ABSTRACT

BOEHM, LAURA FRANCES. Bridge Models and Variable Selection Methods for Spatial Data. (Under the direction of Brian Reich and Montserrat Fuentes.)

In hierarchical Bayesian modeling, spatial random effects are often introduced to capture spatial dependence. This leads to computational simplicity in many cases as observations are independent conditioned on the random effects which facilitates MCMC algorithms. However, when the inferential objectives relate to parameters in the marginal distribution, results of the random effects model can be difficult to interpret. In this thesis, we consider two such cases: spatial logistic regression and spatial variable selection.

Spatially-referenced binary data are common in epidemiology and public health. Owing to its elegant log-odds interpretation of the regression coefficients, a natural model for these data is logistic regression. To account for missing confounding variables that might exhibit a spatial pattern (say, socioeconomic, biological or environmental conditions), it is customary to include a Gaussian spatial random effect. Conditioned on the spatial random effect, the coefficients may be interpreted as log odds ratios. However, marginally over the random effects, the coefficients no longer preserve the log-odds interpretation, and the estimates are hard to interpret and generalize to other spatial regions. To resolve this issue, we propose a new spatial random effect distribution through a copula framework which ensures that the regression coefficients maintain the log-odds interpretation both conditional on and marginally over the spatial random effects. We present simulations to assess the robustness of our approach to various random effects, and apply it to an interesting dataset assessing periodontal health of Gullah-speaking African Americans. The proposed methodology is flexible enough to handle areal or geo-statistical datasets, and hierarchical models with multiple random intercepts.

Previous research has suggested a connection between ambient particulate matter (PM) exposure and acute health effects, but the effect size varies across the United States. Variability in the effect may partially be due to differing community level exposure and health characteristics, but also due to the chemical composition of PM which is known to vary greatly by location and over time. The objective of this paper is to identify particularly harmful components of this chemical mixture. Because of the large number of potentially highly correlated components, we must incorporate some regularization into a statistical model. We assume that at each location the regression coefficients come from a mixture model with the flavor of stochastic search variable selection, but with included coefficients smoothed toward a common mean. We create two extensions of SSVS to model these data. In the first, we use effect modifiers to directly model the inclusion probability as it varies across space. In the second, we utilize a copula to share information about variable inclusion and effect magnitude across locations. These mod-
els differ from current spatial variable selection techniques by accommodating both local and global variable selection. Both models are used to study the association between fine PM (PM $< 2.5 \mu m$) components, measured at 115 counties nationally over the period 2000-2008, and cardiovascular emergency room admissions among Medicare patients.
Bridge Models and Variable Selection Methods for Spatial Data

by
Laura Frances Boehm

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2013

APPROVED BY:

Brian Reich
Co-chair of Advisory Committee

Montserrat Fuentes
Co-chair of Advisory Committee

Ana-Maria Staicu

Yichao Wu

H. Christopher Frey
DEDICATION

To my parents, for all they have done, which is more than words can tell.
To my brother and sister, for always sharing a laugh and kind word.
And to David, my constant supporter and friend, whose encouragement, advice, and love has made this possible.
BIOGRAPHY

The author grew up in Wisconsin, graduating from Menomonie High School in 2004. She attended St. Olaf College in Northfield, Minnesota and received a Bachelor of Arts degree in Mathematics and Art History in 2008. She entered the graduate department of Statistics at North Carolina State University in the fall of 2004, earning her M. Stat. degree in 2010.
ACKNOWLEDGEMENTS

I would like to thank my advisors, Brian Reich and Montserrat Fuentes for their guidance and support. I also thank Francesca Dominici at Harvard School of Public Health for her assistance on the fine particulate matter projects and provision of the Medicare health data, and Dipankar Bandyopadhyay at University of Minnesota for his assistance on the bridge model project and provision of the dental health data. I would also like to thank the members who have served on my committee, Yichao Wu, Ana-Maria Staicu, Chris Frey, Dennis Boos, and Wenbin Lu.
# TABLE OF CONTENTS

**LIST OF TABLES** ................................................. vii

**LIST OF FIGURES** ............................................ ix

**Chapter 1 Introduction** ...................................... 1
  1.1 Copulas .................................................... 1
  1.2 Comparison of the Gaussian and t-copulas ................. 3
  1.3 Stochastic Search Variable Selection ...................... 4

**Chapter 2 Bridging Conditional and Marginal Inference for Spatially-Referenced Binary Data** ................. 6
  2.1 Introduction .............................................. 6
  2.2 The spatial bridge distribution .......................... 9
  2.3 MCMC implementation .................................... 12
  2.4 Simulation Study ......................................... 12
  2.5 Dental Data example ..................................... 19
  2.6 Conclusion .............................................. 24

**Chapter 3 Hierarchical Variable Selection using Effect Modifiers** .................. 27
  3.1 Introduction .............................................. 27
  3.2 Data ....................................................... 29
  3.3 Model ...................................................... 33
    3.3.1 Exchangeable Variable Selection ...................... 33
    3.3.2 Exchangeable Variable Selection with Effect Modifiers .. 34
  3.4 Methods .................................................... 35
    3.4.1 Full Conditionals for the EVS Model ................. 36
    3.4.2 Full Conditionals for VS with Effect Modifiers .......... 36
  3.5 Analysis of PM2.5 Component Data ......................... 37
    3.5.1 Models and Data ..................................... 37
    3.5.2 Results ............................................ 38
  3.6 Conclusion .............................................. 40

**Chapter 4 Spatial Variable Selection Methods for Investigating Acute Health Effects of Fine Particulate Matter Components** .................. 44
  4.1 Introduction .............................................. 44
  4.2 Data ....................................................... 45
  4.3 Model ...................................................... 46
    4.3.1 A Spatial Variable Selection Model ...................... 46
    4.3.2 A comparison of copula and other model solutions .......... 47
  4.4 Simulation study ......................................... 50
    4.4.1 Simulation Results .................................. 51
  4.5 Analysis of Health Data .................................. 52
    4.5.1 Description of data and models ....................... 52
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5.2 Model comparisons</td>
<td>55</td>
</tr>
<tr>
<td>4.5.3 Analysis results</td>
<td>56</td>
</tr>
<tr>
<td>4.5.4 Sensitivity Analysis</td>
<td>57</td>
</tr>
<tr>
<td>4.6 Discussion</td>
<td>59</td>
</tr>
</tbody>
</table>

**REFERENCES**                                        65

**APPENDICES**                                         70
- Appendix A Computing Details for Spatial Bridge Model 71
- Appendix B Supplmental Material for VS with Effects Modifiers 73
  - B.1 Maps of Effect Modifiers Used in PM Component Data Analysis 73
  - B.2 Simulated Example 76
- Appendix C Computing Details for Spatial Variable Selection Model 79
- Appendix D Spatial Variable Selection Model Results 81
Table 2.1  Relative percent bias of population-level coefficients when the underlying true model is Gaussian and bridge. ‘G’ indicates fit of the Gaussian model, ‘B’ indicates the bridge model fit, and ‘GEE’ indicates the GEE fit. Results which show a significant difference from the bridge model are indicated with ‘*’. 17

Table 2.2  $\sqrt{\text{MSE}} \times 100$ when the underlying true model is Gaussian and bridge. ‘G’ indicates fit of the Gaussian model, ‘B’ indicates the bridge model fit, and ‘GEE’ indicates the GEE fit. Results which show a significant difference from the bridge model are indicated with ‘*’. 18

Table 2.3  The relative MSE of the Gaussian and bridge models. 19

Table 2.4  Posterior parameter estimates and 95% credible intervals (C.I.) for tooth specific (T), subject specific (S) and population level (P) fixed-effects for the model with subject and tooth nested random effects. * indicates 95% C.I. that excludes 0. For the $t$-copula, there are $\nu = 2$ degrees of freedom. 23

Table 2.5  Variance components, $\phi_1$, $\phi_2$ and $r$ for the model with subject and tooth nested random effects. * indicates 95% C.I. that excludes 0. $\phi_1$ controls the distribution of $\gamma_i^+$, while $\sigma_1$ is the total subject-specific effect standard deviation given by $\text{sd} (\gamma_i^+)/\phi_2$. For the $t$-copula, there are $\nu = 2$ degrees of freedom. 24

Table 3.1  Median, interquartile range (IQR), and maximum observed value in $\mu g/m^3$ across all 115 sites in the period 2000-2008. Minimum values all approximately zero, and are not shown. The seven most massive components are listed first, with the rest in alphabetical order. 31

Table 3.2  Minimum, median, and maximum for number of enrollees on a given day, number of CVD hospitalizations on a given day, rate per 100,000 on a given day, and county average of daily rate per 100,000 from 2000-2008. 32

Table 3.3  Effect modifiers included in the VS model. 32

Table 3.4  Posterior mean for $\alpha$, $\pi_k$, and average value for $\beta_k^s$. For the models with effect modifiers, $\Phi(\xi_{kt})$ is the posterior probability of inclusion for a site with average values of all effect modifiers. Both $\alpha_k$ and $\beta_k$ are presented as % increase in relative risk of CVD hospitalization for an IQR increase in the pollutant. 39
Table 4.1 Median absolute deviation (MAD) \times 100 for simulation study. MAD averaged across all 50 locations for each of the \( k = 1, \ldots, 9 \) pollutants, and an overall average presented here, for the Spatial (Sp), Exchangeable Variable Selection (EVS), and Spatial Variable Selection (SpVS) models. Simulation settings for \( \alpha_k, \pi_k \) are shown in columns 2 and 3, and abbreviated names of the components used to generate the response are in column 1. For all pollutants, we generate data with \( \omega_k = 0.025 \) and effective range 0 km (Independent), 180 km (Weak Correlation), or 720 km (Strong Correlation). A * indicates a statistically significant difference from the SpVS model.

Table 4.2 Power and Type-I error for spatial model with no variable selection (Sp), exchangeable variable selection model (EVS), and spatial variable selection model (SpVS) for local and global covariates. We calculate power for each setting of the true value of \( \alpha_k \).

Table 4.3 Proportion of posterior median estimates of \( \beta_k(s) \) more than \( \pm 2\omega/\sqrt{C} \) away from zero.

Table 4.4 Number of times each model was selected as “best” by each of three goodness of fit measures: Deviance Information Criterion (DIC), the measure by Gelfand and Ghosh (GG), and log pseudo-maximum-likelihood (LPML).

Table 4.5 Estimated overall relative risk expressed as percent increase \( e^{\alpha_k \pi_k \times 100} - 100 \) and goodness of fit statistics. For Spatial model (no VS), \( e^{\alpha_k \times 100} - 100 \) is reported. For models with spatial correlation, range estimates in km are given. Statistically significant associations (0.05 level) marked by *. Global selection (posterior \( \pi_k > 0.8 \) marked by \( G \) and local selection marked by \( L \) (posterior \( 0.15 < \pi_k < 0.80 \)). We also include goodness of fit criterion, DIC, GG, and LPML. The best model by each measure is highlighted in bold.

Table 4.6 Estimates of mean when included \( e^{\alpha_k \times 100} - 100 \), probability of inclusion \( \pi_k \) from the Spatial VS model with 22 pollutants, and estimate of standard deviation, rescaled \( e^{\omega_k \times 100} - 100 \).
LIST OF FIGURES

Figure 1.1 The spike-and-slab prior used in Stochastic Search Variable Selection, with $\pi = 0.1$, $\omega = 1$, and $C = 150$. ................................................................. 5

Figure 2.1 Marginal log odds ratio for unit increase from $X$ to $X+1$, when the random effects follow a Gaussian or bridge distribution. In both graphs, gray lines have random effects standard deviation $\sigma = 0.2$, and black lines have random effects standard deviation $\sigma = 0.6$. ................................................................. 10

Figure 2.2 Bivariate density for (a) Bridge with Gaussian copula and $\rho = 0.3$ (b) Bridge with $t$-copula, $\rho = 0.3$ and $\nu = 2$ degrees of freedom (c) Bridge with Gaussian copula and $\rho = 0.8$ and (d) Bridge with $t$-copula, $\rho = 0.8$, and $\nu = 2$. In all graphs, $\phi = .95$, for a standard deviation of $\sigma = 0.6$. 15

Figure 2.3 Comparison of correlation by distance for bivariate Gaussian distribution and bridge distribution with $\phi = 0.95$ using a Gaussian and $t_2$ copula. The plots compare (a) the exponential correlation with range 0.1 and (b) the Matern correlation with range 0.1 and smoothness 1.5. .................. 16

Figure 2.4 Posterior estimates of tooth level random effects $\varepsilon_{ij}$ with central 90% credible intervals (a), and posterior probability of a missing or diseased tooth at various tooth locations (b) of an arbitrarily selected subject, obtained from the bridge model with Gaussian copula. The (dark) squares in the lower portion of each figure represent teeth that are missing or diseased. ................................................................. 25

Figure 3.1 Observed level of elemental carbon over the period 2002-2008 in 115 counties. Green indicates highest values, purple the lowest. At right are the median, min, and max values, and the daily averages are plotted along the bottom. ................................................................. 30

Figure 3.2 Average CVD hospitalizations per day per 100,000 enrollees by county. Inset shows Northeast US in greater detail. ................................. 33

Figure 3.3 Posterior estimates of $\beta_k$ and $\gamma_k$ for the EVS model without effect modification ((a)-(b)), the VS with effect modifiers model ((c)-(d)), and the unsmoothed VS with effect modifiers model when there is no shared mean for the $\xi_k$ coefficients ((e)-(f)). ................................................................. 41

Figure 3.4 Estimated RR for elemental carbon by county for EVS model (top) and Effect modifier VS model (bottom) ................................................................. 42

Figure 3.5 Posterior mean and standard deviation of $\xi$ parameters in the effect modifiers model with shared mean (Figures (a) and (c)) and with no shared mean ((b) and (d)). ................................................................. 43
Figure 4.1  One realization from the indicator model and from our copula prior on a 100 x 100 grid. In both, \(\alpha_k = 0.5, \omega_k = 0.2,\) and \(\pi_k = 0.5.\) The continuous part of the two-indicator model are drawn from a GPex(0.5, 0.2, 10) model, such that spatial correlation at neighboring sites is about 0.9. The binary part is \(\gamma_k(s) \sim \text{Bernoulli}\{\Phi[z_k(s)]\}\), where \(z_k(s)\) are GPex(0, 1, 10) and drawn independently from the continuous part. The latent copula \(\theta \sim \text{GPex}(0, 0, 10)\) from the same seed as the continuous part of the two-process model, to give similar pictures. This example is the copula model with \(C = \infty\) to give exact zeros.

Figure 4.2  A comparison of the joint density of \(\beta_k(s)\) and \(\beta_{k}(s')\) under an indicator model (a) and our copula model (b). In (a), \(\beta_k(s) = \gamma_k(s)B_k(s),\) where \(\gamma_k(s) \perp \gamma_k(s')\), and \(B_k(s), B_k(s')\) are multivariate normal with correlation \(\rho.\) In (b), our copula model where \(\gamma_k(s), \beta_k(s')\) joined by Gaussian copula with correlation \(\rho.\) For both plots, the marginal density is the mixture in (4.2), with \(\pi = 0.8, \alpha = 0.5, \omega = 0.5,\) and the correlation is \(\rho = 0.9.\)

Figure 4.3  Estimated county-specific relative risk for an IQR increase in elemental carbon from exchangeable VS model.

Figure 4.4  Estimated county-specific relative risk for an IQR increase in elemental carbon (a) and organic carbon (b) from the SpVS model.

Figure 4.5  Estimated county-specific relative risk for an IQR increase in elemental carbon (a) and organic carbon (b) from the Spatial model.

Figure 4.6  Matrix of all estimated coefficients (in RR per IQR increase) for each of the six models in Table 1. Each row includes coefficients for the labeled pollutant; each column indicates one county. The counties are grouped according to the 8 EPA subregions. The purpose of this plot is to highlight differences in local estimates for each of the models.

Figure B.1  Proportion of population which is Hispanic.

Figure B.2  Proportion of population which is Black.

Figure B.3  Proportion of adults over age 65 in poverty.

Figure B.4  Proportion of population commuting by public transit.

Figure B.5  Proportion of population living in an urban center.

Figure B.6  Median year built for all occupied housing.

Figure B.7  True values for \(\beta_k^s\) and \(\gamma_k^s\) from a simulated example and the resulting estimated parameters from an EVS model and a VS model with effect modifiers.

Figure D.1  Percent relative risk increase for one IQR increase of Sulfate from the SpVS model with 22 pollutants.

Figure D.2  Percent relative risk increase for one IQR increase of Nitrate from the SpVS model with 22 pollutants.

Figure D.3  Percent relative risk increase for one IQR increase of Silicon from the SpVS model with 22 pollutants.

Figure D.4  Percent relative risk increase for one IQR increase of Elemental carbon from the SpVS model with 22 pollutants.
Figure D.5  Percent relative risk increase for one IQR increase of Organic carbon from
the SpVS model with 22 pollutants. .............................................. 83
Figure D.6  Percent relative risk increase for one IQR increase of Sodium from the
SpVS model with 22 pollutants. ...................................................... 84
Figure D.7  Percent relative risk increase for one IQR increase of Ammonium from
the SpVS model with 22 pollutants. .............................................. 84
Figure D.8  Percent relative risk increase for one IQR increase of Aluminum from the
SpVS model with 22 pollutants. ..................................................... 85
Figure D.9  Percent relative risk increase for one IQR increase of Arsenic from the
SpVS model with 22 pollutants. ...................................................... 85
Figure D.10 Percent relative risk increase for one IQR increase of Bromine from the
SpVS model with 22 pollutants. ...................................................... 86
Figure D.11 Percent relative risk increase for one IQR increase of Calcium from the
SpVS model with 22 pollutants....................................................... 86
Figure D.12 Percent relative risk increase for one IQR increase of Chlorine from the
SpVS model with 22 pollutants. ...................................................... 87
Figure D.13 Percent relative risk increase for one IQR increase of Chromium from the
SpVS model with 22 pollutants. ...................................................... 87
Figure D.14 Percent relative risk increase for one IQR increase of Copper from the
SpVS model with 22 pollutants. ..................................................... 88
Figure D.15 Percent relative risk increase for one IQR increase of Iron from the SpVS
model with 22 pollutants. ............................................................. 88
Figure D.16 Percent relative risk increase for one IQR increase of Lead from the SpVS
model with 22 pollutants. ............................................................. 89
Figure D.17 Percent relative risk increase for one IQR increase of Magnesium from the
SpVS model with 22 pollutants. ...................................................... 89
Figure D.18 Percent relative risk increase for one IQR increase of Nickel from the SpVS
model with 22 pollutants. ............................................................. 90
Figure D.19 Percent relative risk increase for one IQR increase of Potassium from the
SpVS model with 22 pollutants. ...................................................... 90
Figure D.20 Percent relative risk increase for one IQR increase of Titanium from the
SpVS model with 22 pollutants. ...................................................... 91
Figure D.21 Percent relative risk increase for one IQR increase of Vanadium from the
SpVS model with 22 pollutants. ..................................................... 91
Figure D.22 Percent relative risk increase for one IQR increase of Zinc from the SpVS
model with 22 pollutants. ............................................................. 92
Chapter 1

Introduction

In this first chapter, we introduce background material and established methodology. We first introduce the idea of a copula, a mathematical structure for inducing correlation between random variables, which we use in Chapters 2 and 4. Secondly, we introduce Stochastic Search Variable Selection (SSVS), a common Bayesian computational approach to variable selection, which we extend in Chapters 3 and 4.

1.1 Copulas

A bivariate copula is a function \( C(u, v) \) from \([0,1] \times [0,1]\) to \([0,1]\) that has certain mathematical properties (see Nelsen (1999)). First discussed by Sklar (Sklar, 1953), a copula binds or joins a joint distribution to its marginals in such a way that the dependence between variables is independent of the marginal distributions. For one of our most familiar joint distributions, the multivariate normal distribution, the joint distribution fully characterizes the marginal distributions of each random variable, as well as their dependence. For normal marginal distributions, however, the usual multivariate distribution is not the only valid joint distribution. Conversely, the dependence structure implied by the multivariate normal distribution can also be applied to random variables with non-normal marginals to create a valid joint density. However, given the marginal distributions and dependence structure implied by the joint distribution of the usual multivariate normal, there is a unique copula (known as the Gaussian copula) that binds this joint distribution to its marginals. These results are concisely reported in Sklar’s Theorem for bivariate densities:

**Theorem 1. Sklar’s Theorem** Given random variables \( X,Y \), with (marginal) distribution functions \( F(x) = P(X < x) \) and \( G(y) = P(Y < y) \), and joint distribution \( H(x, y) = P(X < x, Y < y) \), there exists a copula, \( C_{XY} \) such that \( H(x, y) = C_{XY}[F(x), G(y)] \). If \( F \) and \( G \) are continuous, then \( C \) is unique. Further, given marginal cdfs \( F,G \), and any valid copula \( C \),
\[ H(x, y) = C(F(x), G(y)) \] is a valid joint distribution.

This result readily extends to \( n \)-dimensional multivariate distributions and copulas (Nelsen, 1999). As another example, consider one of the simplest copulas, the product copula. Clearly \( \Pi(u, v) = uv \) defines a valid joint distribution; for variables \( X \) and \( Y \) with marginal cdfs \( F \) and \( G \), the implied cdf is that for independent random variables, \( H(x, y) = \Pi(F(x), G(y)) = F(x)G(y) \). We can write the Gaussian copula described above as
\[
C = \Phi(\Sigma),
\]
where \( \Phi \) is the multivariate normal cdf with covariance matrix \( \Sigma \) and \( C \) refers to the resulting Gaussian copula.

We can see further the usefulness of copulas by considering the following theorem:

**Theorem 2: Invariance to transformation** Let \( X, Y \) be continuous random variables with marginal distribution functions \( F_1, G_1 \) and copula \( C_{XY} \) binding them to their joint distribution \( H_1(x, y) \). Let \( \alpha, \beta \) be strictly increasing functions on \( \text{Range}(X) \) and \( \text{Range}(Y) \) respectively. Then \( C_{XY} = C_{\alpha(X), \beta(Y)} \).

