CARLETON UNIVERSITY

SCHOOL OF
MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: *Fitting Spatially Correlated Data with a Poisson Conditional Autoregressive Model*

AUTHOR: Ventseslav Yordanov

SUPERVISOR: Prof. Jason D. Nielsen

DATE: 29.11.2019

Table of Contents

Table of Figures

§ 1- Introduction

1.1 Preface

Law enforcement officials and criminal psychologists have emphasized that the crime levels of a city are subject to variation according to two predominant factors: (1) socioeconomic environment and (2) geographic region. From a mathematical point of view, a regression framework can be used to statistically model violent crime data and make inference about crime levels in various regions of a city.

However one must pay special attention to the second factor – since the varying violent crime rates across a city may be an indication of spatial dependency among the city's communities. This spatial dependency poses an obstacle, as it violates the assumption of independence among observed data that many classical regression techniques rely on.

It is the purpose of this investigation to demonstrate that a spatial autoregressive model will remedy the problem of dependency among data, as well as provide a quantitative measure of the strength of these spatial relationships. We illustrate this by fitting a (spatial) Poisson intrinsic conditional autoregressive (ICAR) model to homicide data of the city of Chicago, for the year 2012.

In doing so, we will test the degree to which the ICAR model explains the hypothesized spatial dependency between neighbourhoods, while also estimating the impact of several socioeconomic covariates on homicide levels across the city's 77 neighbourhoods.

Among the many candidates for cities whose crime rates are high, the choice of Chicago, IL is justified with two reasons. In 2012, the city earned itself the title of murder capital of the United States, with 515 homicides, while having a population of 3 million which is considerably lower than that of New York, NY, and Los Angeles, CA – both of which had less homicides that year. During the early 2000s the city also saw an aggressive rehousing plan, which was meant to address the problem of gang infestation in many high-rise projects. This resulted in segregated communities in the South, West and East parts of the city, where crime is much higher than anywhere else.

1.2 Guide for the Reader

The reader should note that in the first two sections, there are semi-detailed reviews of concepts in linear regression models (General Linear Models, and General linear Mixed Models). Their role is fundamental in understanding spatial regression, and should be treated as prerequisite knowledge.

Hence, for better understanding the topics in this paper, it is strongly recommended that the reader goes through sections 2 and 3 prior to reading through sections 4 and 5, should they need to recall key concepts in regression.

In Appendix A, the reader can find concepts of regression parameter estimation, as well as the statements and proofs of several important theorems that we will refer to throughout this paper.

§ 2 – Generalized Linear Models Introduction

Our analysis will primarily be focusing on integer-valued count data, and as such we will be working with a type of Poisson generalized linear mixed model, conditional on some randomized effect which inherently adds a hypothesized spatial correlation structure to our model. This randomized effect, $\boldsymbol{\phi}$, follows a multivariate normal distribution. To be more specific, as mentioned in the Preface, we will be building a conditional auto-regressive model (CAR) to fit our data, and we'll discuss this in detail in later chapters. First we will review some of the underlying theory of generalized linear models.

§ 2.1 – Generalized Linear Models Overview

Def 2.1. A generalized linear model (GLM) refers to a broader class of *linear regression models* in which the response vector $\boldsymbol{Y}$ comes from an *exponential family of distributions*, with mean $\boldsymbol{\mu}$, and its relationship to the covariates becomes linear through *a link function* which need not be the identity map. We also have that the errors need not be normally distributed. The model equation is of the form:

$$h(\boldsymbol{\mu}) = [g(\mu_1), \dots, g(\mu_n)]^T = \mathbf{X\beta}. \qquad (2\text{-}1)$$

where $g(.)$ is the link function such that $g(\mu_i) = \boldsymbol{X_i^T\beta}$ and,

- $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$ is the $p+1$ dimensional vector of regression coefficients
- $\boldsymbol{X} = (\mathbf{1}\ X_1 \dots X_p)$ is an $n\ x\ (p+1)$ design matrix, consisting of n-dimensional column vectors $\boldsymbol{X_1}, \dots, \boldsymbol{X_p}$ which represent the p covariates and the $\mathbf{1}$ is a n-dimensional vector of 1's so that an intercept β_0 term can be included in the model
- $\boldsymbol{\mu}$ is the vector of means for the exponential distribution of the response observations.

