12.x Spatial Generalized Linear Models

Many types of data are binary, counts, or positive continuous. Early attempts to model such data relied on transformations to "near normal" so that the methods of classical linear models could be used. For example, a square root transformation was often used for count data. However, Nelder and Wedderburn (1972) introduced a natural extension to linear models using parameteric distributions such as the Poisson distribution for counts, the Bernoulli distribution for binary data, etc., called generalized linear models (GLM, McCullagh and Nelder 1989), which have become very popular and usually preferred to data transformations. A natural extension of GLMs occurs by introducing latent random effects as a linear mixed model to create a class of generalized linear mixed models (GLMM, Breslow and Clayton 1993). The latent random effects are usually assumed to be independent and identically distributed from a normal distribution. However, it is also possible for the latent random effects to be spatially autocorrelated, leading to the spatial generalized linear model (SGLM, Gotway and Stroup 1997; Diggle et al. 1998), which we review here.

Historically, for areal data, such as the models in Chapter 7, there are equivalent models for discrete data, such as those that are binary or counts. These have been termed the autologistic, auto-Poisson, autobinomial, and auto negative binomial models, with obvious connections to their nonspatial distributions (Besag 1974; Cressie 1993). These models have not been very popular, as the conditional specification does not always lead to a recognizable likelihood. For example, for the auto-Poisson, the likelihood may not have a closed form under positive autocorrelation. We will not discuss these models further.

Another approach comes from penalized quasi-likelihood models (Breslow and Clayton 1993; Wolfinger and O'Connell 1993). These models are an extension of GLMs by using the first and second moments of distributions in the regular exponential family and extending their variance structures to include spatial autocorrelation. These models have been implemented in popular software such as the `glmmPQL` function in the `MASS` package in `R` and the `GLIMMIX` package in `SAS`.

A final class of models are based on a hierarchical constuction, where the mean of any of the distributions in GLMs is allowed to vary by using spatial random effects in the mean structure. Here, there are two broad methods of analysis. The most obvious method is to take a Bayesian approach and compute the posterior distribution of all latent spatial variables and parameters. This has been extremely popular beginning with disease-mapping (Clayton and Kaldor 1987) and the introduction of the `WinBUGS` software (Lunn et al. 2000). Less common is a likelihood approach that attempts to estimate covariance parameters, and perhaps fixed effects simultaneously, while integrating out over all spatial random effects. This can be done using Markov chain Monte Carlo methods (e.g., Christensen 2004) or more directly using a Laplace approximation (e.g., Evangelou et al. 2011; Bonat and Ribeiro Jr 2016). Here, we will focus primarily on Bonat and Ribeiro Jr (2016), and improve their methods.

12.x.1 Parametric models for the mean structure

If our linear model is $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$, then generalized linear models establish a link function

between $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$, which we denote as $g(\boldsymbol{\mu}) = \boldsymbol{\eta}$, where $g(\cdot)$ is called the link function. For the Poisson example, $g(\cdot)$ is often the log function. Link functions are monotonic so that $g^{-1}(\cdot)$ is one-to-one with $g(\cdot)$. The log link makes sense for the Poisson example. Recall that the mean of a Poisson distribution must be positive, and if $g(\cdot)$ is the log function, then $g^{-1}(\cdot)$ is the exponential function. Hence $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$ is always positive and $\boldsymbol{\eta}$ is unconstrained on the real line, as is typical for a linear model $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$.

Most GLMs are motivated by the exponential family of distributions, where

$$f(y) = \exp\left\{\frac{y\theta - b(\theta)}{\delta} + c(y, \delta)\right\}.$$

Many common distributions are special cases of the exponential family, including normal, gamma, Poisson, binomial, negative binomial, and beta. The first and second moments are $\mu = b'(\theta)$ and $\text{var}(y_i) = b''(\theta)\delta$. The variance can be related to the mean by solving $\text{var}(y) = a(\mu)\delta$, where $a(\mu)$ is called the variance function. Note that we are now parameterizing the distribution through μ, $a(\mu)$, and δ, rather than $b''(\theta)\delta$. The attaction of this parameterization is that now we can establish the relationship between the mean and the linear model through $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$.

A fully parametric way to create spatially structured dependence for GLMMs is through a hierarchical construction. We will use the notation $[\mathbf{y}|\boldsymbol{\mu}]$ to denote any probability density function of the vector of random variables $\mathbf{y}$ conditional on a vector of parameters, or other fixed variables, $\boldsymbol{\mu}$. We can have a joint distribution on the left side of the conditional bar, and multiple parameter and fixed value vectors on the right, e.g., $[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k]$. For example, let $[\mathbf{y}|\boldsymbol{\mu}]$ be the product of independent Poisson distributions with mean parameters contained in the vector $\boldsymbol{\mu}$. Although not strictly necessary, it makes most sense for $\boldsymbol{\mu}$ to be the mean for $\mathbf{y}$, so $\text{E}(\mathbf{y}) = \boldsymbol{\mu}$. The model for the data $\mathbf{y}$ can have more parameters than just the mean, in which case we write it $[\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\phi}]$. For an example with extra parameters for $\mathbf{y}$, consider the negative binomial distribution, which can be parameterized with a mean, and an extra parameter that allows for overdispersion, which we would write as $[\mathbf{y}|\boldsymbol{\mu}, \phi]$, where ϕ is the overdispersion parameter.

