DISSERTATION

CLIPPED LATENT-VARIABLE SPATIAL MODELS FOR ORDERED CATEGORICAL DATA

Submitted by

Megan Dailey Higgs

Department of Statistics

In partial fulfillment of the requirements

for the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2007

COLORADO STATE UNIVERSITY

September 18, 2007

WE HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER OUR SUPERVISION BY MEGAN DAILEY HIGGS ENTITLED CLIPPED LATENT-VARIABLE SPATIAL MODELS FOR ORDERED CATEGORICAL DATA BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Committee on Graduate Work

Jennifer A. Hoeting (Adviser)

Phillip L. Chapman (Committee Member)

Ryan T. Elmore (Committee Member)

N. LeRoy Poff (Committee Member)

F. Jay Breidt (Department Head)

ABSTRACT OF DISSERTATION

CLIPPED LATENT-VARIABLE SPATIAL MODELS FOR ORDERED CATEGORICAL DATA

Ordered categorical data arise in a variety of scientific disciplines. The ordered categories may be of primary interest, or they may be recorded in an attempt to simplify data collection. When the reason for the categorization is simplification, the categories are typically created based on an underlying continuous variable and a set of threshold values used to define the categories. Information in the continuous variable is clearly lost, but an analysis of ordered categorical data can indirectly involve a latent (unobserved) underlying continuous variable. In fact, this concept creates a convenient platform on which to build models for ordered categorical data, both from an analytical and computational perspective.

Latent variable models, coupled with Bayesian inference, has been an active area of research for independent ordered categorical data (e.g. Albert and Chib, 1993; Cowles, 1996; Nandram and Chen, 1996; Chen and Dey, 1997). Such models rely on the techniques of data augmentation and Markov chain Monte Carlo (MCMC) algorithms. The models for independent data have been extended to include correlation in the response(s), focusing on clustered and repeated measures data (e.g. Chen and Dey, 1996; Chib and Greenberg, 1998; and Ishwaran and Gatsonis, 2000). We, however, are interested in incorporating spatial dependence into the model structure, an extension that has not yet been specifically addressed in the literature. Thus, our goal is to develop models for data collected at point-referenced locations over space. Spatial ordered categorical data result from a variety of research areas, such as ecology, epidemiology, and the social sciences.

Spatial models for binary and count data have received more recent attention (e.g. Diggle et al., 1998; Gelfand et al., 2000; and Christensen et al., 2006). Most of the models for binary and count data remain within the convenient context of one-parameter exponential family distributions, embedding the Gaussian process within the framework of generalized linear mixed models (GLMMs). Such models have been coined generalized linear spatial models. An alternative approach to modeling binary data relies on the concept of a clipped Gaussian random field (De Oliveira, 2000).

In this dissertation, we extend and combine the modeling strategies for independent ordered categorical data and those for binary and count spatial data to develop two models for spatially dependent point-referenced ordered categorical data. The development of these models involves conceptual model building, as well as the development of computational algorithms.

Our proposed models differ in their conceptual origin, in the correlation they induce in the categorical response, and in their generalizability. The first model, termed the Double Latent model, extends the generalized linear spatial model framework described by Diggle et al. (1998), incorporating the latent spatial variable as a random effect. The second model extends the idea of a clipped Gaussian random field as proposed by De Oliveira (2000) and is termed the Clipped Gaussian model. Both approaches rely on the use of an underlying latent Gaussian random field, but differ in how the categorical random field is created from an underlying continuous distribution and at what level the mean structure is defined. They do, however, also have fundamental similarities and can under the probit link assumption be compared as nugget and no-nugget models. Both models are developed under the Bayesian paradigm using Gibbs sampling and MCMC methods, and result in different algorithms used for inference.

We assess and compare the models through analytical, graphical, and simulation-based methods. We assess the primary goal of prediction by formulating

prediction as a decision problem within the Bayesian paradigm, and testing prediction at new locations with the use of hold-out data sets. An in-depth simulation study is used to investigate and compare the behaviors of the two models, demonstrating their success in terms of prediction and estimation on a large variety of realizations of artificial data created under the assumptions of both models and different sets of parameter values. The usefulness of the models to real problems is demonstrated through an application to ordered categorical data describing stream health through an "index of biotic integrity" in Montgomery County, Maryland.

Megan Dailey Higgs
Department of Statistics
Colorado State University
Fort Collins, Colorado 80523
Fall 2007

ACKNOWLEDGEMENTS

The work reported in this dissertation was developed under the STAR Research Assistance Agreement CR-829095 awarded by the U.S. Environmental Protection Agency (EPA) to Colorado State University. This dissertation has not been formally reviewed by the EPA. The views expressed here are solely those of the author and STARMAP, the program they represent. EPA does not endorse any products or commercial services mentioned in this dissertation. This research was also partially supported by the National Science Foundation under Grant DGE-0221595003. I extend many thanks to the STARMAP program and the PRIMES program at Colorado State University for playing a large role in my professional development through allowing me to focus on research and attend a variety of stimulating seminars and conferences.

This dissertation is the culmination of nearly fifteen years of interaction with faculty and students in higher education. In fact, I would have completed my degree in less time had I simply followed my original plan to obtain an M.D. instead of a Ph.D. The path toward a degree in Statistics was cleared for me by science teachers and professors who focused on critical thinking, research methods, and the importance of writing. They kept me interested in the scientific process and fed my desire to continue my education.

I extend specific thanks to my adviser, Dr. Jennifer Hoeting, for her encouragement and expertise, as a statistician and as a working mother. I am grateful to my committee for their time spent reading this dissertation and the resulting comments and suggestions. My adviser from my master's degree, Dr. Alix Gitelman, also

deserves acknowledgement for giving me confidence to initially pursue this doctoral degree, along with her continued support throughout the process.

There are many fellow graduate students to whom I owe thanks, but I especially would like to thank Kathi Irvine for the many hours on the phone "talking stats," along with providing needed motivation and encouragement at crucial moments. And, thanks to Mark Delorey for fielding my questions and for being gracious enough to let me finish first.

My family deserves a great deal of thanks. My parents, Mike and Sue, provided their love and confidence, along with their advice to "stay in school." Their words and actions truly helped to get me to this point, from being wonderful grandparents to editing my writing. My aunt and uncle, Joan and Burt, deserve similar acknowledgement, most recently for their hours spent editing this dissertation. I also want to thank my brother, Jed, for being a constant reminder of the fact that despite the long hours it took to finish this work, I could have been working more; and my always positive sister-in-law Angie for being a nanny at crucial times.

And finally, the completion of this dissertation is very much the result of constant support and understanding from my immediate family. Our dogs, Naya and Hitch, always give me the perfect excuse to get outside. My daughter, Shaden, provides constant perspective regarding the important little things in life, along with giving me a great deal of motivation to finish this degree. And, most of all, I owe thanks to my loving husband, Steve. Without his constant patience and his willingness to make being a good father and my education the priorities in his life, I would not be here. It is because of him that I have been able to realize this goal while still thoroughly enjoying being a mom to our wonderful daughter. And now, we are finally back in Montana to raise our family together, near family.

DEDICATION

This dissertation is dedicated to my daughter, Shaden, in hopes that she will someday appreciate its role in bringing her home to Montana.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION AND PRELIMINARIES

Ordered categorical data arise in a variety of scientific disciplines. The ordered categories may be of primary interest or they may be recorded in an attempt to simplify data collection. For example, time and money can be saved by simply recording a category label describing a particular experimental or observational unit, rather than obtaining and recording a continuous measurement for each unit. When the purpose is simplification, it is often the case that the categories are created based on an underlying continuous variable and a set of cut-point values used to categorize the response. Much of the information in original continuous variable is obviously lost, but an analysis of ordered categorical data can indirectly involve a latent (unobserved) version of the underlying continuous random variable. The concept of categorizing an underlying continuous random variable may also be useful in situations where such a variable is less tangible, if only for computational reasons. Ordered categorical data are also commonly referred to as *ordinal* data, but we use the term *ordered categorical* for its straightforward description of the data type.

Specific examples of such data in an ecological setting are categories obtained from indices, such as the Index of Biotic Integrity (IBI), that describe the overall health of an aquatic area as *poor*, *fair*, *good*, or *excellent*, or a habitat variable, such as health of the riparian vegetative zone, recorded as *poor*, *marginal*, *suboptimal*, or *optimal*. There are many examples of ordered categorical data in the behavioral and social sciences, recorded with labels defined by the Likert scale (Likert, 1932), ranging from *strongly disagree* to *strongly agree*. A less obvious example is that of

size categories of the substrate in streams, recorded as a categorical variable out of convenience at the time of data collection. In this work, we develop models for ordered categorical data that exhibit spatial dependence. We specifically focus on point-referenced data, as opposed to lattice or areal data.

A fair number of models have now been developed for incorporating spatial dependence into models for binary or count data, such as those proposed by Diggle et al. (1998); De Oliveira (2000); Christensen et al. (2000); Kammann and Wand (2003); Paciorek and Ryan (2005); Zhu et al. (2005b), and others. The strategy for fitting these models is straightforward due to the inclusion of the Binomial and Poisson distributions in the one-dimensional exponential family and the now broad development of generalized linear mixed models (GLMM) for such data. However, for point-referenced ordered categorical data, the need for spatial models remains.

This dissertation begins with an introduction to the analysis of ordered categorical data, including a brief review of generalized linear and generalized linear mixed models and their analogous Bayesian formulations. Chapter 1 further presents a review of the literature describing spatial generalized linear models for binary and count data, along with an introduction to Bayesian spatial prediction and Markov chain Monte Carlo (MCMC) methodology used to fit the models we propose in Chapter 2.

1.1 Traditional models for ordered categorical data

We first give an introduction to the analysis of ordered categorical data in the context of independent observations. We define the ordered categorical response variable, Y, as taking values from a set with K ordered elements, $\{r_1, r_2, \ldots, r_K\}$. For convenience, we let $r_k = k$, so that the set of possible values for Y is $\{1, 2, \ldots, K\}$. We restrict our attention to the multi-category situation ($K > 2$) since for $K = 2$ we have the binary case for which the response variable fits into the univariate exponential family and models are already developed. Ordered multi-category response

models are typically formulated as cumulative link generalized linear models (GLM) (McCullagh and Nelder, 1989; Agresti, 2002). As the name suggests, these models are constructed based on cumulative probabilities, $Pr(Y \leq k)$ for $k = 1, \ldots, K$. These cumulative probabilities have practical meaning due to the ordered nature of the categories.

1.1.1 The generalized linear model approach

For ordered categorical data, the typical modeling approach in the context of generalized linear models sets the linear predictor equal to a function, termed the link function, of the cumulative probabilities. The link function can be any monotone increasing function that maps $(0, 1)$ to the real line, but is typically an inverse cumulative distribution function (cdf) (Agresti, 2002). We let F denote a cdf and F^{-1} denote its inverse, such that the model is defined as

$$F^{-1}[Pr(Y \leq k|\mathbf{x})] = \theta_k - \mathbf{x}\boldsymbol{\beta} \qquad k = 1, \ldots, K-1, \tag{1.1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ column vector, $\mathbf{X}$ is a $n \times p$ design matrix, and $\theta_1 < \theta_2 < \ldots < \theta_{K-1}$ to preserve the ordered nature of the categories. We define $\boldsymbol{\beta}$ as a column vector, instead of the typical row vector, for notational convenience. For a single observation, we denote the $1 \times p$ design vector as $\mathbf{x}$. The $\{\theta_k\}$ are generally referred to as the *cut-point* parameters on the scale specified by the link function, and are typically of little interest (McCullagh, 1980). The terminology is explained in Section 1.1.2. We adopt the common convention of setting $\theta_1 = 0$ and including an intercept term in $\boldsymbol{\beta}$ and thus a column vector of ones in $\mathbf{X}$ (Agresti, 2002; Albert and Chib, 1993). For estimability of the parameters, we cannot leave all cut-point parameters to be estimated along with an intercept parameter.

Possible link functions described below include the logit function (inverse standard logistic cdf), the probit function (inverse normal cdf), the inverse Cauchy cdf,

and the log-log function, among others. Linear models of the form (1.1) are qualitatively similar and often produce almost indistinguishable results for different link functions applied to the same data set. Thus, McCullagh (1980), suggests choosing a link function primarily for ease of interpretation. This leads to the common use of the logistic, log-log, and complementary log-log models, as opposed to the probit and inverse Cauchy models. However, it will become clearer later that the probit model has clear computational advantages over the other link functions and should not be easily dismissed, especially when the main goal of the analysis is prediction rather than estimation. There are also alternatives to cumulative models, such as adjacent-category, sequential and two-step models (Agresti, 2002; Albert and Chib, 1998).

The cumulative model in (1.1) assumes that the effects of the covariates are the same for each category. Another way of viewing this assumption is that for two values of a single covariate, x_1 and x_2, either $Pr(Y \leq k|x_1) \leq Pr(Y \leq k|x_2)$ or $Pr(Y \leq k|x_1) \geq Pr(Y \leq k|x_2)$ for *all* k (Agresti, 2002). The convention of using $\theta_k - \mathbf{x}\boldsymbol{\beta}$, as opposed to $\theta_k + \mathbf{x}\boldsymbol{\beta}$, aids in interpretation of the coefficients and also arises logically in the formulation of the model using a latent variable, as will be described in Section 1.1.2. This convention allows the usual interpretation of the sign of a regression coefficient, meaning that a positive β_1 indicates that value of Y tends to increase as the value of the covariate increases. In the context of cumulative probabilities, a positive coefficient indicates that the cumulative probability decreases as x increases, meaning that there is less probability mass for lower values of Y than for the higher values. Thus, larger values of the covariate are associated with larger values of the response.

The cumulative probabilities are assumed to strictly increase in k for a fixed $\mathbf{x}$, corresponding to strictly increasing $\{\theta_k\}$. This restriction reflects the assumption that $Pr(Y = k|\mathbf{x}) > 0$ for all k, and is needed to ensure a non-zero likelihood.

We define $Pr(Y = k|\mathbf{x}) = p_k(\mathbf{x})$, such that the cumulative probability is $Pr(Y \leq k|\mathbf{x}) = \sum_{j=1}^{k} p_k(\mathbf{x})$. Let $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iK})$ be a K-dimensional vector of binary indicators of category for observation i. That is, $u_{ik} = 1$ if $y_i = k$ and the remaining elements in the vector are zero.

The general likelihood function for the cumulative model is given by

$$\begin{aligned} f(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{X}) &= \prod_{i=1}^{n} \left[\prod_{k=1}^{K} p_k(\mathbf{x}_i)^{u_{ik}} \right] \\ &= \prod_{i=1}^{n} \left[\prod_{k=1}^{K} (Pr(Y_i \leq k|\mathbf{x}_i) - Pr(Y_i \leq k-1|\mathbf{x}_i))^{u_{ik}} \right] \\ &= \prod_{i=1}^{n} \left[\prod_{k=1}^{K} (F(\theta_k - \mathbf{x}_i\boldsymbol{\beta}) - F(\theta_{k-1} - \mathbf{x}_i\boldsymbol{\beta}))^{u_{ik}} \right]. \end{aligned} \tag{1.2}$$

We can alternatively express the above likelihood using the observed response vector, $\mathbf{y}$, when each element is the integer label of the categorical response, as previously defined. That is, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ where $y_i = k$ if the ith observation belongs to the kth category for $k = 1, \ldots, K$. Then, $\mathbf{y}$ can be used to index the appropriate cut-point parameters and simplify the likelihood expression to

$$f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{X}) = \prod_{i=1}^{n} (F(\theta_{y_i} - \mathbf{x}_i\boldsymbol{\beta}) - F(\theta_{y_i-1} - \mathbf{x}_i\boldsymbol{\beta})). \tag{1.3}$$

We focus our attention on the cumulative probit model, utilizing the inverse of the standard normal cdf, Φ^{-1}, as the link function. This is a common model for ordered categorical data and is written as

$$\Phi^{-1}[Pr(Y \leq k|\mathbf{x})] = \theta_k - \mathbf{x}\boldsymbol{\beta}. \tag{1.4}$$

The intuition for the model and reasons for choosing it are discussed in further detail in Section 1.1.2. As previously mentioned, cumulative probit models typically result in fits similar to the cumulative logit model (Agresti, 2002).

We also provide a brief description of the cumulative logit model due to its popularity and similarity to the cumulative probit model. The model is defined

using the logit link function, $F^{-1}(\eta) = \log[\eta/(1-\eta)]$, which is the inverse of the standard logistic cdf. The cumulative logit is defined as

$$\text{logit}[Pr(Y \leq k|\mathbf{x})] = \log\left[\frac{Pr(Y \leq k|\mathbf{x})}{1 - Pr(Y \leq k|\mathbf{x})}\right] = \log\left[\frac{Pr(Y \leq k|\mathbf{x})}{Pr(Y > k|\mathbf{x})}\right], \qquad (1.5)$$

for all k, and the model is written as

$$\log\left[\frac{Pr(Y \leq k|\mathbf{x})}{Pr(Y > k|\mathbf{x})}\right] = \theta_k - \mathbf{x}\boldsymbol{\beta}. \qquad (1.6)$$

A model for a single cumulative logit (for one k) is simply a logit model for a binary response where the K categories are grouped into 2 categories via their ordered property. The model in (1.6) simultaneously uses all cumulative logits, and assumes that $\boldsymbol{\beta}$ remains constant for each individual logit. For more than two response categories, it is *not* equivalent to fitting separate logit models for each category since (1.6) constrains the $K-1$ response curves to have the same shape. The difference between cumulative logits for the same category but different levels of the covariate is independent of the specific category (McCullagh, 1980). For interpretation, this translates into the assumption that the log cumulative odds ratio is proportional to the distance between two vectors of the covariates, $\mathbf{x}_1$ and $\mathbf{x}_2$, and is hence given the name *proportional odds model* (Agresti, 2002). For only two response categories, it is equivalent to the logistic model for binary data and is also equivalent to a log-linear model (McCullagh, 1980).

As previously mentioned, a number of other link functions have been investigated. A model defined using the complementary log-log link is termed the proportional hazards model and is written $\log[-\log(1 - Pr(Y \leq k|\mathbf{x}))] = \theta_k - \mathbf{x}\boldsymbol{\beta}$ (McCullagh, 1980). For a single covariate, the model has the property that $Pr(Y > k|x_1) = [Pr(Y > k|x_2)]^{\exp[(x_1 - x_2)\beta_1]}$, resulting in the cumulative probability for category k approaching 1 at a faster rate than it approaches 0 (Agresti, 2002). Others have proposed mixture distributions for link functions, where a mixing

parameter is estimated from the data (Albert and Chib, 1993; Chen and Dey, 1996; Ishwaran and Gatsonis, 2000).

For nominal (unordered) categorical data, an appealing property is permutation invariance, meaning that the categories can be permuted in an arbitrary way without affecting the fit or values of the parameters. This is a general property of log-linear models (McCullagh, 1980), but is not appropriate for ordinal data. Instead, a more fitting requirement is that the model should be invariant under a reversal of category order, which is termed palindromic invariance. This property holds for the logistic, probit, and inverse Cauchy models since reversal of category order simply results in a change in sign of β, along with a change in sign and order of the cut-point parameters, $\boldsymbol{\theta}$.

1.1.2 The latent variable formulation

For ordered categorical data, there is an alternate and equivalent way to conceptualize and obtain the usual cumulative generalized linear model. The foundation of the concept lies in the natural assumption of the existence of an underlying continuous variable that is discretized to create the observed categorical response. This construction aids in model set-up, interpretation, and model fitting. In this section, we provide an introduction to the use of latent (unobserved) variables in the context of cumulative generalized linear models and a further discussion of the cumulative probit model in this context.

The latent variable framework assumes the existence of an underlying continuous variable associated with each response, where the mean of the continuous distribution is modeled as a linear function of the covariate vector for a particular site. We observe only a discretized observation from the continuous distribution. The common effect, $\boldsymbol{\beta}$, described for the models in Section 1.1.1, is motivated by such a regression model for a latent continuous variable. We refer to the latent variable as Z (Agresti, 2002),

and assume that Z has cumulative distribution function $F(z-\mu)$ where μ is a location parameter that depends on the covariates through $\mu(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta}$. We define a vector of cut-points, $-\infty = \theta_0 < \theta_1 < \ldots < \theta_K = \infty$ such that $Y = k$ if $\theta_{k-1} \leq Z < \theta_k$. Therefore, the probability that the response for a particular site falls into category k can be written as

$$Pr(Y_i = k|\mathbf{x}_i) = Pr(\theta_{k-1} \leq Z_i < \theta_k|\mathbf{x}_i) \quad (1.7)$$

$$= F(\theta_k - \mathbf{x}_i\boldsymbol{\beta}) - F(\theta_{k-1} - \mathbf{x}_i\boldsymbol{\beta}), \quad (1.8)$$

where F is the cdf of the link function. Thus, if $Z = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the link function for a cumulative model is defined by specifying the cdf of $\boldsymbol{\epsilon}$. For example, if F is the standard normal cdf, then $F^{-1} = \Phi^{-1}$ is the link function and we obtain the cumulative probit model. That is, the probit model assumes that the underlying latent variable, Z, is normally distributed, and that the parameters can be interpreted in terms of this latent variable (Agresti, 2002). Alternatively, if F is the cdf of the standard logistic distribution then F^{-1} is the logit link and we obtain the proportional odds model previously described. The proportional hazards (or complementary log-log link) model arises from an underlying extreme value distribution for Z.

Use of the standard normal cdf makes the probit model analytically and computationally attractive. Connecting the categorical response to the latent variable potentially allows for more direct interpretation of the model parameters. That is, when the latent variable has clear practical meaning, the parameters can be interpreted in terms of the continuous latent variable, Z, a context in which most researchers are more comfortable interpreting results. Specifically, for the probit model with one covariate, we can conclude that a 1-unit increase in the covariate, X, is associated with an increase in the expected value of the latent variable, $E(Z)$, of β. If the variance of the normal distribution is specified at a value different from 1, then we conclude that a 1-unit increase in X corresponds to a β standard deviation

increase in $E(Z)$ (Agresti, 2002). It should be noted that the value of β is invariant to the choice of cut-points used to discretize the latent variable scale. This is an attractive property making it possible to compare results obtained from studies using different definitions for categorizing a common variable of interest.

The latent variable formulation not only provides insight into the model structure used for ordered categorical data, but it also opens a convenient avenue for fitting the models in a Bayesian framework. The technique involves sampling the latent variable, Z, and falls under the computational method of *data augmentation* (Tanner and Wong, 1987). Likewise, the cumulative probit model provides a convenient and logical starting point for fitting such models due to its use of the standard normal distribution.

1.1.3 Maximum likelihood estimation for cumulative link models

The model given in Equation 1.1 contains $p + K - 1$ parameters to estimate, $(\theta_1, \ldots, \theta_{K-1}, \beta_1, \ldots, \beta_p)$ or equivalently, $(\theta_2, \ldots, \theta_{K-1}, \beta_0, \beta_1, \ldots, \beta_p)$ since both θ_1 and β_0 cannot both be uniquely estimated. For traditional maximum likelihood (ML) estimation, the likelihood is expressed in terms of the cumulative probabilities, and then iterative numerical methods, such as Fisher scoring algorithms, are used to obtain ML estimates (McCullagh and Nelder, 1989; Agresti, 2002).

Cumulative link models can alternatively be viewed as multivariate GLMs (Agresti, 2002; McCullagh, 1980), of which we give a brief description. The multivariate GLM expresses $\mathbf{M}_i = E(\mathbf{U}_i)$ where $\mathbf{U}_i = (U_{i1}, U_{i2}, \ldots, U_{iK})$ is a vector of binary indicators of category for observation i, denoted by $\mathbf{y}_i$ after the data are observed. That is, if the ith observation falls in category k, then the observed $\mathbf{u}_i$ is a vector with 1 in the kth position and zeros in the remaining positions. The general GLM is written as $\mathbf{h}(\mathbf{M}_i) = \mathbf{x}_i\boldsymbol{\beta}$, where $\mathbf{h}$ is a K dimensional vector of link functions. For nominal categorical data, the mean vector is given by probabilities of belonging

to each category, $\mathbf{M}_i = (p_1(\mathbf{x}_i), p_2(\mathbf{x}_i), \ldots, p_{K-1}(\mathbf{x}_i))'$. This multivariate model can, however, also be expressed in terms of cumulative probabilities where we denote the vector of cumulative probabilities as $\boldsymbol{\eta}_i = (\eta_1(\mathbf{x}_i), \eta_2(\mathbf{x}_i), \ldots, \eta_{K-1}(\mathbf{x}_i))'$. Then, for example, the cumulative logit model can be written as

$$h_k(\boldsymbol{\eta}_i) = \log\left[\frac{\eta_k(\mathbf{x}_i)}{1-\eta_k(\mathbf{x}_i)}\right] = \theta_k + \mathbf{x}_i\boldsymbol{\beta}. \tag{1.9}$$

McCullagh (1980) demonstrates that treating cumulative models as multivariate extensions of generalized linear models leads to a straightforward method of obtaining maximum likelihood estimates with rapid convergence of the Newton-Raphson method with Fisher scoring. While a unique maximum is not guaranteed, McCullagh (1980) does not observe problems with multiple modes and/or local maxima, and concludes that non-convergence is usually a reliable indicator of fitting an inappropriate model. McCullagh's method applies also to a multiplicative model used to relax the assumption of constant variance of the underlying distributions and is given by

$$F^{-1}(\eta_{ik}) = \frac{(\theta_k - \mathbf{x}_i\boldsymbol{\beta})}{\tau_i}. \tag{1.10}$$

The idea of specifying models to relax the assumption of constant variance is described further in the following section.

1.1.4 Alternative modeling strategies

When the assumption of a common regression coefficient across categories does not appear to hold, there are alternative models to consider. An obvious extension of the parallel lines model is allowing the $K-1$ regression lines to have different slopes (McCullagh and Nelder, 1989). This model is specified as

$$F^{-1}[Pr(Y \leq k|\mathbf{x})] = \theta_k - \mathbf{x}\boldsymbol{\beta}_k. \tag{1.11}$$

McCullagh and Nelder (1989) point out that a disadvantage of this non-parallel regression model is that the lines must then intersect eventually, potentially leading

to some negative fitted values. Similarly, this model violates the order constraints for the cut-points and the cumulative probabilities (Agresti, 2002). McCullagh and Nelder (1989) maintain that this may not be a fatal flaw in the model as long as the intersections occur in a remote region of the $\mathbf{x}$-space. Agresti (2002) suggests that the simpler parallel lines model should not be automatically discarded based on the significance of a hypothesis test of the assumption, maintaining the importance of investigating practical as well as statistical significance and also adhering to the principle of parsimony.

A non-generalized linear model that incorporates dispersion parameters can also be considered as an alternative to the parallel regressions model. This is a viable alternative since violation of the parallel assumption often results when the dispersion varies with $\mathbf{x}$. In other words, the responses may vary around the same location, but more dispersion occurs at the value of $\mathbf{x}_1$ than $\mathbf{x}_2$ (Agresti, 2002). For example, we may observe $Pr(Y \leq k|\mathbf{x}_1) \leq Pr(Y \leq k|\mathbf{x}_2)$ for the lower categories, while observing $Pr(Y \leq k|\mathbf{x}_1) \geq Pr(Y \leq k|\mathbf{x}_2)$ for the higher categories. As described in McCullagh and Nelder (1989), we can assume that the latent variable, Z, has a continuous distribution with mean $\mathbf{x}\boldsymbol{\beta}$ and variance $\exp(\mathbf{x}\tau)$, instead of the usual unit variance. This is equivalent to specifying the link function (an inverse cdf) based on the distribution of $(Z - \mathbf{x}\boldsymbol{\beta})/\exp(\mathbf{x}\tau)$, or writing the model in the form

$$F^{-1}[Pr(Y \leq k|\mathbf{x})] = \frac{\theta_k - \mathbf{x}\boldsymbol{\beta}}{\exp(\mathbf{x}\boldsymbol{\tau})}. \tag{1.12}$$

In (1.12), $\mathbf{x}\boldsymbol{\beta}$ is the usual linear predictor for the mean and $\mathbf{x}\boldsymbol{\tau}$ is the linear predictor for the variance (or dispersion), describing how the dispersion depends on $\mathbf{x}$. This model constrains the slopes to differ in a systematic fashion, as opposed to the non-systematic specification in (1.11). For example, using the logit link, the model allows a systematic increase or decrease of the odds ratio in k, and is thus useful for testing the proportional odds assumption (McCullagh and Nelder, 1989) without violating

the order constraints for the cumulative probabilities. When $\boldsymbol{\tau} = \mathbf{0}$, the original model is obtained.

Other possible alternatives include fitting a link function with a non-symmetric response curve, such as the complementary log-log link function, adding additional terms to the linear predictor, or permitting separate coefficient parameters for only some of the explanatory variables (Agresti, 2002).

Alternatives to cumulative link models are the *adjacent-category* (or *sequential* model), and the *two-step* model. Both alternatives are fit using Bayesian methods in Albert and Chib (1997). Agresti (2002) states that the choice of model should primarily be based on whether one is interested in interpreting effects in terms of the grouping of categories, as in the cumulative link models, or in terms of the individual categories, as in the adjacent-category model. As previously mentioned, the cumulative model has the advantages of approximate invariance of effect estimates to the construction of response categories and the more obvious connection to an underlying continuous latent variable (Agresti, 2002).

1.1.5 Incorporating random effects in a generalized linear mixed model

Generalized linear mixed models (GLMMs) are simply extensions of generalized linear models (GLMs) that incorporate unobservable random effects to account for additional sources of variability. Traditional GLMMs contain the usual linear predictor of the GLM along with one or more random components, where the random components are assumed to follow a normal distribution. Lee and Nelder (1996) extend the concept in their definition of hierarchical GLMs (HGLMs), allowing the random components to come from arbitrary, non-normal distributions, and thus providing a broader class of models. Further extensions have included models for correlated data, such as those we describe in Section 1.4.5. Models for binary and Poisson data with spatial dependence have also been developed, such as the generalized linear spatial

model described by Diggle et al. (1998), and the ensuing work by Christensen et al. (2000). These models for non-Gaussian spatial data are described in further detail in Section 1.4.

Traditional random effects models for multi-category data use a multivariate GLMM which is a natural extension of the multivariate GLM described previously. However, the likelihood function for these models is not available in closed form, and fitting GLMMs is more difficult than GLMs (Agresti, 2002). Computationally intensive approximate numerical methods are used for maximum likelihood fitting of these multivariate random effect models. It is helpful to take a tiered view of a GLMM, where at the first stage the observations are assumed to be independent given the random effects and thus fall under the definition of a GLM, and at the second stage, the random effects are assumed to be independent and normally distributed (Agresti, 2002). The random effect term then appears as a fixed offset term in the first stage specification. Estimates are typically obtained through the use of pseudo-likelihood methods. See Agresti (2002) for more detail regarding proposed methods to obtain maximum likelihood estimates. Agresti (2002) suggests that further research is needed on GLMMs, for model-fitting, inference, and especially model checking and diagnostics.

The two-stage structure of the GLMM described in the previous paragraph lends itself nicely to hierarchical Bayesian models, which are introduced in Section 1.2. We acknowledge that in a pure Bayesian analysis, all parameters are defined as random variables, versus fixed but unknown parameters, potentially causing criticism of the use of the term *mixed* to refer to a Bayesian model. However, we continue to use the traditional language of *fixed* effects and *random* effects, as we feel these are needed in our discussion of mixed models. Schabenberger and Gotway (2005) generally refer to parameters that pertain to *all* experimental units as *fixed effects* and those that differ among experimental units as *random effects*. We agree with their conclusion

that this distinction is important to the specification of mixed models regardless of whether one is using a Bayesian or frequentist analysis. Furthermore, authors such as Christensen et al. (2000) do refer to "fitting GLMMs in a Bayesian framework." The main difference between a traditional two-stage GLMM specification and a hierarchical Bayesian GLMM approach is that the parameters specified in the second stage of the hierarchy are assigned a prior distribution in the Bayesian approach (Schabenberger and Gotway, 2005). The use of a flat prior distribution results in a posterior distribution that is proportional to the likelihood and thus the mode of the posterior distribution will approximate the maximum likelihood estimate (Agresti, 2002).

1.2 Introduction to Bayesian inference

Bayesian inference proceeds by summarizing a joint probability distribution of all unknown model parameters given an observed set of data (Gelman et al., 2004). The joint probability distribution is termed the joint posterior distribution and reflects the uncertainty in the model parameters after observing the data. The set of unknown parameters can also include unobserved quantities such as predicted values at new locations. Gelman et al. (2004) summarize the process of Bayesian data analysis in three steps: setting up a *full probability model* as a joint distribution for all observable and unobservable quantities; calculating and interpreting the posterior distribution by conditioning on the observed data; and finally evaluating model fit and the implications of the posterior distribution.

For this introduction, we refer to the vector of observed data as $\mathbf{y}$ and the vector of all unknown parameters as $\boldsymbol{\eta}$. The key to Bayesian inference is the conditioning on the observed data, with the final goal being to make probability statements about $\boldsymbol{\eta}$ given $\mathbf{y}$. Traditional statistical methods instead proceed by comparison of procedures used to estimate $\boldsymbol{\eta}$ over the distribution of all possible values of the data conditional

on the unknown true value of $\boldsymbol{\eta}$ (Gelman et al., 2004). As previously mentioned, the Bayesian analysis starts with a model to provide the joint probability distribution for the model unknowns, $\boldsymbol{\eta}$, and the known data, $\mathbf{y}$. Application of Bayes' rule allows this joint probability distribution to be written as the product of a marginal distribution of the unknown parameters and the conditional distribution of the data given the parameters,

$$f(\boldsymbol{\eta}, \mathbf{y}) = f(\boldsymbol{\eta}) f(\mathbf{y}|\boldsymbol{\eta}). \tag{1.13}$$

Another application of Bayes' rule gives the joint posterior density as

$$f(\boldsymbol{\eta}|\mathbf{y}) = \frac{f(\boldsymbol{\eta}, \mathbf{y})}{f(\mathbf{y})} = \frac{f(\boldsymbol{\eta}) f(\mathbf{y}|\boldsymbol{\eta})}{f(\mathbf{y})}. \tag{1.14}$$

Since the observed data, $\mathbf{y}$, are fixed, $f(\mathbf{y})$ is simply considered a normalizing constant. This allows (1.14) to be equivalently written in its unnormalized form as proportional to "the prior times the likelihood," *i.e.*

$$f(\boldsymbol{\eta}|\mathbf{y}) \propto f(\boldsymbol{\eta}) f(\mathbf{y}|\boldsymbol{\eta}). \tag{1.15}$$

We also must define the posterior predictive distribution used to make inference about an unknown, though observable quantities, such as predictions for new observations. Gelman et al. (2004) define the prior predictive distribution as the marginal distribution of $\mathbf{y}$ *before* the data are observed,

$$f(\mathbf{y}) = \int f(\boldsymbol{\eta}) f(\mathbf{y}|\boldsymbol{\eta}) d\boldsymbol{\eta}. \tag{1.16}$$

Then, after the data have been collected let $\tilde{\mathbf{y}}$ refer to unknown, though observable, values. The distribution of $\tilde{\mathbf{y}}$ given the observed data is termed the posterior predictive distribution and is written as

$$f(\tilde{\mathbf{y}}|\mathbf{y}) = \int f(\tilde{\mathbf{y}}, \boldsymbol{\eta}|\mathbf{y}) d\boldsymbol{\eta} = \int f(\tilde{\mathbf{y}}|\boldsymbol{\eta}) f(\boldsymbol{\eta}|\mathbf{y}) d\boldsymbol{\eta}. \tag{1.17}$$

The last equality holds if $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ are conditionally independent given $\boldsymbol{\eta}$. Gelman et al. (2004) point out that the posterior predictive distribution is the average of the conditional predictions over the joint posterior distribution of $\boldsymbol{\eta}$, as is clear by (1.17).

On a more practical and fundamental note, Bayesian inference uses probability as the fundamental measure of uncertainty and describes the state of knowledge about something unknown in terms of probability distributions. Gelman et al. (2004) provide a stimulating discussion of what is meant by a numerical measure of uncertainty and why probability may be a reasonable way to quantify uncertainty.

Our reasons for choosing to apply Bayesian methods to this research problem center around the paradigm's automatic incorporation and quantification of uncertainty in all model parameters, even for complicated models. We feel this is an important advantage when dealing with hierarchical spatial models, and something that has been largely side-stepped in classical geostatistical models where the uncertainty in estimating the correlation function parameters is generally ignored. Gelman et al. (2004) also point to the increased emphasis on interval estimation over hypothesis testing in applied Statistics as a further advantage of Bayesian inference since it allows for the common-sense interpretation of an uncertainty (or probability) interval as a region having high probability of containing the unknown quantity.

1.2.1 Bayesian hierarchical models

Hierarchical models are characterized by the definition of multiple levels or tiers, typically involving the assumption of conditional independence of the observed data given certain parameters. Such models specify a joint probability model that reflects the inherent dependencies between the parameters and are thus particularly useful when multiple parameters are related in some way defined by the structure of the problem. For example, as mentioned in Section 1.1.5, GLMMs can be viewed as a two-stage hierarchical model. The first stage specifies how the data depend on fixed

effects, such as those involving a covariate, while the second stage specifies the distribution of the random effects. This two-stage structure allows for the assumption that the data are independent given the random effects. This conditional independence can greatly simplify the likelihood function and thus facilitate model fitting. In a Bayesian context, such hierarchical models further involve the specification of prior distributions for the parameters of the distribution assigned to the random effects. Inference is then based on the joint posterior distribution of all parameters given the observed data. For the purpose of this dissertation, we only refer to hierarchical spatial models where the random effects are assigned a multivariate normal distribution with covariance structure specified to capture the spatial dependence in the data. An introduction to spatial modeling is given in Section 1.4.

1.2.2 The latent variable formulation and data augmentation

The use of latent variables to facilitate computation falls under the broader heading of data augmentation. The concept was historically used in the EM algorithm to solve maximum likelihood problems until Tanner and Wong (1987) extended the idea to simplify calculation of posterior distributions for a Bayesian analysis. We follow the description of Tanner and Wong (1987) in introducing the concept.

We denote the posterior distribution of interest as $f(\boldsymbol{\eta}|\mathbf{y})$, and augment the observed data, $\mathbf{y}$, by an unobservable latent variable, $\mathbf{Z}$. The augmented posterior distribution, $f(\boldsymbol{\eta}|\mathbf{y},\mathbf{z})$, is calculated assuming that both $\mathbf{Y}$ and $\mathbf{Z}$ are known. The posterior distribution of interest can then be obtained approximately by averaging the augmented posterior distribution over the $\mathbf{z}$'s obtained from their predictive distribution, $f(\mathbf{z}|\mathbf{y})$. Then, the dependence between $f(\boldsymbol{\eta}|\mathbf{y})$ and $f(\mathbf{z}|\mathbf{y})$ leads to an iterative algorithm of successive substitution. The two iterated steps are: (1) generate a sample of m latent data patterns from the predictive distribution of $\mathbf{Z}|\mathbf{y}$ given the current guess of the posterior distribution of $\boldsymbol{\eta}|\mathbf{y}$, and (2) update the

estimate of $\boldsymbol{\eta}|\mathbf{y}$ as the mixture of the m augmented data posteriors. Practically, this algorithm is implemented in situations when it is difficult to sample from $f(\boldsymbol{\eta}|\mathbf{y})$, and easier to sample from $f(\boldsymbol{\eta}|\mathbf{y},\mathbf{z})$ and $f(\mathbf{z}|\boldsymbol{\eta},\mathbf{y})$. van Dyk (2002) similarly points to the usefulness of data augmentation when it is easier to obtain a sample from $f(\boldsymbol{\eta},\mathbf{z}|\mathbf{y})$ than from the posterior of interest, $f(\boldsymbol{\eta}|\mathbf{y})$. He describes the strategy as obtaining a sample from $f(\boldsymbol{\eta},\mathbf{z}|\mathbf{y})$ using $f(\boldsymbol{\eta}|\mathbf{y},\mathbf{z})$ and $f(\mathbf{z}|\boldsymbol{\eta},\mathbf{y})$, and then using the posterior samples of $\boldsymbol{\eta}$ as a sample from $f(\boldsymbol{\eta}|\mathbf{y})$ after simply discarding the sample of $\mathbf{z}$.

1.2.3 Factorization of joint posterior distributions for inference

An important concept in understanding the many methods of fitting Bayesian models is the factorization of joint posterior distributions using Bayes' rule. Without loss of generality, let $\boldsymbol{\eta} = (\eta_1, \eta_2)$ such that the joint posterior distribution is $f(\eta_1, \eta_2|\mathbf{y})$. To make inference about η_1, we desire its marginal posterior distribution, obtained by integrating η_2 out of the joint posterior distribution. It is often the case that such integration of the joint posterior distribution is not analytically possible. It is, however, usually still possible to make inference by simulating realizations from the distribution of interest; factorization of the joint posterior distribution often helps accomplish this. For the two parameter illustration, suppose that we desire the marginal posterior distribution of η_1,

$$f(\eta_1|\mathbf{y}) = \int f(\eta_1|\eta_2,\mathbf{y}) f(\eta_2|\mathbf{y}) d\eta_2. \tag{1.18}$$

The trick is to factor $f(\eta_1, \eta_2|\mathbf{y})$ into $f(\eta_1|\eta_2,\mathbf{y})f(\eta_2|\mathbf{y})$ if we recognize that the factors $f(\eta_1|\eta_2,\mathbf{y})$ and $f(\eta_2|\mathbf{y})$ are relatively easy to simulate from. A sample from the desired marginal posterior distribution can then be obtained using simulation techniques (Gelman et al., 2004). A sample of $\eta_{1,1}, \eta_{1,2}, \ldots, \eta_{1,T}$ is obtained by iteratively generating a value for η_2 from $f(\eta_2|\mathbf{y})$ and then a value of η_1 from $f(\eta_1|\eta_2,\mathbf{y})$ at each of T iterations. The pairs of values, $(\eta_{1,1}, \eta_{2,1}), \ldots, (\eta_{1,m}, \eta_{2,m})$, constitute a

random sample from the joint posterior distribution after convergence, and the set of η_1's constitutes a random sample from the marginal posterior distribution of interest (Gelman et al., 2004). If, however, the factorization still results in non-standard distributions that are difficult to sample from, an alternative to obtaining a sample from the desired distribution is the use of Markov chain Monte Carlo (MCMC) methods described below in Section 1.2.4. Other alternatives include the use of approximations of the desired integral, such as importance sampling or rejection sampling (Givens and Hoeting, 2006).

1.2.4 Introduction to Markov chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC) can be used to obtain a sample from a distribution of interest by generating a dependent sample from the distribution. The method draws values of the parameter of interest from approximate distributions and then corrects the draws to better approximate (and converge to) the desired target distribution (Gelman et al., 2004). A Markov chain is a sequence of random variables such that the distribution of the next value depends only on the present value. Let t denote the iteration number of the Markov chain. Under certain conditions, if we generate $\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \ldots, \boldsymbol{\eta}^{(t)}$ from a Markov chain, then $\boldsymbol{\eta}^{(t)}$ converges in distribution to $f(\boldsymbol{\eta})$ as t approaches infinity. After a certain number of iterations, the chain should converge to a stationary distribution, meaning that the probability of the chain taking on any particular value remains the same. Thus, after convergence, any sequence of observations from the chain represents a sample from the stationary distribution (Schabenberger and Gotway, 2005), meaning that inference can then be made by summarizing the post-convergence sample from a *run* of the chain. Diggle et al. (1998) summarize the goal as setting up a Markov chain with analytically tractable transition probabilities whose stationary distribution has the desired target distribution. A necessity of employing the method for inference is

assessment of whether convergence has been obtained (Gelman, 2004; Givens and Hoeting, 2006). Also, one should be aware that the posterior samples are dependent as a direct consequence of the use of a Markov chain. It has been suggested that an approximately independent random sample can be obtained by using only every kth iteration of the chain. However, if the sequences have reached convergence, they can be used for inference directly without thinning. Gelman (2004) suggests thinning only in applications where computer storage is a problem due to a large number of parameters.

1.2.4.1 Gibbs sampling algorithm

One of the most popular methods of implementing MCMC is use of the Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990). The algorithm also has the more descriptive name of *alternating conditional sampling* (Gelman et al., 2004). We follow Schabenberger and Gotway (2005) and explain the sampler via example, returning to the two-parameter case used in Section 1.2.3. Suppose the posterior distribution of interest is $f(\eta_1, \eta_2|\mathbf{y})$ and that it is not available in closed form. Suppose also that it is possible to derive and sample from the complete (or full) conditional distributions, $f(\eta_1|\eta_2, \mathbf{y})$ and $f(\eta_2|\eta_1, \mathbf{y})$. The Gibbs sampling algorithm then proceeds by choosing a starting value for η_2, denoted $\eta_2^{(0)}$, and then iteratively sampling from the complete conditional distributions. That is, we generate $\eta_1^{(1)}$ from $f(\eta_1|\eta_2^{(0)})$, then $\eta_2^{(1)}$ from $f(\eta_2|\eta_1^{(1)})$, then $\eta_1^{(2)}$ from $f(\eta_1|\eta_2^{(1)})$, then $\eta_2^{(2)}$ from $f(\eta_2|\eta_1^{(2)})$, and so on. If a stationary distribution exists, these sequentially collected samples will, after convergence, represent samples from the desired joint posterior distribution, $f(\eta_1, \eta_2|\mathbf{y})$.

Conditional sampling algorithms are effective tools for simulating from complicated distributions and thus performing Bayesian analysis. The algorithms can, however, be computationally slow when the target distribution contains parameters

that are highly correlated. It is often the case that reparameterizations are needed to help speed convergence (Gelman et al., 2004).

1.2.4.2 Metropolis-Hastings algorithm

The pure Gibbs sampling algorithm relies on the existence of recognizable complete conditional distributions. It is often the case, however, that the complete conditional distribution of one or more parameters may not be recognizable and available in closed form. In such situations, we can use a more general technique to sample from the complete conditional distribution. One such solution is the popular Metropolis-Hastings (M-H) algorithm used to construct a Markov chain with a specific target density. The M-H algorithm is an acceptance/rejection algorithm; it relies on the use of a candidate generating density from which to draw proposed values that are then accepted or rejected based on a rule defined by a M-H ratio defined in (1.19). To introduce the M-H algorithm for sampling from a target density $f(\boldsymbol{\eta}|\mathbf{y})$, denote the candidate generating density as $f_c()$, $\boldsymbol{\eta}^{cur}$ as the current value of the parameters for the chain, and $\boldsymbol{\eta}^{cand}$ as the candidate value of the parameters of the chain. The M-H ratio provides a rule dictating whether at each iteration we retain the current value, $\boldsymbol{\eta}^{cur}$, or move to the candidate value, $\boldsymbol{\eta}^{cand}$. The ratio is calculated as

$$R = \frac{f(\boldsymbol{\eta}^{cand}|\mathbf{y}) f_c(\boldsymbol{\eta}^{cur}|\boldsymbol{\eta}^{cand})}{f(\boldsymbol{\eta}^{cur}|\mathbf{y}) f_c(\boldsymbol{\eta}^{cand}|\boldsymbol{\eta}^{cur})}, \tag{1.19}$$

and the new current value of $\boldsymbol{\eta}$ is then set to the candidate value ($\boldsymbol{\eta}^{cur} = \boldsymbol{\eta}^{cand}$) with probability $\min(R, 1)$ and otherwise the current value of $\boldsymbol{\eta}$ is retained (Gelman et al., 2004). The resulting sample after T iterations, $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_T$, has a target distribution with probability density function $f(\boldsymbol{\eta}|\mathbf{y})$. The Gibbs sampling algorithm is actually a special case of the M-H algorithm where the acceptance probability is always 1, moving the chain to the candidate value at every iteration.

An effective strategy for many problems is to build a conditional sampling Markov chain by combining the Gibbs sampler *and* the Metropolis-Hastings algorithm. As previously mentioned, if some of the complete conditional distributions cannot be sampled from directly, then we can use the M-H algorithm to sample from these distributions while using the simpler Gibbs sampler to sample from the recognizable complete conditionals. This can involve updating the parameters one at a time, as described previously, or it can involve updating several parameters together in a block.

1.3 Bayesian models for ordered categorical data

The latent variable formulation of models for ordered categorical data particularly lends itself to Bayesian inference. There is a good base of literature surrounding the use of Bayesian methods to analyze ordered categorical data for a variety of link functions and model specifications, e.g. Albert and Chib (1993, 1997); Nandram and Chen (1996); Cowles (1996); Johnson and Albert (1999); Chen and Dey (1996). In the Bayesian context, the use of the latent variable is not only a conceptual advantage, but also a computational one. Use of the latent variables improve convergence, especially in situations with dependent data (Besag and Green, 1993; Nandram and Chen, 1996). As previously described, the cumulative logit and cumulative probit models typically give similar estimates of fitted probabilities, but for Bayesian inference, the probit model has an advantage in that it allows for more efficient sampling from the posterior distribution (Johnson and Albert, 1999). Specifically, the cumulative probit model, with its use of the standard normal distribution, results in full conditional distributions that are straightforward to obtain and to sample from, and thus it is the logical starting point for fitting these models.

Bayesian methods for fitting such probit models for independent data have been proposed by authors such as Albert and Chib (1993), Cowles (1996), Nandram and

Chen (1996), and Chen and Dey (1996). The different authors fit the same general model, but build on previous results to improve the efficiency of the MCMC sampling algorithm. We briefly introduce the methods used by these authors in this section and revisit them in further detail throughout this dissertation.

Albert and Chib (1993) initially proposed a standard Gibbs sampling algorithm for the cumulative probit model. In addition to the probit model, they introduce the idea of using a t-distribution to define the link function with the degrees of freedom parameter estimated by the data. Cowles (1996) notes the inefficiency of the standard Gibbs approach in sampling the cut-point parameters and proposes a multivariate Metropolis-Hastings step for sampling the latent variable and the cut-points together in one block. She specifies a truncated multivariate normal proposal distribution for the cut-point parameters, and demonstrates improved convergence with her algorithm over that of Albert and Chib (1993). Nandram and Chen (1996) propose a reparameterization of the cut-point parameters, constraining them to the interval $[0, 1]$. This allows the use of a Dirichlet proposal distribution coupled with Cowles' multivariate Metropolis-Hastings algorithm. Albert and Chib (1997) propose a de-constraint transformation of the cut-point parameters allowing them to fall over the real line. This is coupled with the use of a multivariate-t proposal density with parameters updated as the chain proceeds using optimization of the log full conditional distribution of the transformed cut-point parameters. This is a self-targeting technique similar to a Langevin-Hastings algorithm (Givens and Hoeting, 2006), allowing use of information regarding the target distribution in specification of the proposal distribution . The mean is specified by maximizing the full conditional and the variance is specified by the negative inverse of the second derivative of the log full conditional. Similarly, Chen and Dey (1996) improve upon the algorithm of Nandram and Chen (1996) by proposing a de-constraint transformation of the cut-point parameters, along with a multivariate normal proposal density with the mean and variance specified in the same manner as that of Albert and Chib (1997).

Ishwaran (2000) proposes a reparameterization of the cumulative model to avoid the use of the latent variable and to allow the use of a leapfrog hybrid Monte Carlo method. This model also allows for covariate specific cut-points, and can use different link functions with easy modifications. However, this parameterization has not been explored for correlated data, where the latent variable formulation may afford the largest advantage.

1.4 Spatial models for point-referenced non-Gaussian data

The work in this dissertation focuses on formulating models for ordered categorical data that are point-referenced over space and exhibit spatial dependence. Thus, we give a brief introduction to spatial statistics and spatial modeling for non-Gaussian data in this section before further reviewing the literature surrounding Bayesian models for ordered categorical data. Unless otherwise noted, the results presented here can be found in an introductory text for the modeling of spatial data, such as Cressie (1993) or Schabenberger and Gotway (2005).

1.4.1 Introduction to model-based geostatistics

Geostatistics is a historical term referring to a method of spatial prediction for point-referenced spatial data. By *point-referenced*, we mean that the spatial domain of interest, D, is considered a continuous and fixed set. The continuous property of D allows us to collect observations at *any* fixed location, $\mathbf{s}$, in D. In other words, infinitely many locations can theoretically be placed between any two sample locations, $\mathbf{s}_i$ and $\mathbf{s}_j$ (Schabenberger and Gotway, 2005). Note that the continuity is attached to the domain of interest, before a quantity to be measured at each location is even defined. Thus, in this dissertation, the spatial domain is considered continuous, while the response variable observed at locations within the spatial domain is considered categorical.

The model-based approach incorporates stochasticity through the definition of a random process, $\mathbf{Y}$, indexed by $\mathbf{s}$, over the domain, D,

$$\{\mathbf{Y}(\mathbf{s}) : \mathbf{s} \in D \subset \Re^d\}. \tag{1.20}$$

We also assume $d = 2$ and refer to the random process as a *spatial process* or *random field*, interchangeably. For our 2-dimensional problem, $\mathbf{s}_i$ is a coordinate specified by two numbers, one locating the position in the north-south direction and the other locating the position in the east-west direction. The continuity of the spatial domain often leads to the desire to predict at many new and unsampled locations and prediction is often the primary goal of modeling point-referenced spatial data. For example, the goal may be to produce a map of the response over the entire region (domain) of interest.

1.4.2 Covariance functions in the spatial setting

> Everything is related to everything else, but near things are more related than distant things.
> Tobler's *first law of geography* (Tobler, 1970 as given in Schabenberger and Gotway, 2005)

Correlation functions for point-referenced spatial data reflect the assumption that the correlation between the values at two locations typically decreases as the distance between two locations increases. The covariance function of a spatial process is defined in terms of a distance, $\mathbf{h}$, between two locations,

$$Cov[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})] = E\left([Y(\mathbf{s}) - \mu(\mathbf{s})][Y(\mathbf{s}+\mathbf{h}) - \mu(\mathbf{s}+\mathbf{h})]\right), \tag{1.21}$$

where $\mu(\mathbf{s})$ is the expected value of $Y(\mathbf{s})$. The autocorrelation function is then defined in the usual way,

$$Corr[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})] = \frac{Cov[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})]}{\sqrt{Var[Y(\mathbf{s})]Var[Y(\mathbf{s}+\mathbf{h})]}}. \tag{1.22}$$

Similar to the analysis of autocorrelated data in time series analysis, it is common to make stationarity assumptions about the specified covariance function. For example, it is often assumed that the important information for defining the correlation between values at two locations lies merely in the distance between the locations, and *not* in the the actual coordinates. This, combined with the assumption of a constant mean across the region, is termed *second-order stationarity.* Two important implications of this assumption are that the covariance function can be written as a function of the distance only,

$$Cov[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})] = C(\mathbf{h}), \tag{1.23}$$

and the variance, $Var[Y(\mathbf{s})] = \sigma^2$ is not a function of spatial location, meaning that the correlation can be written as

$$Corr[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})] = \frac{Cov(\mathbf{h})}{\sigma^2}. \tag{1.24}$$

We further discuss stationarity in the Section 1.4.3.