Def 2.2. If Y is a random variable coming from an exponential family of distributions, then its probability density function (or mass function if discrete) has the form:

$$f_y(y\,|\theta,\Phi) = \exp\left[\frac{y\theta - b(\theta)}{\alpha(\Phi)} + d(y,\Phi)\right] \qquad (2\text{-}2)$$

Where $\alpha(\Phi)$, $b(\theta)$, and $d(y,\Phi)$ are some known functions, and

- θ is our parameter of interest
- Φ is a nuisance parameter

Note: When talking about generalized linear models, we must make the distinction from a 'general linear model' which is simply an extension to the univariate case of simple linear regression, where our assumption is the response variables comes from a Gaussian distribution.

There are three parts to any GLM:

- *The response component* – which is characterized by the distribution of the response variable in question. In our case the random component is the 77-dimensional vector of integers which follows a Poisson distribution.
- *The fixed component* – which defines the covariates in the model. More specifically, it defines the linear combination of covariates and their regression weights (as coefficients), which is what we call the *linear predictor* of the model.
- *The link function* – which characterizes the relationship or "link" between the random and fixed components of the model. This function being strictly monotone and continuously differentiable.

The assumptions typically made for GLM's are the following:

- The response data are independently distributed – meaning that each individual data point in the response vector is independent of the others

- Unlike in ordinary linear regression models, the response vector does not need to follow a Gaussian distribution; but it is typically assumed that it follows a distribution from the exponential family (i.e. Poisson, Gamma, Binomial etc.)
- We need not satisfy the homogeneity of variance.
- We do not assume a linear relationship between the response variable and its covariates, however a linear relationship is assumed between the *transformed* response (in terms of the link function) and the covariates (i.e. $g(\boldsymbol{\mu}) = \ln(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ in the case of Poisson regression).
- Estimation of the model parameters is done through maximum likelihood estimation (MLE) rather than ordinary least squares estimation.

§2.2 – Fitting Generalized Linear Models in R (Poisson Regression)

In a Poisson regression model, the response vector **Y** is a vector of integer-valued data points Y_i, which all follow a Poisson distribution with some rate μ_i. This rate is determined by the linear predictors. The link function which relates the rate μ_i and the linear predictors is a *log link* function. Formally, the Poisson regression model for observation i, $i = 1, \ldots, n$ is:

$$P(Y_i = y_i \mid \boldsymbol{X_i}, \boldsymbol{\beta}) = \frac{e^{-\exp[\boldsymbol{X_i^T\beta}]} \exp[\boldsymbol{X_i^T\beta}]}{y_i!} \tag{2-3}$$

where our log link function is given by $g(\mu_i) = \ln(\mu_i) = \boldsymbol{X_i^T\beta}$, consequently $\mu_i = \exp(\boldsymbol{X_i^T\beta})$

The likelihood function for a Poisson regression model where the sample size is n is given by:

$$L(\beta \mid Y_i, X_i) = \prod_{i=1}^{n} P(Y_i = y_i \mid \boldsymbol{X_i}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{e^{-\exp[\boldsymbol{X_i^T\beta}]} \exp[\boldsymbol{X_i^T\beta}]}{y_i!}$$

Taking the log of the likelihood function, we obtain

$$l(\beta) = \sum_{i=1}^{n} y_i X_i \boldsymbol{\beta} - \sum_{i=1}^{n} \exp\{\boldsymbol{X_i^T\beta}\} - \sum_{i=1}^{n} \log(y_i!)$$

$$= \sum_{i=1}^{n} y_i X_i \boldsymbol{\beta} - \exp(\boldsymbol{X_i^T\beta}) - \log(y_i!)$$

Taking the partial derivatives w.r.t **β** and setting to 0 gives

$$\frac{\partial}{\partial\beta}\left(\sum_{i=1}^{n} y_i X_i \boldsymbol{\beta} - \exp\left(\boldsymbol{X_i^T\beta}\right) - \log(y_i!)\right) = 0$$

which does not have a closed form solution, and so, **β** is estimated via iteratively re-weighted least squares (*IRWLS*) technique, which is outlined in Appendix A. R uses this technique to obtain estimates for **β**. Since, in a Poisson model the variance is equal to the mean, obtaining $\widehat{\boldsymbol{\beta}_{MLE}}$ is all we need for fitting data.

Using the ML estimates obtained via IRWLS, we can fit a GLM to an appropriate dataset using R. The software package fits generalized linear models via the built in *glm()* function, where the user must pass in desired parameters in order to specify the type of generalized model. The general form of the function, followed by a description of its arguments is:

```
glm(formula, family = "family_type"(link = linkfunction), data=mydat)
```

where

- *formula* is an object that belongs to the class "formula". Typically the user specifies the relation which they will use to fit the data; i.e. y ~ x1 + x2 + x3 where y is a count and x1, x2, x3 are continuous predictors.
- *family* is the family of distributions specified by the user, which characterize the error and link function which are used in the model; i.e. family=poisson().
- *link* is the user specification for an appropriate link function with respect to the model. We refer to the table below for a summary of conventional link functions.
- *data* is an optional data frame, which contains the variables in the model. If variables not found in *data*, R takes them from the environment that glm() is called from.

