Technical Report RM 680, Department of Statistics and Probability, Michigan State University

Default Bayesian Analysis for Hierarchical Spatial Multivariate Models

Sarat C. Dass, Chae Young Lim and Tapabrata Maiti*
Department of Statistics & Probability
Michigan State University, East Lansing, MI 48824
Email: {sdass,lim,maiti}@stt.msu.edu

Abstract

In recent years, multivariate spatial models have been proven to be an effective tool for analyzing spatially related multidimensional data arising from a common underlying spatial process. Currently, the Bayesian analysis is perhaps the only solution available in this framework where prior selection plays an important role in the inference. The present article contributes towards the development of Bayesian inferential methodology for hierarchical spatial multivariate generalized linear mixed models. The two main contributions of this article are the development of a shrinkage-type default prior and innovative computational techniques for the Gibbs sampling implementation. The default prior elicitation is non-informative but results in a proper posterior on the related parameter spaces. This elicitation not only provides robust inference (with respect to prior choice), but also provides improved estimation. In the computational step, we have developed a transformation of the parameters that avoids sampling from restricted domains, thus providing more stability and efficiency in the Gibbs implementation. The methodology has been extended to the case of missing responses in the multi-dimensional setup. Both simulations and real examples are provided to validate and illustrate the proposed methodology.

Keywords: Generalized linear mixed models, conditional autoregressive models, default Bayesian analysis, health disparity

*Authors' names are in alphabetical order.

1 Introduction

Many spatial problems, particularly those concerning environmental, health and socio-economic variables, are inherently multivariate, meaning that two or more such variables are recorded at each spatial location simultaneously. Multivariate spatial analysis is becoming more relevant and prevalent with the advent of geographic information systems (GIS) that allow the display of spatial data at varying spatial resolutions. Sain and Cressie (2007) viewed the developments of spatial analysis in two main categories: models for geostatistical data (that is, the indices of data points belong in a continuous set) and models for lattice data (data with indices in a discrete or countable set), while specifically mentioning that the latter is not as developed.

The issue of "health disparity" is central to the distribution of federal and state aid based on socio-economic indicators. Health disparity studies analyze how health status of individuals vary across various socio-economic groups and spatial locations, in particular in relation to a specific disease. Multiple response variables are available as indicators of health status, and as a result, models for multivariate spatial lattice data are an indispensable tool for analyzing health disparity data. Recently, Sain and Cressie (2007), Kim *et al.* (2001), Gelfand and Vounatsou (2003), Jin *et al.* (2005) explored multivariate spatial models for lattice data, adopting the Bayesian framework as the natural inferential approach. The only exception, Sain (2009) developed the maximum likelihood estimation procedure for a special case, namely the multivariate conditional autoregressive (CAR) normal model of Sain and Cressie (2007).

Although the Bayesian inferential framework is a natural choice for spatial lattice data, one obvious obstacle is the choice of prior distribution for the model parameters. All previous works mentioned above are based on standard subjective, and at best, vague priors to account for the lack of subjective knowledge. Subjective specification of priors have the obvious drawback of introducing bias in the estimation procedure, the extent of which may not be easy to gauge in applications. A *default*, or non-informative prior, is therefore preferable for the Bayesian approach. However, in the case of hierarchical generalized linear models as in this paper, putting non-informative priors (i.e., improper priors) result in the posterior being improper. This led Natarajan and Kass (2000) to develop default priors in the context of generalized linear mixed models (GLMM) for which the posterior can be established to be proper.

A generalized linear mixed model (GLMM) with a multivariate Gaussian CAR model can be viewed as a special case of a general random effects model with specific restrictions on the structure of the covariance matrix. The Natarajan-Kass prior neither takes into account the special structure of the covariance matrix nor considers the dimension reduction capabilities

of a CAR model in the spatial case. In fact, some simple modifications greatly reduce the number of parameters while maintaining flexibility in modeling (Sain and Cressie, 2007).

This work is motivated from an application of joint mapping of lung cancer mortality incidence and poverty for all counties in the state of Michigan. The questions we seek to address are two-fold: First, whether health and socio-economic disparities are correlated, and second, if so, are the correlations stronger in certain spatial locations compared to others. In this paper, we develop a new default prior for the multivariate spatial lattice data in the context of spatial multivariate GLMM. The standard analysis using subjective and vague priors indicates significant prior sensitivity compared to the proposed prior, which justifies the use of the latter in real applications. Our Bayesian computational scheme is different from previous approaches. We adopt a new parametrization of the multivariate CAR model based on the Cholesky and spectral decompositions of matrices. The enormous advantage gained through this re-parametrization is the removal of constraints on the parameter space induced by the positive definiteness requirement on the inverse covariance matrix. The Bayesian inferential procedure is illustrated through simulation and an application based on SEER data for the state of Michigan.

The rest of the article is organized as follows: Section 2 develops the multivariate GLMM model in the spatial context. Section 3 discusses the motivation for developing the proposed default prior by taking the spatial information into account. The resulting posterior distribution is shown to be proper in this section for both complete and missing data cases. The Gibbs steps are outlined in Section 4 whereas Section 5 gives the numerical findings for both simulated and real data. This is followed by a brief discussion summarizing our findings in Section 6 and the Appendix.

2 Multivariate Generalized Linear Mixed Models

In what follows, we assume that there are n distinct sites on a spatial domain where observations on p variables are recorded. Let the multivariate data consist of the p-dimensional random vector $\mathbf{y}_j \equiv (y_{1j}, y_{2j}, \cdots, y_{pj})'$ for the j-th site, for $j = 1, 2, \cdots, n$. Corresponding to the response y_{ij}, denote by $x_{ij} = (x_{ij1}, x_{ij2}, x_{ijq_i})'$ to be the $q_i \times 1$ vector of explanatory variables. The following two stage hierarchical model is considered for the distribution of the $np \times 1$ vector of all observables $\mathbf{y} \equiv (\mathbf{y}_1', \mathbf{y}_2', \cdots, \mathbf{y}_n')'$: In the first stage of the hierarchy, the variables y_{ij} are independent with density

$$f_{ij}(y \,|\, \eta_{ij}) = C(y) \exp\{\, \eta_{ij} y - h_i(\eta_{ij})\}, \tag{1}$$

where f_{ij} belongs to an exponential family of densities with canonical parameter η_{ij}, and $h_i(\eta_{ij})$ is the normalizing constant satisfying

$$\exp\{\, h_i(\eta_{ij}) \,\} = \int C(y) \exp\{\, \eta_{ij} y \,\} \, dy.$$

It can be shown (see, for example, McCullagh and Nelder (1989)) that $h_i(\cdot)$ is a differentiable function with inverse h_i^{-1}. In the second stage, the canonical parameter η_{ij} is related to x_{ij} via a link function in the usual generalized linear model set-up,

$$\eta_{ij} = x'_{ij}\beta_i + \epsilon_{ij} \tag{2}$$

where β_i is a $q_i \times 1$ vector of regression coefficients and ϵ_{ij} are error random variables. The hierarchical specification is completed by eliciting the distribution for the error component ϵ_{ij}s, namely, $\boldsymbol{\epsilon} \sim N_{np}(0, \boldsymbol{D})$ where $\boldsymbol{\epsilon}_j \equiv (\epsilon_{1j}, \epsilon_{2j}, \cdots, \epsilon_{pj})'$ is the $p \times 1$ error vector at the j-th spatial site, $\boldsymbol{\epsilon} \equiv (\boldsymbol{\epsilon}'_1, \boldsymbol{\epsilon}'_2, \cdots, \boldsymbol{\epsilon}'_n)'$ is the $np \times 1$ vector of all the error variables and $\boldsymbol{D}$ is the covariance matrix (of dimension $np \times np$) of $\boldsymbol{\epsilon}$. Such models based on an unstructured $\boldsymbol{D}$ are called a random-effects GLMs or generalized linear mixed models (GLMMs), and has been the subject of many theoretical as well as practical studies in recent years.

