ABSTRACT

Bayesian Spatial Misclassification Model for Areal Count Data with Applications to COVID-19

Jinjie Chen, Ph.D.

Chairperson: James D. Stamey, Ph.D.

As of December 14, 2020, there have been more than 72.1 million confirmed cases, of which more than 1.61 million have died of COVID-19 globally. In the United States, there are more than 16,200,000 confirmed cases and 299,000 COVID-19-related deaths, the most cases, and deaths of any country. However, even with the huge number of confirmed diagnoses, the public burden of the pandemic is still masked by under-reporting and misclassification. Based on the Bayesian spatial model and Poisson regression, we study two topics, aiming to provide a flexible quantitative approach for simulating and correcting the under-reporting and misclassification of COVID-19 at the US state level. Topic 1 quantifies under-reporting rates with Poisson-logistic regression, combined with the prior information derived from the results of the SARS-CoV-2 antibody sampling study, and then estimates the true case of COVID-19 in each state of the US. Topic 1 also combines the Besag-York-Mollié 2 (BYM2) model to correct the bias of parameter estimation caused by ignoring the spatial auto-correlation. Topic 2 proposes a bivariate Bayesian spatial misclassification model, which can simultaneously calibrate the misclassification of two counts of the same area (for example, state or county). Deaths related to COVID-19 are considered to be misclassified to other causes and vice versa (although the latter case is relatively fewer). In addition, because the number of deaths at the state level shows obvious spatial similarity, BYM2 random effects are included to explain the variability beyond the covariates. Our model was applied to state-level COVID-19 deaths and other deaths, achieving satisfactory results that can be a reference for estimating the true COVID-19 deaths. Topic 3 proposes and discusses the determination of sample size based on skew-normal distribution. This method adopts Bayesian intensive simulation to overcome limitations of closed-form approximation and normality assumption while ensuring sufficient statistical power and nominal coverage of confidence interval (or credible set). Our approach demonstrates good performance and application prospects.

Bayesian Spatial Misclassification Model for
Areal Count Data with Applications to COVID-19

by

Jinjie Chen, B.A., M.S.

A Dissertation

Approved by the Department of Statistical Science

James D. Stamey, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

James D. Stamey, Ph.D., Chairperson

Joon Jin Song, Ph.D.

Jane L. Harvill, Ph.D.

Phil Young, Ph.D.

Accepted by the Graduate School
May 2021

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Throughout the writing of this dissertation, I have received a great deal of support and assistance.

I would first like to thank my supervisors, Dr. Stamey and Dr. Song, whose expertise was invaluable in formulating the research questions and methodology. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level.

Also I would like to thank our program director Dr.Harvill and Dr. Phil Young for being my dissertation committee members.

I would like to acknowledge my colleagues from my internship at Novartis for their wonderful collaboration. I would particularly like to thank my supervisor at Novartis, Dr. John Seaman, for his patient support and for all of the opportunities I was given during my internship.

Finally, I could not have completed this dissertation without the support of my loved family and my friends.

Thank you all.

CHAPTER ONE

A Computationally Efficient Bayesian Hierarchical Spatial Model for Under-reported Count Data

1.1 Introduction

Count data are commonly encountered in a variety of fields such as epidemiology, criminology, and sociology. In many cases, the data recording process is error-prone. Individuals or responsible agencies may fail to report the true count due to a variety of reasons. For instance, under-reporting often occurs when sensitive questions are asked in surveys, such as self-reported use of illegal drugs. But more often, it is attributed to the imperfect process of data collection. The outbreak and fast global spread of COVID-19 in early 2020 have been one of the most far-reaching public health events in modern history. It is believed the cases of COVID-19 were significantly under-reported at the early stage of the outbreak. Some governments thus missed the best opportunity to contract the spread of COVID-19. Under-reporting is problematic because it results in biased statistical inferences and may lead to poor decisions. Therefore, it is essential to correctly estimate the reporting rates and corresponding uncertainty for making a better decision about resource allocation. Recently, Stoner, Economou, and Drummond Marques da Silva (2019) proposed a Bayesian hierarchical approach to model and correct the under-reporting of tuberculosis incidence in Brazil. Their framework relies only on an informative prior distribution for the mean reporting rate. The model is implemented in a fully Bayesian framework, which is highly flexible and provides a posterior predictive distribution for unobserved true counts, thus quantifying the uncertainty of the under-reporting.

Often, observational count data is collected over space. Hence it is also important to account for spatial dependency. A commonly used spatial model for count data is Besag-York-Mollie (BYM) (Besag, York, & Mollié, 1991) which is a lognormal Poisson model that decomposes the regional spatial effect into a sum of a structured ICAR (intrinsic conditional auto-regressive) component $\boldsymbol{\phi}$ and an unstructured component $\boldsymbol{\theta}$ (pure overdispersion). In BYM the structured and unstructured components can not be seen independently from each other and are thus not identifiable. The lack of identifiability often leads to poor convergence and makes the choice of hyperpriors more difficult (e.g., Bernardinelli, Clayton, & Montomoli, 1995; MacNab, 2011; Wakefield, 2007). Simpson et al. (2017) proposed a new modification of the commonly known BYM model, termed the BYM2 model, consisting of a single-precision parameter and one mixing parameter. BYM2 addresses both the identifiability and scaling issue of the BYM model (Riebler, Sørbye, Simpson, & Rue, 2016). Recently Morris et al. (2019) report an effective way to execute BYM2 in Stan language.

In this chapter, we investigate a flexible Bayesian spatial hierarchical model that can correct the under-reporting of areal count data through partial prior information. In addition, we consider the potential spatial confounding effect which is col-linearity between spatial random effect and fixed effect in spatial generalized linear mixed models (SGLMM). It is known that the spatial confounding may lead to biased estimates of the fixed effects and inflated the variances of the estimates. Many researchers have discussed this issue, proposing methods to eliminate spatial confounding, such as the most commonly used restricted spatial regression (RSR), and principal components analysis (PCA) methods. Methods based on RSR constrain the spatial effect to the space orthogonal to fixed effects, thereby eliminating estimation bias and variance inflation. However, as pointed out by Hanks, Schliep, Hooten, and Hoeting (2015), the constraint places an excessively strong and impractical assumption that the spatial effect and the fixed effect are completely orthogonal. In the case of model mismatch (random effects are from SGLMM and correlated with fixed effects), RSR may bias the posterior mean and inappropriately shrink the variances. Hanks et al. (2015) proposed a posterior predictive method to correct the underestimation of the posterior variance and ensure the coverage of the posterior credible sets to achieve nominal coverage.

It is worth noting that PCA methods (variants of RSR) are computationally efficient than SGLMM, especially when the number of spatial areas N is large. However, our simulation shows that the computation time of the BYM2 model (reparameterization of Besag York Mollié Model) implemented in Stan is comparable to PCA methods. Hanks et al. (2015) provides a nice discussion about the selection between SGLMM and PCA.

The outline of the remainder of this chapter is as follows. In Section 1.2, we introduce the Poisson-Logistic (Pogit) model and explain how under-reporting is fully modeled across space. Next, we discuss a solution to the non-identifiability problem in Section 1.3. In Section 1.4, we discuss spatial confounding, explain how it influences the recovery of regression parameters, and discuss approaches to alleviate confounding. In Section 1.5, we describe the proposed full Pogit model. We study the inference and prediction performance of the various approaches via a simulation study in Section 1.6 and apply BYM2 and PCA to a real-world data set in Section 1.7. We conclude with a discussion of our work in Section 1.8.

1.2 Pogit Model For Under-reporting Count Data

1.2.1 Pogit Model with BYM Structure

A popular model to account for under-reporting is the Pogit model proposed by Winkelmann and Zimmermann (1993). The model consists of a binomial component for the observed counts, $\boldsymbol{z}$, conditional on the underlying unobserved true counts $\boldsymbol{y}$

assumed to follow Poisson distribution,

$$z_i \sim \text{Binomial}(\pi_i, y_i) \tag{1.1}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \boldsymbol{W}\boldsymbol{\beta}_s \tag{1.2}$$

$$y_i \sim \text{Poisson}(E_i\lambda_i) \tag{1.3}$$

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X}\boldsymbol{\gamma}_s \tag{1.4}$$

where π_i models the reporting rates for region i, E_i and λ_i represent the offset and true incidence rate, respectively. Vectors $\boldsymbol{\pi} = \pi_1, \cdots, \pi_N$ and $\boldsymbol{\lambda} = \lambda_1, \cdots, \lambda_N$ are related to linear predictors (may or may not include random effects) through logistic and logarithm link functions, respectively. We denote the covariates for the reporting procedure as $\boldsymbol{W}$ and those for the true counts as $\boldsymbol{X}$, and correspondingly denote $\boldsymbol{\beta}_s$ and $\boldsymbol{\gamma}_s$ as the coefficient vectors for logistic and Poisson regressions, respectively.

Count data observed over space often exhibits spatial clustering or similarity. Spatial dependency for adjacent areas sharing similar characteristics that can be modeled by spatial auto-correlation. A commonly used Bayesian spatial model for areal data is the Conditionally Autoregressive (CAR) model proposed by Besag (1974), where the spatial effect of a particular region depends on the effects of all neighboring regions. Besag et al. (1991) further proposed a more flexible model known as BYM, where the spatial random effect is decomposed into two components: a structured component based on the Intrinsic Conditionally Autoregressive (ICAR) model, and an unstructured Gaussian component. Let $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_N)$ denote the ICAR component, and $\boldsymbol{\theta} \sim \text{Normal}(0, \tau)$ be the unstructured effect, then equation (1.2) can be rewritten as,

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X}\boldsymbol{\gamma}_s + \boldsymbol{\phi} + \boldsymbol{\theta}. \tag{1.5}$$

Both CAR and ICAR models smooth spatial random effect by pooling information from neighborhoods. Given a set of N regions with well-defined boundaries, an $N \times N$ adjacency matrix $\boldsymbol{A}$ can be defined to depict the neighbor relationship. To be specific, the (i, j) entry of $\boldsymbol{A}$, denoted as ω_{ij}, is set to 1 if region i and j are neighbors, or set to 0 otherwise. In addition, the diagonal elements of $\boldsymbol{A}$, ω_{ii}, are all 0 as a region is not its own neighbor. Consequently $\boldsymbol{A}$ is a symmetric matrix with all 0 on the diagonal, and is typically sparse. The conditional distribution of the random effect ϕ_i for region i is specified in terms of a weighted average of its neighbors and a overall precision parameter τ,

$$\phi_i | \phi_j, j \neq i, \tau \sim N\left(\alpha \sum_{i \sim j} \omega_{ij} \phi_j, \tau^{-1}\right), \tag{1.6}$$

where $\alpha \in (0, 1)$ is the proximity parameter to remedy the singularity of the overall precision matrix for the joint distribution of $\boldsymbol{\phi}$. ICAR is a special case of CAR when $\alpha = 1$, see Banerjee, Carlin, and Gelfand (2014). The corresponding joint distribution of $\boldsymbol{\phi}$ modeled by CAR is

$$\boldsymbol{\phi} \sim MN\left(\mathbf{0}, [\tau(\boldsymbol{D} - \alpha \boldsymbol{A})]^{-1}\right), \tag{1.7}$$

where $\boldsymbol{D}$ is an $n \times n$ diagonal matrix whose diagonal element d_{ii} equals the number of neighbors for region i. In ICAR, the above condition produces an improper distribution as setting $\alpha = 1$ creates a singular matrix $(\boldsymbol{D} - \boldsymbol{A})$. Furthermore, the joint distribution is non-identifiable as adding any constant to all of the elements of $\boldsymbol{\phi}$ leaves the joint distribution unchanged. Adding a constraint such as $\sum \phi_i = 0$ resolves this problem.

Compared with ICAR, CAR suffers from at least two problems. Firstly, α is difficult to interpret (Carlin, Banerjee, et al., 2003) and needs to take a value above 0.9. Secondly, in the MCMC sampling, α is assigned a prior and updated in each iteration, while for ICAR α is constant and needs not be updated, thus CAR is much more computationally expensive than ICAR especially when N is large. Our simulation studies show that the computation time for sampling from a multivariate Normal distribution with the precision matrix $\boldsymbol{Q} = \boldsymbol{D} - \alpha \boldsymbol{A}$ increases exponentially as the number of regions N does. Morris et al. (2019) also argued that with the Stan language, computing burden of $\det(\boldsymbol{Q})$ for CAR increases at a speed of $O(N^3)$, for example, when $N = 1000$, it will take a billion operations for CAR. In contrast, for ICAR the $\det(\boldsymbol{Q})$ is a constant 0 and thus can be dropped from the log likelihood. It can be shown that for ICAR, the log probability density of $\boldsymbol{\phi}$ is proportional to

$$\frac{n}{2}\log(\det(\boldsymbol{Q})) - \frac{1}{2}\boldsymbol{\phi}'\boldsymbol{Q}\boldsymbol{\phi} \tag{1.8}$$

where $\det(\boldsymbol{Q}) = 0$ as $\boldsymbol{Q}$ is singular, and a constant term can be dropped from the log probability density in Stan, meaning fitting the ICAR model will be much faster than fitting ICAR (Hoffman & Gelman, 2014), in particular when N is large.

1.2.2 A Computationally Efficient Implementation of ICAR in Stan

Morris et al. (2019) proposed an efficient implementation of the ICAR component in Stan. They show that encoding adjacency entries as either 0 or 1 in $\boldsymbol{A}$ is equivalent to an undirected graph with a set of N nodes and a set of edges, one edge per pair of non-zero entries $\{i, j\}$ and $\{j, i\}$. The joint specification of the ICAR random vector $\boldsymbol{\phi}$ centered at 0 with a common variance 1 leads to the pairwise difference formulation:

$$p(\boldsymbol{\phi}) \propto \exp\left(-\frac{1}{2}\sum_{i\sim j}(\phi_i - \phi_j)^2\right). \tag{1.9}$$

To obtain this result, consider the probability density function of ICAR:

$$p(\phi) \propto (2\pi)^{-n/2}|[\boldsymbol{D} - \boldsymbol{A}]^{-1}|^{1/2}\exp\left(-\frac{1}{2}\boldsymbol{\phi}'[\boldsymbol{D} - \boldsymbol{A}]\boldsymbol{\phi}\right), \tag{1.10}$$

dropping the constant terms $(2\pi)^{-n/2}$ and $|[\boldsymbol{D} - \boldsymbol{A}]^{-1}|^{1/2}$ leads to:

$$p(\boldsymbol{\phi}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\phi}'[\boldsymbol{D} - \boldsymbol{A}]\boldsymbol{\phi}\right), \tag{1.11}$$

taking the natural log yields:

$$\begin{aligned}
\log p(\phi) &= -\frac{1}{2}\boldsymbol{\phi}'[\boldsymbol{D} - \boldsymbol{A}]\boldsymbol{\phi} + c \\
&= -\frac{1}{2}\left(\sum_{i,j} \phi_i [\boldsymbol{D} - \boldsymbol{A}]_{i,j} \phi_j\right) + c \\
&= -\frac{1}{2}\left(\sum_{i,j} \phi_i \phi_j \boldsymbol{D}_{i,j} - \sum_{i,j} \phi_i \phi_j \boldsymbol{A}_{i,j}\right) + c \\
&= -\frac{1}{2}\left(\sum_{i} \phi_i^2 \boldsymbol{D}_{i,i} - \sum_{i \sim j} 2\phi_i \phi_j\right) + c \\
&= -\frac{1}{2}\left(\sum_{i \sim j} (\phi_i^2 + \phi_j^2) - \sum_{i \sim j} 2\phi_i \phi_j\right) + c \\
&= -\frac{1}{2}\left(\sum_{i \sim j} (\phi_i - \phi_j)^2\right) + c. \qquad (1.12)
\end{aligned}$$

Equation (1.12) holds because $\boldsymbol{D}$ is a diagonal matrix with $\boldsymbol{D}_{i,i}$ equal to the number of neighbors of region i and the off-diagonal entries equal 0, and the matrix $\boldsymbol{A}$ is a symmetric matrix with entries $\omega_{i,j} = 1$ for neighbors i and j, 0 otherwise. As mentioned above, $\boldsymbol{\phi}$ is non-identifiable as adding any constant to all of the elements of $\boldsymbol{\phi}$ results in the same distribution. Adding the constraint $\sum_i \phi_i = 0$ resolves the non-identifiability problem.

We follow Morris et al. (2019) to encode the neighbor relations as a set of edges and two groups of nodes, indexed as node 1 and node 2: node 1 holds the set of indexes corresponding to ϕ_i and node 2 holds the indexes corresponding to ϕ_j. Encoding ICAR in this way needs less memory and computation time than the typical way of sampling from a multivariate normal distribution when the precision matrix is sparse. To illustrate it, we construct a 3 × 2 regular grid as a simple example as shown in Figure 1.1. This map consists of a single component with neighbor relations: $(1 \sim 2,\ 1 \sim 3,\ 2 \sim 4,\ 3 \sim 4,\ 3 \sim 5,\ 4 \sim 6,\ 5 \sim 6)$. The corresponding adjacency

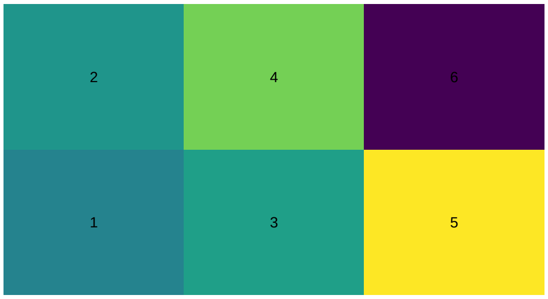


Figure 1.1: A map over 6 regions.

matrix $\boldsymbol{A}$ and neighbor counts matrix $\boldsymbol{D}$ are:

$$\boldsymbol{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}, \boldsymbol{D} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

The precision matrix $\boldsymbol{Q}$ is defined as

$$\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{A} = \begin{pmatrix} 2 & -1 & -1 & 0 & 0 & 0 \\ -1 & 2 & 0 & -1 & 0 & 0 \\ -1 & 0 & 3 & -1 & -1 & 0 \\ 0 & -1 & -1 & 3 & 0 & -1 \\ 0 & 0 & -1 & 0 & 2 & -1 \\ 0 & 0 & 0 & -1 & -1 & 2 \end{pmatrix} \tag{1.13}$$

There are 6 regions labeled 1 to 6 and 7 edges as shown in table 1.1. In this example, encoding equation (1.13) requires a total of 14 elements (see table 1.1), while the precision matrix has 36 elements, 16 of which are 0. In general, if the average number of pairwise neighbors is M for an area consisting of $N \times N$ regions, the spaces to store the nodes are $N \times M$ where $M \leq N$ but the full precision matrix requires

Table 1.1: Example of node 1 and 2 labeling for edges

	edge1	edge2	edge3	edge4	edge5	edge6	edge7
node1	1	1	2	3	3	4	5
node2	2	3	4	4	5	6	6

$N \times N$ spaces. Moreover, M increases much more slowly as N does thus the pairwise encoding becomes more and more efficient. For example $M = 3.6$ for a 10×10 regular grid and increases slightly to 3.87 when the size expands to 30×30.

1.2.3 Reparameterization of BYM Model - BYM2

Fitting the BYM model using MCMC methods is difficult since either component of the model can account for most or all of the individual-level variance. Morris et al. (2019) present an implementation of the BYM2 model (Riebler et al., 2016; Simpson et al., 2017), a reparameterization of the BYM model. In BYM2, the sum of the structured component $\boldsymbol{\phi}$ and the unstructured component $\boldsymbol{\theta}$ are re-parameterized as,

$$\boldsymbol{\phi} + \boldsymbol{\theta} = \sigma \left(\sqrt{\rho}\tilde{\boldsymbol{\theta}} + \sqrt{(1-\rho)/s}\tilde{\boldsymbol{\phi}} \right), \tag{1.14}$$

where σ represents the overall standard deviation, ρ controls the proportion of the variance modeled by the ICAR $\tilde{\boldsymbol{\phi}}$ scaled by s such that $var(\tilde{\phi_i}) \approx 1$ and the heterogeneous effect $\tilde{\boldsymbol{\theta}}$ has a fixed standard deviation 1. A critical condition that $var(\theta_i) \approx var(\phi_i) \approx 1$ is required for the assumption that σ is 1. Riebler et al. (2016) and Morris et al. (2019) recommend scaling the model so the geometric mean of $var(\theta_i)$ is 1. The scaling factor is computed from the adjacency matrix using the "inla.scale.model()" function available from the R-INLA package (Lindgren, Rue, et al., 2015). Function "inla.scale.model()" returns a sparse matrix scaled so the geometric mean of the marginal variances of $\boldsymbol{Q}$ is one.

1.3 Spatial Confounding and Remedies

Spatial confounding is the multicollinearity between spatially varying covariates and spatial random effects. Fitting a spatial regression model usually focuses on estimating the fixed effects of interest while taking into account the spatial correlation, but adding spatial random effects may cause a significant change in the posterior mean and variance of the regression coefficients. Reich, Hodges, and Zadnik (2006) analyzed the scenario and proposed a diagnostic for the posterior variance inflation and an approach to mitigate spatial confounding. Hodges and Reich (2010) noted that the confounding may be strong enough that a significant fixed effect under a non-spatial linear model becomes insignificant when spatial effects are included. Paciorek (2010) shows that spatial confounding can lead to biased estimates, in particular when spatial random effects with a large effective range of spatial autocorrelation are smoothed.