The summary() command gives output of parameter estimates, and tells us if they are significant (i.e. should we keep covariates in the model, how much information about the data remains unexplained by the model, the goodness of fit, etc.). Once the model is fit and estimates are obtained, inference can be made about the data which we're analyzing. This includes building confidence intervals around our coefficients, our mean, and our variance

Figure 2.2.1: The table below provides a useful summary of conventional GLMs

Model	**Random Component**	**Link Function**	**Systematic Component**
Simple Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
Logistic Regression	Binomial	Logit	Mixed
Loglinear Regression	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

source: https://newonlinecourses.science.psu.edu/stat504/node/216/

***Figure 2.2.1*:** Another useful table which summarizes arguments in R with respect to various models used in the *glm()* function

Family	**Default Link Function**
binomial	(link = "logit")
gaussian	(link = "identity")
gamma	(link = "inverse")
inverse gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

source: https://www.statmethods.net/advstats/glm.html

§2.3 – Interpretation of a generalized linear model

Perhaps one of the most important things in statistical analysis is being able to properly interpret our model and its parameters – and in the regression case – the results after we have fit the model. Without this understanding, we are unable to properly make any sort of inference based on a theoretical model, and ultimately, are unable to extract valuable information from the data. As such we will first elaborate on what the parameters of a model are really telling us about our data. To that end, the unknown parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)$ in a generalized linear model represent the effect our p covariates $\boldsymbol{X_j}, j = 1, \dots, p$ have on the response variables $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$. Typically a parameter value β_j that is large in magnitude implies that the corresponding predictor variable has a considerably strong impact on the response.

Before concluding that a particular predictor variable explains our responses, we check if the relationship between the given covariate and response is significance by testing the hypothesis

$$H_o: \beta_j = 0 \; vs. \; H_\alpha: \beta_j \neq 0$$

As a consequence of MLE theory and the delta-method, we use the asymptotically normal test-statistic $z = \frac{\widehat{\beta_j} - \beta_j}{s_{\beta_j}/\sqrt{n}}$, If p-value $2(Z > |z|) < 0.05$ then we reject the null hypothesis that $\beta_j = 0$ and keep the associated predictor variable, concluding that it is significant and it explains part of our response-vector. R computes the p-values for each estimate, and tells us whether its associated covariate is significant or not based on the two sided test above.

Lastly, using the residuals returned by R, we can test to see if the model accurately fits our data. For GLM's, the residuals come into play through the link function, and are fully described by the given exponential distribution of the response variables. Typically, there are two types of residuals to look at in generalized linear models: *deviance residuals* and *Pearson residuals*.

Def 2.3. Pearson Residuals are based on subtracting off the estimated mean of the given distribution of the model from the observed response values, and dividing by the standard deviation. Computed as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{Var(\hat{\mu}_i)}} \tag{2-4}$$

Def 2.4. Deviance Residuals are based on measuring the contribution of each observed data point to the likelihood function of the model. They are typically preferred over (2-4). Computed as:

$$d_i = \pm\sqrt{2\{\log[f_Y(y_i|\hat{\mu}_i)] - l(\hat{\boldsymbol{\beta}})\}} \tag{2-5}$$

where $log\left(f_Y(y_i|\hat{\mu}_i)\right)$ is the log-likelihood function of *saturated model* – the model which has the most general mean structure where μ_i are unconstrained, and $l(\hat{\boldsymbol{\beta}})$ is the log-likelihood function of the proposed model – the GLM in question which we use to fit our data. In general, we usually consider the overall deviance, which is given by

$$D = \sum_{i=1}^{n} d_i^2$$

and we have that $D \sim \chi^2_{N-p}$ where *N* is the number of parameters in the saturated model and *p* is the number of parameters in the proposed model.

For a Poisson GLM, we have that the Pearson residuals are given by $r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$ and the deviance residuals are given by: $d_i = \pm\sqrt{2\left\{y_i \log\left({y_i}/{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right\}}$ and the overall deviance $D = 2\sum_i y_i \log\left({y_i}/{\hat{\mu}_i}\right)$. If the proposed model poorly fits the data, then *D* will be larger than predicted by the χ^2_{N-p} distribution.

