \Sigma &\overset{i.i.d}{\sim} N_2(\mathbf{0}, \Sigma), \\ \varepsilon_{ijk} \mid \sigma^2 &\overset{i.i.d}{\sim} N\left(0, \sigma^2\right), \\ \sigma^2 \mid a, b &\sim IG(a, b), \\ \pi\left(\beta, \sigma_1^2, \sigma_2^2, \rho \mid a_1, a_2, a_3\right) &\propto \left(\sigma_1^2\right)^{-a_1-1} \left(\sigma_2^2\right)^{-a_2-1} \left(1-\rho^2\right)^{-a_3-1}, \end{aligned} \tag{3}$$

where

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix},$$

$IG(a,b)$ represents the inverse gamma distribution, $\varepsilon_{ijk}$ and $\gamma_i$ are independent for all $i, j$, and $k$, and $a, b, a_1, a_2$, and $a_3$ are hyperparameters.

We do not assume that $\rho = 0$, because in some problems it makes sense to allow for the possibility that components of the area-specific random effects $\gamma_i$ are correlated. For example, the SAHIE program models the number of people in the three income to poverty ratio (IPR) categories $IPR \leq 200\%$, $200 < IPR \leq 250\%$, and $IPR > 250\%$ for different geographic regions and demographic subgroups. Intuitively, we should expect strong correlation of the random effects for the first two categories.

*Remark* 2.1. When each $a_i = -1/2$, the priors are the scale invariant priors for the variance components, and when $a_i = -1$, the priors are the improper uniform priors.

*Remark* 2.2. The prior specification for $(\beta, \Sigma)$ in model (3) includes the class of generalized Wishart distributions (Berger and Sun, 2008)

$$\pi_W(\beta, \Sigma) \propto \frac{1}{\sigma_1^{3-c_1} \sigma_2^{2-c_2} \left(1-\rho^2\right)^{2-c_2/2}}$$

and the prior

$$\pi_J(\beta, \Sigma) \propto |\Sigma|^{-c}.$$

*Remark* 2.3. Model (3) is a bivariate extension of the model considered by Ghosh et al. (1998). We consider this extension, rather than using a univariate model for each component of the observation, because when the observations are multivariate, there could be strong correlation which can be exploited to improve the estimators.

The full conditionals for model (3) are

$$\begin{aligned}
\beta \mid \theta,\gamma,\sigma^2,\mathbf{Y} &\sim N_p\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{G}-\mathbf{Z}\gamma\right),\sigma^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right),\\
\gamma_i \mid \theta_i,\beta,\Sigma,\sigma^2,\mathbf{Y} &\sim N_2\left(\left(n_i\mathbf{I}+\sigma^2\Sigma^{-1}\right)^{-1}\sum_{j=1}^{n_i}\left(\mathbf{G}_{ij}-\mathbf{X}_{ij}\beta\right),\left(\frac{n_i}{\sigma^2}\mathbf{I}+\Sigma^{-1}\right)^{-1}\right),\\
\sigma^2 \mid \theta,\beta,\gamma,\mathbf{Y} &\sim IG\left(\sum_{i=1}^{m}n_i+a,\frac{1}{2}\left(\mathbf{G}-\mathbf{X}\beta-\mathbf{Z}\gamma\right)^T\left(\mathbf{G}-\mathbf{X}\beta-\mathbf{Z}\gamma\right)+b\right),\\
\Sigma \mid \gamma &\propto \left(\sigma_1^2\right)^{-a_1-1}\left(\sigma_2^2\right)^{-a_2-1}\left(1-\rho^2\right)^{-a_3-1}\left|\Sigma\right|^{-m/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{m}\gamma_i^T\Sigma^{-1}\gamma_i\right\}.
\end{aligned} \tag{4}$$

$$\begin{aligned}
\theta_{ij} \mid \beta,\gamma_i,\sigma^2,\mathbf{Y} &\propto g_1'\left(\theta_{ij1}\right)g_2'\left(\theta_{ij2}\right)\\
&\quad\times\exp\left\{\theta_{ij}^T\mathbf{t}_{ij}-\psi\left(\theta_{ij}\right)-\frac{1}{2\sigma^2}\sum_{k=1}^{2}\left(g_k\left(\theta_{ijk}\right)-\mathbf{X}_{ijl}\beta-\gamma_{il}\right)^2\right\},
\end{aligned}$$

*Remark* 2.4. If we choose $a_1=a_2=a_3$, then

$$\Sigma \mid \gamma \sim W^{-1}\left(\sum_{i=1}^{m}\gamma_i\gamma_i^T, m+2a_3-1\right),$$

where $W^{-1}(\cdot,\cdot)$ represents the inverse Wishart distribution.

*Remark* 2.5. The conditional distribution of $\theta_{ij}$ given the remaining parameters is not a standard distribution, so a Metropolis-Hastings step must be used. However, if $g_k$ is the identity link for $k=1,2$, then $\pi\left(\theta_{ij} \mid \beta,\gamma_i,\sigma^2,\mathbf{Y}\right)$ is log-concave, and the adaptive rejection algorithm of Gilks and Wild (1992) can be used to sample from this density. This algorithm greatly reduces the complexity of the sampling algorithm, as we do not need to introduce a proposal distribution and the Metropolis-Hastings algorithm.

For the joint posterior distribution to be proper, it is necessary that the full conditional distributions are proper, so that, for example, we must have $m/2+a_k>0$ for all $k$. It is therefore tempting to only check for propriety of the full conditionals. However, the propriety of the full conditionals is not a sufficient condition for the propriety of the joint posterior distribution. Furthermore, it was shown by Hobert and Casella (1996), that when the posterior distribution is improper, the Gibbs Markov chain possesses an invariant measure with infinite mass and thus is null, not positive, recurrent. It follows that if $A$ is any compact set in the parameter space that contains the starting value, the probability of the chain being in the set $A$ after $n$ iterations converges to zero as $n\to\infty$. Therefore, for the inference based on a Gibbs sampler to be valid, we must first check that the joint posterior distribution is proper.