For the hierarchical construction of a generalized linear mixed model, we condition on random effects, and we will change notation to $\mathbf{w}$ rather than $\boldsymbol{\eta}$ to reflect the fact that $\mathbf{w}$ has a probability distribution. Thus we write $[\mathbf{y}|\mathbf{w}]$, where $\text{E}(\mathbf{y}) = \mathbf{w}$, and $\mathbf{w}$ is generally considered to have multivariate normal distribution, which can be denoted $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}]$, where $\mathbf{X}$ is the design matrix of fixed explanatory variables, $\boldsymbol{\beta}$ is a vector of fixed effects parameters, $\text{E}(\mathbf{w}) = \mathbf{X}\boldsymbol{\beta}$, and the vector $\boldsymbol{\theta}$ contains covariance parameters. Thus, we use the notation $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}]$ to indicate the probability density function $\mathbf{w} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. For the hierarchical construction of a SGLM, we simply let $\mathbf{w}$ be a spatial model, which we turn to next.

12.x.2 Spatially structured dependence

To develop a spatial covariance matrix for the moment-based approach of GLMs, which we now call the penalized quasi-likelihood models, we start with a correlation matrix, which in earlier chapters was denoted $\mathbf{R}$. For the stationary case and normally-distributed

data we obtain the full covariance matrix simply by scaling it by an overall variance parameter, $\mathbf{\Sigma} = \sigma^2\mathbf{R}$. A generalization for SGLMs is $\text{var}(\mathbf{y}) = \mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$, where $\mathbf{R}$ is a spatial autocorrelation matrix (minus the overall variance parameter), and $\mathbf{A}$ is a diagonal matrix that contains the variance function $a(\mu_i)\delta$ from our chose model for the exponential family. Note that in the case of normally-distributed data, $a(\mu_i) = 1$ and $\delta = \sigma^2$, so $\mathbf{\Sigma} = \sigma^2\mathbf{R}$ is a special case. As another example, consider the Poisson distribution. Here $\delta = 1$ and $a(\mu_i) = \mu_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$, where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$. In Section 12.x.4, we will develop inference for these models with quasi-likehood and using iterative fitting algorithms for δ and $\boldsymbol{\beta}$.

For the fully parametric, hierarchical models, let

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e},$$

where this model is the same one defined in (1.1), where $\text{var}(\mathbf{e}) = \mathbf{\Sigma}_{\boldsymbol{\theta}}$, and we use the subscript to show the dependence of $\mathbf{\Sigma}$ on $\boldsymbol{\theta}$. Then, a very general model can be constructed hierarchically as,

$$[\mathbf{y}, \mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{X}] = [\mathbf{y}|\mathbf{w}, \boldsymbol{\phi}][\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{X}]. \tag{12.1}$$

As a concrete example, suppose that $[\mathbf{y}|\mathbf{w}]$ is Poisson, and $[\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{X}]$ is multivariate normal, then the joint likelihood is

$$[\mathbf{y}, \mathbf{w}|\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{X}] = \left(\prod_{i=1}^{n} \frac{\exp(w_i)^{y_i}\exp(-\exp(w_i))}{y_i!}\right) \frac{\exp -[(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})]}{(2\pi)^{n/2}|\mathbf{\Sigma}_{\boldsymbol{\theta}}|^{1/2}},$$

and note the use of $\text{E}(y_i) = \mu_i = g^{-1}(w_i) = \exp(w_i)$.

The joint distribution 12.1 forms the basis for inference, with popular choices given by

- putting prior distributions on $\boldsymbol{\phi}$, $\boldsymbol{\beta}$, and $\boldsymbol{\theta}$ and computing, or sampling from, the joint posterior distribution $[\mathbf{w}, \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{X}]$ using any of a variety of Bayesian methods, or
- integrating over $\mathbf{w}$ using a Laplace approximation, and integrating over $\boldsymbol{\beta}$ as in REML, and then using maximum likelihood to estimate $\boldsymbol{\phi}, \boldsymbol{\theta}$ marginally, followed by GLS estimation of $\boldsymbol{\beta}$ and prediction for $\mathbf{w}$.

a Baysian inference is beyond the scope of this book. Below, we briefly outline quasi-likelihood, and then give more details on the Laplace approximation and the marginal maximum likelihood approach.

12.x.3 Parametric models for the covariance structure

In (12.1), we still need parametric models for $\mathbf{\Sigma}$ in $\mathbf{w} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{\Sigma}_{\boldsymbol{\theta}})$. There are no theoretical constraints here, and any valid spatial model for $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ is possible. For example, $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ may be constructed from geostatistical models (Chapter 6), where $\boldsymbol{\theta}$ often contains the partial sill, range, and nugget effect, or $\mathbf{\Sigma}_{\boldsymbol{\theta}}$ may be constructed from spatial weights models (Chapter 7), where $\boldsymbol{\theta}$ often contains an autocorrelation and variance parameter.

12.x.4 Inference using Quasi-likelihood

Qausi-likelihood makes assumptions on the first and second moments, rather than creating a hierarchical model, but the underlying models are similar. Borrowing from the GLM framework, recall that $\mathrm{E}(\mathbf{y}) = g^{-1}(\mathbf{X}\boldsymbol{\beta})$ and $\mathrm{var}(\mathbf{y}) = \mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}$, where $\mathbf{R}$ is a spatial autocorrelation matrix using one of the models in Chapter 6 or Chapter 7 (minus the overall variance parameter), and $\mathbf{A}$ is a diagonal matrix that contains the variance functions of the model.