To prove, let \( F_2(x) = P(\alpha(X) < x) = P(X < \alpha^{-1}(x)) = F_1(\alpha^{-1}(x)) \), and similarly \( G_2(y) \) denote the cdfs of the random variables \( X^* = \alpha(X) \), \( Y^* = \beta(Y) \). Then, the joint distribution of \( \alpha(X), \beta(Y) \) is
\[
H_2(x, y) = C_{\alpha(X), \beta(Y)}(F_2(x), G_2(y)) = P(X < \alpha^{-1}(x), Y < \beta^{-1}(y)) = C_{XY}(F_1(\alpha^{-1}(x)), G_1(\beta^{-1}(y))) = C_{XY}(F_2(x), G_2(y)).
\]

That is, given a set of marginal and joint distributions \( F, G, H \) and their corresponding copula \( C_{XY} \), we can easily find the joint distribution of \( \alpha(X), \beta(Y) \). The same copula that binds the marginal of \( X \) and \( Y \) to their joint distribution \( H_1 \) can be used to bind the marginals of \( \alpha(X) \) and \( \beta(Y) \) to their joint distribution, \( H_2 \). This result also extends to the \( n \)-dimensional case (Nelsen, 1999). This reiterates the point that the dependence structure of the model is specified independently of the marginal distributions.

From the above, it becomes clear that we can use multivariate copulas to extend useful and familiar univariate distributions to valid multivariate joint distributions, with a dependence structure specified by our choice of copula. Within a Bayesian framework, this is easily accomplished within a Markov chain Monte Carlo (MCMC) sampling scheme through latent variables. Let \( \theta = (\theta_1, \ldots, \theta_m) \) be multivariate normal, with mean zero, and correlation \( \Sigma \). If we let \( F_j \) denote the desired marginal distribution for observation \( j \), then \( Y_j = F_j^{-1}[\Phi(\theta_j)] \) are correlated random variables with marginal distribution \( F_j \). If \( F_j \) is continuous, we can write the joint distribution of \( Y_j \) in closed form as a function of \( \theta_j \). If not, the joint distribution is still a valid distribution function by Theorem 2. This MCMC latent variable sampling scheme will be implemented in Chapter 2 as an extension of the bridge distribution for logistic regression with random intercepts, and in Chapter 4 for spatial variable selection.
1.2 Comparison of the Gaussian and t-copulas

The Gaussian copula is often criticized because of the lack of dependence in the tails of the distribution. The t-copula, which is generated from the multivariate-t distribution, is another useful copula, which shares the elliptical properties of the Gaussian copula, but allows for greater dependence in the tails. First, the t-copula is defined as

$$C^T_{\nu, \Sigma}(u_1, \ldots, u_n) = T_{\nu, \Sigma}(t^{-1}_\nu(u_1), \ldots, t^{-1}_\nu(u_n))$$

where $t^{-1}_\nu$ is the inverse cdf of a univariate central t-distribution with $\nu$ degrees of freedom, and $T_{\nu, \Sigma}$ is the multivariate-t cdf. The $d$-dimensional multivariate-t density with $\nu$ degrees of freedom, center $\mu$, and covariance $\Sigma$, is

$$\frac{\Gamma\left(\frac{\nu+d}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\pi\nu^d|\Sigma|}} \left(1 + \frac{(x - \mu)^T\Sigma^{-1}(x - \mu)}{\nu}\right)^{-\frac{\nu+d}{2}}$$

as described by Demarta and McNeil (2005), and we write $x \sim t_d(\nu, \mu, \Sigma)$.

Note that the multivariate-t can be represented as a multivariate normal mixture, such that $x \overset{d}{=} \mu + \sqrt{W}z$, where $z$ is independent of $W$, $z \sim N(0, \Sigma)$ and $1/W \sim \text{gamma}(\nu/2, \nu/2)$. This fact is simply an extension of the univariate case, a result which is familiar from inducing a t distribution in Gibbs’ sampling through a hyperprior on the variance. This fact will also make simulation and MCMC sampling from a t-copula computationally convenient. Note that as the copula is invariant to any monotone transformation (see Theorem 2 above), utilizing a correlation matrix as $\Sigma$ in the t-copula will be sufficient.

The differences between the Gaussian copula and t-copula can be dramatically demonstrated using joint quantile exceedance probabilities, that is, the probability that multiple extreme events will happen simultaneously. The joint exceedance probabilities for the t-copula are greater than those for the Gaussian copula. As one would expect, the ratio of the t to Gaussian exceedance probabilities grows as degrees of freedom decreases (t becomes less like Normal), but also grows with increasing correlation and dimension of the copula. When extreme values are of interest, misspecification of the copula can become extremely important, for instance in hydrology and finance. As an extreme example, Demarta and McNeil (2005) describe finding the probability that 5 correlated stocks will drop to the lowest 1% of their marginal distributions on the same day. When the joint distribution is generated using a $t_4$ copula, and bivariate correlations of 50%, this event will occur approximately every 7 years. However, a Gaussian copula fit to these data suggest it is a 1 in 50 year event.

Another common measure to assess tail dependence is the upper and lower tail coefficient,
and asymptotic conditional probability defined as
\[
\lim_{q \to 1} P(X_2 > F_2^{-1}(q) | X_1 > F_1^{-1}(q)) = \lambda_u
\]
for the upper coefficient, and a similar definition for \( \lambda_l \), the lower coefficient. For elliptically symmetric copulas such as the Gaussian and t-copulas, both coefficients are the same, and we will refer simply to \( \lambda \). For the Gaussian copula, \( \lambda = 0 \), and for the t-copula, \( \lambda = 2t_{\nu+1} \left( -\sqrt{\nu + 1} \sqrt{1 - \rho}/\sqrt{1 + \rho} \right) \), which will always be positive (Demarta and McNeil, 2005), again indicating stronger tail dependence for the t copula. We will use both the t copula and Gaussian copula in Chapter 2, to implement a spatially correlated bridge distribution for random effects in logistic regression.

### 1.3 Stochastic Search Variable Selection

When faced with a large number of potential predictors, variable selection can be a powerful tool, both statistically and scientifically. By reducing the number of variables in the model, standard error of estimates and prediction intervals decrease. From a scientific perspective, variable selection can help identify which variables are most important, or at least most strongly associated with the outcome. From a Bayesian perspective, optimal model selection would involve fitting all possible models, and comparing Bayes factors. As the number of potential models with \( p \) predictors is \( 2^p \), this quickly becomes intractable. Stochastic Search Variable Selection (George and McCulloch, 1993; George and McCulloch, 1997) has become a vastly popular tool for selecting among a large set of potential predictors, as the highest probability models are sampled using MCMC methods. Consider a linear model relating a response \( y_t \) for \( t = 1, \ldots, T \) and covariates \( x_{kt} \) for \( k = 1, \ldots, p \) predictors,

\[
y_t = x_{1t} \beta_1 + \ldots + x_{pt} \beta_p + \varepsilon_t,
\]

where \( \varepsilon_t \) are Gaussian errors. Let \( \beta_k \) be the coefficient on predictor \( k \), and \( \beta_k = \gamma_k B_k \) where, \( \gamma_k \) is a binary random variable, and \( B_k \) continuous. This implies the \( k \)th covariate, \( x_k \), is “in” if \( \gamma_k = 1 \), and “out” if \( \gamma_k = 0 \). The stochastic search model easily allows for calculating the posterior mean of \( \gamma_k \), i.e., the posterior probability that the \( k \)th covariate is in the model, as well as the joint posterior probability of any subset of \( \gamma_1, \ldots, \gamma_p \), thus allowing for an easy way to compare models without having to compute Bayes factors for every possible subset. SSVS is implemented within Gibbs sampling where \( \gamma_k \sim \text{Bernoulli}(\pi) \) and \( B_k \sim N(0, \omega^2) \), where \( \omega^2 \) is some large variance. For the linear model with Gaussian errors the model is fully conjugate.

A more general formulation of the SSVS model writes the distribution of \( \beta_k \) as a Gaussian
mixture with probability $\pi$,

$$
\beta_k \sim \begin{cases} 
N(0, \omega^2) & \text{with probability } \pi \\
N(0, \omega^2/C) & \text{with probability } 1 - \pi
\end{cases}
$$

(1.1)

where $C$ is a large number controlling the ratio of variance when $\beta_k$ is included in the model and when it is excluded. When $C = \infty$, this reduces to the binary-product case above. This prior distribution is shown in Figure 1.1. This prior reflects the belief that many covariates will have no appreciable effect on the response ($\beta_k \approx 0$), and some will have a large effect. In this thesis we consider a hierarchical Bayesian model in which we are interested in selecting amongst $p$ predictors at $n$ sites. In Chapter 3, we consider modeling the inclusion probability $\pi$ as a function of site specific covariates. In Chapter 4, we extend SSVS to a spatial model using a copula.
Chapter 2

Bridging Conditional and Marginal Inference for Spatially-Referenced Binary Data

2.1 Introduction

Spatially-referenced binary data abounds in the fields of epidemiology, public health, geography and image processing, among others. For valid inference on model parameters, it is important to account for correlation indexed by proximity due to the unmeasured spatial factors. While linear mixed models easily allow the incorporation of correlated random effects (REs), logistic regression and other generalized linear mixed models (GLMM) suffer an interpretation problem (i.e. a mismatch between conditional and marginal distributional shapes) due to the nonlinearity of the mean function. Because of this, researchers usually choose between a marginal model for ‘population-averaged’ inference, or a conditional model for ‘subject-specific’ inference for the regression parameters, primarily motivated by inferential objectives. The regression effects can be different at each level due to subject or site level variability, particularly when this variability is large, highlighting the importance of proper interpretation. While model goodness of fit is certainly an important consideration in choosing a random effects distribution, interpretability is essential. In this chapter, we propose a model for spatial binary data that preserves the log odds interpretation of covariate effects both conditional on and marginal over REs.

The motivating data in this study concern the periodontal health of Gullah-speaking African-American Type-2 diabetics (Fernandes et al., 2009; Bandyopadhyay et al., 2010). For each subject, the binary health status (missing/diseased) of each tooth (28 locations, excluding the 4 wisdom teeth) was recorded. The primary objective of this study was to assess the dental health of the population, and evaluate the influence of various subject-level covariates such as age and
gender on the periodontal decay status of each tooth, which requires properly accounting for clustering within subjects and spatial dependence between teeth. Earlier studies (Reich et al., 2007; Reich and Bandyopadhyay, 2010) have shown that periodontal disease markers might be spatially-associated, as a diseased (or decayed) tooth (or sites within a tooth) might be influencing the periodontal health status of a set of neighboring sites or teeth. Therefore, in addition to independent subject-specific RE, we include spatially correlated RE for each tooth nested within subjects.

The usual Gaussian RE logistic regression model for spatially-referenced binary data may not be well-suited here where the regression coefficients are only interpretable conditioned on the spatial REs, and thus in terms of replication at the same location for a given subject. In a more traditional disease mapping setting (with many subjects observed at each spatial location) when researchers are possibly interested in a site-specific interpretation (conditional on county of residence), these conditional parameters have meaning. However, for our data, an interpretation conditioned on tooth REs is purely hypothetical as we cannot take another replicate of the same tooth for the same subject. A similar objection arises for interpreting subject-level covariate effects such as gender, age, etc conditional on the subject. However, coefficients for certain ‘tooth-level’ measures, such as an indicator of molar, could plausibly be interpreted conditional on a subject RE. Hence, we desire a model with interpretable parameters both conditioned on, and marginal over subject and tooth RE.

Standard modeling approaches in a GLMM choose either a marginal or conditional route for inference. For the marginal, one approach is the moment-based generalized estimating equations (GEE) framework (Liang and Zeger, 1986) accommodating a spatial covariance matrix to yield consistent regression estimates which are robust to the misspecification of the underlying correlation structure. However, the GEE not being a likelihood-based method cannot be compared to other likelihood-based methods in terms of efficiency. Likelihood-based methods include direct modeling of the binary response using the multivariate logistic distribution described by O’Brien and Dunson (2004), or the marginalized models described by Heagerty (1999) and Heagerty and Zeger (2000). Both allow for complex correlation structures. While the O’Brien and Dunson method implies correlation directly on the binary responses without using REs, the logistic-normal model of Heagerty (1999) defines the conditional mean through a convolution equation, such that the marginally specified model provides the desired marginal interpretation of the regressors. However, due to the convolution, the conditional log-odds interpretation is lost. Other existing methods focus on conditional, rather than marginal, modeling. The usual conditionally-specified Gaussian REs models yields a conditional log-odds interpretation, as noted above. The marginal log-odds ratio is nonlinear and not in closed form, as will be described in Section 2, which is problematic for interpretation. Another conditional approach which has been applied to these data is the auto-logistic model (Besag, 1972; Bandyopadhyay...
et al., 2009). In this Markovian model, the coefficients have a log-odds interpretation conditioned on a pre-defined neighbor set. Though the full conditional distributions are intuitive, they lead to a complicated joint spatial distribution (Varin et al., 2011), and the marginal effects of covariates are not readily available. Further, it is unclear how to extend this model to geostatistical data defined continuously on a given spatial domain.

Wang and Louis (2003, 2004) made headway in solving the conditional-marginal dilemma for random-intercept logistic regression by proposing a new distribution (for the random intercepts), aptly named the ‘bridge distribution’. The marginal and conditional mean have the same form, and the marginal regression coefficients are proportional to the conditional regression coefficients. It falls under the general definition of marginalized models (Griswold and Zeger, 2004), amenable to a likelihood-based analysis through standard software, but is unique in that it retains the log-odds ratio interpretation both conditionally and marginally.

Motivated by the periodontal data, here we exploit the richness of the Wang and Louis model to study marginal/population-level covariate effects for spatially distributed binary data. This strength becomes further important in richer hierarchical models, such as the nested subject- and site-specific REs we use for the periodontal data, where the coefficients can easily be interpreted at whichever level of the hierarchy is desired. Extension to this setup for bivariate binary responses (Li et al., 2011) and a binary and continuous response (Lin et al., 2010) have been demonstrated. Here we provide a full exploration of a multivariate bridge distribution and its application in a nested REs setting. The spatial bridge distribution is derived using a Gaussian copula, and extensions to a more general $t$-copula are also presented. Although our model development is currently tailored towards the specific structure of dental data, it can be applied to more conventional spatial settings with a single observation at each location, or with multiple subjects observed at the same site. Our model that incorporates both subject- and site-specific REs can simultaneously estimate covariate effects conditioned on both the REs, marginally over subject- and site-specific REs. Identification of high-risk areas through estimation of REs is possible within our unified framework, often precluded in a marginal model. Our hierarchical Bayesian scheme has computational complexity equivalent to the usual Gaussian RE model, and can easily be implemented in standard software such as OpenBUGS.

The chapter proceeds as follows. Section 2.2 introduces the spatial bridge distribution and Section 2.3 presents the associated MCMC computational scheme. In Section 2.4, we conduct simulation studies to evaluate the robustness of our methods under misspecifications of the REs. Section 2.5 provides analysis of the binary dental data. Section 2.6 concludes.
2.2 The spatial bridge distribution

For notational convenience, we specify the model assuming all \( n \) subjects have observations at the same \( m \) spatial locations \( s_1, \ldots, s_m \). Extending to more complex designs is straightforward using our Bayesian hierarchical model. In particular, we do not require replication at each spatial location or a balanced design across subjects. The binary response \( Y_{ij} \) for subject \( i \) at site \( s_j \) is modeled as \( Y_{ij} \mid \varepsilon_{ij} \sim \text{Bernoulli}(\pi_{ij}) \), with \( \text{logit}(\pi_{ij}|X_{ij}, \varepsilon_{ij}) = X_{ij}\beta^S + \varepsilon_{ij} \), where \( \pi_{ij} \) is the Bernoulli success probability, \( X_{ij} \) is the design matrix of covariates, \( \beta^S \) is the regression vector (the superscript ‘\( S \)’ denotes site-specific conditional parametrization of regressors), and \( \varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{im})' \) is the vector of spatial REs for subject \( i \), which are independent and identically distributed.

In addition to the Bernoulli probability conditional on the REs, we are also interested in the Bernoulli probability marginal over the REs, \( \pi^P_{ij} = \int \pi_{ij} \cdot g(\varepsilon_{ij}) \, d\varepsilon_{ij} \). For most RE densities \( g(\varepsilon_{ij}) \), including Gaussian, this marginal probability is unknown and must be computed numerically. In the spirit of Wang and Louis (2003), to preserve the logistic shape both conditionally and marginally we allow \( \varepsilon_{ij} \) to follow the bridge density given by

\[
g(\varepsilon; \phi) = \frac{1}{2\pi} \frac{\sin(\phi \pi)}{\cosh(\phi \varepsilon) + \cos(\phi \pi)},
\]

where \( \cosh(x) = \frac{1}{2}(e^x + e^{-x}) \). We denote this as \( \varepsilon \sim \text{bridge}(\phi) \), where \( 0 < \phi < 1 \) controls the variance and shape of the density. The bridge density is mean 0 and variance \( \sigma^2_b = \frac{\pi^2}{3}(\phi^{-2} - 1) \), and like the normal distribution is symmetric and bell-shaped, though it has heavier tails. We assume \( \phi \) to be common for all observations to assure exchangeability across the spatial units. With such a bridge structure for \( \varepsilon \), the marginal regression model is

\[
\text{logit}(\pi^P_{ij}) = X_{ij}\beta^P,
\]

where \( \beta^P \) is the marginal (population-level) parameter vector and is related to \( \beta^S \) by \( \beta^P = \phi \beta^S = \frac{1}{\sqrt{1 + \frac{1}{2\pi^2} \sigma^2_b}} \beta^S \). This is a development over the marginal interpretation under the Gaussian REs model. The exact marginal log odds for Gaussian REs is nonlinear in \( X \), with nonlinearity increasing with REs variance and the strength of conditional association, \( \beta^S \). Johnson and Kotz (1970) give the approximation \( \text{logit}(\pi^P_{ij}) \approx X_{ij}\beta^P \) where \( \beta^P = \frac{1}{\sqrt{1 + (\frac{\pi^2}{3})^2 \sigma^2}} \beta^S \). Figure 2.1 shows the relationship between the exact and approximate marginal log odds ratio (calculated numerically as a function of \( X \)) for a model with a single predictor. For comparison, the marginal log odds ratio that would result from assuming a bridge random effects distribution with the same variance is also shown. We observe notable differences between these curves, and that as \( X \) approaches either \( \infty \) or \( -\infty \), the log odds for the exact Gaussian model approaches \( \beta^S \).
To extend the bridge model from the exchangeable to the spatial setting, we make use of a copula (Sklar, 1953, 1973; Nelsen, 1999) to capture spatial correlation while preserving the marginal bridge distribution at each location. Let $\theta_i$ be a latent Gaussian process for subject $i$ with zero mean, unit variance and spatial correlation function $\text{cor}([\theta(s), \theta(s')] = \rho(\|s - s'\|)$. One example of a correlation function is the exponential correlation defined by $\rho(\|s - s'\|) = e^{-\|s - s'\|/r}$, where $r > 0$ controls the spatial range of the latent process. Marginal bridge distributions are forced via the probability integral transformation

$$
\varepsilon_{ij} = G^{-1}_\phi\{\Phi[\theta_i(s_j)]\},
$$

where $\Phi$ is the cumulative distribution function (c.d.f) of $N(0, 1)$ and $G^{-1}_\phi$ is the inverse c.d.f of the univariate bridge density available in a closed form (Li et al., 2011) given by $G^{-1}_\phi(y) = \frac{1}{\phi} \log \left[ \frac{\sin(\phi y)}{\sin(\phi(1 - y))} \right]$, for $0 < y < 1$. We fix the variance of $\theta_i(s)$ at one for all $s$, because only the correlation structure is important in the copula specification. If $\theta_i(s)$ has a non-unit variance $\sigma^2$, it would be necessary to define $\varepsilon_{ij} = G^{-1}_\phi\{\Phi[\theta_i(s_j)/\sigma]\}$ to maintain the bridge marginal distribution, eliminating the effect of $\sigma$. 

Figure 2.1: Marginal log odds ratio for unit increase from $X$ to $X+1$, when the random effects follow a Gaussian or bridge distribution. In both graphs, gray lines have random effects standard deviation $\sigma = 0.2$, and black lines have random effects standard deviation $\sigma = 0.6$. 

To extend the bridge model from the exchangeable to the spatial setting, we make use of a copula (Sklar, 1953, 1973; Nelsen, 1999) to capture spatial correlation while preserving the marginal bridge distribution at each location. Let $\theta_i$ be a latent Gaussian process for subject $i$ with zero mean, unit variance and spatial correlation function $\text{cor}([\theta(s), \theta(s')] = \rho(\|s - s'\|)$. One example of a correlation function is the exponential correlation defined by $\rho(\|s - s'\|) = e^{-\|s - s'\|/r}$, where $r > 0$ controls the spatial range of the latent process. Marginal bridge distributions are forced via the probability integral transformation

$$
\varepsilon_{ij} = G^{-1}_\phi\{\Phi[\theta_i(s_j)]\},
$$

where $\Phi$ is the cumulative distribution function (c.d.f) of $N(0, 1)$ and $G^{-1}_\phi$ is the inverse c.d.f of the univariate bridge density available in a closed form (Li et al., 2011) given by $G^{-1}_\phi(y) = \frac{1}{\phi} \log \left[ \frac{\sin(\phi y)}{\sin(\phi(1 - y))} \right]$, for $0 < y < 1$. We fix the variance of $\theta_i(s)$ at one for all $s$, because only the correlation structure is important in the copula specification. If $\theta_i(s)$ has a non-unit variance $\sigma^2$, it would be necessary to define $\varepsilon_{ij} = G^{-1}_\phi\{\Phi[\theta_i(s_j)/\sigma]\}$ to maintain the bridge marginal distribution, eliminating the effect of $\sigma$. 