Let $\mathbf{P}_\mathbf{X}=\mathbf{I}-\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ and $\mathbf{Z}=\oplus_{i=1}^{m}\mathbf{D}_i$, where $\mathbf{D}_i=\left(\mathbf{I}_{2\times2}\cdots\mathbf{I}_{2\times2}\right)^T$ is a matrix consisting of $n_i$ identity matrices and $\oplus$ is the direct sum operator.

**Theorem 2.6.** *The following conditions are sufficient for the propriety of the posterior distribution of model* (3)*:*

1. $\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}$ *is of full rank,*
2. $-m/2<a_k$ *for* $k=1,2$ *and* $a_1+a_2<-m/2$*,*

3. $a_3 < m/2 + a_1 + a_2$,
4. $w = \sum_{i=1}^{m} n_i - p/2 + a_1 + a_2 + a > 0$,
5. $\int \left(\mathbf{G}^T\mathbf{W}\mathbf{G}/2 + b\right)^{-w} \prod_{ijk} g_k'\left(\theta_{ijk}\right) \exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij} - \psi\left(\theta_{ij}\right)\right)\right\} d\theta < \infty$,

*where* $\mathbf{W} = \left(\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{X}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{P}_\mathbf{X}\right)$. *The conditions*

1. $-m/2 < a_k < 0$ *for* $k = 1, 2$,
2. $a_3 < m/2 + a_1 + a_2$,
3. $w = \sum_{i=1}^{m} n_i + a_1 + a_2 - p + a > 0$, *and*
4. $\int \left(\mathbf{G}^T\mathbf{P}_\mathbf{X}\mathbf{G}/2 + b\right)^{-w} \prod_{ijk} g_k'\left(\theta_{ijk}\right) \exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij} - \psi\left(\theta_{ij}\right)\right)\right\} d\theta < \infty$

*are necessary.*

The proof is given in the appendix.

*Remark* 2.7. A popular prior distribution for the variance components is $IG(\varepsilon, \varepsilon)$, with $\varepsilon$ set low, as it is a prior distribution that is proper, makes the full conditionals convenient to work with, and can be thought of as an approximation to a noninformative prior. However, by Theorem 2.6, the $IG(\varepsilon, \varepsilon)$ does not have a proper limiting posterior as $\varepsilon \to 0$, hence, as noted by Gelman (2006), posterior inferences can be very sensitive to $\varepsilon$.

## 3. A multivariate extension

Conceptually, the bivariate small area model introduced in Section 2 can easily be extended to a multivariate model. However, specifying prior distributions on the individual correlation components could quickly become unwieldy for higher-dimensional observations, so in this section, to simplify the discussion, we consider a multivariate extension of Model (3) which assumes independence between the components of the random effects.

As in Section 2, let $\left\{\mathbf{Y}_{ij}, i = 1, \ldots, m, j = 1, \ldots, n_i\right\}$ be conditionally independent random vectors, given a parameter $\theta_{ij}$, with distribution belonging to a natural multivariate exponential family, so that

$$f\left(\mathbf{y}_{ij} \mid \theta_{ij}\right) = \rho\left(\mathbf{y}_{ij}\right) \exp\left\{\theta_{ij}^T\mathbf{y}_{ij} - \psi\left(\theta_{ij}\right)\right\},$$

with $\theta_{ij}^T = \left(\theta_{ij1}, \ldots, \theta_{ijd}\right)$ now assumed to be $d-$dimensional. For $k = 1, \ldots, d$, let $g_k$ be known, monotone, differentiable functions such that

$$g_k\left(\theta_{ijk}\right) = \mathbf{X}_{ijk}\beta + \gamma_{ik} + \varepsilon_{ijk},\ k = 1, \ldots, d.$$

The change in prior specification for the higher-dimensional model is

$$\begin{aligned} \gamma_{ik} \mid \sigma_k^2 &\overset{ind}{\sim} N\left(0, \sigma_k^2\right), \\ \varepsilon_{ijk} \mid \sigma^2 &\overset{i.i.d}{\sim} N\left(0, \sigma^2\right), \\ \sigma^2 \mid a, b &\sim IG(a, b), \\ \pi\left(\beta, \sigma_1^2, \ldots, \sigma_d^2 \mid a_1, \ldots a_d\right) &\propto \prod_{k=1}^{d}\left(\frac{1}{\sigma_k^2}\right)^{a_k+1}. \end{aligned} \tag{5}$$

As before, let $\mathbf{P}_\mathbf{X} = \mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ and let $\mathbf{Z} = \oplus_{i=1}^{m}\mathbf{D}_i$ be a $d\sum_{i=1}^{m} n_i \times md$ matrix, where $\mathbf{D}_i = \left(\mathbf{I}_{d\times d} \cdots \mathbf{I}_{d\times d}\right)^T$ is a matrix consisting of $n_i$ identity matrices and $\oplus$ is the direct sum operator.

**Theorem 3.1.** *The following conditions are sufficient for the propriety of the posterior distribution of model* (5)*:*

1. $t \equiv rank\left(\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}\right) \geq (d-1)m$
2. $(d-1)m - t < 2a_k < 0$ *for all* $k = 1,\ldots,d$,
3. $w = d\sum_{i=1}^m n_i/2 + \sum_{k=1}^d a_k - p/2 + a > 0$,
4. $\int \left(\mathbf{G}^T\mathbf{W}\mathbf{G} + b\right)^{-w} \prod_{ijk} g_k'\left(\theta_{ijk}\right) \exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij} - \psi\left(\theta_{ij}\right)\right)\right\} d\theta < \infty$,

*where* $\mathbf{W} = 1/2\left(\mathbf{P}_\mathbf{X} - \mathbf{P}_\mathbf{X}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}\right)^- \mathbf{Z}\mathbf{P}_\mathbf{X}\right)$*, and* $\left(\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}\right)^-$ *is the generalized inverse of* $\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}$*. The conditions*