Our use of model-based geostatistics requires definition of several other terms, most notably the semi-variogram. The semi-variogram is an integral tool used in classical geostatistics to estimate a correlation function from the data, and is defined as

$$\begin{aligned} \gamma(\mathbf{s},\mathbf{h}) &= \frac{1}{2}\left(Var[Y(\mathbf{s}) - Y(\mathbf{s}+\mathbf{h})]\right) \\ &= \frac{1}{2}\left(Var[Y(\mathbf{s})] + Var[Y(\mathbf{s}+\mathbf{h})] - 2Cov[Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})]\right). \end{aligned} \tag{1.25}$$

Under the assumption of second-order stationarity, the semi-variogram can be written as

$$\gamma(\mathbf{h}) = \frac{1}{2}\left(2\sigma^2 - 2C(\mathbf{h})\right) = C(\mathbf{0}) - C(\mathbf{h}). \tag{1.26}$$

There are several characteristic features of the semi-variogram that warrant definition and are shown in Figure 1.1. The semi-variogram is a continuous and

increasing function starting from the origin and rising smoothly to the value of σ^2. This maximum value is termed the *sill.* The shape of the function is defined by the smoothness and continuity of the spatial process (Schabenberger and Gotway, 2005). The *range* of the spatial process is the lag distance, h_r, such that the covariance between values of two points separated by h_r is 0. If a covariance function is defined such that it only asymptotically achieves the sill, we instead define a *practical range* or *effective range* as the lag distance at which the semi-variogram achieves 95% of the sill, σ^2. It is also common to define a discontinuity at the origin, termed the *nugget.*

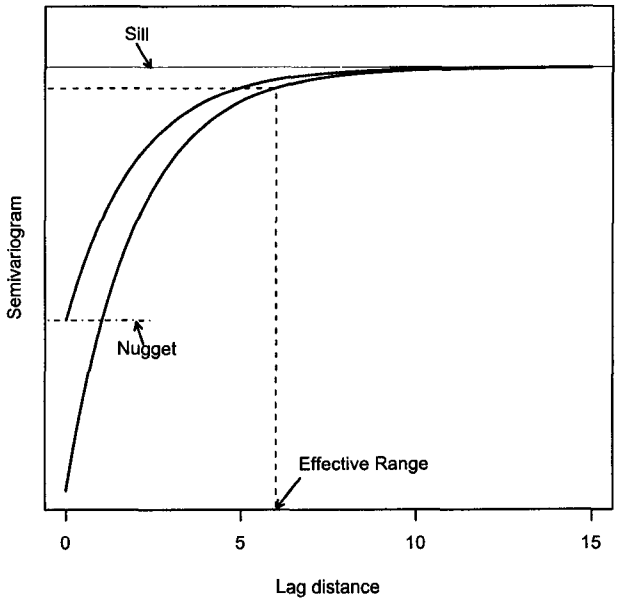

Figure 1.1: This figure, similar to Figure 1.12 in Schabenberger and Gotway (2005), displays the characteristic features of a semi-variogram, namely the *sill, effective range,* and *nugget.*

We use the term *nugget* as a general term to describe any source of variability that would result in a semi-variogram that does not pass through the origin, though there are two main explanations for a practical meaning of the nugget: micro-scale

variation and measurement error (Cressie, 1993). Micro-scale variation is described as that occurring at a lag distance less than the smallest lag observed in the data. Of practical importance is the fact that without replication the variability associated with these two sources cannot be separately estimated and thus the nugget term is usually said to include both. However, it is important to distinguish between the two sources using practical (non-statistical) knowledge because the two sources result in different best spatial predictors (Schabenberger and Gotway, 2005). Even combining the two possible sources, there is difficulty estimating this variance due to the usual lack of replication *and* lack of measurements at "micro" distances. While it is true that empirical semi-variograms often exhibit discontinuities at the origin, there is ensuing discussion regarding the theoretical and practical sense of including a nugget and specifying its source (Schabenberger and Gotway, 2005). Schabenberger and Gotway (2005) point out that mean-square continuity of a spatial process is not compatible with the existence of nugget, making it impossible to apply spatial analyses requiring a continuous covariance function, such as those using the spectral representation and some non-parametric techniques. Banerjee et al. (2004) use the nugget effect only to represent pure error, described as measurement error, or more generally, "noise."

1.4.3 Use of spatial processes and stationarity assumptions

Spatial models for point-referenced data rely on the assumptions of the defined spatial processes. A stochastic process, in general, is a collection of indexed random variables, such as a time series indexed by time or a spatial process indexed by location. Without the use of the latent process, we would be forced to determine the joint distribution for an uncountable number of random variables since $\mathbf{s}$ varies continuously over the region of interest, D (Banerjee et al., 2004). Because a spatial process is a collection of random variables over space, one observation of a spatial

process at n locations is a collection of only one sample from each of n dimensions of the joint distribution. One realization of a spatial process can be visualized as a surface over the spatial domain of interest. One set of spatial data with no replication at the measurement sites reflects the height of the surface at n fixed locations. Thus, a spatial sample of size n is actually only the values of *one* realization of the spatial process at a finite number, n, of locations; there is no replication and it is not possible to obtain a complete sample of even one realization due to the continuity of the spatial domain. This raises the obvious question of how we can confidently make inference with such incomplete data. It is important to consider whether we are interested in long run averages of the spatial process or modeling and predicting incomplete parts of the one realization of the process, for example predicting at new locations and creating an estimated map of the realization (Schabenberger and Gotway, 2005). This is where the importance of stationarity assumptions enter the picture; they allow inference with our incomplete sample size of one (Schabenberger and Gotway, 2005). As Schabenberger and Gotway (2005) note, the stationarity assumption is often criticized since making this assumption in error can lead to invalid inferences and wrong conclusions. However, they argue that such analyses are worthwhile given that a good understanding of the analysis of stationary processes is likely a necessary precursor to the understanding and analysis of non-stationary processes. We agree with this stance.

Recall that we define the spatial process as $\{Y(\mathbf{s}) : \mathbf{s} \in D \subset \Re^d\}$ and the definition of *second-order stationarity* given in Section 1.4.2. A stronger assumption is *strict stationarity*, meaning that the spatial distribution is invariant under translation of the coordinates, implying that the random field repeats itself throughout the spatial domain, D. That is, a random field is strictly stationary if for all n and $\mathbf{h}$,

$$Pr[Y(\mathbf{s}_1) < y_1, Y(\mathbf{s}_2) < y_2, \ldots, Y(\mathbf{s}_n) < y_n] =$$
$$Pr[Y(\mathbf{s}_1 + \mathbf{h}) < y_1, Y(\mathbf{s}_2 + \mathbf{h}) < y_2, \ldots, Y(\mathbf{s}_n + \mathbf{h}) < y_n]. \qquad (1.27)$$

By contrast, second-order stationarity places assumptions on the moments of the spatial distribution, rather than on the distribution itself, and is thus a weaker, though possibly more realistic, assumption. Schabenberger and Gotway (2005) point out that this is equivalent to using only the mean and variance of a random variable in classical estimation techniques, such as least squares methods. Once again, the important characteristics of a second-order stationary spatial process are a covariance function that depends only on the lag distance between locations (not on the coordinate values themselves), along with a constant mean and variance across the domain. This is the spatial analog of a random sample of independent observations with the same mean and variance in classical statistics.

Second-order stationarity restricts the second moments of the process to a function only of the separation distances, that is

$$Cov(Y(\mathbf{s}), Y(\mathbf{s}+\mathbf{h})) = C(\mathbf{h}), \tag{1.28}$$

disregarding the absolute coordinates of the locations. This restriction on the second moment structure does not, however, restrict the covariance function to depend only on separation distance; it can also depend on direction. A covariance function that does *not* depend on direction is termed *isotropic* and one that does depend on direction is termed *anisotropic* (Schabenberger and Gotway, 2005).

An even weaker stationarity assumption is *intrinsic stationarity*. For an intrinsically stationary process, the increments $Y(\mathbf{s}) - Y(\mathbf{s}+\mathbf{h})$ *are* second-order stationary even though $Y(\mathbf{s})$ is not. An intrinsically stationary process is typically defined as having a constant mean and

$$\gamma(\mathbf{h}) = \frac{1}{2} Var[Y(\mathbf{s}) - Y(\mathbf{s}+\mathbf{h})]. \tag{1.29}$$

If a process is intrinsically stationary, but not second-order stationary, then the covariance between $Y(\mathbf{s})$ and $Y(\mathbf{s}+\mathbf{h})$ is not estimable.

Spatial models are typically confined to Gaussian processes and mixtures of Gaussian processes for technical concerns. A spatial process is a Gaussian random field if the joint distribution of the random variables at a set of n locations is a n-dimensional multivariate Gaussian (normal) distribution,

$$Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots Y(\mathbf{s}_n) \sim MVN_n. \tag{1.30}$$

From standard Gaussian distribution theory, this further implies that the marginal distribution of any $Y(\mathbf{s}_i)$ is univariate Gaussian. The many advantages of the Gaussian distribution for classical statistical analysis also hold for spatial analysis using Gaussian processes. Namely, the Gaussian assumption allows for convenient distributional theory since all marginal and joint distributions can be easily obtained once the mean and covariance structure have been specified. Also, in this context, best linear unbiased predictors (BLUPs) for unobserved locations are the best predictors under squared error loss and second-order stationarity implies strict stationarity. Banerjee et al. (2004) further identify that the Gaussian assumption for random effects specified at the second stage of a hierarchical model is consistent with the way that independent random effects with variance components are typically specified in mixed models. Finally, it is hard to criticize the Gaussian assumption since we cannot, without replication, simply test such an assumption. Quoting Banerjee et al. (2004), "With a sample size of one, how can we criticize *any* multivariate distributional specification (Gaussian or otherwise)?" It is, however, possible to use simulation-based methods to investigate the effects of this assumption.

To be practical, the spatial process definition must be used in conjunction with a framework that facilitates statistical inference and the study of estimators, predictors, and the random process itself. Schabenberger and Gotway (2005) describe two possible alternatives, one as the model representation in the spatial domain and one as the spectral representation in the frequency domain. This dissertation focuses on

modeling in the spatial domain with the goal of creating mathematical representations of a data-generating mechanism. As is usual, we think of a statistical model as containing two parts: a structure describing the mean and a structure describing the variation and co-variation of and between the values of the response. As Schabenberger and Gotway (2005) point out, the decomposition is particularly important for the use of second-order stationary random fields. In particular, we can by-pass the need for non-stationary random fields in the presence of a spatially varying (non-constant) mean by modeling the mean structure and the variance-covariance structure separately. This is the notion of de-trending. The stationarity assumption can then be made for the error terms only, after accounting for the spatially varying mean. The hierarchical spatial model with the spatial effects modeled as a zero-mean Gaussian process at the second stage of the hierarchy provides a more indirect method of accomplishing this same goal.

1.4.4 Spatial models for non-Gaussian data

Toward our goal of formulating a model for point-referenced ordered categorical spatial data, we briefly review the spatial models available in the literature for non-Gaussian data. The conventional spatial methods applied when the data can be modeled as Gaussian, such as kriging with Gaussian errors, clearly do not apply to ordered categorical data where the assumption of normality cannot be met through simple transformations. Our strategy in formulating the models proposed in this dissertation is to build upon ideas and methods from the simpler case of spatial models for binary and count data and to combine them with models and methods proposed for independent ordered categorical data. Strategies for fitting models for binary and count data rely on the GLMM framework, coupled with the use of Bayesian inference and novel MCMC strategies to obtain and/or improve convergence of the algorithms. Binary and count data conveniently fall within the one-parameter

exponential family of distributions, making it easier to develop models for this data versus ordered categorical data.

The hallmark paper describing model-based geostatistics using spatial GLMMs is Diggle et al. (1998). They introduce *generalized linear spatial models* by embedding linear kriging methodology within the GLMM framework, as opposed to the trans-Gaussian kriging approach of Cressie (1993) that essentially analyzes non-linearly transformed data. In typical GLM fashion, Diggle et al. (1998) set a function of the mean, M_i, equal to a linear predictor,

$$h(M_i) = \mathbf{x}_i\boldsymbol{\beta} + W(\mathbf{s}_i), \tag{1.31}$$

for covariates $\mathbf{x}_i = \mathbf{x}(\mathbf{s}_i)$. They additionally define an underlying zero-mean Gaussian spatial process, $\mathbf{W}$. More specifically, for the n observed locations, $\mathbf{W}_n \sim MVN(\mathbf{0}, \sigma^2 H(\boldsymbol{\psi}))$ where $[H(\boldsymbol{\psi})]_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$ and ρ is a valid correlation function such that $Corr(W(\mathbf{s}_i), W(\mathbf{s}_j)) = \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$. The values of the spatial process, $\mathbf{W}(\mathbf{s}_1), \mathbf{W}(\mathbf{s}_2), \ldots, \mathbf{W}(\mathbf{s}_n)$, appear as an offset term in the linear predictor, as shown in (1.31). Conditional on $\mathbf{W}$, the Y_i are considered mutually independent with distributions $f_i(y|W(\mathbf{s}_i)) = f(y; M_i)$. It is now clear that for this model, M_i is equal to the conditional mean, $E[Y_i|w(\mathbf{s}_i)]$. The $E[W(\mathbf{s})|\mathbf{y}]$ is termed the *generalized linear predictor* for $W(\mathbf{s})$ (Diggle et al., 1998).

The posterior distribution for the model defined by Diggle et al. (1998) is analytically intractable, leading to the use of Markov chain Monte Carlo (MCMC) methods to simulate from the posterior and posterior predictive distributions. They employ Bayesian inference and note its ease of incorporating parameter uncertainty into the predictions for the spatial process. They implement a Gibbs sampling scheme, sampling in turn from the complete conditional distributions of spatial parameters, the regression parameters, and the spatial process associated with each observation, using the M-H algorithm at each step. They adopt proper uniform priors throughout

to ensure that the joint posterior distribution is proportional to the likelihood surface. Their methods are specifically formulated for data distributions that belong to the one-parameter exponential family, and are thus not directly applicable to the situation of ordered categorical data.

Gelfand et al. (2000) describe a model for binary data that fits within the generalized linear spatial model framework of Diggle et al. (1998), specifying a hierarchical model for point-referenced binary spatial data with spatial random effects at the second stage introduced through a stationary latent Gaussian process. They define a binary stochastic process over the region of interest, D, and then use the concept of a latent variable and a cut-point, as previously described, to model the data. Recall the usual deterministic relationship defined between the continuous latent variable and the categorical response, namely $Y(\mathbf{s}_i) = k$ if $\theta_{k-1} < Z(\mathbf{s}_i) \leq \theta_k$. The latent variable is defined for location i as, $Z(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta} + W(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)$ where $\mathbf{W} = (W(\mathbf{s}_1), W(\mathbf{s}_2), \ldots, W(\mathbf{s}_n))'$ is a $n \times 1$ vector from a realization of a latent second-order stationary Gaussian spatial process over D, defined identically to that described above for Diggle et al. (1998), and $\epsilon(s_i)$ are iid $N(0, \tau^2)$. Then, the $Z(s_i)$ are conditionally independent given $\mathbf{W}$ and $\boldsymbol{\beta}$ such that $\mathbf{Z}|\boldsymbol{\beta}, \mathbf{W} \sim N(\mathbf{x}\boldsymbol{\beta} + \mathbf{W}, \tau^2\mathbf{I})$, where $\mathbf{I}$ is a $n \times n$ identity matrix. This naturally parameterizes a hierarchical model with the conditionally independent $\mathbf{Z}$ at the first stage, $\boldsymbol{\beta}$ and $\mathbf{W}$ modeled at the second stage, and the parameters of the Gaussian spatial process, σ^2 and ψ, modeled at the third stage. Gelfand et al. (2000) focus on the fact that the matrix inversion needed for evaluating the density of the spatial process can be very slow and can result in serious rounding errors. The main contribution of their paper is their proposed use of importance sampling to replace matrix inversion (see Givens and Hoeting (2006) for a review of the method of importance sampling.)

Gelfand et al. (2000) additionally provide an important background discussion of the advantages of modeling the spatial dependence using a latent Gaussian process

and a hierarchical structure. As they point out, this method contrasts with the historically more common method of using a Markov random field specification where the covariance structure is created using conditional distributions based on defined neighborhood structures (Besag, 1974; Geman and Geman, 1984). Some disadvantages of the Markov random field method are that the covariance matrix is typically singular and the density for the spatial random effect is not proper, and obviously not Gaussian (Gelfand et al., 2000). Gelfand et al. (2000) suggest that in many applications, it may be more appropriate to model spatial dependence continuously, in the spirit of classical geostatistics. They also discuss the non-identifiability of the white noise variance, τ^2, and discuss reasons for fixing the parameter at a value of 1. This is an assumption we also adopt and present in more detail in Chapter 3.

Other modeling strategies have been proposed in the context of the generalized linear models for both univariate and multivariate data. Some of these even focus on spatial dependence, but none deal with ordered categories. For example, see the work of Gotway and Stroup (1997); Vounatsou et al. (2000); Zhu et al. (2005a); Paciorek and Ryan (2005); Christensen et al. (2006); Zhao et al. (2006). Here we briefly describe the most recent work by Christensen et al. (2006) and Zhao et al. (2006).

Christensen and others (Christensen et al., 2000; Christensen and Waagepetersen, 2002; Christensen et al., 2006) propose "robust MCMC methods" for combining spatial generalized linear mixed models with Bayesian MCMC approaches. Achieving good mixing and convergence of MCMC algorithms can be very challenging due to the strong dependence and high dimensionality inherent of spatial data sets. Thus, Christensen and others work to develop efficient algorithms, particularly ones that can be applied easily and reliably to more than one data set. They examine the use of Bayesian inference for GLMMs with spatially correlated random effects for binary and count data.

Christensen et al. (2000) extend the work of Diggle et al. (1998) by proposing the use of Langevin-Hastings updates instead of single-site Metropolis-Hastings updates, updating sites simultaneously based on gradient information. They also investigate the extent to which flat priors can be used while maintaining proper posterior distributions. Christensen et al. (2006) address the problem that while performance of an MCMC algorithm may be adequate for one data set, it may fail for another. They work specifically to improve the performance of MCMC algorithms applied to GLMMs, and also to develop more general methods that perform well for different priors and data sets. The key to their "robust" method is the use of data-based reparameterizations, coupled with the blocked algorithm proposed in Christensen and Waagepetersen (2002). Christensen et al. (2006) use a model setup similar to Diggle et al. (1998), with the exception that they define a non-stationary Gaussian process with mean $\mathbf{X}\boldsymbol{\beta}$, instead of the usual mean-zero process. They use the typical hierarchical specification, assuming conditional independence of the Y_i's given the spatial process at n locations. Furthermore, $Y_i|W_i$ has a distribution with density with mean $E[Y_i|W_i] = m_i h^{-1}(W_i)$, where h is the link function and $m_1, m_2, \ldots, m_n$ are scalars. The variance of the Gaussian process is expressed in the typical fashion, $\boldsymbol{\Sigma} = \sigma^2 H(\boldsymbol{\psi})$, where $[H(\boldsymbol{\psi})]_{ij} = \rho(s_i - s_j; \boldsymbol{\psi})$ and ρ is a valid correlation function. Christensen et al. (2006) present examples of the most common GLMMs, the Poisson log-normal and the binomial logit-normal models, staying within the comfort zone of the one-parameter exponential family. They do state that the techniques are generalizable to other link functions and error distributions, but the generalization to ordered categorical data is not straightforward or immediately clear.

Zhao et al. (2006) use a Bayesian MCMC approach for fitting *general design generalized linear mixed models.* The key to these models is allowing the random effects design matrix to have a general structure. They suggest that the main advantage of the general design framework lies in the incorporation of nonparametric

regression through penalized regression splines and maintain that higher-dimensional extensions correspond to the generalized kriging approach proposed by Diggle et al. (1998). Zhao et al. (2006) choose the Bayesian approach for the purpose of capturing the uncertainty in the variance components, a problem that is largely intractable under frequentist distributional theory. Once again, attention is restricted to the canonical one-parameter exponential family structure, which makes direct extension to the ordered categorical case difficult. Examples are given only for binary and count data, using a Bayesian logistic additive mixed model for the binary data and a semi-parametric Poisson spatial model for the count data. For the spatial structure, they use an intrinsic Gaussian autoregressive distribution and define a neighborhood structure based on a cut-off distance. Zhao et al. (2006) give a general formula for the posterior distribution, but it pertains only to one-parameter exponential families. A detailed description of the MCMC algorithm for general design GLMMs is given in Zhao (2003).

De Oliveira (2000) moves away from the GLMM model approach and proposes an approach to more directly map a binary outcome, assuming a binary random field is created directly from the clipping of a Gaussian random field at a fixed threshold level, *i.e.* cut-point. Similar to other methods we have described, he utilizes MCMC methods, incorporating the latent Gaussian random field through data augmentation. The problem is formulated as a spatial prediction problem, with less emphasis on parameter estimation. The concept of modeling categorical data through the clipped Gaussian process allows likelihood-based inference by creating a workable mathematical framework. De Oliveira (2000) computes the probability that a particular site has the characteristic of interest given the data and then creates the binary map using these probabilities, along with calculating a measure of prediction uncertainty. The fact that his method does not rely on the properties of one-parameter exponential families makes it particularly attractive as a solution to the problem of modeling

ordered categorical data, specifically to produce maps of the predicted values for a realization of a categorical random field. De Oliveira (2000) states that the latent Gaussian random field may describe the spatial variation of an unobservable quantity of interest, or it may be used simply as a modeling device. The likelihood is given by

$$f(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\psi}, \sigma^2) = \int_{A(y_1)} \cdots \int_{A(y_n)} \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} |H(\boldsymbol{\psi})|^{-1/2} * \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})' H(\boldsymbol{\psi})^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})\right\} d\mathbf{w}, \tag{1.32}$$

where $A(y_i) = (-\infty, \theta]$ if $y_i = 0$ and (θ, ∞) otherwise. The correlation matrix is specified through the vector of parameters, $\boldsymbol{\psi}$, and denoted as $H(\boldsymbol{\psi})$ as defined earlier in this section.

Since a direct Bayesian analysis using the likelihood in (1.32) is intractable, the method of De Oliveira (2000) again appeals to data augmentation methods. In its most general specification, some parameters of De Oliveira's model are not "(likelihood) identifiable" (De Oliveira, 2000). For one, it is clear that binary data do not contain sufficient information to estimate the variance of the process, σ^2. That is, we can obtain identical binary random fields by clipping infinitely many continuous distributions defined by different variances using a single cut-point parameter. The only information contained in binary data is whether the latent observation is above or below the cut-point, and *not* the magnitude of the distance above or below the cut-point. Thus, De Oliveira (2000) fixes $\sigma^2 = 1$. We revisit this concern in the context of multi-categorical data in Chapter 3.

Additionally, De Oliveira (2000) advises caution when specifying a correlation function to avoid near non-identifiability and related convergence problems. He gives the example of considering the isotropic correlation function of $\rho(d, \boldsymbol{\psi}) = \phi_1^{d^{\phi_2}}$, where d is Euclidean distance and $\boldsymbol{\psi} = (\phi_1, \phi_2)$. The range of correlation is controlled by ϕ_1 and the smoothness (or roughness) of the random field is controlled by $\phi_2 \in (0, 2]$.

He points out that after clipping the Gaussian field, the binary data contain no information about ϕ_2 and thus we cannot realistically hope to estimate it. Thus, De Oliveira (2000) suggests fixing the smoothness parameter at an appropriate value, thus specifying a correlation function depending on only one parameter controlling the range of correlation. After fixing σ^2 and the smoothness parameter, De Oliveira estimates $\boldsymbol{\beta}$ and ϕ_1 using binary data.

De Oliveira (2000) also provides an overview of prediction of the binary variable at new locations using the Bayesian predictive distribution by formulating prediction as a decision problem. This involves specifying a loss function and finding the optimal Bayes predictor by minimizing the Bayesian expected loss. De Oliveira then uses the Bayesian expected loss of the optimal Bayes predictor as a global measure of uncertainty regarding the estimated binary map, thus providing valuable information about the map's "quality". He additionally proposes a simpler "plug-in" approach for use with large datasets that reduces computation, at the expense of ignoring the uncertainty about the model parameters and latent variable. The methods are compared to indicator kriging, which also ignores the uncertainty about correlation parameters. Interestingly, the indicator kriging predictor is optimal only among the class of linear unbiased estimators, while the Bayesian one is optimal among the class of all estimators (De Oliveira, 2000).

In his conclusion, De Oliveira (2000) suggests that his approach can be generalized to model categorical random fields by using a vector of threshold values and extending most of the properties of the clipped Gaussian random field to this more general case. He gives the optimal predictor for a new site as $\hat{Y}(\mathbf{s}_0) = \arg\max_{k\in\{0,1,\ldots,K-1\}}(Pr[Y(\mathbf{s}_0) = k|\mathbf{y}])$, where the conditional probabilities are again computed using the technique of data augmentation. He admits that the MCMC algorithm for such a model would be more complex due to the presence of $K-2$ unknown cut-point parameters.

1.4.5 Methods for other sources of correlation

While the literature is absent for spatial models for point-referenced ordered categorical data, models have been proposed for other sources of dependence, focusing on clustered data and repeated measures. For example, references such as Harville and Mee (1984); Ochi and Prentice (1984); Agresti and Lang (1993); Crouchley (1995); Chib and Greenberg (1998); Ishwaran and Gatsonis (2000); Chen and Dey (1996); Qiu et al. (2002), O'Brien and Dunson (2004); Bradlow and Zaslavsky (1999); and Lawrence et al. (2006) include both Bayesian and maximum likelihood based methods. In this section we provide a brief review of the work of Chen and Dey (1996) and Ishwaran (2000).

Chen and Dey (1996) develop a method for analysis of general correlated ordinal data models using an exact small sample Bayesian approach. They extend the work of Albert and Chib (1993), Cowles (1996), Nandram and Chen (1996) and Cowles et al. (1996) by proposing a latent variable method based on multivariate link functions using scale mixtures of normals, termed the SMMVN-link models. The model is formulated for a multivariate ordered categorical response, Y_{ij}, for the ith observation and the jth variable, indicative of data collected via repeated measures. The vector for each experimental or observational unit is given by $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ})$ where $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n$ are considered independent (an assumption that is not appropriate for spatially correlated data), while $Y_{i1}, Y_{i2}, \ldots, Y_{iJ}$ are dependent. Chen and Dey (1996) introduce n latent J-dimensional random vectors such that $Y_{ij} = k$ if $\theta_{j,k-1} \leq W_{ij} < \theta_{j,k}$, thus allowing different sets of cut-points for each response. The latent variable is assumed to be multivariate normal with mean $\mathbf{x}_i\boldsymbol{\beta}$ and variance-covariance matrix, $\kappa(\lambda)\Sigma$, where $\kappa(\lambda)$ is a positive function of a mixing variable, λ, and Σ is taken to be a $(J \times J)$ correlation matrix with entries $\rho_{j,j'}$ such that $\rho_{jj} = 1$ capturing the correlation in the Y_{ij}'s. For $\boldsymbol{\Sigma} = \mathbf{I}_J$ and $\kappa(\lambda) = 1$, the model reduces to the independent cumulative probit model described by Albert and Chib (1993).

To facilitate computation for difficult sampling problems of the cut-points and the correlation matrix, they extend a reparameterization of the cut-points suggested by Nandram and Chen (1996). Chen and Dey's reparameterization results in fewer unknown cut-points, restricts the cut-point parameters to the interval $[0, 1]$, and results in an unrestricted variance-covariance matrix for the $\mathbf{W}$ that can afford advantages in terms of implementation of the algorithm (Chen and Dey, 1996).

Ishwaran (2000) proposes a re-parameterized version of the cumulative link model. Like that of Chen and Dey (1996), the method considers correlated data from repeated measures on experimental units, but does not apply directly to point-referenced spatial data. He proposes a transformation to provide category specific cut-point parameters. The model proposed by Albert and Chib (1993) is a special case of this model, though Ishwaran (2000) does not utilize latent variable data augmentation to fit the model. He instead uses a hybrid Monte Carlo (HMC) Gibbs sampling method termed the *leapfrog* algorithm, requiring a particular parameterization of the cut-points, as well as selection of priors that ensure a differentiable posterior distribution. He concludes that the HMC method promotes efficient mixing of the chain, as opposed to the conventional MCMC methods used by Albert and Chib (1993), Cowles (1996), and others. However, the method requires computation of the gradient for each step of the algorithm, making it more computationally demanding than traditional Metropolis-Hastings (M-H) steps. Thus, Ishwaran (2000) suggests mixing the M-H algorithm steps with HMC in a preset proportion to reduce computational time, along with only starting the leapfrog algorithm after running the M-H algorithm for several updates of the chain. He concludes that the method leads to efficient mixing of the Markov chain and a more general Gibbs sampler that can be used for more complicated ordinal models than those originally proposed based on latent variable data augmentation. While the combination of M-H and leapfrog

steps may be a strategy to consider in the presence of serious convergence difficulties, increasing computational effort and time is a significant disadvantage given the already computationally intensive nature of the problem.

The models and concepts reviewed in this chapter provide the foundation from which we build our models for point-referenced spatial ordered categorical data. In the following chapters, we propose two models for ordered categorical data with spatial dependence, along with the MCMC algorithms to carry out Bayesian parameter estimation and prediction for these models. We focus on prediction of the categorical response at new unobserved locations, though also discuss estimation of the model parameters. We build on the methods for independent ordered categorical data proposed by Albert and Chib (1993) and others and the methods for spatial data proposed by Diggle et al. (1998) and De Oliveira (2000). This dissertation proceeds by describing the proposed models in detail in Chapter 2, addressing model specification in Chapter 3, discussing prediction using the proposed models in Chapter 4, investigating the models via a simulation study in Chapter 5, applying the models to real data in Chapter 6, and finally concludes with an overview of the work and a discussion of possible future directions of related research in Chapter 7.

Chapter 2

THE DOUBLE LATENT AND CLIPPED GAUSSIAN MODELS

2.1 Introduction to the proposed models

In this chapter, we propose two models for point-referenced ordered categorical data and develop the Bayesian MCMC algorithms used to estimate the model parameters and predict at new locations. Both models rely on the notion of cutting or clipping a latent continuous random variable to model the ordered categorical response. The connection between the categorical random variable and the underlying continuous random variable is defined through a deterministic relationship between the two variables. The deterministic relationship specifies that the probability of the categorical random variable taking on a specific value, say k, is equal to the probability that the continuous random variable falls within a specific interval, whose endpoints we term *cut-point* parameters. The models have conceptual differences in the level at which the clipping occurs and the specification of the mean structure, though also have fundamental similarities. This chapter introduces and compares the two models and provides detailed descriptions of the MCMC algorithms used to fit the models.

We term the first model the *Double Latent* (DL) model. This model builds on the generalized linear spatial model ideas formulated for binary and count data by Diggle et al. (1998), and is also a logical extension of the non-spatial latent variable methods for ordered categorical data described in Section 1.1.1. It is formulated under the assumption that the ordered categorical response is a direct result of clipping a

continuous distribution at specified cut-points and relates cumulative probabilities to a linear function of covariates through a link function in typical GLM fashion. Spatial random effects are introduced through the use of a latent Gaussian process and are incorporated into the linear predictor term of the model. A hierarchical formulation specifies conditional independence of the underlying continuous random variables given the values of the spatial process. Our model extends the ideas of Diggle et al. (1998) and Gelfand et al. (2000) that were applied to binary and count data, but steps outside the convenient context of one-parameter exponential families. Our proposed model has flexibility in that it readily allows for the specification of different link functions, such as the techniques described by Albert and Chib (1993). However, we focus on the probit model, a logical and common starting point for such models due to the mathematical convenience of the standard Gaussian inverse cdf link function. The conditional independence arising with the hierarchical structure facilitates calculation by simplifying the likelihood, and thus inference. Interestingly, the specification of the link function naturally incorporates the equivalent of a nugget variance parameter of a spatial model into the covariance structure of the model. The term *Double Latent* comes from the fact that this model specification requires the definition and use of two latent variables, each of dimension n. The model is described in detail in Section 2.3.

The second model is termed the *Clipped Gaussian* (CG) model, is motivated less by typical linear modeling ideas, and more by the assumption that the observed categorical data come directly from a realization of a categorical random field over the area of interest. Following the latent variable framework, we assume that this categorical random field is created by directly clipping an underlying latent Gaussian spatial process. A visual illustration of this concept is shown in Figure 2.1. The mean of the Gaussian random field is specified by the usual linear predictor term involving covariates. As described in Chapter 1, this idea was originally proposed for

binary data by De Oliveira (2000), and is extended here for ordered categorical data. Like the Double Latent model, the Clipped Gaussian model is written in terms of cumulative probabilities, but lacks the direct connection to generalized linear models since it does not require or allow specification of different link functions. Following De Oliveira (2000), the Clipped Gaussian model is not specified hierarchically and does not benefit from conditional independence, resulting in a more complicated form of the likelihood involving a multi-dimensional integral. The covariance structure of the spatial Gaussian process employed in both the Double Latent and Clipped Gaussian models is assumed *not* to include a nugget term. Recall that an additional random error term in the Double Latent model naturally incorporates a nugget-like variance. This is not the case for the Clipped Gaussian specification. It should be noted that if we were to let the spatial covariance structure include a nugget, the Clipped Gaussian model with nugget could be considered a special case of the Double Latent model with the probit link function. The Clipped Gaussian model requires only one latent variable, and is described in detail in Section 2.4.

This chapter provides the general specification of the models and the algorithmic detail. We further compare, discuss, and investigate the models in the coming chapters. For example, Chapter 3 provides a comparison of the correlation structure of the categorical random fields induced by the two model specifications, along with more practical information regarding model specification such as specification of prior distributions and hyperparameters. The primary goal of the models is prediction at new locations where covariate measurements are possibly available, and thus our main focus of comparison is predictive success on a hold-out data set. We describe the strategy and develop the theory used for prediction in Chapter 4. In Chapter 5, we use simulation and artificial data to apply the models to data created under the assumptions of each model. This allows us to not only assess each model in terms of prediction and estimation, but also in terms of success when applied to data

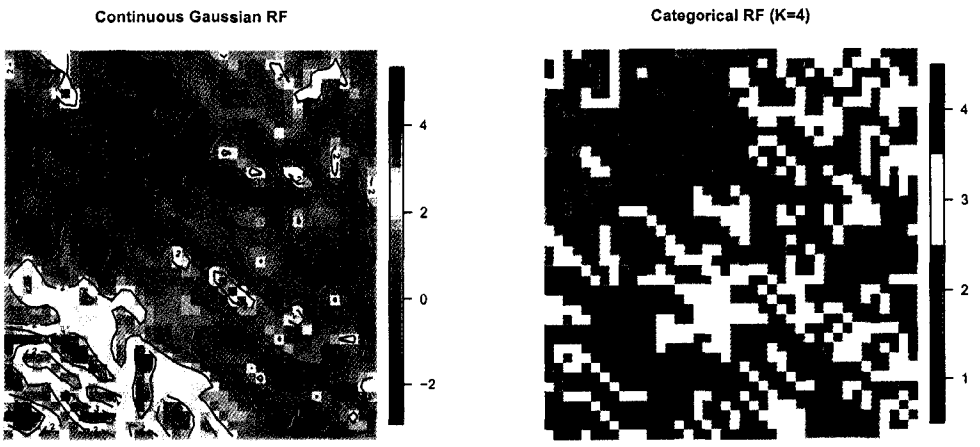

Figure 2.1: The lefthand plot displays one realization from a Gaussian random field, while the right displays a four-category categorical random field created by clipping the continuous random field with a vector of cut-points.

generated under a different model. It further allows assessment of the effectiveness of our measures of predictive ability at choosing the true, and thus correct, model.

2.2 Notation for spatial models

We now briefly reintroduce needed notation and definitions used for spatial modeling. Point-referenced spatial data can be observed at any continuously varying location, $\mathbf{s}$, usually indexed by latitude and longitude, over some region, D, such that $\mathbf{s} \in D \subset \Re^2$. Spatial dependence is introduced into both models through the incorporation of a Gaussian process,

$$\mathbf{W} \sim GP(\boldsymbol{\mu}, \boldsymbol{\Sigma}(\boldsymbol{\gamma})),$$

with mean, $\boldsymbol{\mu}$, and variance-covariance, $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$. The Gaussian process specification for $\mathbf{W}$ defines the joint distribution of any finite set of observations at locations $\{\mathbf{s}_i : \mathbf{s}_i \in D\}$ for $i = 1, \ldots, n$, from one realization of the process, as multivariate

Gaussian. We denote the set of observations as $\mathbf{W}_n = (W(\mathbf{s}_1), W(\mathbf{s}_2), \ldots, W(\mathbf{s}_n))$, and their joint distribution as

$$\mathbf{W}_n \sim MVN_n(\boldsymbol{\mu}(\mathbf{s}), \boldsymbol{\Sigma}(\boldsymbol{\gamma})), \tag{2.1}$$

where MVN_n denotes an n-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}(\mathbf{s})$ and variance-covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma} = (\sigma^2, \boldsymbol{\psi})$. The covariance matrix is written as $\sigma^2 H(\boldsymbol{\psi})$ where $H(\boldsymbol{\psi})$ is a correlation matrix parameterized by parameter vector $\boldsymbol{\psi}$ such that $[H(\boldsymbol{\psi})]_{ij}] = \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$. The correlation function, $\rho(d; \boldsymbol{\psi})$, depends on some measure of separation, d, between two locations, $\mathbf{s}_i$ and $\mathbf{s}_j$, and must be valid in the sense that it results in a positive-definite covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$. Thus, the covariance between $W(\mathbf{s}_i)$ and $W(\mathbf{s}_j)$ is given by a common variance, σ^2, scaled by a value defined through the correlation function, $Cov[W(\mathbf{s}_i), W(\mathbf{s}_j)] = \boldsymbol{\Sigma}_{i,j}(\boldsymbol{\gamma}) = \sigma^2 \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$. We assume a no-nugget specification of the covariance structure.

The expected value of $W(\mathbf{s}_i)$ is defined by the mean of the Gaussian process such that $E[W(\mathbf{s}_i)] = \mu(\mathbf{s}_i)$. For a constant mean process, $E[W(\mathbf{s}_i)] = \mu$ for all $\mathbf{s}_i$. For notational convenience, we use $\mathbf{W}$ in place of $\mathbf{W}_n$ where the context is clear. The proposed Double Latent model utilizes a stationary, mean-zero Gaussian process, while the Clipped Gaussian allows for non-stationarity in the mean structure. Both models, however, assume stationarity of variances, assuming a common variance, σ^2, for all locations.

The vector of ordered categorical random variables at the set of spatial locations, $\{\mathbf{s}_i : \mathbf{s}_i \in D\}$ for $i = 1, \ldots, n$, is given by $\mathbf{Y} = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n))^T$. We define the categorical response variable at one location, $Y(\mathbf{s}_i)$, as taking one of K category labels. For convenience, we use the common convention of integer valued labels such that $Y(\mathbf{s}_i) \in \{1, 2, \ldots, K\}$. For notational simplicity, we often use i in place of $\mathbf{s}_i$ so that $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Again, we use upper-case letters to denote

random variables and lower-case letters to denote the observed values of those random variables.

2.3 The Double Latent model

The Double Latent model builds on the spatial generalized linear mixed model ideas formulated by Diggle et al. (1998). It is specified as a hierarchical model with spatial random effects introduced through a latent Gaussian process. This general hierarchical model was also described for binary data by Gelfand and Ravishanker (1998) and Gelfand et al. (2000) who used importance sampling methods for inference. Under the hierarchical framework, the categorical responses are modeled as conditionally independent given the regression parameters and the latent spatial process. As its name implies, the Double Latent model defines two latent variables found at different levels of the hierarchy. As described in more detail below, the first latent variable, $\mathbf{Z}$, is the continuous variable that is assumed to be cut to create the categorical response and is thus specified at the data level of the hierarchy. It is analogous to the latent variable described for the non-spatial models in Section 1.1.2. The second latent variable, $\mathbf{W}$, is the Gaussian process used to incorporate spatial dependence into the model. The hierarchical structure defines the conditional independence of $\mathbf{Z}$ given $\mathbf{W}$. This conditional independence also holds for the categorical response variable, $\mathbf{Y}$, because of the deterministic relationship between $\mathbf{Z}$ and $\mathbf{Y}$.

Formal model specification begins with the definition of the deterministic relationship between the categorical response, $\mathbf{Y}$, and the underlying continuous latent variable, $\mathbf{Z}$, given for one location by,

$$Y(\mathbf{s}_i) = k \quad \text{if} \quad \theta_{k-1} < Z(\mathbf{s}_i) \leq \theta_k, \tag{2.2}$$

where $k = 1, \ldots, K$, and $\boldsymbol{\theta} = (-\infty = \theta_0 < \theta_1 < \ldots < \theta_{K-1} < \theta_K = \infty)$ is a vector of cut-points.

The first stage of the hierarchical model defines $Z(\mathbf{s}_i)$ in terms of a vector of regression parameters, $\boldsymbol{\beta}$; values of the covariate(s), $\mathbf{x}(\mathbf{s}_i)$; the latent spatial variable, $W(\mathbf{s}_i)$; and an error term, $\epsilon(\mathbf{s}_i)$. That is,

$$Z(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta} + W(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \quad \text{where} \quad \epsilon(\mathbf{s}_i) \sim N(0,1), \tag{2.3}$$

for $i = 1, \ldots, n$, or alternatively

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{W} + \boldsymbol{\epsilon} \quad \text{where} \quad \boldsymbol{\epsilon} \sim MVN_n(\mathbf{0}, \mathbf{I}_n), \tag{2.4}$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix. At the second stage of the hierarchy, $\mathbf{W}$ is assumed to be a Gaussian process with mean zero and variance-covariance matrix, $\boldsymbol{\Sigma}(\gamma)$,

$$\mathbf{W} \sim GP(\mathbf{0}, \boldsymbol{\Sigma}(\gamma)), \tag{2.5}$$

where $\boldsymbol{\Sigma}(\gamma) = \sigma^2 H(\boldsymbol{\psi})$, where $[H(\boldsymbol{\psi})]_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$, and ρ is a valid correlation function with parameter vector $\boldsymbol{\psi}$. The parameters included in γ are modeled at the third stage of the hierarchy through definition of their prior distributions and setting the values of the hyperparameters of those distributions.

The $Z(\mathbf{s}_i)$ are specified as conditionally independent given $\mathbf{W}$ and $\boldsymbol{\beta}$, such that

$$Z(\mathbf{s}_i)|w(\mathbf{s}_i), \boldsymbol{\beta} \sim N(\mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i), 1). \tag{2.6}$$

This also translates into conditional independence for the $Y(\mathbf{s}_i)$ given $\mathbf{W}$ and $\boldsymbol{\beta}$ as a result of the deterministic relationship between $\mathbf{Y}$ and $\mathbf{Z}$. That is, the cumulative probabilities for the categorical random variable are modeled as a function of $\boldsymbol{\beta}$ and $\mathbf{W}$:

$$Pr(Y(\mathbf{s}_i) \leq k|\mathbf{w}, \boldsymbol{\beta}) = h(\mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta} + w(\mathbf{s}_i)). \tag{2.7}$$

The specification of the distribution of $\epsilon(\mathbf{s}_i)$ as standard Gaussian in (2.3), indirectly defines $h(\cdot)$ as the standard normal cdf, therefore specifying an inverse normal cdf link function and the probit model. Specification of different distributions for $\boldsymbol{\epsilon}$, and thus

different link functions, allows for a wide range of models with this same structure, in the typical fashion of generalized linear models. However, we restrict attention to the standard normal (probit) link for this initial investigation of spatial models for point-referenced multi-categorical data. The variance of the white noise process also corresponds to a nugget-like term as described in classical geostatistical modeling (Cressie, 1993). This corresponds to the strategy of incorporating a nugget effect through nesting of models, where a white noise component describes the nugget effect (Schabenberger and Gotway, 2005). In such a context, it makes sense to consider the nugget as measurement error rather than micro-scale variability.

The use of latent variables, combined with the hierarchical structure, lends itself to Bayesian inference, and the absence of closed form posterior distributions leads us to employ MCMC methods. The latent variables are incorporated using the common computational technique of data augmentation to facilitate the MCMC algorithm. Data augmentation typically defines an *extended model* to include the observed data, the latent variables, and all other parameters. We build on methods for ordered categorical data described by Albert and Chib (1993), Albert and Chib (1997), Cowles (1996), Chib and Greenberg (1998), and Nandram and Chen (1996), among others, and on methods for binary spatial data described by authors such as Diggle et al. (1998), Gelfand et al. (2000), and De Oliveira (2000). Our goal is to obtain a sample from the joint posterior distribution of the unknown parameters *and* the latent variables given the observed data. That is, we wish to sample from the posterior distribution

$$f(\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma} | \mathbf{y}) \propto f(\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\mathbf{z}|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}). \quad (2.8)$$

The first term in the factorization of (2.8) is the likelihood conditioned on the latent variable $\mathbf{Z}$. The deterministic relationship between $\mathbf{Y}$ and $\mathbf{Z}$, as defined in (2.2), leads to a very simple form for this distribution. When the value of Z_i is

known, the probability that Y_i belongs to any one of the K categories is simply 0 or 1. That is, $f(y_i|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\theta})$ is merely an indicator of whether the latent z_i falls within the appropriate cut-point interval associated with the response value, Y_i. We refer to these intervals as $A(y_i)$ and use the observed value, y_i, as a subscript to define them such that

$$A(y_i) = (\theta_{y_i-1}, \theta_{y_i}]\,. \tag{2.9}$$

For example, if $y_i = k$, $A(y_i) = (\theta_{y_i-1}, \theta_{y_i}] = (\theta_{k-1}, \theta_k]$. We can then write this augmented likelihood as,

$$\begin{aligned} f(\mathbf{y}|\mathbf{z}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &= \prod_{i=1}^{n} I_{\{\theta_{y_i-1} \le z_i < \theta_{y_i}\}} \\ &= \prod_{i=1}^{n} I_{\{z_i \in A(y_i)\}}, \end{aligned} \tag{2.10}$$

where $I_{\{b \in B\}} = 1$ if $b \in B$ and 0 if $b \notin B$.

We also require the more interesting unaugmented form of the likelihood for the MCMC algorithm. This is obtained by marginalizing over the latent variable, $\mathbf{Z}$, and capitalizing on the conditional independence of the Y_i's given the spatial variable, $\mathbf{W}$,

$$\begin{aligned} f(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &= \prod_{i=1}^{n} Pr[Y_i = k|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}] \\ &= \prod_{i=1}^{n} Pr[\theta_{y_i-1} < Z_i < \theta_{y_i}|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}] \\ &= \prod_{i=1}^{n} [\Phi(\theta_{y_i} - \mathbf{X}_i\boldsymbol{\beta} - w_i) - \Phi(\theta_{y_i-1} - \mathbf{X}_i\boldsymbol{\beta} - w_i)]. \end{aligned} \tag{2.11}$$

The details of the Gibbs sampling algorithm used to sample from the posterior distribution in (2.8) are given in the following subsection, 2.3.1, and reveal how the original likelihood given in (2.11) enters the algorithm.

2.3.1 Gibbs algorithm for the Double Latent model

The Gibbs sampling algorithm proceeds by sampling successively from the complete conditional distributions of the parameters (Gelman et al., 2004). We apply the algorithm for the Double Latent model by iteratively sampling through the following steps. The steps are described in detail in this section.

1. Update the spatial parameters, γ, using the Metropolis-Hastings algorithm.
2. Update the latent spatial variables, $\mathbf{W}$, using the multivariate Gaussian complete conditional distribution.
3. Update each element of the latent variable, $\mathbf{Z}$, using the truncated univariate Gaussian complete conditional distribution.
4. Update the cut-point parameters, θ, via reparameterization and a blocking algorithm.
5. Update the regression parameters, β, using the multivariate Gaussian complete conditional distribution.

Spatial parameters, γ

The vector of spatial parameters for the Double Latent model is $\gamma = (\psi, \sigma^2)$, where ψ is the vector of parameters specifying the correlation function and σ^2 is the common variance. The spatial parameters, $\gamma = (\sigma^2, \psi)$, are specified at a level in the hierarchy that results in conditional independence between γ and both $\mathbf{Y}$ and $\mathbf{Z}$. We also assume independence between γ and (β, θ). Thus, the complete conditional distribution for γ depends only on the latent spatial variable, $\mathbf{W}$, along with its specified prior distribution, and can be written as

$$\begin{aligned} f(\gamma|\mathbf{y},\mathbf{w},\mathbf{z},\beta,\theta) &= f(\gamma|\mathbf{w},\beta,\theta) \\ &\propto f(\mathbf{w}|\beta,\theta,\gamma)f(\gamma|\beta,\theta)f(\beta,\theta) \\ &\propto f(\mathbf{w}|\gamma)f(\gamma). \end{aligned} \tag{2.12}$$

We assume independence between σ^2 and ψ and use proper, vague priors for both parameters resulting in non-standard complete conditional distributions. Specification of the prior distributions is discussed in detail in Chapter 3. We use the Metropolis-Hastings (M-H) algorithm to sample from $f(\gamma|\mathbf{y}, \mathbf{w}, \mathbf{z}, \beta, \theta)$. The parameters σ^2 and ψ can be updated independently with two successive M-H steps. We choose candidate values uniformly over the parameter spaces defined by the priors so that the M-H ratios involve only $f(\mathbf{w}|\gamma) = f(\mathbf{w}|\sigma^2, \psi)$ evaluated at the appropriate values. For reasons discussed in detail in Chapter 3, we fix the value of σ^2, and also restrict our attention to correlation functions with only one parameter to estimate. This includes, however, the use of multi-parameter correlation functions with one or more parameters fixed prior to model fitting.

Spatial latent vector, W

Updating the vector of latent spatial variables can be accomplished in a univariate or multivariate fashion. The univariate method updates by cycling through the univariate conditional distributions of each location. The multivariate method proceeds by drawing from the multivariate complete conditional distribution of all locations. The two methods theoretically result in the same stationary distribution, though may result in computational differences. Diggle et al. (1998) use the univariate method in their algorithm. We describe both methods here, starting with the univariate method.

Let $\mathbf{W}_{-i}$ be the $\mathbf{W}$ vector with the ith element deleted. The full conditional distribution for each component of $\mathbf{W}$ for $i = 1, 2, \ldots, n$ is given by

$$\begin{aligned} f(w_i|\mathbf{w}_{-i}, \mathbf{y}, \mathbf{z}, \beta, \theta, \gamma) &= f(w_i|\mathbf{w}_{-i}, y_i, z_i, \beta, \theta, \gamma) \\ &\propto f(y_i|\mathbf{w}, z_i, \beta, \theta, \gamma) f(w_i|\mathbf{w}_{-i}, z_i, \beta, \theta, \gamma) f(\mathbf{w}_{-i}, Z_i, \beta, \theta, \gamma) \\ &\propto f(y_i|z_i, \beta, \theta, \gamma) f(w_i|\mathbf{w}_{-i}, z_i, \beta, \theta, \gamma) \\ &\propto f(w_i|\mathbf{w}_{-i}, \gamma). \end{aligned} \tag{2.13}$$

Since $\mathbf{W}$ is assumed to follow a Gaussian process, the conditional distribution for one component of $\mathbf{W}$ given all other components and the unknown parameters, $f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = f(w_i|\mathbf{w}_{-i}, \boldsymbol{\gamma})$, is univariate Gaussian. The parameters of this Gaussian distribution are easily calculated using the well known result for specifying the conditional distribution of one element of a vector of jointly normal random variables given other elements of the vector (e.g., Hocking (2003)). This calls for the definition of sub-vectors and sub-matrices. Let $\mathbf{v}_i$ be the ith column of $\boldsymbol{\Sigma}$ with its ith element deleted, and $\boldsymbol{\Sigma}_{-i}$ be the matrix $\boldsymbol{\Sigma}$ with its ith column and ith row removed. Write $\mathbf{W} = (W_i, \mathbf{W}_{-i})'$ and partition the covariance matrix, $\boldsymbol{\Sigma}$, into four parts,

$$\boldsymbol{\Sigma} = \begin{pmatrix} v_{ii} & \mathbf{v}_i' \\ \mathbf{v}_i & \boldsymbol{\Lambda}_{-i} \end{pmatrix}, \tag{2.14}$$

where v_{ii} is the variance of W_i, $\mathbf{v}_i$ is the vector of covariances between W_i and $\mathbf{W}_{-i}$, and $\boldsymbol{\Lambda}_{-i}$ is the variance-covariance matrix of $\mathbf{W}_{-i}$. Then, the complete conditional distribution, $f(w_i|\mathbf{w}_{-i}, \boldsymbol{\gamma})$, is the Gaussian distribution

$$N(\mathbf{v}_i'\boldsymbol{\Lambda}_{-i}^{-1}\mathbf{w}_{-i} \quad , \quad \boldsymbol{\Sigma}_{ii} - \mathbf{v}_i'\boldsymbol{\Lambda}_{-i}^{-1}\mathbf{v}_i), \tag{2.15}$$

resulting in easy sampling of each W_i in turn.

To improve efficiency of the algorithm, we avoid calculating $\boldsymbol{\Sigma}_{-i}^{-1}$ for each i. Instead, we use a matrix result implemented by De Oliveira (1997, 2000) and Christensen et al. (1992). We invert only the original $n \times n$ covariance matrix, $\boldsymbol{\Sigma}$, to obtain $\boldsymbol{\Sigma}^{-1}$. Using the notation in (2.14), we partition $\boldsymbol{\Sigma}^{-1}$ into v_{ii}, $\mathbf{v}_i$, and $\boldsymbol{\Lambda}_{-i}$, and capitalize on the fact that

$$\boldsymbol{\Sigma}_{-i}^{-1} = \boldsymbol{\Lambda}_{-i} - \frac{\mathbf{v}_i\mathbf{v}_i'}{v_{ii}} \tag{2.16}$$

to invert only one matrix, $\boldsymbol{\Sigma}$, for each iteration of the algorithm, instead of n matrices for each iteration.

An equivalent formulation of the algorithm is to draw the vector $\mathbf{W}$ from the appropriate multivariate Gaussian complete conditional distribution at each iteration rather than drawing each element individually from its univariate conditional distribution. The joint specification gives

$$\begin{aligned} f(\mathbf{w}|\mathbf{y},\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) &= f(\mathbf{y}|\mathbf{w},\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{w}|\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{y}|\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{w}|\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{w}|\boldsymbol{\gamma}), \end{aligned} \tag{2.17}$$

where $(\mathbf{W}|\boldsymbol{\gamma})$ is distributed as $MVN_n(\mathbf{0}, \boldsymbol{\Sigma}(\boldsymbol{\gamma}))$. This joint specification is more computationally convenient as it allows us to draw all n elements of $\mathbf{W}$ simultaneously, eliminating the need for consecutive looping though the individual elements of the vector. Theoretically, this results in convergence to the same posterior distribution as use of the marginal univariate distributions (Givens and Hoeting, 2006). We found that the multivariate method resulted in a faster algorithm with improved convergence that also translated into slightly better predictions. Thus, we use multivariate updating for the simulation study in Chapter 5 and to fit the model to real data in Chapter 6.

Latent vector, Z

The non-spatial latent vector, $\mathbf{Z}$, is updated by successively drawing the individual elements, similar to the univariate method described for the spatial latent variable. The complete conditional distribution for each Z_i is an independent truncated Gaussian distribution. It is computationally easier to draw from a truncated univariate distribution than a truncated multivariate distribution, and the independence makes univariate updating a logical choice. Therefore, we are interested in the

complete conditional distribution for one element, Z_i, given by

$$\begin{aligned} f(z_i|\mathbf{z}_{-i}, \mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &= f(z_i|y_i, w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\propto f(y_i|z_i, w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(z_i|w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\propto f(y_i|z_i, w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(z_i|w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}), \end{aligned} \tag{2.18}$$

where $\mathbf{z}_{-i}$ refers to the vector of $\mathbf{z}$ after removing z_i. Recall that $f(y_i|z_i, w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) = I_{\{\theta_{y_i-1} < z_i \leq y_i\}} = I_{\{z_i \in A(y_i)\}}$, and the model is defined such that

$$Z_i|w_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma} \sim N(\mathbf{x}_i\boldsymbol{\beta} + w_i, 1).$$

Thus, the complete conditional for Z_i is the density of a truncated univariate Gaussian distribution,

$$f(z_i|\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \propto \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z_i - \mathbf{x}_i\boldsymbol{\beta} - w_i)^2\right) I_{\{z_i \in A(y_i)\}}, \tag{2.19}$$

and Z_i can be sampled using standard techniques.

Cut-point parameters, $\boldsymbol{\theta}$

We originally consider a vector of cut-points of length $K+1$, defined such that $\theta_0 < \theta_1 < \theta_2 < \ldots < \theta_{K-1} < \theta_K$, where K is the total number of categories. Following the convention of Albert and Chib (1993) and others, we set $\theta_0 = \infty$, $\theta_1 = 0$, and $\theta_K = \infty$, leaving $K-2$ unknown cut-points to estimate. The definition of θ_0 and θ_K as $-\infty$ and ∞, respectively, is purely for notational convenience, allowing us to write one general equation for all k when referring to the probabilities, $Pr(Y_i = k) = Pr(\theta_{k-1} < Z_i \leq \theta_k)$. The cut-point θ_1 is fixed at 0 for identifiability of the parameters, since we include an intercept term in the vector of regression parameters. This issue is discussed further in Chapter 3. We now denote the vector of length $K-2$ as $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_{K-1})$, unless otherwise specified.

The full conditional distribution for the cut-points is derived similarly to that shown in Albert and Chib (1993), assuming a uniform prior for $\boldsymbol{\theta}$ such that

$f(\theta_k|\theta_{k-1},\theta_{k+1}) \propto I_{(\theta_{k-1}<\theta_k<\theta_{k+1})}$. We derive this distribution, up to a proportionality constant, for individual cut-points in $\boldsymbol{\theta}$, resulting in

$$\begin{aligned} f(\theta_k|\mathbf{z},\mathbf{w},\mathbf{y},\boldsymbol{\beta},\boldsymbol{\theta}_{-k},\boldsymbol{\gamma}) &\propto f(\mathbf{y}|\mathbf{z},\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{z}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{y}|\mathbf{w},\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{z}|\mathbf{w},\boldsymbol{\beta})f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\gamma})f(\boldsymbol{\theta}) \\ &\propto f(\mathbf{y}|\mathbf{w},\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\theta}) \\ &= f(\mathbf{y}|\mathbf{w},\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\theta_k|\boldsymbol{\theta}_{-k})f(\boldsymbol{\theta}_{-k}) \\ &\propto f(\mathbf{y}|\mathbf{w},\mathbf{z},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\theta_k|\theta_{k-1},\theta_{k+1}). \end{aligned} \tag{2.20}$$

This can be written out as proportional to

$$\begin{aligned} &\left[\prod_{i=1}^{n} I_{\{\theta_{y_i-1}<z_i\leq\theta_{y_i}\}}\right] I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}} = \\ &\qquad \prod_{i=1}^{n}\left[\sum_{j=1}^{K} I_{\{y_i=j\}} I_{\{\theta_{j-1}<z_i\leq\theta_j\}}\right] I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}} \\ &\propto \prod_{i:y_i\in\{k,k+1\}} \left[I_{\{y_i=k\}} I_{\{\theta_{k-1}<z_i\leq\theta_k\}} + I_{\{y_i=k+1\}} I_{\{\theta_k<z_i\leq\theta_{k+1}\}}\right] I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}} \\ &= \prod_{i:y_i\in\{k,k+1\}} \left[I_{\{\theta_{y_i-1}<z_i\leq\theta_{y_i}\}}\right] I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}}. \end{aligned} \tag{2.21}$$

Note that as it is expressed in Albert and Chib (1998), the complete conditional distribution is always zero for a multi-category problem. To avoid this problem, it is necessary to specify that the product in (2.21) be over i such that $y_i \in \{k, k+1\}$. As described by Albert and Chib (1993), this complete conditional for θ_k is then uniformly distributed over the interval

$$[\max\{\max(z_i; y_i = k), \theta_{k-1}\}\ ,\ \ \min\{\min(z_i; y_i = k+1), \theta_{k+1}\}], \tag{2.22}$$

for $i = 1, \ldots, n$. This is essentially a Uniform distribution over the interval

$$[\max(z_i; y_i = k), \min(z_i; y_i = k+1)] \tag{2.23}$$

while preserving the ordered constraints on the cut-points.