A multivariate Taylor-series approximation to a nonlinear function $g(\mathbf{x})$ around some value $\mathbf{a}$ is

$$g(\mathbf{x}) \approx g(\mathbf{a}) + g'(\mathbf{a})(\mathbf{x} - \mathbf{a}),$$

so let us expand $g^{-1}(\mathbf{X}\boldsymbol{\beta})$ around some value of $\tilde{\boldsymbol{\beta}}$,

$$g^{-1}(\mathbf{X}\boldsymbol{\beta}) \approx g^{-1}(\mathbf{X}\tilde{\boldsymbol{\beta}}) + \boldsymbol{\Delta}\mathbf{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}),$$

where $\boldsymbol{\Delta}$ is a diagonal matrix with diagonal elements

$$\boldsymbol{\Delta} = \frac{\partial g^{-1}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}},$$

evaluated at $\tilde{\boldsymbol{\beta}}$, and recall that $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$. Notice that

$$\mathrm{E}[\mathbf{y} - g^{-1}(\mathbf{X}\tilde{\boldsymbol{\beta}}) - \boldsymbol{\Delta}\mathbf{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})] \approx g^{-1}(\mathbf{X}\boldsymbol{\beta}) - g^{-1}(\mathbf{X}\tilde{\boldsymbol{\beta}}) - \boldsymbol{\Delta}\mathbf{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})],$$

and, with some rearrangement,

$$\mathbf{X}\boldsymbol{\beta} \approx \boldsymbol{\Delta}^{-1}[g^{-1}(\mathbf{X}\boldsymbol{\beta}) - g^{-1}(\mathbf{X}\tilde{\boldsymbol{\beta}})] + \mathbf{X}\tilde{\boldsymbol{\beta}},$$

so this suggest creating pseudo-data from $\mathbf{y}$ as

$$\tilde{\mathbf{p}} = \boldsymbol{\Delta}^{-1}[\mathbf{y} - g^{-1}(\mathbf{X}\tilde{\boldsymbol{\beta}})] + \mathbf{X}\tilde{\boldsymbol{\beta}}, \tag{12.2}$$

whose approximate expectation is $\mathbf{X}\boldsymbol{\beta}$. From our assumed covariance model for $\mathbf{y}$, we have

$$\mathrm{var}(\tilde{\mathbf{p}}) = \boldsymbol{\Delta}^{-1}\mathbf{A}^{1/2}\mathbf{R}\mathbf{A}^{1/2}\boldsymbol{\Delta}^{-1} \equiv \mathbf{V}.$$

Treating $\tilde{\mathbf{p}}$ as data and $\mathbf{V}$ as a covariance matrix with unknown parameters, we can use maximum likelihood or restricted maximum likelihood (Chapter 8) to estimate the covariance parameters of $\mathbf{R}$, which is contained in $\mathbf{V}$, and, upon plugging them into $\mathbf{V}$ we have an estimated covariance matrix that we denote $\tilde{\mathbf{V}}$. From $\tilde{\mathbf{V}}$, we can get updated values,

$$\tilde{\boldsymbol{\beta}} = (\mathbf{X}^T\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{V}}^{-1}\tilde{\mathbf{p}},$$

and from updated $\tilde{\boldsymbol{\beta}}$, we can update the pseudo-data from (12.2). This iterative proceduce continues until convergence in $\tilde{\boldsymbol{\beta}}$ and the covariance parameters in $\mathbf{R}$. The algorithm only needs starting values for $\tilde{\boldsymbol{\beta}}$ which can be obtained by assuming data are independent and

using iteratively reweighted least squares (IRLS), which is the default parameter estimation method for nearly all generalized linear model software. Note that while IRLS converges to the maximum likelihood solution under most common conditions (Green 1984), the quasi-model approach described above does not converge toward any true likelihood, and, in fact, is not guaranteed to converge at all (Boykin et al. 2010; Kleinschmidt et al. 2001; Li et al. 2016).

12.x.5 Marginal MLE for Covariance Parameters

We would like to marginalize the distribution $[\mathbf{w}, \mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}] = [\mathbf{y}|\mathbf{w}, \boldsymbol{\phi}][\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}]$ over both $\mathbf{w}$ and $\boldsymbol{\beta}$ to obtain a distribution of the only the data and variance/covariance parameters,

$$[\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\theta}] = \int_{\mathbf{w}} \int_{\boldsymbol{\beta}} [\mathbf{w}, \mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{\theta}] d\boldsymbol{\beta} d\mathbf{w} = \int_{\mathbf{w}} [\mathbf{y}|\mathbf{w}, \boldsymbol{\phi}] \int_{\boldsymbol{\beta}} [\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}] d\boldsymbol{\beta} d\mathbf{w}.$$

When $[\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}]$ is Gaussian, $\int_{\boldsymbol{\beta}} [\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}] d\boldsymbol{\beta}$ is the likelihood for restricted maximum likelihood estimation (REML) (see Exercise 8.4),

$$[\mathbf{w}|\boldsymbol{\theta}] \equiv \int_{\boldsymbol{\beta}} [\mathbf{w}|\boldsymbol{\beta}, \boldsymbol{\theta}] d\boldsymbol{\beta} = \frac{1}{C_n} \exp[(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})],$$

where $C_n = \sqrt{2\pi^{(n-p)/2} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}| |\mathbf{X}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{X}|}$ and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \mathbf{w}$, and so we only need,

$$[\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\theta}] = \int_{\mathbf{w}} [\mathbf{y}|\mathbf{w}, \boldsymbol{\phi}][\mathbf{w}|\boldsymbol{\theta}] d\mathbf{w}.$$