2.1 Multivariate Gaussian CAR

In the spatial context, the distribution of $\boldsymbol{\epsilon}$, and hence $\boldsymbol{D}$, can be given a more concrete structure based on neighboring dependencies. Following Mardia (1988), we define the multivariate Gaussian CAR model as follows: For some $p \times p$ matrices $\boldsymbol{\Gamma}_j$ and $\boldsymbol{\Lambda}_{jk}$, suppose that the vector $\boldsymbol{\epsilon}_j$ is p-variate Gaussian with

$$E(\boldsymbol{\epsilon}_j \,|\, \boldsymbol{\epsilon}_{-j}) = \sum_{k \in N_j} \boldsymbol{\Lambda}_{jk} \, \boldsymbol{\epsilon}_k, \quad \text{and} \quad Var(\boldsymbol{\epsilon}_j \,|\, \boldsymbol{\epsilon}_{-j}) = \boldsymbol{\Gamma}_j, \quad \text{for } j = 1, 2, \cdots, n, \tag{3}$$

where $\boldsymbol{\epsilon}_{-j} \equiv \{\, \boldsymbol{\epsilon}_k \,:\, k \in N_j \,\}$ denotes the rest of the $\boldsymbol{\epsilon}$ at all locations k in the neighborhood N_j. The existence of the joint distribution for $\boldsymbol{\epsilon}$ is guaranteed by the symmetry and positive definiteness of $\boldsymbol{D}$ which, respectively, translates to

$$\boldsymbol{\Lambda}_{jk} \, \boldsymbol{\Gamma}_k = \boldsymbol{\Gamma}_j \, \boldsymbol{\Lambda}'_{kj} \tag{4}$$

for all pairs (j, k), and Block$(-\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Lambda}_{jk})$ is positive definite with $\boldsymbol{\Lambda}_{jj} = -\mathbf{I}$. Then, from Mardia (1988), $\boldsymbol{\epsilon}$ is $N_{np}(\mathbf{0}, \boldsymbol{D})$ with

$$\boldsymbol{D} = \{\text{Block}(-\boldsymbol{\Gamma}_j^{-1}\boldsymbol{\Lambda}_{jk})\}^{-1}. \tag{5}$$

We introduce some more notation here relevant to the spatial context. The weight matrix $\boldsymbol{W} = ((w_{jk}))$ (of dimension $n \times n$) consists of entries

$$w_{jk} = \begin{cases} 1 & \text{if } j \text{ and } k \text{ are neighbors, and} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

with $w_{jj} \equiv 0$ for $j, k = 1, 2, \cdots, n$. For each site j, define $w_{j+} \equiv \sum_{k \in N_j} w_{jk}$ to be the sum that represents the total number of neighbors of site j, and $\boldsymbol{M}$ to be the $n \times n$ diagonal matrix

$$\boldsymbol{M} = \text{diag}(w_{1+}, w_{2+}, \cdots, w_{n+}). \tag{7}$$

It is convenient both computationally as well as for practical analysis to represent all the $\boldsymbol{\Gamma}_j$ and $\boldsymbol{\Lambda}_{jk}$ in terms of two common $p \times p$ matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{H}$, respectively. The first stage of parametrization is to take

$$\boldsymbol{\Gamma}_j = \frac{\boldsymbol{\Gamma}}{w_{j+}} \quad \text{and} \quad \boldsymbol{\Lambda}_{jk} = \frac{w_{jk}}{w_{j+}} \cdot \boldsymbol{H}. \tag{8}$$

The first part of equation (8) entails a common covariance matrix $\boldsymbol{\Gamma}$ for all sites j, rescaled by the weight factor w_{j+}^{-1}. It follows that sites with a larger number of neighbors will have lower variability, which is reasonable to expect. The second part of equation (8) entails a common set of regression coefficients $\boldsymbol{H}$ rescaled by the inverse of the number of neighbors w_{j+}. The matrix $\boldsymbol{\Gamma}$ must be positive definite since $\boldsymbol{\Gamma}_j$ represents the covariance matrix of $\boldsymbol{\epsilon}_j$ given $\boldsymbol{\epsilon}_{-j}$. Substituting (8) in (4), the symmetry requirement of (4) is equivalent to

$$\boldsymbol{H}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}\boldsymbol{H}' \tag{9}$$

or, in other words, $\boldsymbol{F} \equiv \boldsymbol{\Gamma}^{-1/2}\boldsymbol{H}\boldsymbol{\Gamma}^{1/2}$ should be symmetric, where $\boldsymbol{\Gamma}^{1/2}$ is the (unique) square root matrix of $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}^{-1/2}$ is its inverse.

With the above re-parametrization, the distribution of $\boldsymbol{\epsilon}$ is multivariate normal with mean 0 and covariance matrix $\boldsymbol{D}$, which is now a function of $\boldsymbol{\Gamma}$ and $\boldsymbol{F}$ only. In fact, the inverse of $\boldsymbol{D}$ has the expression

$$\boldsymbol{D}^{-1} = (\boldsymbol{I}_n \otimes \boldsymbol{\Gamma}^{-1/2})(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{F})(\boldsymbol{I}_n \otimes \boldsymbol{\Gamma}^{-1/2}) \tag{10}$$

where $A \otimes B$ represents the Kronecker product of two matrices A and B. To ensure that $\boldsymbol{D}^{-1}$ is positive definite, we state the following theorem:

Theorem 1. *Let $\lambda_1, \lambda_2, \cdots, \lambda_p$ denote the eigenvalues of $\boldsymbol{F} = \boldsymbol{\Gamma}^{-1/2}\boldsymbol{H}\boldsymbol{\Gamma}^{1/2}$. The covariance matrix $\boldsymbol{D}$ is positive definite if $-1 \leq \lambda_k \leq 1$ for all $k = 1, 2, \cdots, p$.*

The reader is referred to the Appendix for the proof. The concatenated vector of all η_{ij}s at the j-th spatial location is denoted by $\boldsymbol{\eta}_j \equiv (\eta_{1j}, \eta_{2j}, \cdots, \eta_{pj})'$. Further, $\boldsymbol{\eta} \equiv (\boldsymbol{\eta}_1', \boldsymbol{\eta}_2', \cdots, \boldsymbol{\eta}_n')'$ represents the $np \times 1$ vector of all η_{ij} variables. Since $\boldsymbol{\epsilon} \sim N_{np}(0, \boldsymbol{D})$, it follows that

$$\boldsymbol{\eta} \sim N_{np}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{D}) \tag{11}$$

where $\boldsymbol{\beta} = (\beta_1', \beta_2', \cdots, \beta_p')'$ is the $q \times 1$ concatenated vector of all regression coefficients with $q = \sum_{i=1}^{p} q_i$, and $\boldsymbol{X}$ is the $p\,n \times q$ design matrix given by

$$\boldsymbol{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}_{(np \times q)}, \qquad \text{with} \qquad \boldsymbol{X}_j = \begin{pmatrix} x_{1j}' & 0 & \cdots & 0 \\ 0 & x_{2j}' & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & x_{pj}' \end{pmatrix}_{(p \times q)} \tag{12}$$

denoting the design matrix at the j-th spatial location for $j = 1, 2, \cdots, n$.

2.2 Handling Partial and Missing Observations

In some applications, part of the $\boldsymbol{y}$ observations can be either missing or partially observed, thus requiring the development of spatial hierarchical GLMMs for partially observed data structures. The sets $\mathcal{M}$, $\mathcal{P}$ and $\mathcal{C}$ represent all pairs (i, j) where y_{ij} is either missing, partially observed or completely observed, respectively. Given $\boldsymbol{\eta}$, the conditional likelihood contribution corresponding to the partially observed data, $\mathcal{D}_{obs}$ say, can be written as the product of two components

$$\ell(\mathcal{D}_{obs} \,|\, \boldsymbol{\eta}) = \prod_{(i,j) \in \mathcal{C}} f_{ij}(y_{ij} \,|\, \eta_{ij}) \prod_{(i,j) \in \mathcal{P}} F_{ij}(P_{ij} \,|\, \eta_{ij}) \tag{13}$$

taken over the sets $\mathcal{C}$ and $\mathcal{P}$, with

$$F_{ij}(P_{ij} \,|\, \eta_{ij}) = \int_{y_{ij} \in P_{ij}} f_{ij}(y_{ij} \,|\, \eta_{ij})\, dy_{ij} \tag{14}$$

denoting the contribution arising from the partial information that y_{ij} belongs to the set P_{ij}. To write down the (unconditional) likelihood for the hierarchical GLMM, we integrate over the distribution of $\boldsymbol{\eta}$ in (11):

$$\ell(\mathcal{D}_{obs} \,|\, \boldsymbol{\beta}, \boldsymbol{F}, \boldsymbol{\Gamma}) = \int_{\boldsymbol{\eta}} \ell(\mathcal{D}_{obs} \,|\, \boldsymbol{\eta})\, f_0(\boldsymbol{\eta} \,|\, \boldsymbol{\beta},\, \boldsymbol{F}, \boldsymbol{\Gamma})\, d\boldsymbol{\eta} \tag{15}$$

where f_0 is the distribution of $\boldsymbol{\eta}$ given in (11). Examples of partially observed data are common in rare diseases. For example, when mapping cancer incidences, certain counties do not report the exact number of incidences if the total number is less than a known threshold. In this case the partial information is $P_{ij} = \{\, y_{ij} < \tau \,\}$ where τ is the known threshold.

3 Default Prior Elicitation

This section discusses the appropriate default priors on the model parameters $\boldsymbol{\beta}$, $\boldsymbol{H}$ and $\boldsymbol{\Gamma}$. Since each β_i represents the regression effects to the mean of the observations y_{ij}, it is natural to elicit a standard "flat" non-informative prior on each β_i for $i = 1, 2, \cdots, p$:

$$\pi_N(\boldsymbol{\beta}) \propto 1 \tag{16}$$

It is also natural to consider the Jeffrey's type non-informative prior on $\boldsymbol{\Gamma}$ of the form $\pi_J(\boldsymbol{\Gamma}) \propto (\det(\boldsymbol{\Gamma}))^{-(p+1)/2}$. We state

Theorem 2. *Let $\pi_0(\boldsymbol{H})$ be a proper prior on $\boldsymbol{H}$. The default prior specification*

$$\pi_0(\boldsymbol{\beta}, \boldsymbol{H}, \boldsymbol{\Gamma}) = \pi_N(\boldsymbol{\beta}) \times \pi_J(\boldsymbol{\Gamma}) \times \pi_0(\boldsymbol{H}) \tag{17}$$

gives a posterior distribution on the parameters that is improper.