The interval in (2.22) is defined by values of z that squeeze around the cut-point of interest, θ_k. As a result, the interval can be quite narrow, leading to slow mixing of the MCMC chain, as first noted by Cowles (1996). To address this problem, Cowles (1996) proposes drawing the latent variable, $\mathbf{Z}$, and cut-points, $\boldsymbol{\theta}$, in one block in an attempt to speed mixing and improve convergence of the chain. That is, she suggests sampling from the joint conditional distribution, $f(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{w}, \gamma, \boldsymbol{\beta})$, rather than the separate full conditional distributions. This technique of "blocking" can reduce the correlation in the MCMC chain, thus accelerating convergence.

The blocking strategy no longer allows for a standard Gibbs sampling algorithm based on the uniform distributions in (2.22), but instead requires a multivariate Metropolis-Hastings updating step. However, the M-H ratio obtained using the multivariate M-H algorithm to sample from this joint distribution is equivalent to that obtained using a M-H step to draw $\boldsymbol{\theta}$ from $f(\boldsymbol{\theta}|\mathbf{y}, \mathbf{w}, \gamma, \boldsymbol{\beta})$ and then drawing $\mathbf{z}$ from its truncated Gaussian complete conditional distribution (Cowles, 1996) . This idea is particularly appealing and straightforward to implement in situations, such as this one, where the full conditional distribution of one of the variables is easy to sample from and the marginal distribution of the other is relatively easy to derive. The trick involves constructing a joint proposal density that is the product of a suitable proposal density for the marginal variable multiplied by the full conditional density for the other variable, in this case $\mathbf{Z}$. This results in the canceling out of the full conditional of $\mathbf{Z}$ from the M-H ratio for updating the marginal variable, and essentially results in sampling the cut-points from their complete conditional distribution after marginalizing over the latent variable, $\mathbf{Z}$ (Cowles, 1996). The idea is more clearly seen by the factorization of the joint conditional distribution,

$$f(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y}, \mathbf{w}, \gamma, \boldsymbol{\beta}) = f(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y}, \gamma, \boldsymbol{\beta}) f(\boldsymbol{\theta}|\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \gamma). \tag{2.24}$$

A number of variations to this approach have been developed to further improve convergence. Several authors have used Cowles' blocking and factorization strategy

in conjunction with re-parameterizations of the cut-point parameters. For example, Albert and Chib (1997, 1998, 2001) additionally propose a de-constraint transformation for the cut-points, along with a multivariate-t proposal distribution for the M-H step. Nandram and Chen (1996) suggest further constraining the cut-points to the unit interval to facilitate use of a Dirichlet proposal density. Chen and Dey (1996) describe yet another transformation, combining the re-scaling transformation of Nandram and Chen (1996) with a de-constraint transformation, similar to that of Albert and Chib (1998).

For the models proposed in this dissertation, we utilize the re-parameterization described by Albert and Chib (1997, 1998, 2001), along with the blocking strategy proposed by Cowles (1996). We re-express $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_{K-1})$ as $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_{K-1})$, where $\alpha_2 = \log(\theta_2)$ and $\alpha_k = \log(\theta_k - \theta_{k-1})$ for $k = 3, \ldots, K-1$. The inverse transformation is then given by $\theta_k = \sum_{i=2}^{k} \exp(\alpha_i)$. There are obvious computational advantages to transforming the constrained cut-points to an unconstrained, real-valued vector, $\boldsymbol{\alpha}$. For example, this approach allows for the use of common multivariate continuous distributions for both the proposal distribution and the prior distribution of $\boldsymbol{\alpha}$. Albert and Chib (1997) report success using a multivariate-t proposal distribution and a multivariate Gaussian prior distribution for $\boldsymbol{\alpha}$, and we also found this to be successful. Following the blocking strategy described above, we sample from the conditional distribution, $f(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, which can be factorized as

$$\begin{aligned} f(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\gamma}) &\propto f(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) f(\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) f(\boldsymbol{\alpha}) \\ &\propto f(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) f(\boldsymbol{\alpha}), \end{aligned} \tag{2.25}$$

assuming independence between $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$. Note that $f(\mathbf{y}|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\gamma})$ is the original likelihood of interest given in (2.11), and $f(\boldsymbol{\alpha})$ can be a multivariate Gaussian prior density.

To further improve convergence, Albert and Chib (1997) chose to match the proposal density to the target conditional density, $f(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \gamma)$. We follow their

suggestion, using a multivariate-t density with ν degrees of freedom and parameters specified by locating the approximate mode, $\hat{\boldsymbol{\alpha}}$, and the inverse of the observed information matrix of $\log(f(\boldsymbol{\alpha}|\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \gamma))$ via several Newton-Raphson iterations. This idea is similar to that of a Langevin-Hastings algorithm. The inverse of the observed information matrix is the negative inverse of the Hessian matrix and is denoted by $\mathbf{V}$. In addition to the automatically updated parameter, $\hat{\boldsymbol{\alpha}}$ and $\mathbf{V}$, the distribution is specified by two fixed tuning parameters, ν and ψ^2. We discuss the tuning of these parameters in Chapter 3. The multivariate-t proposal density is written as

$$f_t(\boldsymbol{\alpha}|\hat{\boldsymbol{\alpha}}, \psi^2\mathbf{V}, \nu) \propto \left((1 + \frac{1}{\nu}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})'(\psi^2)\mathbf{V})^{-1}(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})\right)^{-(\nu-K)/2}. \tag{2.26}$$

Realizations are drawn from this multivariate-t density by first sampling λ from a Gamma$(\nu/2, \nu/2)$ distribution and ϵ from a Normal$(0, \psi^2\mathbf{V})$ distribution and then calculating $\hat{\boldsymbol{\alpha}}+\lambda^{-1/2}\epsilon$ (Albert and Chib, 1997). The algorithm moves to the candidate value, α^{cand}, from the current value, α^{cur}, with a Metropolis-Hastings transition probability

$$\min\left\{\frac{f(\mathbf{y}|\mathbf{w}, \boldsymbol{\alpha}^{cand}, \boldsymbol{\beta}, \gamma)f(\boldsymbol{\alpha}^{cand})f_T(\boldsymbol{\alpha}^{cur}|\hat{\boldsymbol{\alpha}}, \psi^2 V, \nu)}{f(\mathbf{y}|\mathbf{w}, \boldsymbol{\alpha}^{cur}, \boldsymbol{\beta}, \gamma)f(\boldsymbol{\alpha}^{cur})f_T(\boldsymbol{\alpha}^{cand}|\hat{\boldsymbol{\alpha}}, \psi^2 V, \nu)}, 1\right\}. \tag{2.27}$$

To reduce computation time, we modify Albert and Chib's algorithm to obtain $\hat{\boldsymbol{\alpha}}$ and V for the first approximately 10 iterations and subsequently every 50th iteration. This strategy was implemented after observing very little change in $\hat{\boldsymbol{\alpha}}$ and V after the first approximately 10 iterations. The change greatly reduces computation time while not noticeably changing the convergence behavior of the chain.

The model is originally parameterized in terms of $\boldsymbol{\alpha}$ and we place the prior directly on $\boldsymbol{\alpha}$ resulting in no need to include a Jacobian term in the M-H ratio. However, by doing this it is not clear what prior we are indirectly placing on the original vector of cut-points, $\boldsymbol{\theta}$. This fact does not appear to have been investigated

by Albert and Chib (1997, 1998) or anyone else. We devote needed attention to this matter by investigating the prior induced for $\boldsymbol{\theta}$ for the case of four categories in Section 3.3.3.

Regression parameters, $\boldsymbol{\beta}$

The full conditional distribution for the regression parameters is expressed as

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y},\mathbf{z},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma}) &= f(\boldsymbol{\beta}|\mathbf{z},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{z}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta}|\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{z}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta}), \end{aligned} \tag{2.28}$$

assuming independence between $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$. The deterministic relationship between the categorical response, $\mathbf{Y}$, and the latent variable, $\mathbf{Z}$, results in the equality of the full conditional distributions, $f(\boldsymbol{\beta}|\mathbf{y},\mathbf{z},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma})$ and $f(\boldsymbol{\beta}|\mathbf{z},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma})$.

We assume a multivariate Gaussian prior for $\boldsymbol{\beta}$ with mean $\mathbf{b}_0$ and variance-covariance matrix $\mathbf{B}_0$. From (2.6), the Z_i's are conditionally independent Gaussian random variables given $\mathbf{W}$. This results in a multivariate Gaussian complete conditional distribution for $\boldsymbol{\beta}$. The mean vector and variance-covariance matrix of the distribution are derived by collecting terms involving $\boldsymbol{\beta}$, a standard result from linear models (Hocking, 2003). For this derivation, it is helpful to examine the form of $f(\mathbf{z}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})$:

$$\left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\left((\mathbf{z}-\mathbf{w})-\mathbf{X}\boldsymbol{\beta}\right)'\left((\mathbf{z}-\mathbf{w})-\mathbf{X}\boldsymbol{\beta}\right)\right\}. \tag{2.29}$$

In contrast to (2.19), the distribution in (2.29) is not truncated because there is no conditioning on $\mathbf{y}$. Multiplying (2.29) by the multivariate Gaussian prior density for $\boldsymbol{\beta}$ and collecting terms involving $\boldsymbol{\beta}$ results in

$$\boldsymbol{\beta}|\boldsymbol{\theta},\boldsymbol{\gamma},\mathbf{z},\mathbf{w} \sim MVN_p(\tilde{\mathbf{b}},\tilde{\mathbf{B}}) \tag{2.30}$$

where

$$\tilde{\mathbf{B}} = (\mathbf{B}_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1} \tag{2.31}$$

and

$$\begin{aligned} \tilde{\mathbf{b}} &= (\mathbf{B}_0^{-1} + \mathbf{X}^T\mathbf{X})^{-1}(\mathbf{B}_0^{-1}b_0 + \mathbf{X}^T(\mathbf{z} - \mathbf{w})) \\ &= \tilde{\mathbf{B}}(\mathbf{B}_0^{-1}b_0 + \mathbf{X}^T(\mathbf{z} - \mathbf{w})). \end{aligned} \tag{2.32}$$

2.3.2 An alternative mean specification

An alternative way to fit the Double Latent model specifies $\mathbf{X}\boldsymbol{\beta}$ as the mean of the latent Gaussian process, $\mathbf{W}$, rather than including it in the mean structure of the conditionally independent latent variable, $\mathbf{Z}$. This implies that instead of (2.3) and (2.5), we define

$$Z(\mathbf{s}_i) = W(\mathbf{s}_i) + \epsilon(\mathbf{s}_i) \quad \text{where} \quad \epsilon(\mathbf{s}_i) \sim N(0, 1). \tag{2.33}$$

$\mathbf{W}$ is assumed to be a Gaussian process with mean depending on covariates and variance-covariance matrix, $\boldsymbol{\Sigma}(\boldsymbol{\gamma})$,

$$\mathbf{W} \sim GP(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\gamma})). \tag{2.34}$$

We do not fit this parameterization in this dissertation, but in future investigation we will compare it to the originally specified Double Latent model and also to the Clipped Gaussian model which shares the non-zero mean definition of the latent Gaussian process, as is described in the next section. It is possible that one model may have computational and MCMC convergence advantages over the others.

In this alternative specification, $\boldsymbol{\beta}$ appears at a higher level of the hierarchy than in the original Double Latent model, resulting in different complete conditional distributions underlying the Gibbs sampling algorithm. The full conditional distribution

for $\boldsymbol{\beta}$ is now expressed as

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y},\mathbf{z},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma}) &= f(\boldsymbol{\beta}|\mathbf{z},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{z}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{z}|\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta}) \\ &\propto f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta}), \end{aligned} \tag{2.35}$$

again assuming independence between $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$.

Specifying a multivariate Gaussian prior distribution, $\boldsymbol{\beta} \sim MVN(\mathbf{b}_0, \mathbf{B}_0)$, and recalling that $f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})$ is also a multivariate Gaussian density, the complete conditional distribution for $\boldsymbol{\beta}$ is again multivariate Gaussian. Collecting the terms involving $\boldsymbol{\beta}$ gives

$$\boldsymbol{\beta}|\mathbf{z},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma} \sim MVN_p(b^*, \mathbf{B}^*), \tag{2.36}$$

where

$$\mathbf{B}^* = (\mathbf{B}_0^{-1} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}, \tag{2.37}$$

and

$$\begin{aligned} \mathbf{b}^* &= (\mathbf{B}_0^{-1} + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}) \\ &= \mathbf{B}^*(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{w}). \end{aligned} \tag{2.38}$$

Comparing (2.30)-(2.32) to (2.36)-(2.38), we see that the covariance matrix, $\boldsymbol{\Sigma}$, associated with $\mathbf{W}$ appears in the complete conditional distribution of $\boldsymbol{\beta}$ for this parameterization, but not in the original Double Latent specification. Thus, a disadvantage of this approach is that it requires inversion of the $n \times n$ matrix, $\boldsymbol{\Sigma}$, in order to update $\boldsymbol{\beta}$ at each iteration. Also, it involves use of a non-stationary Gaussian process, versus the stationary process specified for the original Double Latent model. Interestingly, Christensen et al. (2006) specify the spatially varying mean in the latent Gaussian process in recent work fitting spatial models for binary

and count data, where it facilitates their data dependent re-parameterizations used to improve convergence and generality of the algorithm. However, it seems that the more work is needed to clarify the advantages and disadvantages of using a latent Gaussian process with a non-zero mean.

2.4 The Clipped Gaussian model

The Clipped Gaussian model is an ordered multi-category extension of the model proposed by De Oliveira (2000) for binary data. We assume the existence of a true categorical random field, $\mathbf{Y}$, over the region of interest and again appeal to the concept of *clipping* to create this categorical random field from an underlying latent Gaussian process, $\mathbf{W}$. There is an underlying conceptual difference between the Double Latent model and the Clipped Gaussian model, along with clear differences in the model formulations. Both models employ the concept of *clipping* or *cutting* to create a categorical random field, but the Double Latent model is formulated from the foundation of generalized linear mixed models whereas the Clipped Gaussian is formulated directly from the idea of thresholding a continuous spatial process. Unlike the Double Latent model, the Clipped Gaussian model is not specified hierarchically, and does not include the additional error component associated with the lower level latent variable, $\mathbf{Z}$. The latent spatial Gaussian process is instead incorporated at the level of the data.

The construction of the connection between a categorical random field and a continuous Gaussian random field provides a means to make likelihood based analysis of a categorical random field feasible (De Oliveira, 1997). Specifically, it allows us to take advantage of the mathematical attractiveness and tractability of the continuous Gaussian random field in order to define and make predictions in the context of the categorical random field. For the Clipped Gaussian model, the deterministic relationship is defined between the categorical response, $\mathbf{Y}$, and the latent process,

$\mathbf{W}$. Recall that the Double Latent model instead defines the deterministic relationship between the categorical response,$\mathbf{Y}$, and the non-spatial latent variable, $\mathbf{Z}$, in (2.2).

To provide the foundation for the Clipped Gaussian model, we parallel the description provided by De Oliveira (1997) for the binary case. We consider our prediction problem as that of estimating a categorical map over space, i.e., mapping a region composed of only K attributes or features. Consider that the region of interest, $D \subset \Re^2$ is partitioned into K disjoint subregions, $D_1, D_2, \ldots, D_K$, where $D = D_1 \cup D_2 \cup \ldots \cup D_K = \cup_{j=1}^{K} D_j$ and $D_k = D - \cup_{j \neq k}^{K} D_j$. A practical example may be a region classified into K mutually exclusive land-use types, ordered according to habitat suitability for a species. The observed ordered categorical response at a location specifies the region, D_k, to which the observation belongs. The goal is to estimate (predict) the regions, $D_1, D_2, \ldots, D_K$ of one realization of the categorical random field using a sample of n observations from that realization. This relies on the assumptions that the set of D_k's represents one realization of a random process, and that the categorical random field is defined continuously at every $\mathbf{s} \in D$. That is, for integer valued labels, $Y(\mathbf{s}) = k$ if $\mathbf{s} \in D_k$.

As previously described, we use the conceptual and mathematical convenience provided by the concept of clipping a Gaussian random field to place the problem of estimating the categorical map in the context of spatial prediction in a continuous random field. This makes likelihood based inference possible by defining a family of finite-dimensional distributions, something that is historically lacking for binary and categorical data (De Oliveira, 1997).

We define the latent Gaussian process, $\mathbf{W}$, as having a mean possibly depending on covariates and the usual spatial covariance matrix. That is,

$$\mathbf{W} \sim GP(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\gamma})) \tag{2.39}$$

where $\Sigma(\gamma) = \sigma^2 H(\psi)$ and $[H(\psi)]_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \psi)$ where ρ is a correlation function that produces a valid covariance matrix. We assume a no-nugget specification of $\Sigma(\gamma)$.

Using integer labels $1, \ldots, K$ for the K categories, the categorical random field is assumed to be obtained through the relationship,

$$\begin{aligned} Y(\mathbf{s}_i) &= \sum_{k=1}^{K} k I_{\{\theta_{k-1} < W(\mathbf{s}_i) \leq \theta_k\}} \\ &= \sum_{k=1}^{K} k I_{\{W(\mathbf{s}_i) \in A(k)\}}. \end{aligned} \tag{2.40}$$

Recall that $A(k) = (\theta_{k-1}, \theta_k]$, and we use the subscript i in place of $\mathbf{s}_i$. The associated probabilities arise from the deterministic relationship defined between $\mathbf{Y}$ and $\mathbf{W}$ giving

$$\begin{aligned} Pr(Y_i = k) &= Pr(\theta_{k-1} < W_i \leq \theta_k) \\ &= Pr(W_i \in A(k)). \end{aligned} \tag{2.41}$$

The cumulative probabilities are defined similarly.

Then, to create the map of a realization of the categorical random field, we use

$$\hat{D}_k = \{\mathbf{s} \in D : \hat{Y}(\mathbf{s}) = k\} \quad \text{for} \quad k = 1, \ldots, K-1, \tag{2.42}$$

where D_K is then given by $D - \cup_{j=1}^{K-1} D_j$. Of course, in practice we can only predict at a finite number of locations, $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m$.

Again, we employ the technique of data augmentation and form an extended model to include the data, the original parameters, *and* the latent variable. As shown in this section, the posterior distribution of interest is not available in closed form and thus we rely on MCMC methods for inference. The methods are facilitated by use of the latent variable data augmentation technique. The joint posterior distribution of the unknown parameters and latent variable, given the observed data, is given by

$$f(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \gamma | \mathbf{y}) \propto f(\mathbf{y} | \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \gamma) f(\mathbf{w} | \boldsymbol{\beta}, \boldsymbol{\theta}, \gamma) f(\boldsymbol{\beta}, \boldsymbol{\theta}, \gamma). \tag{2.43}$$

As a result of the exclusion of both the non-spatial latent variable, $\mathbf{Z}$, and hierarchical structure employed for the Double Latent model, the likelihood for the Clipped Gaussian model is not simplified by the conditional independence. Instead, the joint distribution of the $\mathbf{Y}$ is a n-dimensional integral of a multivariate Gaussian distribution with parameters from the latent Gaussian process. This unaugmented likelihood, first given in (1.32), is

$$f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\theta},\psi,\sigma^2) = \int_{A(y_1)} \cdots \int_{A(y_n)} \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} |H(\psi)|^{-1/2} \cdot \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{w}-\mathbf{x}\boldsymbol{\beta})' H(\psi)^{-1}(\mathbf{w}-\mathbf{x}\boldsymbol{\beta})\right\} d\mathbf{w}. \tag{2.44}$$

It is clear from (2.44) that $f(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta})$ is obtained by integrating $f(\mathbf{y},\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta})$ with respect to $\mathbf{w}$. However, use of the latent $\mathbf{W}(\mathbf{s}_i)$'s within the Gibbs sampler avoids the need for direct evaluation of the multi-dimensional integral.

In contrast to (2.44), recall that the likelihood for the Double Latent model is

$$\begin{aligned} f(\mathbf{y}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) &= \prod_{i=1}^{n} Pr[\theta_{y_i-1} < Z_i < \theta_{y_i}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}] \\ &= \prod_{i=1}^{n} [\Phi(\theta_{y_i} - \mathbf{x}_i\boldsymbol{\beta} - w_i) - \Phi(\theta_{y_i-1} - \mathbf{x}_i\boldsymbol{\beta} - w_i)], \end{aligned}$$

as given in (2.11).

The likelihood of $\mathbf{y}$ given the latent vector $\mathbf{w}$ is simply a product of indicator variables due to the deterministic relationship between $\mathbf{Y}$ and $\mathbf{W}$, which simplifies calculation of many of the complete conditional distributions needed to construct the Gibbs sampling algorithm;

$$\begin{aligned} f(\mathbf{y}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) &= \prod_{i=1}^{n} I_{\{\theta_{y_i-1} \le w_i < \theta_{y_i}\}} \\ &= \prod_{i=1}^{n} I_{\{w_i \in A(y_i)\}}. \end{aligned} \tag{2.45}$$

Once again, the Clipped Gaussian model does not include the white noise process, ϵ, that is incorporated into the first stage of the Double Latent model specification. Recall that the standard normal distribution of the white noise process serves

to define the probit link function in the context of generalized linear models. In contrast, the Clipped Gaussian model is non-hierarchical and completely specified at the one level of the spatial latent variable, $\mathbf{W}$. Under the no-nugget assumption for the covariance of $\mathbf{W}$, the Double Latent model will always differ from the Clipped Gaussian model in its extra source of variability. If one were to allow the spatial covariance structure to include a fixed nugget term, then under the probit link, the Clipped Gaussian model with the nugget fixed at one is a special case of the more general Double Latent model. Incorporation of a nugget term in the specification of $\mathbf{W}$ was not discussed by De Oliveira (1997, 2000). Section 3.1 further elucidates the consequences of this difference in the variance structure through an investigation of the correlation induced by the two models on the categorical response variable.

2.4.1 Gibbs algorithm for the Clipped Gaussian model

The Gibbs sampling algorithm for the Clipped Gaussian model parallels the algorithm proposed by De Oliveira (1997, 2000). Unlike De Oliveira's binary case, however, we must sample the constrained cut-point parameters, thus increasing the level of complexity of the algorithm and leading to additional challenges in terms of convergence. The Gibbs algorithm samples in turn from the complete conditional distributions of the unknown parameters and latent variable. The algorithm iteratively proceeds through the following steps,

1. Update the spatial parameters, γ, using the Metropolis-Hastings algorithm.

2. Update the latent spatial variable, $\mathbf{W}$, using the truncated multivariate Gaussian complete conditional distribution.

3. Update the cut-point parameters, θ, via reparameterization and a blocking algorithm.

4. Update the regression parameters, $\boldsymbol{\beta}$, using their multivariate Gaussian complete conditional distribution.

The details involved in these steps and the derivation of the complete conditionals are now described.

Spatial parameters, $\boldsymbol{\gamma} = (\sigma^2, \psi)$

The spatial parameters, $\boldsymbol{\gamma} = (\sigma^2, \psi)$, are specified at a level in the hierarchy resulting in conditional independence between $\boldsymbol{\gamma}$ and the categorical response, $\mathbf{y}$. Thus, the complete conditional distribution depends only on the latent variable, $\mathbf{W}$, and not on the observed data. It can be factorized as

$$\begin{aligned} f(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}) &= f(\boldsymbol{\gamma}|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}) \\ &\propto f(\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{\beta}, \boldsymbol{\theta}) f(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ &\propto f(\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{\beta}). \end{aligned} \tag{2.46}$$

It is customary to assume independence between the components of $\boldsymbol{\gamma} = (\sigma^2, \psi)$, and to use proper, vague priors for both components. They can then be updated independently using two successive M-H steps to sample from the un-recognizable complete conditional distributions. We choose candidate values uniformly over the parameter spaces defined by the priors so that the M-H ratios only involve $f(\mathbf{w}|\boldsymbol{\beta}, \sigma^2, \psi) f(\sigma^2, \psi)$ evaluated at the appropriate values. As described for the Double Latent model, our algorithm is ultimately simplified by fixing σ^2 and specifying a correlation function with only one parameter to estimate. We provide justification and detail regarding both of these simplifications in Chapter 3, along with a description of our choice of prior distributions and hyperparameter values.

Spatial latent vector, $\mathbf{W}$

The assumption that $\mathbf{W}$ follows a Gaussian process implies that any vector of observations at n locations from one realization of the process has a multivariate

Gaussian distribution,

$$\mathbf{W}_n \sim MVN_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\gamma})). \tag{2.47}$$

Let $\mathbf{W}_{-i}$ be the $\mathbf{W}$ vector with the ith element deleted. The full conditional distribution for each component of $\mathbf{W}$, W_i, is given by

$$\begin{aligned} f(w_i|\mathbf{w}_{-i}, \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &= f(w_i|\mathbf{w}_{-i}, y_i, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\propto f(y_i|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \\ &\propto f(y_i|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) \end{aligned} \tag{2.48}$$

for $i = 1, 2, \ldots, n$.

The latent variable, $\mathbf{W}$, and the categorical response variable, $\mathbf{Y}$, are related via a deterministic function, as given in (2.41). Therefore, when the value of W_i is known, the probability that $Y_i = k$ is either one or zero, specified by an indicator function. That is,

$$\begin{aligned} f(y_i|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) &= I_{\{\theta_{y_i-1} \leq w_i \leq \theta_{y_i}\}} \\ &= I_{\{w_i \in A(y_i)\}}. \end{aligned} \tag{2.49}$$

The multivariate Gaussian distribution of $\mathbf{W}$ specifies a univariate Gaussian conditional distribution for one component of $\mathbf{W}$ given all other components and the unknown parameters, $f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$. The parameters of the distribution are easily calculated using the formula for finding the conditional distribution of one component of jointly normally distributed vector of random variables, given the remaining variables (Hocking, 2003). We use identical notation to that developed for the Double Latent model in Section 2.3.1 when referring to needed sub-vectors and sub-matrices. Let $\mathbf{v}_i$ be the ith column of $\boldsymbol{\Sigma}$ with its ith element deleted, $\boldsymbol{\Sigma}_{-i}$ be the matrix $\boldsymbol{\Sigma}$ with its ith column and ith row removed, and $\mathbf{X}_{-i}$ be the $n \times p$

design matrix with its ith row removed. As usual, $\mathbf{x}_i$ is the ith row of $\mathbf{X}$. Write $\mathbf{W} = (W_i, \mathbf{W}_{-i})'$ and partition the covariance matrix, $\boldsymbol{\Sigma}$, into four parts,

$$\Sigma = \begin{pmatrix} v_{ii} & \mathbf{v}_i' \\ \mathbf{v}_i & \boldsymbol{\Sigma}_{-i} \end{pmatrix}, \tag{2.50}$$

where v_{ii} is the variance of W_i, $\mathbf{v}_i$ is the vector of covariances between W_i and $\mathbf{W}_{-i}$, and $\boldsymbol{\Sigma}_{-i}$ is the variance-covariance matrix of $\mathbf{W}_{-i}$. Then, $f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ is the probability density function of the Gaussian distribution

$$N(\mathbf{x}_i\boldsymbol{\beta} + \mathbf{v}_i'\boldsymbol{\Sigma}_{-i}^{-1}(\mathbf{w}_{-i} - \mathbf{X}_{-i}\boldsymbol{\beta}) \quad , \quad \boldsymbol{\Sigma}_{ii} - \mathbf{v}_i'\boldsymbol{\Sigma}_{-i}^{-1}\mathbf{v}_i). \tag{2.51}$$

It follows that the complete conditional distribution for one component of $\mathbf{W}$ is proportional to

$$f(w_i|\mathbf{w}_{-i}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) I_{w_i \in A(y_i)}, \tag{2.52}$$

which is a truncated univariate Gaussian distribution and can be sampled from with standard techniques.

Cut-point parameters, $\boldsymbol{\theta}$

The algorithm of De Oliveira (1997, 2000) applies only to binary data and simply requires one fixed cut-point, skirting the problem of estimating the set of highly constrained $K - 2$ cut-point parameters required for the ordered categorical data. For the ordered multi-category case, however, we must estimate $K - 2$ cut-point parameters. We accomplish this using the same strategy as described for the Double Latent model, applying the transformation proposed by Albert and Chib (1997) and the blocking algorithm proposed by Cowles (1996). The blocking method samples from the joint conditional distribution of the spatial latent variable, $\mathbf{W}$, and the cut-points, $\boldsymbol{\theta}$. The goal of the blocking is to reduce the correlation in the MCMC chain, thus accelerating convergence, and was introduced in more detail in Section 2.3.1. The strategy essentially results in sampling the cut-points from their complete

conditional distribution after marginalizing over the latent variable, $\mathbf{W}$, in contrast to the Double Latent model where we marginalize over the non-spatial latent variable $\mathbf{Z}$. The factorization of the joint conditional distribution again provides insight into this step,

$$f(\boldsymbol{\theta}, \mathbf{w}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = f(\mathbf{w}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}) f(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}). \tag{2.53}$$

As first described in Section 2.3.1, the de-constraint transformation of the cut-points expresses $\boldsymbol{\theta} = (\theta_2, \ldots, \theta_{K-1})$ as the unconstrained and real-valued vector $\boldsymbol{\alpha} = (\alpha_2, \ldots, \alpha_{K-1})$, where $\alpha_2 = \log(\theta_2)$ and $\alpha_k = \log(\theta_k - \theta_{k-1})$ for $k = 3, \ldots, K-1$. The inverse transformation is then given by $\theta_k = \sum_{i=2}^{k} \exp(\alpha_i)$. The target conditional distribution is

$$f(\boldsymbol{\alpha}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\alpha}). \tag{2.54}$$

The factorization in (2.54) is allowed by the assumed independence between $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$. The density $f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ is the multi-dimensional integral given as the original likelihood function of interest in (2.44). To use the M-H algorithm to sample from this non-standard distribution, we must evaluate both $f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $f(\boldsymbol{\alpha})$ at current and proposed values of the transformed cut-point parameters. These values are readily available from evaluations of multivariate Gaussian cumulative and probability density functions.

The prior and proposal distributions for $\boldsymbol{\alpha}$ are identical to those described for the Double Latent model, leading to the following transition probability defined by the M-H ratio,

$$\min \left\{ \frac{f(\mathbf{y}|\boldsymbol{\alpha}^{cand}, \boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\alpha}^{cand}) f_T(\boldsymbol{\alpha}^{cur}|\hat{\boldsymbol{\alpha}}, \psi^2 V, \nu)}{f(\mathbf{y}|\boldsymbol{\alpha}^{cur}, \boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\alpha}^{cur}) f_T(\boldsymbol{\alpha}^{cand}|\hat{\boldsymbol{\alpha}}, \psi^2 V, \nu)}, 1 \right\}. \tag{2.55}$$

Although it is not directly used in the algorithm because of the aforementioned transformation, we can write out the complete conditional distribution for the untransformed cut-point parameters as a uniform distribution. The derivation proceeds analogously to that described for the Double Latent model in Equations (2.21) and

(2.22), and thus we limit the details here. We again specify a uniform prior for $\boldsymbol{\theta}$ such that $f(\theta_k|\theta_{k-1},\theta_{k+1}) \propto I_{(\theta_{k-1}<\theta_k<\theta_{k+1})}$. First, we factorize the complete conditional distribution as

$$\begin{aligned} f(\theta_k|\mathbf{w},\mathbf{y},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta}_{-k}) &\propto f(\mathbf{y}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{y}|\mathbf{w},\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\gamma})f(\boldsymbol{\theta}) \\ &\propto f(\mathbf{y}|\boldsymbol{\theta},\mathbf{w},\boldsymbol{\beta},\boldsymbol{\gamma})f(\theta_k|\theta_{k-1},\theta_{k+1}). \end{aligned}$$

We can then write $f(\mathbf{y}|\boldsymbol{\theta},\mathbf{w},\boldsymbol{\beta},\boldsymbol{\gamma})f(\theta_k|\theta_{k-1},\theta_{k+1})$ as

$$\begin{aligned} &\prod_{i=1}^{n} I_{\{\theta_{y_i-1}\le w_i<\theta_{y_i}\}} I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}} = \\ &\qquad \prod_{i=1}^{n}\left[\sum_{j=1}^{K} I_{\{y_i=j\}} I_{\{\theta_{j-1}<w_i\le\theta_j\}}\right] I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}} \\ &\propto \prod_{i:y_i\in\{k,k+1\}} \left[I_{\{y_i=k\}} I_{\{\theta_{k-1}<w_i\le\theta_k\}} + I_{\{y_i=k+1\}} I_{\{\theta_k\le w_i<\theta_{k+1}\}}\right] I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}} \\ &\propto \prod_{i:y_i\in\{k,k+1\}} \left[I_{\{\theta_{y_i-1}<w_i\le\theta_{y_i}\}}\right] I_{\{\theta_{k-1}<\theta_k<\theta_{k+1}\}} \end{aligned} \tag{2.56}$$

As similarly described by Albert and Chib (1993), this complete conditional for θ_k is uniform over the interval

$$\left[\max\{\max(w_i; y_i = k), \theta_{k-1}\}\ ,\ \ \min\{\min(w_i; y_i = k+1), \theta_{k+1}\}\right]. \tag{2.57}$$

Regression parameters, $\boldsymbol{\beta}$

The full conditional distribution for the regression parameters, $\boldsymbol{\beta}$, is derived similarly to that described for the Double Latent model;

$$\begin{aligned} f(\boldsymbol{\beta}|\mathbf{y},\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma}) &= f(\boldsymbol{\beta}|\mathbf{w},\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\gamma})f(\boldsymbol{\beta}|\boldsymbol{\theta},\boldsymbol{\gamma}) \\ &\propto f(\mathbf{w}|\boldsymbol{\beta},\boldsymbol{\gamma})f(\boldsymbol{\beta}), \end{aligned} \tag{2.58}$$

assuming prior independence between $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{\gamma}$.

The distribution of $\mathbf{W}|\boldsymbol{\beta}, \boldsymbol{\gamma}$ is multivariate Gaussian by assumption and we further assume a multivariate Gaussian prior distribution such that, $\boldsymbol{\beta} \sim MVN(\mathbf{b}_0, \mathbf{B}_0)$, resulting in a multivariate Gaussian complete conditional distribution. Collecting the terms involving $\boldsymbol{\beta}$, we obtain the standard linear models result,

$$\boldsymbol{\beta}|\boldsymbol{\theta}, \gamma, \mathbf{y}, \mathbf{w} \sim MVN_p(\tilde{\mathbf{b}}, \tilde{\mathbf{B}}) \tag{2.59}$$

where

$$\tilde{\mathbf{B}} = (\mathbf{B}_0^{-1} + \mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}, \tag{2.60}$$

and

$$\begin{aligned} \tilde{\mathbf{b}} &= (\mathbf{B}_0^{-1} + \mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^T\Sigma^{-1}\mathbf{w}) && (2.61) \\ &= \tilde{\mathbf{B}}(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^T\Sigma^{-1}\mathbf{w}). && (2.62) \end{aligned}$$

The algorithms for the Double Latent and Clipped Gaussian models described in this chapter are implemented using R software (R Development Core Team, 2007). The following chapter provides further comparison of the Double Latent and Clipped Gaussian models, along with discussions of the details related to specifying and applying these models in practice.

Chapter 3

MODEL SPECIFICATION IN PRACTICE

Fitting the models in practice requires more information and thought than simply following the details laid out in Chapter 2. This chapter provides discussion of several topics to improve understanding of the models and their differences. It also contains needed information about fitting the models in practice. First, in Section 3.1, we provide a comparison of the models in terms of the second-order structure of the categorical response variables induced by the continuous distributions of the latent variables. This provides insight into how model structure and level of the clipping ultimately translate into correlation in the categorical random field. In Section 3.2, we discuss parameter identifiability for the variance, the spatial correlation scale parameter, and the estimation of the cut-point parameters. We then discuss specification of prior distributions in Section 3.3, specifically those for the correlation scale parameter and the cut-point parameters. Finally, we briefly introduce methods that can be used to check model assumptions and assess convergence of the MCMC algorithms.

3.1 Correlation in the categorical random variable

Properties of the two models can be further elucidated by investigation of the distributions of the latent variables that are ultimately clipped to form the categorical random variables. These distributions directly induce correlation into the categorical random field. The results of this section are not directly used in estimation and prediction, but provide insight into how the structure and assumptions of each model

translate into the spatial correlation structure. Gelfand et al. (2000) and De Oliveira (1997) provide similar, though simpler, investigations for the case of binary data. For both models, the correlation structure of the latent spatial process depends only on the spatial parameters, γ. This is *not* true, however, for the underlying categorical random field, where the correlation structure depends on the spatial parameters, γ, the regression parameters, β, the cut-point parameters, θ, and the values of the covariates, $\mathbf{X}$.

An additional goal of this chapter is to aid in the process of choosing parameter values for the simulation study in Chapter 5. In typical simulation studies involving continuous spatially correlated variables, the correlation between values at different locations is easily changed by varying the scale (or range) parameter and/or smoothness parameter in the correlation function. Our case is more challenging due to the large number of factors ultimately contributing to the correlation. The situation is clearly simplified for the constant mean model, and further for the zero-mean model. De Oliveira (1997) addresses only the constant mean case in detail. However, our model specification and simulation study do involve the incorporation of a non-constant mean and covariate values, and thus we also devote attention to this case. This section is devoted to the derivation of the distributions of the latent continuous variables and to describing the correlation in the categorical random field as a function of the many parameters affecting it. It includes a summary of a graphical investigation of correlation for changing values of the regression parameters, covariate values, cut-point parameters, and distance between locations for both models.

Recall that the Double Latent model is specified hierarchically, with a conditionally independent first stage for the latent variable, $\mathbf{Z}$, that is clipped to give rise to the categorical random variable, $\mathbf{Y}$, where

$$Z(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)\beta + W(\mathbf{s}_i) + \epsilon(\mathbf{s}_i),$$

$\epsilon(\mathbf{s}_i) \sim N(0,1)$, and $\mathbf{W} \sim GP(\mathbf{0}, \Sigma(\gamma))$. Due to the conditional independence of $\mathbf{Z}$ given the spatial process, $\mathbf{W}$, we must investigate the marginalized distribution of $\mathbf{Z}$ to obtain insight into the correlation structure of $\mathbf{Y}$. The zero-mean of the Gaussian spatial process, $\mathbf{W}$, results in a marginal expectation equal to $\mathbf{X}\boldsymbol{\beta}$. The marginal variance of $\mathbf{Z}$ incorporates the variance-covariance of $\boldsymbol{\epsilon}$, which is assumed to be the n-dimensional identity matrix, $\mathbf{I}_n$, and the variance-covariance of the latent Gaussian spatial process, $\mathbf{W}$, given by $\Sigma(\gamma) = \sigma^2 H(\psi)$. Recall that $[H(\psi)]_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \psi) = \rho_{ij}$, where ρ_{ij} is a correlation function that results in a valid variance-covariance matrix. The joint distribution of the latent variable at two different locations is bivariate normal (MVN_2),

$$\begin{pmatrix} Z(\mathbf{s}_i) \\ Z(\mathbf{s}_j) \end{pmatrix} \sim MVN_2\left(\begin{pmatrix} \mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta} \\ \mathbf{x}(\mathbf{s}_j)\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} 1+\sigma^2 & \sigma^2\rho_{ij} \\ \sigma^2\rho_{ij} & 1+\sigma^2 \end{pmatrix}\right), \quad i \neq j. \tag{3.1}$$

In contrast, the Clipped Gaussian model is specified non-hierarchically, with the distribution of interest specified directly through the Gaussian process,

$$\mathbf{W} \sim GP(\mathbf{X}\boldsymbol{\beta}, \Sigma(\gamma)).$$

It is then straightforward to write down the distribution of the latent variable that will be clipped,

$$\begin{pmatrix} W(\mathbf{s}_i) \\ W(\mathbf{s}_j) \end{pmatrix} \sim MVN_2\left(\begin{pmatrix} \mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta} \\ \mathbf{x}(\mathbf{s}_j)\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2\rho_{ij} \\ \sigma^2\rho_{ij} & \sigma^2 \end{pmatrix}\right). \tag{3.2}$$

It is immediately clear from (3.1) and (3.2) that the difference in the underlying correlation structure in the latent variables enters into the common variance of the random variables across the region of interest. This difference clearly stems from the Double Latent model's white noise variance component defined through the chosen link function. Specifically, the variance used to specify the link function results in larger values for the diagonal elements of the marginal variance-covariance matrix, as

shown in (3.1). As previously discussed, this difference is equivalent to the comparison between a nugget and no-nugget model in the language of geostatistics, where the nugget term is fixed at 1.

Thus far, we have assumed that ϵ is a standard normal random variable, specifying the probit link function. A more general Double Latent model can be specified as $\epsilon \sim MVN_n(\mathbf{0}, \tau^2\mathbf{I}_n)$ giving diagonal elements of $(\tau^2 + \sigma^2)$ instead of $(1 + \sigma^2)$ in (3.1). However, care should be taken in ensuring that both of these parameters are estimable. As previously mentioned, the variance component, τ^2, is analogous to the incorporation of a nugget term in classical geostatistical modeling, referring to either microscale variation or measurement error. It is also possible to further generalize the model by specifying a different distribution for ϵ. For example, Albert and Chib (1993) specify a t-distribution and then estimate the degrees of freedom of that distribution to define a "problem specific" link function, theoretically leading to a more robust model. However, for identifiability, we restrict our models to the probit link and also fix τ^2 and σ^2 at 1 for reasons described in Section 3.2, leading to fixed values for the diagonal elements of the variance-covariance matrix of $\mathbf{Y}$; with a value of 2 for the Double Latent model in (3.1) and 1 for the Clipped Gaussian model in (3.2). This specification directly leads to the creation of different categorical random fields after clipping and a different correlation structure in the resulting categorical random field. For example, holding all parameter values constant, data simulated under the Double Latent Model will likely contain more observations in one or both of the end categories since the continuous latent variable has a larger variance than the Clipped Gaussian model.

We now compare the two models with respect to the correlation in the categorical response vector, $\mathbf{Y}$. To proceed, we first assume that $Y(\mathbf{s}_i) = Y_i$ takes integer values from $1, \ldots, K$, where K is the total number of categories. Recall the relationships between Y_i and Z_i for the Double Latent model (2.2) and Y_i and W_i for the Clipped

Gaussian model (2.40). To calculate the correlation between the categorical random variables at two locations, Y_i and Y_j, we must calculate the usual quantities, $E[Y_i]$, $E[Y_j]$, $E[Y_i^2]$, $E[Y_j^2]$, and $E[Y_iY_j]$. In the following calculations, we assume $\tau^2 = 1$ but σ^2 is left to be estimated. To obtain $E[Y_i]$ and the $E[Y_i^2]$ for the Double Latent model, we simply need the unconditional $Pr(Y_i = k)$ for each $k = 1, \ldots, K$;

$$\begin{aligned} Pr(Y_i = k) &= Pr(\theta_{k-1} < Z_i \leq \theta_k) \\ &= \Phi\left(\frac{\theta_k - \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{\theta_{k-1} - \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{1+\sigma^2}}\right), \end{aligned} \tag{3.3}$$

for the cut-point vector $(-\infty, 0, \theta_2, \theta_3, \ldots, \theta_{K-1}, \infty)$. For the Clipped Gaussian model we have,

$$\begin{aligned} Pr(Y_i = k) &= Pr(\theta_{k-1} < W_i \leq \theta_k) \\ &= \Phi\left(\frac{\theta_k - \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{\sigma^2}}\right) - \Phi\left(\frac{\theta_{k-1} - \mathbf{x}_i\boldsymbol{\beta}}{\sqrt{\sigma^2}}\right). \end{aligned} \tag{3.4}$$

Calculation of the covariance requires calculation of the joint probability that $Y_i = k$ and $Y_j = l$ for all $k, l = 1, \ldots, K$, clearly involving the correlation specified by the marginal distributions of $\mathbf{Z}$ and $\mathbf{W}$. For the Double Latent model, the joint probability is given by

$$\begin{aligned} Pr(Y_i = k, Y_j = l) &= Pr(\theta_{k-1} < Z_i \leq \theta_k, \theta_{l-1} < Z_j \leq \theta_l) \\ &= \int_{\theta_{k-1}-\mathbf{x}_i\boldsymbol{\beta}}^{\theta_k-\mathbf{x}_i\boldsymbol{\beta}} \int_{\theta_{l-1}-\mathbf{x}_j\boldsymbol{\beta}}^{\theta_l-\mathbf{x}_j\boldsymbol{\beta}} \phi_2(\boldsymbol{\Sigma}_{z,ij}) dz_i dz_j \qquad (3.5) \\ &= \int_{\frac{(\theta_{k-1}-\mathbf{x}_i\boldsymbol{\beta})}{\sqrt{1+\sigma^2}}}^{\frac{(\theta_k-\mathbf{x}_i\boldsymbol{\beta})}{\sqrt{1+\sigma^2}}} \int_{\frac{(\theta_{l-1}-\mathbf{x}_j\boldsymbol{\beta})}{\sqrt{1+\sigma^2}}}^{\frac{(\theta_l-\mathbf{x}_j\boldsymbol{\beta})}{\sqrt{1+\sigma^2}}} \phi_2(H_{ij}) dz_i dz_j, \qquad (3.6) \end{aligned}$$

where $\phi_2(\cdot)$ is the 2-dimensional multivariate Gaussian probability density function (pdf). Equation (3.6) is specified in terms of the variance-covariance matrix, $\boldsymbol{\Sigma}_{z,ij}$, as given in (3.1), and (3.6) is specified in terms of the correlation matrix $H_{ij} = \rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\psi})$.

Similarly, for the Clipped Gaussian model, the joint probability is given by

$$\begin{aligned} Pr(Y_i = k, Y_j = l) &= Pr(\theta_{k-1} < W_i \leq \theta_k, \theta_{l-1} < W_j \leq \theta_l) \\ &= \int_{\theta_{k-1}-\mathbf{x}_i\boldsymbol{\beta}}^{\theta_k-\mathbf{x}_i\boldsymbol{\beta}} \int_{\theta_{l-1}-\mathbf{x}_j\boldsymbol{\beta}}^{\theta_l-\mathbf{x}_j\boldsymbol{\beta}} \phi(\boldsymbol{\Sigma}_{w,ij}) dw_i dw_j \qquad (3.7) \\ &= \int_{\frac{(\theta_{k-1}-\mathbf{x}_i\boldsymbol{\beta})}{\sigma}}^{\frac{(\theta_k-\mathbf{x}_i\boldsymbol{\beta})}{\sigma}} \int_{\frac{(\theta_{l-1}-\mathbf{x}_j\boldsymbol{\beta})}{\sigma}}^{\frac{(\theta_l-\mathbf{x}_j\boldsymbol{\beta})}{\sigma}} \phi_2(H_{ij}) dw_i dw_j, \qquad (3.8) \end{aligned}$$

where $\boldsymbol{\Sigma}_{w,ij}$ is the variance-covariance matrix of the Gaussian process. Note that H_{ij} is identical for the two models.

The needed expectations for both models are calculated using the above probabilities and the integer values assigned as the category labels. The correlation between Y_i and Y_j is calculated in the usual manner,

$$Corr[Y_i, Y_j] = \frac{E[Y_i, Y_j] - E[Y_i]E[Y_j]}{\sqrt{(E[Y_i^2] - E[Y_i]^2)(E[Y_j^2] - E[Y_j]^2)}}. \qquad (3.9)$$

Thus, the correlations are based on the probabilities defined in (3.3) and (3.6) for the Double Latent model, and (3.4) and (3.8) for the Clipped Gaussian model, clearly demonstrating that the correlation in the categorical random field depends not only on the parameters of the correlation structure in the latent spatial variable, but also on the cut-point parameters, the regression parameters, and values of the covariates.

We are specifically interested in investigating how the correlation defined by (3.9) varies with the values of the regression parameters. The equations for the correlations do not easily simplify to a form that makes the relationship between the regression parameters and the correlation obvious. Thus, we investigate the relationship numerically and graphically for the case of $K = 4$ categories and an exponential correlation function parameterized in terms of scale parameter, ϕ, which is often referred to as the reciprocal of the parameter describing the range of spatial correlation ($\phi = 1/\text{range}$). In terms of previous notation, $\boldsymbol{\gamma} = (\sigma^2, \psi)$ and $\psi = \phi$.

For simplicity, we first investigate a constant mean model and then address the case of a spatially varying mean.

For the case of a constant mean, β_0, the formula for the correlation in the Double Latent model can be written as

$$\frac{\sum_{k=1}^{K}\sum_{l=1}^{K} kl \int_{\theta_{k-1}-\beta_0}^{\theta_k-\beta_0}\int_{\theta_{l-1}-\beta_0}^{\theta_l-\beta_0} \phi(z_i, z_j; H_{ij})dz_i dz_j - \left[\sum_{k=1}^{K} kPr(Y_i = k)\right]^2}{\left[\sum_{k=1}^{K} k^2 Pr(Y_i = k) - \left[\sum_{k=1}^{K} kPr(Y_i = k)\right]^2\right]^2}, \quad (3.10)$$

where

$$Pr(Y_i = k) = \Phi\left(\frac{\theta_k - \beta_0}{\sqrt{1+\sigma^2}}\right) - \Phi\left(\frac{\theta_{k-1} - \beta_0}{\sqrt{1+\sigma^2}}\right). \quad (3.11)$$

For the Clipped Gaussian model, the form of (3.10) remains the same except the probabilities are instead defined as

$$Pr(Y_i = k) = \Phi\left(\frac{\theta_k - \beta_0}{\sigma}\right) - \Phi\left(\frac{\theta_{k-1} - \beta_0}{\sigma}\right). \quad (3.12)$$

Figures 3.1 and 3.2 display the changing correlation between two locations as a function of the constant mean for different values of ϕ and for $\boldsymbol{\theta} = (0, 1, 2)$ and $\boldsymbol{\theta} = (0, 1, 3)$. The horizontal lines represent the value of the correlation in the latent continuous Gaussian random field.

In Figure 3.1, note that the value of β_0 at which the correlation achieves its maximum depends on the vector of cut-point parameters. As expected, it also depends on the model for most values of the cut-points. The Clipped Gaussian and Double Latent models generally produce very similar relationships between the constant mean and the correlation (see Figure 3.1). For $\boldsymbol{\theta} = (0, 1, 2)$, the maximum for both models occurrs at $\beta_0 = 1$. It is immediately clear that the Clipped Gaussian model induces correlations of a higher magnitude than those of the Double Latent model for the same parameter values. Also, there is a clear difference in the shape of the curves for the two models when the cut-point parameters are set to $\boldsymbol{\theta} = (0, 1, 3)$, instead of $\boldsymbol{\theta} = (0, 1, 2)$ (Figure 3.2). The Clipped Gaussian model results in a slightly

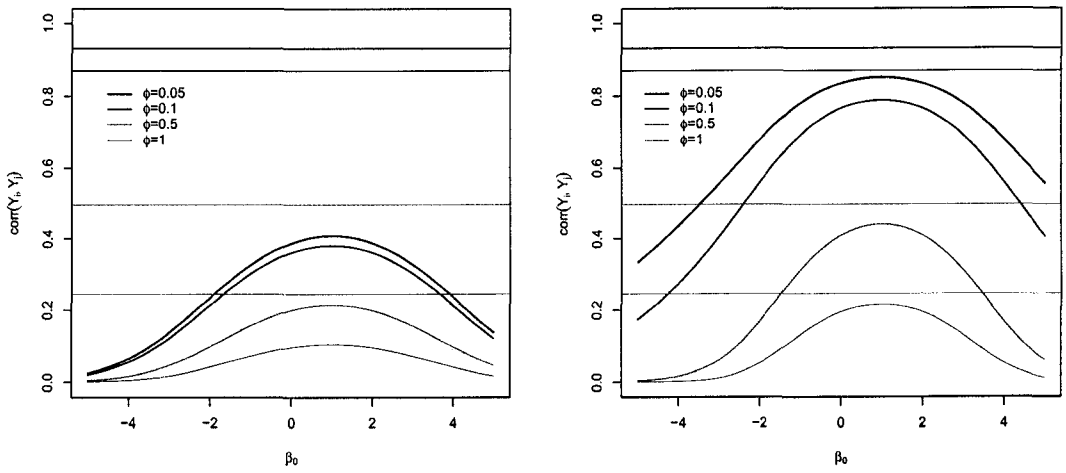

Figure 3.1: The correlation between the categorical random variables, $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$ where $\mathbf{s}_i$ and $\mathbf{s}_j$ are 1.41 units apart for the Double Latent (left) and Clipped Gaussian (right) models. The correlations are given as a function of a constant mean, β_0, for different values of the scale parameter of the exponential correlation function, ϕ. The cut-point parameters are $\boldsymbol{\theta} = (0, 1, 2)$. The horizontal lines show the value of the correlation specified in the continuous spatial process. The maximum correlation occurs at $\beta_0 = 1$ for all values of ϕ and both models.

bimodal shape, whereas the Double Latent maintains the more symmetric, unimodal shape exhibited for $\boldsymbol{\theta} = (0, 1, 2)$. Increasing the distance between the two locations of interest lowers the values of the correlation, as is expected, but retains the general shape of the curves. The same is true for increasing the value of the scale parameter of the correlation function, ϕ, as is demonstrated in the plots.

It is clear from Figures 3.1 and 3.2 that the correlation in the categorical random field is always less than that in the continuous random variable, a fact also observed by De Oliveira (1997), who demonstrated that the correlation structure of his clipped binary random field is always weaker than that of the underlying Gaussian random field. He observes that this effect is very apparent when the absolute value of β_0 gets large, as this results in values for $Pr(Y(\mathbf{s}) = 1)$ that are close to either 1 or 0, representing the extreme cases for realizations of the binary random field. This is displayed in the plots by the unimodal shape of the correlation as a function of

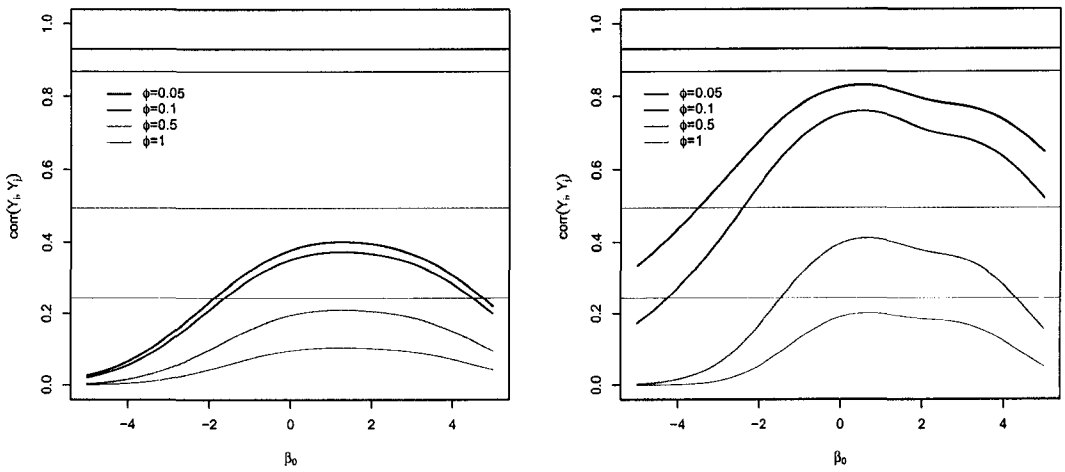

Figure 3.2: The correlation between the categorical random variables, $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$ where $\mathbf{s}_i$ and $\mathbf{s}_j$ are 1.41 units apart for the Double Latent (left) and Clipped Gaussian (right) models. Correlation is plotted as a function of a constant mean, β_0, for different values of the scale parameter of the exponential correlation function, ϕ. The cut-point parameters are $\boldsymbol{\theta} = (0, 1, 3)$. The horizontal lines show the value of the correlation specified in the continuous spatial process. The maximum correlation occurs at approximately $\beta_0 = (1.250, 1.255, 1.256, 1.238)$ for the Double Latent and $\beta_0 = (0.535, 0.548, 0.651, 0.678)$ for the Clipped Gaussian for $\phi = (0.05, 0.1, 0.5, 1)$, respectively.