Let us denote $\ell(\mathbf{w}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \log([\mathbf{y}|\mathbf{w}, \boldsymbol{\phi}][\mathbf{w}|\boldsymbol{\theta}])$, and consider $\int e^{\ell(\mathbf{w}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta})} d\mathbf{w}$. Let $\mathbf{g}$ be the gradient vector with ith element

$$g_i = \frac{\partial \ell(\mathbf{w}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta})}{\partial w_i},$$

and let $\mathbf{H}$ be the Hessian matrix with i, jth element,

$$H_{i,j} = \frac{\partial^2 \ell(\mathbf{w}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta})}{\partial w_i \partial w_j}.$$

Using the multivariate Taylor series expansion around some point $\mathbf{a}$,

$$\int_{\mathbf{w}} e^{\ell(\mathbf{w}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta})} d\mathbf{w} \approx \int_{\mathbf{w}} e^{\ell(\mathbf{a}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta}) + \mathbf{g}^T (\mathbf{w} - \mathbf{a}) + 1/2 (\mathbf{w} - \mathbf{a})^T \mathbf{H} (\mathbf{w} - \mathbf{a})} d\mathbf{w}.$$

Now, if $\mathbf{a}$ is a value for $\ell(\mathbf{a}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta})$ such that $\mathbf{g} = \mathbf{0}$, then

$$\int_{\mathbf{w}} e^{\ell(\mathbf{w}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta})} d\mathbf{w} \approx e^{\ell(\mathbf{a}; \mathbf{y}, \boldsymbol{\phi}, \boldsymbol{\theta})} \int_{\mathbf{w}} e^{-1/2 (\mathbf{w} - \mathbf{a})^T (-\mathbf{H}) (\mathbf{w} - \mathbf{a})} d\mathbf{w}.$$

Let $\mathbf{H_a}$ indicate $\mathbf{H}$ evaluated at $\mathbf{a}$. We know from the normalizing constant of a multivariate Gaussian distribution that

$$\int_{\mathbf{w}} e^{-1/2(\mathbf{w}-\mathbf{a})^T(-\mathbf{H_a})(\mathbf{w}-\mathbf{a})} d\mathbf{w} = (2\pi)^{N/2}|-\mathbf{H_a^{-1}}|^{1/2},$$

so

$$\int_{\mathbf{w}} e^{\ell(\mathbf{w};\mathbf{y},\boldsymbol{\phi},\boldsymbol{\theta})} d\mathbf{w} \approx e^{\ell(\mathbf{a};\mathbf{y},\boldsymbol{\phi},\boldsymbol{\theta})}(2\pi)^{N/2}|-\mathbf{H_a}|^{-1/2} = [\mathbf{y}|\mathbf{a},\boldsymbol{\phi}][\mathbf{a}|\boldsymbol{\theta}](2\pi)^{N/2}|-\mathbf{H_a}|^{-1/2}.$$

A marginal maximum likelihood estimator for $\boldsymbol{\phi}, \boldsymbol{\theta}$, given $\mathbf{a}$, is

$$\{\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}\} = \arg\max_{\boldsymbol{\phi},\boldsymbol{\theta}} [\log[\mathbf{y}|\mathbf{a},\phi] + \log[\mathbf{a}|\theta] - (1/2)\log(|-\mathbf{H_a}(\boldsymbol{\phi},\boldsymbol{\theta})|)] \tag{12.3}$$

where we drop terms that do not contain $\boldsymbol{\phi}$ or $\boldsymbol{\theta}$ and also show the dependence of $\mathbf{H_a}$ on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. The result (12.3) depends on finding $\mathbf{a}$ so that $\mathbf{g} = \mathbf{0}$. To acheive this, we use Newton-Raphson, conditional on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, which we describe next.

Assuming conditional independence of $\mathbf{y}$ on $\mathbf{w}$,

$$\log([\mathbf{y}|\mathbf{w},\boldsymbol{\phi}][\mathbf{w}|\boldsymbol{\theta}]) = \sum_{i=1}^{N} \log[y_i|w_i,\boldsymbol{\phi}] - \frac{1}{2}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}}) + C, \tag{12.4}$$

where C are terms that do not contain $\mathbf{w}$. Let $\mathbf{d}$ be the vector with ith component,

$$d_i \equiv \frac{\partial \log[y_i|w_i,\boldsymbol{\phi}]}{\partial w_i},$$

and

$$\frac{\partial[-\frac{1}{2}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})]}{\partial \mathbf{w}} = -\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{w} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}},$$

so the gradient of (12.4) is

$$\mathbf{g} = \mathbf{d} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{w} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{d} - \mathbf{P}_{\boldsymbol{\theta}}\mathbf{w},$$

where $\mathbf{P}_{\boldsymbol{\theta}}$ was defined in (8.2). For the Hessian, let $\mathbf{D}$ be a diagonal matrix with ith component,

$$D_{i,i} \equiv \frac{\partial^2 \log[y_i|w_i,\boldsymbol{\phi}]}{\partial w_i^2},$$

where all off-diagonal elements are zero because all second partials are 0 when $i \neq j$ due to conditional independence. A table of $\mathbf{d}_i$ and $\mathbf{D}_{i,i}$ for a few common distributions and link functions is given in Table 1.