Proof: We refer the reader to a proof in the Appendix. The consequence of Theorem 2 is that new motivation is required for the development of a default prior on $\boldsymbol{\Gamma}$, which should make the posterior proper. We discuss the development of such a prior in the subsequent paragraphs.

Our justification for the default prior comes from looking at the conditional update of each η_{ij} given the data y_{ij} and the rest of the $\boldsymbol{\eta}$ elements (excluding η_{ij}). In the normal-normal case, one can explicitly derive an expression for the weights that represent the contribution of the data, y_{ij}, and the rest of the $\boldsymbol{\eta}$ to the conditional mean of η_{ij}. However, in the case of non-conjugate GLMMs, it is not possible to obtain a closed form expression for the weights. An approximate approach can be considered based on a quadratic expansion of the exponential family pdf to yield a similar analysis as in the normal-normal model. We consider the prediction of the vector $\boldsymbol{\eta}_j$ given $\boldsymbol{y}_j$ and $\boldsymbol{\eta}_{-j}$ (which are the rest of the η_{ij}s excluding the ones at site j). The GLM approach to estimating $\boldsymbol{\eta}_j$ is iterative and expands the function h_is in (1) around the current estimate $\boldsymbol{\eta}_j^*$. From Taylor's expansion, we have

$$h_i(\eta_{ij}) \approx h_i(\eta_{ij}^*) + (\eta_{ij} - \eta_{ij}^*) h_i^{(1)}(\eta_{ij}^*) + \frac{1}{2}(\eta_{ij} - \eta_{ij}^*)^2 h_i^{(2)}(\eta_{ij}^*). \tag{18}$$

Using the expansion above, the exponential family of densities can be expanded around $\boldsymbol{\eta}_j^*$ similarly as

$$\begin{aligned}
& \prod_{i=1}^{p} \exp\{\, \eta_{ij} y_{ij} - h_i(\eta_{ij}) \,\} \\
\approx \; & \prod_{i=1}^{p} \exp\left\{ \eta_{ij} y_{ij} - h_i(\eta_{ij}^*) - (\eta_{ij} - \eta_{ij}^*) h_i^{(1)}(\eta_{ij}^*) - \frac{1}{2}(\eta_{ij} - \eta_{ij}^*)^2 h_i^{(2)}(\eta_{ij}^*) \right\} \\
\propto \; & \exp\left\{ -\frac{1}{2}(\boldsymbol{\eta}_j - \boldsymbol{\eta}_j^*)' \boldsymbol{H}_j^{(2)} (\boldsymbol{\eta}_j - \boldsymbol{\eta}_j^*) + \boldsymbol{\eta}_j' (\boldsymbol{y}_j - \boldsymbol{h}_j^{(1)}) \right\},
\end{aligned}$$

where $\boldsymbol{h}_j^{(1)} = \left(h_1^{(1)}(\eta_{1j}^*), h_2^{(1)}(\eta_{2j}^*), \cdots, h_p^{(1)}(\eta_{pj}^*)\right)'$ is the $p \times 1$ of first derivatives, and

$$\boldsymbol{H}_j^{(2)} = \text{diag}\left(h_1^{(2)}(\eta_{1j}^*), h_2^{(2)}(\eta_{2j}^*), \cdots, h_p^{(2)}(\eta_{pj}^*)\right) \tag{19}$$

is the diagonal matrix of all second derivatives of h_i, $i = 1, 2, \cdots, p$ evaluated at $\boldsymbol{\eta}_j^*$. The Taylor's expansion above allows us to revert back to the normal-normal case where by completing squares, the observational part for $\boldsymbol{\eta}_j$ can be derived as

$$\exp\left\{-\frac{1}{2}(\boldsymbol{\eta}_j - \boldsymbol{\eta}_i^*(\boldsymbol{y}_i))'\boldsymbol{H}_j^{(2)}(\boldsymbol{\eta}_j - \boldsymbol{\eta}_i^*(\boldsymbol{y}_i))\right\}$$

with $\boldsymbol{\eta}_i^*(\boldsymbol{y}_i) \equiv \left(\boldsymbol{H}_j^{(2)}\right)^{-1}(\boldsymbol{y}_j - \boldsymbol{h}_j^{(1)} + H_j^{(2)}\boldsymbol{\eta}_i^*)$. The contribution from the multivariate Gaussian CAR prior in (11) (i.e., the rest of the $\boldsymbol{\eta}$ elements) is

$$\exp\left\{-\frac{w_{j+}}{2}(\boldsymbol{\eta}_j - \boldsymbol{\eta}_j^{CAR})'(\Gamma^{-1})(\boldsymbol{\eta}_j - \boldsymbol{\eta}_j^{CAR})\right\}$$

where $\eta_j^{CAR} = \boldsymbol{H}\sum_{k\in N_j}\frac{w_{ij}}{w_{j+}}\boldsymbol{\eta}_k$. Combining the likelihood and prior parts, the conditional mean of $\boldsymbol{\eta}_j$ (again by completion of squares) turns out to be

$$(\boldsymbol{H}_j^{(2)} + \boldsymbol{\Gamma}^{-1}\, w_{j+})^{-1}\boldsymbol{H}_j^{(2)}\boldsymbol{\eta}_j^*(\boldsymbol{y}_j) + (\boldsymbol{H}_j^{(2)} + \boldsymbol{\Gamma}^{-1}\, w_{j+})^{-1}\boldsymbol{\Gamma}^{-1}w_{j+}\boldsymbol{\eta}_j^{CAR} \tag{20}$$

with (matrix) weights

$$\boldsymbol{W}_{1j} = (\boldsymbol{H}_j^{(2)} + \boldsymbol{\Gamma}^{-1}\, w_{j+})^{-1}\boldsymbol{H}_j^{(2)} \quad \text{and} \quad \boldsymbol{W}_{2j} = \boldsymbol{I} - \boldsymbol{W}_{1j} \tag{21}$$

corresponding to the direct estimate $\boldsymbol{y}_j$ and the population mean. Our proposal is to induce a prior on $\boldsymbol{\Gamma}$ such that the prior on

$$\boldsymbol{W}_{1j} = \left(\boldsymbol{H}_j^{(2)} + \boldsymbol{\Gamma}^{-1}\, w_{j+}\right)^{-1}\boldsymbol{H}_j^{(2)} = \left(\frac{\boldsymbol{H}_j^{(2)}}{w_{j+}} + \boldsymbol{\Gamma}^{-1}\right)^{-1}\frac{\boldsymbol{H}_j^{(2)}}{w_{j+}}$$

is uniform. A similar technique was used by Daniels (1998) and Natarajan and Kass (2000) in the univariate non-spatial context. Since $\boldsymbol{W}_{1j}$ varies with j, we first replace it by its average across all the sites. Thus, we set

$$w_0 = \frac{1}{np}\sum_j \text{trace}\left(\frac{H_j^{(2)}}{w_{j+}}\right).$$

Substituting $\frac{\boldsymbol{H}_j^{(2)}}{w_{j+}}$ by its average $w_0\,\boldsymbol{I}_p$ in the expression of W_{1j} above, we get the matrix $\boldsymbol{U}$ defined by

$$\boldsymbol{U} \equiv \left(\, w_0\boldsymbol{I}_p + \boldsymbol{\Gamma}^{-1}\right)^{-1} w_0\boldsymbol{I}_p = \left(\, w_0\boldsymbol{\Gamma} + \boldsymbol{I}_p\right)^{-1} w_0\boldsymbol{\Gamma}$$

Note that $0 \leq \boldsymbol{U} \leq \boldsymbol{I}_p$ in terms of positive definiteness with $\det(\boldsymbol{U})$ representing the volume of the weight matrix $\boldsymbol{U}$. We seek a prior on $\boldsymbol{\Gamma}$ that is induced by a default prior on $\det(\boldsymbol{U})$. The class of multivariate Beta distribution given by

$$f(\boldsymbol{U} \,|\, a, b) = C(\det(\boldsymbol{U}))^{a-(p+1)/2}(\det(\boldsymbol{I}_p - \boldsymbol{U}))^{b-(p+1)/2}$$

with $a > (p-1)/2$ and $b > (p-1)/2$ forms a class of prior for $\boldsymbol{U}$. The uniform prior on the weight matrix $\boldsymbol{U}$ is obtained by setting $a = b = (p+1)/2$. The resulting prior on $\boldsymbol{\Gamma}$ corresponding to uniform volume of $\boldsymbol{U}$ is

$$\pi_{UV}(\boldsymbol{\Gamma}) = \det\left(\boldsymbol{I}_p + w_0\boldsymbol{\Gamma}\right)^{-(p+1)}. \tag{22}$$

This is also the prior developed in Natarajan and Kass (2000) leading to shrinkage-type estimators in the non-spatial context. The uniform volume prior is proper from Theorem 2 of Natarajan and Kass (2000).