1. $-m/2 < a_k < 0$ *for all* $k = 1,\ldots,d$,
2. $w = d\sum_{i=1}^m n_i/2 + \sum_{k=1}^d a_k - p/2 + a > 0$, *and*
3. $\int \left(\mathbf{G}^T\mathbf{P}_\mathbf{X}\mathbf{G}/2 + b\right)^{-w} \prod_{ijk} g_k'\left(\theta_{ijk}\right) \exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij} - \psi\left(\theta_{ij}\right)\right)\right\} d\theta < \infty$

*are necessary.*

The reason for the choice of the proper inverse gamma prior for $\sigma^2$ in Models (3) and (5), rather than a noninformative improper of the class considered for the parameters $\sigma_k^2$, is that it is not clear that any improper prior will result in a proper posterior distribution, as can be seen by the following corollary.

**Corollary 3.2.** *If* $g_k(x) = x$ *for all k, and we change the specification in model* (3) *or model* (5) *so that an improper prior of the form* $\pi\left(\sigma^2 \mid a\right) \propto \left(\sigma^2\right)^{-a-1}$ *is used, then the posterior distribution will be improper for* $a \geq -\sum_{k=1}^d a_k$.

**Corollary 3.3.** *If the data are multinomial, so that*

$$\mathbf{Y}_{ij} \mid \theta_{ij} \sim MN\left(N_{ij}; p_{ij1},\ldots,p_{ijd},p_{ij(d+1)}\right),$$

*where* $p_{ijk} = e^{\theta_{ijk}}/\left(1 + \sum_{l=1}^d e^{\theta_{ijl}}\right)$ *for* $k = 1,\ldots,d$ *and* $p_{ij(d+1)} = 1 - \sum_{l=1}^d p_{ijl}$*, and* $g_k(x) = x$ *for all k, then condition 4 in Theorem 3.1 (or condition 5 in Theorem 2.6) is implied by the condition that* $Y_{ijk} > 0$ *for all triples* $(i,j,k)$ *and* $\sum_{k=1}^d Y_{ijk} < N_{ij}$ *for each pair* $(i,j)$.

*Remark* 3.4. A problem that is often encountered in the analysis of multinomial data is "complete separation of points," which results in a likelihood for which the maximum likelihood estimates do not exist and for which the posterior distribution is often improper when improper prior distributions are used (Natarajan and McCulloch, 1995). One consequence of the inclusion of the error terms $\varepsilon_{ijk}$ in the model is that, even if the data shows complete separation of points, we will still have a proper posterior distribution, so long as the conditions of Theorem 3.1 or Corollary 3.3 hold.

*Remark* 3.5. Mixed effects multinomial logistic regression models have been considered before by, for example, Hedeker (2003) and Speckman et al. (2008). Hedeker (2003) used these types of models with proper prior distributions to study observations that are clustered or repeatedly measured. Speckman et al. (2008) showed that propriety of multinomial models with flat prior distributions on the regression coefficients (and no random effects in the model) is equivalent to complete separation of points, which is, in turn, equivalent to the existence of the maximum likelihood estimator.

## 4. Example

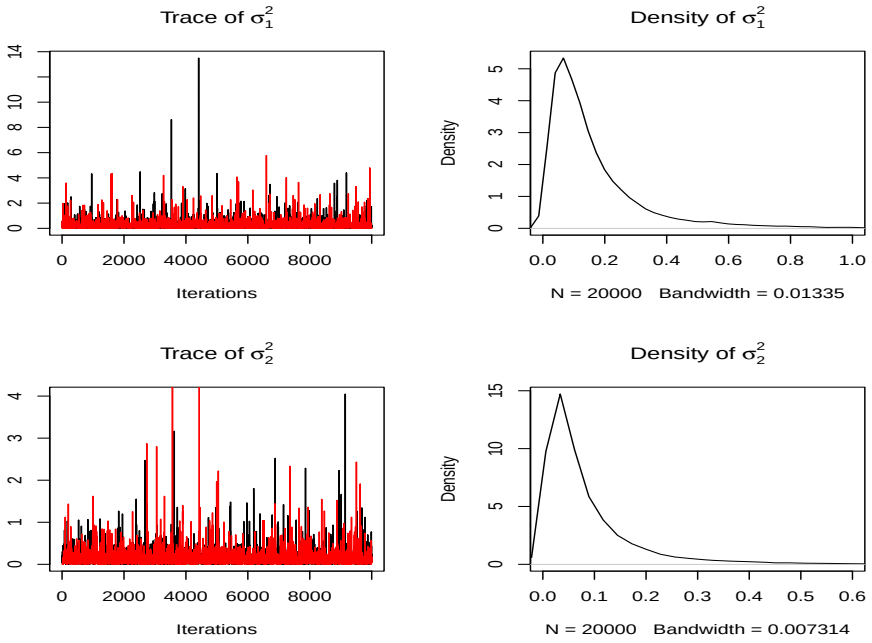
Due to confidentiality issues, data inputs for the SAHIE model cannot be released. Instead, for our example, we use a single simulated data set which has similar characteristics to the SAHIE data set. For this simulated data set, the overall sample size is large; however, for many individual small areas, the sample sizes are small, with a smallest sample size of 4. The median sample size is 206, and the 95th percentile of sample sizes is 1100.