β_0. For the multi-category case, extreme cases also result when the $Pr(Y(\mathbf{s}) = k)$ is close to 1 for the category k. This would result in the prediction of a categorical random field nearly covered with a single category. De Oliveira (1997) provides further intuition by appealing to a result from Wise et al. (1997), stating that this is a demonstration of the fact that a non-linear transformation of a Gaussian process always has a *whitening*, or lessening, effect on the correlation. The Double Latent model results in a level of correlation farther from the value of the latent process than the Clipped Gaussian due to the additional source of noise incorporated by the variance of the second latent variable.

We also investigate the correlation for a spatially varying mean depending on one covariate and thus defined with two regression parameters, β_0 and β_1. We create

3-dimensional plots illustrating the relationship between the regression parameters and the correlation between the values at two locations (Figures 3.3 and 3.4). As with the constant mean case, the general shape of the surface does not change appreciably for different values of the correlation scale parameter, ϕ, or for different distances between the locations. The magnitude of the correlation *does* change, but our conclusion regarding for what value of $\boldsymbol{\beta}$ the correlation obtains an approximate maximum does not change. Thus, we are able to make general conclusions about what combinations of β_0 and β_1 induce high and low correlations for different pairs of covariate values. We use standardized covariates, resulting in covariate values in the range of approximately -2.5 to 2.5. Thus, we restrict our attention to covariate values in this range and find that the general shape of the surface is retained for pairs of covariate values defined by their signs. For example, we obtain one general shape when both covariate values are negative, one when they are both positive, and another for one positive value and one negative value. We give a series of plots to demonstrate these shapes for both the Double Latent model (Figure 3.3) and Clipped Gaussian model (Figure 3.4).

Notice from Figures 3.3 and 3.4 that both positive or both negative values result in a flat ridge running northwest-southeast or southwest-northeast, respectively. One positive value and one negative value results in a more peaked shape running east-west. Thus, as we would expect, the maximum depends on the covariate values. However, we can make some general conclusions to help us choose values of the regression parameters that correspond to different levels of correlation in the categorical response to be used in the simulation study in Chapter 5. We chose to compare data generated from $\boldsymbol{\beta} = (-0.5, 1)$ and $\boldsymbol{\beta} = (2, 1)$, where $\boldsymbol{\beta} = (2, 1)$ results in higher correlation for most combinations of covariate values, the exception being when both are positive.

In conclusion, this section numerically investigated the correlation in a categorical random field created by clipping a continuous random variable. For the constant

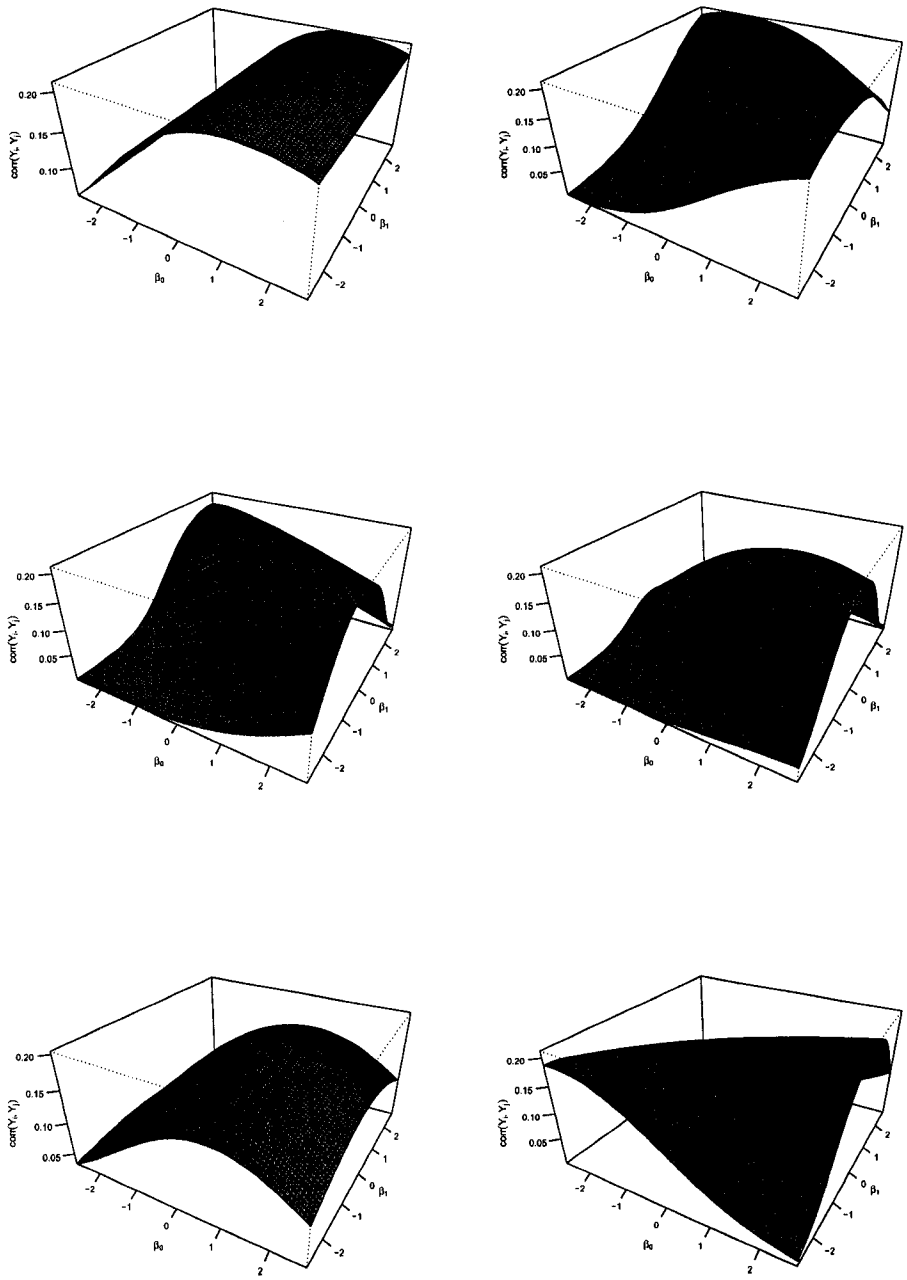

Figure 3.3: *Double Latent model*: The correlation between the categorical random variables, $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$, where $\mathbf{s}_i$ and $\mathbf{s}_j$ are 1.41 units apart, as a function of the regression parameters, β_0 and β_1 for covariate values of $X_i = 0$ and $X_j = 0$ (upper left), $X_i = 1$ and $X_j = 1$ (upper right), $X_i = 1$ and $X_j = 2$ (middle left), $X_i = -1$ and $X_j = 2$ (middle right), $X_i = 0.5$ and $X_j = -0.5$ (lower left), and $X_i = -1$ and $X_j = -2$ (lower right). The cut-point parameters are $\boldsymbol{\theta} = (0, 1, 3)$, and $\phi = 0.5$.

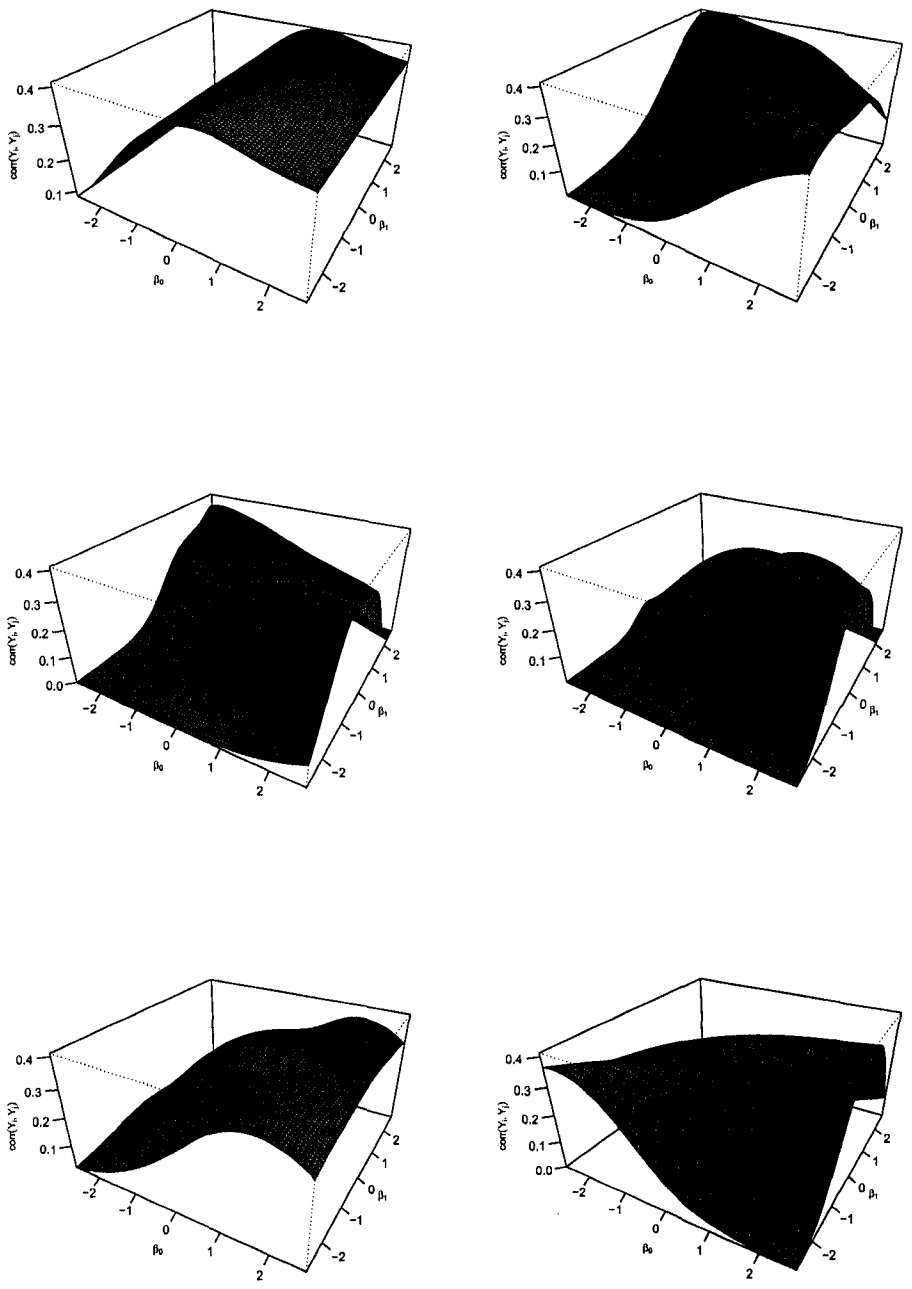

Figure 3.4: *Clipped Gaussian Model*: The correlation between the categorical random variables, $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$, where $\mathbf{s}_i$ and $\mathbf{s}_j$ are 1.41 units apart, as a function of the regression parameters, β_0 and β_1 for covariate values of $X_i = 0$ and $X_j = 0$ (upper left), $X_i = 1$ and $X_j = 1$ (upper right), $X_i = 1$ and $X_j = 2$ (middle left), $X_i = -1$ and $X_j = 2$ (middle right), $X_i = 0.5$ and $X_j = -0.5$(lower left), and $X_i = -1$ and $X_j = -2$(lower right). The cut-point parameters are $\boldsymbol{\theta} = (0, 1, 3)$, except for the last plot where $\boldsymbol{\theta} = (0, 1, 2)$, and the spatial scale parameter is $\phi = 0.5$.

mean case, we produced the characteristic shapes of the correlation as a function of the mean under several spatial scale parameters and two sets of cut-point parameters. This investigation allowed visualization of how the extra variance in the Double Latent model translates into the structure of the categorical random field, and aided in our choice of regression parameters for the simulation study undertaken in Chapter 5. To be more thorough, we also introduce notations and equations for the case when categories are not simply labeled by successive integer values.

3.1.1 Correlation structure for the non-integer labels

In the previous section, we defined our categorical response as taking on successive integer values from 1 to K, from here on referred to as the *integer-label* definition. For the ordered categorical data models we are developing, these labels are not meant to carry specific meaning; it is *not* assumed that the "distance" between $k = 1$ and $k = 2$ is the same as the "distance" between $k = 2$ and $k = 3$. Instead, the integer-label definition is used for notational and mathematical convenience. It provides a logical framework to initiate the investigation of the correlation structure in the categorical random field. De Oliveira (2000) uses the integer-label definition when referring to the multi-category response problem.

However, the problem can be viewed more generally from a multinomial-like perspective, where the response for each location is a vector indicating category membership, rather than an integer from the set $\{1, \ldots, K\}$. We briefly referred to this definition of the response in Chapter 1, and we now term it the *indicator-label* definition. That is, we define the vector

$$U(\mathbf{s}_i) = [I_{\{W(\mathbf{s}_i) \leq \theta_1\}},\ I_{\{\theta_1 < W(\mathbf{s}_i) \leq \theta_2\}},\ \ldots,\ I_{\{\theta_{K-2} < W(\mathbf{s}_i) \leq \theta_{K-1}\}},\ I_{\{\theta_{K-1} < W(\mathbf{s}_i)\}}]. \quad (3.13)$$

For example, if the value at location $\mathbf{s}_i$ falls into the second category of four, then under the integer-label framework $y(\mathbf{s}_i) = 2$, and under the indicator-label framework $u(\mathbf{s}_i) = (0, 1, 0, 0)$. The correlation between $U(\mathbf{s}_i)$ and $U(\mathbf{s}_j)$ is a $K \times K$ matrix

rather than the single value obtained for the correlation between $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$, thus making it more difficult to investigate how the correlation changes for differing values of the regression parameters, covariates, distance, and spatial parameters. As we expect, the analytical form of the correlation is very similar under both definitions.

We let $U_i[k]$ refer to the kth element of the vector $U(\mathbf{s}_i) = U_i$. For the Clipped Gaussian model with $\sigma^2 = 1$, $Corr(U_i[k], U_j[l])$ is equal to

$$\frac{\int_{\theta_{l-1}-\mathbf{x}_j\boldsymbol{\beta}}^{\theta_l-\mathbf{x}_j\boldsymbol{\beta}} \int_{\theta_{k-1}-\mathbf{x}_i\boldsymbol{\beta}}^{\theta_k-\mathbf{x}_i\boldsymbol{\beta}} \phi_2(w_i, w_j; H(\boldsymbol{\psi})_{ij}) dw_i dw_j - [g(k,i)g(l,j)]}{\sqrt{[g(k,i)(1-g(k,i))][g(l,j)(1-g(l,j))]}} \tag{3.14}$$

where

$$g(k,i) = \Phi(\theta_k - \mathbf{x}_i\boldsymbol{\beta}) - \Phi(\theta_{k-1} - \mathbf{x}_i\boldsymbol{\beta}). \tag{3.15}$$

For a constant mean, the expression simplifies to

$$Corr(U_i[k], U_j[l]) = \frac{\int_{\theta_{l-1}-\beta_0}^{\theta_l-\beta_0} \int_{\theta_{k-1}-\beta_0}^{\theta_k-\beta_0} \phi_2(w_i, w_j; H(\boldsymbol{\psi})_{ij}) dw_i dw_j - [m(k)m(l)]}{\sqrt{[m(k)(1-m(k))][m(l)(1-m(l))]}} \tag{3.16}$$

where

$$m(k) = \Phi(\theta_k - \beta_0) - \Phi(\theta_{k-1} - \beta_0). \tag{3.17}$$

We present the above definitions and formulas as a starting point for further investigation into the correlation structure when integer labels for the categories may not be appropriate. For example, it may be that the "distance" between the integer labels $k = 1$ and $k = 2$ is considered very different from the "distance" between $k = 2$ and $k = 3$, such that the integer labels themselves should not enter the calculations.

3.2 Parameter Identifiability

In this dissertation, we stress that our priority lies in predicting the categorical values at unobserved locations. However, investigation of the estimation of the parameters (or the posterior distributions) for simulated data provides needed insight

into the effectiveness of the model and the methods used to fit the models. Therefore, we devote this section to a discussion of potential challenges in fitting our models due to issues of nonidentifiability surrounding several of the parameters. Difficulties in the estimation of spatial covariance parameters for spatial models have been previously documented by several authors (Irvine et al., 2007; Xie and Carlin, 2006; Schmidt et al., 2007; Zimmerman, 2006; Lark, 2000; Zimmerman and Zimmerman, 1991). A possible reason for such estimation problems is lack of identifiability of the parameters.

We now provide a brief introduction to the concept of identifiability in the context of hierarchical Bayesian models. Our discussion closely follows that given by Xie and Carlin (2006). Identifiability is typically referred to as a property of the likelihood function, and is thus the same concept for frequentist and Bayesian inference. We term a parameter unidentifiable if the same likelihood results from two different values of the parameter. In the situation of likelihood unidentifiability, the data do not contain information needed to estimate the unidentifiable parameter. To solve this problem in the context of classical frequentist inference, constraints are added to the model in order to make the parameter identifiable. For Bayesian inference, however, an additional solution lies in the specification of the prior distribution to incorporate more information regarding the parameter. The use of prior distributions to help solve issues of nonidentifiability warrants caution. In this situation, a comparison of the resulting posterior distribution to the prior distribution should clearly be performed. Obviously, if the two distributions are identical, then even though the parameter was technically identifiable after specification of the prior distribution, the data contributed no information to the posterior distribution (and thus no information to estimate the parameter).

To investigate the issue of identifiability, the concepts of *Bayesian learning* and *Bayesian identifiability* have been developed, and are explored in detail by Poirier

(1998). Bayesian learning refers to differences between the posterior distribution of a parameter and its prior distribution, giving a measure of the extent to which the data inform the posterior distribution beyond the information included in the prior distribution. To illustrate this concept, assume that η_1 is a parameter of interest and $\mathbf{y}$ denotes the observed data. If $f(\eta_1|\mathbf{y}) = f(\eta_1)$, then no Bayesian learning took place.

Bayesian identifiability is based on conditional uninformativity of the data rather than marginal uninformativity used to define Bayesian learning. If η_2 is another parameter of interest, then conditional uniformativity is defined by the situation in which $f(\eta_1|\eta_2, \mathbf{y}) = f(\eta_1|\eta_2)$. Traditional nonidentifiability of a parameter implies that the data are conditionally uninformative about the parameter of interest, but conditional uninformativity does not imply the absence of Bayesian learning (Poirier, 1998). That is, since $f(\eta_1|\eta_2, \mathbf{y}) = f(\eta_1|\eta_2)$ does not imply $f(\eta_1|\mathbf{y}) = f(\eta_1)$, the issue of nonidentifiability in Bayesian models is often complicated by the conditional relationships among parameters. Thus, Bayesian learning can still occur even if a parameter is unidentifiable.

To help investigate these issues, Xie and Carlin (2006) propose measures based on differences in precision and Kullback-Leibler divergence and demonstrate their use for familiar Gaussian linear hierarchical models. The measures assess both what can be learned about a parameter given a finite amount of observed data and also what could theoretically be learned in the presence of an infinite amount of data. They suggest for future work that the measures could be useful in more complicated models such as those for non-Gaussian data characterized by one-parameter exponential family distributions.

In this section, we briefly explore the potentially nonidentifiable parameters in the most general specifications of the Double Latent and Clipped Gaussian models. As is typically the case for hierarchical models, these focus on the variance parameters

(Gelfand et al., 2000; Xie and Carlin, 2006). The identifiability issues arise from sources common to Gaussian hierarchical models in general and also due to the use of categorical data via the concept of clipping a continuous distribution. The Double Latent model is the subject of greater discussion due to the additional variance parameter that is naturally defined in its specification.

3.2.1 Identifiability of the variance parameters

For the Double Latent Model, the latent linear model component is defined as

$$Z(\mathbf{s}_i) = \mathbf{x}(\mathbf{s}_i)\boldsymbol{\beta} + W(\mathbf{s}_i) + \epsilon(\mathbf{s}_i). \tag{3.18}$$

Recall that we specify the distribution of $\epsilon(\mathbf{s}_i)$ as standard Gaussian, $N(0,1)$, to obtain the probit link function. A more general, and perhaps attractive, specification defines the variance of $\epsilon(\mathbf{s}_i)$ as τ^2 and leaves this parameter to be estimated. The parameter describes the variability in the latent continuous variable, given the spatial process, and as previously described, can be thought of as playing the role of a nugget term in classical geostatistics. However, there are identifiability problems with estimating this parameter from categorical data.

Lawrence et al. (2006) address this problem for the non-spatial case where the only variance component to estimate is τ^2, pointing out that the observed categorical data are invariant "up to monotone transformations of the underlying latent distribution" Lawrence et al. (2006). This implies that the variance of the underlying continuous random variable cannot be estimated from the information contained in the categorical response. To further explore the non-spatial model, assume that $\mathbf{Z} \sim N(\mathbf{x}\boldsymbol{\beta}, \tau^2\mathbf{I})$ and that we are attempting to estimate τ^2. We can form new regression coefficients, $\boldsymbol{\beta}' = \tau\boldsymbol{\beta}$ and new cut-point parameters, $\boldsymbol{\theta}'_k = \tau\boldsymbol{\theta}_k$ that result in the same observed probabilities for any value of τ. That is, τ will always cancel out of the equation for the probability, leading to $Pr(Y \leq k) = \Phi(\theta_k - \mathbf{x}\boldsymbol{\beta})$ for every

value of τ^2. The spatial model described in Section 2.3 behaves similarly. Thus, for the model to be identifiable, we must fix τ^2.

For spatial models in general, non-identifiability issues surrounding τ^2 are larger than the problem specific to ordered categorical data described in the previous paragraph. Recall that the spatial component of the model, $\mathbf{W}$, is defined through a Gaussian process with variance-covariance matrix $\Sigma(\gamma) = \sigma^2 H(\psi)$, where σ^2 is a common variance. Thus, there are two variance parameters to estimate if τ^2 is left to vary. Spatial models specified by the form (3.18), such as the Double Latent model with $\epsilon(\mathbf{s}_i) \sim N(0, \tau^2)$, cannot fully separate the common spatial variance, σ^2, from the nugget, or white noise, variance, τ^2 (Gelfand et al., 2000). This is not the result of clipping a continuous variable to model ordered categorical data, but is also the case in general for models of the form (3.18) for continuous data. Gelfand et al. (2000) and De Oliveira (1997, 2000) provide the algebraic details to show the non-identifiability of both τ^2 and σ^2, which is straight forward for the binary case. Not surprisingly, the non-identifiability often manifests in estimates that are very sensitive to prior specification (Gelfand et al., 2000). Thus, Gelfand et al. (2000) also set $\tau^2 = 1$, giving what they term a *neutral* specification, meaning that it is actually the ratio, σ^2/τ^2, that is estimated. This is equivalent to working with the "renormalized variogram."

While fixing τ^2 helps solve non-identifiability issues for models for continuous data, it does not eliminate the problem in the context of models for ordered categorical data. The additional issue for ordered categorical data lies in the amount of information available for estimating σ^2, and applies to both the Double Latent model and the Clipped Gaussian model. It is no surprise that categorizing a continuous distribution results in loss of information for estimating this parameter. This loss of information can make it impossible to estimate parameters of the continuous distribution using the categorical data. For the binary case of the Clipped Gaussian

model, De Oliveira (1997) boldly notes that the "binary data contain no information about σ^2." For example, if we think about clipping a Gaussian distribution at one fixed cut-point value to create a binary variable, it is clear that very different variances for the continuous distribution result in the same binary random field; it is the location parameter of the process that changes the binary random field, rather than the variance. In fact, infinitely many values of σ^2 can produce the same binary data. Thus, De Oliveira (2000) advises against attempting to estimate σ^2, instead setting the parameter to a fixed value. This problem does not disappear when dealing with multi-categorical data. As the number of categories increases, the information in the data for estimating σ^2 should increase, but it is not yet clear at what point there may exist adequate information to estimate this variance parameter. This deserves further investigation, but is beyond the scope of this dissertation. We suspect that data consisting of only three or four categories still lack adequate information to estimate the common spatial variance parameter.

Therefore, based on the reasons discussed in this section, we choose to fix both τ^2 and σ^2. We follow the common convention of setting $\tau^2 = 1$ to specify a probit-like model. As part of our simulation study, we investigate the effect of fixing σ^2 at a value different from the true data-generating value. We create artificial data with 4 categories using $\sigma^2_{true} = 1$ and $\sigma^2_{true} = 3$ and fit the models using $\sigma^2_{fix} = 1$ and $\sigma^2_{fix} = 2$. The results are reported in Section 5.2.2.4 and in general reveal that the fixed value has little effect on the predictions or the estimates.

3.2.2 Identifiability of the correlation parameters

De Oliveira (1997, 2000) describes near non-identifiability problems that can arise when estimating the smoothness (or roughness) parameter of the underlying correlation function using binary data. He points out that after clipping the underlying random field, the binary data contain no information about the parameter, even

if we could observe the complete binary realization rather than merely the values of the realization at a finite number of locations. Once again, this problem also exists for categorical data, though to a lesser degree. This is similar to the problem of estimating the spatial variance parameter described in the previous section, Section 3.2.1, in that we are trying to estimate a parameter that describes the underlying continuous distribution from categorical data. For this reason, we restrict the set of correlation functions of interest to those defined by only one scale parameter controlling the range of spatial correlation. We investigate use of the Matèrn correlation function, but fix the value of the smoothness parameter, κ, and estimate the remaining parameter, ϕ. For the simulation study and the real data application, we set $\kappa = 0.5$, a specification equivalent to the exponential correlation function. For his examples, De Oliveira (1997, 2000) used the correlation function, $H(\phi) = \phi^d$, where d is Euclidean distance.

De Oliveira (2000) notes that the Clipped Gaussian model for binary data will perform better in terms of prediction for smooth realizations of binary random fields than rough realizations. One realization can be pictured as a three-dimensional surface, such as a relief map. A smooth surface is characterized by gently rolling hills, whereas a rough surface is characterized by jagged peaks and valleys. In the context of clipping the random field to create a categorical random field, a rougher continuous random field generally results in a more patchy categorical random field. Thus, the terms smooth and less-patchy go together, as well as rough and patchy.

The Gaussian process is inherently smooth due to the unimodal property of the Gaussian distribution. At one extreme, we can envision a very smooth categorical random field characterized by K separate patches of color, one patch for each category. On the other extreme, we envision a very rough random field in which the K colors are interspersed over the entire region in many different patches. When used as a modeling tool, the Gaussian process may result in the prediction of smooth

binary or categorical random fields. Thus, it is possible that a method based on a latent Gaussian process may have better predictive ability if the data are best represented as coming from a smooth (less patchy) random field. In the simulation study of Chapter 5, we seek to cover a range of patchiness in the random fields to assess this idea that the methods may result in improved prediction for data from smoother fields. However, because the correlation in the categorical random field does not only depend on the smoothness parameter of the correlation function, but also the other parameters in the model, this is more difficult. This problem was investigated earlier in this chapter in Section 3.1. Examples of categorical random fields with different degrees of smoothness (or patchiness) generated from our models are shown in Figure 3.5. The upper plots represent smoother (less patchy) realizations than the bottom two plots.

3.2.3 Identifiability of the cut-point parameters and regression coefficients

As previously noted in Section 2.3, it is not possible to estimate all $K-1$ cut-points, $\theta_1, \theta_2, \ldots, \theta_{K-1}$, in addition to an intercept parameter in the vector of regression coefficients, $\boldsymbol{\beta}$. Therefore, it has become common convention to set $\theta_1 = 0$, as we follow this convention. This allows for estimation of the remaining $K-2$ cut-point parameters along with the intercept parameter, β_0. An alternative parameterization allows $K-1$ cut-point parameters to vary, excluding the intercept term. These parameterizations are equivalent models and should result in equivalent model fits and predictions. For example, if we denote the parameters for the no-intercept model with a star, we can write $\theta_1^* = -\beta_0$ and $\theta_k^* = \theta_k - \beta_0$ for $k = 2, \ldots, K-1$.

However, when the models are fit using the MCMC algorithm, there is a difference in how the parameters are estimated. The intercept parameter is estimated as a regression parameter, and thus included in the vector $\boldsymbol{\beta}$. It is updated with the Gibbs step for the regression parameters from a multivariate Gaussian distribution.

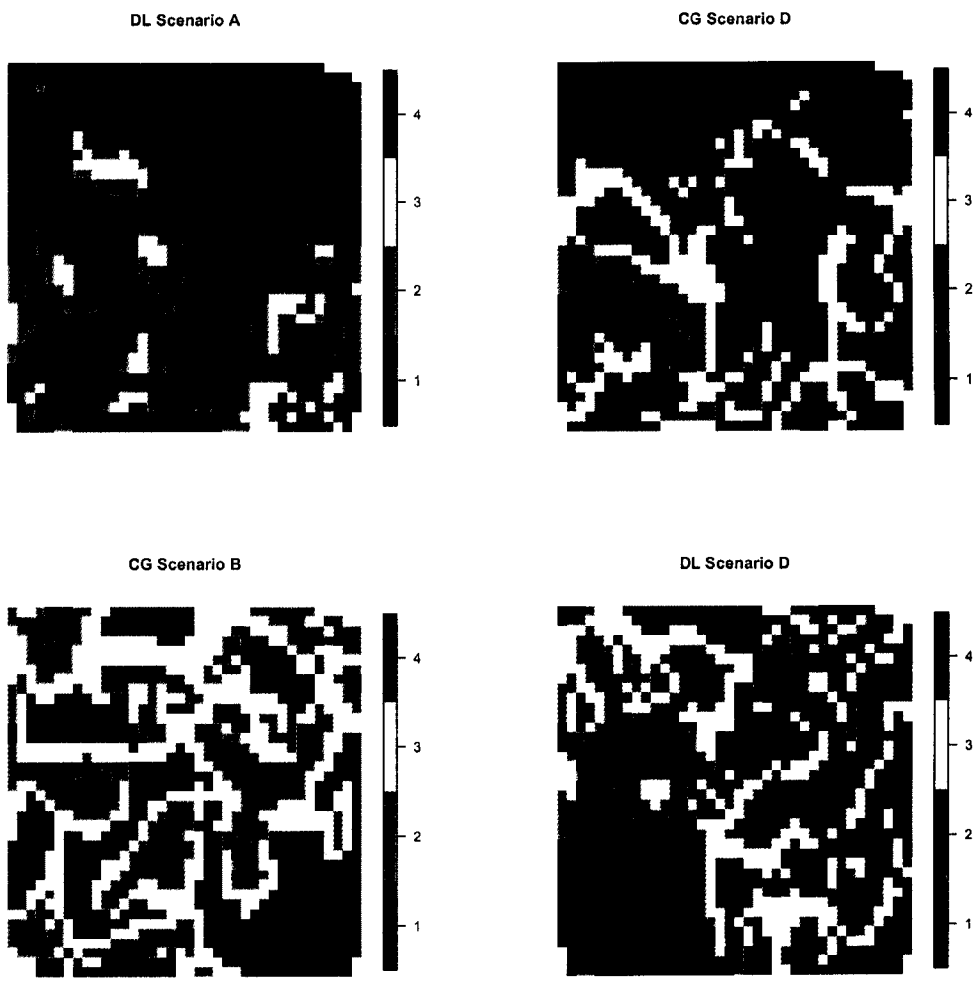

Figure 3.5: Examples of varying degrees of patchiness (or smoothness) in the categorical random fields created by the proposed models.

When the intercept is removed, its equivalent parameter, θ_1, is estimated using a Metropolis-Hastings (M-H) step to sample from the complete conditional of the vector of cut-point parameters. The distribution is not multivariate Gaussian. It seems reasonable that these two different parameterizations, while equivalent models, could result in differences in convergence. They should, however, not change the estimates or predictions appreciably, though we do not specifically investigate this.

3.3 Prior specification

Until this point, we have largely glossed over our choice and specification of prior distributions for the parameters of our models. Our general strategy for choosing and specifying these distributions is to find proper, yet relatively non-informative, distributions. In relation to the previous discussion of nonidentifiability and informativeness of the data, we seek to specify a prior distribution that spreads its density over a reasonable (practical) range of values while letting the data do most of the informing of the posterior distribution. In the terminology of Section 3.2, we desire priors resulting in an adequate degree of Bayesian learning, that is, a posterior distribution that differs substantially from the prior distribution due to the information contained in the data regarding the parameters of interest. We also want to ensure a proper posterior distribution. We steer away from using prior specification to solve problems of indentifiability, as is clear by our decision to fix τ^2 and σ^2, as described in Section 3.2.1. We must, however, still define prior distributions for the scale parameter, ϕ, of the spatial correlation function; the regression parameters, $\boldsymbol{\beta}$; and for the cut-point parameters, $\boldsymbol{\theta}$.

3.3.1 Prior for the spatial scale parameter

To specify the prior distribution for the scale parameter of the exponential correlation function used in the simulation study described in Chapter 5 and for the data application in Chapter 6, we follow the common convention of choosing a gamma

prior distribution. To specify the parameters of the gamma distribution, we follow a strategy described in Banerjee et al. (2004) and Schmidt et al. (2007). The strategy uses the concept of an effective (or practical) range previously described in Section 1.4.2, defined as the intersite distance at which the correlation decreases to 0.05. We denote this effective range as d_0, and for the exponential correlation function it is calculated approximately as $d_0 = 3/\phi$. Specification of the prior mean relies on the assumption that the effective range falls at approximately half the maximum inter-site distance (d_{max}). Note that an approximate value for d_{max} is available as soon as the study area is defined. This assumption results in a reasonable *a priori* guess at the value of the parameter, $\phi_{prior} = 6/d_{max}$, that can be used as the prior mean for the gamma distribution. It is then suggested that the variance of the gamma distribution be fixed at some large value. Since both the mean and variance of the gamma distribution are functions of the same two parameters, these parameters can be solved for after the mean and variance are both specified.

In our investigation of this prior distribution, we found that setting the variance to "some large value" is much too general of statement as small differences in the value can lead to very different shapes of the distribution and it is definitely *not* the case that a larger variance generally leads to a more appropriate or realistic prior distribution. This was also noticed by Schmidt et al. (2007). In fact, we found relatively large variances that led to very obviously inappropriate prior distributions when visually examined. Thus, to choose the value for the variance, we plotted the resulting gamma distribution for a variety of variances, ultimately choosing one value that places most of the probability density over feasible and practical values of ϕ. This range of feasible values is largely determined based on our knowledge of the inter-site distances and the assumption that some degree of spatial correlation does exist. In Section 6.2.1 we provide a description of how this strategy is used to specify the parameters of the gamma distribution for our application to real data.

Another method of specifying the prior distribution for ϕ in the context of a continuous response is suggested by Berger et al. (2001). Berger et al. (2001) investigate "objective" prior distributions for the unknown mean and covariance parameters in spatial models using Gaussian random fields. Their *reference* prior strategy relies on use of the shape of the likelihood function and Fisher information to specify the distribution. The reference prior method has been shown to perform well in simulations and would thus be a logical alternative to explore (Schmidt et al., 2007).

3.3.2 Prior for the regression parameters

Recall from Chapter 2 that we specify a conjugate Gaussian prior distribution for the regression parameters. In the case of a single covariate, $p = 2$, the distribution is specified as

$$\boldsymbol{\beta} \sim MVN_2(\mathbf{b}_0, \mathbf{B}_0). \tag{3.19}$$

Thus, our task is to specify a vector for $\mathbf{b}_0$ and a matrix for $\mathbf{B}_0$. We investigated the effect of many different specifications on the convergence of the parameters which eventually led to the common specification of $\mathbf{b}_0 = \mathbf{0}$ and $\mathbf{B}_0 = \lambda^2\mathbf{I}$ where λ^2 is set at a relatively large value to specify a vague prior. We noticed little difference in the results or convergence of the algorithm for values of $\lambda^2 > 20$ and thus chose to use 20 as the value. Interestingly, we came to this conclusion independent of the findings of De Oliveira (1997) that suggest use of this same value for λ^2 for the binary data problem. The results were very insensitive to the values of $\mathbf{b}_0$ when they were specified within the range of -10 to 10, and thus we chose the convenient zero vector.

3.3.3 Prior for the cut-point parameters

In Chapter 2, we describe the de-constraint transformation applied to the cut-point parameters. Recall that the transformation is

$$\begin{aligned} \alpha_2 &= \log(\theta_2) \\ \alpha_k &= \log(\theta_k - \theta_{k-1}) \text{ for } k = 3, \ldots, K, \end{aligned} \tag{3.20}$$

with an inverse transformation of

$$\theta_k = \sum_{i=1}^{k} \exp(\alpha_i). \tag{3.21}$$

We denote the inverse transformation as $g(\boldsymbol{\alpha})$ in later calculations. Following Albert and Chib (1993), and as given in Chapter 2, one of the advantages of this de-constraint reparameterization is that it allows use of a multivariate Gaussian prior distribution,

$$\boldsymbol{\alpha} \sim MVN_{K-2}(\mathbf{a}_0, \mathbf{A}_0). \tag{3.22}$$

To specify the values of the elements in $\mathbf{a}_0$ and $\mathbf{A}_0$, we investigate how this distribution translates into the prior on the parameter space of the original, untransformed cut-point parameters. This does not appear to have been previously investigated by others. Our simulation study and data application both involve data with 4-categories and thus we use this case for our investigation.

The Jacobian matrix for the transformation in (3.20) for the 4-category case is

$$J = \begin{pmatrix} \frac{\partial}{\partial\theta_2}\log(\theta_2) & \frac{\partial}{\partial\theta_3}\log(\theta_2) \\ \frac{\partial}{\partial\theta_2}\log(\theta_3 - \theta_2) & \frac{\partial}{\partial\theta_3}\log(\theta_3 - \theta_2) \end{pmatrix} = \begin{pmatrix} \frac{1}{\theta_2} & 0 \\ \frac{-1}{\theta_3 - \theta_2} & \frac{1}{\theta_3 - \theta_2} \end{pmatrix}. \tag{3.23}$$

The absolute value of the determinant of the Jacobian is

$$\left| \frac{1}{\theta_2} \frac{1}{(\theta_3 - \theta_2)} \right| = \frac{1}{\theta_2(\theta_3 - \theta_2)} \tag{3.24}$$

since $0 < \theta_2 < \theta_3$. The probability density function for the prior in the original $\boldsymbol{\theta}$ space is then given by

$$\begin{aligned} f_{\boldsymbol{\theta}} &= \frac{1}{\theta_2(\theta_3-\theta_2)} f_{\boldsymbol{\alpha}}(g^{-1}(\boldsymbol{\theta})) \\ &= \frac{1}{\theta_2(\theta_3-\theta_2)} \frac{1}{2\pi} |A_0|^{-1/2} \cdot \\ &\quad \exp\left\{ \frac{-1}{2} \left((g^{-1}(\boldsymbol{\theta}) - \mathbf{a}_0)|A_0|^{-1}(g^{-1}(\boldsymbol{\theta}) - \mathbf{a}_0)' \right) \right\}, \end{aligned} \tag{3.25}$$

where $g^{-1}(\boldsymbol{\theta}) = (\log(\theta_2), \log(\theta_3 - \theta_2))'$ and $\mathbf{a}_0 = (a_{01}, a_{02})$.

To provide more insight into the effect of specification of the prior distribution for $\boldsymbol{\alpha}$ on the resulting prior distribution for the original cut-point parameters, $\boldsymbol{\theta}$, we conduct a graphical investigation by plotting the prior distribution for $\boldsymbol{\theta}$ for varying hyperparameters, $\mathbf{a}_0$ and $\mathbf{A}_0$. We actually specify the elements of $\mathbf{a}_0$ by first specifying the values $\boldsymbol{\theta}_{prior} = (\theta_{2,prior}, \theta_{3,prior})$ and then transforming $\boldsymbol{\theta}_{prior}$ to give $\mathbf{a}_0$. Several plots of the priors in $\boldsymbol{\theta}$ space are shown in Figure 3.6.

The graphical investigation reveals that a diagonal specification of $\mathbf{A}_0$ with variance 0.5 or 1 results in reasonable prior estimates for $\boldsymbol{\theta}$. For larger diagonal elements, the mass is moved toward the origin (see lower left plot of Figure 3.6). This collection of mass near the origin gives a large amount of prior density to small values of θ_2 and θ_3, a fact that leads us to conclude that large values for the diagonal of the variance-covariance matrix should not be used as they do not spread the density over a large enough range of reasonable values. In other words, it may be too informative toward small values of the cut-points, possibly leading to underestimation of the parameters. Recall that the cut-point parameters are constrained to be greater than 0 because $\theta_1 = 0$. Smaller values of the diagonal elements of $\mathbf{A}_0$ lead to accumulation of the mass further from the origin, with values between 0.5 and 1 giving the most reasonable shapes for a non-informative prior. Specifying non-zero values for the off-diagonal elements of $\mathbf{A}_0$ resulted in very little difference in the posterior estimates, but did slow convergence of the algorithm for some values. We conclude that values

of zero for the covariances lead to as or more reasonable results than other values, and thus after much investigation we set $\mathbf{A}_0 = \mathbf{I}$ for both the simulation study in Chapter 5 and the data application in Chapter 6.

As previously mentioned, we choose to specify the prior mean, $\mathbf{a}_0$, through specification of $\boldsymbol{\theta}_{prior}$ to provide a more straightforward way of choosing the values. As we expect, the greater the distance between the two prior cut-points, $\theta_{2,prior}$ and $\theta_{3,prior}$, the more the prior distribution for cut-points is spread out away from the $\theta_1 = \theta_2$ line, as seen in the upper right plot of Figure 3.6. We conclude that a relatively non-informative prior can be obtained by setting the prior cut-points, $\theta_{2,prior}$ and $\theta_{3,prior}$, farther apart than we expect the "true" cut-points to be. It may be difficult to assess this for a new problem, but based on the expected number of observations per category and our assumptions regarding the distribution and variance of the continuous latent variable, it is possible to hazard a good guess.

There is another technique that can be useful in specifying the mean of the prior distribution for the cut-points, along with the intercept regression parameter. The idea is suggested by De Oliveira (1997, 2000) for the regression parameters in the binary data problem, and can be extended for use with the cut-points in a multi-category problem. The strategy relies on prior guesses of the probabilities of sites belonging to the K categories. This is particularly useful when some prior information regarding these probabilities can be elicited. Again, we specify $\mathbf{a}_0$ by first choosing reasonable values for the cut-point parameters and then transforming those values to the $\boldsymbol{\alpha}$ space.

We first describe the strategy for the constant mean case. Denote $Pr(Y_i = k)$ as p_k, and let p_k^* be the *a priori* guess at p_k based on expert or scientific information about the problem. Then, we can use the specification of the model in terms of cumulative probabilities to easily solve for the intercept parameter and the cut-points. In this example, we assume use of the standard Gaussian cdf in the model definition,

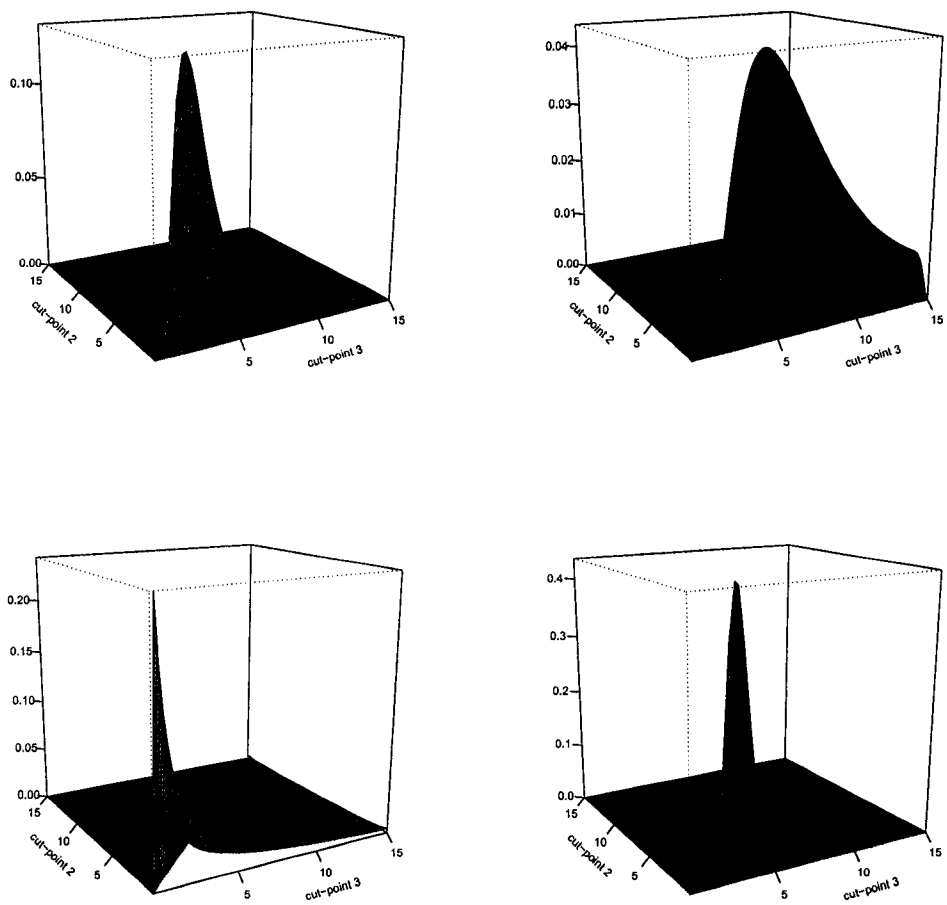

Figure 3.6: Examples of the prior distribution of the original cut-points parameters, $\boldsymbol{\theta}$, induced by the prior specified for the transformed cut-points, $\boldsymbol{\alpha}$, using the hyper-parameters $\mathbf{a}_0$ and $\mathbf{A}_0$. Plots are shown for $\boldsymbol{\theta}_{prior} = (2, 4)$, $\mathbf{A}_0^{diag} = 0.5$ (upper left), $\boldsymbol{\theta}_{prior} = (2, 8)$, $\mathbf{A}_0^{diag} = 0.5$ (upper right), $\boldsymbol{\theta}_{prior} = (2, 4)$, $\mathbf{A}_0^{diag} = 3$ (lower left), and $\boldsymbol{\theta}_{prior} = (2, 4)$, $\mathbf{A}_0^{diag} = 0.1$ (lower right).

but the results can be easily modified for different link functions or relationships between the linear predictor term and the cumulative probabilities. Also recall that $\theta_1 = 0$. We begin by setting

$$Pr(Y \leq k) = \sum_{j=1}^{k} p_k^* \tag{3.26}$$

$$= \Phi(\theta_k - \beta_0), \tag{3.27}$$

giving us $K-1$ equations and $K-1$ unknowns. Solving the equations gives

$$\beta_0 = -\Phi^{-1}(p_1^*) \tag{3.28}$$

$$\theta_2 = \Phi^{-1}(p_1^* + p_2^*) - \Phi^{-1}(p_1^*) \tag{3.29}$$

$$\theta_3 = \Phi^{-1}(p_1^* + p_2^* + p_3^*) - \Phi^{-1}(p_1^*), \tag{3.30}$$

from which the values can be used to specify the means of the prior distribution.

Incorporation of one or more covariates adds at least one additional regression parameter, thus making it impossible to solve the set of equations. However, it may be useful to simply use the values obtained from the constant mean case, assuming that the value of $\mathbf{x}_i\boldsymbol{\beta}$ is a relatively small value. This may not be a bad assumption when using a standardized covariate that varies around zero with unit variance. And, the goal is not to estimate the parameters, only to provide a prior distribution in the correct ballpark that will then be informed by the data. We seek to provide a strategy for specifying the prior that is not dependent on the data, and ignoring the covariate values obviously helps accomplish this goal. We suggest using the constant mean specification to investigate the values obtained for the cut-points, $\boldsymbol{\theta}$, but setting $b_0 = \mathbf{0}$. A graphical analysis regarding the cut-point priors, as described earlier in this section, is also suggested and may possibly lead to the use of values that are farther apart than those obtained with the strategy described in this paragraph.

As an example, suppose that for a four category case we *a priori* $p_1^* = 0.20, p_2^* = 0.60, p_3^* = 0.15$, and $p_4^* = 0.05$ represent *a priori* guesses at the probabilities in a

constant mean model with $\theta_1 = 0$. Then, we set $b_0 = -\Phi^{-1}(p_1^*) = -\Phi^{-1}(0.20) = 0.84$, and subsequently use $\theta_2 = \Phi^{-1}(0.8) + 0.84 = 1.68$, and $\theta_3 = \Phi^{-1}(0.95) + 0.84 = 2.49$. The magnitudes of these values lead to an important point. The preceding calculations are based on the standard Gaussian distribution, meaning that if the values of the probabilities are not extremely small (near 0) or extremely large (near 1), the values for $\mathbf{b}_0$ and $\boldsymbol{\theta}$ will fall nicely between 0 and 3. Thus, we can set a relatively non-informative prior covering this reasonable range of values.

3.4 Assessing model fit and MCMC convergence

In this section, we very briefly cover a possible strategy for assessing model fit, at least for the Double Latent model, and also methods used to evaluate convergence of the MCMC algorithm.

3.4.1 Assessing fit via latent residuals

Albert and Chib (1993, 1995, 1997) suggest the use of latent residuals to assess fit of their non-spatial latent variable models to ordered categorical data. The model of interest is the usual

$$Z_i = \theta_k - \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i. \tag{3.31}$$

Albert and Chib define the latent residual as

$$r_i = Z_i - \mathbf{x}_i\boldsymbol{\beta}, \tag{3.32}$$

where Z_i is the value of the latent variable associated with location i. The latent residual has a standard normal distribution, leading to a truncated normal posterior distribution with density given by

$$f(r_i|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}) \propto \phi(r_i) I_{\{\theta_{(y_i-1)} - \mathbf{x}_i\boldsymbol{\beta} \,<\, r_i \,\leq\, \theta_{y_i} - \mathbf{x}_i\boldsymbol{\beta}\}}, \tag{3.33}$$

where $\phi(\cdot)$ is the pdf of the standard normal distribution and y_i is the observed response value. The posterior density in (3.33) can be approximated by plugging

in the posterior means for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. Conflict between the fitted value of $\mathbf{x}_i\boldsymbol{\beta}$ and the appropriate interval defined by the cut-points parameters, $(\theta_{y_i-1}, \theta_{y_i})$, causes the posterior distribution of the latent residuals to differ from the standard normal distribution.

Albert and Chib (1995) use the latent residuals to detect possible outlying observations that are not fit well by the model. For example, they use the tail probability associated with the standard normal distribution for some constant, g, $Pr(|r_i| > g|y)$, to construct a probability plot. This involves plotting the posterior means of the ordered values of the n latent residuals against the expected order statistics from a sample of size n from the standard normal distribution. Residuals that do not fall on or near the line may be flagged as extreme and may indicate poor fit of the model to the associated observations.

An extension of this residual approach can be developed for the Double Latent model by defining a conditional latent residual as

$$r_i|W_i = Z_i - \mathbf{x}_i\boldsymbol{\beta} - W_i \tag{3.34}$$

where the $r_i|W_i$ are independent standard normal random variables, leading again to a truncated normal posterior density,

$$f(r_i|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{w}) \propto \phi(r_i) I_{\{\theta_{(y_i-1)} - \mathbf{x}_i\boldsymbol{\beta} - w_i \,<\, r_i \,\leq\, \theta_{y_i} - \mathbf{x}_i\boldsymbol{\beta} - w_i\}} \tag{3.35}$$

We can approximate the posterior distribution in (3.35) by plugging in the posterior means for $\boldsymbol{\beta}$, w_i, and $\boldsymbol{\theta}$. Normal probability plots can again be constructed to detect possible outlying observations.

The Clipped Gaussian model does not lend itself to an easy extension of this idea due to the absence of the convenient conditional independence that we find in the Double Latent model, but the idea is worth pursuing in future work.

3.4.2 Posterior predictive checks

Another way to investigate whether the model is consistent with the data is implementation of posterior predictive checks (Gelman et al., 2004). The method compares simulated values from the posterior predictive distribution of replicated data to the observed data, a check of whether the data are consistent with the posterior predictive density. Replicated data values differ from predicted data values for spatial data in that the replicates are values for locations identical to those of the observed data, and predictions are values for new locations. Replicate data are considered data we might see tomorrow under the same model and parameter values.

Gelman et al. (2004) describe the use of "Bayesian p-values" to measure the statistical significance of lack of fit using posterior simulations of (η, y^{rep}), where η is the vector of all unknown parameters and y^{rep} is a replicate response value from the posterior predictive distribution. Tail-area probabilities of a test quantity of interest can be used to obtain a Bayesian p-value, defined as the probability that the replicated data are more extreme than the observed data, given by

$$Pr(T(\mathbf{y}^{rep}, \eta) \geq T(\mathbf{y}, \eta)), \tag{3.36}$$

where $T(\cdot)$ is a function defining a quantity of interest. Practically, the first step in implementing a posterior predictive check is defining the test quantity, $T(\mathbf{Y}, \eta)$, to capture aspects of the data and model that we wish to check. Replicated data are obtained by drawing one y^{rep} from the posterior predictive distribution for each draw of the other unknown parameters, $\eta^{(t)}$. The test quantity is calculated at each iteration for both the observed data and the replicated data, and then the values are compared. The estimated p-value is simply the proportion of the T draws for which the test quantity for the replicated data is as or more extreme than that of the observed data (Gelman et al., 2004).

A useful test quantity for categorical data is the percentage of observations falling into each category,

$$T_k(\mathbf{y}, \boldsymbol{\eta}) = \frac{1}{n} \sum_{i=1}^{n} I_{\{\mathbf{y}=k\}}. \tag{3.37}$$

We devote Chapter 4 to describing the methods used for prediction and the measures used to assess predictive ability of the models using a hold-out data set. Thus, we do not discuss prediction in this section.

3.4.3 Monitoring and assessing convergence

To assess and compare convergence of the different MCMC algorithms, we suggest using the methods described in Cowles and Carlin (1996) and Gelman et al. (2004). The key aspect of such methods is the use of multiple runs of the chain from dispersed starting values. For a cursory investigation, it is very informative to examine both the sample path plots of the values of the chain for each parameter and the lag 1 autocorrelation coefficients for each chain. For those parameters updated via the Metropolis-Hastings algorithm, we also place emphasis on examining the proportion of accepted candidate values, using the range of 0.4 to 0.5 as a target range for these proportions (Givens and Hoeting, 2006).

A more quantitative diagnostic is the potential scale reduction (PSR) (Gelman et al., 2004). The PSR is an estimate of the factor by which the scale of the current distribution of a parameter might be reduced if the chain were run indefinitely long. PSR values close to 1 indicate that each run of the chain produced values close to the target distribution, and thus leads to the conclusion that convergence has been obtained. See Cowles and Carlin (1996) for a thorough review and investigation of other MCMC convergence diagnostics.

3.5 Conclusions

This chapter presented an investigation of the correlation structure in the categorical random fields created under the assumptions of the proposed models. Additionally, it addresses the practical subjects of identifiability and prior specification for Bayesian inference. We also discussed assessment of model fit and MCMC convergence. In the following chapter, we turn to the details of formulating the prediction of the response at new locations as a decision problem, along with definition of the measures we choose to assess predictive ability. As previously stated, the primary goal of applying these models is prediction at new locations. Thus, we devote Chapter 4 to the development and description of the details of spatial prediction using the Double Latent and Clipped Gaussian models.

Chapter 4

PREDICTION AT NEW LOCATIONS

4.1 Prediction as a decision problem

A primary goal of our models, and of spatial models in general, is prediction at new locations. Thus, we evaluate the ability of our models to predict the categorical response at new, unobserved locations, focusing on the observable quantities of interest rather than on the estimation of unobservable parameters. We denote the vector of categorical response values at m new locations as $\mathbf{Y}_0 = (Y(\mathbf{s}_{01}), Y(\mathbf{s}_{02}), \ldots, Y(\mathbf{s}_{0m})) = (Y_{01}, Y_{02}, \ldots, Y_{0m})$. We similarly denote the latent variables for these new locations as $\mathbf{W}_0$ and $\mathbf{Z}_0$. Predictions at the new locations are made using the Bayesian posterior predictive distribution, $f(\mathbf{y}_0|\mathbf{y})$, and are denoted by $\tilde{\mathbf{Y}}_0$. Our goal is to minimize the error in the categorical predictions over the region of interest. This prediction problem can be placed in the context of a decision problem (De Oliveira, 1997, 2000; Berger, 1985), where the definition of a loss function can change to suit the problem of interest. Prediction is based on the Bayesian expected loss, defined as the expected value of the loss function given the data. The optimal Bayes predictor is found by minimizing the Bayesian expected loss.