The prior on $\boldsymbol{H}$ is induced via $\boldsymbol{F}$. The prior on $\boldsymbol{F}$ is taken to be independent of $\boldsymbol{\Gamma}$ and is elicited as follows: Writing the spectral decomposition of $\boldsymbol{F}$ as

$$\boldsymbol{F} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}', \tag{23}$$

we put a uniform prior on $\boldsymbol{Q}$ and in view of Theorem 1, we put a uniform prior $U(-1,+1)$ on the eigenvalues in $\boldsymbol{\Lambda}$. The default prior on $(\boldsymbol{\beta}, \boldsymbol{F}, \boldsymbol{\Gamma})$ is thus

$$\pi_0(\boldsymbol{\beta}, \boldsymbol{F}, \boldsymbol{\Gamma}) = \pi_N(\boldsymbol{\beta}) \times \pi_{UV}(\boldsymbol{\Gamma}) \times \frac{1}{2^p}. \tag{24}$$

For each $i = 1, 2, \cdots, p$, the design matrix corresponding to the i-th response variable is the $n \times q_i$ matrix $\tilde{\boldsymbol{X}}_i = (x_{i1}, x_{i2}, \cdots, x_{in})'$. The submatrix $\tilde{\boldsymbol{X}}_{\mathcal{C}_i}$ is formed by taking all rows j of $\tilde{\boldsymbol{X}}_i$ for which $(i,j) \in \mathcal{C}$. We state

Theorem 3. *Assume that f_{ij} and F_{ij} in (13) are bounded above by a constant independent of η_{ij} for each pair (i,j). Using the default prior (24), the posterior is proper if there exists q_i linearly independent row vectors in $\tilde{\boldsymbol{X}}_{\mathcal{C}_i}$ for each $i = 1, 2, \cdots, p$ such that*

$$\int_{\boldsymbol{\Gamma}} \int_{\boldsymbol{F}} \int_{\boldsymbol{\beta}} \int_{\boldsymbol{\eta}} \left(\prod_{i=1}^{p} \prod_{j=1}^{q_i} f_{ij}(y_{ij} \,|\, \eta_{ij}) \right) f_0(\boldsymbol{\eta} \,|\, \boldsymbol{\beta},\ \boldsymbol{F}, \boldsymbol{\Gamma})\, d\boldsymbol{\eta}\, \pi_0(\boldsymbol{\beta}, \boldsymbol{F}, \boldsymbol{\Gamma})\, d\boldsymbol{\beta}\, d\boldsymbol{F}\, d\boldsymbol{\Gamma}\ < \infty, \tag{25}$$

where f_0 and π_0 is, respectively, the distribution of $\boldsymbol{\eta}$ and the prior, as given in (11) and (24).

Remark 1: Under the assumptions of Theorem 3, $0 \leq f_{ij} \leq A$ and $0 \leq F_{ij} \leq B$, say. In our applications, f_{ij} is taken to be either a Poisson pmf or a normal pdf. For the Poisson

(or, generally for a discrete distribution), it easily follows that the bounds $A = B = 1$, independent of i and j. When f_{ij} is normal with mean η_{ij} and fixed standard deviation σ_0, it follows that $A = 1/\sqrt{2\pi}\sigma_0$ and $B = 1$, thus independent of i and j again. Generally for densities, the bound A needs to be established on a case-by-case basis.

Remark 2: Theorem 3 shows that propriety can be achieved if there are at least q_i sites on the lattice for which the y_{ij}s are completely observed. The only further check that we have to perform is to see if the design matrix corresponding to those sites $\tilde{\boldsymbol{X}}_{\mathcal{C}_i}$ is non-singular. This is a requirement for each i, so the conditions of Theorem 3 can be checked separately for each $i = 1, 2, \cdots, p$.

4 Bayesian Inference

The Gibbs sampler is a natural method for posterior simulations in the case of GLMMs, and is also the method utilized for our spatial models. A slightly different (yet equivalent) parametrization of the spatial multivariate GLMM is considered subsequently. Instead of $\boldsymbol{\Gamma}^{-1/2}$, the lower triangular matrix $\boldsymbol{L}$ obtained from the Cholesky decomposition of $\boldsymbol{\Gamma}^{-1}$ is used. Also, we define the matrix $\boldsymbol{B} \equiv \boldsymbol{LQ}$ with entries $\boldsymbol{B} = ((b_{uv}))_{u,v=1}^{p}$ where $\boldsymbol{Q}$ is the orthogonal matrix from the spectral value decomposition of $\boldsymbol{F}$. The following string of equalities demonstrate that $\boldsymbol{D}$ is a function of $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$: We have

$$\begin{aligned}
\boldsymbol{D}^{-1} &= (\boldsymbol{I}_n \otimes \boldsymbol{L})(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{F})(\boldsymbol{I}_n \otimes \boldsymbol{L}') \\
&= (\boldsymbol{I}_n \otimes \boldsymbol{L})(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}')(\boldsymbol{I}_n \otimes \boldsymbol{L}') \\
&= (\boldsymbol{I}_n \otimes \boldsymbol{L}\,\boldsymbol{Q})(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{\Lambda})(\boldsymbol{I}_n \otimes \boldsymbol{Q}'\,\boldsymbol{L}') \\
&= (\boldsymbol{I}_n \otimes \boldsymbol{B})(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{\Lambda})(\boldsymbol{I}_n \otimes \boldsymbol{B}').
\end{aligned}$$

The enormous advantage of the re-parametrization in terms of $\boldsymbol{B}$ is that the entries of $\boldsymbol{B}$ are unconstrained. Note that it is possible to obtain $\boldsymbol{Q}$ and $\boldsymbol{L}$ uniquely from $\boldsymbol{B}$ using the QR decomposition of $\boldsymbol{B}' = \mathcal{Q}\mathcal{R}$ where $\mathcal{Q}$ is orthogonal and $\mathcal{R}$ is upper triangular. It follows that $\boldsymbol{L} = \mathcal{R}'$ and $\boldsymbol{Q} = \mathcal{Q}'$.

The four main steps of the Gibbs sampler are:

- **Update η_{ij}:** The update of η_{ij} is carried out based on the following (conditional) posterior density of η_{ij} given the rest of the parameters:

$$\pi(\eta_{ij} \mid \cdots) \propto \exp\{\, \eta_{ij} y_{ij} - h_i(\eta_{ij}) - \frac{A_{ij}}{2}(\eta_{ij} - \eta_{ij}^*)^2\} \tag{26}$$

where $\eta^*_{ij} \equiv x'_{ij}\beta_i + \epsilon^*_{ij}$, and ϵ^*_{ij} has the expression $\epsilon^*_{ij} = \frac{1}{\sum_{v=1}^p b^2_{iv}}\,\epsilon_0$,

$$\epsilon_0 = \underbrace{\sum_{v=1}^{p} b^2_{iv}\lambda_v \sum_{k\in N_j} \frac{w_{jk}}{w_{j+}}\,\epsilon_{ik}}_{(1)} - \underbrace{\sum_{v=1}^{p}\sum_{u=1,u\neq i}^{p} b_{iv}b_{uv}\epsilon_{uj}}_{(2)} + \underbrace{\sum_{v=1}^{p}\sum_{u=1,u\neq i}^{p} b_{iv}b_{uv}\lambda_v \sum_{k\in N_j} \frac{w_{jk}}{w_{j+}}\,\epsilon_{uk}}_{(3)} \tag{27}$$

with $\epsilon_{uv} = \eta_{uv} - x'_{uv}\beta_{uv}$ for all $(u, v) = 1, 2, \cdots, p$ except for $(u, v) = (i, j)$, and $A_{ij} = w_{j+}\sum_{v=1}^p b^2_{iv}$. The updating formula in (27) is based on three components: The first component (terms in (1) above) involve η_{ik} values for spatial sites $k \in N_j$ for the same variable index i, the second component involves η_{uj} for the other variables $u \neq i$ but for the same spatial location j, whereas the third component involves η_{uk} for variables other than i and sites in $k \in N_j$. The update of η_{ij} is based on histogramming the conditional posterior density in (26) for each fixed pair (i, j) and cycling through all the combinations of $(i, j) \in \{(1, 1), (1, 2), \cdots, (1, n), \cdots, (p, n)\}$. This is the $\boldsymbol{\eta}_{ij}$ updating step when y_{ij} is completely observed, that is, for $(i, j) \in \mathcal{C}$. When y_{ij} is missing, the first term in equation (26) will be absent. The standard method of analyzing partially observed y_{ij} is to treat it as missing and update the value based on the truncated distribution $f_{ij}(\cdot \,|\, \eta_{ij})$ given that $y_{ij} \in P_{ij}$.