10
Utilizing a Gaussian copula, the joint cumulative distribution of \( \varepsilon_i \), is given by

\[
H(\varepsilon_{i1}, \ldots, \varepsilon_{im}) = \Phi_{\Sigma_r} \left\{ \Phi^{-1}[G_\phi(\varepsilon_{i1})], \ldots, \Phi^{-1}[G_\phi(\varepsilon_{im})] \right\},
\]

where \( \Sigma_r \) is the \( m \times m \) covariance matrix of \( \theta_i = (\theta_{i1}, \ldots, \theta_{im}) \). We denote this model as \( \varepsilon_i \sim \text{bridge}(\phi, \Sigma_r) \). Though motivated here from a geostatistical perspective, any valid correlation matrix, including the correlation implied by a conditionally autoregressive (CAR) model for areal data (Banerjee et al., 2004) could be used in place of \( \Sigma_r \). Under this copula model, the random effects \( \varepsilon_i \) are non-Gaussian, but maintain the Markovian dependency structure of the CAR model.

Although the Gaussian copula is an intuitive way to induce spatial correlation, it may not capture tail dependence (Demarta and McNeil, 2005). A logical alternative is the \( t \)-copula, which can be easily implemented within the Bayesian framework. The \( t \)-copula assumes that \( \theta_i \overset{iid}{\sim} t_\nu(0, \Sigma_r) \), where \( t_\nu \) is the multivariate \( t \) distribution with location 0, scale matrix \( \Sigma_r \) and \( \nu \) degrees of freedom. We can either fix \( \nu \), or assign it an additional hyperprior. Similarly as above, \( \varepsilon_{ij} = G^{-1}_\phi [T_\nu(\theta_{ij})] \), where \( T_\nu \) is the c.d.f of a mean 0 univariate \( t \) distribution with \( \nu \) degrees of freedom. Visual comparison (Figure 2.2) of the bivariate bridge densities for the Gaussian and \( t \)-copula at two levels of correlation reveal differences particularly in the tails. More flexible copulas such as the non-parametric copula as in Fuentes et al. (2013) can be used, but with the limited information available in our binary response, the parametric \( t \)-copula is likely flexible enough to capture important dependence.

To quantify the effect on the correlation function due to the probability integral transformation, Figure 2.3 plots the correlogram assuming exponential and Matern spatial correlation functions, and Gaussian and \( t \) copulas. We find that for both correlation functions and copulas, the correlation function of the copula models has the same general shape as correlation function of the latent Gaussian process, but the magnitude of correlation is slightly lower. This may be useful when specifying informative priors for the spatial correlation parameters.

Spatial prediction of the random effect \( \varepsilon_i(s_0) \) at prediction location \( s_0 \) follows from standard Bayesian Kriging methods. Using the Gaussian copula, we sample \( \theta_i(s_0)|\theta \) using properties of the conditional distribution of a multivariate normal, and then transform to \( \varepsilon_i(s_0) = G^{-1}_\phi \{ \Phi[\theta_i(s_0)] \} \). For the \( t \)-copula, the same approach can be used after exploiting the hierarchical representation of the multivariate \( t \). If we assume that \( \theta_i|\tau_i \sim N(0, \tau_i^2 \Sigma_r) \), where \( \tau_i^{-2} \sim \text{gamma}(\nu/2, \nu/2) \), then \( \theta_i \) follows a multivariate \( t_\nu(0, \Sigma_r) \) marginally over \( \tau_i \). Spatial interpolation can be done conditional on \( \tau_i \) as with the Gaussian copula.

To accommodate the structure of the dental data, we also consider a nested REs model in
which there are REs for both subjects and sites within subjects, given as:

$$\text{logit}\{P(Y_{ij} = 1|X_{ij}, \gamma_i, \varepsilon_{ij})\} = X_{ij}\beta^{T} + \gamma_i + \varepsilon_{ij},$$  \hspace{1cm} (2.3)

where $\gamma_i$ are independent subject-specific REs and $\varepsilon_{ij}$ are tooth-level REs. A slight complication arises with multiple REs because the bridge distribution is not additive or a scale family. As suggested by Wang and Louis (2003), we assume $\gamma_i = \gamma_i^* / \phi_2$, and that $\gamma_i^* \sim \text{bridge}(\phi_1)$ and $\varepsilon_i \sim \text{bridge}(\phi_2, \Sigma_r)$. In this case we are interested in estimating the log odds at each level of the hierarchy. We let the ‘site-within-subject’ level coefficient $\beta^T$ denote the conditional log odds ratio, using ‘T’ to signify a ‘tooth’-level interpretation. After integrating over the tooth REs $\varepsilon_i$, we have logit\{P(Y_{ij} = 1|X_{ij}, \gamma_i^*)\} = X_{ij}\beta^{S} + \gamma_i^*$, and the coefficient $\beta^{S} = \phi_2\beta^{T}$ represents the ‘subject’-level interpretation, in that it can only be interpreted conditional on the subject REs $\gamma_i$. Finally, $\beta^P$ is the ‘population’-level, or completely marginal coefficient, with logit\{P(Y_{ij} = 1|X_{ij})\} = X_{ij}\beta^{P}$. The population level log odds ratios are defined by $\beta^P = \phi_1\phi_2\beta^{T}$. Also, we denote the total subject level standard deviation $\sigma_1 = \text{sd}(\gamma_i)$, and the standard deviation of the bridge REs controlled by $\phi_1$, $\sigma_1^* = \text{sd}(\gamma_i^*)$. Further, $\sigma_2 = \text{sd}(\varepsilon_{ij})$ represents the standard deviation of the site-specific REs, controlled by $\phi_2$. A nested RE structure is also easily implemented in other studies where perhaps $\gamma_i$ are spatially correlated (e.g. county effects) and $\varepsilon_{ij}$ are independent (e.g. subjects within a county).

### 2.3 MCMC implementation

We implement the model using R (R Development Core Team, 2010) (Web Appendix A) for the simulation study, and OpenBUGS (Web Appendix B) for the dental data example. OpenBUGS coding for this model is a very simple extension from the usual Gaussian model, by treating $\varepsilon_{ij}$ as a transformation of a latent Gaussian variable $\theta_{ij}$. Implementation of the model with the $t$-copula requires the $t_\nu$ cdf, which, unlike the Gaussian cdf, is not an available function in OpenBUGS. However, this cdf exists in closed form for $\nu = 2$, and so for the dental analysis with $t$-copula, we hold $\nu$ fixed at 2, rather than include a prior for $\nu$. We ran 10,000 samples with 3000 burn-in for the simulated data, and visually monitored convergence by starting multiple chains from diverse starting values for a selection of datasets. For the dental data analysis, we used 25,000 iterations with 5000 burn-in.

### 2.4 Simulation Study

In this simulation study we compare marginal and conditional estimates of regression coefficients for models using Gaussian and bridge REs. Our objective is to study robustness of the coefficient
estimates to misspecification of the RE distribution and identify the situations that lead to the most sensitivity. Earlier work has already considered the effect of misspecification in logistic regression models with independent, identically distributed random intercepts. Neuhaus et al. (1992) show that for a true RE distribution \( G \), and an assumed RE distribution \( F \), the maximum likelihood estimates under \( F \) minimize Kullback-Leibler divergence of the marginal models implied under \( G \) and \( F \). Asymptotically and in simulation, they show that for any \( F \) and \( G \) the bias of the MLE for the conditional coefficient \( \beta^S \) are expected to be small, but that bias for \( \sigma \) may be large. It is unclear whether these results extend to a correlated intercepts setting and a Bayesian model. Furthermore, because the marginal interpretation depends directly on variance, and Neuhaus et al. (1992) show the variance is heavily biased, it is unclear how this will affect bias in the marginal estimates.

In this study, we consider \( n = 100 \) subjects, each with observations at \( m = 14 \) spatially correlated sites to mimic the structure of our dental data where each subject has 28 teeth. The fourteen spatial locations are equally spaced along a line of length one, representing the structure of one jaw. The data are generated from the nested REs model in (2.3). We consider \( p = 4 \) covariates, two of which are ‘subject level’ and two of which are ‘site level’. The subject level covariates \( X_1 \) and \( X_3 \) take the same value for all sites for subject \( i \), while \( X_2 \) and \( X_4 \) (the site level ones) vary across sites and subjects. All four covariates were drawn independently from standard normal distributions. The values for the conditional coefficients are \( \beta^T = (0.5, 0.5, 1, 1)^T \). Thus, we can simultaneously compare the sensitivity to REs distribution for subject- and site-level covariates, and large and small coefficients. The correlation matrix of the site-specific effects is taken to be exponential with spatial range \( r = 0.1 \) or 0.4. The standard deviation of the subject REs, \( \sigma^*_1 \), is fixed at 1, while the standard deviation of the site-level REs is chosen to be either \( \sigma_2 = 1 \) or 3. The total marginal shrinkage, \( \phi_1 \phi_2 \), is either 0.77 or 0.45. We consider four design settings for data generation:

- **Design 1**: Low Variance and low spatial correlation, i.e., \( \sigma_2 = 1, r = 0.1 \)
- **Design 2**: Low Variance and high spatial correlation, i.e., \( \sigma_2 = 1, r = 0.4 \)
- **Design 3**: High Variance and low spatial correlation, i.e., \( \sigma_2 = 3, r = 0.1 \)
- **Design 4**: High Variance and high spatial correlation, i.e., \( \sigma_2 = 3, r = 0.4 \)

At each design setting, we generate 100 datasets assuming the true REs distribution is either Gaussian or bridge. When generating the data from a Gaussian distribution, we directly generate \( \gamma_i \sim N(0, \sigma_1^2) \). For each dataset, we fit the logistic model with Gaussian REs and bridge REs using the Gaussian copula. We also estimate the marginal effects \( \beta^P \) with GEE, noting only a population level interpretation is possible. We compare methods in terms of relative percent
bias and root mean square error (\(\sqrt{\text{MSE}}\)), each calculated with respect to the posterior mean for \(\beta_T\), \(\beta_S\) and \(\beta_P\). For the marginal \(\beta_P\) in the Gaussian REs models, the ‘true’ value will be the approximation as in Johnson and Kotz (1970) (described in Section 2), applied twice. The model is parameterized in terms of the precision of the REs distributions, so the same priors can be used for either Gaussian or bridge REs. We assume the priors \(1/\sigma_1^2 \sim \text{gamma}(1, 0.25)\) and \(1/\sigma_2^2 \sim \text{gamma}(1, 0.25)\). We use a normal prior with variance 100 for \(\beta\), where the coefficients are independent a priori.

Tables 2.1, 2.2 and 2.3 present the effect of misspecification on marginal inference. The relative percent bias in all models ranges from 0 to \(\pm 10\%\). As expected, when the true RE distribution is Gaussian, the Gaussian fit has slightly smaller bias and \(\sqrt{\text{MSE}}\) than the bridge fit and vice versa. The relative \(\sqrt{\text{MSE}}\)s in Table 2.3 are generally closer to one in the first case where Gaussian model is true and the bridge model is misspecified, compared to the other case where the bridge model is true and the Gaussian model is misspecified. It may be that by including shape parameter \(\phi\) the bridge distribution is more robust than the Gaussian model. The largest difference we observe is for \(\beta_1\) when the bridge model is true. This is a subject-level covariate with a small coefficient; the \(\sqrt{\text{MSE}}\) is 10-25\% greater for the misspecified Gaussian model than the bridge depending on the covariance parameters. The bridge model also compares favorable to the GEE approach. Because the difference in shape of the bridge and Gaussian distribution is more pronounced as variance decreases, we expect larger differences between models for \(\sigma_2 = 1\) and this is generally the case. Frequentist coverage probability (not shown) for all coefficient estimates is close to the nominal 95\%. Across all model fits, we find that the \(\sqrt{\text{MSE}}\) for the coefficients on subject specific covariates \(X_1\) and \(X_3\) is nearly twice that of the site specific covariates \(X_2\) and \(X_4\). The size of the coefficient does not appear to have an effect.

The relatively small differences between the models should not be very surprising as both are symmetric, bell-shaped distributions, and Neuhaus et al. (1992) notes that marginal effects in particular are fairly robust with respect to different distributional assumptions for the conditional REs. In general, conditional effects are less robust compared to marginal effects, which is intuitive since more information is available for estimating marginal effects than conditional effects. It is important to carefully consider the choice of the REs distribution, but it appears little efficiency would be lost by choosing a bridge distribution for easier interpretation, even if a Gaussian distribution were the true correct choice.
Figure 2.2: Bivariate density for (a) Bridge with Gaussian copula and $\rho = 0.3$ (b) Bridge with $t$-copula, $\rho = 0.3$ and $\nu = 2$ degrees of freedom (c) Bridge with Gaussian copula and $\rho = 0.8$ and (d) Bridge with $t$-copula, $\rho = 0.8$, and $\nu = 2$. In all graphs, $\phi = .95$, for a standard deviation of $\sigma = 0.6$. 
Figure 2.3: Comparison of correlation by distance for bivariate Gaussian distribution and bridge distribution with $\phi = 0.95$ using a Gaussian and $t_2$ copula. The plots compare (a) the exponential correlation with range 0.1 and (b) the Matern correlation with range 0.1 and smoothness 1.5.
Table 2.1: Relative percent bias of population-level coefficients when the underlying true model is Gaussian and bridge. ‘G’ indicates fit of the Gaussian model, ‘B’ indicates the bridge model fit, and ‘GEE’ indicates the GEE fit. Results which show a significant difference from the bridge model are indicated with ‘*’.

### Gaussian distribution is ‘true’

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_2 = 1$</th>
<th></th>
<th></th>
<th>$\sigma_2 = 3$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r = 0.10$</td>
<td>$r = 0.40$</td>
<td></td>
<td>$r = 0.10$</td>
<td>$r = 0.40$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>B</td>
<td>GEE</td>
<td>G</td>
<td>B</td>
<td>GEE</td>
</tr>
<tr>
<td>$\beta_1^P$</td>
<td>2.13*</td>
<td>-2.92</td>
<td>0.43*</td>
<td>-3.57*</td>
<td>1.45</td>
<td>-0.34*</td>
</tr>
<tr>
<td>$\beta_2^P$</td>
<td>-3.27*</td>
<td>-6.19</td>
<td>-5.45</td>
<td>-3.56*</td>
<td>-0.48</td>
<td>-3.19</td>
</tr>
<tr>
<td>$\beta_3^P$</td>
<td>2.27*</td>
<td>-1.34</td>
<td>1.12*</td>
<td>-1.96*</td>
<td>1.97</td>
<td>0.99*</td>
</tr>
<tr>
<td>$\beta_4^P$</td>
<td>0.67</td>
<td>-2.03</td>
<td>-0.60*</td>
<td>-2.28*</td>
<td>0.54</td>
<td>-0.94*</td>
</tr>
</tbody>
</table>

### Bridge distribution is ‘true’

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_2 = 1$</th>
<th></th>
<th></th>
<th>$\sigma_2 = 3$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r = 0.10$</td>
<td>$r = 0.40$</td>
<td></td>
<td>$r = 0.10$</td>
<td>$r = 0.40$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G</td>
<td>B</td>
<td>GEE</td>
<td>G</td>
<td>B</td>
<td>GEE</td>
</tr>
<tr>
<td>$\beta_1^P$</td>
<td>4.32</td>
<td>2.80</td>
<td>3.46</td>
<td>3.82</td>
<td>1.95</td>
<td>2.94</td>
</tr>
<tr>
<td>$\beta_2^P$</td>
<td>1.10*</td>
<td>-2.74</td>
<td>-0.58</td>
<td>2.91*</td>
<td>1.51</td>
<td>0.92</td>
</tr>
<tr>
<td>$\beta_3^P$</td>
<td>3.37</td>
<td>3.77</td>
<td>2.68</td>
<td>3.43</td>
<td>2.56</td>
<td>2.63</td>
</tr>
<tr>
<td>$\beta_4^P$</td>
<td>2.97*</td>
<td>2.77</td>
<td>1.83</td>
<td>2.20*</td>
<td>0.85</td>
<td>0.90</td>
</tr>
</tbody>
</table>
Table 2.2: $\sqrt{\text{MSE}} \times 100$ when the underlying true model is Gaussian and bridge. ‘G’ indicates fit of the Gaussian model, ‘B’ indicates the bridge model fit, and ‘GEE’ indicates the GEE fit. Results which show a significant difference from the bridge model are indicated with ‘*’.

### Gaussian distribution is ‘true’

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_2 = 1$</th>
<th></th>
<th>$\sigma_2 = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r = 0.10$</td>
<td>$r = 0.40$</td>
<td>$r = 0.10$</td>
</tr>
<tr>
<td>$\beta_1^P$</td>
<td>11.68</td>
<td>11.70</td>
<td>13.66</td>
</tr>
<tr>
<td>$\beta_2^P$</td>
<td>6.34</td>
<td>6.46</td>
<td>5.89</td>
</tr>
<tr>
<td>$\beta_3^P$</td>
<td>11.13</td>
<td>11.18</td>
<td>12.08</td>
</tr>
</tbody>
</table>

### Bridge distribution is ‘true’

<table>
<thead>
<tr>
<th></th>
<th>$\sigma_2 = 1$</th>
<th></th>
<th>$\sigma_2 = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$r = 0.10$</td>
<td>$r = 0.40$</td>
<td>$r = 0.10$</td>
</tr>
<tr>
<td>$\beta_1^P$</td>
<td>11.05*</td>
<td>8.85</td>
<td>11.00*</td>
</tr>
<tr>
<td>$\beta_3^P$</td>
<td>10.55</td>
<td>11.05</td>
<td>10.57</td>
</tr>
<tr>
<td>$\beta_4^P$</td>
<td>7.75</td>
<td>8.19</td>
<td>7.58</td>
</tr>
<tr>
<td>Gaussian is true</td>
<td>Bridge is true</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_2 = 1$</td>
<td>$\sigma_2 = 3$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\beta_1^P$</td>
<td>1.00</td>
<td>1.25</td>
<td>1.12</td>
</tr>
<tr>
<td>$\beta_2^P$</td>
<td>0.98</td>
<td>1.00</td>
<td>1.02</td>
</tr>
<tr>
<td>$\beta_3^P$</td>
<td>1.00</td>
<td>0.98</td>
<td>0.97</td>
</tr>
<tr>
<td>$\beta_4^P$</td>
<td>0.96</td>
<td>0.96</td>
<td>1.02</td>
</tr>
</tbody>
</table>

### 2.5 Dental Data example

The data consist of dental records for 260 Gullah-speaking African-Americans (Bandyopadhyay et al., 2010) in South Carolina. The location and clinical attachment level (CAL, in mm) were recorded by dental hygienists for each of the 28 adult teeth (excluding wisdom teeth), for each subject. As CAL is measured at multiple sites per tooth, we define a tooth as diseased if the mean CAL $> 3$ mm, indicating moderate to severe periodontitis. The response $Y_{ij} = 1$ if tooth $j$ is missing or diseased for subject $i$, and 0 otherwise. The subjects range in age from 26 to 87, are mostly female (75%), and are from an under-served population with generally poor health outcomes. Over 97% of the subjects have at least one missing or diseased tooth, and 55% have 10 or more missing or diseased teeth. For each subject, age, body mass index (BMI), smoking status, and HbA1c (a measure of blood-glucose level), are recorded, in addition to their dental health records. The objective of our analysis is to assess the oral health of this under-studied population, and identify the effects of these covariates on the (binary) oral health status, both at the population and subject specific level.

We fit the nested model in (2.3) assuming teeth in the upper and lower jaw are independent, and within each jaw, teeth are related by a first-order Markov model (i.e., an exponential covariance). A richer model would allow correlation between jaws where neighbors between jaws and within jaws have potentially different correlation. However, initial model fits suggest the correlation between jaws is negligible after accounting for subject REs, and treating each jaw independently reduces the size of our covariance matrix by half which drastically improves the speed of computation.

We fit this model assuming a Gaussian REs distribution, a bridge REs distribution with Gaussian copula, and a bridge REs distribution with $t$-copula. Here we fix the degrees of freedom $\nu = 2$ for computational purposes because of the closed form of the $t_2$ cdf. A richer model
would allow \( \nu \) to vary; however, because the \( t \) and Gaussian copulas are similar for large \( \nu \), this will also allow for more contrast in the two approaches. For comparison, we also fit these models assuming only a subject RE by setting \( \varepsilon_{ij} = 0 \). The prior for the coefficients \( \beta^T_k \) is \( \text{iid} \sim N(0,100) \), for \( k = 0, \ldots, p \), where \( \beta^T_0 \) is the intercept term. The prior for the range parameter, \( r \sim \text{logNormal}(-2,1) \), where the distance between teeth is standardized so the maximum distance is one unit. As in the simulation study, the prior for both standard deviations are \( 1/\sigma^2_k \sim \text{gamma}(1,0.25) \). In addition to the Bayesian models, we also fit a GEE model (assuming independent subjects and an unstructured working correlation matrix), where the resulting estimates are only available at the (marginal) population level.