4.1.1 Loss functions and Bayesian expected loss

We restrict our attention to a multi-category extension of an additive binary loss function employed by De Oliveira (1997, 2000). For binary data, De Oliveira (1997,

2000) assign zero loss to correct predictions. He measures the loss of predicting $\mathbf{Y}_0$ with $\tilde{\mathbf{Y}}_0$ by

$$\begin{aligned} L(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) &= \sum_{j=1}^{m} L(Y_{0j}, \tilde{Y}_{0j}) \\ &= \frac{1}{m} \sum_{j=1}^{m} \left[d_0 I_{\{\tilde{Y}_{0j}=1, Y_{0j}=0\}} + d_1 I_{\{\tilde{Y}_{0j}=0, Y_{0j}=1\}} \right], \end{aligned} \tag{4.1}$$

where $I_{\{A\}}$ is an indicator of the occurrence of event A, and d_0 and d_1 are fixed values quantifying the amount of loss associated with predicting a 1 when the true value is 0, and a 0 when the true value is 1, respectively. This definition allows the magnitude of the loss to change for different errors.

We extend this loss function for the multi-category case in a straightforward manner. The additive loss function is

$$\begin{aligned} L(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) &= \sum_{j=1}^{m} L(Y_{0j}, \tilde{Y}_{0j}) \\ &= \frac{1}{m} \sum_{j=1}^{m} \left[\sum_{k=1}^{K} \sum_{l \neq k} d_{kl} I_{\{Y_{0j}=k, \tilde{Y}_{0j}=l\}} \right], \end{aligned} \tag{4.2}$$

where $d_{kl} \geq 0$ for all k and l. The d_{kl} are termed *loss coefficients* and, as for the binary case, they serve to quantify the amount of loss incurred when predicting a category l response when the true category is k.

To find the optimal Bayes predictor, we find the predictor the minimizes the Bayesian expected loss. The Bayesian expected loss is defined as the conditional expectation of the loss function given the observed data, $E[L(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0)|\mathbf{y}]$. For the loss function defined in (4.2), the Bayesian expected loss (BEL) is given by

$$\begin{aligned} E[L(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0)|\mathbf{y}] &= E\left[\frac{1}{m} \sum_{j=1}^{m} \left[\sum_{k=1}^{K} \sum_{l \neq k} d_{kl} I_{\{Y_{0j}=k, \tilde{Y}_{0j}=l\}} \right] |\mathbf{y} \right] \\ &= \frac{1}{m} \sum_{j=1}^{m} \left[\sum_{k=1}^{K} \sum_{l \neq k} d_{kl} I_{\{\tilde{Y}_{0j}=l\}} E\left[I_{\{Y_{0j}=k\}} |\mathbf{y} \right] \right] \\ &= \frac{1}{m} \sum_{j=1}^{m} \left[\sum_{k=1}^{K} \sum_{l \neq k} d_{kl} I_{\{\tilde{Y}_{0j}=l\}} Pr\left[Y_{0j} = k|\mathbf{y} \right] \right]. \end{aligned} \tag{4.3}$$

Table 4.1: Bayesian expected loss for a 4-category example. $Pr(Y_{0j} = k|\mathbf{y})$ is denoted as $Pr_j(k)$.

$\tilde{Y}_0$	Bayesian Expected Loss
1	$BEL_1 = d_{21}Pr_j(2) + d_{31}Pr_j(3) + d_{41}Pr_j(4)$
2	$BEL_2 = d_{12}Pr_j(1) + d_{32}Pr_j(3) + d_{42}Pr_j(4)$
3	$BEL_3 = d_{13}Pr_j(1) + d_{23}Pr_j(2) + d_{43}Pr_j(4)$
4	$BEL_4 = d_{14}Pr_j(1) + d_{24}Pr_j(2) + d_{34}Pr_j(3)$

Equation (4.3) can be further simplified to

$$E[L(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0)|\mathbf{y}] = \frac{1}{m}\sum_{j=1}^{m}\left[\sum_{k=1}^{K} Pr[Y_{0j} = k|\mathbf{y}]\left\{\sum_{l\neq k} d_{kl}I_{\{\tilde{Y}_{0j}=l\}}\right\}\right]. \quad (4.4)$$

In (4.4), each location, indexed by j, contributes a non-negative term to the overall sum. Thus, to find the predictor that minimizes the entire expression, the expression can be minimized for each j. Table 4.1 displays the possible expressions for the BEL depending on the predicted value, for one location and a 4-category example. Note that the BEL relies on the predicted value and *not* on the true value and is thus calculable for any new location without the use of a hold-out data set or cross validation. The resulting optimal Bayes predictor clearly depends on the specified values of the loss coefficients. If $d_{kl} = 1$ for all k and l, the loss function defined in (4.2) simplifies to the mis-prediction, or mis-classification, rate and is given by

$$L(\mathbf{Y}_0, \tilde{\mathbf{Y}}_0) = \frac{1}{m}\sum_{j=1}^{m} I_{\{Y_{0j}\neq\tilde{Y}_{0j}\}}. \quad (4.5)$$

Referring again to Table 4.1, minimizing the Bayesian expected loss for the 4-category case leads to a prediction rule specifying that $\tilde{Y}_{0j} = k$ if $\min(BEL_1, BEL_2, \ldots, BEL_K) = BEL_k$. Setting all $d_{jk} = 1$, the optimal Bayes predictor for a single new location reduces to $\tilde{Y}_{0j} = k$ if

$$\max\{Pr(Y_{0j} = 1|\mathbf{y}), Pr(Y_{0j} = 2|\mathbf{y}), \ldots, Pr(Y_{0j} = K|\mathbf{y})\} = Pr(Y_{0j} = k|\mathbf{y}). \quad (4.6)$$

Table 4.2: Bayesian expected loss for 4-categories and loss coefficients of $d_{kl} = |k-l|$. $Pr(Y_{0j} = k|\mathbf{y})$ is denoted as $Pr_j(k)$.

$\tilde{Y}_0$	Bayesian Expected Loss
1	$BEL_1 = 1Pr_j(2) + 2Pr_j(3) + 3Pr_j(4)$
2	$BEL_2 = 1Pr_j(1) + 1Pr_j(3) + 2Pr_j(4)$
3	$BEL_3 = 2Pr_j(1) + 1Pr_j(2) + 1Pr_j(4)$
4	$BEL_4 = 3Pr_j(1) + 2Pr_j(2) + 1Pr_j(3)$

This is more concisely stated as

$$\tilde{Y}_{0j} = \text{argmax}_{k=1,\ldots,K} \{Pr(Y_{0j} = k|\mathbf{y})\}. \tag{4.7}$$

Another natural loss function might take into account some measure of distance between the predicted value and true value, so that greater loss is incurred for predictions that are farther from the truth. For example, we could assign a loss function coefficient of $d_{kl} = |k - l|$ when the labels are integer values. This coefficient describes the number of categories that the prediction falls from the true value. Table 4.2 provides the Bayesian expected loss expressions for this loss function for one location and four categories. The prediction rule is still defined by predicting $\tilde{Y}_{0j} = k$ if $\min(BEL_1, BEL_2, \ldots, BEL_K) = BEL_k$. We do not apply this loss function in this dissertation, but consider this logical future work.

4.1.2 The Bayesian posterior predictive distribution

In general, to make predictions using any version of the loss function in (4.2) we must estimate the conditional probabilities, $Pr(Y_{0j} = k|\mathbf{y})$, for each $k = 1, 2, \ldots, K$. These probability estimates can be obtained as part of our MCMC algorithm, using the *posterior predictive distribution*, $f(\mathbf{y}_0|\mathbf{y})$. As first defined in (1.17) within Section

1.2, the posterior predictive distribution is given in general as

$$\begin{aligned} f(\mathrm{y}_0|\mathrm{y}) &= \int f(\mathrm{y}_0, \eta|\mathrm{y})d\eta \\ &= \int f(\mathrm{y}_0|\eta, \mathrm{y})f(\eta|\mathrm{y})d\eta \qquad (4.8) \\ &= \int f(\mathrm{y}_0|\eta)f(\eta|\mathrm{y})d\eta, \qquad (4.9) \end{aligned}$$

where η represents the vector of all unknown parameters. It is clear from (4.9) that the posterior predictive distribution is an average of predictions given the parameters over the posterior distribution of those parameters (Gelman et al., 2004). The equality between (4.8) and (4.9) holds by the conditional independence of $\mathbf{Y}$ and $\mathbf{Y}_0$ given η.

We first give the details regarding use of the posterior predictive distribution for the Clipped Gaussian model because they are more straightforward than for the Double Latent model due to the single latent variable. The details for the Double Latent model follow analogously, and are subsequently described.

For prediction at m new locations, we again capitalize on the use of the underlying latent variable and deterministic relationship between $\mathbf{Y}$ and $\mathbf{W}$. That is, even though we cannot directly sample from the posterior predictive distribution of $\mathbf{Y}_0$, we *can* sample from the posterior predictive distribution of $\mathbf{W}_0$ to get the conditional probabilities of interest using the equality

$$Pr(Y_{0j} = k|\mathrm{y}) = Pr(\theta_{k-1} < W_{0j} < \theta_k|\mathrm{y}). \qquad (4.10)$$

The $Pr(\theta_{k-1} < W_{0j} < \theta_k|\mathrm{y})$ is computed using $f(\mathrm{w}_0|\mathrm{y})$ and can be estimated using the MCMC algorithm.

To obtain predictions using Bayesian inference, we augment the set of unknown parameters with $\mathbf{W}_0$, the unknown latent variable predictions at the new locations. Thus, when interested in prediction, the desired joint posterior distribution is

$f(\mathbf{w}_0, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y})$, as opposed to $f(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y})$. From this augmented joint posterior distribution, we can obtain the distribution, $f(\mathbf{w}_0|\mathbf{y})$. This posterior predictive distribution is needed to estimate the conditional probabilities, which are then used to make the predictions using a rule defined by the loss function.

The posterior predictive distribution is not available in closed form, but can be sampled from using a MCMC algorithm. As noted in De Oliveira (1997), the algorithm for sampling from this posterior can be divided into two stages due to the factorization,

$$f(\mathbf{w}_0, \mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y}) = f(\mathbf{w}_0|\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y}) f(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y}). \tag{4.11}$$

Thus, we first use the MCMC algorithm described in Section 2.4.1 to obtain T draws (after a burn-in period) from the original posterior distribution, $f(\mathbf{w}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y})$. For each iteration, we obtain a draw from each of the unknown parameters, $(\mathbf{W}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)})$, and then use these values to draw $\mathbf{W}_0^{(t)}$ from $f(\mathbf{w}_0|\mathbf{w}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)})$. We give the results assuming that we have a covariate measured at the unobserved locations. The distribution of $(\mathbf{W}_0|\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ is conveniently multivariate Gaussian and is specified as

$$MVN_m(\mathbf{X}_0\boldsymbol{\beta} + G_\gamma \boldsymbol{\Sigma}_\gamma^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta}), K_\gamma - G_\gamma \boldsymbol{\Sigma}_\gamma^{-1} G_\gamma'). \tag{4.12}$$

Here, G_γ is the $m \times n$ correlation matrix such that $G_{\gamma,ij} = H_\gamma(\mathbf{s}_{0i}, \mathbf{s}_j)$, describing the correlations between the new and observed locations. The matrix K_γ is the $m \times m$ correlation matrix such that $K_{\gamma,ij} = H_\gamma(\mathbf{s}_{0i}, \mathbf{s}_{0j})$, describing the correlations between the new locations. The sequence of draws of $\mathbf{W}_0$ is constructed as a sample from $f(\mathbf{w}_0|\mathbf{y})$.

For prediction at m new locations using the Double Latent model, we again desire summaries of the posterior predictive distribution, $f(\mathbf{y}_0|\mathbf{y})$, and obtain this via the deterministic relationship between $\mathbf{Y}_0$ and the latent variable $\mathbf{Z}_0$. That is, we

use the posterior predictive distribution $f(\mathbf{z}_0|\mathbf{y})$ to obtain the distribution $f(\mathbf{y}_0|\mathbf{y})$, ultimately using the equality $Pr(Y_{0j} = k|\mathbf{y}) = Pr(\theta_{k-1} < Z_{0j} < \theta_k|\mathbf{y})$ to obtain the needed probability estimates. Analogous to that described for the Clipped Gaussian model, the set of unknown parameters is augmented to include the predicted values of the both latent variables, $\mathbf{Z}_0$ and $\mathbf{W}_0$, resulting in the joint posterior distribution $f(\mathbf{z}_0, \mathbf{w}_0, \mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y})$. Sampling from this joint posterior is divided into three stages following its factorization into

$$f(\mathbf{z}_0, \mathbf{w}_0|\mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y}) f(\mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y}) =$$
$$f(\mathbf{z}_0|\mathbf{w}_0, \mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y}) f(\mathbf{w}_0|\mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y}) f(\mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y}). \quad (4.13)$$

We use the MCMC algorithm described in Section 2.3.1 to obtain T draws (after a burn-in period) from the posterior distribution, $f(\mathbf{w}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{y})$. For each one of these draws, $(\mathbf{W}^{(t)}, \mathbf{Z}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)})$, we first draw $\mathbf{W}_0^{(t)}$ from $f(\mathbf{w}_0|\mathbf{w}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \mathbf{y})$ and then $\mathbf{Z}_0^{(t)}$ from $f(\mathbf{z}_0|\mathbf{w}_0^{(t)}, \mathbf{w}^{(t)}, \mathbf{z}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \mathbf{y})$. For the second stage in the factorization, the distribution of $(\mathbf{W}_0|\mathbf{W}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y})$ is multivariate Gaussian and differs from that described for the Clipped Gaussian model only in that it involves a mean zero Gaussian process. The specified distribution is

$$MVN_m(G_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}(\mathbf{w}), K_{\boldsymbol{\gamma}} - G_{\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}G'_{\boldsymbol{\gamma}}). \quad (4.14)$$

As for the Clipped Gaussian specification, $G_{\boldsymbol{\gamma}}$ is the $m \times n$ correlation matrix such that $G_{\boldsymbol{\gamma},ij} = H_{\boldsymbol{\gamma}}(\mathbf{s}_{0i}, \mathbf{s}_j)$ and $K_{\boldsymbol{\gamma}}$ is the $m \times m$ correlation matrix such that $K_{\boldsymbol{\gamma},ij} = H_{\boldsymbol{\gamma}}(\mathbf{s}_{0i}, \mathbf{s}_{0j})$.

For the last stage in the factorization, the distribution of $(\mathbf{Z}_0|\mathbf{W}_0, \mathbf{W}, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\theta}, \gamma, \mathbf{y})$ is multivariate Gaussian with mean $(\mathbf{X}_0\boldsymbol{\beta} + \mathbf{W}_0)$, and the identity matrix, $\mathbf{I}$, as its variance-covariance matrix. That is, for one iteration of the chain and one location, $\mathbf{s}_j$,

$$Z_{0j}^{(t)}|W_{0j}^{(t)}, \boldsymbol{\beta}^{(t)}, \mathbf{y} \sim N(\mathbf{x}_{0j}\boldsymbol{\beta}^{(t)} + W_{0j}^{(t)}, 1). \quad (4.15)$$

The sequence of draws of $\mathbf{Z}_0$ is constructed as a sample from the posterior predictive distribution, $f(\mathbf{z}_0|\mathbf{y})$. Calculation of estimates of the probabilities needed to make the predictions is discussed in detail in Section 4.3, after we provide an introduction and brief discussion regarding the quantities we use to evaluate predictive ability.

4.2 Evaluation of predictive ability

4.2.1 Bayesian expected loss as a measure of predictive uncertainty

An important aspect of prediction is obtaining measures of prediction uncertainty. Estimated Bayesian expected loss ($\widehat{BEL}$) provides a useful measure toward this end, as described by De Oliveira (1997, 2000) for binary data. In this context, the term uncertainty refers to the *decision* to assign a certain value as the prediction, and not to the typical measure of variability in the estimates used to make the predictions. De Oliveira (2000) refers to the Bayesian expected loss of the optimal Bayes predictor as a global measure of uncertainty about an estimated map based on the prediction locations. For the multi-category case, the estimated BEL is obtained by plugging the estimated probabilities into the expression for BEL given in (4.3). That is, for one location,

$$\widehat{BEL} = \sum_{k=1}^{K} \left[\widehat{Pr}(Y_{0j} = k|\mathbf{y}) \sum_{l \neq k} d_{kl} I_{\{\tilde{Y}_{0j}=l\}} \right]. \tag{4.16}$$

To understand how this gives an overall measure of the quality of the predictions, again consider the case where $d_{kl} = 1$ for all k and l. This simplifies the above expression in (4.16) to

$$\begin{aligned} \widehat{BEL} &= \sum_{k=1}^{K} \left[\widehat{Pr}(Y_{0j} = k|\mathbf{y}) I_{\{\tilde{Y}_{0j} \neq k\}} \right] \\ &= \sum_{k \neq \tilde{Y}_{0j}} \widehat{Pr}(Y_{0j} = k|\mathbf{y}). \end{aligned} \tag{4.17}$$

Thus, for one new location, after predicting $\tilde{Y}_{0j} = k^*$, the estimated BEL is the sum of the remaining $K - 1$ conditional probabilities, after excluding that for k^*. This can alternatively be calculated as $[1 - \widehat{Pr}(Y_{0j} = k^*|\mathbf{y})]$, where $\widehat{Pr}(Y_{0j} = k^*|\mathbf{y})$ is the maximum of the K estimated probabilities. Therefore, on a local scale, a smaller maximum estimated probability leads to higher prediction uncertainty for that location. In other words, the decision to assign a certain value as the prediction is considered more uncertain as the total probability is more evenly spread out among the possible groups. De Oliveira (1997, 2000) obtains a global measure of prediction uncertainty by averaging over the local estimated BEL over all new locations. De Oliveira (2000) proposes comparing the predictive quality of two maps through calculation of the mean absolute difference between the corresponding maps of $\widehat{BEL}$.

For more complex loss functions, the Bayesian expected loss takes into account both the remaining probability in the unpredicted categories and the penalty defined by the loss coefficients. The coefficients may increase as distance from the predicted value increases, meaning that estimated probabilities for categories farther from that with the maximum estimated probability are given more weight in the calculation of $\widehat{BEL}$. The magnitude of $\widehat{BEL}$ cannot be directly compared across different loss functions. For the second loss function described in Section 4.1.1 with $d_{kl} = |k - l|$, the expression for the Bayesian expected loss for one location is

$$BEL = \sum_{k=1}^{K} Pr(Y_{0j} = k|\mathbf{y}) \left[\sum_{l \neq k} |k - l| I_{\{\tilde{Y}_0 = l\}} \right], \tag{4.18}$$

which can be further simplified to

$$BEL = \sum_{k=1}^{K} |\tilde{Y}_{0j} - k| Pr(Y_{0j} = k|\mathbf{y}). \tag{4.19}$$

Equation (4.19) clearly shows that the loss coefficients are the absolute values of the differences between the predicted value, $\tilde{Y}_{0j}$, and the labels of the K possible categories. Once again, the estimated BEL is obtained by plugging in the estimates of the probabilities into (4.19).

4.2.2 Measures of mis-prediction

While the estimated Bayesian expected loss is a useful measure and is easily available for any analysis, it is important that the $\widehat{BEL}$ be used in conjunction with measures describing the accuracy of predictions, such as mis-prediction rate. For example, consider the case where the maximum estimated probability is large relative to the other probabilities, but the associated category is *not* the true category. This scenario results in a deceptively small $\widehat{BEL}$ that may not be detected without other measures of predictive ability that rely on true values, such as use of a hold-out data set or cross validation. The scenario is obviously bad behavior for a model; not only is it making wrong predictions, but it is doing so with false confidence. We alternatively hope that wrong predictions occur along side high local $\widehat{BEL}$'s, meaning that the probability is spread out over multiple categories, reflecting the uncertainty in the decision being made.

The use of a hold-out data set allows direct assessment of model's predictive ability. We use the standard mis-prediction rate (MPR), based on an indicator of whether the predicted value is equal to the true value,

$$MPR = \frac{1}{m}\sum_{j=1}^{m} I_{\{Y_{0j}=\tilde{Y}_{0j}\}}. \tag{4.20}$$

Recall that this is the loss under the our chosen loss function when all $d_{kl} = 1$. The MPR measures the total number of wrong predictions, but does not provide information regarding how far the mis-predictions are from the truth. Thus, we also employ the measure of *absolute mis-prediction rate (AMPR)* to take this "distance" into account. The AMPR is defined as

$$AMPR = \frac{1}{m}\sum_{j=1}^{m} |Y_{0j} - \tilde{Y}_{0j}|, \tag{4.21}$$

and typically deserves more emphasis in terms of assessing predictive ability. In the simulation study of Chapter 5 and the application in Chapter 6, we utilize information from all three measures, $\widehat{BEL}$, MPR, and AMPR, to make conclusions regarding

a model's predictive ability. We also investigate the ability of the these measures in selecting the data generating model.

4.3 Calculation of conditional probability estimates

To make predictions, we must estimate the conditional probability that a new location falls in category k given the observed data, $Pr(Y_{0j} = k|\mathbf{y})$, for each $k = 1, \ldots, K$. The rule defined by minimizing the Bayesian expected loss is then applied to assign a categorical value, $\tilde{Y}_{0j}$, to a particular location, $\mathbf{s}_j$. As described in Section 4.1.2, the deterministic relationship between the categorical $\mathbf{Y}$ and the latent continuous variable $\mathbf{W}$ (or $\mathbf{Z}$) allows us to summarize $f(\mathbf{y}_0|\mathbf{y})$ using $f(\mathbf{w}_0|\mathbf{y})$ (or $f(\mathbf{z}_0|\mathbf{y})$). That is, we get $Pr(Y_{0j} = k|\mathbf{y})$ by estimating $Pr(\theta_{k-1} < W_{0j} \leq \theta_k|\mathbf{y})$ for the Clipped Gaussian model or $Pr(\theta_{k-1} < Z_{0j} \leq \theta_k|\mathbf{y})$ for the Double Latent model. We propose two methods for estimating these probabilities: the *Indicator* method and the *Link Function* method. The former follows the method described by De Oliveira (2000) and is based on Monte Carlo integration of the posterior predictive distribution, and the latter allows summary of the posterior distribution of the probability estimates. For each method, we first give details for the Clipped Gaussian model and then extend those results for the Double Latent model.

4.3.1 Indicator method

For the Clipped Gaussian model, the MCMC algorithm results in a sequence of draws, $(\mathbf{W}_{0j}^{(1)}, \ldots, \mathbf{W}_{0j}^{(T)})$, from the posterior predictive distribution, $f(\mathbf{w}_0|\mathbf{y})$, as described in Section 4.1.2. Thus, we can obtain an approximation of the probability $Pr(\theta_{k-1} < W_{0j} \leq \theta_k|\mathbf{y})$ by Monte Carlo integration over T iterations of the chain (Givens and Hoeting, 2006). For the binary case with a single cut-point fixed at zero, this is accomplished by simply calculating the proportion of iterations such that the predicted latent variable is greater than zero (De Oliveira, 2000). Analogously, our vector of cut-points is used to calculate the proportion of predicted latent

variable values in each interval defined by the cut-point parameters via Monte Carlo integration for multiple categories. The Monte Carlo approximation is given by

$$Pr(Y_{0j} = k|\mathbf{y}) \approx \frac{1}{T}\sum_{t=1}^{T} I_{\{\theta_{k-1}^{(t)} \leq w_{0j}^{(t)} < \theta_k^{(t)}\}}, \tag{4.22}$$

where T is the total number of iterations considered after a burn-in period. Probability estimates are obtained directly from this approximation. That is, the probability for one k at one location is estimated by calculating the proportion of posterior samples of W_{0j} falling within the appropriate cut-point interval defined by k. The cut-point intervals are redefined at each iteration by $(\theta_{k-1}^{(t)}, \theta_k^{(t)})$. Similarly, for the Double Latent model we obtain point estimates of the probabilities as

$$\widehat{Pr}(Y_{0j} = k|\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} I_{\{\theta_{k-1}^{(t)} \leq z_{0j}^{(t)} < \theta_k^{(t)}\}}. \tag{4.23}$$

The Indicator method results in one k-dimensional vector of probabilities for each location over one run of the MCMC algorithm. Thus, we obtain a point-estimate of the probability with no direct method to calculate variability in the estimate using our MCMC output. That is, we are not able to summarize an approximate posterior distribution of the probabilities. We consider this a significant shortcoming of the method, especially given the very desirable property of Bayesian inference: the ability to easily quantify uncertainty for estimates of any functional of the parameters.

As previously discussed, De Oliveira (1997, 2000) deals with this disadvantage by focusing on the use of the estimated Bayesian expected loss to capture uncertainty in the predictions. Once again, this should not be confused with capturing the uncertainty in the estimates of the conditional probabilities used to make those predictions. He specifically refers to the $\widehat{BEL}$ as a measure of uncertainty of the map, referring to the uncertainty in the *decision* used to make the prediction, and not in the estimates of the conditional probabilities used to make the decision.

Note that the standard errors of the probability estimates obtained with the Indicator method can be made arbitrarily small by increasing the number of MCMC

iterations, T. This can be seen by applying the usual formula for the standard error of a proportion, $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{T}}$ (Ramsey and Schafer, 2002).

4.3.2 Link Function method

Our desire to quantify uncertainty in the estimated probabilities leads us to propose an alternative method termed the *Link Function* method as a way to obtain a k-dimensional vector of probabilities for *each* iteration of the chain. The probabilities are calculated as a function of the parameters, allowing us to obtain summaries of the posterior distributions of the probabilities. Point estimates are obtained as posterior means and uncertainties in the estimates are obtained via posterior intervals or standard deviations.

The Link Function method takes advantage of the deterministic relationship defined between the categorical response and the latent continuous variable, along with the original distributional assumptions of the latent variables that in part define the two models. As with the Indicator method, we base our inference on the posterior predictive distributions discussed in Section 4.1.2. However, instead of applying Monte Carlo integration over the posterior predictive draws, we use the posterior predictive distribution by calling on our knowledge of the complete conditional distribution of the predicted latent variable, coupled with posterior draws of the remaining parameters. The method is best described by looking at the details for both of our models.

For the Clipped Gaussian model, the method is based on the distributional assumptions of the model and the factorization shown in (4.11). For each iteration, we obtain a posterior sample, $(\mathbf{W}^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \gamma^{(t)})$, from the original joint posterior distribution. We then use this posterior sample to obtain the probability that $\mathbf{W}_{0j}$ falls between two cut-point values by evaluating the cdf of $(\mathbf{W}_0|\mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \gamma, \mathbf{y})$, which is conveniently multivariate Gaussian. We perform the calculations using the conditional distribution for one location and obtain point estimates of the probabilities

as

$$\begin{aligned}\widehat{Pr}(Y_{0j}=k|\mathbf{y}) &= \frac{1}{T}\sum_{t=1}^{T}[Pr(\theta_{k-1}^{(t)} < W_{0i} \leq \theta_k^{(t)}|\mathbf{y})] \\ &= \frac{1}{T}\sum_{t=1}^{T}\left[\int_{\theta_{k-1}^{(t)}}^{\theta_k^{(t)}} f_{W_{0i}|\cdot}^{(t)}(w)dw\right], \end{aligned} \tag{4.24}$$

where $f_{W_{0j}|\cdot}(w)$ is the univariate Gaussian complete conditional distribution of $(W_{0j}|[\mathbf{W}_0]_{-j}, \mathbf{W}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\gamma}, \mathbf{y})$. Averaging over the posterior samples from all iterations effectively integrates out the remaining parameters. To specify this univariate Gaussian distribution, we denote the parameters of the multivariate Gaussian distribution as

$$\boldsymbol{\mu}_{\mathbf{W}_0}^{(t)} = \mathbf{X}_0\boldsymbol{\beta}^{(t)} + G_{\boldsymbol{\gamma}^{(t)}}\Sigma_{\boldsymbol{\gamma}^{(t)}}^{-1}(\mathbf{w}^{(t)} - \mathbf{X}\boldsymbol{\beta}^{(t)}) \tag{4.25}$$

and

$$\Sigma_{\mathbf{W}_0}^{(t)} = K_{\boldsymbol{\gamma}^{(t)}} - G_{\boldsymbol{\gamma}^{(t)}}\Sigma_{\boldsymbol{\gamma}^{(t)}}^{-1}G'_{\boldsymbol{\gamma}^{(t)}}, \tag{4.26}$$

where G_γ is the $m \times n$ correlation matrix such that $G_{\gamma,ij} = H_\gamma(\mathbf{s}_{0i}, \mathbf{s}_j)$ and K_γ is the $m \times m$ correlation matrix such that $K_{\gamma,ij} = H_\gamma(\mathbf{s}_{0i}, \mathbf{s}_{0j})$. To specify the parameters of the univariate Gaussian distribution of $(W_{0j}|[\mathbf{W}_0]_{-j}, \mathbf{y}, \cdot)$, we write $\mathbf{W}_0 = (W_{0j}, [\mathbf{W}_0]_{-j})'$ and partition the covariance matrix, $\Sigma_{\mathbf{W}_0}$, into four parts: v_{0jj} is the variance of W_{0j}, $\mathbf{v}_{0j}$ is the vector of covariances between W_{0j} and $[\mathbf{W}_0]_{-j}$, and $[\Sigma_{\mathbf{W}_0}]_{-j}$ is the variance-covariance matrix of $[\mathbf{W}_0]_{-j}$, giving

$$\Sigma_{\mathbf{W}_0} = \begin{pmatrix} v_{0jj} & \mathbf{v}'_{0j} \\ \mathbf{v}_{0j} & [\Sigma_{\mathbf{W}_0}]_{-j} \end{pmatrix}. \tag{4.27}$$

The $m \times p$ design matrix with the jth row removed is denoted $[\mathbf{X}_0]_{-j}$. Then the Gaussian distribution of $(W_{0j}|[\mathbf{W}_0]_{-j}, \mathbf{y}, \cdot)$ has mean

$$\mu_{W_{0j}} = [\boldsymbol{\mu}_{\mathbf{W}_0}]_j + \mathbf{v}'_{0j}[\Sigma_{\mathbf{W}_0}]_{-j}^{-1}([\mathbf{w}_0]_{-j} - [\mathbf{X}_0]_{-j}\boldsymbol{\beta}), \tag{4.28}$$

and variance,

$$\sigma_{W_{0j}} = v_{0jj} - \mathbf{v}'_{0j}[\Sigma_{\mathbf{W}_0}]_{-j}^{-1}\mathbf{v}_{0j}. \tag{4.29}$$

For one iteration and one location, the K probabilities are obtained as

$$\begin{aligned} Pr(Y_{0j} = k|\mathbf{y})^{(t)} &= \int_{\theta_{k-1}^{(t)}}^{\theta_k^{(t)}} f_{W_{0i}|\cdot}^{(t)}(w)dw \\ &= \int_{\theta_{k-1}^{(t)}}^{\theta_k^{(t)}} N(\mu_{W_{0j}}^{(t)}, \sigma_{W_{0j}}^{(t)})dw \\ &= \Phi\left(\frac{\theta_k^{(t)} - \mu_{W_{0j}}^{(t)}}{\sqrt{\sigma_{W_{0j}}^{(t)}}}\right) - \Phi\left(\frac{\theta_{k-1}^{(t)} - \mu_{W_{0j}}^{(t)}}{\sqrt{\sigma_{W_{0j}}^{(t)}}}\right). \end{aligned} \tag{4.30}$$

Point estimates and measures of variability are then obtained by summarizing over all T iterations. For example, the point estimate given by the posterior mean is simply

$$\widehat{Pr}(Y_{0j} = k|\mathbf{y}) = \frac{1}{T}\sum_{t=1}^{T} Pr(Y_{0j} = k|\mathbf{y})^{(t)}. \tag{4.31}$$

The Link Function method follows the same logic for the Double Latent model, relying on the posterior predictive distribution of $\mathbf{Z}_0$ rather than $\mathbf{W}_0$. The probabilities for one iteration and one location are given by,

$$\begin{aligned} Pr(Y_{0j} = k|\mathbf{y})^{(t)} &= Pr(\theta_{k-1}^{(t)} < Z_{0j} \leq \theta_k^{(t)}|\mathbf{y}, \mathbf{W}_0) \\ &= \int_{\theta_{k-1}^{(t)}}^{\theta_k^{(t)}} f_{Z_{0j}|\mathbf{y},\mathbf{w}}^{(t)}(z)dz, \end{aligned} \tag{4.32}$$

where $f_{Z_{0j}|\mathbf{y},\mathbf{w}}^{(t)}(z)$ is the univariate Gaussian density of the marginal complete conditional distribution of $(Z_{0j}|[Z_0]_{-j}, \mathbf{W}_0^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \mathbf{y})$. For each iteration, we obtain a sample, $(\mathbf{Z}^{(t)}, \mathbf{W}_0^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)})$. This posterior sample is then used to calculate the estimates by plugging these parameter values into the multivariate Gaussian complete conditional distribution of $(\mathbf{Z}_0|\mathbf{W}_0^{(t)}, \boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\gamma}^{(t)}, \mathbf{y})$, specified as

$$MVN_m(\mathbf{X}_0\boldsymbol{\beta}^{(t)} + \mathbf{W}_0^{(t)}, \mathbf{I}). \tag{4.33}$$

The univariate complete conditional distribution for each Z_{0j} is then univariate Gaussian with mean, $\mu_{Z_{0j}}^{(t)} = \mathbf{X}_{0j}\boldsymbol{\beta}^{(t)} + \mathbf{W}_{0j}^{(t)}$ and unit variance. This strategy echoes the

factorization given in (4.13). The probability, $Pr(Y_{0j} = k|\mathbf{y})^{(t)}$ is then calculated for each iteration and one location as

$$\int_{\theta_{k-1}^{(t)}}^{\theta_k^{(t)}} f_{Z_{0i}|\cdot}^{(t)}(z)dz =$$
$$\Phi\left(\theta_k^{(t)} - \mathbf{x}_{0j}\boldsymbol{\beta}^{(t)} - W_{0j}^{(t)}\right) - \Phi\left(\theta_{k-1}^{(t)} - \mathbf{x}_{0j}\boldsymbol{\beta}^{(t)} - W_{0j}^{(t)}\right). \qquad (4.34)$$

Summaries, such as posterior means and posterior intervals are calculated over all T iterations. For example, as for the Clipped Gaussian model, we obtain K point estimates for each location using the posterior mean and obtain measures of uncertainty using posterior intervals.

Table 4.3 gives a brief comparison of the estimated probabilities obtained using the Indicator and Link Function methods for five new (hold-out) locations from one realization of data generated under the Clipped Gaussian model. The estimated probabilities are extremely close in magnitude and lead to the same predictions, with the exception of site 5 where the methods predict different values. Site 5 is near to observed values from both category 1 and category 2, making it a difficult site to predict. Different predicted values occurred for only 2 out of 50 total hold-out locations, with each method predicting correctly for one of the two discrepancies. Table 4.4 further compares the two methods in terms of the minimum, mean and maximum MPR, AMPR, and $\widehat{BEL}$ over 30 realizations for the same model and data. The results of both comparisons are consistent with those obtained using other models and other sets of simulated data. On average, the Link Function method tends to produce lower estimated BEL, meaning that in De Oliveira's terms there is more certainty in the predictive decision. However, since this often does not coincide with lower MPR and AMPR, we do not give much weight to this observation. The MPR and AMPR are generally very close for the two methods, providing no evidence that one method leads to superior predictions over the other. We find it reassuring that the two methods produce comparable results in terms of the estimated probabilities

Table 4.3: Comparison of probability estimates obtained using the Indicator method (I) and Link Function method (LF) for five sites from one realization of artificial data. We denote $\widehat{Pr}(Y_i = k|\mathbf{y})$ as $\widehat{Pr}_k$. The maximum probabilities, and hence those associated with the predicted category, are shown in red for each method and site. These results are from data created under the Clipped Gaussian model and fit using the Clipped Gaussian model with $\sigma^2_{fix} = 1$.

Site	Method	$\widehat{Pr}_1$	$\widehat{Pr}_2$	$\widehat{Pr}_3$	$\widehat{Pr}_4$
1	I	0.838	0.140	0.022	0.000
	LF	0.822	0.159	0.019	0.000
2	I	0.214	0.469	0.316	0.001
	LF	0.213	0.476	0.309	0.002
3	I	0.950	0.048	0.002	0.000
	LF	0.955	0.042	0.002	0.000
4	I	0.176	0.543	0.282	0.000
	LF	0.177	0.539	0.283	0.001
5	I	0.438	0.368	0.193	0.002
	LF	0.293	0.384	0.315	0.009

and prediction, giving us confidence in the Link Function method. The disadvantage of the Link Function method is that it is more computationally demanding because the function evaluated at each iteration of the MCMC algorithm is more complicated than the simple indicator function used for the Indicator method. The computational disadvantage obviously increases with the length of the MCMC chain. However, for practical use, when the models are fit to only one data set, it should be of little hindrance.

It is of interest to note that a standard non-Bayesian and non-spatial maximum likelihood (ML) analysis estimates the probability, $Pr(Y_{0i} = k|\mathbf{x}_{0i})$, by plugging the ML parameter estimates into the original link function. For example, for the probit model with the inverse standard normal cdf link function,

$$\widehat{Pr}(Y_{0j} = k|\mathbf{x}_{0j}) = \Phi(\hat{\theta}_k - \mathbf{x}_{0j}\hat{\boldsymbol{\beta}}) - \Phi(\hat{\theta}_{k-1} - \mathbf{x}_{0j}\hat{\boldsymbol{\beta}}). \tag{4.35}$$

Table 4.4: Comparison of mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and the estimated Bayesian expected loss ($\widehat{BEL}$) for the Indicator method (I) and Link Function method (LF). These are average results over 30 realizations from artificial data created under the Clipped Gaussian model and fit using the Clipped Gaussian model with $\sigma^2_{fix} = 1$. The minimum values for the two methods are shown in red for each summary measure of predictive ability.

	MPR			AMPR			$\widehat{BEL}$		
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
I	0.140	0.370	0.520	0.160	0.465	0.700	0.220	0.313	0.458
LF	0.140	0.373	0.540	0.160	0.473	0.720	0.216	0.310	0.450

This is similar in spirit to the Link Function method in that it utilizes the underlying distributional assumption associated with the defined link function to calculate the probabilities ultimately used for prediction.

In the following chapter, we investigate the predictive ability of the models using artificial data in a simulation study. We use $\widehat{BEL}$, MPR, and AMPR to make comparisons between the Double Latent and Clipped Gaussian models, and also examine the ability of the measures to choose the true, data-generating model. In the remaining chapters, we always refer to the estimated BEL and thus we simply denote $\widehat{BEL}$ as BEL.

Chapter 5

EVALUATION OF METHODS USING SIMULATION

The use of a new statistical model or method in practice should be preceded by simulation studies investigating the effectiveness and validity of the tool using artificial data constructed from known parameter values. Applying the model to many artificial data sets with known parameter values provides a method to evaluate the accuracy and precision of the estimates, as well as the predictive ability of the model, under a wide range of data realizations. A simulation study can give insight into how model behavior relates to characteristics of a particular data set, as well as help distinguish data sets that the models can be applied to with confidence from those that should warrant caution. The simulation results may not glean the underlying reason for successes and failures, but form an important initial investigation. A secondary purpose of the simulation study is to foster an understanding of the variety of realizations that can be expected under the assumption of a given model and set of parameters. A comparison of these realizations to real (or potentially real) data sets can help a researcher decide if the model is reasonable for a particular application.

The proposal of models for ordered categorical data and for binary and count spatial data described in the literature review in Chapter 1 largely excluded simulation studies. For example, the work of Albert and Chib (1993) is applied to real data sets, but not to data sets with known model origins, and that of Cowles (1996) is applied to only one artificial data set. De Oliveira (2000) does employ his method on artificial data for a constant mean model. He creates two realizations of a binary clipped Gaussian random field and then uses two sampling schemes to sample

from the fields. However, he does not investigate the performance of his method over a variety of realizations. The lack of careful assessment of previously proposed models for ordered categorical data using simulation studies is most likely due to the computational expense. We hope to initiate more investigation of models for ordered categorical data and spatial models for non-Gaussian data via artificial data and simulation. In the course of completing the simulation study described in this chapter, we fit the models to nearly 150 different data sets.

In this chapter, we undertake a simulation study to evaluate the ability of our proposed models to predict at new locations and to estimate the true parameters. We are also interested in the relationship between success at prediction and success at parameter estimation. An investigation of the Double Latent and Clipped Gaussian models using simulation could involve an inexhaustible number of permutations of models and parameter values. Thus, we are forced to choose a reasonable sized subset of possible combinations that provide insight into the strengths and weaknesses of the models.

5.1 Creating artificial data

We create data sets under the model assumptions of the Double Latent model and the Clipped Gaussian model. The details of the models are given in Chapter 2 where the models are defined by (2.2), (2.3), and (2.34); and (2.39) and (2.40), respectively. For the simulation study, we create data under different values of the regression coefficient parameters, the cut-point parameters, and the spatial variance. These are incorporated into four different sets of parameter values, labeled as Scenarios A, B, C, and D. The scenarios are described in detail in Section 5.1.3 and the true parameter values are given in Table 5.3. Prior to introducing the details of the four scenarios, in Section 5.1.1 we provide a description of the sampling design used for all artificial data sets and in Section 5.1.2 a general discussion about specifying the

degree of spatial correlation in the data. In Section 5.2, we compare the prediction and estimation results obtained using the different data scenarios within each model (Section 5.2.1), and then compare the results of the two models within each scenario (Section 5.2.2).

5.1.1 Sampling design

The models described in this dissertation are formulated for application to point-referenced spatial data, as opposed to areal, or lattice, data. Thus, the first problem in the simulation study is to determine the number and location of the artificial data sites. The region of interest is assumed to be continuous; that is, a "site" can be located an any point within the region, leading to an uncountable number of spatial designs for any finite sample size. The problem of choosing a "best" sampling design for spatial data is beyond the scope of this research, but has been studied by others such as Zimmerman (2006). Zimmerman (2006) concludes that the objectives of prediction for new locations assuming a known spatial covariance function versus estimation of the parameters of the covariance function lead to different "optimal" designs. He proposes a hybrid criterion in an attempt to investigate optimal designs when both objectives are of interest, as is usually the case with inference for spatial data.

We follow his suggestion in considering both of these objectives in choosing our design layout. One, we desire to have a relatively large number of small inter-site distances since this is advantageous, if not necessary, for estimation of the parameters of the correlation function (Zimmerman, 2006). Second, we desire to provide adequate coverage over the area of interest. This serves the second objective of prediction at new, unsampled locations. In general, combining the two objectives suggests the use of a cluster-type design with more coverage introduced around the edges of the study region. In our experience, cluster designs using a particular radius produce

sampling designs that are indistinguishable from those created by a random design. Investigation of histograms of inter-site distances, as well as visual comparison of site locations can clearly demonstrate this. In our attempt to make this simulation study as general as possible, we seek to cover the middle ground in terms of possible spatial designs. Thus, we use a random design that produces a relatively large number of small inter-site distances as well good coverage of the study region over which we wish to predict. The choice of sample size and area of the study region aid in obtaining such a design. The actual set of locations was chosen at random, but was subject to visual inspection of the map and the histogram of inter-site distances.

We create the artificial ordered categorical data with four categories over the continuous area of a $[0, 10] \times [0, 10]$ square. We initially simulate a random sample of size $N = 175$. The sample is chosen using a Uniform$(0, 10)$ distribution for each dimension to create pairs of values, $\mathbf{s} = (s_x, s_y)$, indexing locations. A hold-out data set of $m = 50$ locations is then chosen from the original sample. These locations will serve to assess prediction at new locations. This leaves a data sample of $n = 125$ to be used for estimation of the model parameters. The specific locations of the observed and hold-out sites remain fixed across all simulations. This allows us to remove the effect of sampling design from the simulation study. The locations of the hypothetical sites are shown in Figure 5.1, with the hold-out locations shown in red. The histogram of inter-site distances is shown in Figure 5.2.

5.1.2 Spatial correlation

It is possible that the success of the models, in terms of both estimation and prediction, may depend on the spatial correlation in the categorical random field. Thus, we aim to use this information in choosing the parameter values for the simulation study. Indirectly, we seek to vary the degree of smoothness, or alternatively patchiness, in the categorical data. For a typical stationary spatial model, the degree

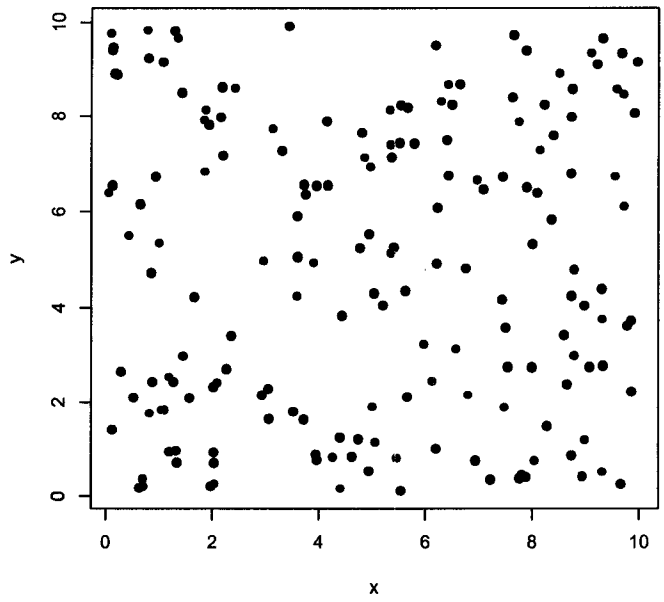

Figure 5.1: Locations of $N = 175$ sampling sites used for all data sets in the simulation study. The $m = 50$ hold-out locations used to evaluate prediction are shown in red, while the remaining $n = 125$ training locations are shown in black.

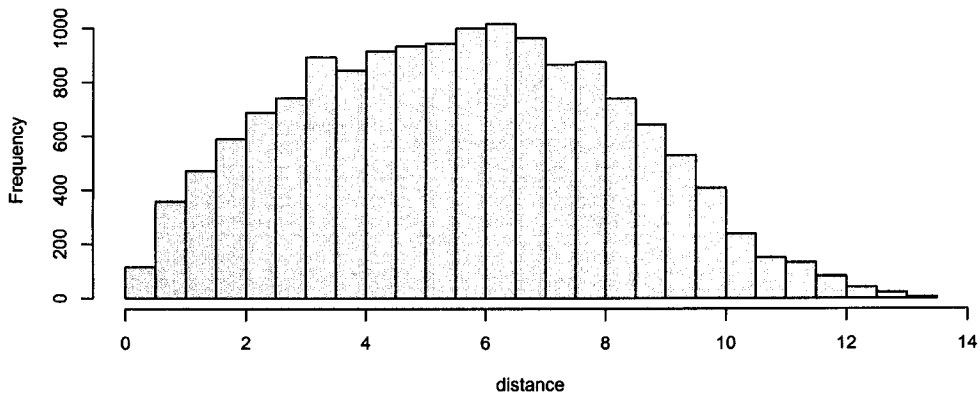

Figure 5.2: Histogram of inter-site distances from the sampling design chosen for the simulation study.

of smoothness is easily altered by varying the range and/or smoothness parameters specifying a particular correlation function. For example, using the Matérn correlation function allows us to vary the scale parameter, ϕ, and the smoothness parameter, κ,

$$\rho(d) = 2^{1-\kappa}(\Gamma(\kappa))^{-1}(d\phi)^{\kappa}K_{\kappa}(d\phi) \tag{5.1}$$

where $K(\cdot)$ is the Bessel function (Abramowitz and Stegun, 1965). The smoothness parameter, κ, is typically fixed at a value to reflect what is known about the smoothness of the underlying process. Note that the exponential correlation function is a special case of the Matérn when $\kappa = 0.5$, giving

$$\rho(d) = \exp(-d\phi). \tag{5.2}$$

If we specify a constant or zero mean for our models, then the smoothness in the continuous latent random field will more directly translate into smoothness in the categorical random field. For the special case of a zero-mean model, we calculate the practical (or effective) range for combinations of the scale and smoothness parameters, shown in Table 5.1. Irvine et al. (2007) provide a similar description of practical range for different parameter values of a correlation function. We aim to avoid fitting multiple models with essentially the same spatial correlation structure. The practical range is defined as the distance at which the correlation drops to 0.05.

The maximum possible distance for the artificial data is 14.14, and the maximum observed distance for the chosen study design is 13.48. Therefore, it does not make sense to include combinations of parameters that result in effective ranges larger than this distance. Table 5.2 gives nine combinations that represent good starting points to cover the spectrum of realistic practical ranges for these data. We visually inspected data arising from these nine sets of parameter values and found that in general it is hard to visually distinguish between realizations created under the different scenarios. The amount of variability among multiple realizations generated under

Table 5.1: Practical (effective) ranges for the Matérn correlation function for differing values of $\phi = \frac{1}{\text{range}}$ and the smoothness parameter, κ.

		κ								
ϕ	$1/\phi$	0.05	0.1	0.3	0.5	0.7	1.0	1.5	2.0	3.0
1.00	1	0.9	1.4	2.4	3.0	3.5	4.0	4.8	5.5	6.6
0.50	2	1.8	2.8	4.8	6.1	7.0	8.0	9.5	10.9	13.2
0.33	3	2.6	4.2	7.2	9.1	10.5	12.0	14.3	16.4	19.8
0.20	5	4.4	7.0	12.1	15.2	17.5	20.0	23.8	27.4	33.0
0.14	7	6.1	9.8	16.9	21.2	24.4	28.0	33.3	38.3	46.2
0.10	10	8.7	14.0	24.1	30.3	34.9	39.9	-	-	-

one set of parameter values appears to be as great as the variability due to the differences in the parameter values. Thus, we turn to devoting more attention to specification of the mean structure which does have a large and obvious affect on the smoothness of the categorical random field. We initially restrict our attention to combinations with $\kappa = 0.5$, corresponding to the exponential correlation function, and fix the scale parameter, ϕ, at 0.5. As shown in Table 5.2, this results in a practical range of approximately 6, which is close to one half the maximum inter-site distance in the data.

For the general specification of our models, however, investigating the effect of smoothness (or patchiness) on estimation and prediction is not as simple as changing only the range and/or smoothness parameters of the correlation function of the Gaussian random field. Recall that the correlation in the categorical random field involves not only the spatial parameters, but also the regression parameters, $\boldsymbol{\beta}$, the covariate, $\mathbf{X}$, and the cut-point parameters, $\boldsymbol{\theta}$ (Section 3.1). That is, the spatially varying mean is incorporated into each model prior to the clipping of the latent variable used to obtain the categorical random field. We use the results of our investigation of the second-order structure in Section 3.1 to set the values of the regression parameters to obtain different levels of correlation for a fixed covariate value, scale

Table 5.2: Practical (effective) ranges for the Matérn correlation function for differing values of ϕ and κ. The reciprocal of the scale parameter, $\frac{1}{\phi}$, is commonly referred to as the *range* parameter. However, all models in this dissertation are parameterized in terms of the *scale* parameter, ϕ.

ϕ	κ	Practical Range
1.00	0.05	0.87
1.00	0.30	2.40
1.00	0.50	3.03
1.00	1.00	4.00
0.50	0.50	6.06
0.50	1.00	8.00
0.33	0.50	9.10
0.33	0.70	10.49
0.20	0.50	15.16

parameter, and set of cut-point parameters. We choose two different $\boldsymbol{\beta}$'s that result in higher and lower correlations for most values of the covariates. We expect, and observe, that higher correlation in the categorical random field corresponds to greater smoothness (less patchiness), while lower correlation corresponds to less smoothness (greater patchiness).

5.1.3 Simulation scenarios

We create a wide variety of realizations of data with four ordered categories labeled with the integers 1 through 4. The data are created under four parameter scenarios, which we term *Scenario A*, *Scenario B*, *Scenario C*, and *Scenario D*. The true parameter values defining the four scenarios are given in Table 5.3. A comparison of Scenarios A and B investigates the effect of varying the values of the regression parameters, and thus the spatially varying mean, while that of Scenarios A and C investigates the effect of varying the amount of spacing between the true values of the cut-point parameters. The comparison of Scenarios C and D investigates the effect of changing the true variance of spatial process. We generated data under each

Table 5.3: True parameter values defining the four scenarios used in the simulation study.

Scenario	β_0	β_1	θ_1	θ_2	θ_3	ϕ	σ^2_{true}
A	-0.5	1	0	1	2	0.5	1
B	2	1	0	1	2	0.5	1
C	-0.5	1	0	0.5	1.5	0.5	1
D	-0.5	1	0	0.5	1.5	0.5	3

parameter scenario for the Double Latent model and the Clipped Gaussian model. Therefore, we refer to fitting the Double Latent model to Double Latent data *or* to Clipped Gaussian data, and vice versa. Image plots of true data created under several of the scenarios is provided in a discussion of prediction and estimation results for individual realizations in Section 5.2.2.4 (see Figures 5.10, 5.11, and 5.12).

Changes in both the regression parameters *and* the spacing of the cut-points affect the relative proportions of observations falling into the four categories. Table 5.4 displays the percentage of observations in each category over five realizations for all combinations of models and scenarios. For example, regarding changes in the intercept parameter, Scenario A produces a high percentage of observations falling in category 1, whereas Scenario B produces a high percentage of observations in category 4. Comparing Scenarios A and C we see that changing the cut-point parameters affects the percentage of observations in category 2. The difference is not dramatic, but may affect a model's success at prediction and/or estimation.

Notice from Table 5.4 that the Clipped Gaussian and Double Latent models result in a very similar break down of their observations into categories, with the only noticeable difference being that data generated under the Double Latent model have a slightly larger proportion of observations in the end categories, 1 and 4. This difference is expected and corresponds to the additional error component incorporated through the latent variable $\mathbf{Z}$ for the Double Latent model. A similar observation

Table 5.4: Proportion of observations in each category for combinations of the parameter scenarios (defined in Table 5.3) and data-generating model (DL = Double Latent and CG = Clipped Gaussian), averaged over five realizations

Scenario	Model	1	2	3	4
A	DL	0.57	0.19	0.12	0.12
	CG	0.58	0.20	0.14	0.07
B	DL	0.16	0.16	0.22	0.46
	CG	0.12	0.19	0.21	0.48
C	DL	0.64	0.09	0.14	0.13
	CG	0.62	0.12	0.15	0.11
D	DL	0.64	0.08	0.12	0.16
	CG	0.64	0.08	0.14	0.14

is made for the data generated under Scenarios C and D, where Scenario D data is generated with a larger variance and thus results in a slightly higher percentage of observations in the end categories.

As previously discussed, an additional consequence of varying the regression and/or cut-point parameters is a change in the spatial correlation between the values of the categorical response after clipping the continuous random variable. Table 5.5 provides the correlation between two categorical responses separated by a distance of 1.41 units for the covariate values provided in the table and Scenarios A, B, and C. Notice that Scenario B data generally possess higher correlations than Scenarios A and C for data generated under both models. The values for Scenario A and Scenario C are very comparable, though for Double Latent data Scenario A values were just slightly lower for every set of covariate values. Thus, we note the correlation difference between Scenario B and both Scenarios A and C, and also the lack of difference between A and C.

Investigating the effect of changing the variance of the spatial process is important because we fix σ^2 in our analyses and need to understand the effect this simplification may have on prediction and estimation. We denote the fixed value of

Table 5.5: Correlation between two categorical response values separated by a distance of 1.41 units for data generated under the Double Latent model and the Clipped Gaussian model. The values of the covariates, (x_1, x_2), are given in the heading of the column. The unique maximum values within each model and covariate combination are shown in red.

Model	Scenario	(0.5, 0.5)	(0.5, -0.5)	(-0.5,-0.5)	(0,0)	(-0.5, 1)	(-1, 0.5)
DL	A	0.198	0.172	0.156	0.180	0.174	0.152
	B	0.180	0.191	0.210	0.198	0.174	0.189
	C	0.195	0.170	0.155	0.179	0.169	0.149
CG	A	0.407	0.332	0.308	0.366	0.324	0.264
	B	0.366	0.380	0.431	0.407	0.324	0.368
	C	0.408	0.326	0.312	0.369	0.305	0.252

σ^2 used for the analysis as σ^2_{fix} and the true value used to generate the data as σ^2_{true}. Recall from Table 5.3 that $\sigma^2_{true} = 1$ for Scenarios A, B, and C and $\sigma^2_{true} = 3$ for Scenario D. This increase in the variance results in a slightly larger proportion of observations in the lowest and highest categories, as compared to Scenario C. We fix the parameter at values greater than, equal to, and less than the true values used to generate the data, running each model/data combination at $\sigma^2_{fix} = 1$ and $\sigma^2_{fix} = 2$.

We also fit the Clipped Gaussian model to data created under the Double Latent model, and vice-versa. This provides the setting to assess the ability of our summary predictive measures to "choose" the correct (data-generating) model and also provides insight into the important practical consideration of how the models will perform on data that was not generated under their exact assumptions, as this will always be the case in real applications.