- **Update $\boldsymbol{\beta}$:** The update of $\boldsymbol{\beta}$ requires a re-ordering of the variables involved. For each $i = 1, 2, \cdots, p$, define $\boldsymbol{\eta}^r_i$ and $\boldsymbol{\epsilon}^r_i$ to be the $n \times 1$ vectors corresponding to the i-th variable, that is, $\boldsymbol{\eta}^r_i = (\eta_{i1}, \eta_{i2}, \cdots \eta_{in})'$ and $\boldsymbol{\epsilon}^r_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots \epsilon_{in})'$. Also, let $\boldsymbol{\eta}^r \equiv ((\boldsymbol{\eta}^r_1)', (\boldsymbol{\eta}^r_2)', \cdots, (\boldsymbol{\eta}^r_p)')$ and $\boldsymbol{\epsilon}^r \equiv ((\boldsymbol{\epsilon}^r_1)', (\boldsymbol{\epsilon}^r_2)', \cdots, (\boldsymbol{\epsilon}^r_p)')$ denote the $np \times 1$ vector of re-ordered entries from $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$, respectively. The covariance matrix of $\boldsymbol{\epsilon}^r$ is subsequently a re-ordered version of $\boldsymbol{D}$ given by

$$(\boldsymbol{D}^r)^{-1} = (\boldsymbol{B} \otimes \boldsymbol{I}_n)(\boldsymbol{I}_p \otimes \boldsymbol{M} - \boldsymbol{\Lambda} \otimes \boldsymbol{W})(\boldsymbol{B}' \otimes \boldsymbol{I}_n). \tag{28}$$

Also, let $\tilde{\boldsymbol{X}} = ((\text{Block Diagonal}\{\tilde{\boldsymbol{X}}_i\}))$ denote the block diagonal matrix consisting of the design matrices for the i-th response variable for $i = 1, 2, \cdots, p$. The conditional posterior distribution of $\boldsymbol{\beta}$ is multivariate normal with mean $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, and covariance matrix $\boldsymbol{S}_{\boldsymbol{\beta}}$ where

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}'(\boldsymbol{D}^r)^{-1}\tilde{\boldsymbol{X}})^{-1}(\tilde{\boldsymbol{X}}'(\boldsymbol{D}^r)^{-1}\boldsymbol{\eta}^r) \qquad \text{and} \qquad \boldsymbol{S}_{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}'(\boldsymbol{D}^r)^{-1}\tilde{\boldsymbol{X}})^{-1}, \tag{29}$$

respectively.

- **Update $\boldsymbol{\Lambda}$:** An enormous advantage of the re-parametrization in terms of $\boldsymbol{B}$ and $\boldsymbol{\Lambda}$ earlier is when updating Λ: The diagonal entries of $\boldsymbol{\Lambda}$ can be updated independently

of each other. Consider the $p \times n$ matrix, $\boldsymbol{\Upsilon}$, constructed by putting ϵ_{ij} in its i-th row and j-th column entry. Define a new matrix $p \times n$ matrix $\boldsymbol{E}$ as

$$\boldsymbol{E} = \boldsymbol{B}'\boldsymbol{\Upsilon} \tag{30}$$

and let e_i' be the i-th row of $\boldsymbol{E}$, for $i = 1, 2, \cdots, p$. The conditional posterior density of λ_k is given by

$$\pi(\lambda \,|\, \cdots) \propto \exp\left\{-\frac{1}{2}\, e_k'(\boldsymbol{M} - \lambda \boldsymbol{W})\, e_k\right\} (\det\,(\boldsymbol{M} - \lambda \boldsymbol{W}))^{1/2} \tag{31}$$

in the range of $-1 \leq \lambda \leq 1$ independently for each $k = 1, 2, \cdots, p$. The update of λ_k is based on histogramming the conditional posterior density in (31) for each fixed k and cycling through all $k = 1, 2, \cdots, p$.

- **Update $\boldsymbol{B}$:** The conditional posterior density of $\boldsymbol{B}$ has the expression

$$\begin{aligned} \pi(\boldsymbol{B} \,|\, \cdots) \;\; \propto \;\; & \exp\left\{-\frac{1}{2}\sum_{k=1}^{p} e_k'(\boldsymbol{M} - \lambda_k \boldsymbol{W})\, e_k\right\} \times \det\,(\, w_0 \boldsymbol{I} + \boldsymbol{B}\boldsymbol{B}'\,)^{-(p+1)} \\ & \times \det(\boldsymbol{B}\boldsymbol{B}')^{(n+1)/2} \end{aligned} \tag{32}$$

where $e_k' \equiv e_k'(\boldsymbol{B})$ is as defined in (30) but now viewed as a function of $\boldsymbol{B}$. The latter part of the conditional density in (32) is the contribution of the default prior on $\boldsymbol{\Gamma}$. The update of $\boldsymbol{B}$ is carried out by updating each entry b_{uv} one at a time. The conditional posterior density of b_{uv} is the same (upto a constant of proportionality) as in (32). The actual update of b_{uv} is performed by histogramming the density in (32) and cycling through all combinations (u, v) with $u, v = 1, 2, \cdots, p$.

5 Experimental Results

5.1 Simulation study

We have conducted extensive simulation studies to check the performance of our methodology. The experimental settings closely mimic county-level data for southern lower Michigan obtained from the SEER database. We took $p = 2$ with y_{1j} and y_{2j} representing binomial responses on $n = 40$ spatial sites with neighborhood structure determined by the adjacency information among Michigan counties. Given $\boldsymbol{\eta}$, the observed data $y_{ij} \sim \text{Binomial}\left(T_{ij}, \frac{e^{\eta_{ij}}}{1+e^{\eta_{ij}}}\right)$, for $i = 1, 2$ and $j = 1, 2, \cdots, 40$, independent of each other; T_{ij} is the total number of trials selected independently from a Uniform$(22, 48)$ distribution. The distribution of $\boldsymbol{\eta}$ is multivariate Gaussian CAR with the following true parameter specifications: $\boldsymbol{\beta}_1 = (-1, 0.3)'$,

$\boldsymbol{\beta}_2 = (-0.5, -0.2)'$,

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.125 \end{pmatrix} \quad \text{and} \quad \boldsymbol{F} = \begin{pmatrix} 0.7 & -0.1 \\ -0.1 & 0.2 \end{pmatrix}.$$

The first components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ correspond to the intercept terms. Here $q_1 = q_2 = 2$ and additional covariate information is gathered from independent normal distributions: $X_1 \sim N(0, \sigma^2 = 0.3)$ and $X_2 \sim N(1, \sigma^2 = 0.5)$.

Two choices of priors were considered for $\boldsymbol{\Gamma}$, namely, (a) the default prior $\pi_{UV}(\boldsymbol{\Gamma}) \propto \det(\boldsymbol{I_p} + \boldsymbol{\omega_0}\boldsymbol{\Gamma})^{-(p+1)}$ and (b) the proper inverse Wishart given by $\pi_{IW}(\boldsymbol{\Gamma}) = IW(\rho, \rho\boldsymbol{A})$, where $\rho \geq p$. The inverse Wishart distribution is a generalization of the inverse gamma for the variance parameter in a multivariate setting. If $\boldsymbol{\Gamma} \sim IW(m, \boldsymbol{\Psi})$, the expectation and variance of entries of $\boldsymbol{\Gamma}$ are given by $E(\boldsymbol{\Gamma}_{kl}) = \Psi_{kl}/(m-p-1)$ and $var(\boldsymbol{\Gamma}_{kl}) = \frac{(m-p+1)\boldsymbol{\Psi}_{kl}^2 + (m-p-1)\boldsymbol{\Psi}_{kk}\boldsymbol{\Psi}_{ll}}{(m-p)(m-p-1)^2(m-p-3)}$ where $\boldsymbol{\Gamma}_{kl}$ and $\boldsymbol{\Psi}_{kl}$ are the (k,l)-th entries of $p \times p$ matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Psi}$, respectively. When ρ is large, $E(\boldsymbol{\Gamma}) \approx \boldsymbol{A}$, and $var(\boldsymbol{\Gamma}_{ij}) \approx 1/\rho$, leading to a high concentration of probability around the initial guess of $\boldsymbol{A}$. Thus, this prior does not represent non-informative prior knowledge. Choice (b) was proposed by Sain and Cressie (2007) as the choice of vague prior for $\boldsymbol{\Gamma}$ when ρ is large, which is not the case (actually, Sain and Cressie (2007) put a prior on $\boldsymbol{\Gamma}^{-1}$ which is Wishart$(\rho, (\rho\boldsymbol{A})^{-1})$, but this is equivalent to choice (b) since $\boldsymbol{\Gamma} \sim IW(m, \boldsymbol{\Psi}) \quad \text{iff} \quad \boldsymbol{\Gamma^{-1}} \sim W(m, \boldsymbol{\Psi^{-1}})$).

The priors (a) and (b) for $\boldsymbol{\Gamma}$ above in turn induce priors on $\boldsymbol{B}$. This is based on the transformations $\boldsymbol{\Gamma} \to \boldsymbol{\Gamma^{-1}}$, $\boldsymbol{\Gamma^{-1}} \to \boldsymbol{LL'}$, and $(\boldsymbol{L}, \boldsymbol{Q}) \to \boldsymbol{B}$. The derivation of the Jacobian for the composition transformation from $(\boldsymbol{\Gamma}, \boldsymbol{Q}) \to \boldsymbol{B}$ is given in the Appendix. The priors on $\boldsymbol{B}$ turn out to be (i) $\pi_{UV}(\boldsymbol{B}) = \det(\boldsymbol{\omega_0 I_p} + \boldsymbol{BB'})^{-(p+1)} \det(\boldsymbol{BB'})^{1/2}$, (ii) $\pi_{IW}(\boldsymbol{B}) = \exp\left\{-\frac{\rho}{2} tr(\boldsymbol{ABB'})\right\} \times \det(\boldsymbol{BB'})^{\frac{\rho-p}{2}}$ respectively, for the priors (a) and (b) for $\boldsymbol{\Gamma}$. Prior choices for $\boldsymbol{\beta}$ are (i) the default non-informative constant prior $\pi_N(\boldsymbol{\beta}) \propto 1$, and (ii) the proper subjective prior $\pi_G(\boldsymbol{\beta}_k) \sim N(0, \sigma_k^2 \boldsymbol{I}_{q_k})$ independently for each $k = 1, 2, \cdots, p$. Using (ii), it is easy to see that the posterior for $\boldsymbol{\beta}$ is $N(\boldsymbol{\mu}, \boldsymbol{S})$ where

$$\boldsymbol{\mu} = \boldsymbol{SX'}(\boldsymbol{D^r})^{-1}\boldsymbol{\eta^r}, \ \boldsymbol{S^{-1}} = \boldsymbol{X'}(\boldsymbol{D^r})^{-1}X + \boldsymbol{\Sigma^{-1}}. \tag{33}$$

Here, $\boldsymbol{\Sigma}$ = Block Diagonal$(\sigma_k^2 \boldsymbol{I}_{q_k})$. Thus, we investigate the following three prior choices: (I) $\pi_{UV}(\boldsymbol{B})$ and $\pi_N(\boldsymbol{\beta})$ and (II) $\pi_{IW}(\boldsymbol{B})$ and $\pi_G(\boldsymbol{\beta}_k)$ with $\rho = 5$, $A = I_p$ and $\sigma_k^2 = 100$, and (III) $\pi_{IW}(\boldsymbol{B})$ and $\pi_G(\boldsymbol{\beta}_k)$ with $\rho = 100,000$, $A = I_p$ and $\sigma_k^2 = 100$.