§2.4 – Example using R to fit Poisson GLM to real data

We now wrap up the discussion regarding fitting and interpretation of generalized linear models, through an example closely related to our crime analysis. We use R to fit a Poisson regression model for our homicide count data for 2012 against 5 socioeconomic factors. In this case, our model components are defined as follows:

- Our response vector **Y** is a 77-dimensional vector which contains the homicide counts for each of the city's 77 districts for the year of 2012.
- Our predictors – in the form of a 77 x 5 design matrix **X** – include respectively, for each of the city's 77 districts:

  → The percent of persons aged 25+ without a high school diploma ($\mathbf{X_1}$)

  → The percent of households below the poverty line ($\mathbf{X_2}$)

  → The percent of persons aged 16+ who are unemployed ($\mathbf{X_3}$)

  → The average per capita income ($\mathbf{X_4}$)

  → The hardship index ($\mathbf{X_5}$), a value between 0-100, calculated as the median of the above socioeconomic factors, as well as crowded housing and government dependency.

Because our response vector $\boldsymbol{Y}$ contains strictly integer-valued data about the number of events (in our case homicides) per some unit (in our case city district), it is appropriate to assume that each of the response entries, Y_i, follow a Poisson(μ_i) distribution. The estimation of these 6 parameters would give us quantitative insight on the impact each of the five socioeconomic factors has on homicides.

Consequently, the model is: $\mu_i = e^{\boldsymbol{X_i^T \beta}}$ and the link function is $g(\mu_i) = \ln e^{\boldsymbol{X_i^T \beta}} = \boldsymbol{X_i^T \beta}$

Where: $$f_y(y_i \mid X_i, \beta) = \frac{e^{-\exp[\boldsymbol{X_i^T \beta}]} \exp[\boldsymbol{X_i^T \beta}]}{y_i!}$$

We implement the following commands in the R console to fit the model and respective data:

Figure 2.4.1. Code used to load and fit Chicago Homicide data in R

```
y <- as.vector(read.table("homicides.txt", header = FALSE)) #load homicide data
X <-as.matrix(read.table("factors.txt". header = FALSE))       #load socioeconomic data

#extract each column of our data matrix
x1 <- X[,-c(2,3,4,5)]
x2 <- X[,-c(1,3,4,5)]
x3 <- X[,-c(1,2,4,5)]
x4 <- X[,-c(1,2,3,5)]
x5 <- X[,-c(1,2,3,4)]

#fit the glm model and return output
glm_pois <- glm(y ~ x1+x2+x3+x4+x5, family = poisson(link="log"))
summary(glm_pois, correlation = TRUE)
```

Figure 2.4.2. Output of Poisson GLM fit for Chicago Homicide data

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.2702  -2.1812  -0.5665   1.2159   6.1876

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.691e-01  3.856e-01   0.698   0.4853
x2          -2.721e-02  1.067e-02  -2.550   0.0108 *
x3          -1.872e-02  9.232e-03  -2.027   0.0426 *
x4           1.533e-02  1.261e-02   1.216   0.2241
x5           1.064e-05  8.478e-06   1.255   0.2096
x6           3.773e-02  9.318e-03   4.049 5.14e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 615.02  on 76  degrees of freedom
Residual deviance: 419.01  on 71  degrees of freedom
AIC: 648.52

Number of Fisher Scoring iterations: 5

Correlation of Coefficients:
   (Intercept) x2    x3    x4    x5
x2 -0.05
x3  0.24        0.59
x4 -0.19        0.65  0.08
x5 -0.91       -0.21 -0.38 -0.05
x6 -0.39       -0.85 -0.75 -0.58  0.58
```

Looking at the summary above, the first piece of information returned to us by R from the *summary()* command are the *deviance residuals*. If our model is specified correctly, then the deviance residuals should be approximately normally distributed. We observe a little bit of skewedness in our case as the median is -0.566, and in the ideal case it should be 0. Next, looking at the coefficients table, the first column gives us the estimated values of the Poisson regression coefficients – i.e. it gives us $\widehat{\boldsymbol{\beta}}$. In the columns to the right, we see the associated std. error for our estimates, the z-value used for testing if the parameter is 0 (Wald's statistic), and the p-value which indicates significance of the estimate (i.e. the coefficients for predictors $\boldsymbol{X_2}$ and $\boldsymbol{X_3}$ in our fit imply that the percent of households below the poverty line, and the percent of persons aged 16+ who are unemployed are significant socioeconomic factors which influence the homicide rate in various districts of the city). The *residual deviance* may be used to perform a goodness of fit test for our model. This residual deviance is a measure of how well the model fits our data; it is the difference between the deviance of our model and the maximum deviance of an ideal theoretical model, where predictors are identical to the observed. Consequently, if the difference between the observed values and their predicted means is small, then we have a model that fits well. Likelihood theory tells us that deviance can be derived as the likelihood ratio test comparing our model to an ideal one, and so under the assumption that our model is correctly specified, the *residual deviance* (not to be confused with the deviance residuals) should follow a chi-squared distribution with degrees of freedom equal to *n-p*. In our summary, the residual deviance is 419.01, there are a total of 77 observations and 6 parameters, so the residual degrees of freedom are 71. We compute the p-value for the chi-squared goodness of fit test as follows:

Figure 2.4.3. Computed p-value for chi-square goodness of fit test for Poisson GLM in R