The models are fit using the algorithms detailed in Chapter 2 for both models, with priors specified as described in Section 3.3. We use the same prior distributions and same starting values for all model and data combinations. The choice of prior distributions was discussed in Section 3.3. Prior to starting the formal simulation study, we investigated the effect of starting values on convergence of the

MCMC chains and resulting estimates and predictions. In general, the algorithms are insensitive to starting values with the specified priors, with the chains moving to the same area of the parameter space well within the first 100 iterations. Thus, we choose starting values of zero for the regression parameters and for the cut-point parameters choose integer values two units apart at 1 and 3. The starting value for the spatial correlation parameter is set at 1. For each realization and model/data combination, we ran the MCMC chain for 3000 iterations, which is the same number used by De Oliveira (1997, 2000). The results are obtained using a burn-in period of 500 iterations since for all runs it was judged that the algorithm appeared to converge well within this number of iterations. That is, all the posterior summaries for prediction and estimation are based on the last 2500 iterations of the chains.

5.2 Simulation results

This section reports and discusses the numerical results of the simulation study. The results for prediction are displayed in Tables 5.6 and 5.7, and those for estimation in Tables 5.8 and 5.9. We first compare results for prediction and estimation for the four different parameter scenarios within each of the two models in Section 5.2.1, and then compare the results from the Double Latent and Clipped Gaussian models within each of the scenarios in Section 5.2.2. We also discuss the ability of the predictive summary measures to choose the correct data-generating model and the effects of fixing the spatial variance parameter at values different from the truth. We also examine characteristics of individual realizations that appear to be associated with successes and failures of the two models. We describe the most relevant and interesting results in detail.

5.2.1 Comparison of scenarios within models

As first described in Section 5.1.3, the three main comparisons of interest when comparing the four parameter scenarios within each model/data combination are:

Scenario A with Scenario B, Scenario A with Scenario C, and Scenario C with Scenario D. We save the comparison of results based on $\sigma^2_{fix} = 1$ versus $\sigma^2_{fix} = 2$ until Section 5.2.2.4 and thus only refer to specific results for $\sigma^2_{fix} = 1$ in the following discussion. Furthermore, we only report results for applying the Double Latent model to Double Latent data and the Clipped Gaussian model to Clipped Gaussian data, saving the comparison of fitting the models to the "wrong" data for Section 5.2.2.3.

We first display results from the Double Latent and Clipped Gaussian models, before discussing each model separately in the following sections. We present a brief graphical comparison of the prediction results obtained from fitting both models under the four parameter scenarios with $\sigma^2_{fix} = 1$. Figure 5.3 displays boxplots of mis-prediction rate (MPR) and absolute mis-prediction rate (AMPR) for the Double Latent model fit to Double Latent data (DL on DL) and the Clipped Gaussian model fit to Clipped Gaussian data (CG on CG) over 30 data realizations for Scenarios A, B, C, and D. Notice that while Scenarios C and D reveal relatively low median=pMPR for DL on DL, Scenario A results in the lowest AMPR, revealing that when mis-predictions are made for Scenarios C and D, they tend to be farther from the true value than those for Scenario A. Also notice the larger variability in the measures for Scenario D. This is consistent with the simulation set-up since $\sigma^2_{true} = 3$ for Scenario D versus $\sigma^2_{true} = 1$ for A, B, and C. This greater variability leads to more varied realizations and a corresponding spread in the measures of success at prediction. CG on CG maintains the same ordering in magnitude for MPR and AMPR across the four scenarios, with noticeably lower AMPR than DL on DL for all scenarios. The variability in the AMPR measure is also clearly lower for CG on CG than DL on DL. Thus, the Clipped Gaussian model applied to Clipped Gaussian data produces more consistent prediction results across different data realizations, possibly because of the lower variability inherent in the Clipped Gaussian model versus the Double Latent model.

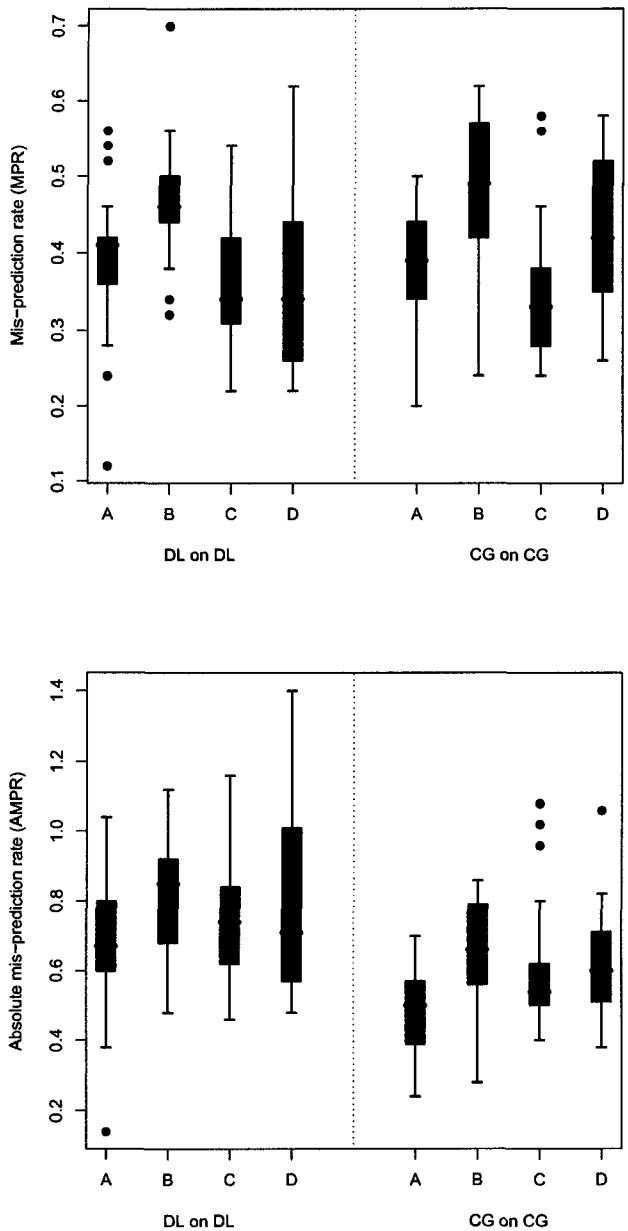

Figure 5.3: Boxplots of mis-prediction rates (MPR) and absolute mis-prediction rates (AMPR) for the Double Latent model fit to Double Latent data (DL on DL) and the Clipped Gaussian model fit to Clipped Gaussian data (CG on CG) over 30 data realizations for Scenarios A, B, C, and D.

5.2.1.1 Double Latent model: Prediction

In Table 5.6 we report the minimum, mean, and maximum values over all realizations for mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and the estimated Bayesian expected loss (BEL) for $\sigma^2_{fix} = 1$. Table 5.7 displays the same measures for $\sigma^2_{fix} = 2$ and is discussed in a later section. The minimum and maximum values are included to flag the existence of realizations that result in particularly good or bad predictions. The observations in this section are based on these summary measures of predictive ability, with most emphasis placed on the values of absolute mis-prediction rate (AMPR) since unlike MPR, it takes into account the magnitude of the prediction error. However, a concurrent look at all the values gives a more complete picture regarding the predictive success of the model on a particular type of data. Recall that lower values of all the predictive measures are indicative of better predictive ability. This section is devoted to results obtained from fitting the Double Latent Model to the Double Latent data, corresponding to the "DL DL" rows of the Table 5.6.

The results indicate that the Double Latent model is more effective in terms of prediction on data generated under Scenario A than Scenario B, with all nine predictive summary measures falling noticeably lower for Scenario A (Table 5.6). This suggests that the Double Latent model is more successful on data with more observations in the lowest category and lower correlation in the categorical random field under the given set of parameter values.

The difference in data generated by Scenario A versus C is not visually obvious, and the results are less conclusive than those for the comparison between A and B. However, Scenario A has lower values for all three summary measures of AMPR, along with minimum MPR. Recall that we place most emphasis on AMPR, thus concluding there is evidence that the Double Latent model is more successful at prediction for Scenario A than Scenario C. In fact, Scenario A exhibits the lowest

Table 5.6: Simulation results for mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and estimated Bayesian expected loss (BEL) for the hold-out set of $m = 50$ new locations for Scenarios A, B, C, and D for $\sigma^2_{fix} = 1$. The results in each row of the table are summarized over 30 realizations.

Scenario	Model	Data	MPR			AMPR			BEL		
			Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
A	DL	DL	0.12	0.40	0.56	0.14	0.68	1.04	0.26	0.37	0.52
	CG	DL	0.14	0.40	0.56	0.20	0.69	1.12	0.20	0.32	0.44
	CG	CG	0.14	0.37	0.52	0.16	0.47	0.70	0.22	0.31	0.46
	DL	CG	0.14	0.36	0.56	0.20	0.55	0.90	0.24	0.35	0.46
B	DL	DL	0.32	0.46	0.70	0.48	0.82	1.12	0.36	0.47	0.59
	CG	DL	0.42	0.55	0.70	0.60	0.79	1.04	0.38	0.44	0.52
	CG	CG	0.24	0.47	0.62	0.36	0.62	0.86	0.28	0.38	0.47
	DL	CG	0.32	0.48	0.70	0.48	0.80	1.14	0.35	0.45	0.57
C	DL	DL	0.22	0.36	0.54	0.46	0.75	1.16	0.26	0.37	0.54
	CG	DL	0.26	0.35	0.50	0.52	0.73	1.14	0.21	0.31	0.41
	CG	CG	0.24	0.35	0.58	0.40	0.61	1.08	0.17	0.30	0.42
	DL	CG	0.20	0.36	0.66	0.36	0.67	1.30	0.24	0.34	0.45
D	DL	DL	0.20	0.37	0.64	0.46	0.84	1.50	0.25	0.36	0.49
	CG	DL	0.20	0.33	0.58	0.40	0.69	1.30	0.18	0.31	0.41
	CG	CG	0.26	0.43	0.58	0.38	0.62	1.06	0.25	0.37	0.48
	DL	CG	0.22	0.38	0.62	0.42	0.80	1.52	0.23	0.36	0.51

Table 5.7: Simulation results for mis-prediction rates for the hold-out set of $m = 50$ new locations for scenarios A, B, C, and D for $\sigma^2_{fix} = 2$. The results are summarized over 30 realizations for Scenarios A, B, and D, and 20 realizations for Scenario C.

Scenario	Model	Data	MPR			AMPR			BEL		
			Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
A	DL	DL	0.12	0.40	0.56	0.14	0.69	1.04	0.26	0.36	0.51
	CG	DL	0.16	0.40	0.54	0.24	0.69	1.08	0.20	0.32	0.44
	CG	CG	0.12	0.36	0.50	0.14	0.48	0.72	0.22	0.32	0.47
	DL	CG	0.14	0.36	0.56	0.20	0.55	0.90	0.25	0.35	0.45
B	DL	DL	0.32	0.47	0.68	0.54	0.84	1.10	0.36	0.47	0.59
	CG	DL	0.32	0.48	0.74	0.52	0.86	1.40	0.30	0.40	0.48
	CG	CG	0.18	0.45	0.58	0.26	0.68	0.94	0.28	0.38	0.46
	DL	CG	0.32	0.48	0.70	0.48	0.80	1.14	0.35	0.45	0.57
C	DL	DL	0.22	0.36	0.54	0.46	0.76	1.22	0.26	0.36	0.52
	CG	DL	0.24	0.35	0.48	0.52	0.71	1.08	0.22	0.31	0.39
	CG	CG	0.22	0.34	0.60	0.40	0.59	1.14	0.18	0.31	0.42
	DL	CG	0.20	0.35	0.66	0.36	0.65	1.30	0.24	0.34	0.44
D	DL	DL	0.20	0.37	0.58	0.44	0.84	1.44	0.25	0.36	0.52
	CG	DL	0.20	0.30	0.40	0.40	0.63	0.88	0.18	0.31	0.39
	CG	CG	0.22	0.37	0.60	0.34	0.58	1.00	0.25	0.35	0.46
	DL	CG	0.22	0.37	0.62	0.48	0.80	1.52	0.23	0.35	0.48

AMPR values for the Double Latent model of the four scenarios, with much lower AMPR values than for either Scenario B or C. Scenarios B and C reveal comparable values for the minimum and mean values, although B results in a lower maximum AMPR. These results do not support the conclusion that the Double Latent model has better predictive ability on data sets with a larger proportion of observations in the lowest category, as was suspected due to its success on Scenario A data. However, this may provide evidence of better predictive performance for lower correlation in the categorical response.

A comparison of results from Scenarios C and D reveals lower mean and maximum values for both MPR and AMPR for Scenario C, with equal values for minimum AMPR. Thus, there is evidence that the model is more predictively successful on Scenario C than Scenario D data, placing the Scenario D data at the bottom of the list of predictive ability. The magnitude of the AMPR values for Scenario D is similar to those of Scenarios B and C, with the largest difference revealed for maximum AMPR where the Scenario D value is 1.50 compared to the next lowest value of 1.16 for Scenario C. Thus, the number of mis-predictions is on average the same for Scenarios C and D, but for at least one realization of Scenario D the mis-predictions are much farther from the true value. The results in this paragraph lend evidence toward the conclusion that the Double Latent model may have better predictive ability on data with smaller variability. However, this explanation may be somewhat naive given that we cannot ignore the fact that these results were obtained for $\sigma^2_{fix} = 1$ which is equal to σ^2_{true} for Scenarios A, B, and C, but not for Scenario D where $\sigma^2_{true} = 3$. However, as is further revealed in Section 5.2.2.4, this does not appear to be the sole cause of the poorer prediction results for Scenario D.

In conclusion, the Double Latent model revealed best predictive ability when applied to data characterized by a large percentage of observations in the lowest category and relatively low correlation in the categorical response.

5.2.1.2 Clipped Gaussian model: Prediction

The observations in this section are now based on the results from fitting the Clipped Gaussian model to Clipped Gaussian data, corresponding to the "CG CG" rows in Table 5.6. In general, the results for the Clipped Gaussian model follow those described in the previous section for the Double Latent model, leading to the same general conclusion. For example, as for the Double Latent model, all nine summary measures of predictive ability for the Clipped Gaussian model are lower for Scenario A than Scenario B, suggesting that the Clipped Gaussian model also tends to be more successful at prediction for data with a larger percentage of observations in the lowest category and/or lower correlation in the categorical random field.

Interestingly, Scenario C revealed lower values for all three measures of BEL, meaning that Scenario C resulted in higher maximum probabilities and thus greater certainty in the predictive decision. However, the relative magnitudes of BEL for Scenario C are in conflict with the scenario's larger AMPR values, indicating that the Clipped Gaussian model is making wrong predictions with false confidence for data generated under Scenario C, as a result of the cut-points that are closer to zero.

The use of the artificial data in De Oliveira (2000) resulted in mis-prediction rates of 0.16 and 0.19 for a very smooth realization, and 0.16 and 0.22 for the patchier realization. These values are consistent with the values we obtained for our better predicted realizations, as demonstrated by the minimum MPR column. We find this encouraging since we are dealing with the more complicated problem of predicting an ordered response rather than a binary response, along with the fact that the realizations to which we apply the models are typically more patchy than those in De Oliveira (2000). He also obtains BEL measures that are consistent with the values we obtain, ranging from 0.21 to 0.41 over his four examples.

5.2.1.3 Double Latent model: Estimation

We also compare results across the four data scenarios in terms of estimation of the parameters. We focus mainly on the average estimation error over the realizations, as summarized in Tables 5.8 and 5.9. Average estimation error is defined as the mean of the estimation errors over L data realizations,

$$\frac{1}{L}\sum_{l=1}^{L}(\tilde{\theta}_l - \theta), \tag{5.3}$$

where $\tilde{\theta}_l$ is the parameter estimate for realization l and θ is the true parameter used to generate the data. For our work, $\tilde{\theta}_l$ is the mean of the draws from the posterior distribution obtained via the MCMC algorithm. Tables 5.8 and 5.9 also display the 5th and 95th percentiles of the distribution of the estimation errors. The mean and the interval together provide information regarding the location and spread of the estimates due to differences in the data sets generated under the same model, thus giving a summary of how the models might behave when applied to real data sets. In the following discussion, we assume that we are actually referring to the absolute value of the estimation error when we refer to the magnitude of the average estimation error. That is, a "small" value should be interpreted as close to zero.

Successful estimation is obviously characterized by small estimation error coupled with small variability. However, interpreting the values as indicating a degree of success at estimation is not simply an exercise in searching for the smallest values and narrowest intervals. For example, it is important to take into account the magnitude of the variability (width of the interval) relative to the magnitude of the estimation error. A very narrow interval relative to the magnitude of the average estimation error indicates that the estimates are biased in a predictable and consistent direction, whereas a slightly larger estimation error with a wider interval may indicate that the estimates are more variable, but center around the true parameter value. We provide discussion regarding the estimation results for both models, and again restrict our attention to the case of $\sigma^2_{fix} = 1$ (Table 5.8) and results from fitting each model to its respective data set. We will refer to Table 5.9 in Section 5.2.2.4.

Table 5.8: Average estimation errors and the 5th and 95th percentiles of the estimation errors over 30 realizations for Scenarios A, B, C, and D. The vector of true values (β_0, β_1, θ_0, θ_2, θ_3, ϕ, σ^2) for each scenario is, A:(-0.5, 1, 0, 1, 2, 0.5, 1), B:(2, 1, 0, 1, 2, 0.5, 1), C:(-0.5, 1, 0, 0.5, 1.5, 0.5, 1), D:(-0.5, 1, 0, 0.5, 1.5, 0.5, 3). The results are for $\sigma^2_{fix} = 1$.

Scenario	Model	Data	β_0	β_1
A	DL	DL	0.05 (-0.46, 0.54)	-0.21 (-0.62, 0.22)
	CG	DL	0.20 (-0.09, 0.48)	-0.52 (-0.88, -0.20)
	CG	CG	0.30 (-0.15, 0.82)	-0.13 (-0.44, 0.23)
	DL	CG	-0.03 (-0.60, 0.59)	-0.01 (-0.45, 0.45)
B	DL	DL	-0.19 (-0.77, 0.69)	-0.14 (-0.54, 0.33)
	CG	DL	-1.06 (-1.31, -0.09)	-0.57 (-0.75, -0.44)
	CG	CG	-0.07 (-0.61, 0.74)	-0.13 (-0.46, 0.20)
	DL	CG	0.34 (-0.66, 1.64)	0.12 (-0.35, 0.57)
C	DL	DL	0.03 (-0.31, 0.48)	-0.25 (-0.45, 0.02)
	CG	DL	0.15 (-0.11, 0.39)	-0.57 (-0.76, -0.30)
	CG	CG	0.11 (-0.30, 0.44)	-0.38 (-0.61, -0.15)
	DL	CG	0.01 (-0.47, 0.49)	0.01 (-0.45, 0.58)
D	DL	DL	0.09 (-0.44, 0.68)	-0.37 (-0.69, -0.03)
	CG	DL	0.24 (-0.07, 0.63)	-0.63 (-0.85, -0.40)
	CG	CG	0.54 (0.00, 1.19)	-0.16 (-0.61, 0.15)
	DL	CG	0.11 (-.059, 0.69)	-0.25 (-0.71, 0.37)

Scen.	Model	Data	θ_2	θ_3	ϕ
A	DL	DL	-0.31 (-0.53, -0.04)	-0.67 (-0.93, -0.33)	-0.05 (-0.09, 0.02)
	CG	DL	-0.56 (-0.71, -0.38)	-1.12 (-1.33, -0.88)	0.69 (0.43, 1.01)
	CG	CG	0.05 (-0.15, 0.30)	0.11 (-0.25, 0.61)	1.02 (0.64, 1.49)
	DL	CG	-0.14 (-0.37, 0.27)	-0.20 (-0.68, 0.39)	-0.06 (-0.12, 0.00)
B	DL	DL	0.11 (-0.25, 0.58)	-0.18 (-0.63, 0.54)	-0.06 (-0.14, 0.01)
	CG	DL	-0.54 (-0.68, -0.35)	-1.09 (-1.23, -0.93)	0.77 (0.46, 1.01)
	CG	CG	0.13 (-0.40, 1.24)	0.18 (-0.29, 1.41)	1.17 (0.74, 1.77)
	DL	CG	0.46 (-0.17, 1.21)	0.34 (-0.39, 1.39)	-0.05 (-0.09, 0.02)
C	DL	DL	-0.14 (-0.27, 0.05)	-0.51 (-0.75, -0.26)	-0.05 (-0.09, 0.01)
	CG	DL	-0.28 (-0.41, -0.16)	-0.86 (-1.08, -0.65)	0.71 (0.39, 1.10)
	CG	CG	-0.20 (-0.31, -0.04)	-0.65 (-0.95, -0.26)	0.59 (0.29, 0.98)
	DL	CG	-0.03 (-0.17, -0.15)	-0.21 (-0.59, 0.15)	-0.05 (-0.09, 0.02)
D	DL	DL	-0.20 (-0.28, -0.09)	-0.69 (-0.87, -0.41)	-0.04 (-0.10, 0.01)
	CG	DL	-0.29 (-0.37, -0.22)	-0.92 (-1.08, -0.78)	0.50 (0.12, 0.96)
	CG	CG	0.48 (0.15, 0.70)	0.49 (-0.06, 0.77)	1.37 (0.91, 1.78)
	DL	CG	-0.18 (-0.31, 0.09)	-0.53 (-0.81, -0.10)	-0.05 (-0.12, 0.00)

Table 5.9: Average estimation errors and the 5th and 95th percentiles of the estimation errors for over 30 realizations for Scenarios A, B, and D and 20 realizations for Scenario C. The vector of true values (β_0, β_1, θ_0, θ_2, θ_3, ϕ, σ^2) for each scenario is, A:(-0.5, 1, 0, 1, 2, 0.5, 1), B:(2, 1, 0, 1, 2, 0.5, 1), C:(-0.5, 1, 0, 0.5, 1.5, 0.5, 1), D:(-0.5, 1, 0, 0.5, 1.5, 0.5, 3). The results are for $\sigma^2_{fix} = 2$.

Scenario	Model	Data	β_0	β_1
A	DL	DL	-0.07 (-0.67, 0.51)	-0.08 (-0.56, 0.43)
	CG	DL	0.06 (-0.33, 0.46)	-0.36 (-0.83, 0.09)
	CG	CG	0.09 (-0.44, 0.59)	0.02 (-0.34, 0.47)
	DL	CG	-0.16 (-0.83, 0.54)	0.14 (-0.36, 0.72)
B	DL	DL	0.17 (-0.60, 1.12)	0.01 (-0.50, 0.58)
	CG	DL	-0.69 (-1.01, -0.38)	-0.33 (-0.63, -0.05)
	CG	CG	-0.33 (-0.90, 0.31)	-0.18 (-0.54, -0.21)
	DL	CG	0.87 (-0.45, 2.55)	0.34 (-0.25, 0.92)
C	DL	DL	-0.04 (-0.55, 0.62)	-0.12 (-0.48, 0.35)
	CG	DL	0.03 (-0.24, 0.33)	-0.41 (-0.66, -0.10)
	CG	CG	-0.01 (-0.62, 0.45)	-0.17 (-0.40, 0.03)
	DL	CG	-0.15 (-0.72, 0.47)	0.21 (-0.42, 0.89)
D	DL	DL	0.01 (-0.63, 0.67)	-0.26 (-0.64, 0.15)
	CG	DL	0.09 (-0.31, 0.53)	-0.51 (-0.74, -0.24)
	CG	CG	0.00 (-0.57, 0.62)	-0.30 (-0.63, -0.00)
	DL	CG	-0.03 (-0.86, 0.64)	-0.10 (-0.61, 0.67)

Sc.	Model	Data	θ_2	θ_3	ϕ
A	DL	DL	-0.23 (-0.49, 0.13)	-0.57 (-0.89, -0.16)	-0.05 (-0.09, 0.02)
	CG	DL	-0.39 (-0.58, -0.16)	-0.78 (-1.09, -0.45)	0.67 (0.44, 0.02)
	CG	CG	0.05 (-0.24, 0.44)	0.10 (-0.30, 0.58)	0.69 (0.34, 1.26)
	DL	CG	-0.04 (-0.34, 0.49)	-0.08 (-0.62, 0.72)	-0.05 (-0.11, 0.00)
B	DL	DL	0.38 (-0.00, 0.92)	0.19 (-0.36, 1.02)	-0.06 (-0.12, -0.00)
	CG	DL	-0.34 (-0.56, -0.06)	-0.68 (-0.99, -0.35)	0.77 (-0.43, 1.04)
	CG	CG	-0.16 (-0.37, 0.08)	-0.38 (-0.59, -0.05)	0.51 (0.17, 0.89)
	DL	CG	0.91 (-0.06, 2.19)	0.84 (-0.20, 2.09)	-0.05 (-0.09, 0.02)
C	DL	DL	-0.09 (-0.24, 0.14)	-0.40 (-0.70, -0.05)	-0.05 (-0.09, 0.02)
	CG	DL	-0.23 (-0.33, -0.13)	-0.67 (-0.89, -0.46)	0.61 (0.37, 0.97)
	CG	CG	-0.04 (-0.22, 0.15)	-0.27 (-0.72, 0.03)	0.61 (0.34, 1.00)
	DL	CG	0.02 (-0.18, 0.23)	-0.07 (-0.49, 0.36)	-0.05 (-0.09, 0.02)
D	DL	DL	-0.15 (-0.25, 0.05)	-0.59 (-0.79, -0.26)	-0.04 (-0.07, -0.01)
	CG	DL	-0.19 (-0.27, -0.11)	-0.65 (-0.78, -0.50)	0.52 (0.19, 0.87)
	CG	CG	-0.20 (-0.35, -0.01)	-0.56 (-0.96, -0.11)	0.37 (0.09, 0.69)
	DL	CG	-0.28 (-0.41, -0.05)	-0.72 (-1.03, -0.38)	-0.04 (-0.11, 0.03)

We now focus on the results for the Double Latent model, corresponding to the "DL DL" rows in Table 5.8. Inspection of the table reveals that nearly all interval widths are relatively wide compared to the magnitudes of the average estimation errors. This is especially true for the intercept parameter, β_0, over all four data scenarios. For example, Scenario A results in a small estimation error of 0.05, with an interval that spans from approximately -0.5 to 0.5. Scenario B results in the widest intervals for all parameters except ϕ, although it also reveals the lowest estimation error for β_1, θ_2, and θ_3, making it difficult to conclude that one scenario demonstrates superior estimation. This result is particularly interesting given that in the previous section Scenario A demonstrated superior predictive ability over Scenario B. Thus, it appears that success at estimation and success at prediction do not necessarily coincide.

For each estimated parameter, we now choose a "best" scenario in terms of estimation, taking into account both the average estimation error and the variability of the estimates as described by the 5th and 95th quantiles for the estimation errors. For β_0, Scenario C data reveals the lowest estimation error (0.03) and the narrowest interval, although there is minimal difference between Scenarios A, C, and D. Recall that the difference between Scenarios A and C is in the true values of the cut-points θ_2 and θ_3 and the difference between Scenario B and the other scenarios is the true value of β_0. Thus, it appears that the models can better estimate β_0 when it is set to -0.5, rather than 2.0. Recall that Scenario B is the only scenario with its highest proportion of observations falling in the highest category rather than the lowest, along with possessing higher correlation in the categorical response.

The estimate of β_1, and particularly the sign, is arguably the most important parameter estimate in terms of interpretation when investigating the relationship between the response and a covariate. For β_1, Scenario B reveals the lowest estimation error at -0.14. Despite the wider interval for Scenario B, we consider it a positive

characteristic that it is better centered around zero than those of the other scenarios. Thus, we choose Scenario B as most successful at estimating this important parameter, although one may argue a lack of practical difference between the results for Scenarios A and B. Recall that Scenario B is characterized by a different value of β_0, resulting in a large proportion of observations in the highest category along with higher correlation in the categorical response.

It is interesting that under all four scenarios and both models, the average estimation bias for β_1 is negative and the intervals are skewed to include more negative values. This indicates a tendency to underestimate the true parameter value. Since the true parameter value is 1 for all scenarios, this indicates that estimates are closer to zero than they should be, indicating that the models may err on the side of failing to detect a significant relationship between the response and the covariate. It is not surprising that the estimation results for this parameter are worst for Scenario D given that the data is generated with higher variability, likely washing out some of the signal from the relationship between the response and the covariate.

We consider estimation of the cut-point parameters, θ_2 and θ_3, together. Interestingly, Scenario B results in parameter estimates with the lowest average estimation error. The intervals for Scenario B are wider than for the other scenarios, but they are also much more centered around zero. In fact, both endpoints of the intervals are negative or nearly negative for the other scenarios. It is interesting that Scenario B seems to result in the worst estimation for the intercept parameter, β_0, but the best for the cut-point parameters given the similar role of these parameters. For Scenarios A, C, and D, the estimation errors are relatively large negative values with relatively narrow intervals, indicating predictable and consistent error in the estimates of these parameters. This may actually be a positive characteristic when considering prediction, moving both cut-points in the same direction a specified amount. However, we expected less displacement of the cut-points to the left for Scenario D where the

latent continuous variable is assumed to possess more variability, and hypothetically leading to cut-point estimates that are more spread out. However, the poor and unexpected results for Scenario D could again be tied to the fact that $\sigma^2_{fix} \neq \sigma^2_{true}$. It is clear, however, that successful estimation of the cut-point parameters on average across realizations is not needed for successful prediction. The parameters are also typically not of interest for interpretation and thus we place little emphasis on them.

The estimates for the spatial scale parameter, ϕ, vary little across the four scenarios. Average estimation error ranges from -0.04 to -0.06 and the interval widths are between 0.10 and 0.15. The intervals are skewed to include more negative values, indicating that the estimates fall consistently below the true value by only a small magnitude. The small difference between the true and estimated value results in little practical difference, since an estimation error of -0.05 translates into a difference in "effective range" of 6.7 for the estimate and 6.0 for the true value. The consistency of these estimates deserves further investigation.

5.2.1.4 Clipped Gaussian model: Estimation

In this section we examine the results obtained from fitting the Clipped Gaussian model to Clipped Gaussian data, corresponding to the "CG CG" rows in Table 5.8. For the Clipped Gaussian model, like the Double Latent, the variability in the estimates is relatively high compared to the magnitudes of the estimation errors. This is again especially true for the intercept parameter, β_0. In contrast to the Double Latent model results, Scenario B reveals the lowest average estimation error at -0.07. It is the only scenario exhibiting a negative value and also has the widest associated interval. The estimation errors for scenarios A and D are large and positive, at 0.30 and 0.54. Taking into account both its small estimation error and relatively narrow interval, we conclude that Scenario C results in the best estimation of β_0.

The average estimation errors and standard deviations of β_1 are equal (after rounding) for Scenarios A and B, and these are selected as the best scenarios based

on lowest estimation error and associated variability. As with the Double Latent, the average estimation errors are negative for all four scenarios, suggesting that both models have a tendency to underestimate this parameter, possibly leading to errors in detecting a significant relationship between the response and a covariate. Once again, while there appeared to be clear differences between Scenarios A and B for prediction, the differences in estimation of the important regression parameter are barely noticeable. It is encouraging that the different sets of parameter values used to generate the data under the four scenarios do not lead to noticeable differences in terms of estimation of this parameter.

We assess the estimation of the two cut-point parameters together. Scenario A results in the lowest average estimation error and and interval width for both θ_2 and θ_3. Thus, Scenario A reveals the best estimation behavior of the cut-point parameters for the Clipped Gaussian model along with good predictive ability. Interestingly, the average estimation error is positive for all scenarios except for Scenario C, where both the interval endpoints are negative. Recall that Scenario C differs from Scenario A in the true values of these cut-point parameters, but it is not clear why one set of true values should be associated with better estimation than the other. It is also noted that for both models there is higher estimation error and variability in the estimates for higher cut-point, θ_3, than the lower cut-point, θ_2. This is likely due to the fact that the interval of possible values for θ_2 is restricted on both ends, whereas the interval for θ_3 has a right endpoint of ∞, thus leading to greater variability in the estimates.

The estimates of ϕ for the Clipped Gaussian model have much greater estimation error and greater variability than those for the Double Latent model for all data scenarios. Scenario C reveals the smallest average estimation error (0.59), the narrowest interval, and the interval that is closest to zero. Recall that Scenario C data has less correlation than Scenario B and approximately the same as Scenario A. We discuss estimation of this parameter in more detail in next section.

5.2.1.5 Estimation for individual realizations

In the interest of brevity, we have until now only reported the estimation error averaged over all realizations. It is of course more informative and of greater interest to view the point estimates and their associated uncertainty intervals for all realizations. This provides the additional necessary information regarding the width of the posterior intervals and their coverage of the true parameter values. We choose several plots displaying these point estimates and intervals that are representative of those observed for other model/data combinations. The plots are included in Figures 5.4, 5.5, 5.6, and 5.7, and display results for both the Double Latent model on Double Latent data (upper plot) and the Clipped Gaussian model on Clipped Gaussian data (lower plot). In this section, we provide a brief discussion of conclusions from each plot.

Figure 5.4 displays point estimates and 95 posterior intervals for β_0 for Scenario A with $\sigma^2_{fix} = 1$. The results for the Double Latent model on Double Latent data (DL on DL) are given in the upper plot and those for the Clipped Gaussian model on Clipped Gaussian data (CG on CG) are given in the lower plot. As given in Table 5.8, it is clear from the plots that the magnitude of the average estimation error is greater for Clipped Gaussian model (lower plot) than the Double Latent model (upper plot). DL on DL has noticeably wider 95% posterior intervals with widths of approximately 1 versus 0.5 for the CG on CG. The wider widths, in addition to being better centered around the true value, resulted in better coverage of the true value for the Double Latent (26/30 realizations) compared to CG on CG (18/30 realizations).

Figure 5.5 displays the point estimates and posterior intervals for the important regression parameter, β_1. The results for the Double Latent model on Double Latent data (DL on DL) are again given in the upper plot and those for the Clipped Gaussian model on Double Latent data (CG on DL) are given in the lower plot. DL on

DL results in wider posterior intervals than those obtained by fitting the Clipped Gaussian model to the same data, and also in clearly better coverage of the true value $\beta_1 = 1$. In practice we might use the 95% posterior interval to help make a conclusion regarding significance of the relationship between the response and the covariate. That is, the exclusion of zero is taken as evidence of a significant relationship. Both models capture the significance of the true positive relationship for most realizations, with 25/30 and 26/30 intervals excluding zero for DL on DL and CG on DL respectively. Thus, the fact that the posterior interval fails to include the true data-generating value does not mean that it fails to detect the existence and direction of the relationship between response and the covariate.

While Figure 5.5 displays the results of fitting both models to Double Latent data, Figure 5.6 displays the results for β_1 of fitting both models to Clipped Gaussian data, (CG on CG vs DL on CG). The results in this figure are for $\sigma^2_{fix} = 2$ instead of $\sigma^2_{fix} = 1$. Notice that the posterior intervals for DL on CG (lower plot) are appreciably wider than those for CG on CG (upper plot). Coverage of the true value is actually better for DL on CG, again probably due to the large width of the intervals. All intervals for both situations exclude zero, indicating that the true positive relationship between the response and the covariate is captured for every realization. Thus, similar to the results in Figure 5.5, even when the posterior interval fails to include the true data-generating value, it leads to the correct conclusion regarding the existence and direction of the relationship between response and the covariate.

Figure 5.7 displays estimation results for the correlation parameter, ϕ, for each model applied to its respective data set, DL on DL (upper plot) and CG on CG (lower plot) for $\sigma^2_{fix} = 1$. The Double Latent model clearly results in lower estimation errors and superior coverage of the true value as compared to CG on CG. Despite the poor performance of CG on CG, notice that for realization 28 it is relatively successful.

This realization has a very large proportion of observations in the lowest category, resulting in a very smooth realization with a category composition of $(126, 38, 10, 1)$ for the $N = 175$ total observations. This is in contrast to realization 26 that results in poor estimation of the parameter. The observations for realization 26 are more spread out across the categories with a composition of $(96, 43, 25, 11)$. It does make sense that the estimate of ϕ should be smaller for smoother (less patchy) realizations, corresponding to a larger range parameter translating into higher correlation between close locations.

5.2.1.6 Estimation of the regression parameter

We now return to a discussion of estimation of the regression parameter, β_1. As briefly described in the previous section, an important check on the models in terms of practical use is that they are able to capture the potential significance of a regression parameter by accurately describing the sign and magnitude of the relationship between the response and the covariate. Recall that the data were generated by setting $\beta_1 = 1$ and thus we evaluate if parameter estimates and posterior intervals result in accurate conclusions regarding the existence and sign of this relationship used to create the data. In general, both models were successful at describing the true positive relationship, even when applied to data generated under the other model (as demonstrated by Figure 5.5.) In particular, 95% posterior intervals for the Clipped Gaussian model on Clipped Gaussian data exclude zero and provide evidence that β_1 is different from zero for every realization, for all data Scenarios, and for both fixed values of σ^2_{fix}. The Clipped Gaussian model was less successful on the Double Latent data, ranging from no intervals covering zero for Scenario B to 7 intervals covering zero for Scenario D. We expect performance to be worst on Scenario D for the reasons alluded to earlier. For the Double Latent model on Double Latent data, the number of 95% posterior intervals including zero ranged from zero on Scenario C data to 5 on Scenario D data. The widths of the posterior intervals for β_1 are

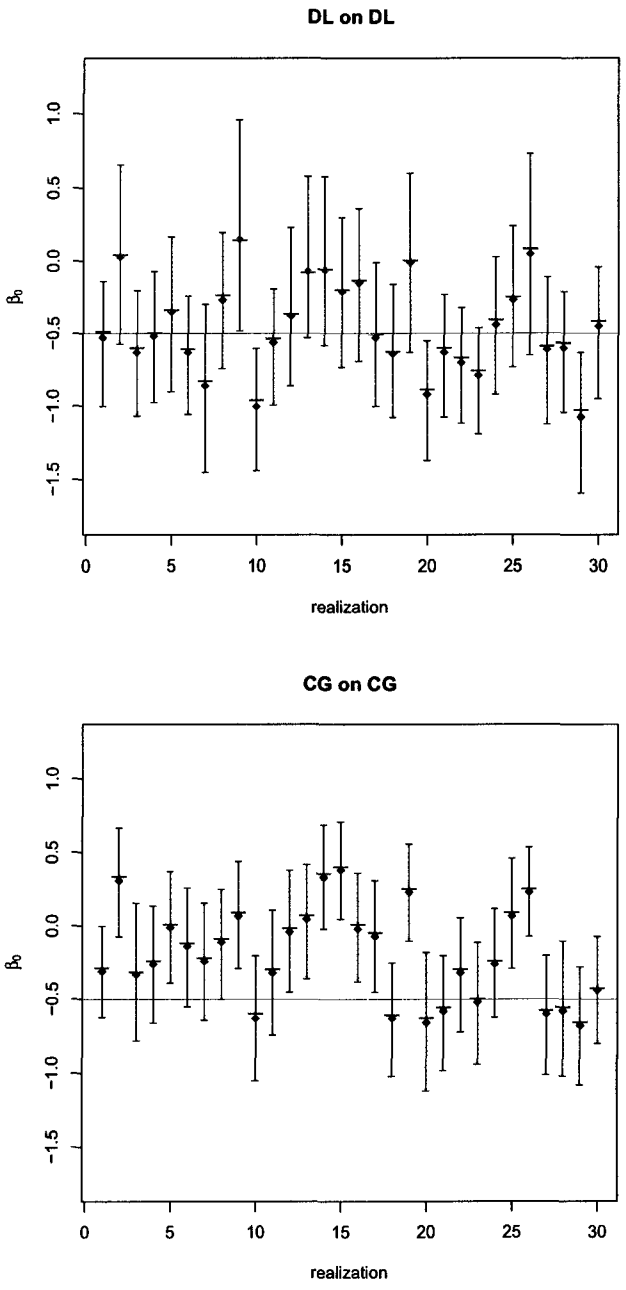

Figure 5.4: Point estimates and 95% posterior intervals for β_0 for Scenario A and $\sigma^2_{fix} = 1$. The red line shows the true value of $\beta_0 = -0.5$. Means are denoted by the red triangles and medians by the green bars. The top plot displays the results for the Double Latent model on Double Latent data (DL on DL) and the bottom plot displays the results for the Clipped Gaussian model on Clipped Gaussian data (CG on CG).

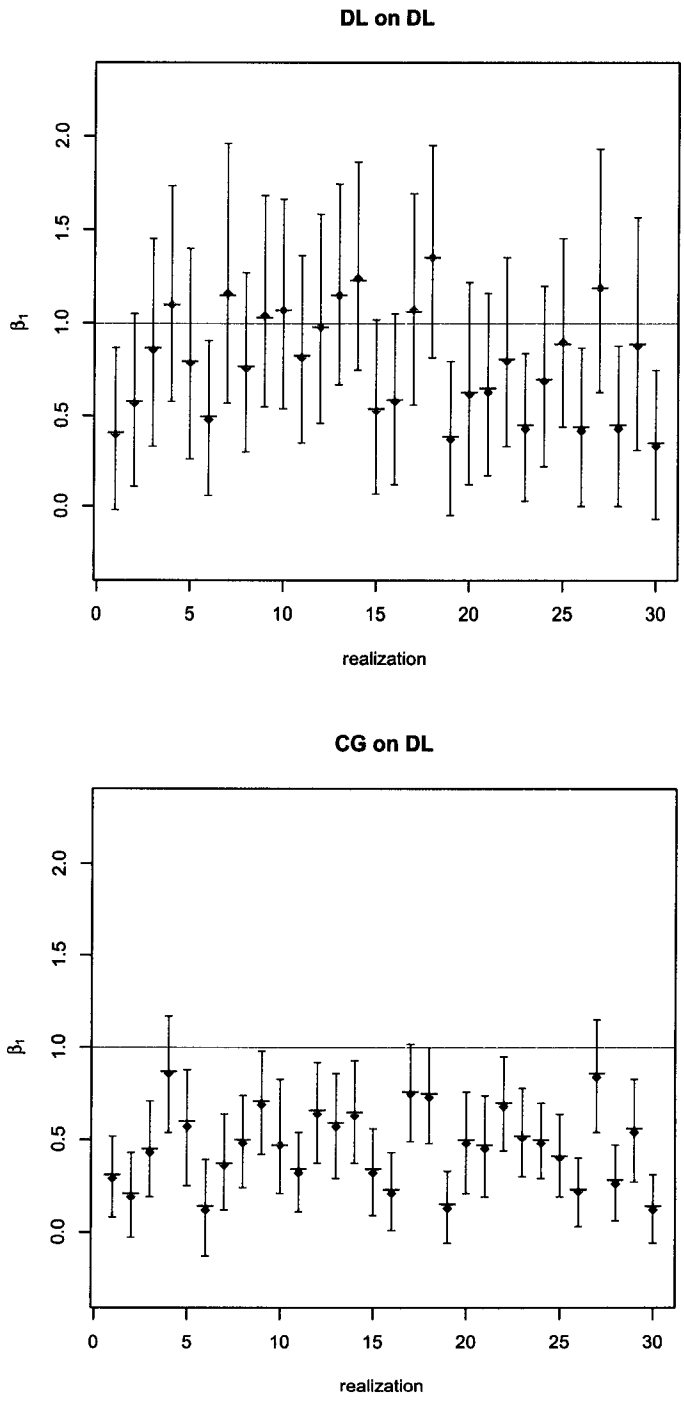

Figure 5.5: Point estimates and 95% posterior intervals for β_1 for Scenario A and $\sigma^2_{fix} = 1$. The red line shows the true value of $\beta_1 = 1$. Means are denoted by the red triangles and medians by the green bars. The top plot displays the results for the Double Latent model on Double Latent data (DL on DL) and the bottom plot displays the results for the Clipped Gaussian model on Double Latent data (CG on DL).

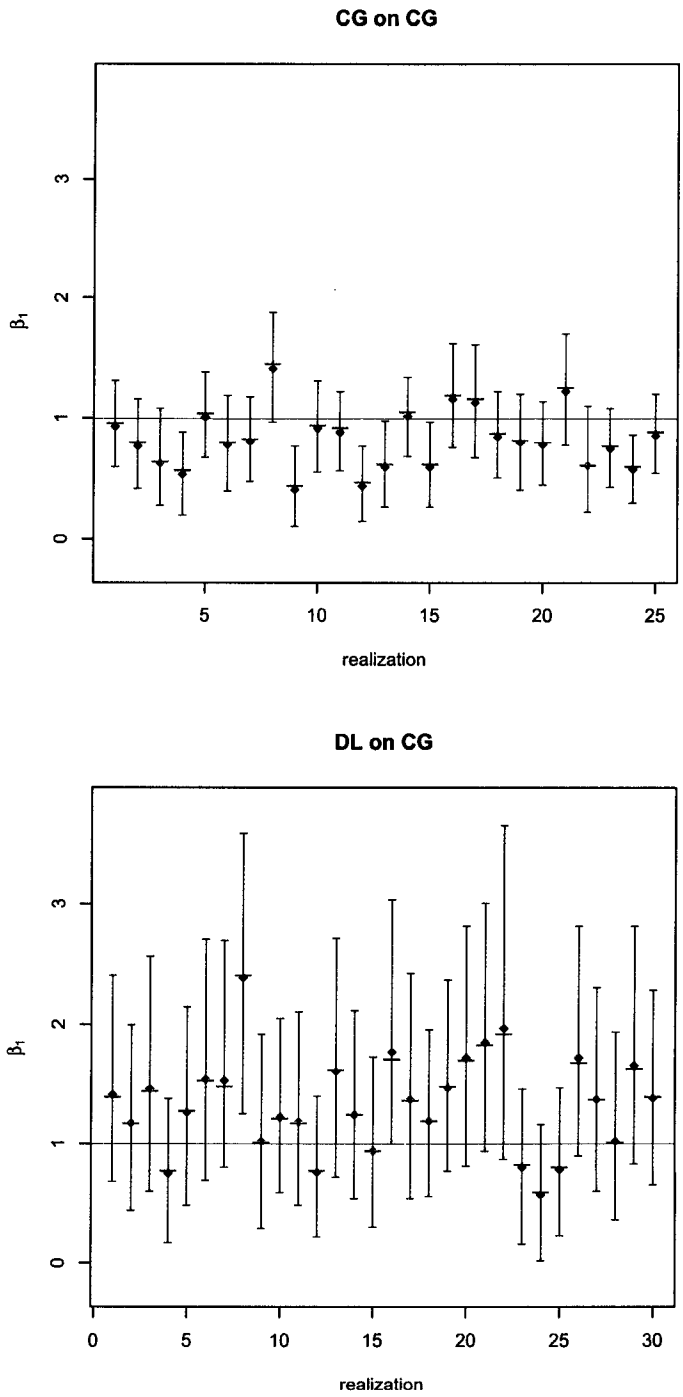

Figure 5.6: Point estimates and 95% posterior intervals for β_1 for Scenario B and $\sigma^2_{fix} = 2$. The red lines shows the true values of $\beta_1 = 1$. Means are denoted by the red triangles and medians by the green bars. The top plot displays the results for the Clipped Gaussian model on Clipped Gaussian data (CG on CG) and the bottom plot displays the results for the Double Latent model on Clipped Gaussian data (DL on CG).

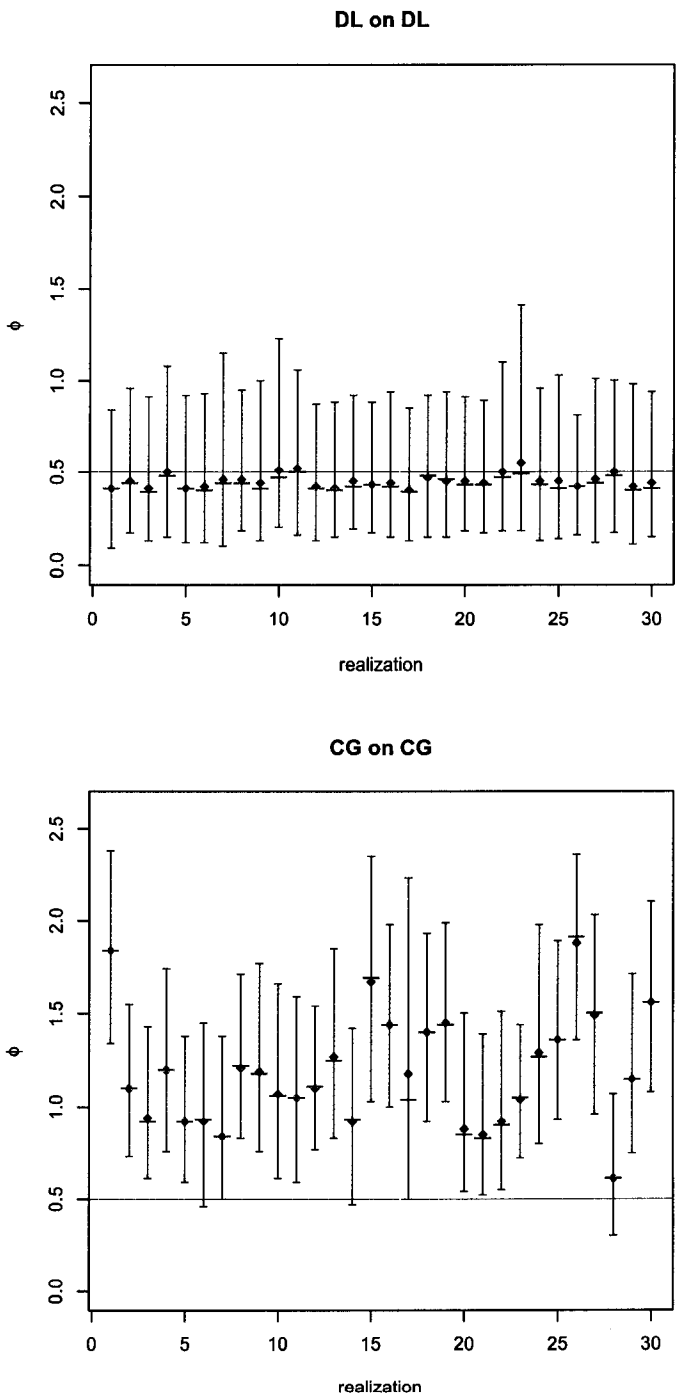

Figure 5.7: Point estimates and 95% posterior intervals for ϕ for Scenario A and $\sigma^2_{fix} = 2$. The red line shows the true value of $\phi = 0.5$. Means are denoted by the red triangles and medians by the green bars. The top plot displays the results for the Double Latent model on Double Latent data (DL on DL) and the bottom plot displays the results for the Clipped Gaussian model on Clipped Gaussian data (CG on CG).

generally larger when applying the Double Latent model than the Clipped Gaussian model, regardless of the source of the data. This is expected due to the extra source of variability in the Double Latent model. Interestingly, applying the Double Latent model to Clipped Gaussian data actually improves the significance rate for Scenarios A and B to zero non-significant intervals. Again, this is evidence that it is more difficult for the models to detect the significant relationship in the Double Latent data due to the introduction of the white noise variance on top of the relationship between the covariate and the response.

5.2.2 Comparison of models within scenarios

This section focuses on a comparison of the Double Latent and Clipped Gaussian models within each parameter scenario. It centers around an investigation of the relationship between the model used to fit the data and the model used to generate the data for both prediction and estimation. In Sections 5.2.2.1 and 5.2.2.2, we begin with a comparison of the model/data combinations within each parameter scenario, first comparing the models when fit to their respective data sets (DL on DL versus CG on CG) and then comparing the models when applied to the other's data sets (CG on DL versus DL on CG). Then, in Section 5.2.2.3, we address the question of whether applying the data-generating model to the data results in superior prediction over applying the non-data-generating model to the same data set; that is, we compare DL on DL versus CG on DL and CG on CG versus DL on CG. This question can alternatively be formed as whether the reported predictive measures based on hold-out data "choose" the true data-generating model. We envision two extreme situations related to this question: one where the measures of predictive ability are always lower for the data-generating model and another where one model out-predicts the other regardless of the data-generating mechanism. Of course, the results fall somewhere between these extremes and thus require interpretation and discussion.

Granted, the predictive measures of MPR, AMPR, and BEL were not designed for Bayesian model selection, but are relevant in this context where the modeling goal is prediction at new locations. The use of Bayesian model selection criteria such as Deviance Information Criteria (DIC) of Spiegelhalter et al. (2002) are not pursued here as their relevance to the modeling objective of prediction at new locations is not clear. We base most of the observations and conclusions given in the section on the results reported in Tables 5.6 and 5.8, where $\sigma^2_{fix} = 1$.

5.2.2.1 A comparison of predictive ability

In this section, we first compare the predictive results within each scenario using the four rows for each scenario in Table 5.6: DL on DL, CG on DL, CG on CG, and DL on CG. We compare the results for fitting each model to its respective data (DL on DL versus CG on CG), and then compare the results when applying each model to data generated under the other (CG on DL versus DL on CG). We save additional comparisons for Section 5.2.2.3. Our conclusions are formed with an emphasis on absolute mis-prediction rate (AMPR).

Scenario A, CG on CG, reveals lower values for mean and maximum MPR and AMPR along with all three BEL measurements than DL on DL. Scenario B and CG on CG revealed lower values than DL on DL for all measures except mean MPR. Likewise, Scenario C also reveals better prediction for CG on CG, with the exception of one realization that led to higher MPR and AMPR for CG on CG. Scenario D resulted in lower measures of AMPR for CG on CG, though higher values for minimum and mean MPR. This apparent conflict between the values of MPR and AMPR indicates that for many realizations CG on CG had a slightly greater number of mis-predictions, but that these wrong predictions were closer to the true value than those for the Double Latent model. Overall, we conclude that CG on CG generally results in superior prediction over the DL on DL for the scenarios considered here. However, this does have exceptions for certain data realizations and should not be

taken as a sweeping conclusion as in some cases the measures of predictive ability are rather inconclusive.

When the models are applied to the "wrong" data, the results are more ambiguous. For Scenario A, DL on CG results in lower mean MPR, mean AMPR, and maximum AMPR, suggesting that the Double Latent model is more successful on Clipped Gaussian data than the Clipped Gaussian model is on Double Latent data. However, under Scenario B, DL on CG results in lower or equal measures for mean MPR and minimum AMPR, but slightly higher values for mean and maximum AMPR. The values for mean AMPR are practically equivalent, making it difficult to conclude that one model outperforms the other. Under Scenario C, CG on DL results in nearly equivalent mean MPRs, but DL on CG results in lower mean AMPR. Thus, the mis-predictions for DL on CG are closer to the true value than those for CG on DL. In contrast, Scenario D reveals that all nine measures are lower for CG on DL versus DL on CG. Thus, combining the results for Scenarios B, C, and D we would conclude that the Clipped Gaussian model is more successful on Double Latent data than the Double Latent model is on Clipped Gaussian data. The results for Scenario A are contradictory to this. The reason for this difference is not clear.

Up to this point, we have largely glossed over the BEL results. It is clear from the results in Table 5.6 that application of the Clipped Gaussian model nearly always results in lower values for the estimated Bayesian expected loss (BEL). Thus, this value appears to be highly dependent on the model used and generally does not correspond to the relative magnitude of the MPR and AMPR values. Because of this, we place little emphasis on BEL in terms of assessing predictive ability and providing information to help "choose" the data-generating model. It may be a useful measure to assess the uncertainty in the decision process of prediction using a single model, but is not recommended for comparing the quality of predictions between models.

5.2.2.2 A comparison of estimation ability

We now compare the two models in terms of estimation within each data scenario using results from Table 5.8. We first provide conclusions regarding the comparison of the two models, each fit to its respective data set. The results are equivocal, making it difficult to provide broad conclusions. For example, for β_0 DL on DL exhibits better estimation for Scenarios A, C, and D, but for β_1 the estimation errors and their 5th and 95th percentiles are closer with CG on CG revealing slightly lower estimation errors and narrower intervals. Thus, the Clipped Gaussian model appears to be more successful at estimating the most important parameter in terms of interpretation, β_1, the exception being Scenario C. The results are similarly ambiguous for θ_2 and θ_3. The results are very unequivocal, however, for the correlation parameter, ϕ, with the Double Latent model clearly resulting in better estimates of the parameter, regardless of the data-generating model in the scenarios considered here.

As in the above section, we also compare the results obtained from fitting each model to the "wrong" data. This provides information regarding the behavior of each model when applied to data generated by a different model. This is obviously an important practical consideration since we will never encounter real data that have been exactly generated by one of these models. Thus, this may give a more realistic look at the model's performance in terms of prediction and estimation.