Results for the nested REs models appear in Tables 2.4 and 2.5. We compare these models using the Deviance Information Criterion (or DIC, Spiegelhalter et al. (2002)), popularly used in Bayesian inference, and defined as \( \text{DIC} = \bar{D} + p_D \), where \( \bar{D} \) is the posterior mean of the deviance and \( p_D \) is the effective number of parameters. Smaller values of DIC are preferred. A subject-only REs model appears to be inadequate here (DIC=6385 for bridge model) as expected, since we believe the implicit assumption of independent tooth effects is violated. The bridge model with both subject and tooth REs with a Gaussian copula has the lowest DIC of 4157. In addition to DIC, we calculated the Brier score, where we fit the model using 90% of the data randomly selected across teeth and subject, and estimate the posterior mean \( \hat{\pi}_{ij} \) for the remaining 10% (hold-out set). The Brier score (Brier, 1950) is then defined by \( \frac{1}{N} \sum_{i,j} (Y_{ij} - \hat{\pi}_{ij})^2 \), where \( N \) is the number of held out teeth (728). The Brier score ranges from 0 to 1, and smaller numbers are better. We further calculate the misclassification indices \( CV_1 = N_1^{-1} \sum_{(i,s)} \{ Y_{ij} \hat{\pi}_{ij} \} \) and \( CV_0 = N_0^{-1} \sum_{(i,s)} \{ (1 - Y_{ij}) \hat{\pi}_{ij} \} \), where \( N_1 = \sum_{i,j} Y_{ij} \), or the number of held out ones, and \( N_0 = \sum_{i,j} (1 - Y_{ij}) \), or the number of held out zeros. We note that while the Brier score, \( CV_1 \) and \( CV_0 \) are similar for all three models displayed in Table 2, the Brier score is also minimized by the bridge model with Gaussian copula. To assess the statistical significance of these subtle differences in the Brier score, we repeated this test set validation ten times on a subset of 100 subjects and found the bridge model had smaller Brier score than the Gaussian model for seven of the ten splits, and a paired t-test on the difference in Brier had p-value 0.02. Also, the Brier score is higher for the subject only models (0.151 for Bridge, and 0.150 for Gaussian) than the spatial models in Table 2 (0.132 for the bridge and 0.134 for the Gaussian), again suggesting the nested REs models produce a better fit.

The advantage of the bridge model is that we can get an exact interpretation of the coefficients at any level. In all models, the strongest predictor is the indicator for molar. For all Bayesian models, age and HbA1c are significant, while BMI and smoking are not significant in any model. Tooth specific parameters \( \beta^T \) are hard to interpret here as they represent the comparison of two observations from the same tooth for the same subject, and this type of replication is not possible in the periodontal setting. After integrating over the tooth random
effects, we are left with a subject specific model with coefficients denoted by ‘S’ in Table 2.4. These coefficients are interpretable conditioned on the subject REs. For example, in the bridge model, we see the odds of a molar being missing or diseased are $e^{2.15} = 8.58$ (95% credible set [7.17, 10.18]) times higher than other teeth for the same subject. Age is also interpretable at both the subject and population level. For a given individual, the odds of diseased/missing teeth increases 90% with each standard deviation increase in age - about 10.9 years. At the population level, however, the odds of a diseased/missing teeth increase by 58%, the attenuation due to averaging over between subject variability characterized by $\sigma_1$ in our model. It is important to emphasize that the subject- and population-level coefficients for the Gaussian REs model displayed in Table 2.4 are approximations, while in the bridge model these represent the exact log odds ratio. The GEE estimates also represent an exact marginal log odds ratio, and remain comparable to the population level estimates for both the Bayesian models. The 95% intervals for the GEE analysis are generally wider than those from bridge model with Gaussian copula, and unlike the Bayesian models the GEE analysis produces a (counterintuitive) significant protective effect of smoking on tooth loss.

Although various parameter estimates remain comparable across models, there are some notable differences. The nested REs model, the choice of RE distribution and copula has a large impact on the estimates. This is not surprising since different choices of REs/copula yield different scales, on which regression effects are estimated. For example, the population level estimate of the odds ratio corresponding to gender is $e^{-0.39} = 0.68$ (0.49, 0.88) for the Gaussian model, as compared to 0.57 (0.40, 0.76) for the bridge model with Gaussian copula. It should be noted that the impact is mainly on the magnitude of effect estimates, though the inference remains consistent across all three models, i.e., Age, Sex, HbA1c and the molar effect remain significant. Because of the relatively high degree of variation in the subject and tooth REs, there remain substantial differences in tooth-, subject-, and population-level interpretations of the coefficients. For example the estimated odds ratio for gender in the bridge model with Gaussian copula is $e^{\beta_T} = 0.15$ at the tooth level, 0.45 at the subject level, and 0.57 at the population level. This difference in the conditional and marginal interpretations highlights the importance of correctly choosing conditional or marginal inference. There are also differences in the covariance parameter estimates across models. The between tooth variability is estimated much lower for the t-copula model than the Gaussian copula model, resulting in a lesser degree of marginal shrinkage. For the subject-and-site models, note that $\sigma_1 = \sigma_1^*/\phi_2$ displayed in Table 2.5 is the total standard deviation of the subject REs $\gamma_i$, which depends on both $\phi_1$ and $\phi_2$. This is much greater than the standard deviation found in the subject only model. The range parameter is about 0.15, regardless of copula, suggesting a correlation of 0.6 between adjacent teeth.

In the Bayesian setting, we can also easily estimate the REs for each subject and tooth.
This may be of particular interest in epidemiological settings, where we may, for instance, want to assess disease risk for certain states or counties, or look for areas with unusually high or low risk. To demonstrate this, Figure 2.4 display the estimated REs and posterior probabilities of a missing or diseased tooth at various tooth locations for an arbitrary subject. The posterior probabilities tend to be higher where teeth are missing.
Table 2.4: Posterior parameter estimates and 95% credible intervals (C.I.) for tooth specific (T), subject specific (S) and population level (P) fixed-effects for the model with subject and tooth nested random effects. * indicates 95% C.I. that excludes 0. For the $t$-copula, there are $\nu = 2$ degrees of freedom.

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Bridge: Gauss. copula</th>
<th>Bridge: $t$-copula</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIC</td>
<td>5283</td>
<td>4157</td>
<td>5490</td>
<td></td>
</tr>
<tr>
<td>pD</td>
<td>2578</td>
<td>1931</td>
<td>1692</td>
<td></td>
</tr>
<tr>
<td>Brier Score</td>
<td>0.134</td>
<td>0.132</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td>CV$_0$, CV$_1$</td>
<td>0.226, 0.670</td>
<td>0.225, 0.670</td>
<td>0.223, 0.677</td>
<td></td>
</tr>
<tr>
<td>Int</td>
<td>-1.61 (-2.54,-0.66)</td>
<td>-1.64 (-2.72,-0.60)</td>
<td>-1.42 (-2.18,-0.58)</td>
<td></td>
</tr>
<tr>
<td>T:Age</td>
<td>1.17 (0.76, 1.61)*</td>
<td>1.48 (0.84, 2.23)*</td>
<td>1.14 (0.73, 1.59)*</td>
<td></td>
</tr>
<tr>
<td>T:Sex</td>
<td>-1.21 (-2.09,-0.45)*</td>
<td>-1.89 (-3.65,-0.80)*</td>
<td>-1.28 (-2.13,-0.60)*</td>
<td></td>
</tr>
<tr>
<td>T: BMI</td>
<td>-0.06 (-0.45, 0.29)</td>
<td>0.06 (-0.35, 0.57)</td>
<td>0.04 (-0.31, 0.39)</td>
<td></td>
</tr>
<tr>
<td>T: Smoking</td>
<td>0.38 (-0.08, 0.79)</td>
<td>0.37 (-0.14, 1.08)</td>
<td>0.32 (-0.05, 0.68)</td>
<td></td>
</tr>
<tr>
<td>T: HbA1c</td>
<td>0.42 (0.08, 0.88)*</td>
<td>0.48 (0.05, 0.97)*</td>
<td>0.39 (0.12, 0.71)*</td>
<td></td>
</tr>
<tr>
<td>T: Molar</td>
<td>4.12 (3.62, 4.75)*</td>
<td>4.96 (3.68, 7.37)*</td>
<td>3.78 (3.32, 4.30)*</td>
<td></td>
</tr>
<tr>
<td>S: Age</td>
<td>0.68 (0.43, 0.94)*</td>
<td>0.64 (0.42, 0.84)*</td>
<td>0.91 (0.59, 1.21)*</td>
<td></td>
</tr>
<tr>
<td>S: Sex</td>
<td>-0.71 (-1.26,-0.25)*</td>
<td>-0.80 (-1.26,-0.37)*</td>
<td>-1.17 (-1.71,-0.65)*</td>
<td></td>
</tr>
<tr>
<td>S: BMI</td>
<td>-0.04 (-0.25, 0.17)</td>
<td>0.02 (-0.17, 0.22)</td>
<td>0.02 (-0.25, 0.29)</td>
<td></td>
</tr>
<tr>
<td>S: Smoking</td>
<td>0.23 (-0.04, 0.49)</td>
<td>0.16 (-0.06, 0.48)</td>
<td>0.13 (-0.17, 0.42)</td>
<td></td>
</tr>
<tr>
<td>S: HbA1c</td>
<td>0.24 (0.05, 0.50)*</td>
<td>0.21 (0.02, 0.39)*</td>
<td>0.34 (0.08, 0.60)*</td>
<td></td>
</tr>
<tr>
<td>S: Molar</td>
<td>2.40 (2.23, 2.57)*</td>
<td>2.15 (1.97, 2.32)*</td>
<td>2.78 (2.52, 3.07)*</td>
<td></td>
</tr>
<tr>
<td>P: Age</td>
<td>0.37 (0.22, 0.53)*</td>
<td>0.46 (0.29, 0.60)*</td>
<td>0.57 (0.37, 0.76)*</td>
<td>0.50 (0.33, 0.66)*</td>
</tr>
<tr>
<td>P: Sex</td>
<td>-0.39 (-0.71,-0.13)*</td>
<td>-0.57 (-0.92,-0.27)*</td>
<td>-0.74 (-1.07,-0.41)*</td>
<td>-0.52 (-0.92,-0.12)*</td>
</tr>
<tr>
<td>P: BMI</td>
<td>-0.02 (-0.14, 0.10)</td>
<td>0.01 (-0.12, 0.16)</td>
<td>0.01 (-0.16, 0.18)</td>
<td>-0.04 (-0.25, 0.16)</td>
</tr>
<tr>
<td>P: Smoking</td>
<td>0.12 (-0.02, 0.28)</td>
<td>0.12 (-0.04, 0.33)</td>
<td>0.08 (-0.11, 0.27)</td>
<td>-0.43 (-0.80, -0.06)*</td>
</tr>
<tr>
<td>P: HbA1c</td>
<td>0.13 (0.03, 0.27)*</td>
<td>0.15 (0.02, 0.28)*</td>
<td>0.21 (0.05, 0.38)*</td>
<td>0.19 (0.04, 0.33)*</td>
</tr>
<tr>
<td>P: Molar</td>
<td>1.31 (1.15, 1.46)*</td>
<td>1.54 (1.40, 1.68)*</td>
<td>1.76 (1.58, 1.94)*</td>
<td>1.73 (1.58, 1.88)*</td>
</tr>
</tbody>
</table>
Table 2.5: Variance components, $\phi_1$, $\phi_2$ and $r$ for the model with subject and tooth nested random effects. * indicates 95% C.I. that excludes 0. $\phi_1$ controls the distribution of $\gamma_i^*$, while $\sigma_1$ is the total subject-specific effect standard deviation given by $\text{sd}(\gamma_i^*)/\phi_2$. For the $t$-copula, there are $\nu = 2$ degrees of freedom.

<table>
<thead>
<tr>
<th></th>
<th>Gaussian</th>
<th>Bridge: Gauss. copula</th>
<th>Bridge: $t$-copula</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi_1$</td>
<td>0.54 (0.48, 0.61)</td>
<td>0.72 (0.67, 0.76)</td>
<td>0.63 (0.57, 0.69)</td>
</tr>
<tr>
<td>$\phi_2$</td>
<td>0.58 (0.51, 0.65)</td>
<td>0.45 (0.29, 0.59)</td>
<td>0.63 (0.44, 0.78)</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>5.14 (4.02, 6.72)</td>
<td>4.07 (2.87, 6.17)</td>
<td>3.65 (2.70, 5.13)</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>2.69 (2.26, 3.25)</td>
<td>3.76 (2.49, 5.94)</td>
<td>2.33 (1.46, 3.71)</td>
</tr>
<tr>
<td>$r$</td>
<td>0.15 (0.12, 0.18)</td>
<td>0.15 (0.10, 0.18)</td>
<td>0.15 (0.11, 0.21)</td>
</tr>
</tbody>
</table>

2.6 Conclusion

In this chapter, we extend the bridge distribution to the spatial setting using a copula. Either a Gaussian or $t$-copula model can be implemented in standard software such as R or OpenBUGS. From our simulation study, we see that under REs misspecification, utilizing a bridge model did not result in a dramatic loss of efficiency. While comparing the fit to the dental dataset (see Table 2 in Section 5), the cross-validated measures (Brier score, $CV_0$, $CV_1$) are very close, although the bridge model (with the Gaussian copula) resulted in a lower DIC as compared to the Gaussian model. In any model selection problem the goodness of fit of the REs density should always remain the primary criterion in identifying the appropriate model to be utilized. However, (as in our dental data example) when both models produce similar fit, utilizing the bridge model allows for easy interpretability of regression coefficients at any level of a hierarchical model and the ability to estimate REs. In practice, there is often not enough data information to assess the distributional assumption of the (conditional) REs. Therefore, the bridge distribution may be viewed as a ‘vehicle’ to assess and compare multilevel effects, and variations across data levels may reveal insights on level-specific heterogeneity. The information is often practically important since the source of variation is often of great interest.

A similar copula framework could be used to extend the other bridge specifications, such as for the complimentary-log-log link function for binary data (Wang and Louis, 2003), or other bridge-like distributions such as the positive stable distribution for failure time models (Hougaard, 1986), to a spatial setting. Generalizations to other copulas are also readily available. The Gaussian copula has been criticized in certain modeling applications, particularly when tail behavior is important. The copula implies symmetry of correlation - high-high combinations are as likely as low-low combinations. More concerning, extreme values are essentially independent.
Figure 2.4: Posterior estimates of tooth level random effects $\varepsilon_{ij}$ with central 90% credible intervals (a), and posterior probability of a missing or diseased tooth at various tooth locations (b) of an arbitrarily selected subject, obtained from the bridge model with Gaussian copula. The (dark) squares in the lower portion of each figure represent teeth that are missing or diseased.
of each other, which for binary data may impede adequately modeling large clusters of zeros or ones. There is no readily apparent way to extend the approach to models with random slope coefficients, as Wang and Louis (2003) point out for the univariate bridge distribution. It would be of interest to further explore the robustness of the bridge distribution to misspecified correlation structures or missing covariates, as well as choice of copula.

Finally, we stress that although our model was developed in the highly-structured dental data setting with independent subjects and a regular grid of locations, the spatial bridge model is widely-applicable to other kinds of spatial binary data. For example, a common setup in ecology is to measure the presence or absence of a species at many spatial locations. Two common objectives are to produce a spatial map of prevalence throughout the region of interest, and to estimate the effects of spatial covariates such as distance to a roadway or land-use on the presence of the species. These objectives can be addressed easily and simultaneously using the spatial bridge model, by interpolating the spatial random effects $\varepsilon$ using standard Bayesian Kriging methods, and studying the posterior of the population parameters $\beta^p$. 

-
Chapter 3

Hierarchical Variable Selection using Effect Modifiers

3.1 Introduction

Particulate matter (PM) is a complex mixture of airborne particles from both primary and secondary anthropogenic and natural sources. Major sources include combustion of fossil fuels and biomass, dust from industrial, construction, and mining operations, wildfires, and lightning strikes (Schlesinger et al., 2006). Particles are either emitted directly into the air (primary pollution) or formed by chemical interactions of gases and primary pollutant particles in the air (secondary pollution). Ambient levels of this complex mixture are currently regulated by the Environmental Protection Agency (EPA) based solely on particle size, not chemical composition or source. Yearly average and hourly maximum levels of the total amount of both coarse (PM10; particles < 10µm in aerodynamic diameter) and fine (PM2.5; particles < 2.5µm) are restricted.

Numerous studies have quantified the relationship of fine particulate matter exposure and human health using multi-site time series studies (Dominici et al., 2000, 2006; Choi et al., 2009) comparing ambient measures of particulate matter to hospital admissions and mortality data at the county and city level. Both spatial and temporal heterogeneity in health effect estimates have been observed (Bell et al., 2007, 2008; Zhou et al., 2011; Ito et al., 2011). Variation in effect between sites is often partially explained using local health and exposure characteristics but evidence from both epidemiological and laboratory studies suggests that chemical components of PM2.5 vary in toxicity and the observed variance in effect may be due to differing chemical composition across space and time (Fuentes et al., 2006; Schlesinger et al., 2006; Bell et al., 2007; Peng et al., 2009; Bell et al., 2009; Levy et al., 2012). While PM2.5 is currently regulated by total mass, deeper understanding could yield more targeted regulation by component or sources.
A growing body of literature has investigated the potential health impact of specific PM2.5 components. While carbon fractions, including elemental carbon, black carbon, and organic carbon matter, are shown to have positive associations in a variety of health studies, no components have been ruled out in all epidemiological and toxicological studies (Rohr and Wyzga, 2012). Studying individual chemical components does not eliminate between site variability in the health effect estimates. There is some variability in particle size within component depending on the source (Schlesinger, 2007), and also variability in particle chemistry and acidity, as these components may be part of larger molecules (Schlesinger et al., 2006). Site-specific effects of individual components may also vary due to population health characteristics which modify susceptibility or by lifestyle or housing characteristics which modify exposure to ambient air. Dominici et al. (2002, 2003) investigate this possibility using socioeconomic and demographic information as effect modifiers in the relationship of mortality and PM10. Zeka et al. (2005) show city characteristics such as population density and average winter temperature are important effect modifiers in PM10 and cause-specific mortality relationships.

Previous epidemiological studies investigating components of fine particulate matter tend to focus on a few pollutants or pollutant groups. Peng et al. (2009) investigate the relationship of the seven most massive PM2.5 components and cardiovascular disease (CVD) hospitalizations. They use a hierarchical model to estimate national effects, but assume the effects at every site are independent. The county specific effects and their statistical significance vary across the country. Some city-specific analyses investigate a larger number of pollutants (Ito et al., 2011; Zhou et al., 2011), but these results are hard to generalize to other locations, and different pollutants are selected as most significant. To our knowledge, a model which accomplishes variable selection across all sites within one model has not previously been attempted for these data. In this chapter, we propose a hierarchical Bayesian model for the probability of a non-null effect in terms of county-specific features such as racial distribution, housing and commuting characteristics, and socioeconomic status.

Using one hierarchical model allows us to pool information across sites, which is especially valuable when effect estimates are very small. Using variable selection techniques is an additional improvement over traditional analysis, as we are able to include a much larger list of pollutants without overfitting the model. By shrinking less significant predictors toward zero, we improve the stability of the estimates of the other pollutant coefficients. In our model, we allow for a different set of pollutants to be selected at every site, while assuming a hierarchical model on both the inclusion probability and effect estimate magnitude. This work is an extension of Stochastic Search Variable Selection (George and McCulloch, 1993; George and McCulloch, 1997) described in Chapter 1.

We implement both an exchangeable variable selection model and our variable selection model with effect-modifiers using a two-stage approximation to the fully Bayesian model that
is computationally efficient. Though a two-stage model is commonly used in air pollution studies (Dominici et al. (2000) and Peng et al. (2009) among others) it is the first time to our knowledge it has been used with variable selection.

3.2 Data

Of the approximately 50 speciated PM2.5 components measured by the EPA, we selected $p = 22$ components of interest. Each contributes at least 1% of total mass to PM2.5, or the literature has suggested a potential link with health outcomes, or both. The components and summary statistics are shown in Table 3.1. We include Figure 3.1 to demonstrate the variability in components over time and location. These speciated PM measurements are taken from the EPA’s Air Quality System (AQS) and AirExplorer databases (www.epa.gov/ttn/airs/airsaqs/, www.epa.gov/airexplorer/). The AQS data include raw monitor values and daily averages, while AirExplorer is a processed data product designed for use by health and epidemiology research. For twenty of the components we use the AQS data. Because of a high proportion of missingness for elemental carbon (EC) and organic carbon matter (OCM) in the AQS database, we use the AirExplorer data for these components. Following Peng et al. (2009), for counties that had more than one active monitor on a given day, an average was taken using 10% trimmed mean if more than 10 stations; for 3–10 stations, minimum and maximum values were excluded from the mean; and for 2 stations, we use the mean. All components are measured in $\mu g/m^3$ except EC, which is measured in inverse megameters, a measure of light extinction in haze.

We only used information from non-source-oriented monitors, and exclude values flagged by the EPA for data quality issues. Source-oriented monitors are placed with the intention of monitoring a known large pollutant source, and may not be representative of population exposure. To avoid biased pollution measurements, we exclude these and focus on non-source-oriented monitors, which are placed with the purpose of estimating the exposure in populated areas. We include 117 counties in the US with at least 100,000 residents and PM2.5 components monitors active on at least 150 days in the time period 2000-2008. Of these we exclude two California counties with data quality issues. In these counties half the pollution measurements were 1000 times larger than expected based on nearby counties and measurements on preceding days. In total we include 115 counties as shown in Figure 3.2.

As recommended by Gelfand et al. (2003) covariates are scaled, but not centered. We use the 90th percentile value to scale rather than standard deviation because of the skewness of the pollution data. We also remove extreme pollution values. We define a value as extreme if a measurement is more than double the second highest value for that pollutant and site. This results in the removal of 293 observations, or just less than 0.5% of the observed data values. This is the same criterion for extremeness used by Peng et al. (2009) for their study of the seven
Figure 3.1: Observed level of elemental carbon over the period 2002-2008 in 115 counties. Green indicates highest values, purple the lowest. At right are the median, min, and max values, and the daily averages are plotted along the bottom.
major components.

The health data includes Medicare beneficiary enrollment and Medicare Part-A inpatient records, aggregated to daily county totals. We count the number of patients hospitalized with a principal ICD-9 diagnosis code related to cardiovascular disease (CVD). These include heart failure, ischemic heart disease, and cerebrovascular events among others. Table 3.2 and Figure 3.2 show the average CVD hospitalization rates vary greatly by county, ranging from 8.5 to 32 per 100,000 Medicare enrollees per day.