A commonly used approach to mitigate potential spatial confounding is the restricted spatial regression (RSR) models, in which the spatial random effect $\boldsymbol{\phi}$ is constrained to be orthogonal to the fixed effects in $\boldsymbol{X}$ (Guan & Haran, 2018; Hodges & Reich, 2010; Hughes & Haran, 2013; Reich et al., 2006). Let $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$, then the linear predictor of the SGLMM is rewritten as

$$\begin{aligned}
f(\boldsymbol{\mu}) &= \boldsymbol{X\beta} + \boldsymbol{\phi} + \boldsymbol{\theta} && (1.15)\\
&= \boldsymbol{X\beta} + \boldsymbol{P\phi} + (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\phi} + \boldsymbol{\theta}\\
&= \boldsymbol{X\beta} + \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'\phi} + (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\phi} + \boldsymbol{\theta}\\
&= \boldsymbol{X}\left[\boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'\phi}\right] + (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\phi} + \boldsymbol{\theta}\\
&= \boldsymbol{X\delta} + (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\phi} + \boldsymbol{\theta} && (1.16)\\
&= \boldsymbol{X\delta} + \boldsymbol{\phi}^* + \boldsymbol{\theta}, && (1.17)
\end{aligned}$$

where $\boldsymbol{\phi}^* \sim N(\boldsymbol{0}, \tau_s\boldsymbol{Q}), \boldsymbol{\theta} \sim N(\boldsymbol{0}, \tau_h\boldsymbol{I})$, and

$$\boldsymbol{\delta} \equiv \boldsymbol{\beta} + (\boldsymbol{X'X})^{-1}\boldsymbol{X'\phi} \qquad (1.18)$$

are referred to as the unconditional regression coefficients by Hanks et al. (2015) as opposed to the conditional regression coefficients $\boldsymbol{\beta}$. $\boldsymbol{\phi}^* \equiv (\boldsymbol{I}_n - \boldsymbol{P})\boldsymbol{\phi}$ are the spatial random effects that are orthogonal to $\boldsymbol{X}$. Hanks et al. (2015) demonstrate that the RSR model imposes strong assumptions on the fixed effects in the spatial model. They show that under model misspecification, the posterior credible sets for $\boldsymbol{\delta}$ obtained by RSR will be inappropriately underestimated when the true model is SGLMM. They also propose a posterior predictive approach to expand RSR credible intervals based on equation (1.16). If we denote $\widetilde{\boldsymbol{\delta}}^{(t)}$ as the t^{th} MCMC posterior sample of $\boldsymbol{\delta}$, then the t^{th} posterior sample for $\boldsymbol{\beta}$ is

$$\widetilde{\boldsymbol{\beta}}^{(t)} = \widetilde{\boldsymbol{\delta}}^{(t)} - (\boldsymbol{X'X})^{-1}\boldsymbol{X'}\widetilde{\boldsymbol{\phi}^*}^{(t)}. \tag{1.19}$$

Hughes and Haran (2013) noticed that the RSR model also reduces the number of model parameters from $n + p + 1$ to $n + 1$. To see this, consider the spectral decomposition of $\boldsymbol{I} - \boldsymbol{P}$, yielding the orthogonal basis $\boldsymbol{L}_{n\times(n-p)}$, equation (1.15) can be written as

$$f(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{L\eta} + \boldsymbol{\theta}. \tag{1.20}$$

The prior for the random effect $\boldsymbol{\eta}$ reduces to

$$p(\boldsymbol{\eta}|\tau) \propto \exp\{-\frac{\tau}{2}\boldsymbol{\eta}'\boldsymbol{Q}_R\boldsymbol{\eta}\}, \tag{1.21}$$

where $\boldsymbol{Q}_R = \boldsymbol{L'QL}$.

The RSR model can only slightly reduce the dimension of the random effects, and ignores the underlying graph that represents the spatial structure (labeled by M). To this end, Hughes and Haran (2013) proposed a principal component analysis approach based on the Moran operator and re-parameterization of RSR model. The Moran operator is modified from the numerator of Moran's I statistic, a popular measure of spatial dependence proposed by Moran (1950),

$$\boldsymbol{I_X}(\boldsymbol{A}) = \frac{n}{\boldsymbol{1'A1}} \frac{\boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{11'}/n)\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{11'}/n)\boldsymbol{y}}{\boldsymbol{y'}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}}.$$

Replacing $\mathbf{I} - \mathbf{1}\mathbf{1}'/n$ with $\boldsymbol{I} - \boldsymbol{P}$ results in the Moran operator for $\boldsymbol{X}$ with respect to M,

$$\boldsymbol{I_X}(\boldsymbol{A}) = \frac{n}{\mathbf{1}'\boldsymbol{A}\mathbf{1}} \frac{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{A}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}}{\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}}. \tag{1.22}$$

Boots and Tiefelsdorf (2000) showed that the eigenvectors of the Moran operator exhaust all possible spatial patterns that can arise on the underlying graph M. Moreover, since the repulsion of random effects is not desired in most real world applications, the eigenvectors corresponding to negative eigenvalues can be discarded directly, resulting in an almost 50% reduction of the dimension. Hughes and Haran (2013) showed that a much greater reduction is possible in practice, with 50-100 eigenvectors being enough for most data sets. By only retaining $r \ll n$ eigenvectors of the Moran operator, the so-called sparse areal mixed model (SAMM) is

$$f(\boldsymbol{\mu}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{R}_{n\times r}\boldsymbol{\eta}_r + \boldsymbol{\theta},$$

where $\boldsymbol{\eta_r}$ is an r-dimensional random vector. Here, $\boldsymbol{\eta_r}$ has a multivariate normal prior with mean $\mathbf{0}$ and precision matrix $\tau\boldsymbol{Q}_r$, where $\boldsymbol{Q}_r = \boldsymbol{R}'\boldsymbol{Q}\boldsymbol{R}$, and $\boldsymbol{R}$ contains the first $r \ll n$ eigenvectors of the Moran operator. The computation speed of SAMM is expected to improve significantly compared with RSR.

1.4 Full Pogit Model with BYM, BYM2, PCA and PPD

The model adopted by Stoner et al. (2019) ignores that the spatial confounding may have an impact on the recovery of the regression coefficients and the estimation of the reporting rate, as they may focus more on the accuracy of prediction rather than coefficient inference. R package NIMBLE was used to implement the MCMC sampling via AFSS sampler to optimizes the random walk. The mixing of the chains requires 800K iterations and a 400K burn-in which is pretty slow. Here, we make use of Stan language for all MCMC sampling, and use BYM2 or SAMM to replace BYM. Our simulations show that only a few thousand iterations are needed to achieve convergence. Moreover, we compared the performances of the Pogit with BYM2 (Pogit-BYM2) and Pogit with SAMM (Pogit-SAMM) in the presence and absence of spatial confounding. For Pogit-BYM2, equation (1.4) becomes

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X}\boldsymbol{\gamma}_s + \sigma\left(\sqrt{\rho}\tilde{\boldsymbol{\theta}} + \sqrt{(1-\rho)/s}\tilde{\boldsymbol{\phi}}\right), \tag{1.23}$$

while for Pogit-SAMM equation (1.4) is replaced by

$$\log(\boldsymbol{\lambda}) = \boldsymbol{X}\boldsymbol{\gamma}_s + \boldsymbol{R}_{n\times r}\boldsymbol{\eta}_r + \boldsymbol{\theta}. \tag{1.24}$$

Pogit model equations (1.1)-(1.4) treat the true but unknown counts as latent variables to account for the bias when estimating the Poisson regression coefficients. This leads to slow-mixing of the MCMC chains. In addition, since the true count is unknown, (1.1)-(1.4) is difficult to express directly in Stan language. Conveniently, the following reparameterization is recommended based on summing out the latent variables:

$$z_i \sim Poisson(\pi_i\lambda_i) \tag{1.25}$$

$$y_i - z_i \sim Poisson((1-\pi_i)\lambda_i) \tag{1.26}$$

Equation (1.25) is much more efficient in terms of mixing speed, and samples of the latent variable $\boldsymbol{y}$ can be generated using equation (1.26). However, equation (1.25) is over parameterized as the same observed counts z_i could result from either a high reporting rate, π_i, multiplied by a low Poisson mean, λ_i, or vise versa. This means infinite groups of parameters will satisfy the same likelihood function of $\boldsymbol{z}$ in the absence of prior information or any completely reported observations.

If additional information in the data is available, for example, a gold standard for a small set of completely reported counts, informative priors can be derived to eliminate non-identifiability. On the other hand, it is often impractical or even impossible to obtain any complete data. In such cases, expert opinion or results of similar studies can be used to determine informative priors. In the following simulations, we will assume a mildly informative prior for the intercept of the logistic regression.

1.5 A Simulation Study

In this section, we apply models (1.25) and (1.26) to simulated under-reported Poisson data. 100 data sets were simulated based on the same spatial structure, which is a 20 by 20 regular grid. The design matrix $\boldsymbol{X} = [\boldsymbol{x}\ \boldsymbol{y}]$, where $\boldsymbol{x} = (x_1, \cdots, x_{400})'$, $\boldsymbol{y} = (y_1, \cdots, y_{400})'$ are the standardized vertices of x- and y-coordinates. We let $\boldsymbol{\gamma_s} = (2, 2)$. Under this experimental design, $[\boldsymbol{x}\ \boldsymbol{y}]$ will be confounded with ICAR random effects. For the under-reporting (logistic) regression, we set $\boldsymbol{\beta_s} = (0, 2)$ with a single covariate drawn from $U(-1, 1)$.

We first simulated random effects from the BYM model (ICAR spatial effects). Since ICAR is improper, that is, the precision matrix $\tau\boldsymbol{Q}$ is a singular matrix, we can not get random effects from ICAR directly. Instead we will follow the method of Guan and Haran (2018) to simulate ICAR effects using the eigen components of $\tau\boldsymbol{Q}$. Let (λ_i, e_i) denote the eigenpairs of $\tau\boldsymbol{Q}$, and we simulate $a_i \sim N(0, \lambda_i^{-1})$ for $\lambda_i \neq 0$. Then $\boldsymbol{\phi} = \sum_i a_i e_i$ has the desired distribution. We set $\tau = 4$ for spatial effects, and $\tau_e = 9$ for the overall precision of the random errors. A level plot of simulated spatial random effects is shown in Figure 1.2. Since in this case, ICAR is the true model for spatial effects, we will adopt the posterior prediction recommended by Hanks et al. (2015) to adjust for the inappropriate narrowness of posterior credible sets for γ_s when using the PCA approach.

We fit the simulated data set with five models: the standard Pogit with no spatial effects (NS for short), the Pogit-BYM (BYM for short), the Pogit-BYM2 (BYM2 for short), the Pogit-PCA (PCA for short) and the Pogit-PCA-PPD model (PPD for short) with 100 eigenvectors. The simulation result for one single data set with confounded ICAR effects is listed in Table 1.2. Figure 1.3 shows the corresponding

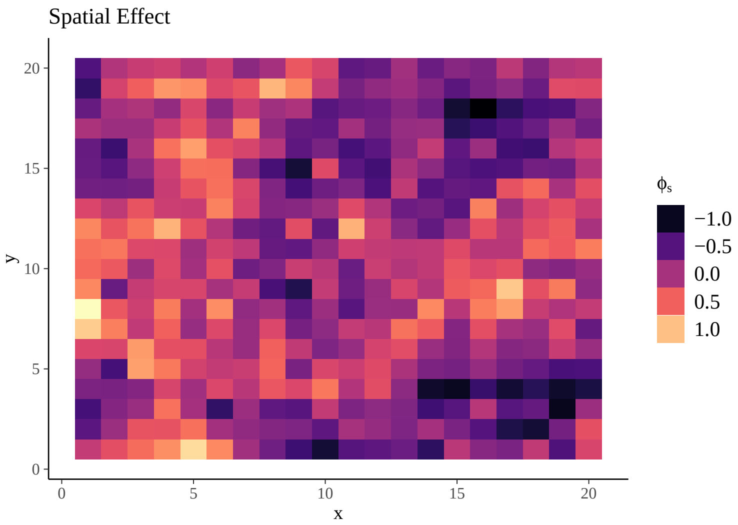


Figure 1.2: Spatial random effects of one simulated data.

point estimates and 95% confidence intervals (CIs) of the coefficients. Firstly, for the regression coefficient parameters (β_s, γ_s), the 95% CIs obtained by NS are too narrow, while those obtained by BYM are inappropriately too wide (in particular for γ_s) due to the existence of spatial confounding. Secondly, PCA and PPD models have almost same estimates of β_s, but for γ_s, PPD corrects the inappropriate narrowness. Thirdly, the result obtained by the BYM2 model is surprisingly comparable to that of PPD, which indicates that the reparameterization of BYM may benefit the correction for spatial confounding naturally. In order to further compare the performance of different models in the case of confounding, Table 1.3 lists results across 100 data sets. Two measures are considered for comparison: mean square error (MSE) and 95% CI coverage. BYM2 and PPD are comparable and perform best. Table 1.4 lists the average CI length where we see BYM2 outperforms PPD by providing more accurate confidence interval (shorter length).

Table 1.2: Model comparison, point and 95% CI estimates for one single data set with confounding spatial effect.

			Mean (95% CI)		
	NS	BYM	BYM2	PCA	PPD
β_0	-0.12 (-0.46, 0.24)	0.24 (-0.47, 0.90)	0.26 (-0.30, 0.77)	0.35 (-0.22, 0.90)	0.35 (-0.22,,0.90)
β_1	1.82 (1.48, 2.20)	2.25 (1.63, 2.93)	2.17 (1.66, 2.72)	2.27 (1.70, 2.84)	2.27 (1.70,,2.84)
γ_1	2.18 (2.04, 2.33)	2.03 (0.72, 3.38)	1.97 (1.48, 2.47)	1.93 (1.74, 2.15)	2.17 (1.36, 3.04)
γ_2	2.13 (1.98, 2.28)	1.68 (0.34, 3.04)	1.89 (1.40, 2.38)	1.95 (1.75, 2.18)	2.03 (1.14, 2.91)

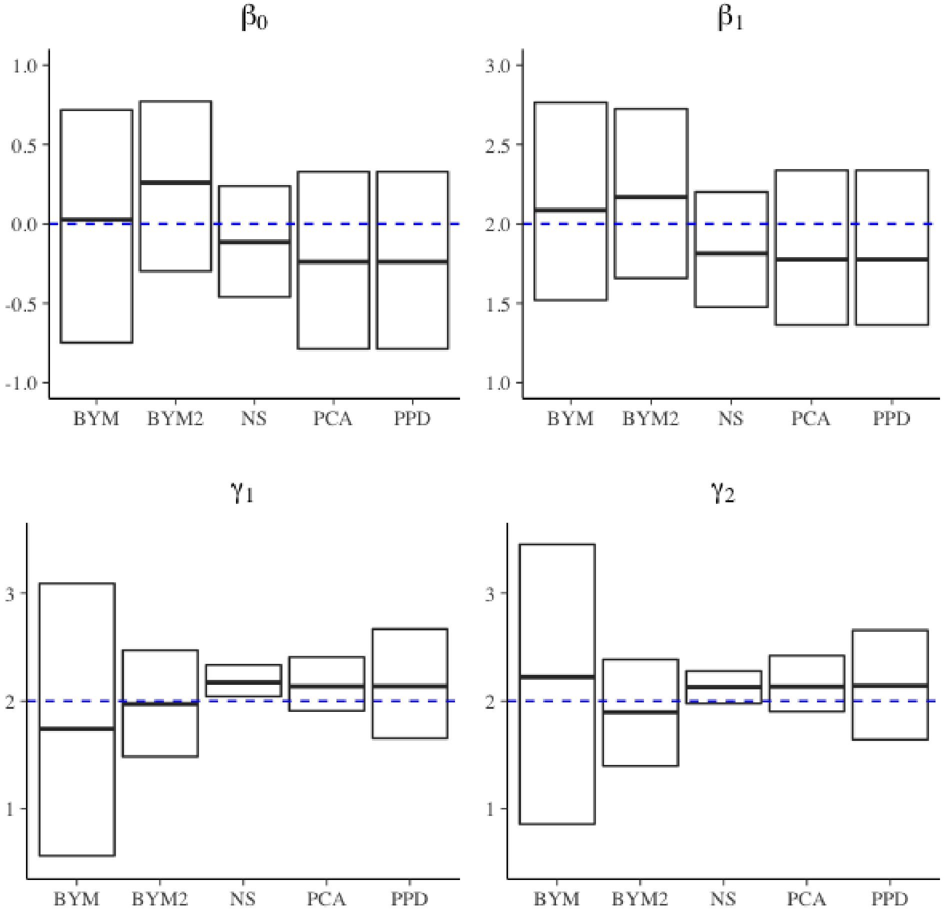


Figure 1.3: Boxplots illustrating inference for a simulated Pogit dataset with ICAR effects: under-reporting coefficients in the first row; Poisson regression coefficients in the second row.

Table 1.3: Model comparisons (1): coverage and mean squared error(MSE) for 100 confounded Pogit data sets with $n = 400$.

	NS	BYM	BYM2	PCA	PPD
β_0(coverage)	0.048(39%)	-0.052(97%)	0.019(95%)	-0.021(94%)	-0.021(94%)
β_0 MSE	0.679	0.299	0.308	0.309	0.309
β_1(coverage)	2.113(48%)	2.036(97%)	2.015(94%)	2.021(93%)	2.021(93%)
β_1 MSE	0.506	0.265	0.262	0.266	0.266
γ_1(coverage)	2.082(25%)	1.981(100%)	2.028(89%)	2.018(50%)	2.009(97%)
γ_1 MSE	0.518	0.425	0.407	0.412	0.426
γ_2(coverage)	2.074(16%)	1.982(100%)	2.028(94%)	2.003(44%)	2.005(96%)
γ_2 MSE	0.562	0.407	0.397	0.412	0.449

Table 1.4: Model comparisons (2): CI length for 100 confounded Pogit data sets with $n = 400$.

	NS	BYM	BYM2	PCA	PPD
β_0	0.571	1.549	1.154	1.248	1.248
β_1	0.613	1.258	0.995	1.055	1.055
γ_1	0.265	2.908	1.343	0.545	1.945
γ_2	0.266	2.913	1.348	0.544	1.943

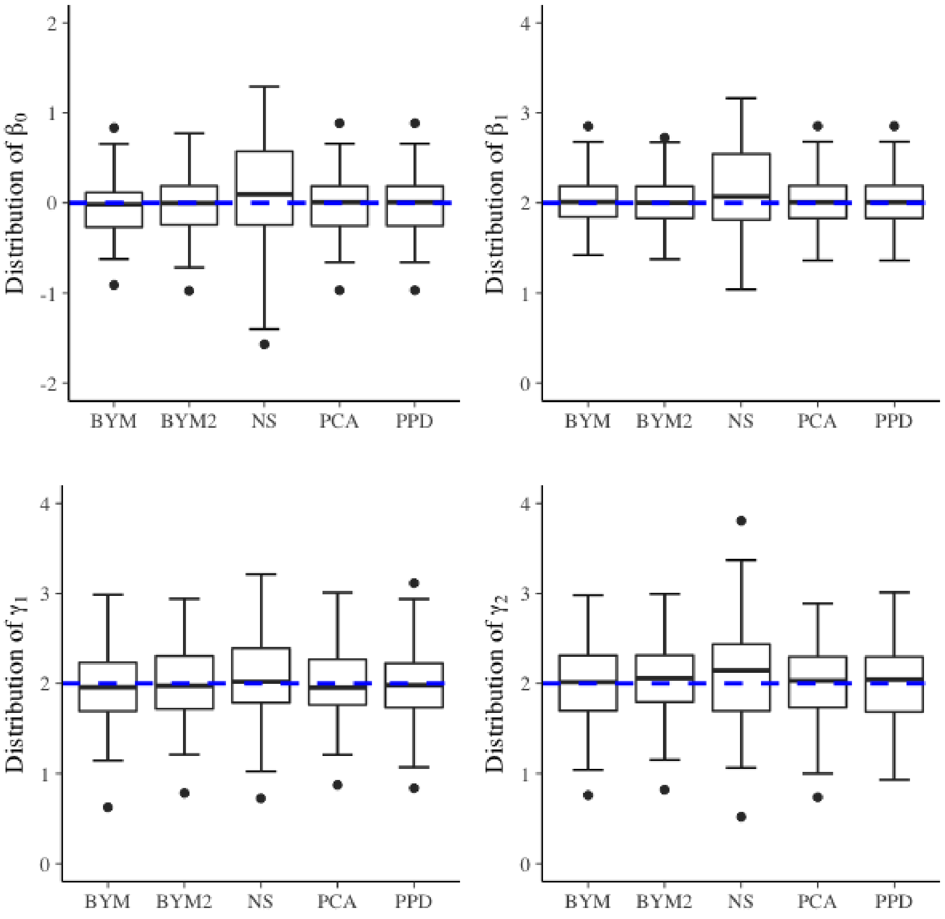


Figure 1.4: Simulation study of 100 Pogit datasets with confounded ICAR effects, $n = 400$: distribution of $\boldsymbol{\beta}_s$ and $\boldsymbol{\gamma}_s$ for Pogit-BYM2, Pogit-SAMM, Pogit-PPD and Pogit-NS. All distributions center around the truth. Pogit-BYM2, Pogit-SAMM and Pogit-PPD are comparable, while Pogit-NS is inferior in the sense of larger variance and bias for all parameters.

Table 1.5 shows the results of 100 simulated non-confounded data sets. We see that in terms of MSE and coverage, the performance of BYM and BYM2 are both inferior to PCA and PPD. PCA is better than PPD in coverage (the former is closer to the nominal value of 95%). In addition, Table 1.6 shows that the length of CI given by the PCA method is also better than that of PPD and BYM2. Therefore, we believe that the PCA model is optimal in a non-confounded situation. The result is as expected since the spatial effect and X are completely orthogonal. The correction in the PPD model is redundant and only increases the uncertainty of the posterior samples. Figures 1.5 and 1.6 show that although BYM2 does not perform as well as PCA and PPD in this case, it still presents a significant improvement over BYM.

Table 1.5: Model comparisons I: coverage and MSE for non-confounded Pogit data with $n = 400$.

	NS	BYM	BYM2	PCA	PCA-PPD
β_0(coverage)	0.019(54%)	-0.062(100%)	-0.010(100%)	-0.043(95%)	-0.043(95%)
β_0 MSE	0.401	0.257	0.263	0.240	0.240
β_1(coverage)	2.022(62%)	0.000(100%)	1.993(97%)	1.978(94%)	1.978(94%)
β_1 MSE	0.334	0.232	0.24	0.214	0.214
γ_1(coverage)	2.062(59%)	1.985(100%)	2.023(100%)	1.998(97%)	1.999(100%)
γ_1 MSE	0.179	0.178	0.143	0.102	0.098
γ_2(coverage)	2.072(71%)	1.983(100%)	2.020(100%)	2.015(94%)	2.019(100%)
γ_2 MSE	0.185	0.197	0.162	0.108	0.114

Table 1.6: Model comparisons 2: CI length for 100 non-confounded Pogit data sets with $n = 400$.