Table 1: Distributions, inverse link functions, and first and second partial derivative with respect to w_i for the data model part of the loglikelihood.

Distribution	$\mu = g^{-1}(\eta)$	$\mathbf{d}_i$	$\mathbf{D}_{i,i}$
Binomial	$\mu = \frac{\exp(\eta)}{1+\exp(\eta)}$	$y_i - \frac{n_i \exp(w_i)}{1+\exp(w_i)}$	$-\frac{n_i \exp(w_i)}{(1+\exp(w_i))^2}$
Poisson	$\mu = \exp(\eta)$	$y_i - \exp(w_i)$	$-\exp(w_i)$
Negative Binomial	$\mu = \exp(\eta)$	$\frac{\phi(y_i - e^{w_i})}{\phi + e^{w_i}}$	$-\frac{\phi e^{w_i}(\phi + y_i)}{(\phi + e^{w_i})^2}$

In Table 1, the alternative parameterization for the negative binomial was used,

$$[y|\mu,\phi] = \frac{\Gamma(y+\phi)}{\Gamma(\phi)y!}\left(\frac{\mu}{\mu+\phi}\right)^{y}\left(\frac{\phi}{\mu+\phi}\right)^{\phi},$$

where $\mathrm{E}(y) = \mu$ and $\mathrm{var}(y) = \mu + \mu^2/\phi$.

Next, notice that

$$\frac{\partial^2[-\frac{1}{2}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})]}{\partial\mathbf{w}\partial\mathbf{w}^T} = -\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}(\mathbf{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} = -\mathbf{P}_{\boldsymbol{\theta}},$$

and so the Hessian of (12.4) is

$$\mathbf{H} = \mathbf{D} - \mathbf{P}_{\boldsymbol{\theta}}. \tag{12.5}$$

Conditional on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, a Newton-Raphson update is,

$$\mathbf{w}^{[k+1]} = \mathbf{w}^{[k]} - \mathbf{H}^{-1}\mathbf{g},$$

and upon convergence we set $\mathbf{a} = \mathbf{w}$ in (12.3) for any evaluation of the likelihood for given $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Notice that this makes the marginal MLE doubly iterative, as we solve for $\mathbf{a}$ while optimizing for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. It is possible to use other maximization routines, such as the EM algorithm, but, generally, the Newton-Raphson algorithm converges rapidly (often around 10 iterations in our experience). However, on occasion, the stepsize needs to be adjusted so that $\mathbf{g}$ does not diverge. For example, it is easy and fast to check $\mathbf{g}^{[k+1]} = \mathbf{d} - \mathbf{P}_{\boldsymbol{\theta}}\mathbf{w}^{[k+1]}$, and if $\mathbf{g}^{[k+1]}$ is "larger" than $\mathbf{g}$ by some criterion (e.g., largest or average element of $\mathbf{g}$), then take

$$\mathbf{w}^{[k+1]} = \mathbf{w}^{[k]} - \alpha\mathbf{H}^{-1}\mathbf{g},$$

where $0 < \alpha < 1$. In the simulation below, we check $\mathbf{g}^{[k+1]}$ in the manner described above, and set $\alpha = 0.1$ if the largest element of $\mathbf{g}^{[k+1]}$ is larger than the largest element of $\mathbf{g}$. The advantage of using Newton-Raphson is that it provides $\mathbf{H}$, which is required to make adjustments to estimation and prediction, as we describe next.

12.x.6 Estimation of Fixed Effects and Prediction

In order to estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, it was necessary to optimize the likelihood for $\mathbf{w}$, which we called $\mathbf{a}$, using Newton-Raphson, for each evaluation of the likelihood. Upon convergence in estimating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, we also have optimized values for $\mathbf{w}$, and let us denote them as $\hat{\mathbf{w}} = \mathbf{a}$. Also, recall that, in contrast to Bonat and Ribeiro Jr (2016), we integrated over $\boldsymbol{\beta}$, so an estimator is needed.

An obvious estimator of $\boldsymbol{\beta}$ is to consider $\hat{\mathbf{w}}$ as if they were observed data, and then use the generalized least squares estimator $\hat{\boldsymbol{\beta}} = \mathbf{B}\hat{\mathbf{w}}$, where $\mathbf{B} = (\mathbf{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$. However, $\mathbf{w}$ contains predictions of an unobserved, latent random variables, rather than observed values.

We will need to make some adjustments in order to estimate the variance of $\hat{\boldsymbol{\beta}}$. It will be convenient to condition on $\mathbf{w}$ as if we had observed them, and then we can use

$$\text{var}(\mathbf{B}\hat{\mathbf{w}}) = \text{E}_{\mathbf{w}}[\text{var}(\mathbf{B}\hat{\mathbf{w}}|\mathbf{w})] + \text{var}_{\mathbf{w}}[\text{E}(\mathbf{B}\hat{\mathbf{w}}|\mathbf{w})].$$

We will assume that $\hat{\mathbf{w}}$ is unbiased for $\mathbf{w}$, i.e., $E(\hat{\mathbf{w}}|\mathbf{w}) = \mathbf{w}$, so $\text{var}_{\mathbf{w}}[\text{E}(\mathbf{B}\hat{\mathbf{w}}|\mathbf{w})] = \mathbf{B}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\mathbf{B}^T$, which simplifies to $\mathbf{C}_{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}$, the usual variance-covariance matrix of fixed effects when using generalized least squares (5.2). We can use the inverse of the observed Fisher-Information to obtain $\text{var}(\hat{\mathbf{w}}|\mathbf{w})$, which is $-\mathbf{H}^{-1}$. Notice that this depends on $\mathbf{w}$ through the diagonal elements of $\mathbf{D}$ in (12.5). As an approximation, we will replace $\mathbf{w}$ with their predicted values $\hat{\mathbf{w}} = \mathbf{a}$, and so an estimator of the covariance matrix of fixed effects is