In terms of estimation, we devote our discussion to the most interpretable parameter, β_1. It is our hope that the models have the ability to pick up the direction and significance of an underlying relationship between the response and the covariate even if the mechanism by which the association entered the data is different from that assumed under the model. The numerical results are displayed in Table 5.8, corresponding to the "CG DL" and "DL CG" rows. For all four data scenarios, the Double Latent model applied to Clipped Gaussian data results in lower average estimation errors than that for the Clipped Gaussian model applied to Double Latent

data. In fact, in most cases the values are lower than all other model-data combinations for Scenarios A, B, and C, with that for D only being greater than CG on CG. Low estimation errors for DL on CG is also observed in general for the other parameters. Based on the magnitudes and signs of the estimation errors, DL on CG tends to overestimate the regression parameter, as compared to the typical underestimates resulting from the other model/data combinations. This means that the point estimates are generally greater than the true value of 1 and would, on average, be less likely than the other combinations to fail to detect a significant relationship between the response and the covariate. The exception is Scenario D, where the estimation error is negative, which could again be due to the discrepancy between σ^2_{fix} and σ^2_{true}, or just the larger value of σ^2_{true}.

5.2.2.3 Success at choosing the data-generating model

Obtaining MPR and AMPR after application of both models to data generated by one of the models provides an assessment of whether fitting the "better" model to the data generally leads to better prediction. An alternative way to phrase this comparison is an assessment of whether MPR and AMPR "choose" the correct, data-generating model by exhibiting lower values for the true models. Recall that both of these measures of predictive ability are based on a hold-out data set. We initially base our comparison on the average of the predictive measures over all realizations, and then also provide a brief investigation of the values for individual realizations. In general, the average values follow the results for particular realizations giving us more confidence in making conclusions based on the average values. However, there are exceptions, and these have practical importance since for real data we will obtain values for only one realization. Table 5.10 translates the numerical results of Tables 5.6 and 5.7 into a format that fits the context of the comparisons discussed in this section. We define two values to be equivalent (tied) if the difference is less than or

equal to 0.03, corresponding to a difference of approximately one mis-prediction over 30 realizations.

For Scenario A, all three summary measures of AMPR are lower for the true models, with those of MPR being more inconclusive. For the Double Latent data, the values for mean AMPR are considered equivalent, but overall the results provide evidence that AMPR may be useful in choosing the data-generating model, especially for the Clipped Gaussian data and if all three predictive measures are taken into account. As previously mentioned, all three BEL summary measures are typically lower for the Clipped Gaussian model, regardless of the data and thus we pay little attention to BEL in this discussion.

For Scenario B, all nine measures are lower for the data-generating model when both models are applied to the Clipped Gaussian data. The results for the Double Latent data are more equivocal, with lower mean MPR for the true Double Latent model, but lower AMPR for the Clipped Gaussian model. The mean AMPR values are close enough to be considered equivalent, but the combination of MPR and AMPR indicate that while there are fewer mis-predictions for DL on DL, some tend to be farther from the true value. For Scenario C data, the measures are more successful at choosing the true model for the Clipped Gaussian data than for the Double Latent data, particularly the mean and maximum AMPR.

The AMPR measure is again successful for the Clipped Gaussian data for Scenario D, though unsuccessful for the Double Latent data. MPR is likewise unsuccessful for Scenario C. The failure of the predictive measures to indicate the data-generating model for Scenario D is also true for the $\sigma^2_{fix} = 2$, and thus does not improve as σ^2_{fix} moves one unit closer to σ^2_{true}. The results for the other scenarios for $\sigma^2_{fix} = 2$ lead to the same general conclusions, as is clear from comparing the tables in Table 5.10.

While Table 5.10 and the previous discussion provide meaningful information about whether the average values of the predictive measures over all realizations

Table 5.10: This table translates the numerical results from Tables 5.6 and 5.7 into symbolic results describing for what scenario/data combinations the predictive measures are lower for when the true data-generating model is applied. In other words, the table displays the combinations for which the true model was selected based on the predictive measures. The results are based on simulation results for the hold-out set of $m = 50$ new locations for scenarios A, B, C, and D for $\sigma^2_{fix} = 1$ (upper table) and $\sigma^2_{fix} = 2$ (lower table). A $*$ indicates that the value resulting from application of the true model to the data was lower than that for applying the wrong model, and thus the measure chose the correct model. A $-$ indicates that the value resulting from application of the true model is greater than that for applying the wrong model, and thus the measure failed to choose the correct model. A T indicates a tie defined as a difference between the two values of less than or equal to 0.03, approximately corresponding to a difference of one mis-prediction over 30 realizations (bottom table).

$\sigma^2_{fix} = 1$

Scenario	Data	MPR			AMPR			BEL		
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
A	DL	T	T	T	$*$	T	$*$	$-$	$-$	$-$
	CG	T	T	$*$	$*$	$*$	$*$	T	$*$	T
B	DL	$*$	$*$	T	$*$	T	$-$	T	T	$-$
	CG	$*$	T	$*$	$*$	$*$	$*$	$*$	$*$	$*$
C	DL	$*$	T	$-$	$*$	T	T	$-$	$-$	$-$
	CG	$-$	T	$*$	$-$	$*$	$*$	$*$	$*$	T
D	DL	T	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
	CG	$-$	$-$	$*$	$*$	$*$	$*$	T	T	T

$\sigma^2_{fix} = 2$

Scenario	Data	MPR			AMPR			BEL		
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
A	DL	$*$	T	T	$*$	T	$*$	$-$	$-$	$-$
	CG	T	T	$*$	$*$	$*$	$*$	T	T	T
B	DL	T	T	$*$	T	T	$*$	$-$	$-$	$-$
	CG	$*$	T	$*$	$*$	$*$	$*$	$*$	$*$	$*$
C	DL	T	T	$-$	$*$	$-$	$-$	$-$	$-$	$-$
	CG	T	T	$*$	$*$	$*$	$*$	$*$	T	T
D	DL	T	$-$	$-$	$-$	$-$	$-$	$-$	$-$	$-$
	CG	T	T	T	$*$	$*$	$*$	T	T	T

choose the correct value, it is more informative from a practical standpoint to assess whether the summary measures are lower for the true model for individual realizations, since in practice we only possess data from one realization. In Figure 5.8, we display the results for Scenario A and $\sigma^2_{fix} = 1$. In the figure, we connect the values of MPR (or AMPR) arising from the same realization of the data. The value on the left is obtained by fitting the true, data-generating model and the value on the right is the value obtained by fitting the "wrong" model to the data. A positive slope (red line) indicates that the measure is lower for the true model, corresponding to a correct identification of the data-generating model. A negative slope (black line) indicates that the measure is higher for the true model, corresponding to an incorrect identification of the data-generating model, and a horizontal blue line clearly indicates that the measures are equal. The figure clearly shows a greater number of red lines for AMPR than MPR.

To supplement Figure 5.8, Table 5.11 reports the number of realizations for which the predictive measure is lower, equal to, and higher than that for the correct data-generating model. Toward the same end, Table 5.12 reports the rate that the predictive measures, and combinations of predictive measures, chose the correct model over 30 realizations for Scenario A and $\sigma^2_{fix} = 2$. Higher rates mean that the measures were successful for a higher percentage of the realizations. This table further reveals better success of the measures on Clipped Gaussian data than Double Latent data. It clearly demonstrates the apparent uselessness of BEL in choosing the data-generating model for Double Latent data, along with its apparent success for some realizations of the Clipped Gaussian data. Notice that when taking MPR, AMPR, *and* BEL into account, the correct model was chosen for 97% of the realizations for the Clipped Gaussian data.

5.2.2.4 Effects of fixing the spatial variance parameter

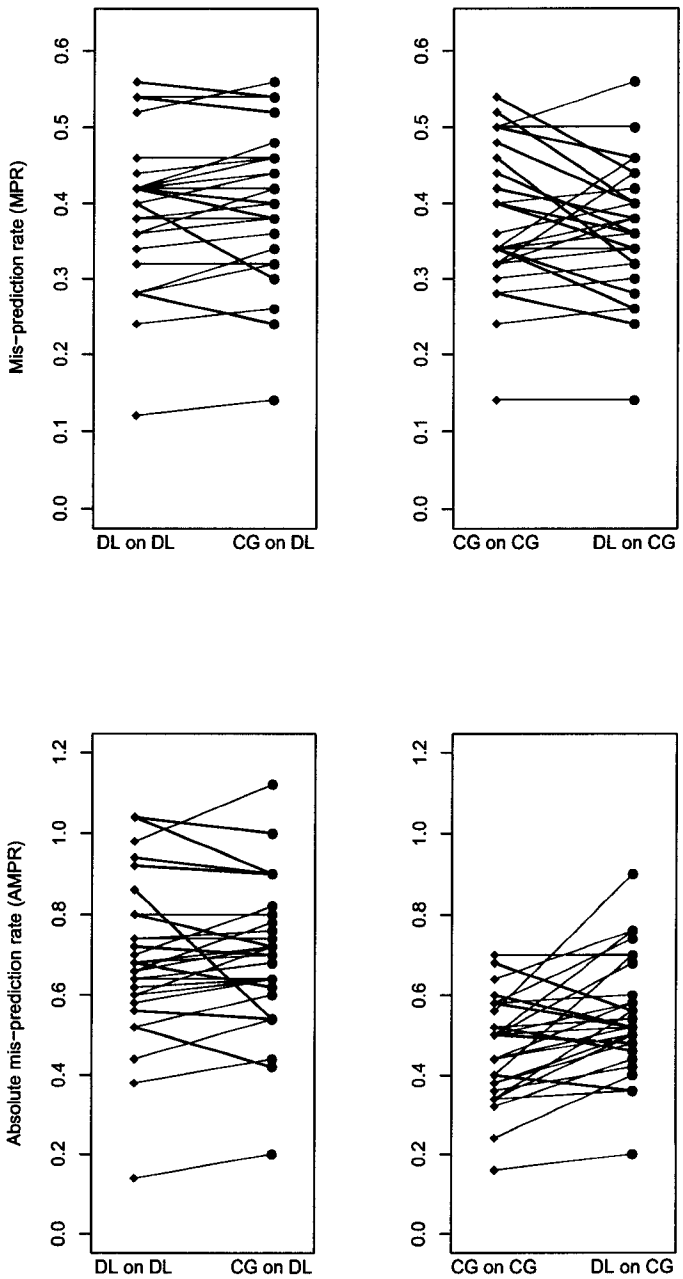

Figure 5.8: Plots connecting the values from a single realization of the data for MPR (top) and AMPR (bottom). A red line (positive slope) indicates correct identification of the true model, a blue horizontal line indicates that the measures were equal, and a black line (negative slope) indicates the measure was lower for the incorrect model and thus incorrectly identified the model. These results are under Scenario A and $\sigma^2_{fix} = 1$.

Table 5.11: The number of realizations for which the predictive measure is lower (*correct*), equal to (*equal*), and higher (*incorrect*) for the correct data-generating model versus the incorrect model (DL on DL vs. CG on DL and CG on CG vs. DL on CG). The results go with Figure 5.8 and are for Scenario A and $\sigma^2_{fix} = 1$. There are 30 total realizations.

Data		Correct	Equal	Incorrect
DL	MPR	16	7	7
	AMPR	16	4	10
CG	MPR	12	6	12
	AMPR	23	1	6

Table 5.12: Rate that each measure of predictive ability is lowest for the correct model out of 30 realizations for data created under Scenario A with $\sigma^2_{fix} = 2$. The numbers reported for MPR/AMPR are the rates that *both* MPR and AMPR are lowest for the correct model. Likewise, MPR/AMPR/BEL gives the rates that all three predictive measures are lowest for the correct model.

Data	MPR	AMPR	BEL	MPR/AMPR	MPR/AMPR/BEL
DL	0.73	0.60	0.0	0.73	0.73
CG	0.63	0.83	0.9	0.83	0.97

Overall, there is very little difference between the predictive measures for fitting the models with $\sigma^2_{fix} = 1$ versus $\sigma^2_{fix} = 2$, as revealed by a comparison of the values in Tables 5.6 and 5.7. In fact, nearly half of the values are equal after rounding to two decimal places, with many more exhibiting a difference of less than 0.02. The most notable difference is found when applying the Clipped Gaussian model to Scenario B, where most MPR values are lower for $\sigma^2_{fix} = 2$ but the mean and maximum AMPR are actually higher, meaning that fewer sites are mis-predicted, but the mis-predictions tend to be farther from the true value. This result can be seen more clearly in Figure 5.9, where the plots compare MPR and AMPR values for $\sigma^2_{fix} = 1$ and $\sigma^2_{fix} = 2$ for Scenarios A and B when the models are applied to their respective data sets.

For the purpose of choosing an appropriate value of σ^2, it appears that AMPR is most successful, as we would suspect. Most of the AMPR summary measures are lower when $\sigma^2_{fix} = \sigma^2_{true}$ under Scenarios A, B, and C, but these differences are minimal and probably not of practical importance. The fact that fixing the value of σ^2 at a value different from the true value does not appear to negatively affect the predictive ability of these models is encouraging for their practical use since for real data the parameter will surely be fixed at an incorrect value. However, this also means that the predictive measures are not successful at "choosing" the correct, data-generating value of σ^2, at least for the small difference investigated here. There is evidence, however, as described below for Scenario D, that the predictive measures can in some cases be useful in this sense if the true value and fixed values are far enough apart. For example, for Scenario D, there is a difference between the results for $\sigma^2_{fix} = 1$ and $\sigma^2_{fix} = 2$, but less so between $\sigma^2_{fix} = 2$ and $\sigma^2_{fix} = \sigma^2_{true} = 3$. In general, the results are more equivocal for the Double Latent data than the Clipped Gaussian data.

Scenario D deserves attention due to its true value of $\sigma^2_{true} = 3$ versus $\sigma^2_{true} = 1$ for the other scenarios. Thus, setting $\sigma^2_{fix} = 2$ is closer to the truth than setting

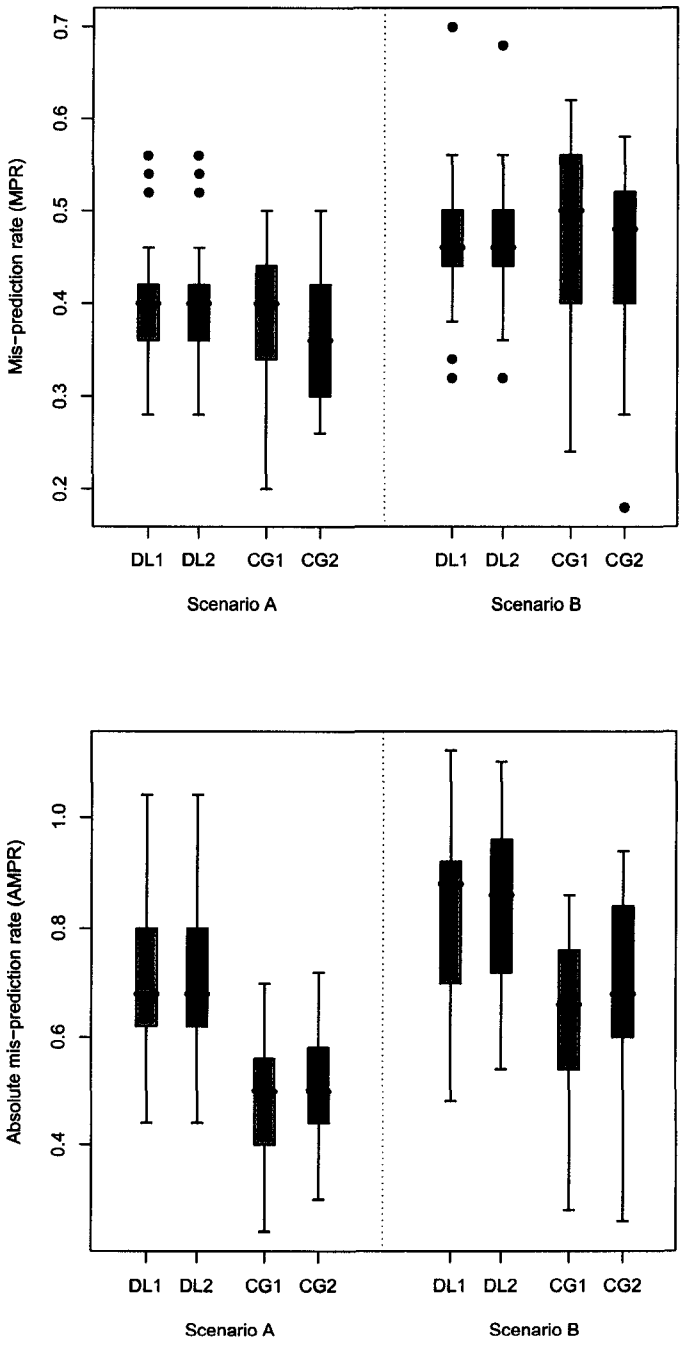

Figure 5.9: A comparison of the mis-prediction rate (MPR) and the absolute mis-prediction rate (AMPR) values for $\sigma^2_{fix} = 1$ (green) and $\sigma^2_{fix} = 2$ (blue) for DL on DL and CG and CG under Scenarios A and B. Recall that the data were generated using $\sigma^2_{true} = 1$.

$\sigma^2_{fix} = 1$, in contrast to the other scenarios. The summary measures of AMPR for $\sigma^2_{fix} = 2$ are lower than or equal to those for $\sigma^2_{fix} = 1$ for all 4 model/data comparisons for Scenario D data. We also compared these results to those obtained for $\sigma^2_{fix} = 3$ for the Double Latent model on Double Latent data and do not find a similar decrease in AMPR, with resulting AMPR values equal to or greater than those obtained for $\sigma^2_{fix} = 2$. This suggests there may be a range around the true value for which we can fix the parameter without appreciably impacting the predictions. Based on this observation, we suggest initially fitting the models with different values of σ^2_{fix} greater than one unit apart.

We also investigate the effect of fixing the magnitude of σ^2_{fix} on the estimation of the other parameters by comparing the results in Tables 5.8 and 5.9, focusing here on Scenarios A and B. When combining information contained in the average estimation error and the 5th and 95th quantiles of those errors, it is clear that no appreciable differences in estimation result from using $\sigma^2_{fix} = 1$ versus $\sigma^2_{fix} = 2$. However, there are some interesting and mostly unexpected results worth mentioning. For one, CG on CG under Scenario A generally revealed lower average estimation errors for $\sigma^2_{fix} = 2$ than $\sigma^2_{fix} = 1$, though the intervals are generally narrower for $\sigma^2_{fix} = 1$. The same is true in terms of average estimation error for DL on DL under Scenario A, with the exception of β_0. Scenario B revealed a nearly opposite result for both model/data combinations. As previously mentioned, we are mainly interested in the estimation of β_1 to describe the direction and magnitude of the relationship between the response and the covariate. The estimation error for β_1 is actually lower for $\sigma^2_{fix} = 2$ for all model/data comparisons in Scenario A, and for estimation on the DL data in Scenario B. Additionally, the width of the 95% posterior intervals are generally wider when $\sigma^2_{fix} = 2$ for nearly all model/data/scenario combinations.

For Clipped Gaussian model on Double Latent data, we observe better estimation of β_1 for $\sigma^2_{fix} = 2$, which follows our hypothesis that the results for applying the

Clipped Gaussian model to data generated under the Double Latent model may be better for $\sigma^2_{fix} = 2$ than $\sigma^2_{fix} = 1$ because the extra variability incorporated into the Double Latent data translates into $\sigma^2_{fix} = 2$ being closer to the true amount of variability than $\sigma^2_{fix} = 1$. For example, for Scenario A, the average estimation error for $\sigma^2_{fix} = 2$ is -0.36 with a (5th, 95th) interval of $(-0.83, 0.09)$, compared to -0.52 and $(-0.88, -0.20)$ for $\sigma^2_{fix} = 1$. This translates into estimates that on average are not only closer to the true value, but that will be more likely to capture the significance of a relationship between the response and the covariate. The results for Scenario B are in the same direction though the difference is probably negligible.

5.2.3 A look at specific realizations

To investigate characteristics of data sets that lead to successes and failures in terms of prediction, we take a closer look at the individual realizations for which the two models were most successful in terms of prediction, as well as those for which the models displayed the poorer predictive performance. For the Double Latent model, the lowest minimum MPR and AMPR were observed on a realization from Scenario A, at 0.12 and 0.14 respectively. This corresponds to correct prediction of 44 out of the 50 hold-out sites, with only one prediction falling more than one category away from the true value. This realization had the largest number, 43 out of 50, of hold-out observations in category 1, and the model predicted that 48 of the 50 belonged to category 1. The true category composition is $(43, 4, 2, 1)$ and the predicted table is $(48, 0, 0, 2)$. An image plot of the true values of this realization, along with those from realizations resulting in the best prediction for Scenario B, are shown in Figure 5.10.

In terms of estimation, the Scenario A realization had the second largest average estimation error for β_0. The average estimation error for the parameter over all realizations was 0.045 and the estimation error for this particular realization was

-0.57. Estimation error was also slightly larger than average for θ_2 and θ_3. However, for the important regression parameter, β_1, the estimation error of -0.12 is smaller than the average of -0.21.

The lowest minimum MPR and AMPR for the Clipped Gaussian model are observed on the Clipped Gaussian version of the same realization from Scenario A. The values are 0.14 and 0.16, respectively, meaning that it correctly predicted 43 out of the 50 hold-out sites, with only one prediction falling more than one category from the true value. The true category composition is $(42, 5, 1, 2)$ and the predicted is $(43, 5, 2, 0)$. The predicted table looks closer to the true table for the Clipped Gaussian than for the Double Latent even though the MPR and AMPR are slightly lower for the Double Latent results. An image plot of the true values for the realization are shown in the top two plots of Figure 5.10. In terms of estimation, the estimates for the Clipped Gaussian realization reveal less error than those from the Double Latent model. For both β_0 and β_1, the estimation errors are below average, but the error is of the opposite sign compared to the average.

We also look more closely at realizations resulting in relatively poor predictions for both models compared to other realizations. Figure 5.11 displays image plots of four such realizations, two from Scenario A and two from Scenario B. Notice from a comparison of the plots in Figures 5.10 with those in Figure 5.11 that the realizations resulting in better predictions are less patchy (more smooth) than those resulting in poorer prediction. De Oliveira (2000) also came to this conclusion using his two realizations of artificial data that differed by their degree of smoothness. This is not a surprising result, as prediction in less patchy environments is a simpler task for nearly any predictive method. The MPR and AMPR corresponding to the Double Latent Scenario A realizations are 0.54 and 1.04, while those for the Clipped Gaussian model are 0.44 and 0.56. This corresponds to correct prediction of only 23 out of the 50 sites for the Double Latent and 28 for the Clipped Gaussian.

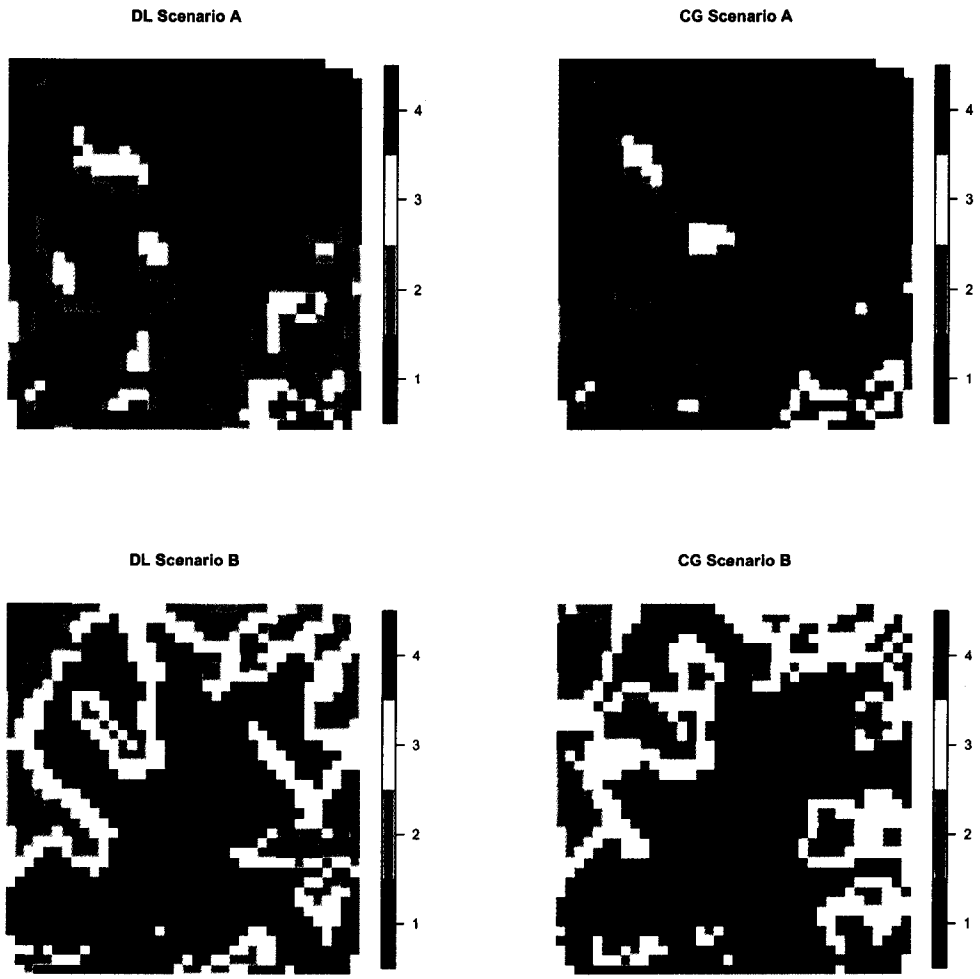

Figure 5.10: Image plots of the true values for Scenario A realization 29 (top) and Scenario B realization 11 (bottom) generated under the Double Latent model (left) and the Clipped Gaussian model (right). These represent realizations that lead to successful prediction.

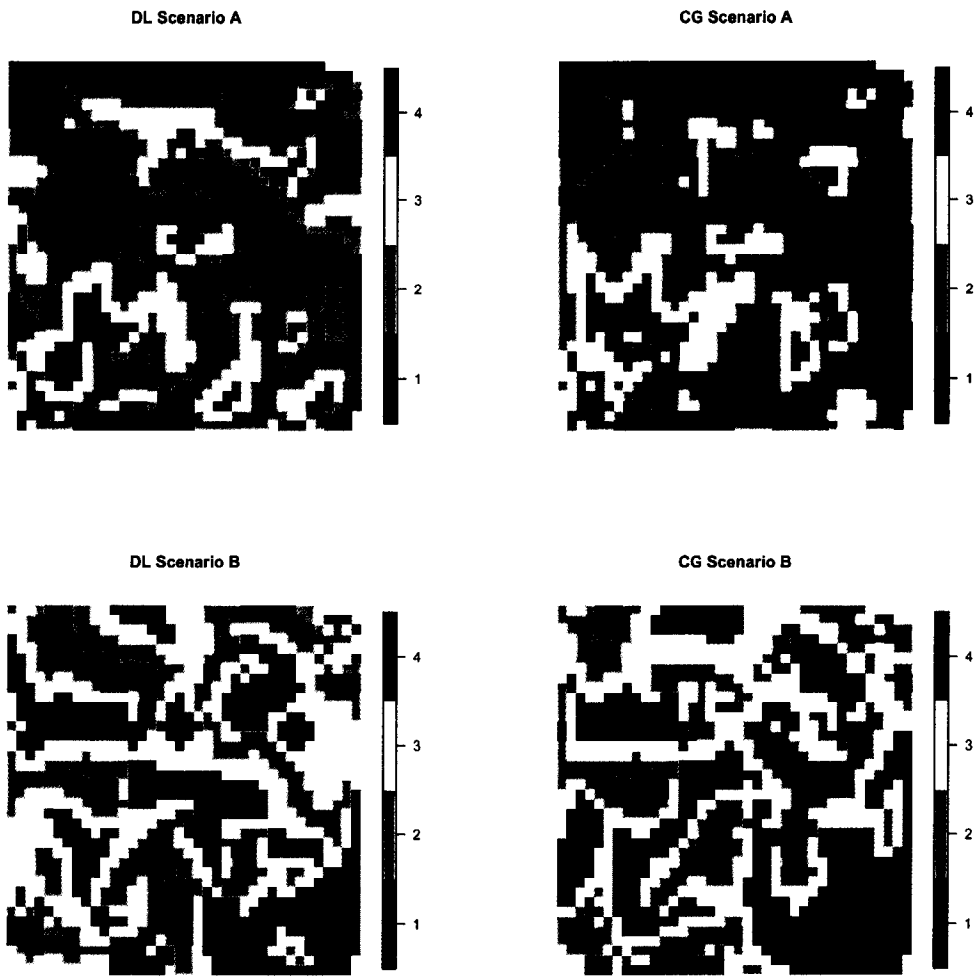

Figure 5.11: Image plots of the true values for Scenario A realization 19 for Double Latent (top left) and Clipped Gaussian (top right) data, and for Scenario B realization 5 for Double Latent (bottom left) and Clipped Gaussian (bottom left) data. These are realizations that result in relatively poor predictions compared to other realizations.

In terms of estimation for the realizations representing poor predictions, the Double Latent model on the Double Latent data results in relatively poor estimates of both β_0 and β_1, with estimation errors of 0.49 and -0.63, respectively. Compared the the "good" prediction realization, the estimation error is lower for β_0, but higher for β_1 where the "good" realization estimation error is only -0.12. For the Clipped Gaussian model, the results are similar, with the estimation error for β_0 equal to 0.73 and that for β_1 being 0.34, much higher than the average error of -0.13 and of the opposite sign. This sign change was also observed for the "good" realization, but the error was much smaller at 0.09.

Therefore, while in general, estimation success and predictive success do not have to occur together, more accurate estimation of β_1 does appear to be related to success in prediction, as we would expect. However, this conclusion is not true to the degree that we observe the *best* estimates of β_1 for the realizations resulting in the best prediction. To further explore this relationship between prediction and the estimation of the regression parameter, we create plots displaying realizations resulting in both particularly good and poor estimation of the regression parameter, β_1. These plots are displayed in Figure 5.12 for each model applied to its respective data set. The figure provides a general idea of characteristics of the categorical composition and spatial layout leading to successful estimation of the parameter. For example, notice that the realizations resulting in better estimation of the β_1 show a more distinct pattern of change from the northwest corner to the southeast corner which is indicative of true spatial layout of $\mathbf{X}\boldsymbol{\beta}$. The model parameters of Scenario D resulted in increased error added on top of the spatially varying covariate, thus washing out this underlying spatial layout and possibly making it more difficult to estimate the regression parameter, β_1, as expected.

5.3 Discussion and conclusions

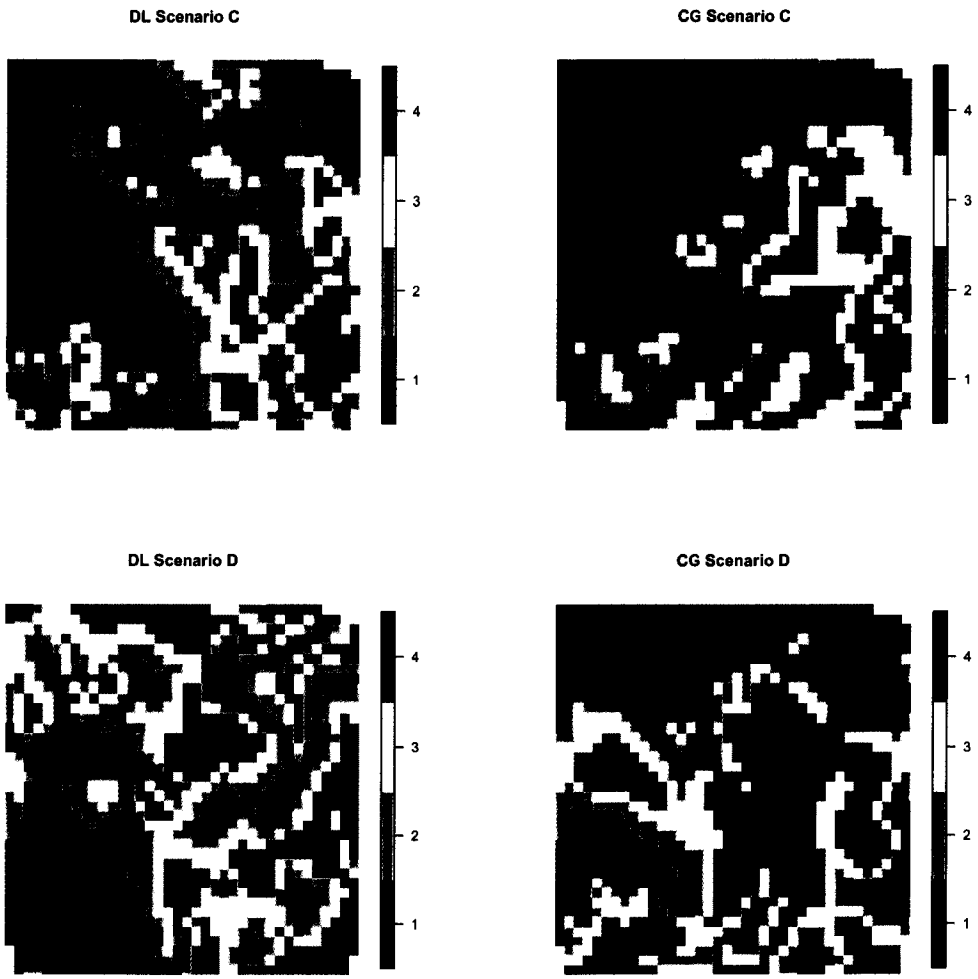

Figure 5.12: Image plots of the true values displaying realizations that result in *good* prediction of β_1 for the Double Latent model on Double Latent data (DL on DL) (upper left) and the Clipped Gaussian model on Clipped Gaussian data (CG on CG) (upper right), and those resulting in *bad* prediction for DL on DL (lower left) and CG on CG (lower right). The data are respectively from scenario(realization) C(8), C(19), D(17), and D(13).

Overall, the simulation study reveals that the models are effective in terms of prediction over a wide variety of data realizations. In the process of completing the simulation study and other preliminary investigations, we fit the models to nearly 150 different sets of data. The models correctly predicted the values of on average 65% of the hold-out sites, and as many as 88% for specific realizations. The predictions were rarely more than one category away from the true value. The models were less successful in terms of parameter estimation when applied to some data realizations, but this is less of a concern given that our primary goal is prediction at new locations. In this section, we highlight some of the most important or interesting observations gleaned from the results of this simulation study.

First, in general it does not appear that success at estimation and success at prediction must necessarily occur together. In fact, the data generated under Scenario A demonstrate this by revealing superior prediction coupled with poorer estimation of most of the parameters compared to the other scenarios. Estimation of the regression parameter, β_1, describing the relationship between the response and covariate, appears to be the parameter whose accurate estimation is most related to good prediction. This is not a surprising observation since better estimation of β_1 leads to better use of the information available in the covariate.

As far as a comparison of the two models in terms of prediction, the models appear to respond to the different scenarios and data sets in a similar way leading to the same general conclusions regarding their use on data generated under the four different scenarios. Thus, it is difficult to provide a recommendation regarding the type of data that each model may be more successful on. However, it is possible that such a conclusion could be reached with the help of additional simulation studies with differing category compositions. One possibly important characteristic of the Double Latent model that may deserve further attention is its tendency to predict values for the end categories, while more rarely producing predictions of the middle

categories. The Clipped Gaussian model, on the other hand, more readily makes predictions in all categories. Thus, it would be an interesting future comparison to fit the models to data containing a larger proportion of their observations in the middle categories. It is also of further interest to investigate how this behavior might change as the number of categories increases.

As previously mentioned, the Double Latent model capitalizes on the conditional independence which leads to the advantage of being computationally faster. The Double Latent model typically runs in approximately one third the time of the Clipped Gaussian model. The comparison of the marginal distributions of the clipped continuous latent variables reveals that, marginally, the difference between the two models is the common variance, with the Double Latent having increased variance due to its extra white noise component. We thus expected that the results for applying the Clipped Gaussian model to data generated under the Double Latent model may be better for $\sigma^2_{fix} = 2$ than $\sigma^2_{fix} = 1$ since some of the extra variability in the data might be accounted for through the larger fixed parameter value. In fact, for this case, the diagonal elements of the covariance matrix would be equal to that of the data-generating model. For prediction, this hypothesis was not found to hold true for any of the data scenarios. However, the estimation results *did* provide evidence toward this conclusion, revealing smaller estimation errors and often variability for $\sigma^2_{fix} = 2$ for most of the parameters and all four data scenarios.

As far as usefulness of the predictive measures to select the more appropriate model, we feel comfortable recommending the use of MPR and AMPR, particularly together, to help choose the more appropriate model. We also feel confident in recommending against the use of estimated Bayesian expected loss in choosing the data-generating model since for data generated under the Double Latent model it failed to choose the correct model nearly 100% of the time. With real data, we have little chance of knowing if it was generated by a mechanism similar enough to that of

the Double Latent model that the measure will likely fail. Also, if the primary goal is prediction, then we are less concerned with finding a model that represents the data-generating mechanism, than finding a model possessing good predictive ability.

In conclusion, the simulation study presented in this chapter provided a necessary initial investigation regarding prediction and estimation using the proposed models. It allowed us to make conclusions regarding particular successes of the models and situations that warrant caution and/or further investigation. It also provided a setting to examine the measures of predictive ability we chose to employ. Additionally, it begins to fill a void in the literature regarding simulation studies for latent variable methods applied to ordered categorical data. The results from the study point to future directions of investigation to better understand the behavior of these models, as discussed in Chapter 7.

Chapter 6

PREDICTING STREAM HEALTH IN MONTGOMERY COUNTY, MARYLAND

The public's concern for maintaining or improving the quality of our nation's waters ultimately led to the Clean Water Act of 1977. The goal of the act is "restoring and maintaining the chemical, physical, and biological integrity of the nation's waters" (U.S. Environmental Protection Agency (EPA), 2003). Toward this end, the streams and rivers must be monitored to both assess their current state of health, and also to provide reference values to compare with future observation. It is primarily the responsibility of the individual states to conduct monitoring programs. In the initial years after the enactment of the Clean Water Act, work focused on the chemical integrity, but has more recently also emphasized biological integrity (U.S. Environmental Protection Agency (EPA), 2003). We apply our proposed models to biological stream data collected in Montgomery County, Maryland by Montgomery County's Department of Environmental Protection and as part of the Maryland Biological Stream Survey (MBSS) for the Maryland Department of Natural Resources. A goal of the monitoring program is to rate stream areas according to their overall health as compared to reference streams.

One common measure of stream health is the *Index of Biotic Integrity*, or IBI. Biological (or biotic) integrity is defined by the U.S. Environmental Protection Agency (EPA) as "the condition of the biological communities (usually benthic macroinvertebrate and/or fish) of a waterbody based on a comparison to a reference that is a relatively undisturbed system and represents the best quality to be expected for

the ecoregion" (U.S. Environmental Protection Agency (EPA), 2007). They define IBI as a "combination of measures, or metrics, that describe community structure, function, and pollution sensitivity and are used to assess the health of an aquatic system," and view the index as providing a picture of overall ecological stream health. Separate IBIs are constructed for benthic macroinvertebrates and fish and then categorized using a set of cut-point values based on the reference streams to create the IBI *narrative* describing the overall health of a site as *poor*, *fair*, *good*, or *excellent*. A rating of *poor* indicates that the stream (or stream segment) is considered unhealthy compared to the reference streams. A rating of *good* implies that the stream is comparable to the highest quality reference streams in the region and a rating of *fair* implies that it is comparable to the remainder of the reference streams. *Excellent* rated streams are considered in better health than the highest quality reference streams in the region (U.S. Environmental Protection Agency (EPA), 2007). We use the four-category benthic macroinvertebrate IBI (bIBI) narrative for our analysis. Benthic macroinvertebrates are aquatic animals without backbones that dwell on the bottom of aquatic environments and are larger than 1/2 millimeter. Examples include crayfish and many types of aquatic insect larvae (U.S. Environmental Protection Agency (EPA), 2007).

The goal of our analysis is to predict the categorical benthic IBI for new, unsampled locations within the study region. The streams of Montgomery County were densely sampled over this relatively small region, providing an attractive data set for which to apply spatial models (Figure 6.1). As shown on the map, stations outside of the county were sampled in some cases when the streams eventually enter Montgomery County.

6.1 Description of the data

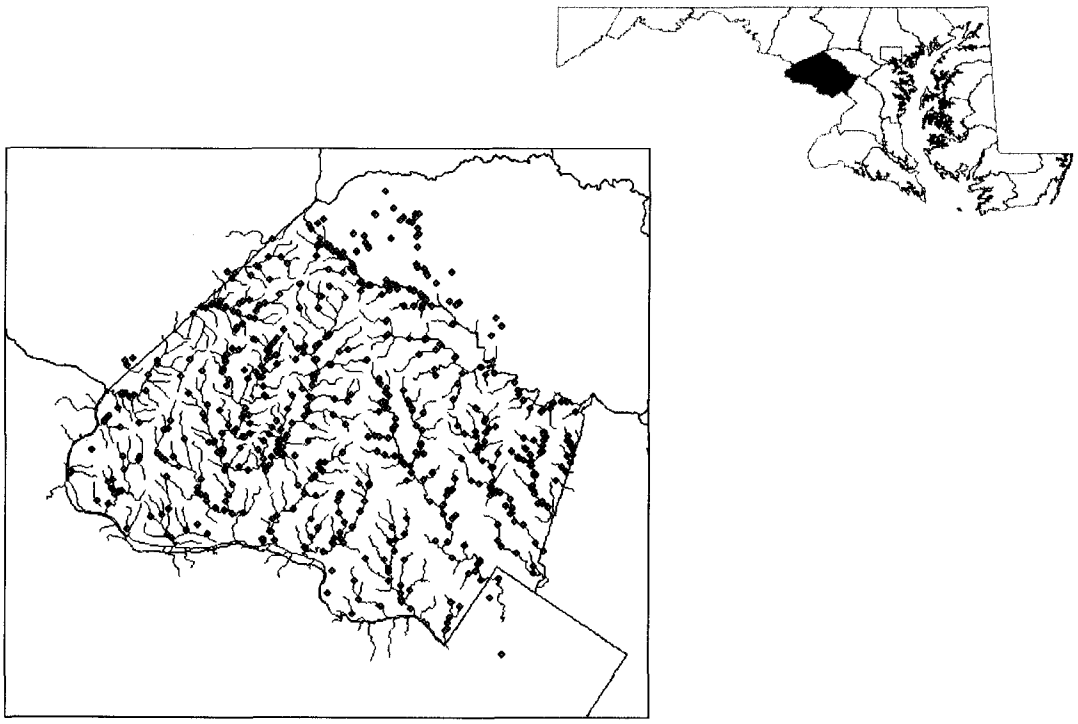

Figure 6.1: The map shows the location of Montgomery County within the state of Maryland and the enlargement identifies the locations of the stations for the data set.

6.1.1 Exploratory data analysis

We utilized data from Montgomery County, Maryland collected at $N = 299$ stations over 8 years, from 1996 to 2003, with the majority of observations coming from 1996 to 1999. Montgomery County covers an area of 497 square miles (1287 square kilometers), had a population of 855,000 in the year 2000, and lies adjacent to Washington D.C. Examination of categorical benthic IBI for single locations over time reveals relatively little variation in the response as the result of time. Therefore, we maintain that for the purpose of this analysis it is reasonable to assume that the response is constant over time. We instead focus on the greater source of variability in the data, that due to spatial location. The spatial distribution of the benthic IBI narratives for this sample is shown in Figure 6.2, along with a map of population density of Montgomery County in year 2000 from U.S. Census Bureau data. Note that benthic IBI tends to be lower in areas of higher population density. This is to be expected if benthic IBI reflects overall stream health and if higher population density is associated with lower water quality or stream health. The image plot of the IBI narrative was created using linear interpolation between the sampled sites.

The one covariate chosen for use in this analysis is *conductivity*, or *specific conductance*. The following information on specific conductance can be found in Murphy (2007). Conductivity is a measure of how well water conducts electricity, measured using a one centimeter cube of aqueous solution, and is used as an indirect measure of the total dissolved solids in the water. Conductivity increases as the amount and mobility of ions in the water increase, thus providing an indirect measure of the presence of dissolved solids that are linked to water pollution. Examples of such solids are chloride, nitrate, sulfate, phosphate, and iron. Conductance is the reciprocal of resistance, and the unit of measure is typically a *mho*, derived from spelling *ohm*, the unit of measure of resistance, backwards. It is usually reported in micro mho's per centimeter ($\mu mho/cm$). Conductance changes with temperature and

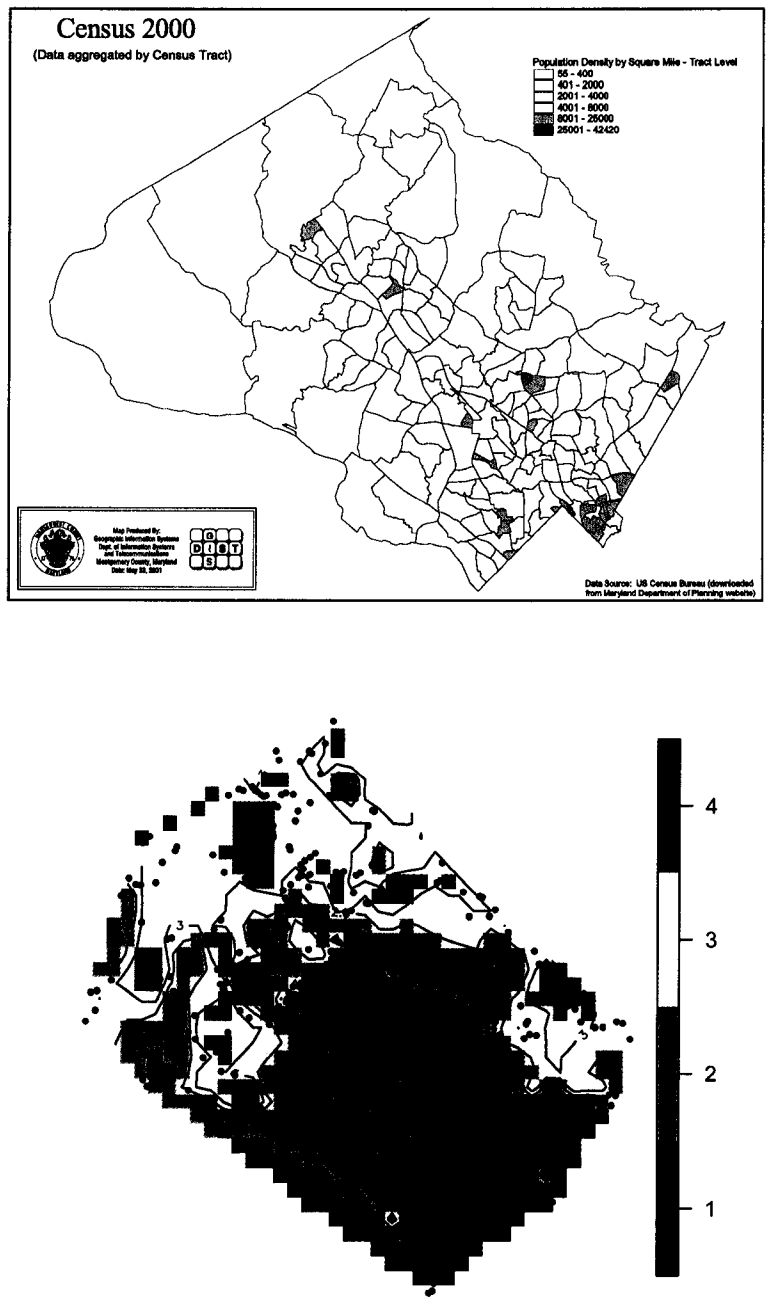

Figure 6.2: A comparison of the plot of the population density from Census 2000 data to that of benthic IBI narrative for Montgomery County, Maryland. The integer value labels for the benthic IBI scores correspond to 1 = *Poor*, 2 = *Fair*, 3 = *Good*, and 4 = *Excellent*.

thus measurements are typically converted to those representing room temperature (25 degrees Celsius). Conductivity is also affected by the rock and soil type of the stream and so is also a surrogate measure for the geology and soil type at or upstream of the station. It is often an indicator of agricultural runoff containing phosphate and nitrate from fertilizers and leached salts from soil, as well as of road runoff that introduces chemicals leaked from automobiles and from de-icing. The EPA does not issue regulations based on specific conductance, but they do give standards for drinking water for total dissolved solids (TDS) of less than $500mg/L$. It is common to attempt to establish a linear relationship between conductance and TDS such that TDS can be estimated from conductivity, typically multiplying by factors in the range of $0.55 - 0.75$. Thus, for drinking water, the approximate standard range for specific conductance is $275 - 375\mu mho/cm$.

Conductivity was chosen as a covariate for our analysis based on its relationship with the benthic IBI and because it is easily obtained and generally useful for detecting contaminants in water (Murphy, 2007). As shown in Figure 6.3, higher values of conductivity appear to be associated with a lower benthic IBI scores. The image plot of conductivity over the county is shown in Figure 6.4.

We also explored the spatial correlation in the data. The spatially dense sampling led to a relatively large number of small inter-site distances, as shown in the histogram of inter-site distances (Figure 6.5), which is advantageous for estimating spatial parameters, as discussed in Chapter 3. Universal Transverse Mercator (UTM) coordinates are used for the analysis and referred to as *eastings* and *northings*. We scale the original coordinates to obtain distances in a numerically convenient range, resulting in a maximum distance between stations on our scale of 4.92. This corresponds to a distance of 49.2 kilometers, or 30.6 miles.

An empirical semi-variogram for the benthic IBI scores is shown in Figure 6.6. Notice that before accounting for any covariates, the spatial correlation appears

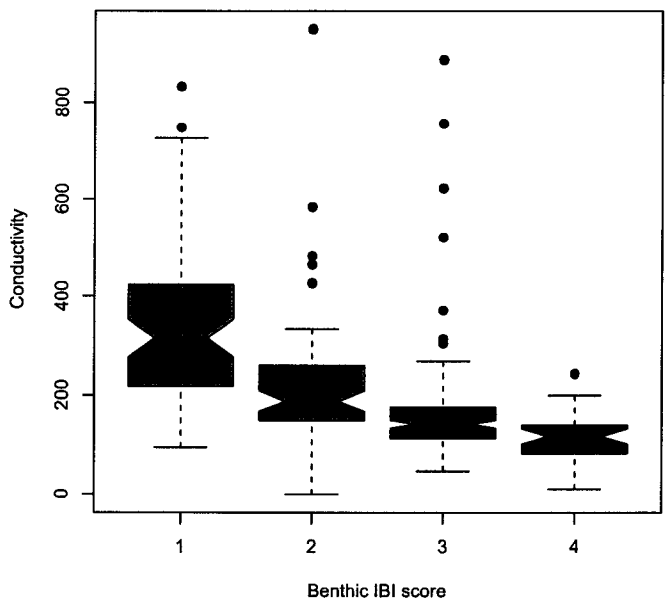

Figure 6.3: This plot displays the observed relationship between benthic IBI score and the value of conductivity (or specific conductance) measured at the station. The integer value labels for the IBI scores correspond to 1 = *Poor*, 2 = *Fair*, 3 = *Good*, and 4 = *Excellent*.

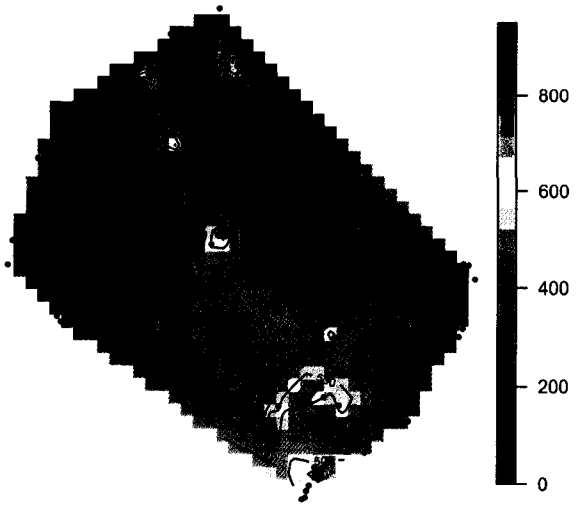

Figure 6.4: This image plot displays the linearly interpolated values of the conductivity (or specific conductance) measured at the sites in Montgomery County for the 299 stations used in this analysis.

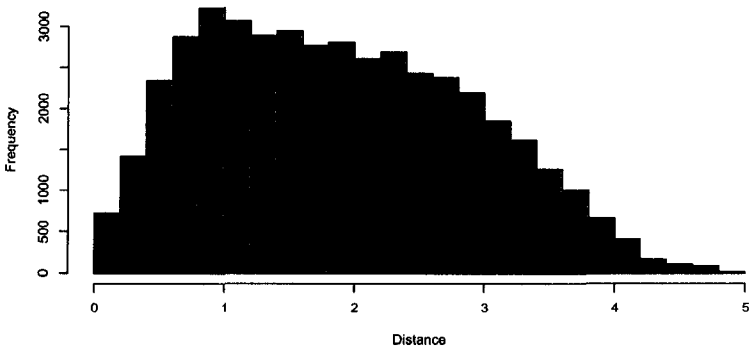

Figure 6.5: A histogram of the inter-site distances for the $N = 299$ stations used for this analysis.

to exhibit a large range relative to the maximum inter-site distance in the data. We use the term *range* to mean the distance at which the correlation between two points is effectively zero, approximately corresponding to the distance when the semi-variogram nears the estimated sill or asymptote. See Chapter 1, Figure 1.1 for a review of the terminology associated with a semi-variogram.

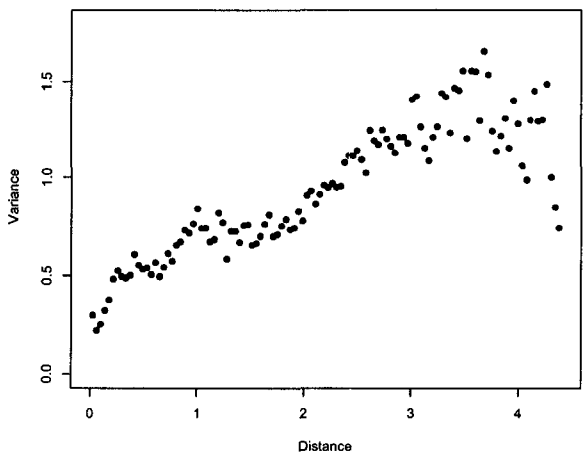

Figure 6.6: An empirical semi-variogram displaying the relationship between covariance of benthic IBI values and inter-site distance for the $N = 299$ stations.

6.2 Applying the models

To evaluate the predictive ability of our models, we select a hold-out data set of $m = 100$, leaving a sample size of $n = 199$ from which to estimate the model parameters. Figure 6.7 displays the locations of the stations and the sites that were used as a hold-out data set. Figure 6.8 displays the image plot for the remaining sites after removing the hold-out data set, and thus the sites used for fitting the model.

6.2.1 Prior specification

As in the simulation study described in Chapter 5, we fix the value of σ^2 and adopt a one parameter correlation function. We fit the model using both $\sigma^2 = 1$

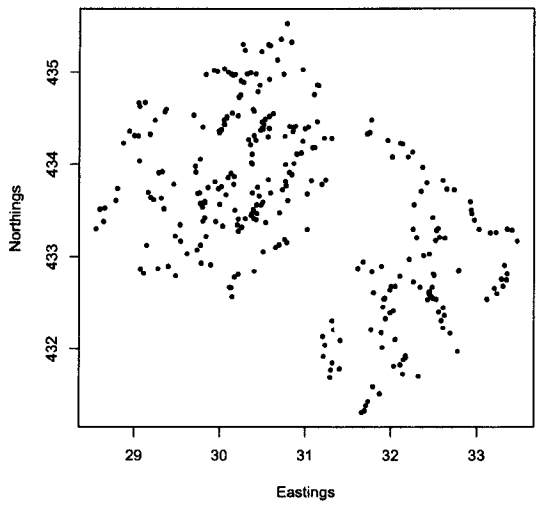

Figure 6.7: This plot displays the spatial layout of the 299 stations used in this analysis. The sites shown in red represent the hold-out data set of $m = 100$ used to evaluate the predictive ability of the models.

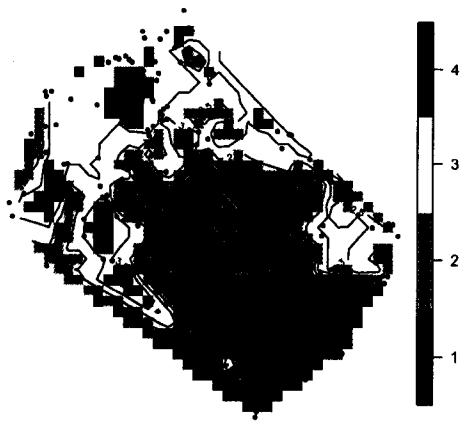

Figure 6.8: Image map of the values of the benthic IBI over the Montgomery County for the 199 training stations used in this analysis. The integer value labels correspond to 1 = *Poor*, 2 = *Fair*, 3 = *Good*, and 4 = *Excellent*. Linear interpolation was used to create the image plot.

and $\sigma^2 = 2$, and compare the results. As expected based on the results from the simulation study, we do not find a noticeable difference between the results and thus report only the results for $\sigma^2 = 1$.