	NS	BYM	BYM2	PCA	PPD
β_0	0.581	1.580	1.180	0.986	0.986
β_1	0.613	1.267	1.012	0.890	0.890
γ_1	0.270	2.896	1.483	0.438	0.658
γ_2	0.270	2.892	1.481	0.439	0.658

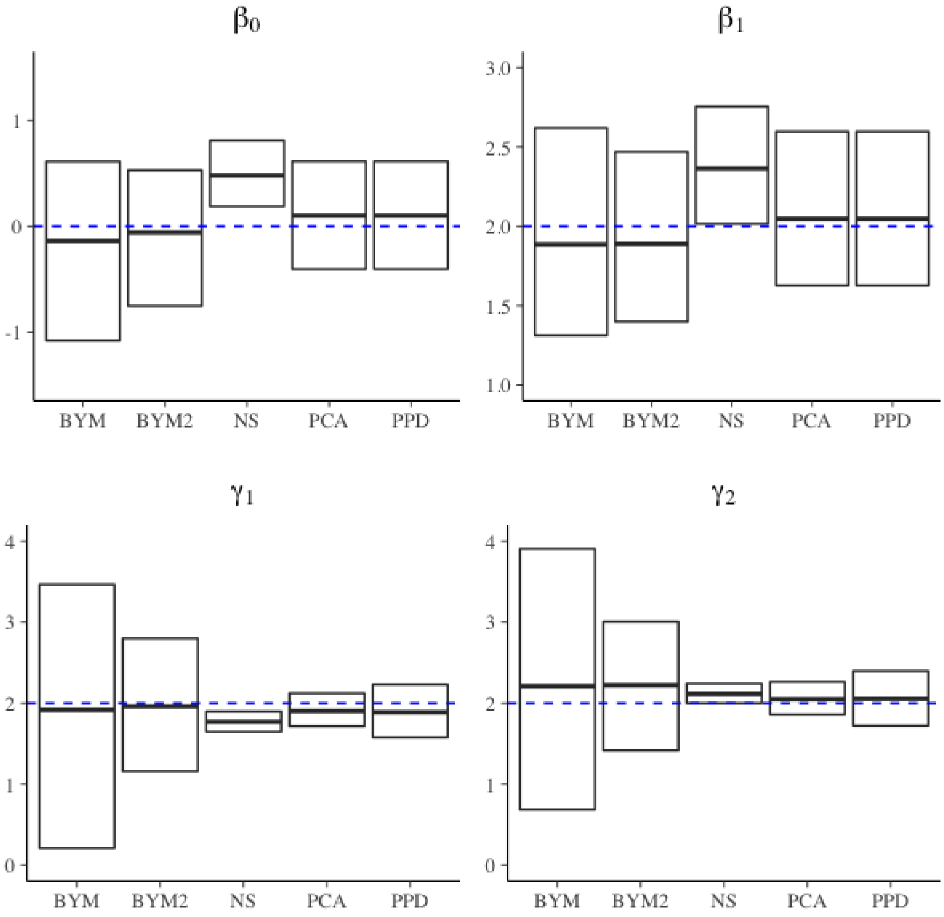


Figure 1.5: Boxplots illustrating inference for a simulated Pogit dataset with orthogonal spatial effects: Under-reporting coefficients in the first row; Poisson regression coefficients in the second row.

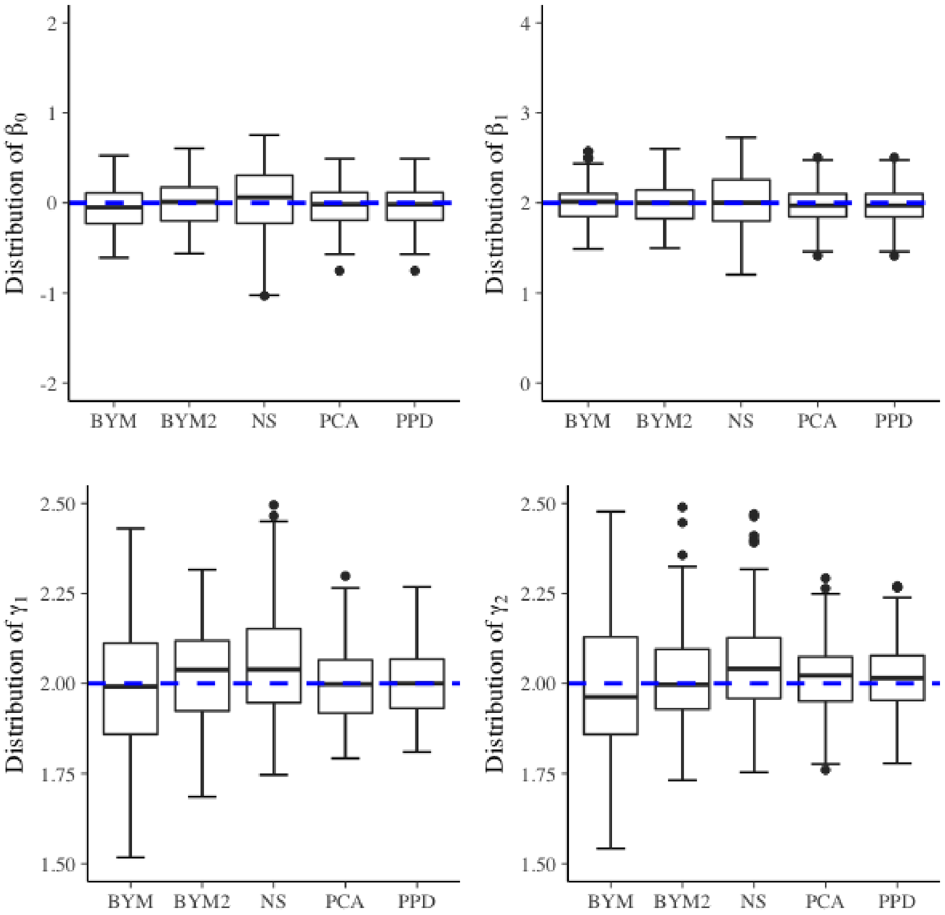


Figure 1.6: Simulation study of 100 Pogit datasets with non-confounded spatial random effects, $n = 400$: distribution of $\boldsymbol{\beta}_s$ and $\boldsymbol{\gamma}_s$ for Pogit-BYM2, Pogit-SAMM, Pogit-PPD and Pogit-NS. All distributions center around the truth. Pogit-BYM2, Pogit-SAMM and Pogit-PPD are comparable, while Pogit-NS is inferior in the sense of larger variance and bias for all parameters.

We conclude our simulation study with a brief discussion. Pogit- BYM2, Pogit-PCA, and Pogit-PPD are all variants of the Pogit-BYM model used by Stone et al. Due to the complexity of the Pogit model itself, the spatial confounding(if exists) will make the problem even more complicated. Our simulation study shows that if spatial confounding does exist, Pogit-BYM2 is better than other models, while in spite of compelling advantages in computation speed, Pogit-PCA introduces a too strong assumption, resulting in biased estimates of regression coefficients. The bias can be corrected by PPD though slightly inferior to Pogit-BYM2. In contrast, if the spatial confounding does not exist (that is, the spatial effect and the fixed effect are orthogonal), the incorrect adoption of Pogit-PCA will lead to underestimation of the impact of spatial covariates on the response, as the correlation between spatial covariates and spatial random effects might inflate the variance. In the absence of a compelling reason to assume that the random effects should be orthogonal to the fixed effects, conservative Pogit-BYM2 or Pogit-PPD are recommended to avoid the risk arisen by model misspecification. Overall, Pogit-BYM2 shows satisfactory performance under both situations, so in the next section, we will apply the Pogit-BYM2 model to a real-world data set.

1.6 Application

Addressing COVID19 is a pressing health concern. Inadequate knowledge about the extent of the coronavirus disease 2019 (COVID-19) epidemic has challenged public health responses and planning. Most reports of confirmed cases rely on polymerase chain reaction-based testing of symptomatic patients. These estimates of confirmed cases may fail to account for individuals who have unwittingly recovered from the infection, individuals with mild or no symptoms, and individuals with symptoms who have not been tested due to limited availability of tests (Sood et al., 2020). Time-lag bias and failure of proactive contact tracing and containment will also lead to underestimation of the number of people who are infected with the virus. According to Rajgor, Lee, Archuleta, Bagdasarian, and Quek (2020), one unique situation allowed for an accurate estimate of COVID-19 incidence rate, specifically, the outbreak of COVID-19 among passengers on board the Diamond Princess cruise ship who were quarantined between Jan 20, and Feb 29, 2020. This scenario provided a population living in a defined territory without most other confounders, such as imported cases, defaulters of screening, or lack of testing capability. 3711 passengers and crews were on board, of whom 705 became sick and tested positive for COVID-19 and seven died. Accordingly, the infection rate is as high as 19% in this particular situation. Of course, the special structure of the cruise ship and the dense crowds could have accelerated the spread of COVID-19, thus the infection rate, in this case, is estimated to be higher than community transmission on average, but it still indicates that the infection rate of COVID-19 is much higher than the reported numbers in many countries including the US, in particular in the early stage of the Pandemic. Many studies support this argument. For example, in a community seroprevalence study in Los Angeles County, the prevalence of antibodies to SARS-CoV-2 was 4.65% (bootstrap CI $2.52\% - 7.07\%$) indicating that approximately $367,000$ ($198,890 - 557,998$) adults had SARS-CoV-2 antibodies, which is substantially greater than the 8430 cumulative number of confirmed infections in the county by April 10 (Sood et al., 2020). In other words, the case reporting rate was only 2.30% ($1.51\% - 4.24\%$). Another antibodies study of SARS-CoV-2 by Bendavid et al. (2020) implies that by early April, the case reporting rate is approximately 1.85% ($1.10\% - 4.00\%$) in Santa Clara County, CA. Hortacsu, Liu, and Schwieg (2020) estimate $4\% - 14\%$ ($1.5\% - 10\%$) of actual infections that had been reported in US up to March 16, accounting for an assumed reporting lag of 8(5) days. Ribeiro, Bernardes, et al. (2020) estimate a 12.99% reporting rate in Brazil by March 20.

We collected the state-level daily accumulated COVID-19 cases of U.S. from `https://covidtracking.com/`. Figure 1.7 shows the state-level map of total confirmed COVID-19 cases per 10,000 people in the USA by April 30,2021, 2020. A total of 1,071,003 confirmed cases was reported in the 48 contiguous US states and Washington DC. Among the studied states, the highest incidence of COVID-19 was in New York State (304,372 cases in total, 15.4 per 100,000 population). According to the global Moran's I statistic (Moran's $I = 0.21$, $p = 0.00006$), the state-level COVID-19 prevalence presents highly positive auto-correlations or clustered patterns. We also collected state-by-state risk factors from `https://www.americashealthrankings.org`. Table 1.7 gives a detailed description for each variable. It is believed that insufficient COVID-19 testing capability at the early stage of the pandemic was the key factor leading to under-reporting. Historical testing data by states are also collected from `covidtracking.com`. Moreover, box plots of the risk factors presented in Figure 1.8 show that there are few extreme cases in the covariates.

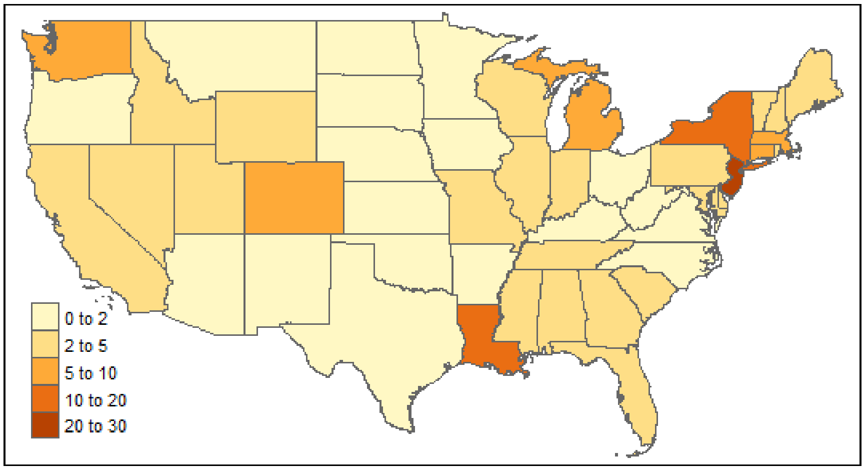


Figure 1.7: States-level confirmed COVID-19 cases/10K population by March 31, 2020.

The different scales among the covariates can cause convergence problems and other difficulties in model fitting. We decided to standardize the variables before fitting, mainly for the following purposes: 1. optimizing the MCMC sampling, our experience shows that standardization can accelerate the model convergence process, in particular for a complex model like ours; 2. easing interpretation, for example, after standardizing the "Testing" variable, the intercept of the logit part can be directly explained as the reporting rate at the logit level given mean testing level; 3. facilitating the identification of important risk factors; 4. reducing multicollinearity.

Table 1.7: Risk factors used for US COVID-19 cases.

Variable description	Variable name
Total Population	Pop
Population Density (per Sq. Mile)	Popdensity
Percent of Persons Without Insurance	Uninsured
Percent Physically Inactive Persons (20 Years and Over)	Inactive
Percent Obese Persons (20 Years and Over)	Obesity
Percent Current Smokers (Persons 18 Years and Over)	Smoking
Percentage of adults who reported binge drinking	Excessive_Drinking
Air pollution - particulate matter	AirPollution
Number of deaths due to drug injury per 100K	Drug_death
Multidimensional Deprivation Index	MDI
The total number of testing	Testing

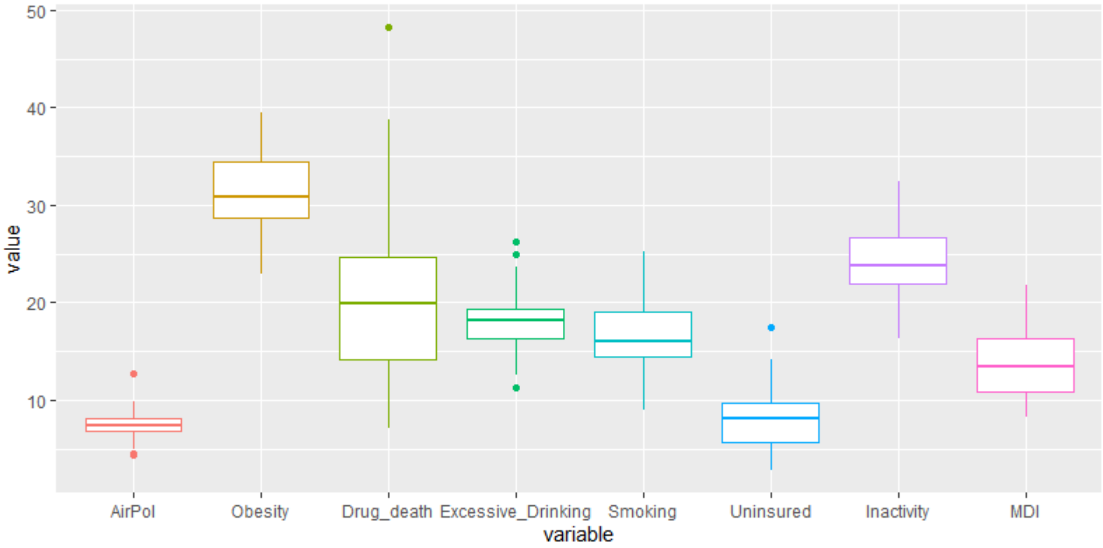


Figure 1.8: Distributions of the potential covariates.

Table 1.8: Summary statistics of potential covariates.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
AirPol	4.400	6.800	7.400	7.476	8.200	12.800	1.450
Obesity	22.900	28.700	30.900	31.461	34.400	39.500	3.864
Drug_death	7.200	14.200	19.900	20.782	24.700	48.300	8.866
Excessive_Drinking	11.300	16.400	18.200	18.171	19.400	26.300	3.101
Smoking	9.000	14.500	16.100	16.606	19.000	25.200	3.324
Uninsured	2.800	5.700	8.100	8.090	9.700	17.500	2.988
Inactivity	16.400	22.000	23.800	24.165	26.700	32.400	3.836
MDI	8.300	10.800	13.400	13.980	16.400	21.800	3.914
Popdensity	6.000	52.000	106.000	424.327	231.000	11011.000	1566.865

1.6.1 Models for Comparison

Besides the proposed full model (equations 1.1-1.4) labeled as M4, three reduced models (labeled as M1, M2 and M3 below) are also fitted for comparison,

$$\begin{aligned} \text{M1 Naive Poisson: } z_i &\sim \text{Poisson}(E_i\lambda_i), \\ \log(\boldsymbol{\lambda}) &= \boldsymbol{X}\boldsymbol{\gamma}_s + \boldsymbol{\theta}. \end{aligned}$$

$$\begin{aligned} \text{M2 Under-reporting only: } z_i &\sim \text{Binomial}(\pi_i, y_i), \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \boldsymbol{W}\boldsymbol{\beta}_s, \\ y_i &\sim \text{Poisson}(E_i\lambda_i), \\ \log(\boldsymbol{\lambda}) &= \boldsymbol{X}\boldsymbol{\gamma}_s + \boldsymbol{\theta}. \end{aligned}$$

$$\begin{aligned} \text{M3 Spatial only: } z_i &\sim \text{Poisson}(E_i\lambda_i), \\ \log(\boldsymbol{\lambda}) &= \boldsymbol{X}\boldsymbol{\gamma}_s + \boldsymbol{\phi} + \boldsymbol{\theta}. \end{aligned}$$

1.6.2 Priors

For the regression parameters, we assume diffuse normal prior

$$\boldsymbol{\beta}_s \sim N(0, 10^2), \ s = 1, \cdots, k-1 \tag{1.27}$$

$$\boldsymbol{\gamma}_s \sim N(0, 10^2), \ s = 0, 1, \cdots, j-1 \tag{1.28}$$

where, k and j are the column numbers of $\boldsymbol{W}$ and $\boldsymbol{X}$, respectively. A relative informative prior is needed for the intercept of the logistic regression β_0. We assume that the mode of reporting rates π_i is around 0.1, and that $Pr(\pi_i > 0.3)$ is tiny (i.e. $1 - Pr(\pi_i <= 0.3) = 0.0001$) which results in a $beta(7, 55)$ prior on π_i. If covariates in $\boldsymbol{W}$ are all standardized, β_0 can be interpreted as the log-odds of reporting rate at the average level of covariates. In this case if we assume $\text{logit}(p_0) = \beta_0$, it is reasonable to directly assign $beta(7, 55)$ prior on p_0, and implicitly induce a prior on β_0. Note that a Jacobian adjustment is required for the log likelihood statement in Stan language (Carpenter et al., 2017).

1.6.3 Results

We fitted the four models described in section 1.6.1 to the state-level Covid-19 cases. Point estimates and 90% CIs of regression coefficients are presented in Figure 1.9. As M1 and M2 ignore the spatial dependency, the confidence interval of the regression coefficient is inappropriately narrow, leading to the conclusion that all covariates are statistically significant, although M2 takes into account the under-reporting of the data, resulting in significantly different point estimates from M1. In contrast, both M3 and M4 have spatial structure (BYM2), so the confidence intervals is much wider than those of M1 and M2, which is in line with expectations. In this case, only obesity (M3 and M4) and death from drug use (M4) are statistically significant at 90% confidence level, reflecting the huge influence of the spatial dependency on the estimation of regression coefficients. In addition, we found that both M2 and M4 models support a positive correlation between the covariate "Testing" and the reporting rate, with coefficient estimate of 0.54 (90% CI: 0.20-0.95) by M4. In other words, it is reasonable to believe that as of April 2020, there exist varying degrees of under-reporting in continental states of the United States, and the severity of under-reporting is negatively correlated with the standardized number of tests per capita.

To see it, we present the posterior estimates of reporting rate for all 49 states in Figure 1.10.

The full model (M4) indicates that at the early stage of the pandemic (as of April 31, 2020), Rhode Island has the highest reporting rate, about 48.75% (95% CI: 13.97%-89.07%), followed by Massachusetts, with a reporting rate of about 39.36% (95% CI: 12.15%-75.28%), New York State follows with a reporting rate of approximately 31.74% (95% CI: 10.83%-62.69%). By contrast, Kansas State has the lowest reporting rate of 7.75% (95% CI: 3.64-13.44%), while South Carolina has the second-lowest rate of 7.97% (95% CI: 3.76-13.69%).

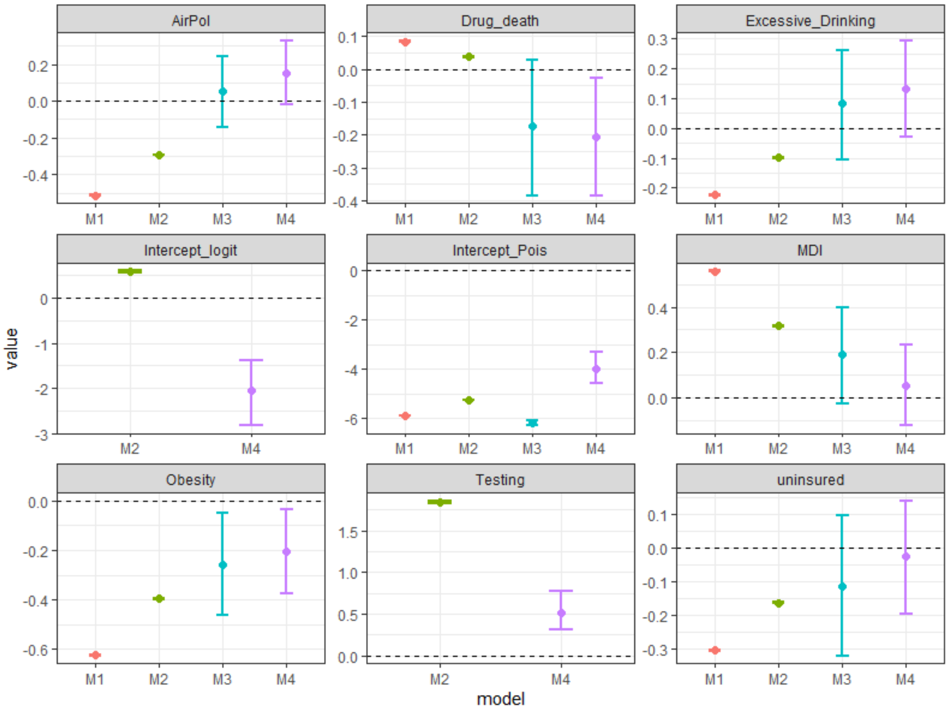


Figure 1.9: Point estimates and 90% CIs of coefficients fitted by four models described in section 1.6.1.