The Gibbs sampler was run for $10,000$ iterations and checked for convergence using traceplots and the R-statistic of Gelman and Rubin (1992). We established convergence for all the experiments by $5,000$ iterations; all diagnostic plots are satisfactory but are suppressed to save space. Outputs from the Gibbs chains were used to compute different statistics to validate and compare the proposed Bayesian methodology. The three prior

choices are compared in terms of their ability to derive consistent estimation and prediction results. The deviation measures of comparisons are (1) relative mean absolute deviation (RMAD), (2) mean square error (MSE), (3) empirical 90% HPD coverage probabilities (CP), and (4) width of the 90% HPD set (W). Formulas for a generic quantity θ are given by $RMAD_\theta = (E(\theta) - \theta_0)/\theta_0$, $MSE_\theta = E(\theta - \theta_0)^2$, $CP_\theta = P\{\theta_0 \in HPD(\theta)\}$ and W is the width of $HPD(\theta)$; in the RHS of each expression, θ represents a sample (or samples) from the posterior, $HPD(\theta)$ is the 90% HPD set calculated based on θ samples and θ_0 is the true value set in the simulation experiments. We used 500 replications in each experiment and report the averages of the deviation measures above. The computational time for all the 500 replications in each experimental setup is approximately $20 - 25$ hours. All computations were carried out using an HP ProLiant DL160 machine (a cluster of nodes) with 8 Intel Xeon cores and 24GB of memory at the High Performance Computing Center (HPCC) at Michigan State University.

Table 1 reports a summary of all deviation measures. For convenience of understanding the results, we report averages over specific components of the unknown parameters; for example, the $\boldsymbol{\beta}$ column reports the average over all β components, $\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}$, and similarly for the other parameters $\boldsymbol{\Gamma}$ and $\boldsymbol{F}$. The last column reports averages over all $40 \times 2 = 80$ predicted values for $\boldsymbol{\eta}$; the deviation measures are calculated based on $\boldsymbol{\eta}_0$, the true values generated in each replication. Entries in Table 1 clearly shows the sensitivity of standard prior distributions (i.e., priors (II) and (III)) used in the literature. For example, for a wrong choice of ρ and $\boldsymbol{A}$ in π_{IW}, the coverage can be even 0 along with high MSE and RMAD. This effect can be reduced with a more sensible prior choice, for example, choice (II). On the other hand, π_{UV} always provides sensible results. One might notice that the sensitivity of priors (II) and (III) is highest for $\boldsymbol{\Gamma}$ compared to the other columns in Table 1. This is due to the fact that the prior for $\boldsymbol{\Gamma}$ changes significantly for the three choices (I-III) whereas we always use the default uniform prior for $\boldsymbol{F}$. The regression parameters $\boldsymbol{\beta}$ are less affected by the prior choice compared to $\boldsymbol{\Gamma}$ due to the fact that $\boldsymbol{\beta}$ is related to the mean parameter with large prior variance while $\boldsymbol{\Gamma}$ is related to dispersion. Nevertheless, the standard choice of Gaussian prior on $\boldsymbol{\beta}$ also appears to be somewhat sensitive, but not to the extent of $\boldsymbol{\Gamma}$. Although the $\boldsymbol{\eta}$ components are not fixed model parameters (i.e., they vary from county to county), their inference can also be sensitive to the different prior choices. To explain the discrepancies in the $\boldsymbol{\Gamma}$ entries, Figure 1 plots the posterior densities of Γ_{22}, the $(2,2)$-th entry of $\boldsymbol{\Gamma}$, corresponding to the three different prior choices for a arbitrarily chosen replicate. Note that under prior (III), the prior mean is I_2 whereas the prior variance is 10^{-5} making it highly concentrated on a value different from the true Γ_{22}; while we understand that the small prior variance is unreasonable, this choice is not uncommon (see, for example, Sain and Cressie, 2007). The situation improves under prior (II) where the prior mean is the

same but the prior variance is 0.2 which gives a comparatively higher prior mass around the true Γ_{22}. Overall, the proposed default prior π_{UV} performed well in all respects. This prior is thus a robust choice. We have also explored with another choice of $\mathbf{\Gamma}$, namely

$$\mathbf{\Gamma} = \begin{pmatrix} 10 & 6 \\ 6 & 5 \end{pmatrix}.$$

The results are similar to one discussed here and is therefore not presented. The componentwise univariate spatial analysis was carried out and, as expected, the multivariate analysis had superior performance.

We also performed similar experiments with 10% missing observations. The results are reported in Table 2. Comparative trends similar to the complete data case with priors (I-III) are also observed here.

5.2 Real data examples

The Bayesian inferential framework developed in this paper is applied to study bivariate dependence of a number of health-socio-economic indicators in the state of Michigan. Two studies are conducted with different pairs of response variables: (1) lung cancer mortality incidence and poverty, and (2) lung cancer mortality and air quality index (AQI) measurements. Study (1) and (2) illustrate the complete and missing data applications, respectively. The source of our socio-economic data is SEER (*URL: seer.cancer.gov*). In each application, the Gibbs sampler was run for $10,000$ iterations and checked for convergence as in the simulated data. Posterior samples are obtained from the Gibbs chains for computing the mean, standard deviation and 90% highest posterior density (HPD) intervals for all the parameters.

5.2.1 Complete Data Example: Study 1

The mortality of lung cancer, the first component of the response variable in Study 1, is rare enough relative to the population in the 68 counties of lower Michigan that a Poisson distribution is appropriate. We write the model (conditional on η_{1j}) as

$$y_{1j} \overset{\text{ind}}{\sim} \text{Poisson}(E_j e^{\eta_{1j}}),\ i = 1, 2 \text{ and } j = 1, 2, \cdots, 68,$$

where y_{1j} measures the observed number of deaths in county j and E_j is the estimated population at risk; we assume E_j is known and give a way for calculating them later. The poverty count, y_{2j}, is taken to be the second component of the response variable in Study 1. We model y_{2j} as a Binomial with the number of trials being the total county population and success probability $\frac{e^{\eta_{2j}}}{1+e^{\eta_{2j}}}$. The associated covariates for y_{1j} are the intercept, PM25

(particulate matter with size < 2.5 μm obtained from EPAs NEI database), and the extents of urbanization, non-white population and non-industry (these are measured as proportions). Covariates for poverty are the intercept, and the extents of urbanization and non-industry. Thus, $q_1 = 5$ and $q_2 = 3$.

To calculate E_j, we take each county's age distribution into account, which is available from U.S. Census 2000. The expected age-adjusted number of deaths due in county j is

$$E_j = \sum_{k=1}^{m} \omega^k N_j^k,$$

for $j = 1, 2, \cdots, 68$ where $\omega^k = \sum_{j=1}^{68} D_j^k / \sum_{j=1}^{68} N_j^k$ is the age-specific death rate due to lung cancer for age group k and N_j^k is the total population at risk in county j for age group k. The county level maps of the age-adjusted standardized mortality ratios (SMRs), Y_{1j}/E_j for lung cancer shown in Figure 2 exhibit evidence of correlation over space. Figure 2 also gives the spatial distribution of poverty levels which can be seen to be highly spatially correlated with lung cancer (simple correlation between lung cancer and poverty is around 0.4).

We present summary conclusions of our analysis. The standard errors of the parameters estimates for prior (I) are smaller compared to priors (II) and (III). For example, the average standard error of regression coefficients for lung cancer is 0.61 under prior (I), and 0.65 and 1.28 for (II) and (III), respectively. For the variance component parameters, the standard errors and widths of HPD sets are comparable under priors (I) and (II). Note that the inference for $\mathbf{\Gamma}$ is highly misleading for prior (III) since most of the posterior probability is concentrated around the prior mean. Another difference is that the covariate PM25 (related to the particle matter in the air) for lung cancer incidence is positive and significant under (I), whereas it is insignificant under priors (II) and (III). For prior (III), the posterior mean takes a negative value which is not very realistic. For brevity, other statistics along with the smooth map of $\boldsymbol{\eta}$ are suppressed.