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \mathbf{B}(-\mathbf{H}_{\mathbf{a}}^{-1})\mathbf{B}^T + \mathbf{C}_{\boldsymbol{\beta}}. \tag{12.6}$$

We proceed in a similar fashion for predictions and prediction variances. Suppose that we want to predict the latent spatial variable W_j at unobserved locations for $j = 1, \ldots, J$, and we denote this vector of random variables as $\mathbf{u}$. Note that from (9.4), we can write $\hat{\mathbf{u}} = \boldsymbol{\Lambda}\hat{\mathbf{w}}$ where $\boldsymbol{\Lambda} = \mathbf{X}_{\mathbf{u}}^T\mathbf{B} + \boldsymbol{\Sigma}_{\mathbf{wu}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Sigma}_{\mathbf{wu}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}\mathbf{B}$ and where $\boldsymbol{\Sigma}_{\mathbf{wu}}^T$ is the covariance matrix between $\mathbf{w}$ and $\mathbf{u}$ at the observed and unobserved locations, respectively. As for fixed effects, we need an estimator of the mean-squared-prediction errors, also called the prediction variance, which is $\text{var}(\hat{\mathbf{u}} - \mathbf{u}) = \text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$. Now, if we had observed $\mathbf{w}$, rather than predicting $\hat{\mathbf{w}}$, then $\text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$ is given by (9.5), but again we need to make some adjustments because we are estimating $\hat{\mathbf{w}}$. Conditioning on $\mathbf{w}$ and $\mathbf{u}$, we have

$$\text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}) = \text{E}_{\mathbf{w},\mathbf{u}}[\text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w}, \mathbf{u})] + \text{var}_{\mathbf{w},\mathbf{u}}[\text{E}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w}, \mathbf{u})].$$

As we did earlier, we will assume that $\hat{\mathbf{w}}$ is unbiased for $\mathbf{w}$, so $\text{E}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w}, \mathbf{u}) = \boldsymbol{\Lambda}\mathbf{w} - \mathbf{u}$, and the variance of this will be the same as if we had observed $\mathbf{w}$, so $\text{var}_{\mathbf{w},\mathbf{u}}(\boldsymbol{\Lambda}\mathbf{w} - \mathbf{u})$ is given by (9.5). Conditionally, $\text{var}_{\hat{\mathbf{w}}}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$ does not depend on $\mathbf{u}$, so $\text{E}_{\mathbf{w},\mathbf{u}}[\text{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w}, \mathbf{u})] = \text{E}_{\mathbf{w}}[\boldsymbol{\Lambda}(-\mathbf{H}_{\mathbf{w}}^{-1})\boldsymbol{\Lambda}^T]$, and, to take expectation, we simply replace $\mathbf{w}$ in $\mathbf{H}$ with its estimator $\mathbf{a}$. Putting them together, we obtain

$$\widehat{\text{var}}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}) = \boldsymbol{\Lambda}(-\mathbf{H}_{\mathbf{a}}^{-1})\boldsymbol{\Lambda}^T + \boldsymbol{\Sigma}_{\mathbf{uu}} - \boldsymbol{\Sigma}_{\mathbf{wu}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\mathbf{wu}} + \mathbf{K}\mathbf{C}_{\boldsymbol{\beta}}\mathbf{K}^T, \tag{12.7}$$

where $\boldsymbol{\Sigma}_{\mathbf{uu}}^T$ is the covariance matrix of $\mathbf{u}$ and $\mathbf{K} = \mathbf{X}_{\mathbf{u}}^T - \boldsymbol{\Sigma}_{\mathbf{wu}}^T\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}$.

How well do all of these approximations work? We will illustrate with a simulation so that we know the true values. We created a square grid of 20×20 locations equally spaced on a $(0, 1) \times (0, 1)$ domain. Let $\mathbf{X} = \mathbf{1}$ and a single overall mean parameter $\beta_0 = 2$. We generated $\mathbf{w}$ from a spatially-autocorrelated model with an exponential autocorrelation function (Table 6.2) where $\alpha = 1$, and the autocovariance model was $\text{cov}(w(\mathbf{s}), w(\mathbf{s} + r)) = \exp(-r) + 0.0001\mathcal{I}(r = 0)$, where $\mathcal{I}(\cdot)$ is the indicator function, equal to one if its argument is true, otherwise it is zero. The 400 simulated $\mathbf{w}$ values are shown in Figure 1B. Conditional on the $\mathbf{w}$, at each spatial location we independently simulated a Poisson random variable with mean equal to $\exp(w_i)$, which are shown in Figure 1A.