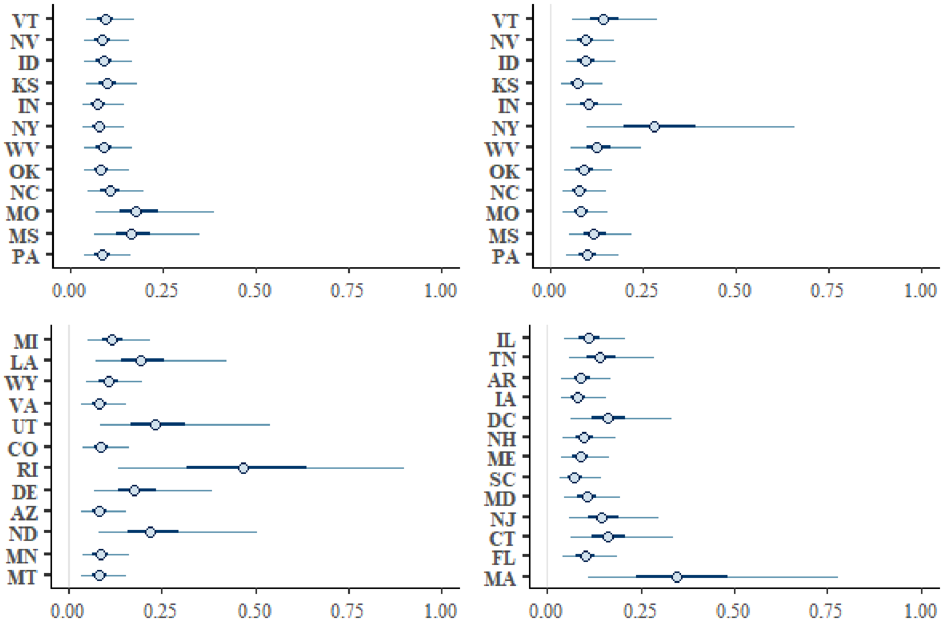


Figure 1.10: Point estimates and 95% CIs of state-level Covid-19 reporting rates.

Figure 1.11 shows a clear, monotonically increasing (estimated) relationship between covariate "Testing" and the probability of reporting π_i. The 95% credible interval does not incorporate a horizontal line that implies no relationship. Overall, states with very low testing z-scores (< -1) have approximately one-sixth of the reporting probability of ones with high testing z-scores (> 3).

Table 1.9 lists reported and estimated COVID-19 cases in 10 states (5 most and 5 least). It suggests that potentially tens of thousands of cases were unreported. For example, 304,372 COVID-19 cases were reported in New York state as of April of 2020, while M4 estimates over 1.25 million people had been infected by then.

Figure 1.12a compares the empirical distribution of the observed COVID-19 cases to the distributions of 200 replicated data sets, '$yrep$', from the posterior predictive distribution, and the 200 posterior predictive samples are centered around the empirical density, showing no systematic divergence. This is indicative of a good fit for the model. Figure 1.12b compares the median of the observed COVID-19 cases

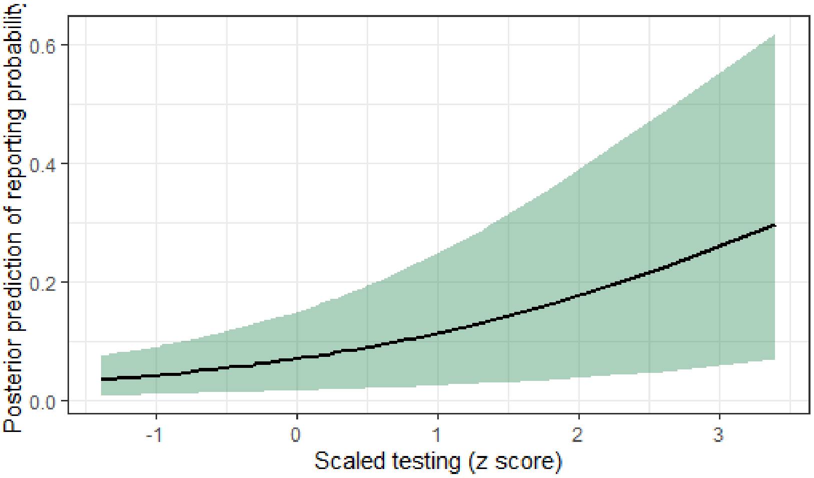


Figure 1.11: Posterior mean predicted effect of testing ability on the reporting probability of COVID-19, with associated 95% credible interval.

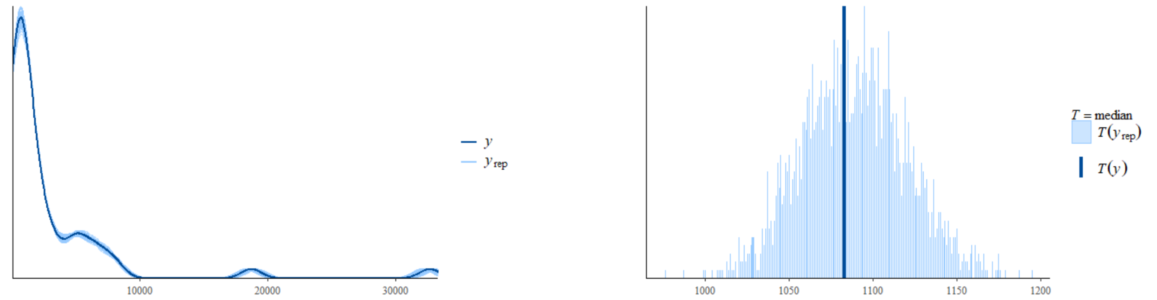
Table 1.9: Comparison between reported and predicted cases for 5 most and 5 least states

State	Reported	Estimate	2.50%	97.50%
North Dakota	1067	5592	2413	12496
Wyoming	559	5922	2781	12788
Montana	453	6347	2978	13474
Vermont	866	6937	3228	15170
West Virginia	1118	9976	4704	21851
Pennsylvania	45763	512674	241643	1085573
Illinois	52918	534930	252330	1136623
California	48917	609516	288815	1301314
New Jersey	118652	918559	427005	1990627
New York	304372	1254965	540881	2829135

y to medians of 200 replicated data '*yrep*' from posterior samples. We see $T(y)$ is roughly centered at the histogram of 200 predicted medians, again indicating a good fit.

1.7 Conclusion

In this chapter, we investigate a flexible Bayesian hierarchical spatial model for correcting the under-reporting of count data and apply it to the state-level COVID-19 case in U.S. as of April 30, 2020. The model proposed here is based on "Pogit" model,



(a) Compare the empirical distribution of the data y to the distributions of 200 replicated data $yrep$.

(b) Compare the median the data y to medians of 200 replicated data $yrep$ from posterior samples.

Figure 1.12: Model checking.

which combines Poisson regression and logistic regression the former of which is used to represent the true counting process, and the latter is used to quantify the reporting rate of each state. One key flexibility of the model is that it can estimate the true reporting rate of any sub-region by borrowing information from other regions, given moderately informative prior information on average reporting rate. In addition, the model considers the spatial dependency among states, through the BYM2 structure. As a reparameterization of the commonly used BYM structure, the BYM2 model solves the unidentifiable issue of spatial random effects and random errors, resulting in great improvement of MCMC convergence.

Chapter 1 also discusses the impact of spatial confounding on the Pogit model. Through simulation, the pros and cons of the four different approches are compared with and without spatial confounding. We conclude that the model mismatch significantly reduce the accuracy of parameter estimation. We suggest that the BYM2-based spatial structure is most appropriate when it is uncertain whether spatial confounding exists.

When applying the Pogit-BYM2 model (M4) to state-level COVID-19 data, we confirmed that there are varying degrees of under-reporting in various states in U.S. , and the degree of under-reporting is negatively correlated with the standardized number of tests per capita. Furthermore, we found that most states have low reporting rates. It should not be surprising, as most states lacked sufficient testing capabilities in the early stages of the pandemic. The low number of reports in turn made people ignore the seriousness of the disease and slowed down the effective measures taken by governments, enterprises and households, leading to the rapid spread of the epidemic. It is worth noting that the statistically significant risk factors identified by M4 should not be over-interpreted, as our model is not entirely based on epidemiological disease transmission patterns, nor has it considered all possible factors, plus the state-level data is aggregated. All in all, the main contribution of this application is to estimate the under-reporting severity of various states in U.S. based on the data of serological study, expecting that governments can pay more attention to similar infectious diseases during early stage of epidemic in future.

CHAPTER TWO

A Bayesian Bi-variate Spatial Model for Correcting Diagnostic Misclassification between Two Counts

2.1 Introduction

The analysis of count data subject to misclassification is an important problem in many epidemiological, medical and environmental applications. Failing to account for misclassification can bias estimates and underestimate standard errors, leading to underestimation of some risk factors and overestimation of others (Stamey, Young, & Seaman Jr, 2008). Two approaches are commonly used to correct for misclassification. If a gold standard measurement exists, a main-study large sample plus a small validation sample can improve estimation, see Lyles (2002); in contrast, if no gold standard is available, a Bayesian approach with subjective prior information on at least some subset of the parameters is an alternative, for example, Joseph, Gyorkos, and Coupal (1995). Sposto, Preston, Shimizu, and Mabuchi (1992) extend the likelihood approach used for under-reported counts by Whittemore and Gong (1991), to model misclassified counts via Poisson regression. A Bayesian approach implemented by Stamey et al. (2008) provides an alternative method for modeling misclassified counts with Poisson regression.

Potential spatial dependency in misclassified counts has received considerably less attention. Though the spatial structure, if it exists, might have been modelled by the known fixed effects, it is common for some spatial structure to remain in the residuals. There are two potential sources for the remaining spatial auto-correlation. One is from unmeasured or unknown covariates that are spatially correlated and the other is inherent neighborhood and/or clustering effects (Lee, 2013).

For a single count response variable, univariate CAR or ICAR is commonly used to account for spatial dependency. In the case of misclassified counts, two responses are investigated simultaneously (for instance, deaths due to cancer and deaths due to non-cancer). Therefore, a bivariate spatial model is necessary to account for the correlation across responses (Jin, Carlin, & Banerjee, 2005). It is intuitive to extend the univariate CAR (Besag, 1974) to a multivariate version. For instance, Gelfand and Vounatsou (2003) and Carlin et al. (2003) provide generalizations of the proper Multivariate CAR (MCAR) model that allow for different propriety parameters for each response.

In this chapter, we extend the model of Stamey et al. (2008) by incorporating a Multivariate CAR structure. This chapter is organized as follows. In section 2.2, the hierarchical Bayesian model with misclassification and spatial structure is discussed in detail. We investigate the way to incorporate correlated spatial effects into the bivariate misclassification in section 2.3. In section 2.4 we conducted a simulation study. In section 2.5 we apply the model to a real-world data set. We provide concluding comments in section 2.6.

2.2 Misclassification Model of Bivariate Count Data With Independent Spatial Random Effects

The model of Stamey et al. (2008) has two fallible samples, $\mathbf{y_1} = (y_{11}, y_{21}, \cdots, y_{N1})'$ and $\mathbf{y_2} = (y_{12}, y_{22}, \cdots, y_{N2})'$, assuming $y_{i1} \sim \text{Poisson}(E_i\mu_{i1})$ and $y_{i2} \sim \text{Poisson}(E_i\mu_{i2})$. The quantity N is the sample size for both groups and E_i are offsets. Parameters $\mu_{i1} = \lambda_{i1}(1 - p_1) + \lambda_{i2}p_2$ and $\mu_{i2} = \lambda_{i2}(1 - p_2) + \lambda_{i1}p_1$ are the Poisson rates of the observed (fallible) data in the ith unit. The parameters λ_{i1} and λ_{i2} are the true incidence rates. The quantity p_1 is defined as the probability that a count truly from group 1 is misclassified as belonging to group 2 while p_2 is the probability that a count of group 2 is incorrectly labeled to group 1. To facilitate the description of the model, we will label group 1 as, for example, death due to cancer, and group 2 as death due to non-cancer.

Let $\mathbf{X} \equiv [\mathbf{x}_i]$ be the matrix of covariates, where $\mathbf{x}_i$ is the vector of covariates for the ith unit. The true incidence rates are assumed to depend on the covariates through the log link function

$$\log(\lambda_{ig}) = \mathbf{x}'_i\boldsymbol{\beta}^g, \quad i = 1, \cdots, N, \quad g = 1, 2 \tag{2.1}$$

where $\boldsymbol{\beta}^g$ is the vector of regression parameters for count response g. We extend model (2.1) by adding spatial random effects to linear predictors of λ_{ig}. Assuming $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ are independent spatial random effects for each response, respectively, then (2.1) becomes

$$\log(\lambda_{ig}) = \mathbf{x}'_i\boldsymbol{\beta}^g + \phi_i^g, \tag{2.2}$$

The joint likelihood of the observable data subject to misclassification is proportional to

$$\begin{aligned}\prod_{i=1}^{N}&[\lambda_{i1}(1-p_1) + \lambda_{i2}p_2]^{y_{i1}}[\lambda_{i2}(1-p_2) + \lambda_{i1}p_1]^{y_{i2}} \\ &\times \exp\{-E_i[\lambda_{i1}(1-p_1) + \lambda_{i2}p_2] - E_i[\lambda_{i2}(1-p_2) + \lambda_{i1}p_1]\}\end{aligned} \tag{2.3}$$

As demonstrated by Daniel Paulino, Soares, and Neuhaus (2003) and McInturff, Johnson, Cowling, and Gardner (2004), equation (2.3) is over-parameterized because of the presence of misclassification parameters. In the Bayesian paradigm, informative priors, validation data, or both are needed for valid estimation. Following Stamey et al. (2008), independent Beta priors are used for p_1 and p_2. Specifically, we assume that $p_1 \sim \text{Beta}(a_1, b_1)$ and $p_2 \sim \text{Beta}(a_2, b_2)$. If validation data are also available, we assume they are random samples from two independent binomial distributions. For example, n_1 subjects known to have died from cancer are diagnosed with both a gold standard and a fallible method, and m_1 are mislabled by the fallible method. Similarly, in n_2 deaths due to non-cancer diagnosed by the gold standard, m_2 are misclassified to the cancer group. Thus, we have $m_1 \sim binomial(n_1, p_1)$ and $m_2 \sim binomial(n_2, p_2)$. Combining the validation data with Beta priors, Bayes theorem yields posterior distributions $p_1 \sim \text{Beta}(m_1 + a_1, n_1 - m_1 + b_1)$ and $p_2 \sim \text{Beta}(m_2 + a_2, n_2 - m_2 + b_2)$. These are then used as prior distributions for the main study data. For the regression coefficients, we assume a multivariate normal prior,

$$\boldsymbol{\beta}^g \sim \mathbf{N}_p(\mathbf{b}_g, \mathbf{B_g}), \tag{2.4}$$

where $\mathbf{b}_g$ is the vector of prior means for the regression coefficients $\boldsymbol{\beta}^g$ and $\mathbf{B_g}$ is the covariance matrix. A commonly used non-informative prior can be fashioned by setting $\mathbf{b}_g = (0, 0, \cdots, 0)'$ and

$$\mathbf{B}_g = \begin{bmatrix} \sigma^2_{1g} & 0 & \cdots & 0 \\ 0 & \sigma^2_{2g} & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2_{kg} \end{bmatrix} \tag{2.5}$$

with large positive values for $\sigma^2_{ig}, i = 1, 2, ..., k$ and $g = 1, 2$. However, we find in some examples that convergence of the Markov Chain Monte Carlo (MCMC) algorithm used to obtain the needed posterior distributions is slow if the diagonal value $\boldsymbol{B}_g$ are too large.

In order to estimate the parameters, a common approach for models with misclassification is to augment the observed data with the unobserved misclassified counts. Using a latent variable approach can ease the derivation of the the full conditionals. We define U_{i1} to be the counts mislabled as being from non-cancer and U_{i2} to be the counts incorrectly classified as being from the cancer group. Note that $U_{i1} \leq y_{i2}$ and $U_{i2} \leq y_{i1}$.

Augmenting equation (2.3) with the unobserved misclassified counts and multiplying by the prior densities, we obtain the joint posterior density

$$
\begin{aligned}
p(&\boldsymbol{\beta_1},\boldsymbol{\beta_2}, p_1, p_2, \boldsymbol{U_1}, \boldsymbol{U_2}|\boldsymbol{y_1}, \boldsymbol{y_2}, m_1, m_2)\\
&\propto \prod_{i=1}^{N} \lambda_{i1}^{y_{i1}-U_{i2}+U_{i1}} \exp\{-E_i\lambda_{i1}\} p_1^{U_{i1}} (1-p_1)^{y_{i1}-U_{i2}} p_1^{m_1+a_1-1} (1-p_1)^{n_1-m_1+b_1-1}\\
&\times \lambda_{i2}^{y_{i2}-U_{i1}+U_{i2}} \exp\{-E_i\lambda_{i2}\} p_2^{U_{i2}} (1-p_2)^{y_{i2}-U_{i1}} p_2^{m_2+a_2-1} (1-p_2)^{n_2-m_2+b_2-1}\\
&\times \exp[\frac{1}{2}(\boldsymbol{\beta_1}-\boldsymbol{b_1})\boldsymbol{B_1}^{-1}(\boldsymbol{\beta_1}-\boldsymbol{b_1})] \exp[\frac{1}{2}(\boldsymbol{\beta_2}-\boldsymbol{b_2})\boldsymbol{B_2}^{-1}(\boldsymbol{\beta_2}-\boldsymbol{b_2})]\\
&\times \tau_\phi^{\frac{1}{2}N} \exp\left[-\frac{1}{2}\tau_\phi \sum\sum_{i\neq j}(\phi_i-\phi_j)^2\omega_{ij}\right] \tau_\eta^{\frac{1}{2}N} \exp\left[-\frac{1}{2}\tau_\eta \sum\sum_{i\neq j}(\eta_i-\eta_j)^2\omega_{ij}\right],
\end{aligned}
\tag{2.6}
$$

where vectors $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ are independent spatial effects for λ_{i1} and λ_{i2}. We assume $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ follow the Intrinsic Conditional Auto-regressive (ICAR) model centered at zero by placing constraints: $\sum_{i=1}^{N}\phi_i = 0$ and $\sum_{i=1}^{N}\eta_i = 0$.

The full conditionals are

(1)

$$p_1|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, p_2, \boldsymbol{U_1}, \boldsymbol{U_2}, \boldsymbol{y_1}, \boldsymbol{y_2}, m_1 \propto p_1^{\sum U_{i1}+m_1+a_1-1}(1-p_1)^{\sum(y_{i1}-U_{i2})+n_1-m_1+b_1-1}$$

$$\sim Beta(\sum U_{i1}+m_1+a_1, \sum(y_{i1}-U_{i2})+n_1-m_1+b_1).$$

(2)

$$p_2|\boldsymbol{\beta_1}, \boldsymbol{\beta_2}, p_1, \boldsymbol{U_1}, \boldsymbol{U_2}, \boldsymbol{y_1}, \boldsymbol{y_2}, m_2 \propto p_2^{\sum U_{i2}+m_2+a_2-1}(1-p_2)^{\sum(y_{i2}-U_{i1})+n_2-m_2+b_2-1}$$

$$\sim Beta(\sum U_{i2}+m_2+a_2, \sum(y_{i2}-U_{i1})+n_2-m_2+b_2).$$

(3)

$$
\begin{aligned}
U_{i1}|\text{others} \propto & \lambda_{i1}^{U_{i1}} p_1^{U_{i1}} \lambda_{i2}^{y_{i2}-U_{i1}} (1-p_2)^{y_{i2}-U_{i1}} \\
= & (\lambda_{i1}p_1)^{U_{i1}} [\lambda_{i2}(1-p_2)]^{y_{i2}-U_{i1}} \\
= & [\frac{\lambda_{i1}p_1}{\lambda_{i1}p_1+\lambda_{i2}(1-p_2)}]^{U_{i1}} [\frac{\lambda_{i2}(1-p_2)}{\lambda_{i1}p_1+\lambda_{i2}(1-p_2)}]^{y_{i2}-U_{i1}} \\
& \times [\lambda_{i1}p_1+\lambda_{i2}(1-p_2)]^{U_{i1}+y_{i2}-U_{i1}} \\
\propto & [\frac{\lambda_{i1}p_1}{\lambda_{i1}p_1+\lambda_{i2}(1-p_2)}]^{U_{i1}} [\frac{\lambda_{i2}(1-p_2)}{\lambda_{i1}p_1+\lambda_{i2}(1-p_2)}]^{y_{i2}-U_{i1}} \\
& \pi_{i1}^{U_{i1}} (1-\pi_{i1})^{y_{i2}-U_{i1}}
\end{aligned}
$$

$\sim \mathbf{Bin}\,(y_{i2}, \pi_{i1})$, where $\pi_{i1} = \frac{\lambda_{i1}p_1}{\lambda_{i1}p_1+\lambda_{i2}(1-p_2)}$.

(4)

$$
\begin{aligned}
U_{i2}|\text{others} \propto & [\frac{\lambda_{i2}p_2}{\lambda_{i2}p_2+\lambda_{i1}(1-p_1)}]^{U_{i2}} [\frac{\lambda_{i1}(1-p_1)}{\lambda_{i2}p_2+\lambda_{i1}(1-p_1)}]^{y_{i1}-U_{i2}} \\
= & \pi_{i2}^{U_{i2}} (1-\pi_{i2})^{y_{i1}-U_{i2}}
\end{aligned}
$$

$\sim \mathbf{Bin}\,(y_{i1}, \pi_{i2})$, where $\pi_{i2} = \frac{\lambda_{i2}p_2}{\lambda_{i2}p_2+\lambda_{i1}(1-p_1)}$.