5.2.2 Missing Data Example: Study 2

In Study 2, AQI is taken as the second component of the response variable in place of poverty. AQI information is obtained from the EPA AQI report site. Air pollutant monitoring stations are sparsely located in 32 out of 68 lower Michigan counties, and thus, constitutes missing information. The covariates for AQI are the intercept and the non-industrialization status of the county ($q_2 = 2$). We take y_{2j} to be normally distributed with mean η_{2j} and fixed standard deviation σ_0, estimated using our data and set at 0.1.

Results for the standard errors and width of HPD sets for the parameters are similar to the complete data case. There are two striking features in this application. First, the extent of urbanization for lung cancer incidence is negative and significant under (I) whereas it is

positive under the other two priors (which may not be reasonable). Second, the regression coefficient for racial segregation (non-white) is significant under (I) and (II) but not under (III). This shows the sensitivity of the subjective elicitation under the missing data setup as well.

6 Conclusion

The Bayesian inferential framework is perhaps the only solution available for analyzing hierarchical spatial multivariate data. In the absence of reliable subjective information, the use of Jeffreys type non-informative priors or diffuse conjugate priors is popular. However, in the context of the hierarchical spatial multivariate GLMMS, we have shown that none of these priors will work; the Jeffreys prior yields a posterior that is improper whereas the diffuse conjugate prior is highly sensitive. This characteristic has also been observed in Natarajan and Kass (2000) in a simpler setup, namely in univariate GLMMs without any spatial components. This led us to elicit priors on the model parameters that will be close to Jeffreys but still yield a proper posterior for inference.

The development of prior elicitation can be thought of as an extension of Natarajan and Kass (2000) in the spatial context. Besides the prior development, we propose some innovative computational techniques for the Gibbs implementation. Suitable transformations are made on the parameters which avoid sampling from restricted domains, thus providing more stability and efficiency in the Gibbs steps. The methodology has been extended to the case of missing responses in the multi-dimensional setup.

We have carried out extensive simulation studies to establish the superiority of the proposed methodology. As we have mentioned in the Introduction, the motivation of this work came from Michigan SEER data analysis. We have provided two real examples briefly merely for illustration. Both examples support the use of the newly proposed prior rather than commonly used Jeffreys or diffuse conjugate priors.

References

Gelfand, A. and Vounatsou, P. (2003), "Proper multivariate conditional autoregressive models for spatial data analysis", *Biostatistics*, 4, 11-25.

Gelman, A., and Rubin, D. (1990). "Inference from iterative simulation using multiple sequences (with discussion)." *Statistical Science*, 7, 457-511.

Jin, X., Carlin, B.P., and Banerjee, S. (2005), "Generalized hierarchical multivariate CAR models for areal data", *Biometrics*, 61, 950-961.

Kim, H.D. Sun, D. and Tsutakawa, R. (2001), "A bivariate Bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model", *Journal of the American Statististical Association*, 96, 1506-1521.

Mardia, K.S. (1988), "Multi-dimensional multivariate Gaussian Markov random fields with application to image processing", *Journal of Multivariate Analysis*, 24, 265-284.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.), London: Chapman and Hall.

Muirhead (1982). *Aspects of Multivariate Statistical Theory*, John Wiley & Sons Canada.

Natarajan, R. and Kass, R (2000), "Reference Bayesian methods for generalized linear mixed models", *Journal of Multivariate Analysis*, 95, 227-237.

Sain, S. (2009), "Parameter estimation for multivariate spatial lattice models", To appear in *Biometrics*.

Sain, S. and Cressie, N. (2007), "A spatial model for multivariate lattice data", *Journal of Econometrics*, 140, 226-259.

Appendix

Proof of Theorem 1: A necessary and sufficient condition for $\boldsymbol{D}$ to be positive definite (pd) is the expression for $\boldsymbol{D}^{-1}$ to be positive definite. Since $\boldsymbol{\Gamma}^{-1/2}$ is pd and hence non-singular, it follows from (10) that $(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{F})$ should be positive definite. The eigenvalues of $(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{F})$ is the same as the eigenvalues of $(\boldsymbol{M} \otimes \boldsymbol{I}_p - \boldsymbol{W} \otimes \boldsymbol{\Lambda})$, which is the collection of all eigenvalues of $\boldsymbol{M} - \lambda_k \boldsymbol{W}$, for $k = 1, 2, \cdots, p$. Now, requiring that $\boldsymbol{M} - \lambda_k \boldsymbol{W}$ be diagonally dominant (which implies positive definiteness), it follows that

$$|\lambda_k| \sum_{l \in N_j} w_{jl} \leq w_{j+} \quad \Rightarrow \quad |\lambda_k| \, w_{j+} \leq w_{j+} \quad \Rightarrow \quad |\lambda_k| \leq 1 \tag{34}$$

for all $k = 1, 2, \cdots, p$.

Proof of Theorem 2: In order to show that the posterior is improper, it is enough to show that the marginal of $\mathbf{y} \equiv (\mathbf{y}_1', \mathbf{y}_2', \cdots, \mathbf{y}_n')'$ does not exist; that is, the integration with

respect to the variable $\boldsymbol{\eta}$ and the parameters $\boldsymbol{H}$, $\boldsymbol{\Gamma}$ and $\boldsymbol{\beta}$ is infinity. The marginal of $\mathbf{y}$ has the expression

$$\begin{aligned}\mathbf{m}(\mathbf{y}) &= \int_{\boldsymbol{\eta}}\int_{\boldsymbol{H}}\int_{\boldsymbol{\Gamma}}\int_{\boldsymbol{\beta}}\left(\prod_{i=1}^{p}\prod_{j=1}^{n} f_i(y_{ij}\,|\,\eta_{ij})\right)\frac{1}{(2\pi)^{np/2}}(\det(\mathbf{D}))^{-1/2}\;\times \\ &\quad\times\exp\left\{-\frac{1}{2}(\boldsymbol{\eta}-\mathbf{X}\boldsymbol{\beta})'\mathbf{D}^{-1}(\boldsymbol{\eta}-\mathbf{X}\boldsymbol{\beta})\right\}\,d\boldsymbol{\eta}\,\pi(\boldsymbol{H})\,d\boldsymbol{H}\,\frac{1}{(\det(\boldsymbol{\Gamma}))^{(p+1)/2}}d\boldsymbol{\Gamma}\,d\boldsymbol{\beta}\end{aligned}$$

We make a change of variable from $\boldsymbol{\eta}\rightarrow\boldsymbol{\epsilon}$ using the transformation $\boldsymbol{\epsilon}=\boldsymbol{\eta}-\boldsymbol{X\beta}$. Next, write $\det(\boldsymbol{D})=\det(\boldsymbol{\Gamma})^n\times g_0(\boldsymbol{H})$ for some function g_0 of $\boldsymbol{H}$, and note that the expression within the exponent can be simplified to $-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Gamma}^{-1}\boldsymbol{S})$ where

$$\boldsymbol{S}=\sum_{j,l=1}^{n}\boldsymbol{H}_{jl}\boldsymbol{\epsilon}_j\boldsymbol{\epsilon}_l'$$

with $\boldsymbol{H}_{jl}\equiv -\,w_{jl}\,\boldsymbol{H}$ if $j\neq l$, and $\boldsymbol{H}_{jj}=w_{j+}\,\boldsymbol{I}_p$. Now, integrating with respect to $\boldsymbol{\Gamma}$, the marginal reduces to

$$\boldsymbol{m}(\boldsymbol{y})=\int_{\boldsymbol{H}}\int_{\boldsymbol{\beta}}\int_{\boldsymbol{\epsilon}}\left(\prod_{i=1}^{p}\prod_{j=1}^{n} f_i(y_{ij}\,|\,\epsilon_{ij}+x_{ij}'\beta_i)\right)\frac{1}{(\det(\boldsymbol{S}))^{n/2}}\,d\boldsymbol{\epsilon}\,d\boldsymbol{\beta}\,(g_0(\boldsymbol{H}))^{-1}\,\pi(\boldsymbol{H})\,d\boldsymbol{H},$$

ignoring proportionality constants. We make a change of variable from $\boldsymbol{\epsilon}\rightarrow\boldsymbol{u}$ defined by $\epsilon_{11}=u_{11}$ and $\epsilon_{ij}=u_{ij}u_{11}$ for $(i,j)\neq(1,1)$ with an associated Jacobian of u_{11}^{np-1}. With this substitution,

$$\boldsymbol{S}=u_{11}^2\,\boldsymbol{S}^*$$

where $\boldsymbol{S}^*=\sum_{j,l=1}^{n}\boldsymbol{H}_{jl}\boldsymbol{U}_{jl}$ where $\boldsymbol{U}_{jl}=V_j\,V_l'$ with $V_1=(1,u_{21},u_{31},\cdots,u_{p1})'$ and $V_j=(u_{1j},u_{2j},\cdots,u_{pj})$ for $j\geq 2$. It follows that

$$\det(\boldsymbol{S})=u_{11}^{2p}\det(\boldsymbol{S}^*)$$

Subsequently,

$$\boldsymbol{m}(\boldsymbol{y})=\int_{\boldsymbol{H}}\int_{\boldsymbol{\beta}}\int_{\boldsymbol{u}}\left(\prod_{i=1}^{p}\prod_{j=1}^{n} f_i(y_{ij}\,|\,u_{ij}^*u_{11}+x_{ij}'\beta_i)\right)\frac{1}{u_{11}}\,\frac{1}{(\det(\boldsymbol{S}^*)^{n/2})}\,d\boldsymbol{u}\,d\boldsymbol{\beta}\,(g_0(\boldsymbol{H}))^{-1}\,\pi(\boldsymbol{H})\,d\boldsymbol{H},$$

where $u_{ij}^*=u_{ij}$ for $(i,j)\neq(1,1)$, $u_{11}^*=1$ and $\boldsymbol{S}^*=\sum_{j,l=1}^{n}\boldsymbol{H}_{jl}\boldsymbol{U}_{jl}$). It follows that the integral with respect to u_{11} diverges around $u_{11}=0$ proving that $\boldsymbol{m}(\boldsymbol{y})=\infty$.