Table 3.1: Median, interquartile range (IQR), and maximum observed value in \(\mu g/m^3\) across all 115 sites in the period 2000-2008. Minimum values all approximately zero, and are not shown. The seven most massive components are listed first, with the rest in alphabetical order.

<table>
<thead>
<tr>
<th>Element</th>
<th>Median</th>
<th>IQR</th>
<th>Max</th>
<th>Element</th>
<th>Median</th>
<th>IQR</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfate</td>
<td>2.43</td>
<td>2.89</td>
<td>39.90</td>
<td>Chlorine</td>
<td>0.00</td>
<td>0.02</td>
<td>5.78</td>
</tr>
<tr>
<td>Nitrate</td>
<td>0.83</td>
<td>1.54</td>
<td>45.30</td>
<td>Chromium</td>
<td>0.00</td>
<td>0.00</td>
<td>1.17</td>
</tr>
<tr>
<td>Silicon</td>
<td>0.07</td>
<td>0.10</td>
<td>10.10</td>
<td>Copper</td>
<td>0.00</td>
<td>0.00</td>
<td>0.55</td>
</tr>
<tr>
<td>Elemental Carbon</td>
<td>0.54</td>
<td>0.52</td>
<td>11.00</td>
<td>Iron</td>
<td>0.07</td>
<td>0.08</td>
<td>24.10</td>
</tr>
<tr>
<td>Organic Carbon</td>
<td>3.30</td>
<td>3.42</td>
<td>101.18</td>
<td>Lead</td>
<td>0.00</td>
<td>0.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Sodium Ion</td>
<td>0.09</td>
<td>0.13</td>
<td>33.40</td>
<td>Magnesium</td>
<td>0.00</td>
<td>0.01</td>
<td>2.72</td>
</tr>
<tr>
<td>Ammonium Ion</td>
<td>1.12</td>
<td>1.43</td>
<td>21.70</td>
<td>Nickel</td>
<td>0.00</td>
<td>0.00</td>
<td>0.91</td>
</tr>
<tr>
<td>Aluminum</td>
<td>0.01</td>
<td>0.04</td>
<td>4.01</td>
<td>Potassium</td>
<td>0.05</td>
<td>0.05</td>
<td>12.80</td>
</tr>
<tr>
<td>Arsenic</td>
<td>0.00</td>
<td>0.00</td>
<td>0.16</td>
<td>Titanium</td>
<td>0.00</td>
<td>0.01</td>
<td>0.56</td>
</tr>
<tr>
<td>Bromine</td>
<td>0.00</td>
<td>0.00</td>
<td>0.19</td>
<td>Vanadium</td>
<td>0.00</td>
<td>0.00</td>
<td>0.17</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.04</td>
<td>0.05</td>
<td>4.90</td>
<td>Zinc</td>
<td>0.01</td>
<td>0.01</td>
<td>3.11</td>
</tr>
</tbody>
</table>

In addition to these time-varying data, we incorporate county-specific covariates that are used to model the probability a pollutant component is included in the model at a given county. These covariates are measured only once per site, and hence are not considered time dependent. These include county demographic, economic, and housing information from the 2000 census (factfinder2.census.gov). These variables and their descriptions are presented in Table 3.3. The spatial variation in these coefficients are shown in the appendix. We indicate the vector of these effect modifiers at site \(s\) as \(W^s = (W^s_0, W^s_1, \ldots, W^s_m)'\), where \(1, \ldots, m\) index effect modifiers and \(W^s_0 = 1\) to include an intercept term. All effect modifiers are centered at their mean and scaled by standard deviation.
Table 3.2: Minimum, median, and maximum for number of enrollees on a given day, number of CVD hospitalizations on a given day, rate per 100,000 on a given day, and county average of daily rate per 100,000 from 2000-2008.

<table>
<thead>
<tr>
<th></th>
<th>Enrollees</th>
<th>CVD hospitalizations</th>
<th>Daily Rate (per 100,000)</th>
<th>County Avg. Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>min</td>
<td>10,227</td>
<td>0</td>
<td>0</td>
<td>8.5</td>
</tr>
<tr>
<td>median</td>
<td>51,593</td>
<td>9</td>
<td>17</td>
<td>18</td>
</tr>
<tr>
<td>max</td>
<td>601,716</td>
<td>167</td>
<td>98</td>
<td>32</td>
</tr>
</tbody>
</table>

Table 3.3: Effect modifiers included in the VS model.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Hispanic</td>
<td>Percent of total population which is Hispanic</td>
</tr>
<tr>
<td>% Black</td>
<td>Percent of total population which is Black</td>
</tr>
<tr>
<td>% Poverty 65+</td>
<td>Percent of population over 65 below the poverty level</td>
</tr>
<tr>
<td>% Commute by Public Transit</td>
<td>Percent of population over 18 commuting by public transit</td>
</tr>
<tr>
<td>% Urban</td>
<td>Percent of total population living in an urban center</td>
</tr>
<tr>
<td>Year Housing</td>
<td>Median year built of all occupied housing units</td>
</tr>
</tbody>
</table>
Figure 3.2: Average CVD hospitalizations per day per 100,000 enrollees by county. Inset shows Northeast US in greater detail.

3.3 Model

We first introduce a variable selection model in which the prior probability of inclusion is constant across space. To differentiate this from our new model, and to maintain consistency with Chapter 4, we call this the exchangeable variable selection (EVS) model. We then introduce covariates which modify the probability of inclusion. This model we call “VS with effect modifiers.”

3.3.1 Exchangeable Variable Selection

We observe \( Y_{st} \) CVD cases at county \( s \) on day \( t \), in addition to a vector of \( p = 22 \) pollutants, \( x_{st} \), and confounding variables \( z_{st} \). The confounding variables include indicator functions for the day of the week and smooth functions of time, temperature, and dewpoint. To adjust for long term and seasonal trends, we include natural cubic spline smoothing functions of time with 7 degrees of freedom (df) per year as a predictor in the model. We also use splines for daily average temperature (6 df) and the 3-day lag mean dewpoint (3 df) to control for confounding with weather. We chose these numbers of degrees of freedom for consistency with Dominici et al. (2002). We discuss the sensitivity of the choice of degrees of freedom for these data in Chapter
4. We utilize the spatially varying coefficients model described by Gelfand et al. (2003),
\[
Y^s_t | \beta^s, \eta^s \sim \text{Poisson} \left[ N^s_t \exp \{ \beta^s_0 + x^s_t \beta^s + z^s_t \eta^s \} \right]
\]  \hspace{1cm} (3.1)
where \(N^s_t\) is a known offset for the number of Medicare patients at county \(s\) on day \(t\) and \(\beta^s_0\) is a spatially varying intercept. We include varying coefficients \(\beta^s = (\beta^s_1, \ldots, \beta^s_p)'\) and \(\eta^s = (\eta^s_1, \ldots, \eta^s_{q_s})'\). Note that the dimension of \(\eta^s\), \(q_s\), depends on the county due to the use of 7 knots per year for the spline function of time, and that each county may be observed over a different time interval. We control for unmeasured confounding independently within each location and then borrow information across locations for the health estimates.

We assume the pollutant coefficients of interest are independent and follow the mixture distribution
\[
\beta^s_k \sim \begin{cases} 
N (\alpha_k, \omega^2_k) & \text{if } \gamma^s_k = 1 \\
N (0, \omega^2_k/C) & \text{if } \gamma^s_k = 0
\end{cases}
\]  \hspace{1cm} (3.2)
for each covariate \(k\) and county \(s\). That is, when the binary random variable \(\gamma^s_k = 1\), the covariate \(x_k\) is “in” the model at site \(s\). For variables included in the model (\(\gamma^s_k = 1\)), the effects have a normal prior centered on \(\alpha_k\) with standard deviation \(\omega_k\). The covariate \(x_k\) is out of the model at site \(s\) if \(\gamma^s_k = 0\). In this case, the prior is centered on zero, with small standard deviation, \(\omega_k/\sqrt{C}\). This is similar to the spike and slab prior of SSVS introduced in Chapter 1, but extended to include selection across both covariates and sites.

In the EVS model, \(\gamma^s_k \sim \text{Bernoulli}(\pi_k)\), so that there is a common prior inclusion probability across all sites. That is, for sites with covariate \(x_k\) included in the model, the coefficients \(\beta^s_k\) are centered around a common mean \(\alpha_k\) rather than zero. This promotes sharing of information across sites. Here we assume that \(\beta^s_k\) are conditionally independent across sites, and we discuss the extension to spatially correlated \(\beta^s_k\) in Chapter 4. We refer to this model as exchangeable rather than independent because with unknown parameters \(\alpha_k, \pi_k\) and \(\omega_k\), the sites share information in our hierarchical Bayesian model.

### 3.3.2 Exchangeable Variable Selection with Effect Modifiers

Our contribution is to model the probability of inclusion using covariates, thus adding another level to the hierarchical model, and allowing the probability of inclusion, \(\pi_k\) to vary by site. Denoting \(\Phi\) as the standard normal cdf, \(P(\gamma^s_k = 1) = \pi^s_k\) where \(\pi^s_k\) is a function of covariates, \(\Phi^{-1}(\pi^s_k) = W^s \xi_k\). The pollutant specific \(m + 1\)-vector of coefficients \(\xi_k = (\xi_{k0}, \xi_{k1}, \ldots, \xi_{km})'\) are used to model these site specific inclusion probabilities.

We let \(\xi_{kj} \sim N(\mu_j, \sigma_j^2)\) for \(k = 1, \ldots, p\), imposing some shrinkage toward a common effect modification for each covariate. Intuitively, we assume that covariates such as poverty would
have a similar effect on health effects of all pollutants; if the effects vary considerably across pollutants, our estimate for \( \sigma_j \) will be high and this case can also be handle by our model.

Because we center and scale the covariates in \( W \), the intercept term \( \xi_{k0} \) has the interpretation that \( \Phi(\xi_{k0}) \) is the probability of including covariate \( k \) at a site with average levels of covariates \( W \). We expect that this intercept is very different across pollutants, and therefore fix the prior \( \mu_0 = 0 \), and \( \sigma_0^2 \) as a large variance. We also consider a model in which \( \mu_j = 0 \) and \( \sigma_j \) is large for all \( j = 1, \ldots, m \), which implies no sharing of information about the effect of covariate \( j \) across pollutants. We refer to the first model with shared mean as “VS with effect modifiers”, and the second as the “unsmoothed VS with effect modifiers” model.

### 3.4 Methods

Computational savings are created by using an approximation to the first stage of the model described in (3.1). Following Daniels and Kass (1998), an approximation to this likelihood, based on the conditional posterior of \( \beta^s_k \), is

\[
\hat{\beta}^s \sim N(\beta^s, \hat{V}^s)
\]  

(3.3)

where \( \hat{\beta}^s \) is the maximum likelihood estimate of the \( p \)-vector of site-specific coefficients \( \beta^s \) in the Poisson regression model in (3.1) and \( \hat{V}^s \) is the \( p \times p \) estimated covariance matrix of \( \hat{\beta}^s \) calculated as the inverse of the information matrix. Similar two-stage models have been implemented for studying air pollution data across sites (Dominici et al., 2000; Berger et al., 2001; Dominici et al., 2002, 2006; Peng et al., 2009). This is however the first time to our knowledge such an approximation has been used in a variable selection setting.

By using the two stage model, we are able to do all computation using Gibbs’ sampling. The approximated model and priors for the two-stage EVS model without covariates are as follows:

\[
\hat{\beta}^s \sim N(\beta^s, \hat{V}^s) \text{ for } s \text{ in } 1, \ldots, n \\
\beta_k^s \sim \gamma_k^s N(\alpha_k, \omega_k^2) + (1 - \gamma_k^s) N(0, \omega_k^2/C) \\
\gamma_k^s \sim \text{Bernoulli}(\pi_k^s) \\
\alpha_k \sim N(0, \tau^2) \\
1/\omega^2 \sim \text{Gamma}(a_\omega, b_\omega)
\]

For the EVS model without covariates, \( \pi_k^s = \pi_k \), and \( \pi_k \sim \text{beta}(a_\pi, b_\pi) \). In this study, we let \( a_\pi = b_\pi = 1 \), to give a uniform distribution on \( (0, 1) \). The VS with effect modifiers model sets \( \pi_k^s = \Phi(W^s \xi_k) \). Both models exhibit conjugate full-conditional distributions resulting in simple closed form updates in Gibbs’ sampling.
3.4.1 Full Conditionals for the EVS Model

The update for the vector of county-specific coefficients \( \beta^s \) is

\[
\beta^s | \gamma^s, \alpha, \omega \sim N_p \left( [\hat{V}^{s-1} + D_1^{-1}]^{-1} [\hat{V}^{s-1} \beta^s + D_2^{-1} \alpha], [\hat{V}^{s-1} + D_1^{-1}]^{-1} \right)
\]

where \( D_1 \) and \( D_2 \) are diagonal matrices, with diagonal elements \( d_{1kk} = \gamma_k^s \omega_k^2 + (1 - \gamma_k^s)\omega_k^2/C \) for \( D_1 \) and \( d_{2kk} = \gamma_k^s \omega_k^2 \) for \( D_2 \). The prior parameters \( \alpha_k \) and \( \omega_k \) are updated as follows.

\[
1/\omega_k^2 | \beta_k, \gamma_k, \alpha_k \sim \text{Gamma} \left\{ a, b + 1 + \frac{1}{2} \sum_{s=1}^{n} \gamma_k^s (\beta_k^s - \alpha_k)^2 + (1 - \gamma_k^s)C \beta_k^{2s} \right\} \\
\alpha_k | \beta_k, \omega_k, \tau \sim N \left\{ \frac{\sum_{s=1}^{n} \gamma_k^s}{\omega_k}, \frac{1}{\tau^2} \right\}^{-1} \sum_{s=1}^{n} \gamma_k^s \beta_k^s, \left( \frac{\sum_{s=1}^{n} \gamma_k^s}{\omega_k} + \frac{1}{\tau^2} \right)^{-1} \\
\]

Finally, in the EVS model, \( \gamma_k^s \) and \( \pi_k \) also have full conditional updates as written below, where \( \phi(x) \) is the standard normal pdf,

\[
\gamma_k^s | \beta_k^s, \pi_k, \alpha_k, \omega_k \sim \text{Bernoulli} \left\{ \frac{\pi_k \phi \left( \frac{\beta_k^s - \alpha_k}{\omega_k} \right)}{\pi_k \phi \left( \frac{\beta_k^s - \alpha_k}{\omega_k} \right) + (1 - \pi_k) C \phi \left( \frac{\beta_k^s}{\omega_k} \right)} \right\} \\
\pi_k | \gamma_k \sim \text{Beta} \left\{ 1 + \sum_{s=1}^{n} \gamma_k^s, 1 + \sum_{s=1}^{n} (1 - \gamma_k^s) \right\}.
\]

3.4.2 Full Conditionals for VS with Effect Modifiers

For VS with effect modifiers, we define \( \pi_k^s = \Phi(W^s\gamma_k) \). The model remains (conditionally) conjugate through the introduction of auxiliary Gaussian random variables \( z_k^s \). We define

\[
z_k^s \sim N(W^s \gamma_k, 1) \\
\gamma_k^s = I(z_k > 0).
\]

Thus \( \pi_k^s = P(\gamma_k^s = 1) = P(z_k^s > 0) = \Phi(W^s \gamma_k) \). When \( \gamma_k \) has a Gaussian prior, we can easily see that the conditional distribution \( \gamma_k | z_k \) is conjugate, and is also a normal distribution,

\[
\gamma_k | z_k \sim N \left( \left[ W^s W + \text{diag}(\sigma^2) \right]^{-1} [W^s z_k + \text{diag}(\sigma^2)^{-1} \mu], 
\left[ W^s W + \text{diag}(\sigma^2)^{-1} \right]^{-1} \right),
\]

where \( \text{diag}(\sigma^2) \) is a diagonal matrix with the vector \( \sigma \) on the diagonal. The Gibbs’ sampling update for \( z_k^s \) is a truncated normal distribution. If we denote \( N_{tr}(\mu, \sigma^2, a, b) \), as the truncated normal distribution with mean \( \mu \), variance \( \sigma^2 \), and limits \( a \) and \( b \), \( z_k^s | \gamma_k^s = 1 \sim N_{tr}(W^s \gamma_k, 1, 0, \infty) \) and similarly, \( z_k^s | \gamma_k^s = 0 \sim N_{tr}(W^s \gamma_k, 1, -\infty, 0) \). That is, \( z_k^s \) is re-
stricted to greater than 0 when $\gamma^s_k = 1$, and is restricted to less than 0 when $\gamma^s_k = 0$. This auxiliary variable method (Albert and Chib, 1993) is commonly used in probit regression in a Bayesian setting. In that case however, the binary responses are fixed and do not need to be updated. As $\gamma^s_k|z^s_k$ is clearly fixed, we update $\gamma^s_k$ by considering the joint distribution $\gamma^s_k, z^s_k|\xi_k$, and marginalize over $z^s_k$ to find the distribution of $\gamma^s_k|\xi_k$ directly. The Gibbs’ update is then the conjugate conditional distribution

$$
\gamma^s_k|\xi, \beta^s_k, \alpha_k, \omega_k \sim \text{Bernoulli}
\left\{ \frac{\Phi \left( W^{st}\xi_k \phi \left( \frac{\beta_k^s - \alpha_k}{\omega_k} \right) \right)}{\Phi \left( W^{st}\xi_k \phi \left( \frac{\beta_k^s - \alpha_k}{\omega_k} \right) \right) + \left[ 1 - \Phi \left( W^{st}\xi_k \right) \right] C \phi \left( \frac{\beta_k^* C}{\omega_k} \right)} \right\}.
$$

As the updates for $\beta^s_k, \alpha_k$, and $\omega_k$ depend on $\gamma^s_k$, the full conditionals for these parameters are the same in the VS with effect modifiers model as in the EVS model. In the appendix, we include a simulated example which demonstrates our computational algorithm.

### 3.5 Analysis of PM2.5 Component Data

#### 3.5.1 Models and Data

We fit the EVS, VS with effect modifiers, and unsmoothed VS with effects modifiers models to the CVD and pollution data. The CVD responses are modeled using a Poisson regression model as described in Section 3.3.

The prior for $\alpha_k$ is $N(0, 0.1^2)$, and the prior for $1/\omega^2_k$ is $\text{Gamma}(0.1, 0.001)$. The prior variance on $\alpha_k$ allows for up to 10% relative risk increases, while the prior for $1/\omega^2_k$ puts 45% of the mass on $\omega_k < 0.01$. A beta(1, 1) prior on $\pi_k$ is used for the ordinary EVS model. In the VS effect modification model, we fix $\sigma_0 = 1$, with a gamma(0.001, 0.001) prior on $\sigma_j = 1, \ldots, m$, and let $\mu_j \sim N(0, 3)$. In the unsmoothed VS model, we let fix $\mu_j = 0$ and $\sigma_j = 1$ for $j = 1, \ldots, m$. As the effect modifiers are scaled, they range from approximately -2 to 2; a $\xi = 1$ would indicate an estimated inclusion probability varying from 0 to 1 along the range of the covariate, for all other covariates at their mean, and an intercept $\xi_0 = 0$. For less informative priors, we found $\xi$ did not converge in the unsmoothed VS model.

We fit the EVS model with different choices for the variance ratio $C$, $C=100$, 225, 400, and 900, and found that $C = 100$ had the lowest (best) Deviance Information Criterion (DIC; Spiegelhalter et al. (2002)). DIC includes a model goodness of fit term based on the likelihood, and a penalty term for model complexity. We calculate the DIC using the normal likelihood approximation. Results presented here are for the model with $C = 100$, and we also use $C = 100$ in both the smoothed and unsmoothed VS models with effect modifiers.
3.5.2 Results

In Table 3.4 we display the posterior estimates of $\alpha_k$, the effect size when selected in the model, $\pi_k$, the national level inclusion probability, and the average of $\beta_k^s$ across all sites. For the effect modifier models, $\Phi(\xi_{k0})$ represents the inclusion probability when the covariates $W$ are equal to their mean value, as all the covariates are centered. Figure 3.3 displays the site specific effect estimates $\beta_k^s$ and posterior site-specific inclusion probabilities, $\hat{\gamma}_k^s$. The county specific effects are grouped by EPA subregions to give some spatial context. The regions are Northeast (NE), Mid Atlantic (Mid Atl), South Atlantic (S Atl), East South Central (ESC), West South Central (WSC), East North Central (ENC), West North Central (WNC), Mountain (Mtn), and Pacific (Pac). We see that elemental carbon has the largest effect size in the EVS model, and it is selected across all three models, although probability of inclusion is much lower for the unsmoothed VS with effect modifiers model. The smoothed VS with effect modifiers model also includes ammonium and a small negative effect of sulfate.

We find a significant effect for elemental carbon which varies across locations. The average relative risk increase of CVD hospitalization for an IQR increase in ambient elemental carbon is estimated to be 0.38 in the VS with effect modifiers model, and 0.66 for the model with no covariates. The parameter $\alpha_4$ estimates the average effect in counties with elemental carbon selected in the model, and this is estimate to be 0.88 in the effect modifier model, and 0.96 in the model without covariates. We can see that including covariates in effect modification leads to higher prior and posterior variability in probability of inclusion, and also greater range of posterior estimates for $\beta$. The site specific estimates range from about 0.1 to 1.3 % increase in relative risk, as seen in Figure 3.4, and have an apparent regional pattern. We note that Peng et al. (2009) used a similar two-stage approximation without variable selection on the same data over the time period 2000-2006, and found a national average 0.8% relative risk increase for an IQR increase in elemental carbon, which is consistent with our result.