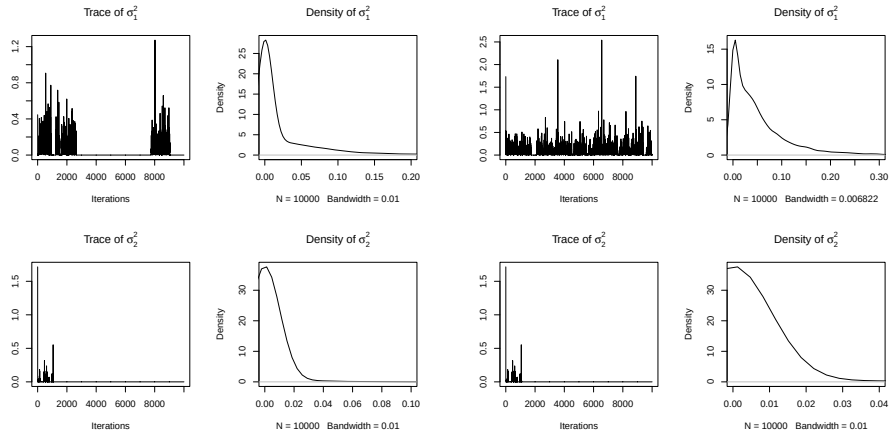
For our example, we model direct estimates of the proportion of people in each income category, the IPR categories $0-200\%$, $200\%-250\%$, and $>250\%$. We compare the bivariate model (3) with model (5) with the parameter $\rho$ set to 0, both using the multinomial likelihood. The covariate information used for the regression component of the model are main effects for age (0 – 18, 19 – 39, 40 – 49, and 50 – 64), race/ethnicity (Hispanic, White not Hispanic, Black not Hispanic, and Other not Hispanic), and gender. Estimates are made for each demographic subgroup for each of 7 states, so that there are a total of $2*4*4*2*7 = 224$ estimates to be made.

*Remark* 4.1. We note that there are important differences between the SAHIE model and the small area models that we consider in this paper. The SAHIE model is more complicated, and has more parameters. In addition, the SAHIE model is able to make use of administrative records, which strengthen estimates. The models we consider here can be thought of as simplified versions of the SAHIE model. See Bauder and Luery (2010) for a complete description of the SAHIE methodology.

We first investigate the behavior of the Markov chains for the two different models for different values of the hyperparameters. In all the examples that follow, we set $a = b = 0.01$, so that we are using a vague, but proper prior distribution for $\sigma^2$. All computations were done using R. For Model (5), we ran the Gibbs sampler for three cases: 1) $a_1 = a_2 = -1$, corresponding to improper uniform prior distributions on the variance components, leading to a proper posterior distribution (Theorem 3.1), 2) $a_1 = a_2 = 0$, corresponding to values of the hyperparameters that result in an improper posterior, but that are on the boundary of the

**Figure 1**: Trace and density plot when $a_1 = a_2 = -1$ for Model (5)






(a) Chain 1 (b) Chain 2

**Figure 2**: Sensitivity to starting value when $a_1 = a_2 = 0$ for Model (5)

parameter space of parameter values that lead to an improper posterior, and 3) $a_1 = a_2 = 1$, which are values which more clearly lead to an improper posterior distribution.

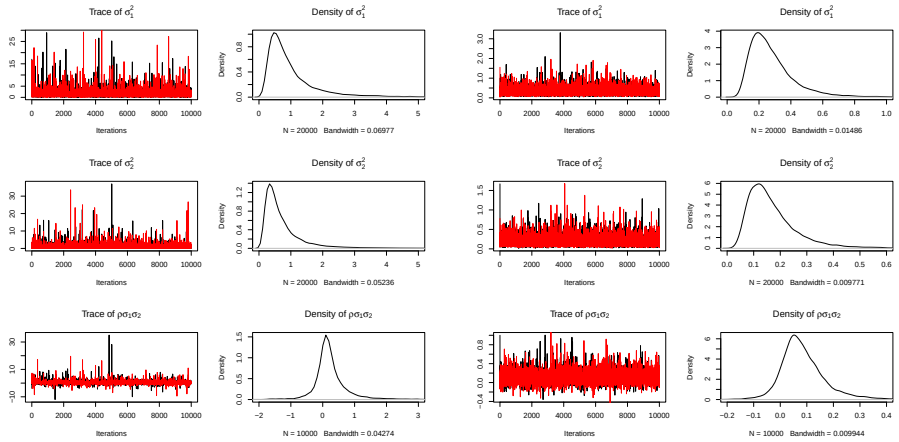
Figure 1 shows the trace and density plots for the parameters $\sigma_1^2$ and $\sigma_2^2$ when $a_1 = a_2 = -1$, based on two separate chains with randomly chosen starting values. Since we know that these values of $a_1$ and $a_2$ result in a proper posterior distribution, we can use this figure as a baseline for comparison.

Figure 2 shows the trace and density plots for the parameters $\sigma_1^2$ and $\sigma_2^2$ when $a_1 = a_2 = 0$. Figures 2(a) and 2(b) are the trace and density plots for the Gibbs sampler when two different randomly chosen starting values were used. By Theorem 3.1, these values of the hyperparameters result in an improper posterior distribution. However, the necessary and sufficient conditions for propriety given by Theorem 3.1 are that $a_1$ and $a_2$ are less than 0, so that these values are as close as possible to values that will give us a proper posterior, while still resulting in an improper posterior.

It is clear from the behavior of the trace plots of Figure 2 that something is wrong. The chain for $\sigma_1^2$ shown in Figure 2(b) exhibits seemingly reasonable behavior. However, the chain for $\sigma_1^2$ shown in Figure 2(a) gives a clearer indication of bad behavior. The chain has several regions where it seems to get stuck in a neighborhood of 0, followed by jumps into a more reasonable region of the parameter space. This shows the sensitivity of the analysis to the starting values. Both chains for $\sigma_2^2$ did seem to concentrate near 0 without the jumps shown in the chains for $\sigma_1^2$.

When we set the values of the hyperparameters to 1, lack of convergence to a proper limiting distribution was clear. After a number of iterations of the Gibbs sampler, the values of $\sigma_1^2$ and $\sigma_2^2$ settled in a neighborhood of 0, so that $\sigma_k^2$ became computationally indistinguishable from 0, and the algorithm crashed. In this example, propriety or impropriety of the posterior was clearly evident from the results of the Gibbs sampler for different values of the hyperparameters for model (5). However, the story was not the same when we repeated this procedure for Model (3). We ran the Gibbs sampler two times, with values of the hyperparameters set to $a_1 = a_2 = a_3 = 0$ for the first run, and $a_1 = a_2 = a_3 = 7/2$ for the second run.