(5)

$$
p(\boldsymbol{\beta_1}|\text{others}) \propto \exp[\frac{1}{2}(\boldsymbol{\beta_1}-\boldsymbol{b_1})\boldsymbol{B_1}^{-1}(\boldsymbol{\beta_1}-\boldsymbol{b_1})] \prod_{i=1}^{N} \lambda_{i1}^{y_{i1}-U_{i2}+U_{i1}} \exp\{-E_i\lambda_{i1}\}
$$

(6)

$$
p(\boldsymbol{\beta_2}|\text{others}) \propto \exp[\frac{1}{2}(\boldsymbol{\beta_2}-\boldsymbol{b_2})\boldsymbol{B_2}^{-1}(\boldsymbol{\beta_2}-\boldsymbol{b_2})] \prod_{i=1}^{N} \lambda_{i2}^{y_{i2}-U_{i1}+U_{i2}} \exp\{-E_i\lambda_{i2}\}
$$

(7)

$$\begin{aligned}
\phi_i|\text{others} \propto \exp&\left[-\frac{1}{2}\tau_\phi \sum_{i\neq j}(\phi_i - \phi_j)^2\omega_{ij}\right] \\
= \exp&\left[-\frac{1}{2}\tau_\phi \sum_{i\neq j}(\phi_i^2 - 2\phi_i\phi_j + \phi_j^2)\omega_{ij}\right] \\
= \exp&\left[-\frac{1}{2}\tau_\phi \sum_{i\neq j}(\phi_i^2 - 2\phi_i\phi_j + \phi_j^2)\omega_{ij}\right] \\
\propto \exp&\left[-\frac{1}{2}\tau_\phi \left(k\phi_i^2 - 2\sum_{i\neq j}(\omega_{ij}\phi_j\phi_i)\right)\right] \\
= \exp&\left[-\frac{1}{2}k\tau_\phi(\phi_i^2 - 2\sum_{i\neq j}(\omega_{ij}\phi_j\phi_i)/k_i)\right] \\
\propto \exp&\left[-\frac{1}{2}k_i\tau_\phi(\phi_i - \sum_{i\neq j}(\omega_{ij}\phi_j)/k_i)^2\right]
\end{aligned}$$

$\sim N\left(k_i^{-1}\sum_{i\neq j}(\omega_{ij}\phi_j), k_i\tau_\phi\right)$, where $k_i = \sum_j \omega_{ij}$.

(8)

$$\eta_i|\text{others} \sim N\left(k_i^{-1}\sum_{i\neq j}(\omega_{ij}\eta_j), k_i\tau_\eta\right)$$

where $k_i = \sum_j \omega_{ij}$

We will use Stan (a probabilistic programming platform that does full Bayesian inference using Hamiltonian Monte Carlo (HMC)) to simulate the bi-variate misclassification model discussed here and fit the model for COVID-19 deaths versus deaths due to other causes in Section 2.6. Stan can efficiently fit our model through 2000 iterations with 1000 warmups. As a comparison, the misclassification model ignoring complex spatial structure required 500,000 iterations via Gibbs sampling (Stamey et al., 2008).

2.3 Bivariate Count Data With Correlated Spatial Random Effects

In Section 2.2, two individual uni-variate CAR effects are added to each response respectively. In other words, we assume $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ are independent. However, it is more appropriate to assume that $\boldsymbol{\phi}$ and $\boldsymbol{\eta}$ are correlated, since the two responses (here cancer and non-cancer) may be correlated.

Let $\boldsymbol{S}' = (S_1', S_2', \cdots, S_n')$ where $\boldsymbol{S}$ is $np \times 1$ with each S_i being a p dimensional vector. We consider a multivariate normal distribution for $\boldsymbol{S}$ of the form

$$p(\boldsymbol{S}) = (2\pi)^{-np}|\boldsymbol{B}|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{S}^T\boldsymbol{B}\boldsymbol{S}\right) \tag{2.7}$$

where

$$\boldsymbol{B} = (\boldsymbol{D}_W - \alpha\boldsymbol{W}) \otimes \boldsymbol{\Lambda} \tag{2.8}$$

$\boldsymbol{B}$ is an $np \times np$ symmetric positive definite matrix, α is known as the propriety or spatial smoothness parameter and is restricted to $(0, 1)$, and $\boldsymbol{\Lambda}$ is a $p \times p$ positive definite matrix. $\boldsymbol{B}$ can be seen as the Kronecker product of two partial precision matrices: one for spatial dependency $(\boldsymbol{D}_W - \alpha\boldsymbol{W})$, and the other for the covariance across responses, given by $\boldsymbol{\Lambda}$. This model is denoted as $MCAR(\alpha, \boldsymbol{\Lambda})$ (Gelfand & Vounatsou, 2003). The full Bayesian hierarchical model can be implemented by placing appropriate priors on α ($Beta$ distribution) and $\boldsymbol{\Lambda}$ (for instance, a $Wishart(\rho, \boldsymbol{\Lambda_0})$). Carlin et al. (2003) recommended a relatively vague $Wishart(\rho = 2, \boldsymbol{\Lambda_0} = Diag(25, 25))$ prior on $\boldsymbol{\Lambda}$ and a $Beta(18, 2)$ on α.

In this study, for simplicity, only two responses are considered. Suppose a common design matrix $\boldsymbol{X} \equiv [\boldsymbol{x}_i]$ is available for reponses $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$, and we define $\boldsymbol{S}_1^T = (s_{11}, s_{12}, \cdots, s_{1n})$ and $\boldsymbol{S}_2^T = (s_{21}, s_{22}, \cdots, s_{2n})$ as the spatial random effects for each response. Following Carlin et al. (2003), the vector of joint spatial effects after being re-arranged as side-by-side vectors of group effects, follow a $2 \times n$ dimensional multivariate normal distribution centered at $\mathbf{0}$ with a precision matrix $\boldsymbol{B} = \begin{pmatrix} \boldsymbol{R_1'}\boldsymbol{R_1}\Lambda_{11} & \boldsymbol{R_1'}\boldsymbol{R_2}\Lambda_{12} \\ \boldsymbol{R_2'}\boldsymbol{R_1}\Lambda_{21} & \boldsymbol{R_2'}\boldsymbol{R_2}\Lambda_{11} \end{pmatrix}$, i.e.

$$\begin{pmatrix} \boldsymbol{S_1} \\ \boldsymbol{S_2} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{R_1'}\boldsymbol{R_1}\Lambda_{11} & \boldsymbol{R_1'}\boldsymbol{R_2}\Lambda_{12} \\ \boldsymbol{R_2'}\boldsymbol{R_1}\Lambda_{21} & \boldsymbol{R_2'}\boldsymbol{R_2}\Lambda_{11} \end{pmatrix}^{-1}\right) \tag{2.9}$$

where $\boldsymbol{R}_k'\boldsymbol{R}_k = \boldsymbol{D} - \alpha_k\boldsymbol{W}$, $k = 1, 2$, i.e. $\boldsymbol{R}_k$ is the upper-triangular matrix of the Cholesky decomposition of $\boldsymbol{D} - \alpha_k\boldsymbol{W}$. Matrix $\boldsymbol{\Lambda} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix}$ controls the non-spatial correlation between the two variables for any unit i. The propriety of the spatial structure can be easily met as long as the Cholesky decomposition exists and $\boldsymbol{\Lambda}$ is posititive definite. A sufficient condition for the existense of the above Cholesky decomposition is $|\alpha_1| < 1$ and $|\alpha_2| < 1$.

By replacing independent CAR effects $\boldsymbol{\Phi}$ and $\boldsymbol{\eta}$ in equation (2.2) with $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, we obtain a MCAR version of the misclassification model. Usually such a generalized linear regression with spatial random effects (uni- or multi-variate) is called Spatial Generalized Linear Mixed Models (SGLMM).

2.3.1 Sparse Areal Mixed Model for Multivariate Outcomes

SGLMM face two main challenges: spatial confounding as discussed in Section 1.3 and intensive computational burden. An approach proposed by Musgrove, Young, Hughes, and Eberly (2019) addresses both of these challenges by extending the model developed by Hughes and Haran (2013). The Multivariate Sparse Areal Mixed Model (MSAMM) by Musgrove et al. (2019) employs the same orthogonal, multi-resolution spatial basis described by Hughes and Haran (2013). We first recall the univariate SAMM model

$$f(\boldsymbol{\mu}) = \boldsymbol{X\beta} + \boldsymbol{R}_{n\times r}\boldsymbol{\eta}_r \tag{2.10}$$

Model (2.10) can be extended to a multivariate version (MSAMM) as developed by Musgrove et al. (2019). We assume the multiple outcomes observed at each areal unit have a common design matrix, $\boldsymbol{X}$. Specifically, for $g \in 1, \cdots, G$ we have outcomes $\boldsymbol{y}^g_{n\times 1}$, regression coefficients $\boldsymbol{\beta}^g$ and spatial effects $\boldsymbol{\phi}^g$, then

$$f(\boldsymbol{\mu^g}) = \boldsymbol{X\beta}^g + \boldsymbol{\phi}^g. \tag{2.11}$$

We then convert $\boldsymbol{\phi} = [\phi^1, \cdots, \phi^G]$ into a vector $\boldsymbol{\phi}_{NG\times 1}$ and specify its density as

$$p(\boldsymbol{\phi}|\boldsymbol{\Sigma}) \propto \exp\{-\frac{1}{2}\boldsymbol{\phi}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{Q})\boldsymbol{\phi}\}, \tag{2.12}$$

where $\boldsymbol{\Sigma}$ is a $G \times G$ co-variance matrix with the gth diagonal entry proportional to the variance of the spatial effects corresponding to the gth outcome and the (g, g') off-diagonal entry proportional to the covariance between the gth and g'th spatial effects. Similar to univariate SAMM, the construction of MSAMM is based on principal component analysis of Moran's I operator (1.22). Let $\boldsymbol{R}$ be a matrix whose columns are the first r eigenvectors of Moran's I operator, then the precision matrix can be approximated by $\boldsymbol{Q}_r = \boldsymbol{R}'\boldsymbol{Q}\boldsymbol{R}$, and the MSAMM can be specified as

$$f(\boldsymbol{\mu}^g) = \boldsymbol{X}\boldsymbol{\beta}^g + \boldsymbol{R}\boldsymbol{\eta}_r^g, \tag{2.13}$$

where $\boldsymbol{\eta}_r^g$ is a $r \times 1$ vector, and if we set $\boldsymbol{\Delta} = (\eta_r^{1'}, \cdots, \eta_r^{G'})'$,

$$p(\boldsymbol{\Delta}|\boldsymbol{\Sigma}) \propto exp\{-\frac{1}{2}\boldsymbol{\Delta}'(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{Q}_r)\boldsymbol{\Delta}\}, \tag{2.14}$$

where $\boldsymbol{\Sigma}$ is the $G \times G$ covariance matrix, and $\boldsymbol{Q}_r$ is the reduced $r \times r$ precision matrix. The Krockneckor product in (2.14) is computationally expensive when fitting the model with either traditional MCMC or Hamiltonian Sampling (Stan). Musgrove et al. (2019) showed computation can be eased considerably as follows. Let $\boldsymbol{C}_r$ be the upper Cholesky triangle of $\boldsymbol{Q}_r$, and let $\boldsymbol{W}_r = \boldsymbol{C}_r^{-1}$ such that $\boldsymbol{W}_r\boldsymbol{W}_r' = \boldsymbol{Q}_{\boldsymbol{r}}^{\boldsymbol{-1}}$. Then, for $\boldsymbol{\Psi} = (\boldsymbol{\psi}_r^{1'}, \cdots, \boldsymbol{\psi}_r^{G'})'$, each $\boldsymbol{\psi}_r^g$ is also $r \times 1$, and if $\boldsymbol{\Psi}|\boldsymbol{\Sigma} \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_r)$, it can be shown that $(\boldsymbol{I}_g \otimes \boldsymbol{W}_s)\boldsymbol{\Psi}$ and $\boldsymbol{\Delta}$ have the same distribution, with

$$\begin{aligned} E[(\boldsymbol{I}_g \otimes \boldsymbol{W}_s)\boldsymbol{\Psi}] &= (\boldsymbol{I}_g \otimes \boldsymbol{W}_s)E[\boldsymbol{\Psi}] = \boldsymbol{0}, \\ Cov[(\boldsymbol{I}_g \otimes \boldsymbol{W}_s)\boldsymbol{\Psi}] &= (\boldsymbol{I}_g \otimes \boldsymbol{W}_s)(\boldsymbol{\Sigma} \otimes \boldsymbol{I}_r)(\boldsymbol{I}_g \otimes \boldsymbol{W}_s)' = \boldsymbol{\Sigma} \otimes \boldsymbol{Q}_r^{-1}, \end{aligned}$$

then (2.13) can be written as

$$f(\boldsymbol{\mu}^g) = \boldsymbol{X}\boldsymbol{\beta}^g + \boldsymbol{R}\boldsymbol{W}_r\boldsymbol{\psi}_r^g, \quad g = 1, \cdots, G \tag{2.15}$$

where $p(\boldsymbol{\Psi}|\boldsymbol{\Sigma}) = N(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_r)$, and $\boldsymbol{I}_r$ is the r-dimensional identity matrix.

2.3.2 Poisson Regression with Sparse Areal Mixed Model for Mis-classified Counts

The proposed full model for misclassified counts with independent SAMM or Joint MSAMM can be expressed as

$$\begin{aligned}
\boldsymbol{y}^1 &\sim Poisson(E\boldsymbol{\mu}^1)\\
\boldsymbol{y}^2 &\sim Poisson(E\boldsymbol{\mu}^2)\\
\boldsymbol{\mu}^1 &= \boldsymbol{\lambda}^1 * (1-p^1) + \boldsymbol{\lambda}^2 * p^2\\
\boldsymbol{\mu}^2 &= \boldsymbol{\lambda}^2 * (1-p^2) + \boldsymbol{\lambda}^1 * p^1\\
\log(\boldsymbol{\lambda}^1) &= \boldsymbol{X}\boldsymbol{\beta}^1 + \boldsymbol{R}\boldsymbol{\eta}_r^1\\
\log(\boldsymbol{\lambda}^2) &= \boldsymbol{X}\boldsymbol{\beta}^2 + \boldsymbol{R}\boldsymbol{\eta}_r^2.
\end{aligned}$$

For the regression coefficients, we assume independent normal priors,

$$\boldsymbol{\beta}^g \sim N(\boldsymbol{0}, 100\boldsymbol{I}), \quad g = 1, 2,$$

for the independent SAMM, the reduced spatial effects are assumed

$$\boldsymbol{\eta}_r^g \sim N(\boldsymbol{0}, \tau_g \boldsymbol{R}'\boldsymbol{Q}\boldsymbol{R}), \quad g = 1, 2,$$

and a gamma$(0.5, 0.0005)$ prior is assigned to the hyper-parameter τ_g.
For MSAMM, the regressions become

$$\begin{aligned}
\log(\boldsymbol{\lambda}^1) &= \boldsymbol{X}\boldsymbol{\beta}^1 + \boldsymbol{R}\boldsymbol{W}_r\boldsymbol{\psi}_r^1\\
\log(\boldsymbol{\lambda}^2) &= \boldsymbol{X}\boldsymbol{\beta}^2 + \boldsymbol{R}\boldsymbol{W}_r\boldsymbol{\psi}_r^2,
\end{aligned}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\psi_r^{1\prime}}, \boldsymbol{\psi_r^{2\prime}})' \sim N(\boldsymbol{0}, \boldsymbol{\Sigma} \otimes \boldsymbol{I}_r)$, and $\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}$ is the 2×2 covariance matrix between the two responses. Several appealing choices for the prior distribution of $\boldsymbol{\Sigma}$ exist. For example, Gelman and Hill (2006) suggest using a scaled inverse Wishart prior for $\boldsymbol{\Sigma}$, motivated primarily by its conjugacy to the multivariate likelihood function and thus simplifies Gibbs sampling. The prior for the covariance matrix can be decomposed into a scale and a correlation matrix, and it can be implemented in a more natural way. To be specific, in our bivariate case, we define

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{pmatrix} \tag{2.16}$$

where σ_1 and σ_2 are the variance scaling parameters for each response and $\mathbf{\Omega} = \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix}$ is a correlation matrix, such that the following relations hold,

$$\begin{aligned} \sigma_i^2 &= \mathbf{\Sigma}_{ii} \\ \mathbf{\Omega}_{i,j} &= \frac{\mathbf{\Sigma}_{i,j}}{\sigma_i \sigma_j}, \quad i = 1, 2, \quad j = 1, 2. \end{aligned}$$

For each standard deviation we assign a weakly informative half-Cauchy distribution with a small scale parameter,

$$\sigma_i \sim \text{Cauchy}(0,\ 2.5), \quad \sigma_i > 0.$$

For the correlation matrix $\mathbf{\Omega}$, we assume an LKJ prior with shape parameter $\eta \geq 1$

$$\mathbf{\Omega} \sim \text{LKJCorr}(\eta)$$

The LKJ correlation distribution is defined by

$$\text{LKJCorr}(\mathbf{\Sigma}|\eta) \propto \det(\Sigma)^{\eta-1}$$

Alternative priors for $\mathbf{\Sigma}$ include hierarchical half-t prior distribution(Huang, Wand, et al., 2013), and the covariance matrix separation strategy (Barnard, McCulloch, & Meng, 2000). We fit the model via Hamilton Monte Carlo Sampling with the Stan Language (Carpenter et al., 2017).

2.4 Simulation

We carried out simulation studies to investigate the performance of our proposed model. We will compare three different models: MIS-MSAMM (misclassification with bivariate sparse areal mixed model), MIS-SAMM (misclassification with two independent univariate sparse areal mixed model), and MIS-RE(mis-classification with just random normal error). The criteria for comparison is average bias, mean square error and coverage probability of the 95% posterior credible intervals. We start with simulating bi-variate count data based on 10 × 10 grids (sub-regions) as shown in Figure 2.1.

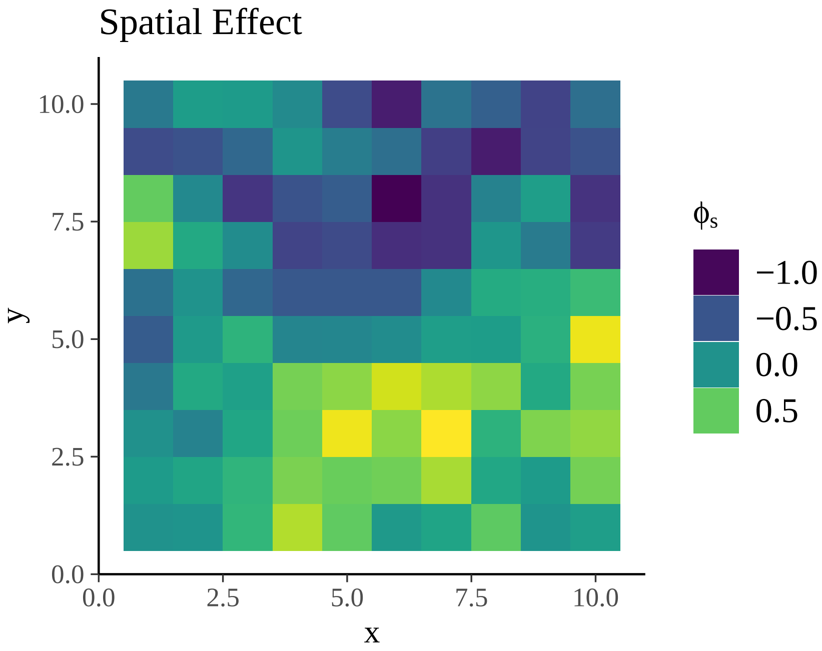


Figure 2.1: Sampling and fitting priors for skewness parameter

For each grid, we simulated two Poisson responses. For the regression coefficients, we select $\boldsymbol{\beta}^1 = (-1, 1.8, 0.15)'$, and $\boldsymbol{\beta}^2 = (-0.6, 1.5, 0)'$, respectively. Thus we have two covariates, both of which have non-zero coefficients for the first group, while only one covariate has a non-zero coefficient for the second group. Furthermore, we assign moderate misclassification rate $p_1 = 0.4$ of first-group mislabeled to second group and fairly low misclassification $p_2 = 0.05$ of second-group occurrences in the first group. The co-variance matrix is set as

$$\boldsymbol{\Sigma} = \begin{pmatrix} 4 & \sqrt{32}\rho \\ \sqrt{32}\rho & 8 \end{pmatrix}, \tag{2.17}$$

where the correlation coefficient $\boldsymbol{\rho} = (0.0, 0.2, 0.4, 0.6, 0.8)$. For the MIS-RE model, random normal errors without spatial dependency were introduced to account for extra variation in the generalized linear model such that we can make a fair comparison.

Table 2.1 and 2.2 compare MIS-MSAMM and MIS-SAMM models when applied to 100 simulated data sets consisting of misclassified and correlated bivariate responses (correlation ρ varies from 0 to 0.8). Model performances are evaluated by three measures: bias, the root of means square error, and coverage.

When the correlation between the two variables is minor or moderate (ρ=0 to 0.6), the performance of the two models is comparable. When the correlation is strong ($\rho = 0.8$), MIS-MSAMM performs better than MIS-SAMM (in the sense of smaller bias and RMSE), though the difference is small. Overall both models can recover regression coefficients, misclassification rates, and correlation coefficients pretty well. In the case of similar performance, we typically prefer parsimonious models. Table 2.3 compares MIS-MSAMM (full model) and three reduced models. It can be seen that the full model outperforms the other three in all aspects (with smaller deviation and RMSE; CI coverage comparable to nominal level). Ignoring misclassification and spatial structure, the Naive model results in the largest bias and RMSE, and the smallest coverage rate (all less than 10%). The spatial model corrects the inappropriately small posterior variance and thus improves the coverage, but fails to reduce bias and RMSE. The misclassification model significantly improves CI coverage (except for the two intercepts), but the overall performance is still inferior to MIS-MSAMM. In summary, the MIS-MSAMM model we proposed here is a flexible option when both the spatial structure and misclassification is significant. In section 2.5 we will apply the MIS-MSAMM model to a real data set.

Table 2.1: Results for simulation study comparing the MIS-MSAMM to MIS-SAMM fits. RMSE represents root mean square error. 100 datasets are simulated for small to moderate correlation values ($\rho = 0.0, 0.2, 0.4$). Variances for two responses are set to 4 and 8 respectively.