We also consider the effect modification parameters $\xi$. Of the covariates $W$, only the intercept terms were statistically significant. For the remaining six covariates, posterior probability of being greater than zero is 19 to 77%, indicating no significant effects. The overall effects $\mu_j$ also lack statistical significance. In Figure 3.5 we compare the effect estimates and posterior standard deviation of these coefficients for the smoothed and unsmoothed model. In the smoothed model, the posterior estimates of $\sigma_j$ were very small, indicating strong shrinkage toward $\mu_j$, and ultimately toward 0, having much smaller but more consistent effects across pollutants than in the unsmoothed model. Posterior standard deviations in the unsmoothed model tend to be very high. Though none of the parameters are statistically significant, we do see some interesting trends. For example, the coefficients for percent Hispanic are negative except for elemental carbon. It is not surprising that an effect different from the overall mean
is suggested for this pollutant, as elemental carbon shows up with strongest significance in the model. It is logical that the strongest signal, though still not statistically significant in this case, would be detected for this pollutant, and that for the others, which all are included with probability near zero, would follow the overall mean. Though the effects are not statistically significant, we see a positive association of inclusion probability with poverty, the percent of the population living in an urban center, and the median age of housing. The percent commuting by public transit, percent Hispanic, and percent Black are all negatively associated with inclusion probability.

Table 3.4: Posterior mean for $\alpha_k$, $\pi_k$, and average value for $\beta_k$. For the models with effect modifiers, $\Phi(\xi_{k0})$ is the posterior probability of inclusion for a site with average values of all effect modifiers. Both $\alpha_k$ and $\beta_k$ are presented as % increase in relative risk of CVD hospitalization for an IQR increase in the pollutant.

<table>
<thead>
<tr>
<th></th>
<th>EVS</th>
<th>VS with Effect Modifiers</th>
<th>Unsmoothed VS Effect Modifiers</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_k$</td>
<td>0.79</td>
<td>0.94</td>
<td>1.40</td>
</tr>
<tr>
<td>$\pi_k$</td>
<td>0.13</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>$\Phi(\xi_{k0})$</td>
<td>-0.21</td>
<td>0.44</td>
<td>0.58</td>
</tr>
<tr>
<td>$\text{avg } \beta$</td>
<td>0.23</td>
<td>0.33</td>
<td>0.38</td>
</tr>
<tr>
<td>Sulfate</td>
<td>0.33</td>
<td>0.07</td>
<td>0.25</td>
</tr>
<tr>
<td>Nitrate</td>
<td>-0.09</td>
<td>0.04</td>
<td>0.28</td>
</tr>
<tr>
<td>Silicon</td>
<td>0.96</td>
<td>0.72</td>
<td>0.66</td>
</tr>
<tr>
<td>Organic carbon</td>
<td>0.51</td>
<td>0.12</td>
<td>0.83</td>
</tr>
<tr>
<td>Sodium</td>
<td>0.51</td>
<td>0.07</td>
<td>0.33</td>
</tr>
<tr>
<td>Ammonium</td>
<td>1.24</td>
<td>0.18</td>
<td>0.91</td>
</tr>
<tr>
<td>Aluminum</td>
<td>-0.44</td>
<td>0.03</td>
<td>0.62</td>
</tr>
<tr>
<td>Arsenic</td>
<td>-1.18</td>
<td>0.06</td>
<td>0.11</td>
</tr>
<tr>
<td>Bromine</td>
<td>-0.29</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.91</td>
<td>0.07</td>
<td>0.62</td>
</tr>
<tr>
<td>Chlorine</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>Chromium</td>
<td>-0.68</td>
<td>0.02</td>
<td>-0.15</td>
</tr>
<tr>
<td>Copper</td>
<td>-0.08</td>
<td>0.03</td>
<td>0.11</td>
</tr>
<tr>
<td>Iron</td>
<td>0.74</td>
<td>0.07</td>
<td>0.33</td>
</tr>
<tr>
<td>Lead</td>
<td>-0.04</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Magnesium</td>
<td>-0.32</td>
<td>0.03</td>
<td>-0.35</td>
</tr>
<tr>
<td>Nickel</td>
<td>-0.23</td>
<td>0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>Potassium</td>
<td>-0.08</td>
<td>0.09</td>
<td>0.05</td>
</tr>
<tr>
<td>Titanium</td>
<td>-0.16</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>Vanadium</td>
<td>0.43</td>
<td>0.03</td>
<td>0.18</td>
</tr>
<tr>
<td>Zinc</td>
<td>0.46</td>
<td>0.04</td>
<td>0.21</td>
</tr>
</tbody>
</table>
3.6 Conclusion

In this chapter, we introduced a model which attempts to explain variability across counties in the pollutants deemed to be significant predictors of CVD hospital admissions. The two-stage approximation to the hierarchical variable selection model is a computationally efficient way to combine information across sites. A traditional hierarchical variable selection procedure such as EVS assumes that the prior probability of inclusion is the same across space. We more flexibly model the prior site-specific probability of inclusion, allowing it to vary across space as a function of spatially varying effect modifiers.

In our data analysis, we find that elemental carbon is associated with increased CVD hospitalization risk amongst Medicare patients. Using a hierarchical variable selection model, we pool information across sites to select the best model. In this case, effect modification does not dramatically change model fit or estimation. This may be because none of the effect modifiers selected are strongly related to the relationship of pollution and health, or simply because the estimated inclusion probabilities do not vary dramatically across space. The largest range for $\gamma_k^s$ is about 0.5 to 1. This type of model may be more effective in a situation in which inclusion probability is known to differ greatly across space. For example, Scheel et al. (2013) predict insurance claims as a function of weather variables in southern Norway. They utilize a spatial variable selection model to select important weather variables by site. The selected covariates vary dramatically across the region of study, and posterior inclusion probabilities for the covariates range from nearly 0 to nearly 1. In this or other similar cases, effect modifiers could be used to model this heterogeneity directly. For the pollution and CVD hospitalization case, though effect modifiers do not seem to describe the heterogeneity in probability inclusion, there does appear to be a spatial pattern in both posterior inclusion probability and effect size. An extension to the EVS model focused on exploiting spatial correlation follows in Chapter 4.
Figure 3.3: Posterior estimates of $\beta_k$ and $\gamma_k$ for the EVS model without effect modification ((a)-(b)), the VS with effect modifiers model ((c)-(d)), and the unsmoothed VS with effect modifiers model when there is no shared mean for the $\xi_k$ coefficients ((e)-(f)).
Figure 3.4: Estimated RR for elemental carbon by county for EVS model (top) and Effect modifier VS model (bottom)
Figure 3.5: Posterior mean and standard deviation of $\xi$ parameters in the effect modifiers model with shared mean (Figures (a) and (c)) and with no shared mean ((b) and (d)).
Chapter 4

Spatial Variable Selection Methods for Investigating Acute Health Effects of Fine Particulate Matter Components

4.1 Introduction

As discussed in Chapter 2, a large body of literature seeks to quantify the relationship of fine particulate matter components and human health outcomes. In Chapter 2 we model the inclusion probability as a function of site-specific covariates. We assume that the latent indicator of inclusion for covariate $k$ in the model at county $s$, $\gamma_k(s)$, are independent, and that the coefficient for covariate $k$ at site $s$, $\beta_k(s)$, are also independent across sites. In this chapter, we introduce spatial correlation into the model. We apply our method to the same PM.5 component and Medicare health data as in Chapter 3.

Previous work on spatial variable selection have taken either a global or local approach. Reich et al. (2010) consider multiple predictors globally, that is, included or excluded from the model at all sites. The coefficients for the included variables are allowed to vary by site with spatial priors. Smith and Fahrmeir (2007) propose local selection for a single predictor. The predictor may be included in the regression at some sites and not others. This is accomplished by treating the coefficient of interest as the product of a spatially correlated binary random field to determine covariate inclusion and exclusion, and an independent Gaussian field to determine the effect size given inclusion. Similar work with multiple predictors in generalized linear models have been undertaken by Lum (2012) and Scheel et al. (2013). In all these works, the Gaussian
field assumes the magnitudes of the coefficients are independent across sites, and there is no cross dependence between the magnitude and inclusion probability implied in the prior. Since the continuous components are independent, the resulting covariate process is not smooth across space, which may be undesirable in some cases.

We therefore extend this approach to include multiple predictors and spatial correlation in the effect size, while simultaneously allowing global selection as a special case. Rather than model the covariate processes as products of binary and continuous spatial processes, we fix the marginal distribution to be the familiar mixture distribution often used in Bayesian variable selection, and join the mixture marginal distributions using a Gaussian copula. This approach contains the independent variable selection model as a special case, permits both local and global variable selection, and provides a smooth spatial process for each covariate effect.

We again apply our method to the PM2.5 component data, to find pollutants associated with CVD hospitalization amongst medicare patients. Our model is an extension of stochastic search variable selection (SSVS) which allows the coefficient for each pollutant to vary by site, allowing for further unexplained heterogeneity due to health or exposure characteristics, and assumes these effects are spatially correlated. The coefficients will be shrunk toward zero at some or all counties, while incorporating spatial information.

In the remainder of this chapter, we begin with details of the PM2.5 and health data (Section 4.2), then further describe and motivate this spatial variable selection model for CVD hospital admissions (Section 4.3). In Section 4.4, we evaluate the performance of our model using a simulation study. We then apply the model to CVD data and find Elemental Carbon is significantly associated with CVD admissions (Section 4.5). In Section 4.6 we conclude.

4.2 Data

The data are the same as in Chapter 3. We select \( p = 22 \) components of PM2.5 which contribute at least 1% to total mass, or have been potentially linked with health outcomes in previous scientific literature. These speciated PM measurements are taken from the EPA’s Air Quality System (AQS) and AirExplorer databases (www.epa.gov/ttn/airs/airsaqs/, www.epa.gov/airexplorer/). We use the same inclusion and exclusion criteria as previously, resulting in a total of 115 counties with data observed in the time period 2000-2008. Because of the skewness of the pollutant data, we scale each pollutant by its 90th percentile. For consistency with previous published work using these data, we rescale all results to be reported as relative risk per IQR increase. We exclude any pollution values that are more than double the 5th-max at a given county, resulting in about 1.5% of the observations being removed.

The health data includes Medicare beneficiary enrollment and Medicare Part-A inpatient records, aggregated to daily county totals. We count the number of patients hospitalized with
a principal ICD-9 diagnosis code related to cardiovascular disease (CVD).

4.3 Model

We observe \( Y_t(s) \) CVD cases at county \( s \) on day \( t \), in addition to a vector of pollutants, \( x_t(s) \), and confounding variables \( z_t(s) \). As we treat the coefficients as a realization of a continuous spatial process, we use function notation rather than superscripts as in Chapter 3. The confounding variables include indicator functions for the day of the week and smooth functions of time, temperature, and dewpoint. We utilize the spatially varying coefficients model described by Gelfand et al. (2003),

\[
Y_t(s) | \beta(s), \eta(s) \sim \text{Poisson} \left[ N_t(s) \exp \{ \beta_0(s) + x_t(s)' \beta(s) + z_t(s)' \eta(s) \} \right]
\]

(4.1)

where \( N_t(s) \) is a known offset for the number of Medicare patients at county \( s \) on day \( t \) and \( \beta_0(s) \) is a spatially varying intercept. We include spatially varying coefficients \( \beta(s) = [\beta_1(s), \ldots, \beta_p(s)]^T \) and \( \eta(s) = [\eta_1(s), \ldots, \eta_q(s)]^T \). Note that the dimension of \( \eta(s), q_s \), depends on the county \( s \) due to the use of 7 knots per year for the spline function of time, and that each county may be observed over a different time interval. We control for unmeasured confounding independently within each location and then borrow information across locations for the health estimates. Therefore we select a low precision normal density as the prior for the confounding effects \( \eta(s) \), which is independent across counties. We let \( \beta_k(s) \) denote the coefficient on pollutant \( k \), for \( k = 1, \ldots, p \), and county \( s \), for \( s = 1, \ldots, n \). The prior on \( \beta_k(s) \) implies both spatial correlation and variable selection.

4.3.1 A Spatial Variable Selection Model

In the application to PM2.5 constituent data, we anticipate that some pollutants are important everywhere or nowhere (global), and some only have significant effects at certain locations. We define global behavior as the case where a given pollutant exhibits similar distribution across the entire space – either centered around the overall mean effect \( \alpha_k \), or around 0 – where the center is the same for all counties. A pollutant exhibiting local behavior may have an effect near the mean \( \alpha_k \) for some counties, but near zero for others. The marginal prior distribution for each coefficient \( \beta_k(s) \) is a mixture density,

\[
f_k[\beta_k(s)] = \pi_k N(\alpha_k, \omega_k^2) + (1 - \pi_k) N(0, \omega_k^2/C),
\]

(4.2)

where \( N(m, s^2) \) is the Gaussian density function with mean \( m \) and variance \( s^2 \). Global behavior is exhibited when \( \pi_k = 1 \) or 0. Coefficients are smoothed either toward \( \alpha_k \), the overall mean
when $\beta_k$ is important, or toward 0. Thus $\pi_k\alpha_k$ is the overall mean effect across all counties. Variability is controlled by $\omega_k$, with $C$ representing the ratio of variance for coefficients “in” and “out” of the model. One interpretation of $C$, is that if $\beta_k(s)$ is within $\pm 3\omega_k/\sqrt{C}$ of 0, it may be safely replaced by zero (George and McCulloch, 1997). If this tuning parameter $C = \infty$, the second part of the mixture distribution is a point mass at zero. This is the formulation of SSVS used by Smith and Fahrmeir (2007) and Reich et al. (2010), in which the coefficients are modeled as the product of binary and continuous fields. In the following analysis and simulation study, we use the value $C = 100$. In Section 4.5.4 we discuss the sensitivity of the model to this choice.

In our model we apply a Gaussian copula (Nelsen, 1999; Sklar, 1973, 1953) to the marginal distribution in (4.2). A copula is a general technique for modeling dependence between random variables while preserving desired marginal distributions. To implement the copula model, we introduce latent variables, $\theta_k(s)$, which follow a mean zero Gaussian process with correlation function $\rho(s, s')$. Here we use an exponential spatial correlation function,

$$
\rho(s, s') = \text{cor} \left[ \theta_k(s), \theta_k(s') \right] = \exp \left( \frac{|s - s'|}{r} \right).
$$

The range parameter $r$ has the interpretation that at distance $3r$ correlation is about 0.05. Although we choose an exponential correlation structure, any correlation function may be used, including a generalization to non-stationary areal correlation structure such as that implied by a conditional autoregressive model (CAR). We fix the variance of $\theta_k$ to one for identifiability purposes in the Gaussian copula. We write a Gaussian process with mean $m$, standard deviation $v$, and exponential spatial correlation with range $r$ as GPex($m$, $v$, $r$). Hence, $\theta_k(s) \sim \text{GPex}(0, 1, r)$. We force $\beta_k(s)$ to have the desired marginal distribution by the transformation

$$
\beta_k(s) = F_k^{-1} \left\{ \Phi[\theta_k(s)] \right\},
$$

where $\Phi$ is the standard normal cdf and $F_k$ is the marginal cdf of $\beta_k(s)$ defined in (4.2) and dependent on $\pi_k$, $\alpha_k$, and $\omega_k$. Note that for the special case of $C = \infty$ we can write the inverse cdf $F_k^{-1}$ in closed form (see Appendix) allowing us to easily generate correlated data for simulation purposes.

### 4.3.2 A comparison of copula and other model solutions

As noted in the introduction, an alternative approach is to define each coefficient as the product of a binary field and a continuous field. In a non-spatial setting with $p$ potential predictors, we define $\beta_k$ as the coefficient on predictor $k$. One approach to stochastic search variable selection defines $\beta_k = \gamma_k B_k$, where $\gamma_k$ is a binary random variable and $B_k$ is continuous, as described in
Figure 4.1: One realization from the indicator model and from our copula prior on a 100×100 grid. In both, $\alpha_k = 0.5$, $\omega_k = 0.2$, and $\pi_k = 0.5$. The continuous part of the two-indicator model are drawn from a GPex(0.5, 0.2, 10) model, such that spatial correlation at neighboring sites is about 0.9. The binary part is $\gamma_k(s) \sim \text{Bernoulli}\{\Phi[z_k(s)]\}$, where $z_k(s)$ are GPex(0, 1, 10) and drawn independently from the continuous part. The latent copula $\theta \sim \text{GPex}(0, 0, 10)$ from the same seed as the continuous part of the two-process model, to give similar pictures. This example is the copula model with $C = \infty$ to give exact zeros.

The introduction. In this section, we extend this notation to include spatial variation, in which $\beta_k(s) = \gamma_k(s)B_k(s)$ is a realization of a spatial process at site $s$, and is the product of a binary spatial process $\gamma_k(s)$, and $B_k(s)$ is a continuous process.

The global approach by Reich et al. (2010) focuses on selecting covariates for a spatial regression model with multivariate response, where coefficients are selected “in” or “out” of the model at all locations simultaneously. Using our notation, their model can be summarized as $\gamma_k(s) = \gamma_k \sim \text{Bernoulli}(\pi)$ and $B_k(s)$ is a Gaussian Process with spatial correlation. Because selection is global, it is very possible that if a covariate only has an effect at one or a few locations, it will be selected “out” of the model. This type of model structure would be appropriate when it is expected that covariates have similar, but not equivalent, effects across sites.

The local approach by Smith and Fahrmeir (2007) is an application to functional Magnetic Resonance Imaging (fMRI) data, in which brain activity is measured in response to a stimulus. The primary focus is on finding voxels activated by the stimuli, and these are expected to occur in small regions or clusters. This model has $p = 1$ predictor. The binary activity indicator varies spatially according to an Ising (auto-logistic) prior, which is a binary analog to the Conditionally
Figure 4.2: A comparison of the joint density of $\beta_k(s)$ and $\beta_k(s')$ under an indicator model (a) and our copula model (b). In (a), $\beta_k(s) = \gamma_k(s)B_k(s)$, where $\gamma_k(s) \perp \gamma_k(s')$, and $B_k(s)$, $B_k(s')$ are multivariate normal with correlation $\rho$. In (b), our copula model where $\gamma_k(s)$, $\beta_k(s')$ joined by Gaussian copula with correlation $\rho$. For both plots, the marginal density is the mixture in (4.2), with $\pi = 0.8$, $\alpha = 0.5$, $\omega = 0.5$, and the correlation is $\rho = 0.9$.

Autoregressive (CAR) model used for continuous data measured on a lattice. The continuous part, measuring the effect size, is assumed to vary across sites, but no spatial correlation is directly imposed in the prior. Using our notation, we can write $\gamma_k(s) = \gamma(s) \sim \text{Ising}$, and $B_k(s) = B(s) \sim N(0, \sigma^2)$.

Treating $\beta_k(s)$ as the product of two fields or variables requires choosing a correlation and cross-correlation structure for the two processes. Figure 4.2 (a) illustrates the bivariate density function of coefficients at nearby counties for one choice of binary and continuous fields. This is as in the Smith and Fahrmeir (2007) model, in which the binary field is correlated, and the continuous field is an independent Gaussian process. The density of our copula model is shown in Figure 4.2 (b). Notice that in the indicator model, when one coefficient is non-zero and the other is zero, there is no shrinkage of the nonzero coefficient as there is in the copula model. While careful tuning of the correlation structures may yield a density similar to the copula density, the copula model has the advantage of creating a continuous prior surface, and naturally incorporating the intuition that if a component has no effect within a given county, nearby counties are unlikely to have large coefficients. The realization of the spatial process on a spatial grid in Figure 4.1 shows large areas with $\beta(s) = 0$ (light blue) corresponding to regions of null effects. The effect is near zero for sites near null regions, and $\beta(s)$ varies smoothly across sites in non-null regions. Figure 4.2 illustrates that the copula model automatically introduces...
4.4 Simulation study

In the simulation study, we investigate the performance of our model in terms of mean absolute deviation (MAD), power, Type I error, and ability to correctly identify null coefficients compared to a spatial model without variable selection and a variable selection model without spatial correlation. We compare performance for varying degrees of spatial correlation in the coefficients $\beta_k(s)$. We generate coefficients $\beta_k(s)$ from multivariate normal latent variables $\theta_k(s)$ with exponential correlation using the transformation in (4.3) where $F$ is the cdf of (4.2) with $C = \infty$. Though we fit the continuous version of the model where $C < \infty$, we chose to generate the data for the $C = \infty$ case so that the true values of the coefficients are exactly zero when they are out of the model, whether globally or locally.

Each model has $p = 9$ covariates. By varying $\pi_1, \ldots, \pi_9$ we determine the behavior for each of the nine pollutant coefficient vectors. We define “global” behavior, as being generated with $\pi_k = 1$ and thus $\beta_k(s)$ are drawn from a multivariate normal distribution, $\beta_k(s) \sim N(\alpha_k, \omega_k^2 \Sigma)$, where $\Sigma$ is an exponential correlation matrix. Globally “null” covariates have $\pi_k = 0$ and hence $\beta_k(s) = 0$ for all $s$. Covariates exhibiting “local” behavior have a non-zero mean $\alpha_k$, but $0 < \pi_k < 1$ so that the marginal distribution at each site is a mixture distribution.

For computational purposes, we restrict the simulation study to 48 counties in the Eastern United States. The maximum distance between these counties is 1360 km. We include the components which contribute the greatest mass to total PM2.5: sulfate, nitrate, elemental carbon, organic carbon, silicon, and sodium, as well as arsenic, bromine, and calcium. By using the actual pollutant data to generate responses, we evaluate our under the more realistic scenario in which covariates of interest are moderately to strongly correlated within site across time. Table 4.1 specifies the components and true parameters $\alpha_k$ and $\pi_k$ used to generate the data. The overall mean $\alpha$ is 0, 0.05, or 0.10, and the inclusion probability $\pi_k$ is 0, 0.3, 0.7, or 1, indicating null, local, or global inclusion. For all coefficients we set $\omega_k = 0.025$. We use the Poisson regression model (4.1) and the observed pollution data $x_t(s)$ to generate values of $Y_t(s)$ based on our draws of $\beta(s)$. That is, for each of $M$ datasets we generate $\beta_k(s)$ and then $Y_t(s)|\beta(s), x_t(s)$. For simplicity, we do not include any confounding variables in the simulation study.