Proof of Theorem 3: Without loss of generality, we take the first q_i rows of $\tilde{\boldsymbol{X}}_{\mathcal{C}_i}$ to be the linearly independent rows. It follows that the marginal of $\boldsymbol{y}$,

$$\boldsymbol{m}(\boldsymbol{y})\leq C_0\int_{\boldsymbol{F}}\int_{\boldsymbol{\Gamma}}\int_{\boldsymbol{\beta}}\int_{\boldsymbol{\eta}}\left(\prod_{i=1}^{p}\prod_{j=1}^{q_i} f_{ij}(y_{ij}\,|\,\epsilon_{ij}+x_{ij}'\beta_i)\right)\,f_0(\boldsymbol{\eta}\,|\,\boldsymbol{\beta}\,\boldsymbol{F},\boldsymbol{\Gamma})\,d\boldsymbol{\beta}\,d\boldsymbol{\eta}\,\pi(\boldsymbol{F})\,d\boldsymbol{F}\,\pi_{UV}(\boldsymbol{\Gamma})\,d\boldsymbol{\Gamma}$$

where C_0 is a constant depending on A and B and the submatrix $X_i^* = (x_{i1}^{'}, x_{i2}^{'}, \cdots, x_{iq_i}^{'})'$ is of dimension $q_i \times q_i$ with full rank q_i. Making a change of variable from $\beta_i \rightarrow r_i = (\epsilon_{i1}, \epsilon_{i2}, \cdots, \epsilon_{iq_i})' + X_i^* \beta_i$ for $i = 1, 2, \cdots, p$, condition (25) implies that

$$\begin{aligned} \boldsymbol{m}(\boldsymbol{y}) &\leq \prod_{i=1}^{p} \det(X_i^*) \int_{\boldsymbol{\epsilon}} \int_{\boldsymbol{F}} \int_{\boldsymbol{\Gamma}} f_0(\boldsymbol{\epsilon} \,|\, \boldsymbol{F}, \boldsymbol{\Gamma})\, d\boldsymbol{\epsilon}\, \pi(\boldsymbol{F})\, d\boldsymbol{F}\, \pi_{UV}(\boldsymbol{\Gamma})\, d\boldsymbol{\Gamma} \\ &\leq \prod_{i=1}^{p} \det(X_i^*) < \infty \end{aligned}$$

since the integrands all integrate to finite numbers; f_0 integrates to 1 with respect to $\boldsymbol{\epsilon}$, and $\pi(\boldsymbol{F})$ and $\pi_{UV}(\boldsymbol{\Gamma})$, respectively integrates to 1 with respect to $\boldsymbol{F}$ and $\boldsymbol{\Gamma}$ since they are proper priors.

Derivation of the Jacobian from $(\boldsymbol{\Gamma}, \boldsymbol{Q}) \rightarrow \boldsymbol{B}$: We consider three transformation steps: $(\boldsymbol{\Gamma}, \boldsymbol{Q}) \overset{(a)}{\rightarrow} (\boldsymbol{\Gamma^{-1}}, \boldsymbol{Q}) \overset{(b)}{\rightarrow} (\boldsymbol{L}, \boldsymbol{Q}) \overset{(c)}{\rightarrow} \boldsymbol{B}$, where (a) is an inverse transformation, (b) is Cholesky decomposition and (c) is QR decomposition. The Jacobian of each transformation can be obtained as follows (Muirhead (1982)): (a) $d\boldsymbol{\Gamma} = \det(\boldsymbol{\Gamma^{-1}})^{-(p+1)}\, d\boldsymbol{\Gamma}^{-1}$, (b) $d\boldsymbol{\Gamma^{-1}} = 2^p \prod_{i=1}^{p} L_{ii}^{p+1-i}\, d\boldsymbol{L'}$, and (c) $d\boldsymbol{B} = \prod_{i=1}^{p} L_{ii}^{p-i}\, d\boldsymbol{L'}(\boldsymbol{Q}\, d^*\boldsymbol{Q'})$, where $(\boldsymbol{Q}\, d^*\boldsymbol{Q'})$ defines the Haar measure on the set of $p \times p$ orthogonal matrices. Thus, defining $d\boldsymbol{Q} \propto (\boldsymbol{Q} d^* \boldsymbol{Q'})$, we have $d\boldsymbol{\Gamma} d\boldsymbol{Q} = \det(\boldsymbol{\Gamma^{-1}})^{-(p+1)}\, 2^p \prod_{i=1}^{p} L_{ii}^{p+1-i}\, d\boldsymbol{L'} d\boldsymbol{Q} \propto \det(\boldsymbol{\Gamma^{-1}})^{-(p+1)} \prod_{i=1}^{p} L_{ii} d\boldsymbol{B} \propto \det(\boldsymbol{\Gamma^{-1}})^{-p-1/2} d\boldsymbol{B} \propto \det(\boldsymbol{B B'})^{-(p+1/2)} d\boldsymbol{B}$.

θ	$\boldsymbol{\beta}$			$\boldsymbol{\Gamma}$			$\boldsymbol{F}$			$\boldsymbol{\eta}$		
Prior	(I)	(II)	(III)	(I)	(II)	(III)	(I)	(II)	(III)	(I)	(II)	(III)
RMAD	0.253	0.250	0.258	0.799	1.660	2.994	0.956	0.982	0.990	0.557	0.533	0.558
MSE	0.017	0.017	0.017	0.045	0.113	0.336	0.166	0.167	0.169	0.066	0.071	0.085
CP	0.861	0.894	0.925	0.896	0.618	0.000	0.881	0.859	0.859	0.874	0.933	0.932
W	0.383	0.430	0.478	0.546	0.642	0.027	1.236	1.161	1.107	0.787	0.943	1.040

Table 1: Deviations measures for averages over $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{F}$ and $\boldsymbol{\eta}$ for the complete data case.

θ	$\boldsymbol{\beta}$			$\boldsymbol{\Gamma}$			$\boldsymbol{F}$			$\boldsymbol{\eta}$		
Prior	(I)	(II)	(III)	(I)	(II)	(III)	(I)	(II)	(III)	(I)	(II)	(III)
RMAD	0.256	0.254	0.263	1.037	1.782	2.992	0.916	0.928	0.929	0.539	0.559	1.132
MSE	0.018	0.018	0.018	0.074	0.143	0.336	0.155	0.157	0.158	0.071	0.075	0.090
CP	0.911	0.936	0.951	0.923	0.634	0.000	0.931	0.890	0.882	0.889	0.936	0.934
W	0.445	0.495	0.534	0.710	0.746	0.026	1.294	1.210	1.157	0.840	0.977	1.059

Table 2: Deviations measures for averages over $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{F}$ and $\boldsymbol{\eta}$ the missing data case.

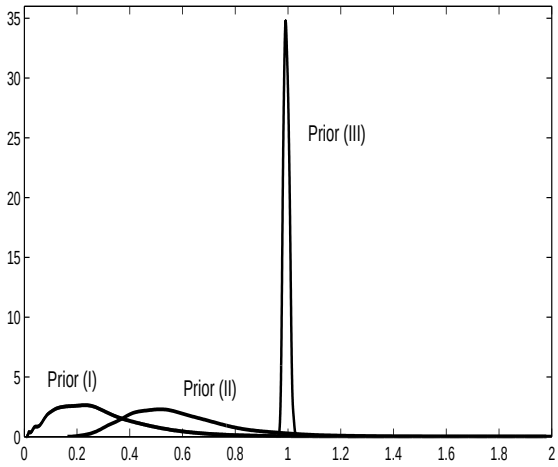
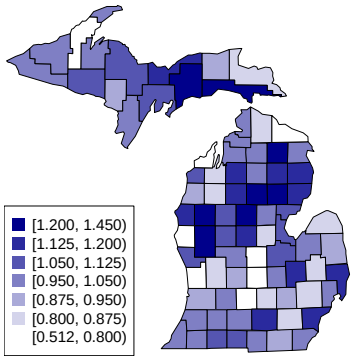
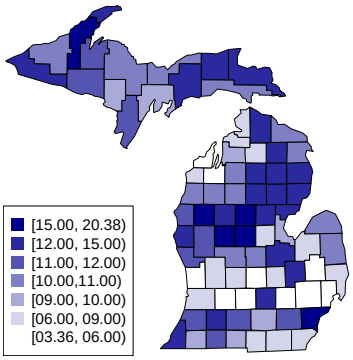


Figure 1: Posterior densities for Γ_{22} corresponding to the three prior choices.

Figure 2: Observed SMR of Lung Cancer and Poverty in Michigan



(a) Obs SMR



(b) Poverty(%)