Figure 3(a) shows the trace plot and the density plot for two chains with randomly chosen starting values (the first shown in black, and the second in red) for $\sigma_1^2$, $\sigma_2^2$, and $\rho\sigma_1\sigma_2$ when we use Model (3) with $a_1 = a_2 = a_3 = 0$. By Theorem 2.6, these values of $a_k$ lie on



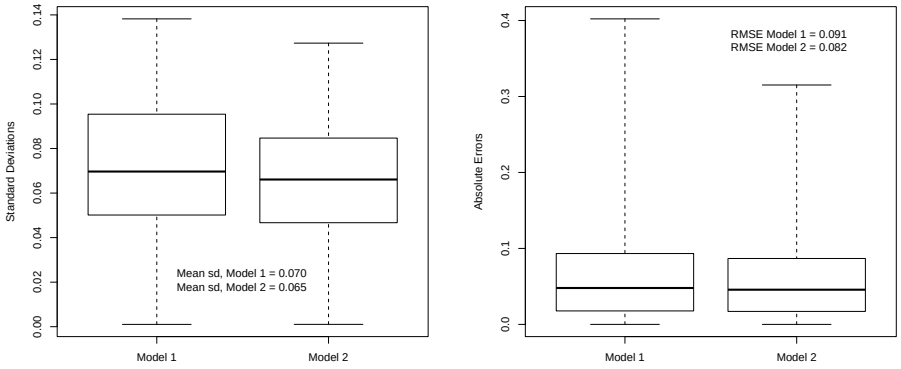

(a) $a_1 = a_2 = a_3 = 0$ (b) $a_1 = a_2 = a_3 = 7/2$

**Figure 3**: Gibbs sampling chains for improper posterior distributions for Model (5)

the boundary of the parameter space separating the hyperparameters which correspond to a proper posterior distribution with those corresponding to an improper posterior distribution.

A visual inspection of Figure 3(a) gives no indication that there may be a problem with the analysis based on this model specification. In addition, the usual convergence diagnostics, such as the Geweke statistic and the Heidelberger and Welch statistic (Robert and Casella, 2004; Plummer et al., 2010) indicate "convergence" of the Markov chain to the stationary distribution, so that it is not clear how to detect lack of propriety through the output of the Gibbs sampler in this case.

Figure 3(b) shows the trace and density plots when $a_1 = a_2 = a_3 = 7/2$, which, by Theorem 2.6 is more clearly in the part of the parameter space which results in an improper posterior distribution. Still, we do not see any clear lack of convergence through the plots or the convergence statistics. What we do see in comparing Figures 3(b) to 3(a) is that the samples for $\sigma_1^2$ and $\sigma_2^2$ in the second case are more tightly concentrated around 0 than they are in the first case. It appears that the inclusion of the parameter $\rho$ holds together the output for $\sigma_1^2$ and $\sigma_2^2$ enough, at least in the first 10000 iterations of the Gibbs chain, to




(a) (b)

**Figure 4**: Comparison of models (3) and (5)

prevent either $\sigma_1^2$ or $\sigma_2^2$ from diverging quickly, making it more difficult to detect lack of convergence or lack of propriety of the posterior distribution.

Finally, we do a comparison of the two models to try to understand if the inclusion of the parameter $\rho$ is important for our analysis. The inclusion of $\rho$ makes Model (3) more complicated than Model (5), makes the conditions for propriety of the posterior distribution less clear (compare Theorem 2.6 to Theorem 3.1) and, as the previous examples suggest, makes the Markov chains from the Gibbs sampler less indicative of poor convergence when the posterior distribution is truly improper.

Figure 4(a) compares the box plot of the 224 estimated posterior standard deviations for the main parameter of interest $p_{ij}$, the proportion in income category $i$ in age / race / sex / region category $j$ for each of the two models. The average standard deviation for Model 3 is 0.065, while the average standard deviation for Model 5 is 0.070. Perhaps more importantly, the 75th and 95th percentiles are significantly reduced when we use Model 3.

Figure 4(b) compares the box plot of the 224 absolute errors for the estimated proportions $p_{ij}$ using the two different models. Model (3) reduces the root mean squared error from 0.091 to 0.082, and also lowers the 75th and 95th percentiles compared to Model (5). Taken together, Figure 4 suggest that despite additional analytic difficulties, inclusion of the correlation parameter $\rho$ can be useful for improving the analysis.

## 5. Appendix: Proofs

**Lemma 5.1** (Fiedler 1971)**.** *Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be* $n \times n$ *positive semidefinite matrices and let* $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A}) \geq 0$ *and* $\lambda_1(\mathbf{B}) \geq \cdots \geq \lambda_n(\mathbf{B}) \geq 0$ *be the eigenvalues of* $\mathbf{A}$ *and* $\mathbf{B}$*. Then*

$$\prod_{i=1}^{n} (\lambda_i(\mathbf{A}) + \lambda_i(\mathbf{B})) \leq |\mathbf{A}+\mathbf{B}| \leq \prod_{i=1}^{n} (\lambda_i(\mathbf{A}) + \lambda_{n-i+1}(\mathbf{B})).$$

**Lemma 5.2.** *The eigenvalues of the inverse of the covariance matrix*

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

*satisfy*

$$0 \leq \lambda_2(\Sigma) \leq \frac{\min(\sigma_1^2, \sigma_2^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)} \leq \frac{\max(\sigma_1^2, \sigma_2^2)}{\sigma_1^2\sigma_2^2(1-\rho^2)} \leq \lambda_1(\Sigma) \leq \frac{\sigma_1^2+\sigma_2^2}{\sigma_1^2\sigma_2^2(1-\rho^2)}.$$