Following the procedures outlined above, we generate $M = 50$ datasets using each of the following designs for the correlation in $\beta_k(s)$:

1. **No Spatial Correlation** The true coefficients for each county are independent.
2. **Moderate Spatial Correlation** The true coefficients for each county are exponentially
spatially correlated with range $r = 60$ km (effective range 180 km).

3. **Strong Spatial Correlation** The true coefficients for each county are exponentially spatially correlated with range $r = 240$ km (effective range 720 km).

In all cases we also include a random intercept $\beta_0(s)$, with $E[\beta_0(s)] = -8$ yielding an approximate average risk similar to the median risk observed. For the model with no spatial correlation, the intercepts are also uncorrelated. The intercept in designs 2 and 3 are exponentially spatially correlated with $r = 120$ km, which implies an effective range of 360 km. The standard deviation of $\beta_0(s)$, $\omega_0 = 1$, is similar to the between county standard deviation of average CVD rate.

We investigate the following three models:

1. **Spatial Model** (Sp) $\beta_k(s) \sim \text{GPex}(\alpha_k, \omega_k, r)$, without variable selection (i.e., $\pi_k = 1, k = 1, \ldots, p$).

2. **Exchangeable VS Model** (EVS) $\beta_k(s)$ are independent and follow the distribution in (4.2).

3. **Spatial VS Model** (SpVS) The full model in Section 4.3. $\beta_k(s)$ marginally follow a mixture distribution described in (4.2), and dependence is induced with a Gaussian copula.

In the simulation study, we use relatively uninformative priors, with $\alpha_k \sim N(0, 1)$, $\pi_k \sim \text{Beta}(1, 1)$, and $1/\omega_k^2 \sim \text{Gamma}(0.001, 0.001)$. For each dataset we run the model for 5500 iterations discarding the first 1000. We compare these three models using MAD, power, type-I error, and the ability to identify null coefficients.

### 4.4.1 Simulation Results

As shown in Table 4.1, MAD is lower for the VS models than the spatial model, and when there is spatial correlation, the SpVS model has lower MAD than the EVS model. For the globally “in” covariates, the spatial and SpVS models have similar MAD as expected. For locally significant coefficients, both VS models are superior as they correctly shrink coefficients toward zero. The differences are particularly dramatic when $\pi_k = 0.3$ and when $\alpha_k = 0.1$, as these create a more bimodal distribution of $\beta_k(s)$ which the spatial model is unable to accommodate. As spatial correlation increases, the differences between the SpVS and EVS models become apparent. Though both identify null coefficients well, the SpVS model has improved MAD for the coefficients which are included in the model. As expected, the SpVS model shows its superior performance the most for local coefficients when spatial correlation is strong. In this case, MAD for the SpVS model is 50% lower than the spatial model and 13% lower than the EVS model.
for null coefficients, and is nearly 20% lower than the EVS model for non-null coefficients. Tests of statistical significance (0.05 level) are carried out using a frequentist paired t-test. Absolute deviations are averaged across all 48 sites. The differences between MAD for each of the 50 datasets are then used to calculate a t-statistic.

Type-I error rates are comparable across all models, and are near the nominal 0.05 level. Type-I error are much less than 0.05 for all three models for the globally zero coefficients. For the local covariates, there is not a significant difference in power when \( \pi_k = 0.3 \) and when \( \pi_k = 0.7 \), therefore power is only reported for each level of \( \alpha_k \) in Table 4.2. When the coefficients are uncorrelated, the spatial model has the highest power, with the EVS and SpVS models having slightly lower power very similar to each other. When the coefficients have weak or strong correlation, both spatial models significantly outperform the EVS model.

We also investigate the ability of each model to correctly identify when coefficients are near zero. Because the probability of observing an estimated \( \beta_k(s) = 0 \) is zero, we instead calculated the proportion of times the posterior median estimate of \( \beta_k(s) \) is within a small neighborhood of zero. In Table 4.3, we use the neighborhood \( \pm 2\omega/\sqrt{C} \), in this case \( \pm 0.005 \). As expected, the Spatial model does very poorly at this task, and the SpVS and EVS models perform similarly well. When spatial correlation is strong, the SpVS model is an improvement over EVS for identifying local coefficients.

Finally, in Table 4.4 we check the ability of three goodness of fit measures, Deviance Information Criterion (DIC), a measure by Gelfand and Ghosh (1998) (GG), and log pseudo marginal likelihood (LPML) to pick the correct model. Deviance Information Criterion (DIC) combines a goodness of fit term and a model complexity penalty (Spiegelhalter et al., 2002). The measure described by Gelfand and Ghosh (1998) (denoted GG in this chapter) under squared error loss can be viewed as the sum of a predictive goodness of fit measure and a model complexity penalty. As with DIC, smaller values are preferred. Log pseudo marginal likelihood (LPML) as described in Carlin and Louis (2009) is used to evaluate a model’s predictive power. Larger values (closer to zero) are preferred. We see that DIC and LPML most often pick the correct “best” model, while GG is less accurate. These measures are described in Section 4.5.2.

### 4.5 Analysis of Health Data

#### 4.5.1 Description of data and models

We use the CVD hospitalization and PM2.5 component data described in Section 4.2. We fit the Spatial, EVS, and Spatial VS Poisson regression models described in Section 4.4. We use 22 components as predictors in this model. As noted in Section 4.3, we included indicator functions for the day of the week, and smooth functions of temperature, dewpoint, and time.
Table 4.1: Median absolute deviation (MAD) × 100 for simulation study. MAD averaged across all 50 locations for each of the \( k = 1, \ldots, 9 \) pollutants, and an overall average presented here, for the Spatial (Sp), Exchangeable Variable Selection (EVS), and Spatial Variable Selection (SpVS) models. Simulation settings for \( \alpha_k, \pi_k \) are shown in columns 2 and 3, and abbreviated names of the components used to generate the response are in column 1. For all pollutants, we generate data with \( \omega_k = 0.025 \) and effective range 0 km (Independent), 180 km (Weak Correlation), or 720 km (Strong Correlation). A * indicates a statistically significant difference from the SpVS model.

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>( \alpha )</th>
<th>( \pi )</th>
<th>Independent Sp EVS SpVS</th>
<th>Weak Correlation Sp EVS SpVS</th>
<th>Strong Correlation Sp EVS SpVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulf.</td>
<td>0.05</td>
<td>1</td>
<td>1.39* 1.66* 1.48</td>
<td>1.36 1.65* 1.38</td>
<td>1.10 1.55* 1.12</td>
</tr>
<tr>
<td>Nitr.</td>
<td>0.10</td>
<td>1</td>
<td>1.40* 1.46 1.45</td>
<td>1.37* 1.59* 1.41</td>
<td>1.16 1.43* 1.14</td>
</tr>
<tr>
<td>Sili.</td>
<td>0.05</td>
<td>0.3</td>
<td>1.34* 0.91 0.95</td>
<td>1.41* 0.99* 1.11</td>
<td>1.17* 0.95 0.91</td>
</tr>
<tr>
<td>El C.</td>
<td>0.05</td>
<td>0.7</td>
<td>1.91* 2.10 2.08</td>
<td>1.79 2.04* 1.78</td>
<td>1.49* 1.83* 1.44</td>
</tr>
<tr>
<td>Sodi.</td>
<td>0.10</td>
<td>0.3</td>
<td>1.53* 0.80 0.86</td>
<td>1.52* 0.79 0.84</td>
<td>1.21* 0.70 0.66</td>
</tr>
<tr>
<td>Or C.</td>
<td>0.10</td>
<td>0.7</td>
<td>2.30* 2.02 2.04</td>
<td>2.30* 2.01 1.95</td>
<td>1.91* 2.05* 1.60</td>
</tr>
<tr>
<td>Arse.</td>
<td>0</td>
<td>0</td>
<td>0.53* 0.12 0.12</td>
<td>0.50* 0.16 0.19</td>
<td>0.46* 0.13* 0.18</td>
</tr>
<tr>
<td>Brom.</td>
<td>0</td>
<td>0</td>
<td>0.60* 0.15 0.13</td>
<td>0.56* 0.14* 0.21</td>
<td>0.50* 0.12* 0.24</td>
</tr>
<tr>
<td>Calc.</td>
<td>0</td>
<td>0</td>
<td>0.63* 0.15 0.15</td>
<td>0.69* 0.15* 0.25</td>
<td>0.54* 0.12* 0.26</td>
</tr>
<tr>
<td>Avg. Local, ( \beta_k(s) = 0 )</td>
<td></td>
<td></td>
<td>1.61* 1.19* 1.23</td>
<td>1.64* 1.24 1.24</td>
<td>1.26* 0.72 0.63</td>
</tr>
<tr>
<td>Avg. Local, ( \beta_k(s) \neq 0 )</td>
<td></td>
<td></td>
<td>1.93* 1.72 1.73</td>
<td>1.87* 1.67* 1.59</td>
<td>1.62 2.04 1.67</td>
</tr>
</tbody>
</table>

as confounding variables. We use natural cubic smoothing spline functions to model these relationships, and examine sensitivity to these choices in Section 4.5.4. As is also discussed in Section 4.5.4, we fit the spatial variable selection model with different choices the value of \( C \), the ratio of included/excluded variance.

Due to the large number of confounding variables per site, we chose to fit a two-stage approximation to the fully Bayesian model, as in Chapter 3. The Poisson likelihood function in 4.1 is approximated by a normal likelihood, \( \hat{\beta}(s) \sim N_p[\hat{\beta}(s), \hat{V}(s)] \), where \( \hat{\beta}(s) \) is the maximum likelihood estimator and \( \hat{V}(s) \) is the estimated covariance matrix. This two-stage model has the advantage of much faster computation time and model fit does not depend on MCMC convergence of the confounding coefficients, as the approximation is fit marginally over the confounding coefficients \( \eta(s) \).

In all of the models, we choose a flat prior between 0 and 1 for \( \pi_k \), but note that one could use a beta prior for \( \pi_k \) to encourage global selection by choosing both beta parameters to be less than one. Local selection (0 < \( \pi_k < 1 \)) could also be weighted more heavily in the prior,
Table 4.2: Power and Type-I error for spatial model with no variable selection (Sp), exchangeable variable selection model (EVS), and spatial variable selection model (SpVS) for local and global covariates. We calculate power for each setting of the true value of $\alpha_k$.

<table>
<thead>
<tr>
<th>$\alpha_k$</th>
<th>Independent</th>
<th>Weak Correlation</th>
<th>Strong Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sp  EVS SpVS</td>
<td>Sp  EVS SpVS</td>
<td>Sp  EVS SpVS</td>
</tr>
<tr>
<td>Local</td>
<td>0.05</td>
<td>0.41 0.30 0.34</td>
<td>0.48 0.31 0.49</td>
</tr>
<tr>
<td>Local</td>
<td>0.10</td>
<td>0.80 0.76 0.75</td>
<td>0.84 0.74 0.81</td>
</tr>
<tr>
<td>Global</td>
<td>0.05</td>
<td>0.74 0.55 0.61</td>
<td>0.75 0.55 0.73</td>
</tr>
<tr>
<td>Global</td>
<td>0.10</td>
<td>1.00 0.92 0.92</td>
<td>0.99 0.88 0.95</td>
</tr>
</tbody>
</table>

and greater prior weight could be placed on inclusion or exclusion by choosing an asymmetric beta distribution for the prior. The prior for $\alpha_k$ is normal with standard deviation 0.1, which is approximately five to ten times larger than the effect sizes we expect to see. Recalling that we model the log relative risk, 0.1 would correspond to a more than 10% increase in risk of CVD hospitalization for a 90th percentile increase in pollution, which previous work suggests is quite large. We have found that model fit is somewhat sensitive to the choice of the prior on $\omega_k$. When $\omega_k$ is large it becomes difficult to differentiate between the 0 and $\alpha_k$ modes of the distribution, and thus $\pi_k$ and $\alpha_k$ become difficult to estimate. Though $\alpha_k$ and $\pi_k$ are therefore somewhat sensitive to the choice of prior on $\omega_k$, we found that estimates of $\beta_k(s)$ are more stable over different choices of prior. We therefore choose a gamma(0.1, 0.001) prior on $\omega_k^2$ to be relatively uninformative while encouraging $\omega_k$ to be small. We note in the simulation study that larger effect sizes are more robust to an uninformative prior on $\omega_k^2$, and that this prior is equivalent to fitting gamma(0.1, 0.1), a vague prior, on data with 10 times larger effect sizes, roughly equivalent to the effect sizes of the simulation study.

We fit the spatial VS model with different choices for the value of $C$; $C=100, 225, 400,$ and 900. DIC was nearly identical for the first three choices, and highest for $C=900$. Here we present the model with $C = 400$. The EVS model presented here also has $C = 400$ for consistency. As noted in Chapter 3, $C = 100$ minimizes DIC for the EVS model, however, the model fit is very similar.

We generate 6500 MCMC samples, discarding the first 1500 for burn in. For consistency,
Table 4.3: Proportion of posterior median estimates of $\beta_k(s)$ more than $\pm 2\omega/\sqrt{C}$ away from zero.

<table>
<thead>
<tr>
<th></th>
<th>Independent</th>
<th>Weak Correlation</th>
<th>Strong Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sp  EVS  SpVS</td>
<td>Sp  EVS  SpVS</td>
<td>Sp  EVS  SpVS</td>
</tr>
<tr>
<td>Global, $\beta \neq 0$</td>
<td>0.99 0.97 0.98</td>
<td>1.00 0.97 0.98</td>
<td>0.99 0.97 0.99</td>
</tr>
<tr>
<td>Global, $\beta = 0$</td>
<td>0.48 0.02 0.02</td>
<td>0.47 0.02 0.07</td>
<td>0.39 0.02 0.08</td>
</tr>
<tr>
<td>Local, $\beta \neq 0$</td>
<td>0.98 0.87 0.87</td>
<td>0.98 0.89 0.91</td>
<td>0.98 0.88 0.93</td>
</tr>
<tr>
<td>Local, $\beta = 0$</td>
<td>0.74 0.21 0.22</td>
<td>0.74 0.19 0.28</td>
<td>0.67 0.20 0.27</td>
</tr>
</tbody>
</table>

Table 4.4: Number of times each model was selected as “best” by each of three goodness of fit measures: Deviance Information Criterion (DIC), the measure by Gelfand and Ghosh (GG), and log pseudo-maximum-likelihood (LPML).

<table>
<thead>
<tr>
<th></th>
<th>Independent</th>
<th>Weak Correlation</th>
<th>Strong Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sp  EVS  SpVS</td>
<td>Sp  EVS  SpVS</td>
<td>Sp  EVS  SpVS</td>
</tr>
<tr>
<td>DIC</td>
<td>0  25  25</td>
<td>0  12  38</td>
<td>0  2  48</td>
</tr>
<tr>
<td>GG</td>
<td>15  17  18</td>
<td>14  32  4</td>
<td>6  43  1</td>
</tr>
<tr>
<td>LPML</td>
<td>0  24  26</td>
<td>0  13  37</td>
<td>0  2  48</td>
</tr>
</tbody>
</table>

we fit all three models using the same metropolis algorithm, with range fixed at 0 for the EVS model, and $\pi_k$ fixed at 1 for the spatial model.

4.5.2 Model comparisons

We compare models using DIC, LPML, and GG, the goodness of fit measures described in Section 4.4.1. We calculate LPML and GG statistics using the profile likelihood, the Poisson likelihood in 4.1 with $\eta_k(s)$ replaced by its maximum likelihood estimate, $\hat{\eta}_k(s)$. We calculate the DIC using both the normal likelihood approximation.

The goodness of fit measures for each model are shown in Table 4.5. We find that the spatial VS model is selected as the best model by DIC and LPML. The spatial model without variable selection is selected as the best model by GG, but as shown in the simulation study, this result is questionable. The EVS model is second best according to DIC and LPML. As noted, we present the EVS model with $C=400$ here for consistency; though the EVS model with $C=100$
had slightly lower DIC, it is still not as low as the spatial VS model. We also note the estimated spatial correlation is very strong, further evidence that the spatial models are a better choice than the exchangeable model.

4.5.3 Analysis results

![Figure 4.3: Estimated county-specific relative risk for an IQR increase in elemental carbon from exchangeable VS model.](image)

Across all models, we consistently find positive overall effects for elemental carbon. In the EVS model, elemental carbon is a global effect, and the overall effect estimate $\alpha_4$ is statistically significant (95% posterior credible interval excludes 0). For the spatial and spatial VS models, posterior probability that $\alpha_4 > 0$ is 0.88. No other overall effects are statistically significant, though some county specific relative risk estimates are significant. These results are in general agreement with Peng et al. (2009), who found 0.8% increase in CVD hospitalization risk for an IQR increase in elemental carbon, over the time period 2002-2006 in a six pollutant model. In all models, we see a spatial pattern in that the effects of elemental carbon are highest in the Northeastern United States and along the West coast. The overall effects for all pollutants are reported in Table 4.5, where local covariates are defined as $0.15 < \pi_k < 0.80$ and global
covariates as $\pi_k > 0.80$. Individual county effects for all pollutants are shown in Figure 4.6. Each figure is divided into sections for each of the eight EPA subregions, to give some spatial context. Selected maps of individual county effects are shown here; maps for all pollutants are in the appendix. All effects are reported as percent increase in relative risk for an IQR increase in the pollutant. In Table 4.6 we present the posterior estimates for the overall mean when included $\alpha_k$, the inclusion probability $\pi_k$, and the random effects standard deviation, $\omega^2_k$ for our spatial VS model.

In addition to the effect of elemental carbon, we see local effects of organic carbon in the spatial VS model. Though the overall effect estimate is small, some counties have significant effects. In particular, in Figure 4.6 and in Figures 4.4 and 4.5 we see that elemental and organic carbon have differing spatial patterns, with the elemental carbon effects being highest on the coasts, and the organic carbon effects highest in Ohio, Michigan, and western Pennsylvania. In the spatial model, we see a large negative effect for sulfate, and a large positive effect for ammonium. These two pollutants are highly correlated, with correlation up to 0.9 or higher in some counties. Neither pollutant has a strong effect in the VS models, seeming to indicate these results are due to multicollinearity in the spatial model.

### 4.5.4 Sensitivity Analysis

In Table 4.5 and Figure 4.6, there is clearly some sensitivity of both overall and individual effects between models. In addition to these model choices, we investigated the sensitivity of the results to a number of other modeling choices. The first is the choice of model for the confounding variables, including time, temperature and dewpoint. We account for as much variability as possible using the confounding variables to isolate the effects of the PM components. To allow for a flexible and nonlinear relationship to the response, cubic splines are a natural choice, and have been used frequently in the air pollution health effects literature. In their analysis of PM2.5 components and CVD hospitalizations, Peng et al. (2009) use 8 degrees of freedom (df) per year, plus 6 df for temp and 6 df for 3-day lag temp, and 3 for dewpoint and 3-day lag dewpoint. Temperature and dewpoint are highly correlated with each other. By plotting the temperature and dewpoint effect estimates from this rich model as a function of time, it appears the effects are canceling each other out, and that using 18 df for the temp/dewpoint model could be done more concisely. Dominici et al. (2002) use a model with 6 df for temperature and 3 df for 3-day lag dewpoint, and we choose this model for the weather variables. Dominici et al. (2002) use 7 df per year for the time function. We fit the two-stage Spatial VS model with $C = 400$ with 4, 7, 8, and 12 df per year. We found that the DIC was lowest for the 4 df model, and second lowest for our model presented in this Chapter, the model with 7 df per year. In the 7, 8, and 12 df models, elemental carbon had a consistently strong effect, while in the 4
Table 4.5: Estimated overall relative risk expressed as percent increase \( e^{\alpha_k \pi_k} \times 100 - 100 \) and goodness of fit statistics. For Spatial model (no VS), \( e^{\alpha_k} \times 100 - 100 \) is reported. For models with spatial correlation, range estimates in km are given. Statistically significant associations (0.05 level) marked by *. Global selection (posterior \( \pi_k > 0.8 \)) marked by \( ^G \) and local selection marked by \( ^L \) (posterior \( 0.15 < \pi_k < 0.80 \)). We also include goodness of fit criterion, DIC, GG, and LPML. The best model by each measure is highlighted in bold.