		MSAMM				SAMM			
Para	Truth	Avg.Est	Bias	Rmse	CP	Avg.Est	Bias	Rmse	CP
β_0	-1.000	-0.981	0.019	0.223	0.970	-0.981	0.019	0.218	0.960
β_1	1.800	1.798	-0.002	0.037	0.920	1.798	-0.002	0.036	0.940
β_2	0.150	0.147	-0.003	0.040	0.970	0.147	-0.003	0.040	0.980
γ_0	-0.600	-0.587	0.013	0.339	0.970	-0.588	0.012	0.338	0.960
γ_1	1.500	1.498	-0.002	0.058	0.950	1.498	-0.002	0.058	0.940
γ_2	0.000	-0.001	-0.001	0.061	0.960	-0.002	-0.002	0.061	0.990
σ_1^2	4.000	4.198	0.198	0.964	0.980	**3.983**	**-0.017**	**0.906**	0.950
σ_2^2	8.000	8.275	0.275	1.749	0.960	**7.920**	**-0.080**	**1.677**	0.950
$\boldsymbol{\rho}$	**0.0**	0.019	0.019	0.160	0.930	-	-	-	-
p_1	0.4	0.401	0.001	0.005	0.92	0.401	0.001	0.005	0.95
p_2	0.05	0.051	0.001	0.008	0.95	0.051	0.001	0.008	0.95
β_0	1.0	-0.957	0.043	0.292	0.950	**-1.024**	**-0.024**	0.265	0.940
β_1	1.8	1.794	-0.006	0.034	0.940	**1.801**	**0.001**	0.037	0.960
β_2	0.15	0.146	-0.004	0.052	0.960	0.153	0.003	0.054	0.940
γ_0	-0.6	**-0.604**	**-0.004**	0.380	0.950	0.583	0.017	0.352	0.940
γ_1	1.5	**1.507**	**0.007**	0.055	0.930	0.153	0.003	0.054	0.940
γ_2	0	-0.001	-0.001	0.074	0.950	-0.003	-0.003	0.073	0.950
σ_1^2	4	**4.098**	**0.098**	0.916	0.950	3.856	-0.144	0.821	0.920
σ_2^2	8	7.724	-0.276	1.959	0.930	**7.313**	**1.063**	11.244	0.940
$\boldsymbol{\rho}$	**0.2**	0.180	-0.020	0.164	0.920	-	-	-	-
p_1	0.4	0.400	<0.001	0.005	0.95	0.402	0.002	0.024	0.950
p_2	0.05	0.050	<0.001	0.007	0.95	0.049	-0.001	0.007	0.960
β_0	-1.000	-1.006	-0.006	0.260	0.960	**-1.000**	**0.000**	0.264	0.960
β_1	1.800	1.795	-0.005	0.036	0.910	1.795	-0.005	0.037	0.920
β_2	0.150	**0.149**	**-0.001**	0.065	0.950	0.147	-0.003	0.067	0.950
γ_0	-0.600	-0.594	0.006	0.392	0.990	**-0.603**	**-0.003**	0.418	0.980
γ_1	1.500	1.501	0.001	0.052	0.950	1.501	0.001	0.054	0.950
γ_2	0.000	-0.012	-0.012	0.088	0.970	-0.012	-0.012	0.092	0.970
σ_1^2	4.000	**3.960**	**-0.040**	0.918	0.930	3.789	-0.211	0.924	0.920
σ_2^2	8.000	**8.125**	**0.125**	1.893	0.960	7.847	-0.153	1.880	0.970
$\boldsymbol{\rho}$	**0.4**	0.349	-0.051	0.151	0.920	-	-	-	-
p_1	0.4	0.400	<0.001	0.005	0.96	0.399	-0.001	0.005	0.930
p_2	0.05	0.049	-0.001	0.007	0.94	0.049	-0.001	0.007	0.940

Table 2.2: Results for simulation study comparing the MIS-MSAMM to MIS-SAMM fits. RMSE represents root mean square error. 100 datasets are simulated for moderately strong to strong correlation values ($\rho = 0.6, 0.8$). Variances for two responses are set to 4 and 8 respectively.

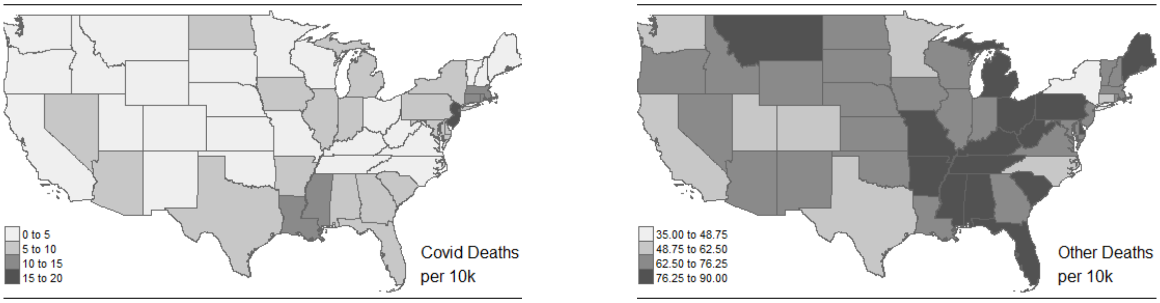
			MSAMM				SAMM		
Para	Truth	Avg.Est	Bias	Rmse	CP	Avg.Est	Bias	Rmse	CP
β_0	-1.000	-0.997	0.003	0.192	0.960	-1.004	-0.004	0.190	0.950
β_1	1.800	1.798	-0.002	0.030	0.950	1.798	-0.002	0.030	0.940
β_2	0.150	0.151	0.001	0.038	0.950	0.153	0.003	0.038	0.940
γ_0	-0.600	-0.630	-0.030	0.312	0.940	**-0.604**	**-0.004**	0.315	0.950
γ_1	1.500	1.499	-0.001	0.038	0.950	1.498	-0.002	0.037	0.950
γ_2	0.000	0.007	0.007	0.062	0.920	**0.003**	**0.003**	0.061	0.950
σ_1^2	4.000	4.105	0.105	0.914	0.960	**3.992**	**-0.008**	0.908	0.930
σ_2^2	8.000	8.067	0.067	1.728	0.960	7.880	-0.120	1.740	0.950
$\boldsymbol{\rho}$	**0.600**	0.577	-0.023	0.109	0.950	-	-	-	-
p_1	0.4	0.399	-0.001	0.003	0.970	0.398	-0.002	0.004	0.950
p_2	0.05	0.049	-0.001	0.006	0.980	0.049	-0.001	0.006	0.960
β_0	-1.000	-1.036	-0.036	0.257	0.970	**-1.031**	**-0.031**	0.265	0.990
β_1	1.800	1.799	-0.001	0.030	0.920	1.798	-0.002	0.031	0.970
β_2	0.150	0.159	0.009	0.061	0.990	0.159	0.009	0.062	0.980
γ_0	-0.600	**-0.659**	**-0.059**	0.358	0.950	-0.684	-0.084	0.390	0.960
γ_1	1.500	**1.508**	**0.008**	0.044	0.920	1.510	0.010	0.048	0.950
γ_2	0.000	**0.008**	**0.008**	0.075	0.970	0.013	0.013	0.080	0.950
σ_1^2	4.000	**3.931**	**-0.069**	0.879	0.960	3.894	-0.106	0.900	0.930
σ_2^2	8.000	**7.795**	**-0.205**	1.738	0.950	7.764	-0.236	1.786	0.960
$\boldsymbol{\rho}$	**0.800**	0.771	-0.029	0.072	0.920	-	-	-	-
p_1	0.4	0.400	<0.001	0.004	0.97	0.398	-0.002	0.004	0.940
p_2	0.05	0.049	-0.001	0.005	0.99	0.048	-0.002	0.006	0.950

Table 2.3: Results for simulation study comparing the MIS-BSAMM to Naive Poisson regression fits, Spatial-only and MisOnly fits. RMSE represents root mean square error. 100 datasets are simulated for moderately correlation ($\rho = 0.6$).

			Full				Naive		
Para	Truth	Avg.Est	Bias	Rmse	CP	Avg.Est	Bias	Rmse	CP
β_0	-1	-0.997	0.003	0.192	0.960	-0.636	0.364	1.448	0.020
β_1	1.8	1.798	-0.002	0.030	0.950	1.829	0.029	0.237	0.070
β_2	0.15	0.151	0.001	0.038	0.950	0.143	-0.007	0.131	0.060
γ_0	-0.6	-0.630	-0.030	0.312	0.940	0.316	0.916	1.435	0.030
γ_1	1.5	1.499	-0.001	0.038	0.950	1.707	0.207	0.286	0.020
γ_2	0	0.007	0.007	0.062	0.920	0.080	0.080	0.143	0.030
σ_1^2	4	4.105	0.105	0.914	0.960	-	-	-	-
σ_2^2	8	8.067	0.067	1.728	0.960	-	-	-	-
ρ	0.6	0.577	-0.023	0.109	0.950	-	-	-	-
p_1	0.4	0.399	0.001	0.003	0.970	-	-	-	-
p_2	0.05	0.049	-0.001	0.006	0.980	-	-	-	-
			Spatial				Mis		
β_0	-1	-1.126	-0.126	0.347	0.340	-1.104	-0.104	1.236	0.620
β_1	1.8	1.786	-0.014	0.056	0.460	1.786	-0.014	0.145	0.930
β_2	0.15	0.134	-0.016	0.039	0.450	0.151	0.001	0.096	0.970
γ_0	-0.6	-0.070	0.530	0.853	0.150	-0.831	-0.231	0.241	0.720
γ_1	1.5	1.674	0.174	0.198	0.030	1.528	0.028	0.284	0.940
γ_2	0	0.084	0.084	0.111	0.110	-0.024	-0.024	0.162	0.960
σ_1^2	4	3.402	-0.598	0.822	0.080	-	-	-	-
σ_2^2	8	5.173	-3.077	3.221	0.090	-	-	-	-
ρ	0.6	0.360	0.240	0.192	0.000	-	-	-	-
p_1	0.4	-	-	-	-	0.412	0.012	0.077	0.800
p_2	0.05	-	-	-	-	0.061	0.011	0.077	0.930

2.5 Application

As an example, we consider 2020 US state-level deaths associated with COVID-19 (group 1) and deaths due to other causes (group 2) by Oct 31. The sensitivity and specificity of COVID tests are both less than 1 and cause of death is often misclassified for many other types of deaths. According to the global Moran's I statistic (Moran's $I = 0.33$, $p = 0.0001$ for COVID-19 deaths per 10K population, and $I = 0.187$, $p = 0.02$ for other deaths), the state-level mortality of COVID-19 or other cause in the USA both had positive auto-correlations or clustered patterns, which also can be easily seen from Figure 2.2. The predicted number of deaths due to COVID-19 for some states is of particular medical interest. It can be computed from the posterior predictive distribution and easily implemented in Stan.



(a) Covid-19 deaths per 10K.

(b) Other deaths per 10K.

Figure 2.2: States-level mortality distribution across US.

Five models will be applied to the data for comparison,

M1: Naive Poisson Regression

M2: Samm without misclassification

M3: Samm with misclassification

M4: Msamm without misclassification

M5: Msamm with misclassification

2.5.1 Covariates and Priors

Two covariates will be used: "Obesity" and "Drug_death". "Smoking" is another covariate of interest, but we found "Smoking" is highly correlated with "Obesity" (0.77 correlation), therefore we ignore "Smoking" first. Although we believed the misclassification rate p_2 is less than p_1, in other words, the probability that deaths due to COVID-19 are misassigned to other causes is greater than the possibility of misassigning other deaths to COVID-19, we still place the same moderately informative prior beta$(10, 40)$ on p_1 and p_2. The beta(10, 40) allows for a wide range of misclassification probabilities since it has a 95 percent prior interval of (0.10, 0.32). For the regression parameters, we assign normal$(0, 10)$ as priors.

2.5.2 Results

Figure 2.3 presents the point estimates and 90% CIs of coefficients fitted by five models. Compared with the other four spatial models (M2-M5), the CIs of the regression coefficient given by naive Poisson (M1) is inappropriately short. The regression coefficient estimates given by models M2-M4 are slightly different, but the CI lengths are almost the same. M2 and M4 are closer as both ignore misclassification, while M3 and M5 are more similar as both incorporate the misclassification. The estimates of the misclassification probability p_1 obtained by M3 and M5 are almost the same, but as for the estimation of p_2, the CI length given by M5 is more than 4 times wider than that given by M3, although both support p_1 to be greater than p_2. Here we prefer using M3 because the results of M5 indicate that the correlation between the two counts is insignificant.

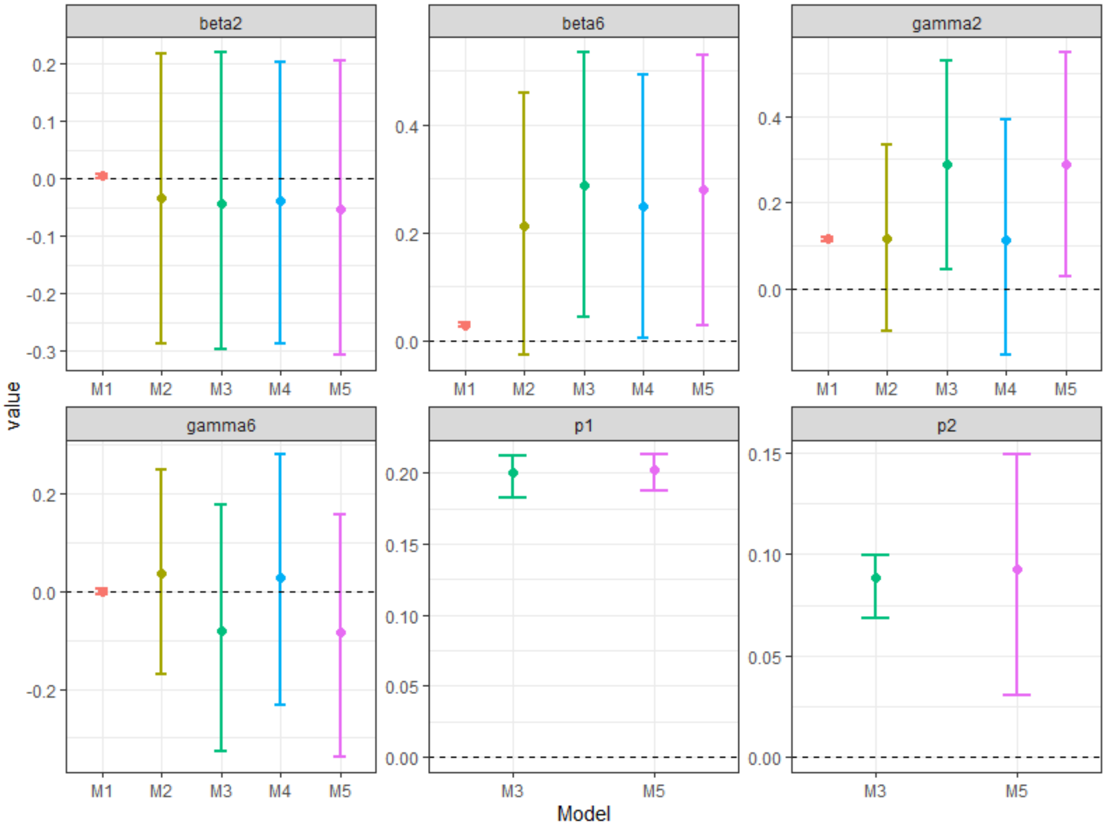


Figure 2.3: Point estimates and 90% CIs of coefficients fitted by five models."beta2" and "gamma2" correspond to "Obesity", "beta6" and "gamma6" correspond to "Drug_death".

Table 2.4 presents partial fitting results of the five models. Taking into account the misclassification, "obesity" is statistically relevant to COVID-19 death, while "drug_death" is significantly related to the incidence of other deaths (classified as common pneumonia or influenza). The model estimates that approximately 20.1% (90% CI: 0.183-0.212) of COVID-19 deaths were incorrectly classified as other causes, while approximately 8.9% (90% CI: 0.069-0.099) of other deaths were incorrectly attributed to COVID-19. We believe that a reasonable explanation could be that due to the limited early detection capabilities, some patients who died of COVID-19 were not confirmed by nucleic acid tests, while other deaths were not misclassified as much is also due to the insufficient nucleic acid testing of COVID-9. Of course, the nucleic acid test is not 100% accurate. In short, p_2 less than p_1 is in line with our perceptions and expectations.

Table 2.4: Partial fitting results of M1-5.

paras	mean	5%	95%	Model
β_2	0.005	0.001	0.009	M1
β_2	-0.033	-0.286	0.217	M2
β_2	-0.045	-0.297	0.219	M3
β_2	-0.038	-0.287	0.202	M4
β_2	-0.054	-0.307	0.205	M5
β_6	0.029	0.025	0.033	M1
β_6	0.214	-0.027	0.459	M2
β_6	0.288	0.044	0.534	M3
β_6	0.248	0.006	0.493	M4
β_6	0.281	0.029	0.530	M5
γ_2	0.116	0.111	0.120	M1
γ_2	0.115	-0.099	0.332	M2
γ_2	0.290	0.046	0.529	M3
γ_2	0.113	-0.153	0.390	M4
γ_2	0.289	0.028	0.547	M5
γ_6	-0.001	-0.006	0.005	M1
γ_6	0.036	-0.170	0.248	M2
γ_6	-0.081	-0.326	0.177	M3
γ_6	0.029	-0.233	0.280	M4
γ_6	-0.082	-0.340	0.158	M5
p_1	0.201	0.183	0.212	M3
p_1	0.202	0.187	0.213	M5
p_2	0.089	0.069	0.099	M3
p_2	0.092	0.030	0.149	M5
ρ	-0.008	-0.859	0.874	M4
ρ	-0.144	-0.938	0.866	M5

2.6 Conclusion

In this chapter, we investigate a hierarchical Bayesian model for correcting misclassification between two count responses. In addition to misclassification, the areal spatial correlation and the possible correlation between the two responses are also studied. The proposed model is an extension of Stamey et al. We avoid the complex multivariate CAR model, instead, we adopt SAMM or MSAMM, leading to significant improvement of fitting efficiency and computation speed. The other main advantages of the proposed model include flexibility and scalability; no requirement of a gold standard; insensitivity to the increase of N.

Simulation shows that the performance of MSAMM is equivalent to that of SAMM, especially when the correlation between the two responses is minor or moderate. In other words, in most cases, using SAMM is good enough. Thus we suggest one use MSAMM to fit the data first in practical applications. Once the posterior correlation is statistically insignificant, one can turn to SAMM to get more accurate results.

We apply the misclassification model with spatial structures to the cumulative number of deaths due to COVID-19 and other causes (pneumonia or flu) reported in each state of the U.S. as of the end of October 2020. Our result shows that about 20% of COVID-19 deaths were misclassified to other causes, and about 8.9% of other deaths were misattributed to COVID-19. We believe it is mainly due to insufficient testing capabilities for COVID-19, especially in the early stages of the pandemic. Some symptoms of COVID-19 are very similar to pneumonia or influenza, which also increases the probability of misclassification between the two groups. It is worth noting that the estimation of the regression coefficient of the covariate should not be over-interpreted as there might exist other confounding variables that have not been considered.

CHAPTER THREE

Bayesian Sample Size Determination for Skew Normal Data

3.1 Introduction

The selection of appropriate sample size is one of the most important parts of an experimental design, particularly for clinical trials where resources are limited. Considerable attention has been paid to sample size determination for comparing the means of two samples from normal distributions. Examples can be found in many texts (Kempthorne (1952) and Rosner (2015)). However, less effort has been given to the problem of computing sample size for the case of comparing two sample means of skew-normal populations.

Incorrectly assuming the data follows a normal distribution can have a considerable impact on the probabilities of Type I and Type II errors, and thus lead to a biased inference. Commonly used statistical techniques, such as the t-test, ANOVA, and linear regression, assume normality of errors. Thanks to the central limit theorem, moderate violations of the normality assumption may only have a minimal impact on the bias and efficiency of coefficient estimates, in particular, when the violation is moderate and the sample size is sufficiently large.

In cases where the impact of the normality violation is not negligible, robust techniques, such as non-parametric methods, are often applied. However, robust methods suffer from lower statistical power. Another commonly used alternative is to perform nonlinear transformations before the analysis. Examples of such types of transformations include arc-sine, logarithm, square root, and Box-Cox (of which, logarithmic and root transformation are special cases). However, many studies have shown that transformations can be problematic (Martin & Williams, 2017): non-linear transformations can meaningfully change the inference results; on the transformed scale, the inference might be reversed; on the original scale, the estimates are biased; in addition, interpretation of the regression coefficients can be meaningfully altered and not intuitive.

The assumption of normality mainly serves two purposes: one is for theoretical derivation and simplification, and the other is to facilitate computation. However, the rapid increase in computing power has enabled us to fit more complicated models, thus the convenience brought by the normality assumption has become less important. Quite often, estimators for parameters in complicated models may have no closed-form and need to rely on computer approximation such as Markov chain Monte Carlo (MCMC).

The skew-normal (SN) distribution proposed by Azzalini (1985) has been thought of as a good alternative to non-parametric methods or transformations when data are asymmetric. The SN allows for either positive or negative moderate skewness through a third parameter, the shape parameter, with the normal family as a special case of the SN when the shape parameter equals 0. Some examples of using the skew-normal include an application to a strength-stress model in reliability analysis (Gupta & Brown, 2001), modeling HIV-RNA in blood and in seminal plasma (Ghosh, Branco, & Chakraborty, 2007), an application to IQ scores and heights of Australian athletes (Hasanalipour & Sharafi, 2012), and an application to a cork stopper's process production (Figueiredo & Gomes, 2013).

In this chapter, we investigate the Bayesian sample size determination for skewed normal data. The Bayesian approach we apply is modeled after Wang and Gelfand (2002) and Brutti et al (2008). The simulation-based algorithm seeks to find the optimal sample size required to obtain a pre-specified power for a hypothesis test for the mean difference in a Bayesian context. The chapter is organized as follows. In Section 2, we overview important aspects of the skew-normal distribution. In Section 3 we discuss sample size determination from the Bayesian perspective. In Section 4 we give an example of the simulation-based scheme. We provide concluding comments in Section 5.

3.2 Skew Normal Distribution

We first overview properties of the skew normal (SN) distribution. This family of distributions has a shape parameter that defines the direction of the asymmetry of the distribution, also called the skewness parameter. The skew normal distribution is a three-parameter continuous probability distribution family with probability density function (PDF)

$$f(y;\xi,\omega,\alpha) = \frac{2}{\omega}\phi\left(\frac{y-\xi}{\omega}\right)\Phi\left[\alpha\left(\frac{y-\xi}{\omega}\right)\right], \xi \in \mathbb{R}, \omega > 0.$$

where ξ, ω, and α denote the location, scale, and shape parameters, respectively. $\phi(y)$ is the pdf of the standard normal and $\Phi(y)$ is the cumulative density function (CDF) of the standard normal distribution. We denote a random variable that has a skew normal distribution as $Y \sim SN(\xi, \omega, \alpha)$. A positive value of α indicates right skewness, and a negative α corresponds to left skewness. Moreover, it can be verified that the normal distribution is recovered when $\alpha = 0$.

The mean and variance of the SN are

$$E(Y) = \xi + \omega\gamma\sqrt{\frac{2}{\pi}}$$
$$Var(Y) = \omega^2\left(1 - \frac{2\gamma^2}{\pi}\right),$$

where $\gamma = \frac{\alpha}{\sqrt{1+\alpha^2}}$.

The skew normal distribution allows for either positive or negative skewness by introducing a skew (sometimes referred to as the shape) parameter α, which makes it flexible and useful in modelling real data sets that potentially violate the normality assumption. Lack of normality is often addressed through transformations, for example, the logarithm or square root transformations. The SN is a reasonable alternative for the normal distribution in the sense that it generalizes the normal distribution.