*Proof.* Straightforward. □

*Proof of Theorem 2.6.* We use the notation $[x \mid y]$ to represent a general conditional density of a random variable $X$ given $Y$. The joint posterior distribution is

$$\left[\theta, \beta, \gamma, \Sigma, \sigma^2 \mid \mathbf{Y}\right] = [\mathbf{Y} \mid \theta]\left[\theta \mid \beta, \gamma, \sigma^2\right]\left[\gamma \mid \Sigma, \sigma^2\right][\beta, \Sigma \mid a_1, a_2, a_3]\left[\sigma^2 \mid a, b\right]/m(\mathbf{Y})$$

where

$$m(\mathbf{Y}) = \int \left[\theta, \beta, \gamma, \Sigma, \sigma^2 \mid \mathbf{Y}\right] d\beta\, d\gamma\, d\Sigma\, d\sigma^2\, d\theta.$$

Clearly, the joint posterior distribution is proper if and only if $m(\mathbf{Y})$ is finite, so we need to

check that the function

$$\exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij}-\psi(\theta_{ij})\right)\right\}\left(\prod_{ijk}g_k'\left(\theta_{ijk}\right)\right)\left(\frac{1}{\sigma_1^2}\right)^{a_1+1}\left(\frac{1}{\sigma_2^2}\right)^{a_2+1}\left(\frac{1}{1-\rho^2}\right)^{a_3+1}$$
$$\times\left(\frac{1}{\sigma^2}\right)^{\Sigma_i n_i+a+1}\exp\left\{-\frac{1}{2\sigma^2}\sum_{ijk}\left(g_k\left(\theta_{ijk}\right)-\mathbf{X}_{ijk}\beta-\gamma_{ik}\right)^2\right\}$$
$$\times|\Sigma|^{-m/2}\exp\left\{-\frac{b}{\sigma^2}\right\}\exp\left\{-\frac{1}{2}\sum_{i=1}^m\gamma_i^T\Sigma^{-1}\gamma_i\right\}$$

is integrable. To simplify notation, we write

$$\sum_{ijk}\left(g_k\left(\theta_{ijk}\right)-\mathbf{X}_{ijk}\beta-\gamma_{ik}\right)^2=\left(\mathbf{G}-\mathbf{X}\beta-\mathbf{Z}\gamma\right)^T\left(\mathbf{G}-\mathbf{X}\beta-\mathbf{Z}\gamma\right),$$

where $\mathbf{G}\in\mathbb{R}^{2\Sigma_{i=1}^m n_i}$ and $\gamma\in\mathbb{R}^{2m}$ are vectors with elements $g\left(\theta_{ijk}\right)$ and $\gamma_{ik}$, respectively.

The calculations are similar to those used in the proof of Theorem 1 in Hobert and Casella (1996), so we give only a sketch of the proof. First, it is straightforward to show that

$$\begin{aligned}\iint\exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{G}-\mathbf{X}\beta-\mathbf{Z}\gamma\right)^T\left(\mathbf{G}-\mathbf{X}\beta-\mathbf{Z}\gamma\right)-\frac{1}{2}\sum_{i=1}^m\gamma_i^T\Sigma^{-1}\gamma_i\right\}d\beta\,d\gamma\\ \propto\left(\sigma^2\right)^{m+p/2}|\mathbf{M}|^{-1/2}\exp\left\{-\frac{1}{2\sigma^2}\mathbf{G}^T\left(\mathbf{P}_\mathbf{X}-\mathbf{P}_\mathbf{X}\mathbf{Z}\mathbf{M}^{-1}\mathbf{Z}^T\mathbf{P}_\mathbf{X}\right)\mathbf{G}\right\},\end{aligned}\tag{6}$$

where $\mathbf{M}=\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}+\sigma^2\widetilde{\Sigma}$ and $\widetilde{\Sigma}=\oplus_{i=1}^m\Sigma$.

Next, note that $\widetilde{\Sigma}$ is positive definite and $\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}$ is positive semidefinite, so

$$\sigma^2\widetilde{\Sigma}^{-1}+\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}\geq\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}\geq\mathbf{0},$$

which implies that

$$\begin{aligned}\exp\left\{-\frac{1}{2\sigma^2}\mathbf{G}^T\mathbf{P}_\mathbf{X}\mathbf{G}\right\}&\leq\exp\left\{-\frac{1}{2\sigma^2}\mathbf{G}^T\left(\mathbf{P}_\mathbf{X}-\mathbf{P}_\mathbf{X}\mathbf{Z}\mathbf{M}^{-1}\mathbf{Z}^T\mathbf{P}_\mathbf{X}\right)\mathbf{G}\right\}\\&\leq\exp\left\{-\frac{1}{2\sigma^2}\mathbf{G}^T\mathbf{W}\mathbf{G}\right\},\end{aligned}\tag{7}$$

where $\mathbf{W}=\mathbf{P}_\mathbf{X}-\mathbf{P}_\mathbf{X}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{P}_\mathbf{X}$.

Let $\lambda_1$ be the largest eigenvalue of $\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}$ and $\lambda_2$ the smallest eigenvalue of $\mathbf{Z}^T\mathbf{P}_\mathbf{X}\mathbf{Z}$, which is assumed to be non-zero. Using Lemma 5.1 and Lemma 5.2 we obtain the following bounds on $|\mathbf{M}|$:

$$\begin{aligned}\lambda_2^m\left(\frac{\sigma^2\max\left(\sigma_1^2,\sigma_2^2\right)}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}+\lambda_2\right)^m&\leq|\mathbf{M}|\\&\leq\left(\frac{\sigma^2\min\left(\sigma_1^2,\sigma_2^2\right)}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}+\lambda_1\right)^m\left(\frac{\sigma^2\left(\sigma_1^2+\sigma_2^2\right)}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}+\lambda_1\right)^m.\end{aligned}\tag{8}$$