Using the values in Figure 1A, we assumed an unknown mean and covariance function $\text{cov}(w(\mathbf{s}), w(\mathbf{s}+r)) = \sigma_1^2 \exp(-r/\alpha) + \sigma_0^2 \mathcal{I}(r=0)$, where recall that σ_1^2 is the partial sill, α is the range parameter, and σ_0^2 is the nugget effect. Optimizing the likelihood for (12.3) for $\boldsymbol{\theta} = (\sigma_1^2, \alpha, \sigma_0^2)$ we obtain the values $\hat{\sigma}_1^2 = 0.950$, $\hat{\alpha} = 0.894$, and $\hat{\sigma}_0^2 = 0.046$. The likelilhood surface for σ_1^2 and α is shown in Figure 1C. A pronounced ridge shows the positive association in the likelihood between σ_1^2 and α, which we saw in Chapter 8; e.g., Figure 8.4. The estimation of $\boldsymbol{\theta} = (\sigma_1^2, \alpha, \sigma_0^2)$ also provided $\hat{\mathbf{w}}$, which are shown in Figure 1D, and it appears that we were able to recover the spatial patterning of the true simulated $\mathbf{w}$ quite well. A similar example using Bernoulli data is provided in the `R` code that accompanies this book.

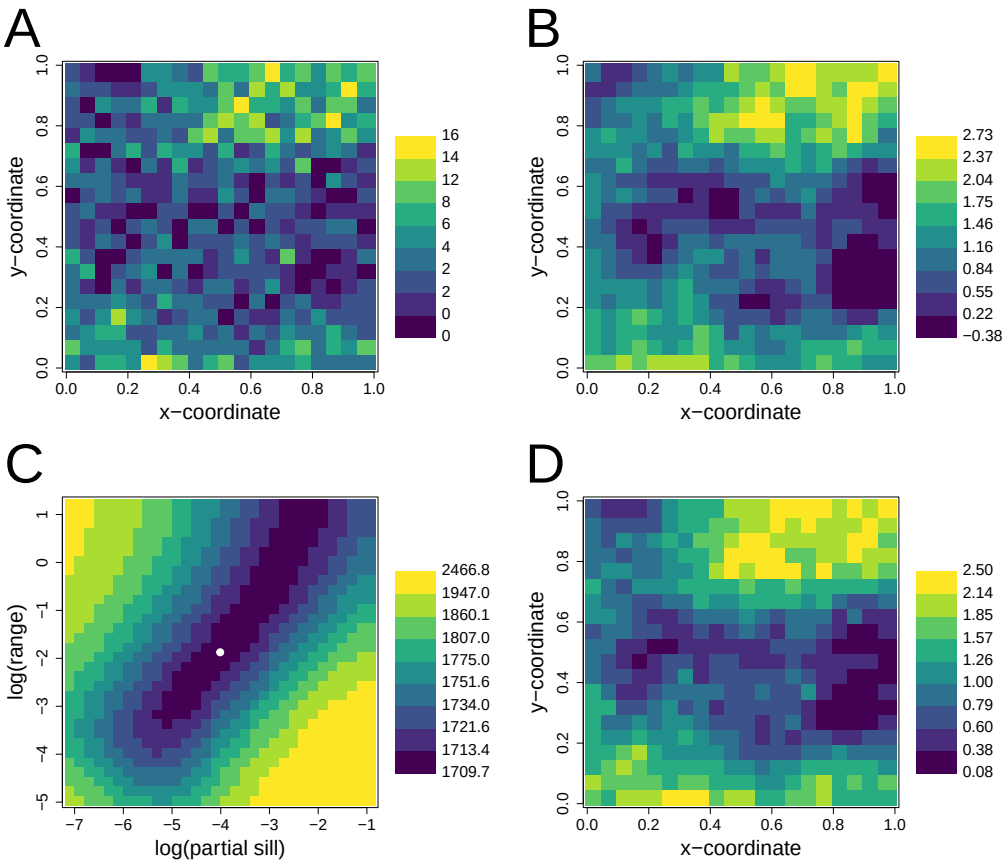

Figure 1: Estimation for simulated data. A. Simulated count data using the model described in the text. B. The true simulated $\mathbf{w}$ values. C. The likelihood surface of the simulated data. The white circle shows the estimated value. D. The estimated $\hat{\mathbf{w}}$ values.

Of course, this is just one simulation. How does it work on average? Are the estimators and predictors unbiased, and are we estimating their variance correctly so that they have the proper confidence and prediction interval coverage? In order to answer these questions, we did a computer simulation experiment. Here, we created 200 locations randomly on

the unit square. We used the same autocovariance model as above, $\text{cov}(w(\mathbf{s}), w(\mathbf{s}+r)) = \exp(-r) + 0.0001\mathcal{I}(r=0)$. For the mean structure, we used

$$\text{E}(w_i) = \beta_0 + \beta_1 x_i + \beta_2 \tau_i + \beta_3 (x{:}\tau)_i$$

where x_i were randomly simulated from $\text{N}(0,1)$, τ_i were randomly simulated bernoulli variables with probability $p = 0.5$, and $(x{:}\tau)_i$ was in interaction between the normally-distributed and bernoulli-distributed explanatory variables. We set $\beta = (0.5, 0.5, -0.5, 0.5)$. We also created 100 prediction locations where we used a 10×10 square grid of equally-spaced prediction points throughout the unit square. Explanatory variables were also simulated at the prediction locations, and so 300 w_i values were simulated from $\text{N}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$. We then created the observed data as counts from a Poisson distribution conditional on the $\mathbf{w}$, where at each spatial location we independently simulated the Poisson random variable with mean equal to $\exp(w_i)$. We simulated 2000 data sets to assess bias and confidence/prediction interval coverage.