However, a practical problem with usage of the SN in applications is the possibility that the maximum likelihood estimator of the parameter which regulates skewness diverges (Azzalini & Arellano-Valle, 2012). The difficulty is due to the following (Bayes & Branco, 2007):

- The maximum likelihood estimator for α can be infinite,
- The Fisher information matrix is singular when $\alpha = 0$,
- The profile-likelihood function for α has a stationary point at $\alpha = 0$, independent from the observed sample.

The second problem above can be addressed with a reparameterization first proposed by Azzalini (1985), while the first and third problems are more serious and arise intrinsically from the likelihood shape. The main approach suggested to fix this problem is using a weight function to calibrate the likelihood, for example, see (Sartori, 2003). A more natural way to calibrate the likelihood is with the Bayesian approach which incorporates a prior distribution in place of the weight functions. Liseo and Loperfido (2006) demonstrated that the unbounded Jeffreys prior for the skewness parameter is proper but not a convenient expression (Azzalini, 1985), and thus is difficult to use directly. To make use of the Jeffrey's prior in modern MCMC software such as JAGS or Stan, an explicit expression for the prior distribution is needed. Bayes and Branco (2007) have shown that the Jeffreys' prior can be well approximated by a Student's t distribution $t(0, \pi^2/4; 1/2)$. This guarantees the finiteness of the shape (skew) parameter estimator resulting from the mode of the posterior distribution (Azzalini & Genton, 2008). Bayes and Branco (2007) also investigated using a uniform prior bounded between $[-1, 1]$ for $\gamma = \frac{\alpha}{\sqrt{1+\alpha^2}}$ which leads to a student's $t(0, 1/2; 2)$ for α through variable transformation. Figure 3.1 presents the density plot of the two priors for shape parameter mentioned above, it can be seen that the approximate Jeffreys' piror is less informative than the prior induced from a uniform prior for γ. As for prior specification to the location (ξ) and scale (ω) parameters, we discuss in more detail in Section 3.4.

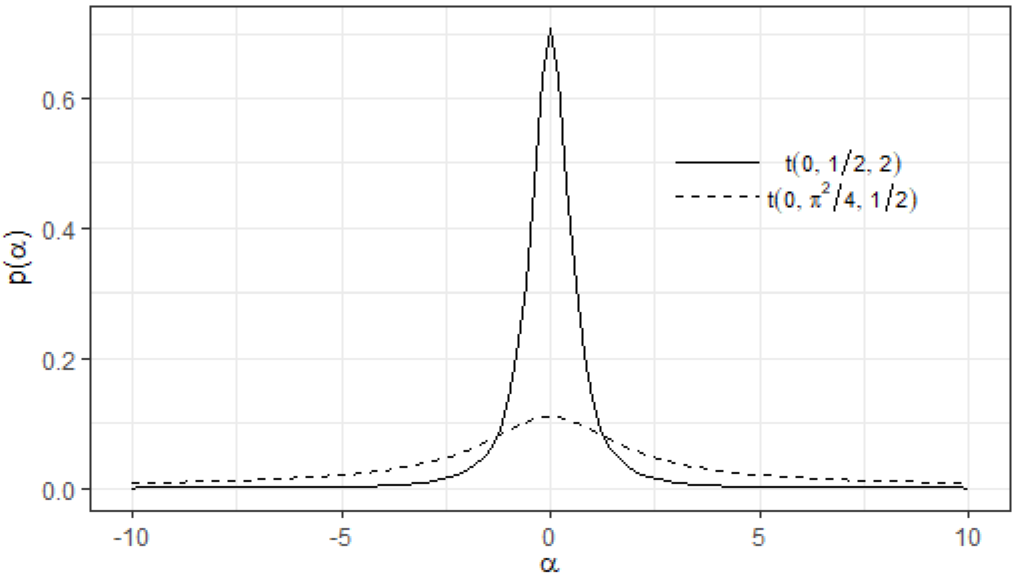


Figure 3.1: Prior density function for α.

3.3 Sample Size Determination

In clinical trials, sample size determination (SSD) plays an important role in the trade-off between budget and statistical power. According to a study published in JAMA Internal Medicine, clinical trials cost a median of \$41,117 per patient and \$3,562 per patient visit (Moore, Zhang, Anderson, & Alexander, 2018), thus sample size optimization is essential to controlling costs. On the other hand, an insufficient sample size might lead to the failure of discovering a true effect, that is, the study might be underpowered. Prior to computational advancements allowing more use of Bayesian methods, frequentist approaches were most commonly used for sample size determination. For many standard models, frequentist SSD is straightforward and easy to use as quite often closed-form formulae are available under certain assumptions and/or asymptotic approximations. As a simple example of frequentist SSD we consider the case of testing the difference between two sample means $\delta = \mu_1 - \mu_2$. We would like to test $H_0 : \delta = \delta_1$ against $H_a : \delta > \delta_1$. In a clinical trial, δ_1 would typically represent the clinical relevance that the researcher would consider important. For a balanced design under the normality assumption, the sample size needed for each group is determined by the following formula,

$$n = 2(Z_\alpha + Z_\beta)^2\sigma^2/\delta_1^2, \tag{3.1}$$

where α and β are the significance level and 1 - power specified by the researchers; Z_α is the critical value of the standard normal distribution at α, and Z_β is the critical value at β; σ^2 is the population variance (the two populations are assumed to have equal variance). There are some obvious limitations based on approaches like 3.1. First of all, one needs a point estimate of σ prior to the design. Though it is possible to estimate nuisance parameters from similar studies or by doing a small pilot study, the uncertainty of σ^2 is often not accounted for. Second, the normality and equal variance assumptions are often violated in real world data. Third, it is difficult to incorporate prior information and/or expert opinion into the design. In contrast, Bayesian SSD approaches provide a way to remedy these limitations, for example, the uncertainty of σ^2 can be incorporated into the process through an appropriate prior distribution.

Bayesian SSD can be thought of as a type of pre-posterior analysis (Wang, Gelfand, et al., 2002). For most problems of moderate complexity, Bayesian SSD is based on computationally demanding simulations, where synthetic data are drawn from a hypothetical distribution based on parameters generated from a design prior distribution. Consequently, the design prior is required to be proper, informative, and capable of capturing the experimenter's knowledge and uncertainty about the parameter of interest. We assume vector $\boldsymbol{\theta}$ contains the parameters of the proposed model, and the quantity of interest is the difference between the two sample means, δ. We assume that δ is truly greater than δ_1. Under this assumption, we expect that by drawing a synthetic data set of the minimal size of n_{min}, we can demonstrate the expected value of some posterior probability measure meets a specified criterion.

When the synthetic data set is generated based on the design prior distribution in each run of the simulation, a second, often relatively diffuse prior distribution will be applied to obtain the posterior distribution (or MCMC samples) for $\boldsymbol{\theta}$ and δ. In other words, usually two sets of prior distributions are needed : a sampling (design) prior $\pi_D(\boldsymbol{\theta})$ for data generation and a fitting (analysis) prior $\pi_A(\boldsymbol{\theta})$ for model fitting. $\pi_D(\boldsymbol{\theta})$ reflects the designers' beliefs or expectations about the experiment, and therefore is used for sampling synthetic data; $\pi_A(\boldsymbol{\theta})$ is often required to be relatively non-informative. The analysis prior $\pi_A(\boldsymbol{\theta})$ is also the one that we expect to use for model fitting once the real trial data is obtained.

Let $\boldsymbol{\Theta}$ denote the parameter space for $\boldsymbol{\theta}$, assume that the design under consideration will generate independent and identically distributed (iid) observations $\boldsymbol{X} = \{x_1, \cdots, x_n\}$ from density $f(x|\delta, \boldsymbol{\theta})$, where n is the sample size and let $\boldsymbol{S}$ be the sample space of $\boldsymbol{X}$. The prior predictive distribution of $\boldsymbol{X}$ is the marginal distribution of the data averaged over the design prior, $\pi_D(\boldsymbol{\theta})$, as follows,

$$m_{\pi_D}(\boldsymbol{X}) = \int_{\boldsymbol{\Theta}} f(\boldsymbol{X}|\boldsymbol{\theta}) \cdot \pi_D(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{3.2}$$

and the posterior distribution of θ given $\boldsymbol{X}$ and the analysis prior, $\pi_A(\theta)$, is

$$f(\theta|\boldsymbol{X}) \propto f(\boldsymbol{X}|\theta)\pi_A(\theta), \tag{3.3}$$

where $f(\boldsymbol{X}|\boldsymbol{\theta})$ is the likelihood of $\boldsymbol{X}$. Following Brutti, De Santis, and Gubbiotti (2008), without loss of generality, the design will be considered successful if the posterior probability that $\delta > \delta_1$ is larger than a predefined threshold γ

$$P_{\pi_A}(\delta > \delta_1|\boldsymbol{X}) > \gamma,\ \gamma \in (0, 1), \tag{3.4}$$

where $P_{\pi_A}(\cdot|\boldsymbol{X})$ is some posterior probability measure. Different SSD criteria can be defined based on formula 3.4. Here, we focus on the Predictive Probability Criterion (PPC). The PPC is based on the predictive probability of obtaining a successful experiment

$$p_n = P_{m_{\pi_D}}[S_n] = \int_{S_n} P_{\pi_A}(\delta > \delta_1) \cdot m_{\pi_D}(\boldsymbol{X}) d\boldsymbol{X}, \tag{3.5}$$

where S_n denotes those values of the sample space of $\boldsymbol{X}$ that lead to rejection of the hypothesis when the sample size is n. We are seeking the smallest sample size n such that p_n exceeds a threshold, say $1-\beta_p$. In other words, we wish to design a study with the smallest size n_{min} such that the posterior probability that $Pr(\delta > \delta_1|\boldsymbol{X}) > 1-\alpha$ on average is at least $1-\beta_p$, i.e,

$$E\left[I\{Pr(\delta > \delta_1|\boldsymbol{X}) > 1-\alpha\}\right] = 1-\beta_p, \tag{3.6}$$

where $I\{\cdot\} = 1$ if the posterior probability exceeds $1-\alpha$, and 0 otherwise. α is comparable to the significance level of frequentist approaches, and generally selected to be 0.1, 0.05, or 0.01. $1-\beta_p$ is equivalent to the statistical power, and typically set to 0.8, 0.85, or 0.9 (Beavers & Stamey, 2012). For complex models, the average power in equation (3.6) is computed through computationally intense simulation often with MCMC sampling to approximate the posterior distribution.

One important advantage that the Bayesian approach has over frequentist methods for sample size determination is flexibility. Recall that equation (3.1) holds under the normality assumption which is often violated in practice. In Bayesian SSD, the normality assumption is easily relaxed. Here we will adopt the skewed normal distribution to demonstrate the advantage of Bayesian SSD over frequentist approaches when applied to a scenario of comparing the means for two populations where the two underlying populations are potentiallhy skewed.

Suppose we are seeking to find the minimal sample size required for a clinical trial when the desired difference between the means of the treatment group and the control group is δ. If evidence from similar studies or pilot studies show that the data distribution is obviously asymmetric, or there are a moderate number of outliers, the frequentist approaches mentioned earlier can be biased. Here, we relax the assumptions of normality and homoscedasticity. Specifically, we assume that the two populations follow skewed normal distributions and allow for different variances.

Suppose $\boldsymbol{y_1}$ and $\boldsymbol{y_2}$ are sampled from $SN(\xi_1, \omega_1, \alpha_1)$ and $SN(\xi_2, \omega_2, \alpha_2)$, respectively, with equal sample size $n_1 = n_2 = n$. For simplicity, we assume $\alpha_1 = \alpha_2$, and $\xi_2 = \xi_1 + \delta$, where $\delta > 0$. The assumption on the α's can be relaxed. The quantity of interest is

$$\begin{aligned}\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1 &= \xi_1 + \delta + \omega_2 \gamma \sqrt{\frac{2}{\pi}} - (\xi_1 + \omega_1 \gamma \sqrt{\frac{2}{\pi}}) \\ &= \delta + (\omega_2 - \omega_1) \gamma \sqrt{\frac{2}{\pi}}\end{aligned}$$

We see from above that under this scenario it is difficult to calculate the minimal sample size with a closed form. The Bayesian SSD approach we propose here is simulation based.

The Bayesian SSD approach based on the PPC will compute the average power for each sample size for a set of design priors specified by the researchers. The mean power is approximated for each sample size by averaging over B repetitions. In the following sequence, the subscript $j \in \{1, \cdots, B\}$ represents the j^{th} repetition for each sample size n_i. The following steps are repeated,

```
Simulate $\xi_{1j}$, $\omega_j$, and $\alpha_j$ from their respective design priors;
Generate synthetic data $\boldsymbol{y}_{j1}$ and $\boldsymbol{y}_{j2}$ with size $n_1 = n_2 = n$, from
  $SN(\xi_{1j}, \omega_j, \alpha_j)$ and $SN(\xi_{1j} + \delta_1, \omega_j, \alpha_j)$, respectively;
Obtain the posterior probability that $\delta_j = \xi_{1j} - \xi_{2j} > \delta_1$;
Repeat steps 1-3 $B$ times for each sample size $n$, storing $Pr(\delta_j > \delta_1 | \boldsymbol{y}_{j1}, \boldsymbol{y}_{j2})$;
Compute the mean power by averaging the binary outcome of whether the
  posterior probability $Pr(\delta_j > \delta_1 | \boldsymbol{y}_{j1}, \boldsymbol{y}_{j2}) > 1 - \alpha$ as below
  $p^{(n)} = \frac{1}{B} \sum_{j=1}^{B} \boldsymbol{I}\{Pr(\delta_j > \delta_1 | \boldsymbol{y}_{j1}, \boldsymbol{y}_{j2}) > 1 - \alpha\}$;
Repeat steps 1-5 for the series of sample size and plot $p^{(n)}$ by $n$;
```

Algorithm 1: Predictive probability criterion.

3.4 A Simulation Study

When a condition such as atherosclerotic disease of the arteries occurs in the lower extremeties, walking impairment often results. (Nicolaï, Teijink, Prins, et al.,

2010) detail a multicenter randomized trial to determine if exercise therapy versus exercise therapy with feedback differed in terms of walking distance. Walking distance for patients with this sort of condition tends to result in right skewed data. So, we use this scenario as a motivation for our sample size determination procedure.

In Nicolaï et al. (2010), the number of meters walked was typically between 200 meters and 600 meters. We use this data to determine the parameters of our design priors, $\pi_D(\boldsymbol{\theta})$, for the model parameters $\boldsymbol{\theta}_j = (\xi_{1j}, \omega_j \alpha_j)$. Specifically, we assign each component of $\pi_D(\boldsymbol{\theta})$ as

$$\begin{aligned}
\xi_D &\sim N(0,1),\\
\omega_D &\sim Gamma(2, rate = 1),\\
\alpha_D &\sim N(3,1).
\end{aligned}$$

Next $\boldsymbol{y}_{j1}$ and $\boldsymbol{y}_{j2}$ are sampled from two SN distributions given $\boldsymbol{\theta}_j$ and δ. Figure 3.2 shows the histograms of a single set of $\boldsymbol{y}_{j1}$ and $\boldsymbol{y}_{j2}$ for $\delta = 0.25$. Under the assumption of normality, equation 3.1 yields the sample size needed for 95% confidence level and 80% power to be approximately 164 for each group.

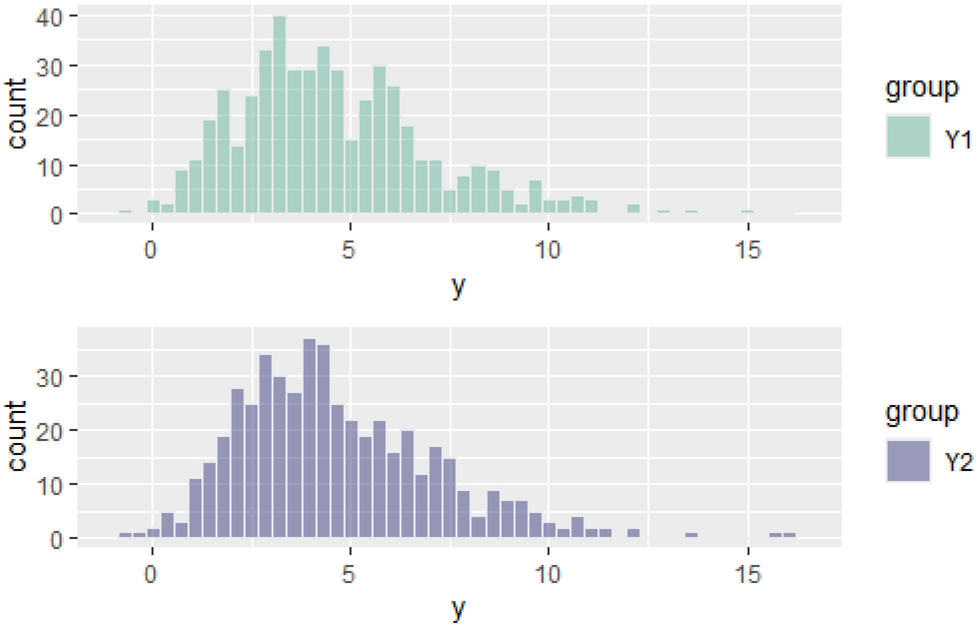


Figure 3.2: Histograms of a single set of $\boldsymbol{y}_{j1}$ and $\boldsymbol{y}_{j2}$ for $\delta = 0.25$

To investigate the approach in more detail, five values of δ are considered $(0.15, 0.20, 0.25, 0.30, 0.35)$. For each data set, the posterior probability of $\delta > 0$ is recorded. The repetition j is labeled as a success, or 1, if that probability is at least $1 - \alpha$, and a failure, or 0, otherwise. Relatively diffuse priors are used for the data analysis. Specifically, the analysis priors are as follows,

$$\begin{aligned}
\xi_A &\sim N(0, 10), \\
\omega_A &\sim N(0, 10), \\
\alpha_A &\sim t(0, \pi^2/4; 1/2), \\
\delta_A &\sim N(0, 10).
\end{aligned}$$

Partial comparisons of $\pi_D(\boldsymbol{\theta})$ and $\pi_A(\boldsymbol{\theta})$ are shown in Figure 3.3 and Figure 3.4.

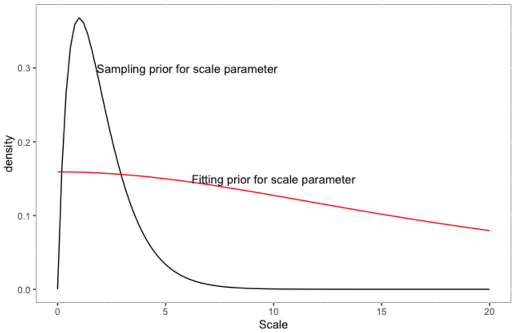


Figure 3.3: Sampling (design) and fitting (analysis) priors for ω_1 and ω_2.

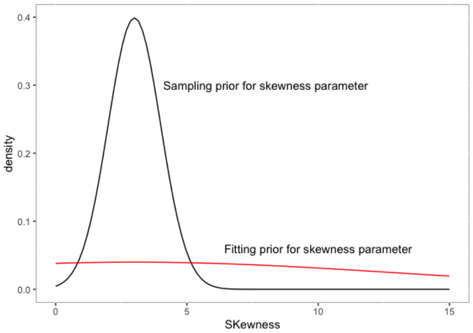


Figure 3.4: Sampling (design) and fitting (analysis) priors for α.

Typically, the grid search over increasing sample sizes would be performed until the sample size n_{min} is found so that for all sample sizes larger than or equal to n_{min} the PPC criterion in equation (3.6) is fulfilled. Normally 200 or more repetitions are needed for each n such that the expectation on the left side of equation (3.6) is sufficiently precise. Here, we show a set of Bayesian statistical powers against sample size and find the appropriate sample through interpolation. Specifically, once we obtain average powers for each sample size from, for example, 50 to 1500 (step by 50), an imputation method will be adapted to fit a smoothing power curve. The power curve is expected to monotonically increase and asymptotically approximate to a horizontal line at 1.0, thus we would fit the power curve with an asymptotic regression model which describes a limited growth for a quantity (powers here) against its predictor, which is the sample size n. The particular parameterization for the asymptotic regression model that we use is

$$Y = Asym + (R_0 - Asym) * exp\left[-exp(lrc) * x\right]$$

where $Asym$ is the asymptotic limit of Y as x goes to infinity, R_0 is the mean estimate of Y at $x = 0$, and lrc represents the natural logarithm of the rate constant.

The R function "*SSasymp*( )" in the "*stats*" package will be used to fit the above formula and then the "*geom_smooth*( )" function from the R package "*ggplot*" is used to plot the resulting output, given the argument $method = nls$, where "nls" represents non-linear smoothing. Simulation results (scatters) and corresponding smoothed curves are presented in Figure 3.5. As an example, for a δ of 0.15, then at least $n = 800$ observations from each group is required to obtain 80% power. Another property as shown in 3.5 is that as δ increases, the power curve moves up as a whole, which is as expected.

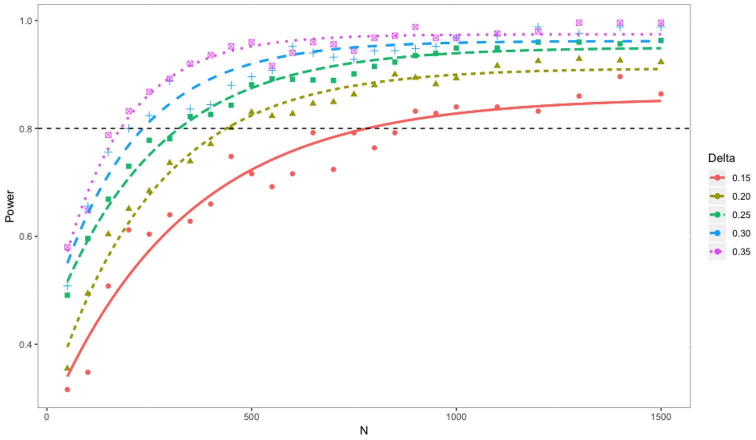


Figure 3.5: Bayesian powers vs N by δ, $\alpha = 0.05$

3.5 Conclusion

Sample size determination is a key step in many studies, such as costly human clinical trials. When the primary endpoint of the study is a continuous variable, the normal approximation is usually used to calculate the required sample size. However, if the normal distribution assumption is severely violated, such kind of approximation may underestimate the required sample size, resulting in a low-power design. Data transformation is an option, but as discussed in many works of literature, data transformation also could be problematic, for example, data transformation can not cope with negative skewness (left-skewed). The Bayesian sample size determination (SSD) based on the skew-normal distribution discussed in this chapter is a good alternative. The approach does not require normality assumption nor data transformation and allows for left skewness. In addition, the two groups of responses are allowed to have different skewness. Moreover, Bayesian SSD is able to incorporate prior information if applicable, thereby reducing the required sample size. The main shortcoming of Bayesian SSD is that it is based on intensive simulation, and using grid search to determine the sample size is relatively time-consuming and requires more computing resources. Our simulation experiments show that the cost of Bayesian SSD is completely affordable. In our example, an ordinary 8-core PC was used and all calculations were completed in 4 days with parallel computing. If more cores are available, the simulation can be completed in a short time. Therefore, we believe that the computational burden is negligible compared to the benefits of this approach.