Sufficient conditions for the propriety of the joint posterior distribution can be obtained

by combining equations (6), (7), and (8) to get the following upper bound:

$$\begin{aligned}\left[\theta,\Sigma,\sigma^2 \mid \mathbf{Y}\right] \leq C\left(\sigma^2\right)^{-\sum_{i=1}^m n_i+p/2+m-a-1}\exp\left\{-\frac{b}{\sigma^2}\right\}\left(\prod_{ijk} g_k'\left(\theta_{ijk}\right)\right)\\ \times\exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij}-\psi\left(\theta_{ij}\right)\right)\right\}\exp\left\{-\frac{1}{2\sigma^2}\mathbf{G}^T\mathbf{W}\mathbf{G}\right\}\left(\sigma_1^2\right)^{-m/2-a_1-1}\\ \times\left(\sigma_2^2\right)^{-m/2-a_2-1}\left(1-\rho^2\right)^{-m/2-a_3}\left(\frac{\sigma^2\max\left(\sigma_1^2,\sigma_2^2\right)}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}+\lambda_2\right)^{-m/2}.\end{aligned}$$

The integral

$$\begin{aligned}\int_{-1}^1\int_0^\infty\int_0^\infty\left(\frac{1}{1-\rho^2}\right)^{m/2+a_3+1}\left(\frac{1}{\sigma_1^2}\right)^{m/2+a_1+1}\left(\frac{1}{\sigma_2^2}\right)^{m/2+a_2+1}\\ \times\left(\frac{\sigma^2\max\left(\sigma_1^2,\sigma_2^2\right)}{\sigma_1^2\sigma_2^2\left(1-\rho^2\right)}+\lambda_2\right)^{-m/2}d\sigma_1^2\,d\sigma_2^2\,d\rho\end{aligned}$$

can be calculated by considering separately the regions $\sigma_1^2>\sigma_2^2$ and $\sigma_1^2<\sigma_2^2$, and is finite and of order $\left(\sigma^2\right)^{-m-a_1-a_2}$ if and only if $-m/2<a_1$, $-m/2<a_2$, $a_1+a_2<-m/2$, and $m/2+a_2+a_2-a_3>0$. This gives

$$\begin{aligned}\left[\theta,\sigma^2 \mid \mathbf{Y}\right] \leq C\left(\prod_{ijk} g_k'\left(\theta_{ijk}\right)\right)\exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij}-\psi\left(\theta_{ij}\right)\right)\right\}\\ \times\left(\sigma^2\right)^{-\sum_{i=1}^m n_i+p/2-a_1-a_2-a-1}\exp\left\{-\frac{b}{\sigma^2}\right\}\exp\left\{-\frac{1}{2\sigma^2}\mathbf{g}^T\mathbf{W}\mathbf{g}\right\}.\end{aligned}\tag{9}$$

As a function of $\sigma^2$, the right hand side of equation (9) is the kernel of an inverse gamma density, so long as $w=\sum_{i=1}^m n_i-p/2+a_1+a_2+a>0$. Hence

$$\left[\theta \mid \mathbf{Y}\right] \leq C\left(\mathbf{G}^T\mathbf{W}\mathbf{G}/2+b\right)^{-w}\left(\prod_{ijk} g_k'\left(\theta_{ijk}\right)\right)\exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij}-\psi\left(\theta_{ij}\right)\right)\right\}.$$

We therefore require that

$$\int\left(\mathbf{G}^T\mathbf{W}\mathbf{G}/2+b\right)^{-w}\left(\prod_{ijk} g_k'\left(\theta_{ijk}\right)\right)\exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{T}_{ij}-\psi\left(\theta_{ij}\right)\right)\right\}d\theta<\infty.$$

Necessary conditions for the propriety of the joint posterior distribution can be found by first using equations (6), (7), and (8) to get a lower bound for $\left[\theta,\Sigma,\sigma^2 \mid \mathbf{Y}\right]$, and then repeating the previous arguments used to prove sufficiency. □

The proof of Theorem 3.1 is similar to the proof of Theorem 2.6 so is omitted.

*Proof of Corollary 3.2.* With the change in model specification, condition 3 in the set of necessary conditions of Theorem 3.1 becomes

$$\begin{aligned}\int\left(\theta^T\mathbf{P}_\mathbf{X}\theta\right)^{-w}\exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{Y}_{ij}-\psi\left(\theta_{ij}\right)\right)\right\}d\theta\\ \equiv\int\left(\theta^T\mathbf{P}_\mathbf{X}\theta\right)^{-w}\exp\left\{\theta^T\mathbf{Y}-\widetilde{\psi}\left(\theta\right)\right\}d\theta<\infty.\end{aligned}$$

The matrix $\mathbf{P_X}$ is symmetric, idempotent, and of rank $q = d\sum_{i=1}^{m} n_i - p$. We can therefore use the decomposition $\mathbf{P_X} = \mathbf{P}\Lambda\mathbf{P}^T$, where $\mathbf{P}$ is an orthogonal matrix and $\Lambda$ is a diagonal matrix with the first q diagonal elements equal to 1 and the remaining diagonal elements equal to 0. Let $\mathbf{B}$ be a compact neighborhood of $\mathbf{0}$. Making the change of variables $\theta = \mathbf{P}\xi$, the integral is proportional to

$$\begin{aligned} I &= \int (\xi\Lambda\xi)^{-w} \exp\left\{\xi^T\mathbf{P}^T\mathbf{Y} - \widetilde{\psi}(\mathbf{P}\xi)\right\} d\xi \\ &= \int \left(\xi_1^2 + \cdots + \xi_q^2\right)^{-w} \exp\left\{\xi^T\mathbf{P}^T\mathbf{y} - \widetilde{\psi}(\mathbf{P}\xi)\right\} d\xi \\ &\geq \int_{\mathbf{B}} \left(\xi_1^2 + \cdots + \xi_q^2\right)^{-w} \exp\left\{\xi^T\mathbf{P}^T\mathbf{y} - \widetilde{\psi}(\mathbf{P}\xi)\right\} d\xi \\ &\geq \left(\inf_{\xi\in\mathbf{B}} \exp\left\{\xi^T\mathbf{Y} - \widetilde{\psi}(\mathbf{P}\xi)\right\}\right) \int_{\mathbf{B}} \left(\xi_1^2 + \cdots + \xi_q^2\right)^{-w} d\xi \\ &= C\int_{\mathbf{B}} \xi_1^{-2w} \left(1 + \left(\frac{\xi_2}{\xi_1}\right)^2 + \cdots + \left(\frac{\xi_q}{\xi_1}\right)^2\right)^{-w} d\xi. \end{aligned}$$