Based on our experience fitting the proposed models to simulated data, we use the exponential correlation function, given by $\rho(d) = \exp(-d\phi)$, and choose a proper, though non-informative, prior for the scale parameter specified simply by our knowledge of the maximum inter-site distance and likely values that arise in practice given this distance. Note that the quantity $1/\phi$ is often referred to as the *range* parameter, and we parameterize the exponential correlation function in terms of the reciprocal of the range parameter. Our goal is that the prior serve to keep the chain in the ball-park of the true value while still allowing the data to drive the posterior distribution of the parameter. For the Montgomery County data, this maximum distance is 4.92. We employ the strategy described in Section 3.3.1, relying on a proper Gamma distribution with mean and variance fixed to define the prior (Banerjee et al., 2004; Schmidt et al., 2007). We assume that some spatial autocorrelation does exist and that the range should not practically be less than 1/20th of the maximum inter-site distance, or approximately 0.25. Thus, it seems reasonable to place most of the prior density for the *range* parameter over the interval $(0.25, 5)$, corresponding to $\phi \in (0.2, 4)$.

We fit the models using two different priors based on specifying different values for the variance for the Gamma distribution. The first, *Prior 0.6*, specifies the variance of the Gamma prior distribution to be 0.6, resulting in a seemingly appropriate and non-informative prior distribution adequately covering the interval of plausible values for ϕ. The probability density function of this Gamma distribution, with shape parameter equal to 2.48 and rate parameter equal to 2.03, is plotted in Figure 6.9. Second, we fix the Gamma distribution variance to be 0.05, referred to as *Prior 0.05*. A comparison of the results assesses sensitivity to the prior distribution. *Prior*

0.05 results in a more informative and seemingly less appropriate prior distribution, placing most of its density over the range of $(0.5, 2.0)$, giving less weight to both small and large values of ϕ.

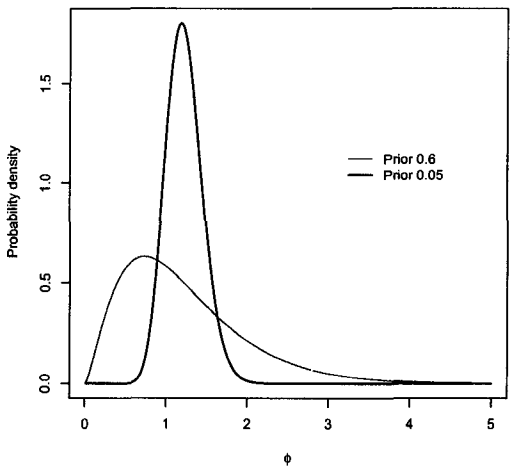

Figure 6.9: The probability density functions for the two prior distributions specified for the scale parameter, ϕ, of the exponential correlation function. The first, shown in red, is termed *Prior 0.05* and corresponds to a Gamma$(2.48, 2.03)$. The second, shown in black, is termed the *Prior 0.60* and corresponds to a Gamma$(29.7, 24.4)$.

For the remaining parameters, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we also use proper, though non-informative, values that were found to work well during our simulation study. These are described in detail in Sections 3.3.2 and 3.3.3.

6.2.2 A comparison with realizations generated under the models

In this section, we provide a visual comparison of data generated by the assumed models with the real data. This provides a strategy for investigating whether the models provide adequate representations of reality and suffice to be data generating mechanisms. The plots in Figure 6.10 display data from the simulation study in Chapter 5 that are somewhat similar to the real data. These artificial data sets were not created with the goal of matching the true data, but serve to show that even

Table 6.1: Tabled counts of the four categories from the real data and realizations from artificial data created for the simulation study in Chapter 5. The labels for the scenario are given as Scenario(realization number).

Model	Scenario	Counts	Percentages
-	bIBI data	(71, 67, 128, 33)	(0.24, 0.22, 0.43, 0.11)
CG	A(9)	(79, 50, 30, 16)	(0.45, 0.28, 0.17, 0.09)
DL	A(15)	(89, 34, 30 ,22)	(0.51, 0.19, 0.17, 0.13)
CG	B(5)	(18, 37, 49, 70)	(0.10, 0.21, 0.28, 0.40)
DL	B(4)	(30, 23, 43, 79)	(0.17, 0.13, 0.24, 0.45)

with obviously different parameter values, the models seem to produce adequate representations of the observed data. We denote the number of observations in category k by n_k and give the tabled count of the number of observations per category for the four category case as (n_1, n_2, n_3, n_4). Analogously, we give the percentage of the total observations in each category by (p_1, p_2, p_3, p_4). Table 6.1 displays the tabled count of true values for the real data compared with those realizations shown in Figure 6.10.

6.3 Results

In this section, we compare the results of the Double Latent and Clipped Gaussian models in terms of estimation and prediction. We also compare the results to those obtained using a non-spatial cumulative probit model. The non-spatial model was fit using maximum likelihood through the *polr* function in the R statistical package (R Development Core Team, 2007).

6.3.1 Prediction

The primary goal of this analysis is the prediction of the categorical benthic IBI, as an indicator of stream health, for unobserved locations in Montgomery County. Thus, we first evaluate the predictive ability of the model for our hold-out data set of $m = 100$ locations, shown in red in Figure 6.7. The results are based on one

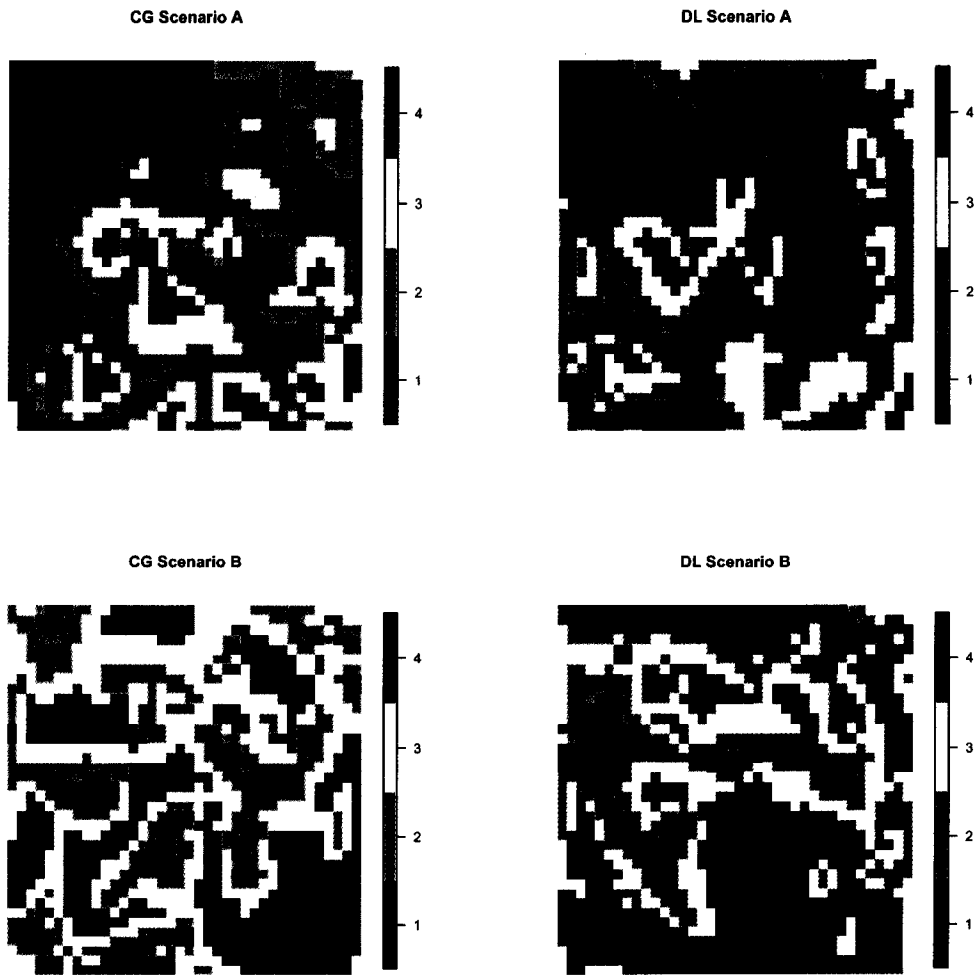

Figure 6.10: Artificial data created for the Chapter 5 simulation study, including Clipped Gaussian, Scenario A (realization 9) (upper left); Double Latent, Scenario A (realization 15) (upper right); Clipped Gaussian, Scenario B (realization 5) (lower left); and Double Latent, Scenario B (realization 4) (lower right). See Table 6.1 for the respective category counts and associated percentages.

Table 6.2: Evaluation of predictive ability for the three models, Clipped Gaussian (CG), Double Latent (DL), and Non-spatial (NS) using mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and the estimated Bayesian expected loss (BEL) using *Prior 0.6*.

Model	σ^2	MPR	AMPR	BEL
DL	1	0.51	0.66	0.56
CG	1	0.39	0.42	0.32
DL	2	0.47	0.58	0.59
CG	2	0.40	0.44	0.33
NS	-	0.55	0.74	0.49

run of the MCMC algorithm for 10,000 iterations using a burn-in period of 2000 iterations. Table 6.2 reports the prediction results for each model using $\sigma^2_{fix} = 1$ and $\sigma^2_{fix} = 2$ using the *Prior 0.6* for ϕ. The results for *Prior 0.05* are given in Table 6.3. Notice that there is very little difference in predictive performance between the two priors. In fact, the MPR values are identical and the AMPR and BEL for the Clipped Gaussian are off by only 0.01, meaning that one mis-prediction using *Prior 0.05* was one category farther from the true value than that for *Prior 0.6*. Similarly small differences were observed for use of other priors.

Recall that the lower values of MPR, AMPR, and BEL are indicative of better predictions. Application of the Clipped Gaussian model results in superior predictive ability in terms of MPR, AMPR, and BEL for this data set, as shown in Tables 6.2 and 6.3. The values are lower for all three measures, with the Clipped Gaussian model correctly predicting 61 out of the 100 hold-out sites, versus only 49 correct predictions for the Double Latent and 45 for the non-spatial. The small difference between the MPR and AMPR for the Clipped Gaussian model indicates that the mis-predictions were rarely more than one category away from the true value. The difference between MPR and AMPR is greater for the Double Latent and even greater for the non-spatial model, indicating that the predictions are farther from the truth. The estimated Bayesian expected loss (BEL) is lower for the Clipped Gaussian model

Table 6.3: Evaluation of predictive ability for the three models, Clipped Gaussian (CG), Double Latent (DL), and Non-spatial (NS) using mis-prediction rate (MPR), absolute mis-prediction rate (AMPR), and the estimated Bayesian expected loss (BEL). The CG and DL models were fit using $\sigma^2 = 1$ and using *Prior 0.05*.

Model	MPR	AMPR	BEL
DL	0.51	0.66	0.56
CG	0.39	0.43	0.33
NS	0.55	0.74	0.49

which we expect based on the simulation results from Chapter 5 showing that BEL is nearly always lower when applying the Clipped Gaussian versus the Double Latent. A larger BEL is indicative of a larger maximum estimated probability corresponding to the predicted category, and can thus be associated with increased certainty in the predictive decision.

The simulation study in Chapter 5 reveals that a comparison of the prediction results using MPR and AMPR generally choose the data-generating model, as described in Section 5.2.2.3. Since the primary goal of the model is prediction, it makes practical sense to select the model with the lowest number of mis-predictions and absolute mis-predictions, which is clearly the Clipped Gaussian model in this case. Image plots of the prediction results, including the training data, for the Clipped Gaussian, Double Latent, and non-spatial models are shown in the plots within Figure 6.11. We also include the image plot of the true data for comparison. The plots clearly show the superior prediction of the Clipped Gaussian and Double Latent models over the naive non-spatial cumulative probit model. Notice that the Clipped Gaussian model appears to predict a slightly smoother surface with larger patches than the Double Latent. It is difficult to conclude from the plots that the Clipped Gaussian model results in correct prediction of 12 more sites than the Double Latent along with its mis-predictions being closer to the true value.

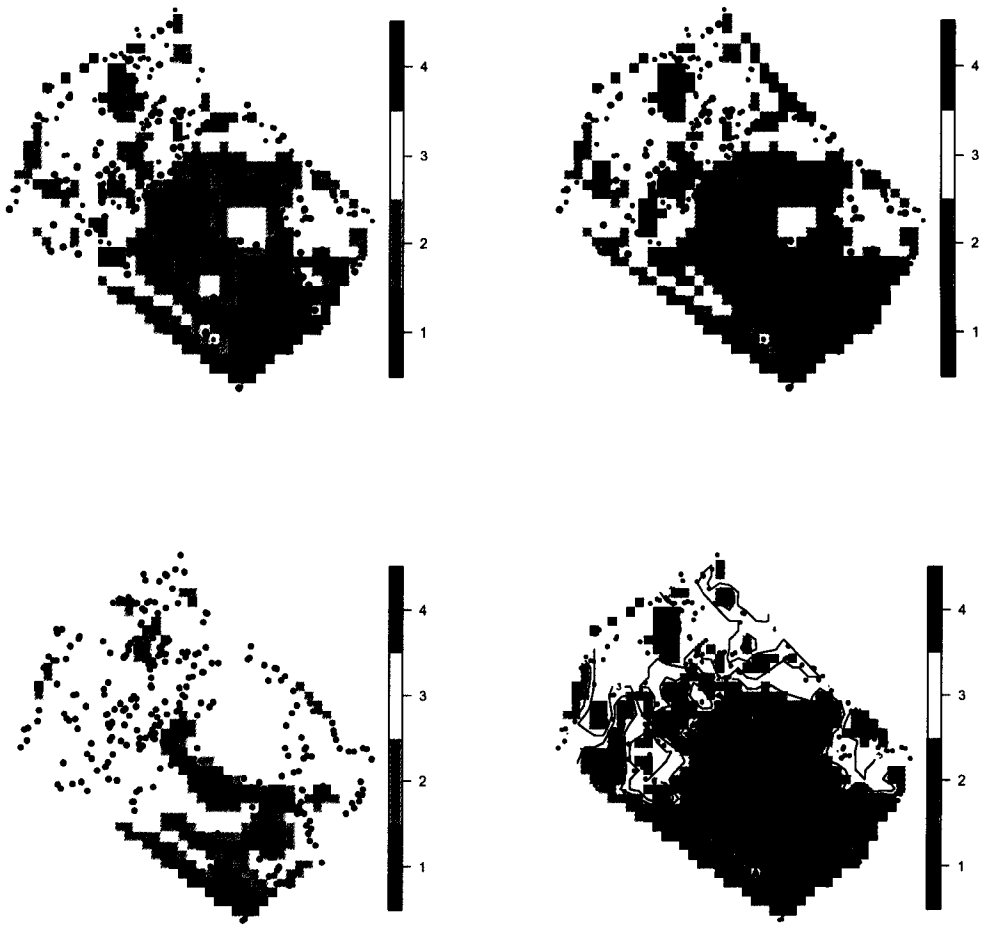

Figure 6.11: A comparison of the image plots of benthic IBI predictions from the Double Latent model with $\sigma^2_{fix} = 1$ (upper left); the Clipped Gaussian model with $\sigma^2_{fix} = 1$ (upper right); and non-spatial cumulative probit model (lower left); along with the true data (lower right). The hold-out sites are denoted by the larger symbols.

Table 6.2 also includes a comparison of results for two values of σ^2_{fix}. For the Double Latent model, the predictive measures are slightly lower for $\sigma^2_{fix} = 2$ than $\sigma^2_{fix} = 1$, showing a difference of 4 mis-predictions. However, for the Clipped Gaussian model, the measures are slightly higher, though only revealing a difference of one prediction. Both the Clipped Gaussian and Double Latent models have better MPR and AMPR than the non-spatial model for both values of σ^2_{fix}, and fitting the Clipped Gaussian model with the $\sigma^2_{fix} = 1$ gives the lowest values.

Our models provide estimates of the probabilities of category membership for each location. If we apply the Link Function method of obtaining these conditional probabilities described in Section 4.3.2 of Chapter 4, they also provide measures of uncertainty in these probability estimates using posterior standard deviations. Figures 6.12 and 6.13 display image plots of the category probability estimates and associated standard deviations for $Pr(Y_i = \text{Poor})$, $Pr(Y_i = \text{Fair})$, $Pr(Y_i = \text{Good})$, and $Pr(Y_i = \text{Excellent})$ obtained using the Clipped Gaussian model.

6.3.2 Estimation of model parameters

Although of secondary interest to our analysis, we report the estimates of model parameters with their associated 95% posterior intervals and compare the estimates given by the two models. Tables 6.4 and 6.5 display and compare results for both the Double Latent (DL) and Clipped Gaussian (CG) models, along with those from the non-spatial cumulative model. Results are also summarized in Figure 6.14. Given the differences in the underlying structure of the three models, we do not expect the estimates of the corresponding parameters to coincide, though we do expect that the practical conclusions should not change between the models.

Notice that all three models produce a negative estimate for β_1, indicating that the benthic IBI score tends to be lower for sites with higher measures of conductivity. The 95% posterior interval for the Double Latent model does not cover zero,

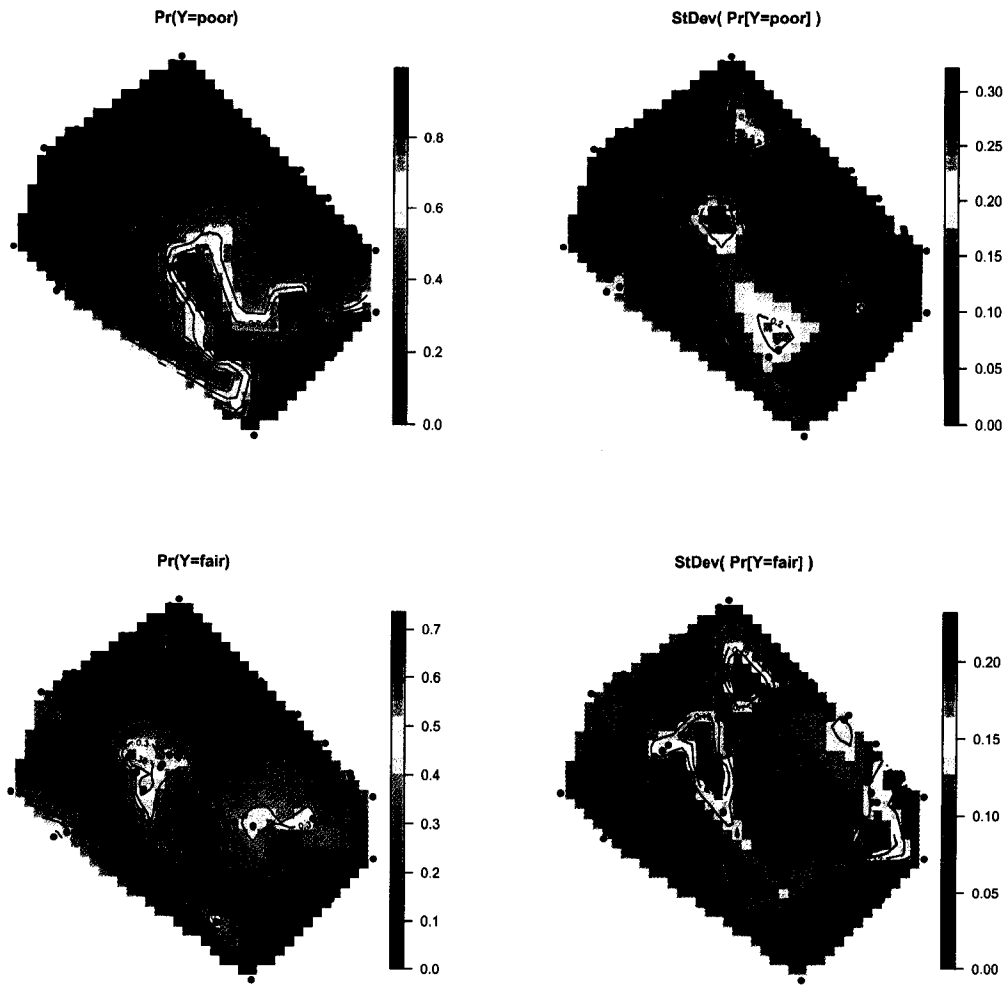

Figure 6.12: Image plots of the category probability posterior means and and standard deviations for $Pr(Y_i = \text{Poor})$ and $Pr(Y_i = \text{Fair})$, obtained using the Clipped Gaussian model with $\sigma^2 = 1$ and *Prior0.6*.

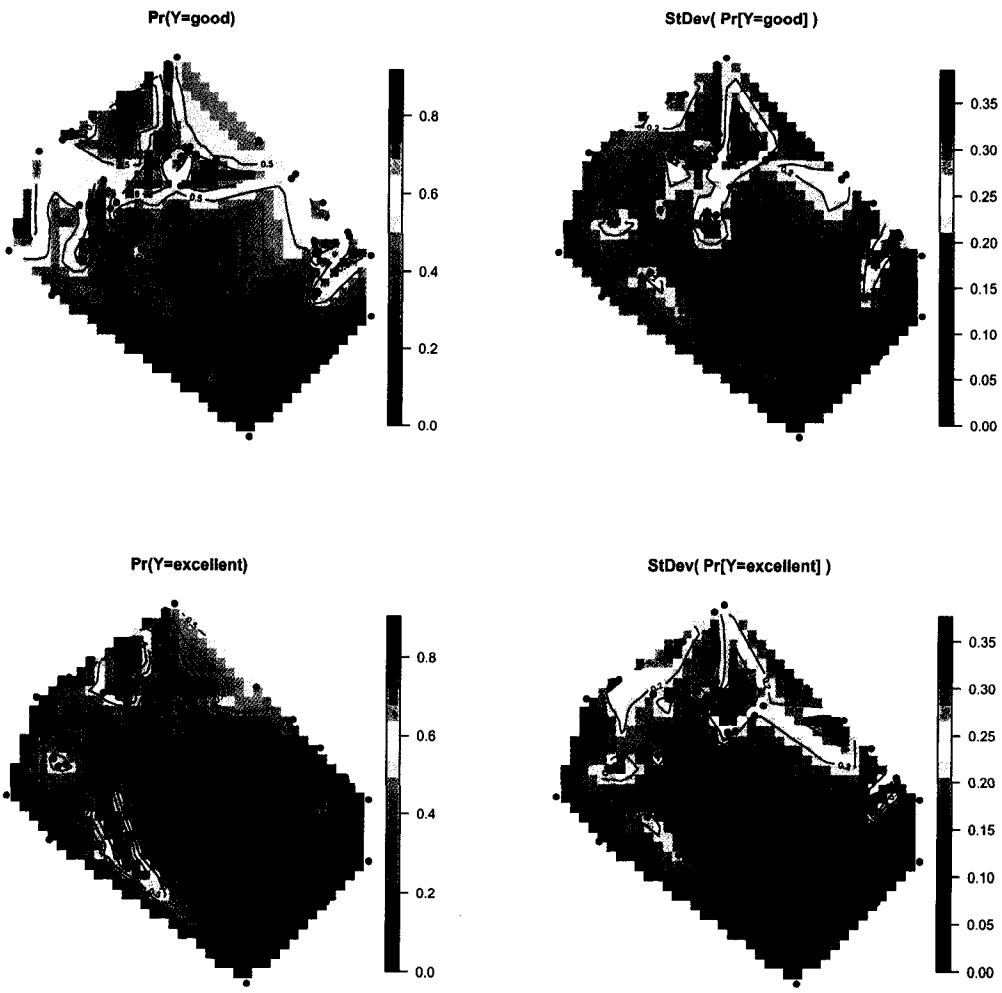

Figure 6.13: Image plots of the category probability posterior means and and standard deviations for $Pr(Y_i = \text{Good})$ and $Pr(Y_i = \text{Excellent})$, obtained using the Clipped Gaussian model with $\sigma^2 = 1$ and *Prior0.6*.

indicating significant evidence for this negative relationship. The same is true for the approximate 95% confidence interval constructed for the non-spatial model, although we expect this interval to be too narrow because the model ignores the lack of independence in the data. The Clipped Gaussian model reveals 95% posterior intervals with an upper end falling very close to, if not on, zero, providing less evidence of a negative relationship between benthic IBI and conductivity.

We use the estimate of the correlation scale parameter, ϕ, to make a conclusion regarding the effective range of the spatial correlation. The model is parameterized using the scale parameter, ϕ, defined as the reciprocal of what is commonly termed the *range* parameter since it is used to estimate the range of the spatial correlation. The practical range for the exponential correlation function is given by $3/\phi$. From the empirical semi-variogram shown in Figure 6.6, we visually estimate the effective range at slightly greater than 3 before accounting for the spatial correlation explained by conductivity. The estimates of ϕ result in practically reasonable estimates of the effective range for both models and both priors for $\sigma^2_{fix} = 1$. The Double latent model estimates an effective range of 3.3 and the Clipped Gaussian model estimates a shorter effective range of 1.79. However, when $\sigma^2_{fix} = 2$ the Clipped Gaussian model estimate leads to an effective range of 13, a value that does not seem reasonable given the exploratory data analysis. Interestingly, the seemingly less appropriate prior (*Prior 0.05*) results in as, or more, reasonable estimates for both models with $\sigma^2_{fix} = 1$, estimating an effective range of approximately 2.5 after accounting for relationship between benthic IBI response and conductivity.

Table 6.4 also includes a comparison of results for the two values of σ^2_{fix}. The estimates for the Clipped Gaussian model under the two different values reveal a greater difference than those of the Double Latent model. For $\sigma^2_{fix} = 2$, the Clipped Gaussian point estimates are closer to zero. For β_1, the posterior interval is also narrower and includes zero, whereas that for $\sigma^2_{fix} = 1$ does not include zero with an

Table 6.4: Posterior means of the model parameters with 95% posterior intervals for the Double Latent (DL), Clipped Gaussian (CG) for *Prior 0.6*, along with estimates and approximate 95% Wald's confidence intervals for the non-spatial (NS) model.

Model	σ^2	β_0	β_1
DL	1	1.16 (-0.04, 2.77)	-0.82 (-1.37, -0.38)
CG	1	0.39 (-0.39, 1.05)	-0.10 (-0.18, -0.02)
DL	2	1.20 (-0.39, 3.13)	-0.90 (-1.56, -0.34)
CG	2	0.05 (-1.58, 1.65)	-0.03 (-0.08, 0.02)
NS	-	-0.88(-1.09, -0.66)	-0.63 (-0.79, -0.47)

Model	σ^2	θ_2	θ_3	ϕ
DL	1	1.13 (0.19, 3.22)	3.01 (1.14, 5.46)	0.91 (0.21, 2.08)
CG	1	0.44 (0.24, 0.60)	1.48 (0.80, 1.89)	1.68 (0.48, 2.43)
DL	2	1.19 (0.12, 3.47)	3.39 (0.94, 6.62)	1.02 (0.15, 2.25)
CG	2	0.15 (0.06, 0.33)	0.49 (0.21, 1.09)	0.23 (0.05, 0.83)
NS	-	0.75 (0.56, 0.94)	2.31 (2.05, 2.57)	-

upper endpoint of -0.02. The estimates from the Double Latent model were less affected by the value of $\sigma^2_{fix} = 1$.

6.4 Conclusions and discussion

For the regression parameters and the cut-point parameters, both models exhibit little sensitivity to different parameter specifications for the prior distributions and virtually no sensitivity to different starting values. However, the Clipped Gaussian model does show sensitivity to the prior specification for the spatial scale parameter, ϕ, of the exponential correlation function. The prediction results reveal less sensitivity, which is advantageous given the main goal of applying these models. Interestingly, the mis-prediction rate (MPR) and absolute mis-prediction rate (AMPR) are slightly lower for the naive *Prior 0.05* than the seemingly better *Prior 0.6*. Future work should include a more thorough sensitivity analysis for the scale parameter.

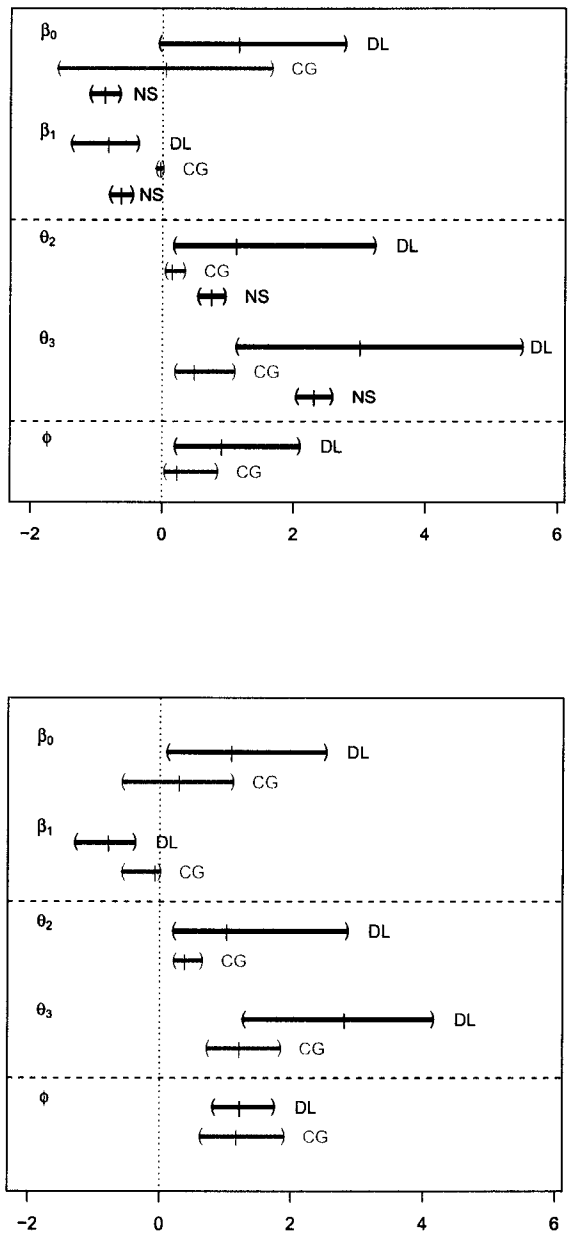

Figure 6.14: The top plot displays posterior means and 95% posterior intervals for the Double Latent(DL) and Clipped Gaussian (CG) models for the benthic IBI data for *Prior 0.6*, along with estimates and approximate 95% Wald's confidence intervals for the non-spatial (NS) model. The bottom plot displays the results for *Prior 0.05*.

Table 6.5: Posterior means and 95% posterior intervals for the Double Latent (DL) and Clipped Gaussian (CG) models for *Prior 0.05*, along with estimates and approximate 95% Wald's confidence intervals for the non-spatial (NS) model.

Model	σ^2	β_0	β_1
DL	1	1.09 (0.12, 2.51)	-0.78 (-1.28, -0.39)
CG	1	0.29 (-0.56, 1.10)	-0.07 (-0.56, -0.01)
NS	-	-0.88 (-1.09, -0.66)	-0.63 (-0.79, -0.47)

Model	σ^2	θ_2	θ_3	ϕ
DL	1	1.02 (0.22, 2.84)	2.81 (1.28, 5.14)	1.22 (0.82, 1.73)
CG	1	0.38 (0.23, 0.63)	1.21 (0.73, 1.82)	1.17 (0.63, 1.88)
NS	-	0.75 (0.56, 0.94)	2.31 (2.05, 2.57)	-

In our investigation of sensitivity to the value of σ^2_{fix}, we found that the Clipped Gaussian model was more sensitive to the value in terms of estimation, whereas the Double Latent model was more affected in terms of prediction. Neither of these effects are dramatic, but it is something that should be further explored and indicates the importance of using a hold-out data set or cross-validation to check the predictive ability for different values of σ^2, as it is likely that different data sets will respond differently to changes in this value.

Thus, we conclude that in terms of prediction, the Clipped Gaussian model results in superior performance. However, for estimation, the Double Latent model produces estimates that are less sensitive to prior specification and fixed parameter values, as well as seemingly more reasonable point estimates and associated uncertainty intervals.

6.4.1 Applying the methods to stream data

Our models are most directly applied to a region of interest with an easily defined study area, D, such that any point within the boundary is a possible location of interest. We usually think of the continuous domain, D, as a continuous block

of area on the surface of the earth, such as a particular study area, an ecoregion, or a county. However, continuous domains can also have a less traditional and less convex shape, such as those covering river and stream beds, where it only makes practical sense to think of locations that lie on the stream network and not at any arbitrary point within the region of interest. When mapping a stream network over a large region, the shape of the domain can become very complex. With the use of geographic information system (GIS) technology, it may be possible to better define such a region and produce maps that are more suited to this application. However, this is beyond the scope of this research.

In the case of predictions over entire stream network, it is common to construct a continuous prediction map over the region of interest with the basic understanding that the values are only meaningful at points that fall on the stream network. This perhaps over-simplification does result in visually pleasing maps for decision makers and the public sector. In our application to Montgomery County, Maryland, the area is relatively small and densely sampled, making us feel more comfortable applying these models as if the entire region were of interest. However, we only obtain covariate values for locations on the stream networks and thus are only interested in predicting at these sites. Care should be taken when applying these models to stream and river data in general, especially over networks spread across a larger area and with less dense sampling.

Another possible limitation in our application to stream data is the measure of distance we use to define the spatial correlation function. We naively use Euclidean distance which does not take into account the networked structure and flow inherent of streams. It would be more appropriate to use a distance measure that takes into account the network structure in addition to the direction and volume of stream flow (Ver Hoef et al., 2006). Ver Hoef et al. (2006) discuss valid models for stream data and show that some commonly used functional forms for modeling spatial autocovariance

are not valid for use with stream distance, such as the spherical and linear-with-sill models. The exponential correlation function is valid when used with stream distance. Ver Hoef et al. (2006) develop valid spatial autocovariance models for stream networks that incorporate stream distance and flow using a method based on moving averages. They apply their models to stream sulfate concentration data from a stream network in Maryland collected by the U.S. EPA for the Maryland Biological Stream Survey (MBSS). The incorporation of flow into the models resulted in reduced root mean square prediction error from cross validation as compared to a model based only on stream distance. Ideally, we would use their proposed method of incorporating both flow and stream distance when constructing a model for the spatial autocovariance, and consider this future work. Ver Hoef et al. (2006) also make predictions over a dense set of locations along the stream network to create a stream "map" of the predicted values. Each prediction is represented by a circle whose size is proportional to the level of precision in the prediction and the color indicates the value of the prediction. This is an idea that could again be incorporated into our work with better use of GIS tools.

Cressie et al. (2006) further formalize the ideas of Ver Hoef et al. (2006) by modeling the spatial dependence through the definition of a one-dimensional process convolution based on Brownian motion on the stream network. They choose a kernel that is nonnegative, linearly decreasing, and that gives zero weight to stream distances greater than some fixed value. They create maps of the stream or river network following the strategy of Ver Hoef et al. (2006), using variable colors to represent magnitude and variable sizes to represent level of precision.

Despite the limitations of our models when applied to stream data, this application reveals that they can be successful in predicting a categorical variable at unobserved locations. The importance of including both the Double Latent and Clipped Gaussian models in this work is clearly demonstrated by the difference in predictive

ability for the two models for the real data. In this case, the Clipped Gaussian model is superior for prediction, but based on simulation results we strongly suspect that the Double Latent model will be superior for some data sets.

Chapter 7

CONCLUSIONS AND FUTURE WORK

In this dissertation, we develop models and methods for prediction of an ordered categorical response in a spatial setting. The models naturally incorporate the spatial dependence through the use of latent Gaussian processes with a spatial covariance structure. The connection between the use of continuous latent variables and the ordered categorical response is accomplished by "clipping" the continuous variable using a vector of cut-point, or threshold, values. We propose two models utilizing this concept of clipping, the Double Latent model and the Clipped Gaussian model. These models fill a void in the literature regarding spatial models for point-referenced ordered categorical data.

The Double Latent model is developed in the spirit of generalized linear mixed models (GLMMs), with a zero-mean latent Gaussian spatial process introduced as a random effect term, in addition to a typical white-noise variance component. The mean is specified as a function of covariates, and the variance-covariance of the Gaussian process is specified through a correlation function resulting in a valid covariance matrix.

The Clipped Gaussian model is developed as an extension of the clipped Gaussian random field approach proposed by De Oliveira (1997, 2000) for binary data. This model assumes that the latent Gaussian random field is clipped directly to obtain the ordered categorical response. The mean is again specified as a function of covariates, but now at the level of the Gaussian process. The variance-covariance matrix is defined in the same manner as that for the Double Latent model. This

model does not include the additional white noise component defined in the GLMM structure of the Double Latent model.

7.1 Conclusions

In Chapter 1, we briefly introduced the two models and provided needed background information in the form of a literature review of models for independent ordered categorical data and spatial models proposed for binary and count data, along with introductions to spatial modeling and terminology, Bayesian inference, and Markov chain Monte Carlo (MCMC) methods.

Chapter 2 provided the details of each model along with a description of the Gibbs sampling algorithm used to fit each of the models. Derivations of the complete conditional distributions used to implement the algorithm and details of transformations and computational improvements made to the standard Gibbs sampling approach were also included.

We provided comparisons of the models in Chapter 3 where we addressed a variety of more practical issues related to fitting the models. This included an investigation into how the specification of the models ultimately translates into correlation in the categorical response. Identifiability of the parameters and strategies to deal with potential non-identifiability were also discussed. Specification of prior distributions for the parameters of the models was addressed and specific recommendations given. Investigation of the second-order structure of the categorical random variable reveals that data generated under the Clipped Gaussian model possess correlations of higher magnitude than those created under the Double Latent model, a direct result of the extra white noise variance component incorporated into the Double Latent model.

The primary goal of the models is prediction at unobserved locations, and Chapter 4 outlined the details of making and evaluating the predictions. This included

an introduction to formulating prediction as a decision problem within the Bayesian paradigm, along with a description of two proposed methods for estimating the probabilities needed to make the predictions.

Chapter 5 provided an in-depth simulation study using artificial data simulated under known models and specifications with the purpose of assessing the predictive ability of the models and their success at estimation. We fit the models to four different data scenarios defined by different parameter values and also apply each model to data generated under the assumptions of the other model, along with the use of two different fixed values for σ^2 for each model. Results for prediction and estimation are summarized and discussed, providing necessary information regarding the behavior of the models and serving to pinpoint areas and issues that deserve further investigation.

The study demonstrates the success of the models on a wide variety of data realizations and provides confidence in the practical usefulness of the models. Here we highlight just a few of the interesting conclusions gleaned from the simulations. First, the Clipped Gaussian model generally results in superior predictive ability, even for data generated under the Double Latent model. This is most likely an effect of the extra variance component incorporated into the Double Latent model. Additionally, the prediction results for both models are very dependent on the individual data realizations, displaying excellent predictions for some realizations and less accurate prediction for others. The predictive success appears to be related to the smoothness (or patchiness) of the realization, with smooth realizations generally resulting in better prediction. A smoother realization is characterized by larger and fewer patches of each category over the region of interest. This result was hypothesized in Chapter 3 and was also suggested by De Oliveira (2000).

In terms of estimation, we find that both models are able to accurately capture the relationship between the response and the covariate defined in the generation

of the data. This is true even with the relatively large variability in the posterior estimates of the regression parameter, β_1 over the many realizations. The 95% posterior interval are wider for the Double Latent model regardless of the data generating mechanism.

Finally, our models are applied to solve a real problem in Chapter 6. The ordered categorical response of interest is the benthic Index of Biotic Integrity (benthic IBI), interpreted as a measure of stream health. The study region is Montgomery County, Maryland, where spatially dense sampling of the stream networks results in the need for models incorporating spatial dependence. Using a hold-out data set of 100 observations, we find that the Clipped Gaussian model is more successful in terms of prediction at new sites compared to the Double Latent model, and that both of our proposed models are more successful than a naive non-spatial cumulative probit model. Both models provide posterior estimates for the regression parameter that indicate moderate to conclusive evidence of a negative relationship between the covariate of conductivity and the benthic IBI category.

7.2 Future work

We devote this section to an introduction and discussion of specific areas and directions of future research surrounding the proposed models. Many of these points were briefly mentioned in previous chapters and are expanded upon here.

7.2.1 Computation

The computational expense of fitting these models is relatively high. Therefore, one area for future research lies in the improvement of the efficiency of the algorithms. We chose to implement a Gibbs sampling algorithm as a logical first step in the development of these methods. Now that we have demonstrated success with the more traditional algorithm, other possibly computationally superior methods can be investigated and compared to the results obtained with our algorithm. Possibilities

include the use of methods to improve the efficiency of matrix inversion, especially for large data sets. One strategy is the use of a spectral representation such that the covariance matrix can be embedded into a block-circulant matrix, allowing use of methods such as fast Fourier transforms (FFTs) to compute the inverse and square root of the variance-covariance matrix (Wikle, 2002; Royle and Wikle, 2005; Paciorek and Ryan, 2005). This method requires that the spatial locations be forced into a lattice structure, which for point-referenced data is usually a disadvantage. Manipulation of the covariance matrix may also be simplified by use of banded matrices that are formed based on the assumption that $\rho(\mathbf{s}_i, \mathbf{s}_j; \phi) = 0$ for pairs of locations $(\mathbf{s}_i, \mathbf{s}_j)$ that are sufficiently far apart (Christensen et al., 2006).

Another strategy is to directly improve the convergence behavior, and thus run time, of the MCMC chain. For example, Meng and van Dyk (1999) propose a "marginal augmentation" which is shown to converge faster than the straight Gibbs algorithm of Albert and Chib (1993) for independent ordered categorical data. Ishwaran (2000) also proposes a "leapfrog hybrid MC" algorithm to improve convergence for non-spatial cumulative link models, although it is not clear that this would reduce computational time.

Finally, we should not rule out the use of approximate Bayesian methods to decrease computational time by the requirement of more theoretical work before fitting the models. For example, Gelfand et al. (2000) describes the use of importance sampling for spatial binary data, and Christensen et al. (2006) propose data dependent reparameterizations to improve and automate convergence of Bayesian spatial models for binary and count data. It is not immediately clear how their methods can be extended to the ordered categorical case, but this is a direction that is worth further study. Additionally, Eidsvik et al. (2006) propose another method for approximate Bayesian inference for spatial GLMMs, in the context of one-parameter exponential family problems, based on the use of the marginal posterior density of the model

parameters and the marginal posterior distribution of the latent spatial variable at a single location.

In a non-Bayesian context, Kammann and Wand (2003) propose *geoadditive models* that merge geostatistical and additive models using a nonparametric mixed model representation employing penalized splines. Such methods are described as "reduced knot" or "low rank" kriging and provide a solution to the inversion of large matrices. They are typically fit using (restricted) maximum likelihood. Note that Christensen et al. (2006) also employ a nonparameteric approach to estimating the covariance function.

7.2.2 Prior specification and formal sensitivity analysis

For this dissertation, considerable effort was spent, before undertaking the formal simulation study, fitting the models for different prior distributions and hyperparameter values. This work is not summarized in a formal prior sensitivity analysis, but this is something that should be incorporated in future work, now that we have demonstrated the usefulness of these models. As discussed in Chapter 3, there is room for more in-depth research related to choice of the prior distributions and specification of their parameters, particularly those related to parameters specifying the spatial correlation function and those for the cut-point parameters.

7.2.3 Additional simulation studies

Given the possible number of parameter and model combinations and the computational time required to fit the models of interest, we were restricted in the number of comparisons we could make. Thus, a logical future direction of research is continued investigation of model behavior using artificial data. For example, it is of immediate interest to investigate performance on data with a more varied categorical composition, particularly that with a higher proportion of observations in the middle categories, rather than the end categories. The effect of varying the scale parameter

of the correlation function and employing different correlation functions is also of interest. Additionally, an investigation of fixing the value of σ^2 to values at varying distances from the true value is also needed.

7.2.4 Use of the predictive measures

It is also possible that we can create a more informative measure of predictive ability by combining the information in mis-prediction rate (MPR) and absolute mis-prediction rate (AMPR) with that contained in the estimated Bayesian expected loss (BEL). For example, we could compare the average BEL over locations with correct predictions to those with incorrect predictions, thus providing valuable information regarding whether higher probability estimates coincide with correct predictions, as we hope they would. For example, we would consider it a positive characteristic of the model if the BEL for correct predictions was greater than the BEL for mis-predictions. This idea can be carried further to compare the BEL for locations in which predictions are off by one category to that over locations in which the predictions are off by more than one category. A decrease in the average BEL with an increase in prediction error would indicate that BEL truly is providing a useful measure of the uncertainty in the predictive decision. Conversely, if the average BEL values for the three cases are equal, this would indicate that the model is producing probability estimates of the same magnitude regardless of the location, thus providing no meaningful information regarding uncertainty in the decision.

7.2.5 Alternative link functions

The use of the probit link function, defined by the inverse standard normal cumulative distribution function, provides a useful and logical formulation of the model for both theoretical and computational reasons. However, as these models see more practical use, it is likely that different link functions will be desired, particularly the logit link defined by the standard logistic cumulative distribution function. The

logit link is particularly attractive due to its advantage in terms of interpretation, allowing conclusions to be framed in the context of odds ratios. The incorporation of this link function into the Double Latent model is straightforward, simply accomplished by changing the distribution of the white noise process from standard normal to standard logistic. The Clipped Gaussian model is not as adaptable, although we suggest looking into the use of multivariate logistic distribution to define the spatial process, taking the place of the latent Gaussian random field. In fact, O'Brien and Dunson (2004) propose a new likelihood for multivariate logistic regression based on the idea of finding multivariate distributions that have marginal logistic distributions, and their work may provide a starting place for logistic spatial models for ordered categorical data. There is also the possibility of using a link function defined by the t-distribution with estimable degrees of freedom, such as that described by Albert and Chib (1993) for independent data, to add flexibility and robustness to the models.

In Chapter 2 we briefly described an alternative version of the Double Latent model that specifies the linear function of covariates, $\mathbf{X}\boldsymbol{\beta}$, as the mean of the spatial Gaussian process rather than of the first-stage latent variable, $\mathbf{Z}$. This reparameterization may provide insight into the advantages and disadvantages of specifying the mean at different levels of the hierarchy. Both levels of mean specification are observed in the literature. For example, Diggle et al. (1998) incorporates this mean at the first-stage, whereas Christensen et al. (2006) includes $\mathbf{X}\boldsymbol{\beta}$ at the second-stage in the mean of the Gaussian random field. Knowledge of the computational and/or theoretical advantages to either method would be useful in a variety of applications involving hierarchical spatial models.

7.2.6 Alternative loss functions

In Chapter 4 we described two loss functions, but only used one in the remainder of the dissertation. Use of the second loss function, defined to provide a greater

penalty for predictions that fall farther from the true value, is also of immediate interest. The method of defining prediction in terms of decision theory allows the loss function to change for different applications, thus providing possibly advantageous flexibility in terms of maximizing the predictive ability using a problem-specific strategy. We suspect that a comparison of different loss functions for a particular type of data may provide interesting insights into this method of spatial prediction and also into the use of the predictive measures we employ in Chapters 4, 5, and 6.

7.2.7 Identifiability of parameters

Another direction of future research was discussed in Chapter 3 regarding the estimation of variance parameters used to specify a latent continuous variable or process. The amount of information available in categorical data to estimate such a parameter remains unknown, and depends on the problem-specific number of categories. A greater number of categories may often lead to values of the cut-points that extend farther into the tails of the continuous distribution, which should provide greater information regarding the spread of this latent distribution. It may be possible to recommend a minimum number of categories such that the parameter can be estimated. However, the number of categories would most likely have to be coupled with an appropriate coverage of the cut-points over the range of the continuous variable. This future work should rely on both analytical work and simulation studies.

In Chapter 4 we developed the method of prediction under the assumption that the categories of the response are given successive integer labels, termed the integer label framework. We also briefly describe the problem under the indicator-label framework for cases when integer labels are not deemed appropriate. It would be interesting to pursue this direction of specifying the prediction problem in this multivariate, and more general, fashion. The two methods should be compared in terms of the correlation they induce in the categorical response variable.

7.2.8 Conclusion

In conclusion, our research fills a void in the current literature surrounding spatial models for point-referenced ordered categorical data. We have proposed two concurrent models and pointed to areas of future research, related both to the proposed models and to the analysis of spatially point-referenced non-Gaussian data in general. It is our hope that this work will spark additional discussion and research in this area, and eventually lead to techniques that can be readily applied by researchers collecting ordered categorical data in a spatial setting. The need for such models arises in a variety of scientific disciplines where spatial dependence is an integral and useful characteristic of the data, and the data are either intrinsically categorical or transformed to categorical in order to simplify data collection.

Bibliography

Abramowitz, M. and Stegun, I. A., editors (1965). *Handbook of mathematical functions.* Dover, New York, New York, USA.

Agresti, A. (2002). *Categorical Data Analysis.* John Wiley & Sons, Inc., Hoboken, New Jersey, USA.

Agresti, A. and Lang, J. B. (1993). A proportional odds model with subject-specific effects for repeated ordered categorical responses. *Biometrika*, 80:527–534.

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88:669–679.

Albert, J. and Chib, S. (1995). Bayesian residual analysis for binary response regression models. *Biometrika*, 82:747–759.

Albert, J. and Chib, S. (1997). Bayesian methods for cumulative, sequential, and two-step ordinal data regression models. Technical report, Bowling Green State University.

Albert, J. and Chib, S. (1998). Sequential ordinal modeling with applications to survival data. Technical report, Bowling Green State University.

Albert, J. and Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics*, 57:829–836.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

Berger, J. . (1985). *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, New York, USA.

Berger, J. O., Oliveira, V. D., and Sanso, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96:1361–1374.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B*, 36:192–236.

Besag, J. and Green, P. J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society B*, 55:25–102.

Bradlow, E. T. and Zaslavsky, A. M. (1999). A hierarchical latent variable model for ordinal data from a customer satisfaction survey with "no answer" responses. *Journal of the American Statistical Association*, 94:43–52.

Chen, M.-H. and Dey, D. K. (1996). A unified Bayesian approach for analyzing correlated ordinal response data. *Brazilian Journal of Probability and Statistics*, 14:87–111.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, 85:347–361.

Christensen, O. F., Moller, J., and Waagepetersen, R. (2000). Analysis of spatial data using generalized linear mixed models and Langevin-type Markov chain Monte Carlo. Technical report, Department of Mathematical Sciences, Aalborg University.

Christensen, O. F., Roberts, G. O., and Skold, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 15:1–17.

Christensen, O. F. and Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, 58:280–286.

Christensen, R., Johnson, W., and Pearson, L. M. (1992). Prediction diagnostics for spatial linear models. *Biometrika*, 79:583–391.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized linear models. *Statistics and Computing*, 6:101–111.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91:883–904.

Cowles, M. K., Carlin, B. P., and Connett, J. E. (1996). Bayesian tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association*, 91:86–98.

Cressie, N., Frey, J., Harch, B., and Smith, M. (2006). Spatial prediction on a river network. *Journal of Agricultural, Biological, and Environmental Statistics*, 11:127–150.

Cressie, N. A. C. (1993). *Statistics for Spatial Data.* John Wiley & Sons Inc., New York, New York, USA.

Crouchley, R. (1995). A random-effects model for ordered categorical data. *Journal of the American Statistical Association*, 90:489–498.

De Oliveira, V. (1997). *Prediction in some classes of non-Gaussian random fields.* PhD thesis, University of Maryland at College.

De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *Computational Statistics and Data Analysis*, 34:299–314.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, 47:299–350.

Eidsvik, J., Martino, S., and Rue, H. (2006). Approximate Bayesian inference in spatial generalized linear mixed models. Technical report, Department of Mathematical Sciences, Norwegian University of Science and Technology.

Gelfand, A. E. and Ravishanker, N. (1998). Inference for Bayesian variogram models with large data sets. Technical report, Department of Statistics, University of Connecticut.

Gelfand, A. E., Ravishanker, N., and Ecker, M. D. (2000). *Generalized Linear Models: A Bayesian Perspective*, chapter Modeling and inference for point-referenced binary spatial data, pages 373–386. Marcel Decker Inc.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Gelman, A. (2004). Parameterization and Bayesian modeling. *Journal of the American Statistical Association*, 99:537–545.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

Geman, A. and Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Givens, G. and Hoeting, J. A. (2006). *Computational Statistics*. John Wiley & Sons, Inc., Hoboken, New Jersey.

Gotway, C. A. and Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, 2:157–178.

Harville, D. A. and Mee, R. W. (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics*, 40:393–408.

Hocking, R. R. (2003). *Methods and applications of linear models: Regression and analysis of variance.* John Wiley & Sons, Inc., Hoboken, New Jersey.

Irvine, K. M., Gitelman, A. I., and Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: effects on covariance estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, to appear.

Ishwaran, H. (2000). Univariate and multirater ordinal cumulative link regression with covariate specific cutpoints. *The Canadian Journal of Statistics*, 28(4):715–730.

Ishwaran, H. and Gatsonis, C. A. (2000). A general class of hierarchical ordinal regression models with applications to correlated roc analysis. *The Canadian Journal of Statistics*, 28:731–750.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling.* Springer-Verlag, New York, New York, USA.

Kammann, E. E. and Wand, M. P. (2003). Geoadditive models. *Applied Statistics*, 52:1–18.

Lark, R. M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, 51:717–728.

Lawrence, E., Bingham, D., Liu, C., and Nair, V. N. (2006). Bayesian inference for ordinal data using multivariate probit models. *Technometrics*, submitted.

Lee, Y. and Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58:619–678.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, B*, 42:109–142.

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

Meng, X.-L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika*, 86:301–320.

Murphy, S. F. (2007). *General information on specific conductance.* Boulder Area Sustainability Information Network (BASIN).

Nandram, B. and Chen, M.-H. (1996). Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation*, 54:129–144.

O'Brien, S. M. and Dunson, D. B. (2004). Bayesian multivariate logistic regression. *Biometrics*, 60:739–746.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, 71:531–543.

Paciorek, C. J. and Ryan, L. (2005). *Computational techniques for spatial logistic regression with large datasets.* Number 32 in Harvard University Biostatistics Working Paper Series. The Berkeley Electronic Press.

Poirier, D. J. (1998). Revising beliefs in nonidentified models. *Econometric Theory*, 14:483–509.

Qiu, Z., Song, P.-K., and Tan, M. (2002). Bayesian hierarchical models for multi-level repeated ordinal data using WinBUGS. *Journal of Biopharmaceutical Statistics*, 12:121–135.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramsey, F. and Schafer, D. (2002). *The Statistical Sleuth: A Course in Method of Data Analysis*. Duxbery Press, Pacific Grove, CA, USA.

Royle, J. A. and Wikle, C. K. (2005). Efficient statistical mapping of avian count data. *Environmental and Ecological Statistics*, 12:225–243.

Schabenberger, O. and Gotway, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC Press, Boca Raton, Florida, USA.

Schmidt, A. M., Conceicao, M. G., and Moreira, G. A. (2007). Investigating the sensitivity of Gaussian processes to the choice of their correlation function and prior specifications. *Journal of statistical computation and simulation*, to appear.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and der Linde, A. V. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64:583–616.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540.

U.S. Environmental Protection Agency (EPA) (2003). *Introduction to the Clean Water Act*. U.S. Environmental Protection Agency (EPA).

U.S. Environmental Protection Agency (EPA) (2007). *Mid-Atlantic Integrated Assessment - The state of Maryland's freshwater streams.* U.S. Environmental Protection Agency (EPA).

van Dyk, D. A. (2002). *Statistical challenges in modern astronomy III*, chapter Hierarchical models, data augmentation, and Markov chain Monte Carlo, pages 41–56. Springer-Verlag, New York, New York, USA.

Ver Hoef, J. M., Peterson, E., and Theobald, D. M. (2006). Spatial statistical models that use flow and stream distance. *Environmental and Ecological Statistics*, 13:449–464.

Vounatsou, P., Smith, T., and Gelfand, A. E. (2000). Spatial modelling of multinomial data with latent structure: an application to geographical mapping of human gene and haplotype frequencies. *Biostatistics*, 1:177–189.

Wikle, C. K. (2002). *Spatial Cluster Modelling*, chapter Spatial modelling of count data: A case study in modelling breeding bird survey data on large spatial domains, pages 199–209. Chapman and Hall/CRC, Boca-Raton, Florida, USA.

Xie, Y. and Carlin, B. P. (2006). Measures of Bayesian learning and identifiability in hierarchical models. *Journal of Statistical Planning and Inference*, 136:3458–3477.

Zhao, Y. (2003). *General design Bayesian generalized linear mixed models with applications to spatial statistics.* PhD thesis, Department of Biostatistics, Harvard University.

Zhao, Y., Staudenmayer, J., Coull, B. A., and Wand, M. P. (2006). General design Bayesian generalized linear mixed models. *Statistical Science*, 21(1):35–51.

Zhu, J., Eickhoff, J. C., and Yan, P. (2005a). Generalized linear latent variable models for repeated measures of spatially correlated multivariate data. *Biometrics*, 61:674–683.

Zhu, J., Huang, H.-C., and Wu, J. (2005b). Modeling spatial-temporal binary data using Markov random fields. *Journal of Agricultural, Biological, and Environmental Statistics*, 10(2):212–225.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics*, 17(6):635–652.

Zimmerman, D. L. and Zimmerman, M. B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, 33(1):77–91.