APPENDICES

APPENDIX A

Stan Codes for Chapter I

A.1 Code for BYM2 Model

```
functions{
  real icar_normal_lpdf(vector phi, int N, int[] node1, int[]
     node2) {
    return -0.5 * dot_self(phi[node1] - phi[node2])
    + normal_lpdf(sum(phi) | 0,0.001 * N);
  }
}
data {
  int<lower=0> N;
  int<lower=0> N_edges;
  int<lower=1, upper=N> node1[N_edges];  // node1[i], node2[i]
  int<lower=1, upper=N> node2[N_edges];  // node1[i] < node2[i]
  int<lower=0> y[N];
  vector<lower=0>[N] E; //exposure
  int<lower=1> K;//number of covariates
  matrix[N,K] x;//design matrix
  real<lower=0> scaling_factor;//
}
transformed data{
  vector[N] logE=log(E);
}
parameters {
  real beta0;
  vector[K] betas;
  real logit_rho;//proportion of spatial effects
  vector[N] phi; //spatial effects
  vector[N] theta; //heterogeneous rand effects
  real<lower=0> sigma; //overall standard deviation
}
transformed parameters{
  real<lower=0,upper=1> rho=inv_logit(logit_rho);
  vector[N] convolved_re = sqrt(1 - rho) * theta + sqrt(rho /
     scaling_factor) * phi;
}
model {
  //likelihood
  y ~ poisson_log(logE + beta0 + x*betas + convolved_re*sigma);
  //priors
  beta0 ~ normal(0,10);
  betas ~ normal(0,10);
  logit_rho ~ normal(0,1);
  sigma ~ normal(0,1);
  theta ~ normal(0,1);
  phi ~ icar_normal(N, node1, node2);
}

generated quantities{
  vector[N] eta = logE + beta0 +x*betas + convolved_re*sigma;
  vector[N] mu = exp(eta);
  int y_rep[N];
  if(max(eta)>20){
    for(n in 1:N) y_rep[n] = -1;
  }else{
    for(n in 1:N) y_rep[n] = poisson_log_rng(eta[n]);
  }
}
```

A.2 Code for Full Model via BYM2

```
functions{
  real icar_normal_lpdf(vector phi, int N, int[] node1, int[]
     node2) {
    return -0.5 * dot_self(phi[node1] - phi[node2])
    + normal_lpdf(sum(phi) | 0,0.001 * N);
  }
}
data {
  int<lower=0> N;
  int<lower=0> N_edges;
  int<lower=1, upper=N> node1[N_edges];
  int<lower=1, upper=N> node2[N_edges];
  int<lower=0> z[N];//Observed counts
  vector<lower=0>[N] E;//exposure
  int<lower=1> K;
  matrix[N,K] X;//design matrix for counts
  int<lower=1> J;
  matrix[N,J] W;//design matrix for reporting rate
  real<lower=0> scaling_factor;
}
transformed data{
  vector[N] logE=log(E);
}
parameters {
  //coefficients for Poisson
  real beta0;
  vector[K] betas;
  //coefficients for Logistic
  real gamma0;
  vector[J] gammas;

  real logit_rho;
  vector[N] phi;// spatial random effects
  vector[N] theta; //unstructured random effects
  real<lower=0> sigma;
}
transformed parameters{
  real<lower=0,upper=1> p0;//reporting rate when W take average
  real<lower=0,upper=1> rho = inv_logit(logit_rho);
  vector[N] convolved_re;
  convolved_re = sqrt(1-rho)*theta + sqrt(rho/scaling_factor)*
     phi;
  vector<lower=0,upper=1>[N] p;
  vector<lower=0>[N] lambda;
  vector<lower=0>[N] mu;
  lambda = exp(logE + beta0 + x*betas + convolved_re*sigma);
  p0 = inv_logit(gamma0);
  p = inv_logit(gamma0 + w*gammas);
  mu = lambda .* p;
}
model {
  //likelihood
  z ~ poisson(mu);
  target += gamma0-2*log(1+exp(gamma0));
  //priors
  p0 ~ beta(2.42,35.25);//will induce prior on gamma0
  gammas ~ normal(0,10);
  beta0 ~ normal(0,10);
  betas ~ normal(0,10);
  logit_rho ~ normal(0,1);
  sigma ~ normal(0,1);
  theta ~ normal(0,1);
  phi ~ icar_normal(N, node1, node2);
}

generated quantities{
  vector[N] eta = logE + beta0 +x*betas + convolved_re*sigma;
  vector[N] lambda_rep = exp(eta);
  vector[N] p_rep = inv_logit(gamma0 + w*gammas);
  int y_rep[N];
  int z_rep[N];
  if(max(eta)>20){
    for ( n in 1:N) {
      y_rep[n] = -1;
      z_rep[n] = -1;
    }
  }else{
    for ( n in 1:N) {
      y_rep[n] = poisson_log_rng(eta[n]);
      z_rep[n] = poisson_rng(lambda_rep[n]*p_rep[n]);
    }
  }
}
```

APPENDIX B

Stan Codes for Chapter II

B.1 Stan Code for Misclassification Only

```
data {
  int<lower=0> N; // number of areas;
  int<lower=0> y1[N]; // response1
  int<lower=0> y2[N]; // response2
  int<lower=1> K; // number of covariates
  matrix[N,K] X; // design matrix
  vector<lower=0>[N] E; //exposure
}

transformed data{
  vector[N] logE=log(E);
}

parameters {
  vector[N] theta1; //unstructured random effect 1
  vector[N] theta2; //unstructured random effect 2
  real<lower=0> epsilon1;//hyper parameter for theta1
  real<lower=0> epsilon2;//hyper parameter for theta2
 real<lower=0,upper=1> p1;//Misclassified rate 1
 real<lower=0,upper=0.2> p2;//Misclassified rate 2
 vector[K] betas; //coefficients for y1
 vector[K] gammas; // coefficients for y2
}

// parameters transformation
transformed parameters{
  vector<lower=0>[N] mu1; //Poisson rate 1 with
     misclassification
  vector<lower=0>[N] mu2; //Poisson rate 2 with
     misclassification
  vector<lower=0>[N] l1; //True Poisson rate 1
  vector<lower=0>[N] l2; //True Poisson rate 2

  l1 = exp(X*betas  + logE + theta1);
  l2 = exp(X*gammas + logE + theta2);

  mu2 = l1*p1 + l2*(1-p2);
  mu1 = l1*(1-p1) + l2*p2;
}

model {
  //likelihood
  y1 ~ poisson(mu1);
  y2 ~ poisson(mu2);
  //priors
  theta1 ~ normal(0,epsilon1);
  theta2 ~ normal(0,epsilon2);
  epsilon1~normal(0,1);
  epsilon2~normal(0,1);
  //priors for coefficients and mis-rates
  betas ~ normal(0,1);
  gammas ~ normal(0,1);
  p1 ~ beta(10,30);
  p2 ~ beta(10,30);
}
```

B.2 Stan Code for MSAMM Only

```
functions {
  matrix kronecker_prod(matrix A, matrix B) {
    matrix[rows(A) * rows(B), cols(A) * cols(B)] C;
    int m;
    int n;
    int p;
    int q;
    m = rows(A);
    n = cols(A);
    p = rows(B);
    q = cols(B);
    for (r in 1:m) {
      for (s in 1:n) {
         for (v in 1:p) {
             for (w in 1:q){
             C[p*(r-1)+v,q*(s-1)+w] = A[r,s]*B[v,w];
             }
         }

      }
    }
    return C;
  }
}
data {
  int<lower=0> N;
  int<lower=1> q;
  int<lower=0> y1[N];
  int<lower=0> y2[N];
  int<lower=1> K;
  matrix[N,K] X;
  matrix[N,q] M;
  matrix[q,q] Ls;
  matrix[q,q] I
  vector<lower=0>[N] E;
}

transformed data{
  vector[N] logE=log(E);
}

parameters {
  vector[N] theta1;
  vector[N] theta2;
  real<lower=0> epsilon1;
  real<lower=0> epsilon2;
  corr_matrix[2] Omega;//correlation matrix between diseases
  vector[K] betas; //coefficient vector for y1
  vector[K] gammas; // coefficient vector for y2
  vector[2*q] phi; // joint spatial effects
  vector<lower=0>[2]ksi;//scaled vector for standard deviations
}

// parameters transformation
transformed parameters{
  cov_matrix[2] LAM;// covariance matrix between y1 and y2
  cov_matrix[2*q] Sigma; //covariance matrix for phi
  //vector<lower=0>[N] mu1; //rate 1
  //vector<lower=0>[N] mu2; //rate 2
  vector<lower=0>[N] l1; //True rate 1
  vector<lower=0>[N] l2; //True rate 2

  l1 = exp(X*betas + M*Ls1*phi[1:q] + logE + theta1);
  l2 = exp(X*gammas + M*Ls1*phi[(1+q):(2*q)] + logE + theta2);
  LAM = quad_form_diag(Omega, ksi);
  Sigma = kronecker_prod(LAM, I2);
}

model {
  //likelihood
  y1 ~ poisson(l1);
  y2 ~ poisson(l2);
  //prior for joint spatial effect
    phi~ multi_normal(rep_vector(0,2*q), Sigma);
  theta1 ~ normal(0,epsilon1);
  theta2 ~ normal(0,epsilon2);
  epsilon1~normal(0,1);
  epsilon2~normal(0,1);
  //priors for coefficients
  betas ~ normal(0,1);
  gammas ~ normal(0,1);
  //prior for scaled vector of sd
  ksi ~ normal(0, 2.5);
  //prior for correlation matrix
  Omega ~ lkj_corr(1);
}

generated quantities{
  vector[N] l1_rep;
  vector[N] l2_rep;
  int y1_rep[N];
  int y2_rep[N];
  l1_rep = exp(X*betas + M*Ls1*phi[1:q] + logE + theta1);
  l2_rep = exp(X*gammas + M*Ls1*phi[(1+q):(2*q)] + logE +
     theta2);
  for (i in 1:N){
    if(log(l1_rep[i])>20) l1_rep[i] = 9999999;
    if(log(l2_rep[i])>20) l2_rep[i] = 9999999;
  }
  y1_rep=poisson_rng(l1_rep);
  y2_rep=poisson_rng(l2_rep);
}
```

B.3 Stan Code for Full Model

```
data {
  int<lower=0> N; // num of areas;
  int<lower=1> q; // dimension for spatial effects after
     reduction.
  int<lower=0> y1[N]; // response1
  int<lower=0> y2[N]; // response2
  int<lower=1> K; // number of covariates
  matrix[N,K] X; // design matrix
  matrix[N,q] M; // eigenvector matrix
  matrix[q,q] Ls;// transformation matrix
  matrix[q,q] I; // q by q identity matrix

  vector<lower=0>[N] E; //exposure
}

transformed data{
  vector[N] logE=log(E);
}

parameters {
 real<lower=0,upper=0.4> p1;//mis-rate 1
 real<lower=0,upper=0.2> p2;//mis-rate 2
  corr_matrix[2] Omega;//correlation matrix between y1 and y2
  vector[K] betas; //coefficients for y1
  vector[K] gammas; // coefficients for y2
  vector[2*q] phi; // joint spatial effects
  vector[N] theta1;// unstructured RE 1
  vector[N] theta2;// unstructured RE 2
  real<lower=0> epsilon1; //hyper-para 1
  real<lower=0> epsilon2; //hyper-para 2
  vector<lower=0>[2] ksi;//scaled vector for standard
     deviations
}

// parameters transformation
transformed parameters{
  cov_matrix[2] LAM;//covariance matrix between y1 and y2
  cov_matrix[2*q] Sigma; //covariance matrix for phi
  vector<lower=0>[N] mu1; //rate 1
  vector<lower=0>[N] mu2; //rate 2
  vector<lower=0>[N] l1; //True rate 1
  vector<lower=0>[N] l2; //True rate 2

  l1 = exp(X*betas + M*Ls1*phi[1:q] + logE + theta1);
  l2 = exp(X*gammas + M*Ls1*phi[(1+q):(2*q)] + logE + theta2);

  mu2 = l1*p1 + l2*(1-p2);
  mu1 = l1*(1-p1) + l2*p2;

  LAM = quad_form_diag(Omega, ksi);
  Sigma = kronecker_prod(LAM, I2);
}

model {
  //likelihood
  y1 ~ poisson(mu1);
  y2 ~ poisson(mu2);
  //prior for joint spatial effect
  phi~ multi_normal(rep_vector(0,2*q), Sigma);
  theta1 ~ normal(0,epsilon1);
  theta2 ~ normal(0,epsilon2);
  epsilon1 ~ normal(0,1);
  epsilon2 ~ normal(0,1);
  //priors for coefficients
  betas ~ normal(0,1);
  gammas ~ normal(0,1);
  //Priors for mis-rates
  p1 ~ beta(10,40);
  p2 ~ beta(10,40);
  //prior for scaled vector of sd
  ksi ~ normal(0, 2.5);
  //prior for correlation matrix
  Omega ~ lkj_corr(1);
}

generated quantities{
  vector[N] l1_rep;
  vector[N] l2_rep;
  vector[N] mu1_rep;
  vector[N] mu2_rep;
  int y1_rep[N];
  int y2_rep[N];
  l1_rep = exp(X*betas + M*Ls1*phi[1:q] + logE + theta1);
  l2_rep = exp(X*gammas + M*Ls1*phi[(1+q):(2*q)] + logE +
     theta2);
  for (i in 1:N){
    if(log(l1_rep[i])>20) l1_rep[i] = 9999999;
    if(log(l2_rep[i])>20) l2_rep[i] = 9999999;
  }
  mu2_rep = l1_rep*p1 + l2_rep*(1-p2);
  mu1_rep = l1_rep*(1-p1) + l2_rep*p2;
  y1_rep=poisson_rng(mu1_rep);
  y2_rep=poisson_rng(mu2_rep);
}
```

REFERENCES

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Azzalini, A., & Arellano-Valle, R. B. (2012). *Maximum penalized likelihood estimation for skew-normal and skew-t distributions.*

Azzalini, A., & Genton, M. G. (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, *76*(1), 106–129.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data.* CRC press.

Barnard, J., McCulloch, R., & Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 1281–1311.

Bayes, C. L., & Branco, M. D. (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics*, 141–163.

Beavers, D. P., & Stamey, J. D. (2012). Bayesian sample size determination for binary regression with a misclassified covariate and no gold standard. *Computational Statistics & Data Analysis*, *56*(8), 2574–2582.

Bendavid, E., Mulaney, B., Sood, N., Shah, S., Ling, E., Bromley-Dulfano, R., ... others (2020). Covid-19 antibody seroprevalence in santa clara county, california. *MedRxiv*.

Bernardinelli, L., Clayton, D., & Montomoli, C. (1995). Bayesian estimates of disease maps: how important are priors? *Statistics in medicine*, *14*(21-22), 2411–2431.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192–225.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, *43*(1), 1–20.

Boots, B., & Tiefelsdorf, M. (2000). Global and local spatial autocorrelation in bounded regular tessellations. *Journal of Geographical Systems*, *2*(4), 319–348.

Brutti, P., De Santis, F., & Gubbiotti, S. (2008). Robust bayesian sample size determination in clinical trials. *Statistics in Medicine*, *27*(13), 2290–2306.

Carlin, B. P., Banerjee, S., et al. (2003). Hierarchical multivariate car models for spatio-temporally correlated survival data. *Bayesian statistics*, *7*(7), 45–63.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, *76*(1).

Daniel Paulino, C., Soares, P., & Neuhaus, J. (2003). Binomial regression with misclassification. *Biometrics*, *59*(3), 670–675.

Figueiredo, F., & Gomes, M. I. (2013). The skew-normal distribution in spc. *REVSTAT-Statistical Journal*, *11*(1), 83–104.

Gelfand, A. E., & Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, *4*(1), 11–15.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Ghosh, P., Branco, M. D., & Chakraborty, H. (2007). Bivariate random effect model using skew-normal distribution with application to hiv-rna. *Statistics in Medicine*, *26*(6), 1255–1267.

Guan, Y., & Haran, M. (2018). A computationally efficient projection-based approach for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *27*(4), 701–714.

Gupta, R. C., & Brown, N. (2001). Reliability studies of the skew-normal distribution and its application to a strength-stress model. *Communications in Statistics-Theory and Methods*, *30*(11), 2427–2445.

Hanks, E. M., Schliep, E. M., Hooten, M. B., & Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, *26*(4), 243–254.

Hasanalipour, P., & Sharafi, M. (2012). A new generalized balakrishnan skew-normal distribution. *Statistical papers*, *53*(1), 219–228.

Hodges, J. S., & Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, *64*(4), 325–334.

Hoffman, M. D., & Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, *15*(1), 1593–1623.

Hortacsu, A., Liu, J., & Schwieg, T. (2020). Estimating the fraction of unreported infections in epidemics with a known epicenter: an application to covid-19. *medRxiv*. Retrieved from https://www.medrxiv.org/content/early/2020/04/17/2020.04.13.20063511 doi: 10.1101/2020.04.13.20063511

Huang, A., Wand, M. P., et al. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, *8*(2), 439–452.

Hughes, J., & Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*(1), 139–159.

Jin, X., Carlin, B. P., & Banerjee, S. (2005). Generalized hierarchical multivariate car models for areal data. *Biometrics*, *61*(4), 950–961.

Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American journal of epidemiology*, *141*(3), 263–272.

Kempthorne, O. (1952). The design and analysis of experiments.

Lee, D. (2013). Carbayes: an r package for bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, *55*(13), 1–24.

Lindgren, F., Rue, H., et al. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, *63*(19), 1–25.

Liseo, B., & Loperfido, N. (2006). A note on reference priors for the scalar skew-normal distribution. *Journal of Statistical Planning and Inference*, *136*(2), 373–389.

Lyles, R. H. (2002). A note on estimating crude odds ratios in case–control studies with differentially misclassified exposure. *Biometrics*, *58*(4), 1034–1036.

MacNab, Y. C. (2011). On gaussian markov random fields and bayesian disease mapping. *Statistical Methods in Medical Research*, *20*(1), 49–68.

Martin, S. R., & Williams, D. R. (2017). Outgrowing the procrustean bed of normality: The utility of bayesian modeling for asymmetrical data analysis.

McInturff, P., Johnson, W. O., Cowling, D., & Gardner, I. A. (2004). Modelling risk when binary outcomes are subject to error. *Statistics in medicine*, *23*(7), 1095–1109.

Moore, T. J., Zhang, H., Anderson, G., & Alexander, G. C. (2018). Estimated costs of pivotal trials for novel therapeutic agents approved by the us food and drug administration, 2015-2016. *JAMA internal medicine*, *178*(11), 1451–1457.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37*(1/2), 17–23.

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., & DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. *Spatial and spatio-temporal epidemiology*, *31*, 100301.

Musgrove, D., Young, D. S., Hughes, J., & Eberly, L. E. (2019). A sparse areal mixed model for multivariate outcomes, with an application to zero-inflated census data. In *Modern statistical methods for spatial and multivariate data* (pp. 51–74). Springer.

Nicolaï, S. P., Teijink, J. A., Prins, M. H., et al. (2010). Multicenter randomized clinical trial of supervised exercise therapy with or without feedback versus walking advice for intermittent claudication. *Journal of Vascular Surgery*, *52*(2), 348–355.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical science: a review journal of the Institute of Mathematical Statistics*, *25*(1), 107.

Rajgor, D. D., Lee, M. H., Archuleta, S., Bagdasarian, N., & Quek, S. C. (2020). The many estimates of the covid-19 case fatality rate. *The Lancet Infectious Diseases*, *20*(7), 776–777.

Reich, B. J., Hodges, J. S., & Zadnik, V. (2006). Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, *62*(4), 1197–1206.

Ribeiro, L. C., Bernardes, A. T., et al. (2020). Estimate of underreporting of covid-19 in brazil by acute respiratory syndrome hospitalization reports. In *Ideas.*

Riebler, A., Sørbye, S. H., Simpson, D., & Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, *25*(4), 1145–1165.

Rosner, B. (2015). *Fundamentals of biostatistics.* Nelson Education.

Sartori, N. (2003). *Bias reduction of maximum likelihood estimates: skew normal and skew t distributions* (Tech. Rep.). Technical Report, Universita di Padova, Italy.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, *32*(1), 1–28.

Sood, N., Simon, P., Ebner, P., Eichner, D., Reynolds, J., Bendavid, E., & Bhattacharya, J. (2020). Seroprevalence of sars-cov-2–specific antibodies among adults in los angeles county, california, on april 10-11, 2020. *Jama.*

Sposto, R., Preston, D. L., Shimizu, Y., & Mabuchi, K. (1992). The effect of diagnostic misclassification on non-cancer and cancer mortality dose response in a-bomb survivors. *Biometrics*, 605–617.

Stamey, J. D., Young, D. M., & Seaman Jr, J. W. (2008). A bayesian approach to adjust for diagnostic misclassification between two mortality causes in poisson regression. *Statistics in medicine*, *27*(13), 2440–2452.

Stoner, O., Economou, T., & Drummond Marques da Silva, G. (2019). A hierarchical framework for correcting under-reporting in count data. *Journal of the American Statistical Association*, 1–17.

Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, *8*(2), 158–183.

Wang, F., Gelfand, A. E., et al. (2002). A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, *17*(2), 193–208.

Whittemore, A. S., & Gong, G. (1991). Poisson regression with misclassified counts: application to cervical cancer mortality rates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *40*(1), 81–93.

Winkelmann, R., & Zimmermann, K. F. (1993). *Poisson-logistic regression.* Volkswirtschaftl. Fakultät d. Ludwig-Maximilians-Univ. München.