We make the transformation $u_1 = \xi_1$ and $u_i = \xi_i/\xi_1$ for $i = 2,\ldots,q$. Let $\widetilde{\mathbf{B}}$ be a neighborhood of $\mathbf{0}$ that is a subset of the transformed space. The Jacobian of this transformation is $u_1^{q-1}$, which leads to

$$I \geq C\int_{\widetilde{\mathbf{B}}} u_1^{-2w+q-1} \left(1 + u_2^2 + \cdots + u_q^2\right)^{-w} d\mathbf{u}.$$

This integral diverges if $-2w + q \leq 0$, or equivalently if $a \geq -\sum_{k=1}^{d} a_k$. □

*Proof of Corollary 3.3.* We need to check that

$$\int \left(\theta^T\mathbf{W}\theta + b\right)^{-w} \exp\left\{\sum_{ij}\left(\theta_{ij}^T\mathbf{Y}_{ij} - N_{ij}\log\left(1 + \sum_{k=1}^{d} e^{\theta_{ijk}}\right)\right)\right\} d\theta < \infty.$$

Since $w > 0$ and $\mathbf{W}$ is non-negative definite, we can use the bound $\left(\theta^T\mathbf{W}\theta + b\right)^{-w} \leq b^{-w}$, and check that

$$\int \exp\left\{\theta_{ij}^T\mathbf{Y}_{ij} - N_{ij}\log\left(1 + \sum_{k=1}^{d} e^{\theta_{ijk}}\right)\right\} d\theta_{ij} < \infty$$

for each pair $(i,j)$. Making the change of variables $p_{ijk} = e^{\theta_{ijk}}/\left(1 + \sum_{l=1}^{d} e^{\theta_{ijl}}\right)$ for $k = 1,\ldots,d$ gives (suppressing the subscript $(i,j)$)

$$\int\limits_{p_k>0, \sum_{k=1}^{d} p_k<1} p_1^{Y_1-1}\cdots p_d^{Y_d-1}\left(1 - p_1 - \cdots - p_d\right)^{N-\sum Y_i-1} d\mathbf{p},$$

which is finite so long as $Y_i > 0$ and $\sum_{i=1}^{d} Y_i < N$, since then the integrand is the kernel of a Dirichlet distribution. □

## References

Bauder, M. and Luery, D. (2010), "Small area estimation of health insurance coverage in 2007," Tech. rep., Small Area Methods Branch, Data Integration Division, U. S. Census Bureau, available at `http://www.census.gov/did/www/sahie/methods/20062007/index.html`.

Berger, J. O. and Sun, D. (2008), "Objective priors for the bivariate normal model," *Ann. Statist.*, 36, 963 – 982.

Datta, G. S., Day, B., and Maiti, T. (1998), "Multivariate Bayesian small area estimation: an application to survey and satelite data," *Sankhyā, Ser. A*, 60, 344 – 362.

Fay, R. E. (1987), "Application of multivariate regression to small domain estimation," in *Small Area Statistics*, eds. Platek, R., Rao, J. N. K., Särndal, C. E., and Singh, M. P., New York: Wiley, pp. 91 – 102.

Fiedler, M. (1971), "Bounds for the determinant of the sum of Hermitian matrices," *Proc. Amer. Math. Soc.*, 30, 27 – 31.

Fisher, R. (2003), "Errors-in-variables model for county-level poverty estimation," Tech. rep., Housing and Household Economic Statistics Division, U. S. Census Bureau, available at `http://www.census.gov/did/www/saipe/publications/files/tech.report.5.pdf`.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models," *Bayesian Anal.*, 1, 515 – 533.

Ghosh, M., Natarajan, K., Stroud, T. W. F., and Carlin, B. P. (1998), "Generalized linear models for small-area estimation," *J. Amer. Statist. Assoc.*, 93, 273 – 282.

Gilks, W. R. and Wild, P. (1992), "Adaptive rejection sampling for Gibbs sampling," *Appl. Statist.*, 41, 337 – 348.

Hedeker, D. (2003), "A mixed-effects multinomial logistic regression model," *Statist. Med.*, 22, 1433 – 1446.

Hobert, J. P. and Casella, G. (1996), "The effect of improper priors on Gibbs sampling in Hierarchical linear mixed models," *J. Amer. Statist. Assoc.*, 91, 1461 – 1473.

Natarajan, R. and Kass, R. E. (2000), "Reference Bayesian methods for generalized linear mixed models," *J. Amer. Statist. Assoc.*, 95, 227 – 237.

Natarajan, R. and McCulloch, C. E. (1995), "A note on the existence of the posterior distribution for a class of mixed models for binomial responses," *Biometrika*, 82, 639 – 643.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2010), *coda: Output analysis and diagnostics for MCMC*, r package version 0.13-5.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rao, J. N. K. (2003), *Small Area Estimation*, Hoboken, New Jersey: Wiley.

Robert, C. P. and Casella, G. (2004), *Monte Carlo Statistical Methods*, New York, New York: Springer, 2nd ed.

Speckman, P. L., Lee, J., and Sun, D. (2008), "Existence of the MLE and propriety of posterior for a general multinomial choice model," *Statist. Sinica*, 19, 731 – 748.

Sun, D., Tsutakawa, R. K., and He, Z. (2001), "Propriety of posteriors with improper priors in hierarchical linear mixed models," *Statist. Sinica*, 11, 77 – 95.