For each simulated data set, we first estimated the covariance parameters using (12.3), using an exponential autocorrelation model as before. Then, we used the estimated covariance parameters as plugin values for the autocovariance model to obtain $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$, and along with the estimated $\hat{\mathbf{w}}$, we estimated fixed effects $\hat{\boldsymbol{\beta}} = \mathbf{B}\hat{\mathbf{w}}$. To estimate bias, we took the average of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ over all 2000 simulated data sets. We also formed 90% confidence intervals as $\hat{\boldsymbol{\beta}} \pm 1.645\widehat{\text{se}}(\hat{\boldsymbol{\beta}})$, where $\widehat{\text{se}}(\hat{\boldsymbol{\beta}})$ were the square roots of the diagonal elements of (12.6). Over the 2000 simulations, we computed the proportion of times that the confidence interval contained the true value. If we are estimating the variances of $\hat{\boldsymbol{\beta}}$ well, the coverage should be close to 90%. We also computed the confidence interval coverage based on the naive unadjusted $\mathbf{C}_{\boldsymbol{\beta}}$.

The results are shown in Table 2, where we see that there is very little bias in estimating any of the parameters in $\boldsymbol{\beta}$. The confidence interval coverage for β_0 is slightly low, but the confidence interval coverages for β_1 - β_3 are very close to 90% when using (12.6), but they are much too short when using $\mathbf{C}_{\boldsymbol{\beta}}$.

We also used the estimated covariance parameters in $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}}$ and the estimated $\hat{\mathbf{w}}$ to make predictions at all 100 values for each simulated data set using $\hat{\mathbf{u}} = \boldsymbol{\Lambda}\hat{\mathbf{w}}$. To estimate bias, we took the average of $\hat{\mathbf{u}} - \mathbf{u}$ for each simulated data set, where recall that $\mathbf{u}$ contains 100 simulated values, and then averaged those across the 2000 simulated data sets. We also formed 90% prediction intervals as $\hat{\mathbf{u}} \pm 1.645\widehat{\text{se}}(\hat{\mathbf{u}})$, where $\widehat{\text{se}}(\hat{\mathbf{u}})$ were the square roots of the diagonal elements of (12.7). Over the 2000 simulations, we computed the proportion of times that the prediction intervals contained the true values, which should be about 90%. Table 2 gives the prediction results, which show little indication of bias, and coverage was very close to 90%.

Table 2: Bias and coverage for estimation of fixed effects $\boldsymbol{\beta}$ and for prediction of $\mathbf{u}$ at unobserved locations. Coverage is for 90% confidence and prediction intervals, and CI90_c used the corrected versions in (12.6) and (12.7), while CI90_u shows coverage for the uncorrected standard-error estimator based on $\mathbf{C}_{\boldsymbol{\beta}}$ and the uncorrected prediction standard errors using (9.5).

effect	bias	CI90_u	CI90_c
β_0	-0.006	0.865	0.875
β_1	-0.003	0.360	0.905
β_2	0.002	0.310	0.898
β_3	-0.004	0.331	0.896
$\hat{\mathbf{u}}$	0.034	0.692	0.894

References

Besag, J. (1974), "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *Journal of the Royal Statistical Society, Series B*, 36, 192–236.

Bonat, W. H. and Ribeiro Jr, P. J. (2016), "Practical likelihood analysis for spatial generalized linear mixed models," *Environmetrics*, 27, 83–89.

Boykin, D., Camp, M. J., Johnson, L., Kramer, M., Meek, D., Palmquist, D., Vinyard, B., and West, M. (2010), "Generalized linear mixed model estimation using PROC GLIMMIX: results from simulations when the data and model match, and when the model is misspecified," in *Conference on Applied Statistics in Agriculture*, pp. 137–156 DOI:10.4148/2475–7772.1064.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Christensen, O. F. (2004), "Monte Carlo maximum likelihood in model-based geostatistics," *Journal of Computational and Graphical Statistics*, 13, 702–718.

Clayton, D. and Kaldor, J. (1987), "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping," *Biometrics*, 43, 671–681.

Cressie, N. A. C. (1993), *Statistics for Spatial Data, Revised Edition*, New York: John Wiley & Sons.

Diggle, P. J., Tawn, J. A., and Moyeed, R. A. (1998), "Model-based geostatistics (with discussion)," *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 47, 299–326.

Evangelou, E., Zhu, Z., and Smith, R. L. (2011), "Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation," *Journal of Statistical Planning and Inference*, 141, 3564–3577.

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.

Green, P. J. (1984), "Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives," *Journal of the Royal Statistical Society: Series B (Methodological)*, 46, 149–170.

Kleinschmidt, I., Sharp, B. L., Clarke, G. P. Y., Curtis, B., and Fraser, C. (2001), "Use of generalized linear mixed models in the spatial analysis of small-area malaria incidence rates in KwaZulu Natal, South Africa," *American Journal of Epidemiology*, 153, 1213–1221.

Li, L., Brumback, B. A., Weppelmann, T. A., Morris Jr, J. G., and Ali, A. (2016), "Adjusting for unmeasured confounding due to either of two crossed factors with a logistic regression model," *Statistics in Medicine*, 35, 3179–3188.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000), "WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility," *Statistics and Computing*, 10, 325–337.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models, 2nd Edition*, Chapman & Hall Ltd.

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized linear models," *Journal of the Royal Statistical Society, Series A: General*, 135, 370–384.

Wolfinger, R. and O'Connell, M. (1993), "Generalized linear mixed models: A pseudo-likelihood approach," *Journal of Statistical Computation and Simulation*, 48, 233–243.