<table>
<thead>
<tr>
<th></th>
<th>SpVS</th>
<th>Spatial</th>
<th>EVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfate</td>
<td>0.579</td>
<td>0.059</td>
<td>0.122</td>
</tr>
<tr>
<td>Nitrate</td>
<td>-0.101</td>
<td>-0.006</td>
<td>0.150</td>
</tr>
<tr>
<td>Silicon</td>
<td>0.233</td>
<td>-0.053</td>
<td>0.003</td>
</tr>
<tr>
<td>Elemental carbon</td>
<td>0.535</td>
<td>0.647</td>
<td>0.755^G^*</td>
</tr>
<tr>
<td>Organic carbon</td>
<td>0.020</td>
<td>-0.226</td>
<td>0.039</td>
</tr>
<tr>
<td>Sodium</td>
<td>-0.121</td>
<td>-0.107</td>
<td>0.025</td>
</tr>
<tr>
<td>Ammonium</td>
<td>-0.358</td>
<td>0.630</td>
<td>0.032</td>
</tr>
<tr>
<td>Aluminum</td>
<td>0.119</td>
<td>-0.084</td>
<td>-0.080</td>
</tr>
<tr>
<td>Arsenic</td>
<td>-0.136</td>
<td>-0.580</td>
<td>-0.018</td>
</tr>
<tr>
<td>Bromine</td>
<td>-0.148</td>
<td>0.046</td>
<td>0.096</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.083</td>
<td>0.022</td>
<td>0.017</td>
</tr>
<tr>
<td>Chlorine</td>
<td>0.107</td>
<td>-0.013</td>
<td>0.011</td>
</tr>
<tr>
<td>Chromium</td>
<td>0.004</td>
<td>-0.166</td>
<td>0.002</td>
</tr>
<tr>
<td>Copper</td>
<td>-0.136</td>
<td>0.026</td>
<td>0.000</td>
</tr>
<tr>
<td>Iron</td>
<td>-0.108</td>
<td>0.383</td>
<td>-0.016</td>
</tr>
<tr>
<td>Lead</td>
<td>0.242</td>
<td>-0.074</td>
<td>-0.064</td>
</tr>
<tr>
<td>Magnesium</td>
<td>0.195</td>
<td>-0.081</td>
<td>-0.011</td>
</tr>
<tr>
<td>Nickel</td>
<td>0.101</td>
<td>0.065</td>
<td>-0.046</td>
</tr>
<tr>
<td>Potassium</td>
<td>0.320</td>
<td>-0.121</td>
<td>-0.027</td>
</tr>
<tr>
<td>Titanium</td>
<td>0.068</td>
<td>0.114</td>
<td>-0.047</td>
</tr>
<tr>
<td>Vanadium</td>
<td>0.377</td>
<td>-0.141</td>
<td>-0.010</td>
</tr>
<tr>
<td>Zinc</td>
<td>0.041</td>
<td>0.166</td>
<td>-0.044</td>
</tr>
<tr>
<td>Range (km)</td>
<td>1143</td>
<td>1175</td>
<td></td>
</tr>
<tr>
<td>DIC</td>
<td>2630</td>
<td>2731</td>
<td>2678</td>
</tr>
<tr>
<td>GG</td>
<td>27.333</td>
<td>27.295</td>
<td>27.341</td>
</tr>
<tr>
<td>LPML</td>
<td>-155457</td>
<td>-155612</td>
<td>-155519</td>
</tr>
</tbody>
</table>

df model no pollutants had strong effects. In the models with at least 7 df, the overall effect of elemental carbon remains consistent, but individual county effects are somewhat variable. Correlation between county effect estimates with 7 and 8 df is 0.60, and correlation between county effect estimates with 7 and 12 df is 0.50.
Table 4.6: Estimates of mean when included \((e^{\alpha_k} \times 100 - 100)\), probability of inclusion \(\pi_k\) from the Spatial VS model with 22 pollutants, and estimate of standard deviation, rescaled \((e^{\omega_k} \times 100 - 100)\).

<table>
<thead>
<tr>
<th></th>
<th>(\alpha)</th>
<th>(\pi)</th>
<th>(\omega)</th>
<th></th>
<th>(\alpha)</th>
<th>(\pi)</th>
<th>(\omega)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfate</td>
<td>1.41</td>
<td>0.44</td>
<td>1.47</td>
<td>Chlorine</td>
<td>0.96</td>
<td>0.07</td>
<td>1.72</td>
</tr>
<tr>
<td>Nitrate</td>
<td>-0.71</td>
<td>0.13</td>
<td>1.82</td>
<td>Chromium</td>
<td>-0.11</td>
<td>0.07</td>
<td>1.15</td>
</tr>
<tr>
<td>Silicon</td>
<td>1.43</td>
<td>0.13</td>
<td>1.86</td>
<td>Copper</td>
<td>-1.29</td>
<td>0.06</td>
<td>2.54</td>
</tr>
<tr>
<td>Elemental carbon</td>
<td>0.85</td>
<td>0.57</td>
<td>1.22</td>
<td>Iron</td>
<td>-0.74</td>
<td>0.10</td>
<td>3.07</td>
</tr>
<tr>
<td>Organic carbon</td>
<td>-0.03</td>
<td>0.45</td>
<td>1.59</td>
<td>Lead</td>
<td>0.90</td>
<td>0.08</td>
<td>2.08</td>
</tr>
<tr>
<td>Sodium</td>
<td>-0.80</td>
<td>0.09</td>
<td>1.50</td>
<td>Magnesium</td>
<td>2.50</td>
<td>0.09</td>
<td>1.83</td>
</tr>
<tr>
<td>Ammonium</td>
<td>-2.00</td>
<td>0.15</td>
<td>3.12</td>
<td>Nickel</td>
<td>0.92</td>
<td>0.06</td>
<td>2.19</td>
</tr>
<tr>
<td>Aluminum</td>
<td>-0.04</td>
<td>0.13</td>
<td>1.38</td>
<td>Potassium</td>
<td>3.06</td>
<td>0.09</td>
<td>4.16</td>
</tr>
<tr>
<td>Arsenic</td>
<td>-0.28</td>
<td>0.14</td>
<td>1.99</td>
<td>Titanium</td>
<td>0.91</td>
<td>0.05</td>
<td>3.59</td>
</tr>
<tr>
<td>Bromine</td>
<td>-0.91</td>
<td>0.21</td>
<td>2.43</td>
<td>Vanadium</td>
<td>2.68</td>
<td>0.09</td>
<td>1.37</td>
</tr>
<tr>
<td>Calcium</td>
<td>0.79</td>
<td>0.07</td>
<td>1.39</td>
<td>Zinc</td>
<td>0.42</td>
<td>0.12</td>
<td>1.79</td>
</tr>
</tbody>
</table>

We also fit the EVS and SpVS models with several choices for the variance ratio \(C\). We chose \(C=100, 225, 400\), and 900. For \(C=900\), no variables are selected into either model. For \(C<900\) in the EVS model, elemental carbon is chosen as the most significant variable in all models. DIC for the EVS model is lowest for \(C=100\). For the SpVS models, EC is selected as a local covariate for \(C < 900\). Organic carbon, sulfate and nitrate are selected as local covariates but not consistently across models. Individual county effects of elemental carbon are relatively robust across choices for \(C\), but for \(C = 100\) or 225, the estimated posterior inclusion probability \((\pi_4)\) for elemental carbon is small, indicating no separation of a mode near \(\alpha_4\) and a mode near zero. DIC for the SpVS model is highest for \(C = 900\) and nearly equivalent for \(C=100, 225\) and 400.

### 4.6 Discussion

In this chapter, we develop a statistical model which allows for local variable selection of spatially correlated coefficients. The copula framework creates a smooth prior surface reflecting the intuition that effect estimates near counties where the effect is zero will also be small. A simulation study confirms that including spatial correlation in the model, while adding computational complexity, can result in better model fit and improved effect estimation than the spatial model or the exchangeable variable selection model. While the MAD for globally included covariates
are similar between spatial and spatial VS models as expected, variable selection dramatically improves MAD for local and globally excluded covariates as effect estimates are accurately shrunk toward zero. The most significant gains are seen when inclusion probability is small and the effect size when included is large. The spatial VS model also improves upon the exchangeable model in terms of both MAD and power. While both VS models show improved MAD over the spatial model, spatial VS has lower MAD when effect estimates are correlated. Importantly, the spatial VS model has much higher power than the exchangeable model, particularly as correlation increases. These results demonstrate the gains in model fit and power by sharing information across space.

We use this Spatial VS model to investigate which components of fine particulate matter may have an association with human health effects. This investigation can help us understand how observed spatial and temporal variations in the health effects of total PM2.5 concentration may be related to variation in chemical components. No previous study to our knowledge has utilized variable selection techniques on such a large list of components, and few have incorporated spatial correlation. In our analysis, we see a positive impact of elemental carbon consistent with the work by Peng et al. (2009). The overall effect of elemental carbon is statistically significant in our EVS analysis, but not for SpVS at the 0.05 level. Elemental carbon effects tend to be highest in the Northeastern US and West coast, and there is some evidence of an effect of organic carbon in Ohio and Western Pennsylvania. The spatial variable selection model fits the data better than the exchangeable model or a naive spatial model according to multiple goodness of fit criterion.

When interpreting these overall effect estimates, it is also important to note that these numbers reflect only the average risk increase or decrease for the 115 counties included in the study, and are not a reflection of an average national risk. The counties selected for the study are generally densely populated urban counties which may differ significantly in pollution levels, demographic, socio-economic and health characteristics from other US counties. Also, while epidemiological studies such as this one can point to interesting relationships between pollutant levels and health effects, we must be careful in interpreting effects as causal links. As discussed by Thomas et al. (2007), the indicated pollutants may be correlated with other pollutants not included with the study, or in fact may be proxying for another pollutant which is included. Because we do not observed the relationship between ambient and personal exposure, the indicated pollutants may be signaling due to higher association with personal exposure, though another pollutant is more correlated with health. Measurement error, particularly that of using ambient pollution as a proxy for personal exposure, may also play a role, especially for pollutants which are extremely spatially heterogeneous. A spatially homogenous pollutant may have stronger association between ambient and personal exposure, and thus may mask the more heterogeneous pollutant which has a relationship between personal exposure and health.
Previous work studying measurement error and coarse PM10 (diameter 2.5–10 \( \mu m \)) (Chang et al., 2011) and total PM10 (Zeger et al., 2000) suggest relatively consistent estimates of that effect under different measurement error scenarios, but spatial heterogeneity of individual components of PM2.5 may be appreciably greater, and individual components vary in their level of heterogeneity (Bell et al., 2011). Further investigation of how spatial heterogeneity and measurement error may change health effect estimates is warranted, particularly in variable selection models.

Further research into the reason for heterogeneity in pollutant effects using both epidemiological and toxicological approaches is also needed. Heterogeneity may be due to differences in particle chemistry, interaction effects, differing population exposure characteristics, or the measurement error issues described above. Though we have divided PM into its chemical components, the physical and toxicological properties are by no means homogenous. For example, sulfate and nitrate particles are mostly secondary pollutants created by reactions of sulfur and nitrogen oxides with other gases and particles to form a diverse set of molecules, including sulfite and nitrite, which varies in acidity (Schlesinger et al., 2006). Toxicological studies suggest this acidity, not just the components, may be an important factor in health outcomes (Schlesinger et al., 2006). Individual chemical components may come from diverse primary sources as well, resulting in differing size and associated particles. For example, potassium becomes airborne from burning biomass, from blown crustal material (dust), and salts from sea spray which may have differing biological effects (Schlesinger, 2007). An extension of this SpVS model to a space-time setting would be useful for fully understanding possible seasonal effects. Though we adjust for seasonal confounding, we do not consider interactions of pollutant effect and season in the current model.
Figure 4.4: Estimated county-specific relative risk for an IQR increase in elemental carbon (a) and organic carbon (b) from the SpVS model.
Figure 4.5: Estimated county-specific relative risk for an IQR increase in elemental carbon (a) and organic carbon (b) from the Spatial model.
Figure 4.6: Matrix of all estimated coefficients (in RR per IQR increase) for each of the six models in Table 1. Each row includes coefficients for the labeled pollutant; each column indicates one county. The counties are grouped according to the 8 EPA subregions. The purpose of this plot is to highlight differences in local estimates for each of the models.
REFERENCES


67


Appendix A

Computing Details for Spatial Bridge Model

This section contains WinBUGS code for a nested random effects model with 4 covariates for \( n \) subjects with \( m \) locations recorded for each subject. The subject random effects \( \gamma_i \) are independent Bridge random variables, and the location (“tooth”) random effects \( \varepsilon_{ij} \) are multivariate bridge, with exponential spatial correlation matrix \( \Sigma \). The constants \( m, n, \) the matrix of distances between locations (DIST), and the mathematical constant \( \pi \), need to be entered as data in addition to \( X \) and \( Y \).

```r
model{
  for (i in 1:n){ for (j in 1:m){

    Y[i,j] ~ dbern(pi[i,j])

    logit(pi[i,j]) <- alpha+X1[i,j]*beta1+ X2[i,j]*beta2+
    X3[i,j]*beta3+ X4[i,j]*beta4+gam[i]+eps[i,j]

    # get PHI(theta), P(z<theta)
    ptheta[i,j]<-phi(theta[i,j]) #phi is N(0,1) cdf
    eps[i,j]<- log( sin(bphi*pi*ptheta[i,j])/
    sin(bphi*pi*(1-ptheta[i,j])))/bphi2

  }#endj
}
```
# Draw the subject random effects from Bridge distribution
u[i]~dunif(0,1)
gam[i]<- log( sin(bphi1*pi*u[i])/sin(bphi1*pi*(1-u[i])))/(bphi1*bphi2)

# Draw the latent location random effects from
# a multivariate normal distribution
theta[i,1:m]~dmvnorm(0, V[,])

} # end i

# hyperpriors
range~dlnorm(-2,1)
beta1~dnorm(0,btau)
beta2~dnorm(0,btau)
beta3~dnorm(0,btau)
beta4~dnorm(0,btau)

# Define precision matrix for theta, e.g. exponential
# (could also use spatial.exp instead of dmvnorm above)
for(k in 1:m){ for(l in 1:m){
  Sigma[k,l]<-exp(-DIST[k,l]/range)
}
V[1:m,1:m]<-inverse(Sigma[,])

# Prior for precision in RE dist'ns is gamma.
tau1~dgamma(1,.125)
tau2~dgamma(1,.125)

# define the bridge parameters phi... use 'bphi' since 'phi' is a function.
bphi1<-1/sqrt(3/(tau1*pi^2)+1)
bphi2<-1/sqrt(3/(tau2*pi^2)+1)
}
Appendix B

Supplemental Material for VS with Effects Modifiers

B.1 Maps of Effect Modifiers Used in PM Component Data Analysis

Figure B.1: Proportion of population which is Hispanic.
Figure B.2: Proportion of population which is Black.

Figure B.3: Proportion of adults over age 65 in poverty.
Figure B.4: Proportion of population commuting by public transit.

Figure B.5: Proportion of population living in an urban center.
B.2 Simulated Example

To demonstrate the benefit of effect modification in variable selection we consider a simple example. For simplicity, we generate \( \hat{\beta}, \beta, \) and \( \gamma \) directly rather than also generating data from the first stage Poisson model. We consider a vector of \( p = 5 \) covariates of interest at \( n = 100 \) sites. The true value for \( \beta^s \) is either 0, 1, 1.5, or 2, as shown in Figure B.7(a). Two effect modifiers \( W \) are generated independently as standard normal random variables. True values for \( \xi \) are selected to give varying levels of inclusion. The values for \( \xi_1 \) were 3, 0, 2, -3, and -3, while for \( \xi_2 = \xi_3 = -1, \) and \( \xi_3 = -1. \) Thus, the probability \( x_2 \) is included depends upon \( W_1 \) only, and inclusion of \( x_3 \) depends upon \( W_1 \) and \( W_2. \) The inclusion indicators \( \gamma_k \) are generated as Bernoulli random variables with probability \( \Phi(W^T \xi_k) \). The values of \( \hat{\beta} \) are \( \beta^s \) plus Gaussian random noise with variance 0.5. This emulates the regression coefficient estimates from a linear model with five independent covariates.

As seen in Figure B.7, the posterior estimates of \( \gamma_k \) are better in the model with covariates included. The estimates of \( \beta_k \) have lower MSE for all but covariate 2 for the VS with effect modifiers model as well. Experimentation confirms that effect modifiers can show more improvement in the model as the effect sizes for \( \xi \) increase and as variance in \( \hat{\beta}_k^s \) increases. In addition
to illustrating the potential benefits of including effect modifiers in the model, this example demonstrates the effectiveness of our computational algorithm for estimating $\beta_k(s)$ and $\gamma_k^s$. 
Figure B.7: True values for $\beta^s_k$ and $\gamma^s_k$ from a simulated example and the resulting estimated parameters from an EVS model and a VS model with effect modifiers.
Appendix C

Computing Details for Spatial Variable Selection Model

Let $\theta_k(s)$ follow a multivariate normal distribution with some spatial correlation matrix, $\theta_k = (\theta_k(1), \ldots, \theta_k(n)) \sim N_n(0, \Sigma_k)$. In our study, we use an exponential spatial correlation, however any geostatistical or areal correlation matrix is acceptable, including the correlation matrix implied by the Conditional Autoregressive (CAR) model popular for areal data.

Define

$$\beta_k(s) = F_k^{-1}[\Phi(\theta_k(s))]$$

where $F_k$ is the marginal cdf of $\beta_k(s)$ defined in (4.2) and dependent on $\pi_k, \alpha_k,$ and $\omega_k$.

Note that for the special case of $C=\infty$, we can easily write the inverse cdf in closed form as:

$$F_{\psi_k}^{-1}(y) = \begin{cases} 
\alpha_k + \omega_k \Phi^{-1} \left( \frac{y}{\pi_k} \right) & \text{if } y < \pi_k \Phi \left( \frac{-\alpha_k}{\omega_k} \right) \\
0 & \text{if } \pi_k \Phi \left( \frac{-\alpha_k}{\omega_k} \right) < y < 1 - \pi_k + \pi_k \Phi \left( \frac{-\alpha_k}{\omega_k} \right) \\
\alpha_k - \omega_k \Phi^{-1} \left( \frac{1-y}{\pi_k} \right) & \text{if } y \geq 1 - \pi_k + \pi_k \Phi \left( \frac{-\alpha_k}{\omega_k} \right)
\end{cases}$$

for $y \in [0, 1]$, and $\psi_k$ indexes the other parameters ($\alpha_k, \omega_k, \pi_k$). This allows us to easily generate data for the simulation study. Unfortunately, when $C$ is any finite number, we can’t write the inverse cdf in closed form, and must use an approximation if it is necessary for computation (e.g. to generate data).

However, when $C$ is finite, we can use a simple variable transformation to find the joint distribution of $\beta_k$ directly, as

$$f(\beta_k(1), \ldots, \beta_k(n)) = \frac{\phi_\Sigma [\theta_k(1), \ldots, \theta_k(n)]}{\prod_{i=1}^n \phi [\theta_k(i)]} \prod_{i=1}^n f [\beta_k(i)],$$
where \( \theta_k(i) = \Phi^{-1} \{ F[\beta_k(i)] \} \). We use this joint density in MCMC sampling.

To generate the data, we will utilize the observed pollutant data, and generate the response as described in Chapter 4. For the simulation, we ignore other confounding variables.

The Exchangeable variable selection (EVS) model is fit using the same model with range parameters fixed at zero. The Spatial (Sp) model with no variable selection could be implemented using Gibbs’ sampling for \( \alpha \) and \( \omega \), however for consistency we fit within the same MCMC as the Spatial VS model with \( \pi_k \) fixed at 1.

The Spatial VS model utilizes Metropolis sampling for \( \alpha \) with a normal candidate distribution. The updates for \( \omega \) and \( \pi \) are Metropolis-Hastings. We let the candidate distribution for \( \omega_k \) be defined by \( \log \omega_k^* \sim N(\omega_k, \sigma_{\text{TUNE}}^2 \omega) \), and the candidate distribution for \( \pi_k \) be defined on the scale of \( \mu_k = \Phi^{-1}(\pi_k) \), such that \( \mu_k^* \sim N(\mu_k, \min \left\{ \frac{\sigma_{\text{TUNE}}^2 \omega}{\pi_k (1 - \pi_k)}, 1 \right\}) \).

All models use Gibbs’ sampling for \( \alpha_0 \) and \( \omega_0 \), and Metropolis updates for \( \beta_0, \beta \). We update these one at a time using a normal candidate distribution.

For two-stage analysis, we use the approximate model described in Chapter 3, and \( \beta_k(s) \) are updated using Metropolis steps.
Appendix D

Spatial Variable Selection Model
Results

In this appendix we include the county specific effect estimates for each of the 22 pollutants from our spatial variable selection model described in Chapter 4.

Figure D.1: Percent relative risk increase for one IQR increase of Sulfate from the SpVS model with 22 pollutants.
Figure D.2: Percent relative risk increase for one IQR increase of Nitrate from the SpVS model with 22 pollutants.

Figure D.3: Percent relative risk increase for one IQR increase of Silicon from the SpVS model with 22 pollutants.
Figure D.4: Percent relative risk increase for one IQR increase of Elemental carbon from the SpVS model with 22 pollutants.

Figure D.5: Percent relative risk increase for one IQR increase of Organic carbon from the SpVS model with 22 pollutants.
Figure D.6: Percent relative risk increase for one IQR increase of Sodium from the SpVS model with 22 pollutants.

Figure D.7: Percent relative risk increase for one IQR increase of Ammonium from the SpVS model with 22 pollutants.
Figure D.8: Percent relative risk increase for one IQR increase of Aluminum from the SpVS model with 22 pollutants.

Figure D.9: Percent relative risk increase for one IQR increase of Arsenic from the SpVS model with 22 pollutants.
Figure D.10: Percent relative risk increase for one IQR increase of Bromine from the SpVS model with 22 pollutants.

Figure D.11: Percent relative risk increase for one IQR increase of Calcium from the SpVS model with 22 pollutants.
Figure D.12: Percent relative risk increase for one IQR increase of Chlorine from the SpVS model with 22 pollutants.

Figure D.13: Percent relative risk increase for one IQR increase of Chromium from the SpVS model with 22 pollutants.
Figure D.14: Percent relative risk increase for one IQR increase of Copper from the SpVS model with 22 pollutants.

Figure D.15: Percent relative risk increase for one IQR increase of Iron from the SpVS model with 22 pollutants.
Figure D.16: Percent relative risk increase for one IQR increase of Lead from the SpVS model with 22 pollutants.

Figure D.17: Percent relative risk increase for one IQR increase of Magnesium from the SpVS model with 22 pollutants.
Figure D.18: Percent relative risk increase for one IQR increase of Nickel from the SpVS model with 22 pollutants.

Figure D.19: Percent relative risk increase for one IQR increase of Potassium from the SpVS model with 22 pollutants.
Figure D.20: Percent relative risk increase for one IQR increase of Titanium from the SpVS model with 22 pollutants.

Figure D.21: Percent relative risk increase for one IQR increase of Vanadium from the SpVS model with 22 pollutants.
Figure D.22: Percent relative risk increase for one IQR increase of Zinc from the SpVS model with 22 pollutants.