```
> pchisq(glm_pois$deviance, df=glm_pois$df.residual, lower.tail=FALSE)
[1] 8.525694e-51
```

In this test, the null hypothesis is that our model is correctly specified, however, looking at Figure 2.4.3 above, we have overwhelmingly strong evidence to reject this hypothesis. In essence, we see that our model does not accurately fit the data. This suggests that our model is missing something major – like a correlation structure. As such, we may want to use a *generalized linear mixed model* (GLMM), which will be the topic of discussion in the next chapter.

§ 3 –Generalized Linear Mixed Models

As seen in the investigation in the last chapter where homicide count data was fit using a GLM (Poisson Regression model), the model did not accurately fit our data based on the chi-squared goodness of fit test. Loosely speaking, this was likely due to the assumption that data points were independent. It makes a lot more sense to assume there is some sort of correlation between areas of high crime rate, whether it be because they're neighbouring districts or because they have similar socioeconomic environments. This correlation must somehow be accounted for in our model, and so this leads us into the discussion of *generalized linear mixed models* (GLMM's).

These types of mixed models are historically known to be very useful in describing and fitting presumably related data points which are observed in clusters. These models are advantageous over non-mixed approaches in regression analysis, as they allow one to circumvent the restriction in analysis, posed by dependence among observations. We should make the distinction between linear mixed models, and generalized linear mixed models first. As was the case in the non-mixed models, the GLMM is an extension of the LMM, which allows for the response variable to come from an exponential family of distributions, and relates the responses to the predictors in a linear relationship only through the use of a *link function* as defined in chapter 2. As an e.g. in a Poisson GLMM, the response is related to the predictors by

$$g(E\{Y_i|X\}) = g(\mu_i) = \ln(\mu_i) = \ln e^{X_i^T \beta + Z_i^T \phi} = \boldsymbol{X_i^T \beta} + \boldsymbol{Z_i^T \phi}$$

§ 3.1 –Generalized Linear Mixed Models

Def 3.1. In regression analysis, a generalized linear mixed model is a generalized linear regression model which has both a *fixed component* and a *random component*, both of which are related to the response, **y**, through a non-identity link function. As was the case in GLMs, GLMM's also have a three part specification, however the difference here is that

$\boldsymbol{\eta} = \big(g(\mu_1), \ldots, g(\mu_n)\big)^{\boldsymbol{T}} = \boldsymbol{X\beta} + \boldsymbol{Z}\phi$ where $g(\mu_i) = \boldsymbol{X_i^T\beta} + \boldsymbol{Z_i^T\phi}$

- $\boldsymbol{\beta}$ is an unknown p-vector of fixed effects (non-random)
- $\boldsymbol{\phi}$ is an unknown q-vector of random effect coefficients, where E($\boldsymbol{\phi}$) = **0** and Var($\boldsymbol{\phi}$) = $\boldsymbol{G}_{\phi}$ is its (q x q) variance-covariance matrix. It is often assumed that is normally distributed.
- $\boldsymbol{X}$ is the (n x p) design matrix containing the *p* fixed predictor variables, relating the responses in $\boldsymbol{y}$ to the vector $\boldsymbol{\beta}$
- $\boldsymbol{Z}$ is the (n x q) design matrix containing the *q* random effects, relating the responses in $\boldsymbol{y}$ to the vector $\boldsymbol{\phi}$ – often assumed to be the identity matrix

The reader should note that when we say fixed effects, we mean the portion of the model that contains non-random information and data, and likewise when we say random effects, we are talking about the part of the model which contains information about the data in the form of a random vector. As is the case in non-mixed models, here we can also have linear mixed models (LMMs), which are a narrower class of regression models where the responses and errors are assumed to come from a Gaussian distribution. As always, one of the main

goals in regression models is to estimate the unknown parameters in order to gather more information about our responses and relationships amongst predictors. The method used to get estimates for $\boldsymbol{\beta}$ and $\boldsymbol{u}$ in LMMs done via solving *Henderson's mixed model equations* (MME), which we outline next.

§ 3.2 Estimating GLMM Parameters

Linear Mixed Model (LMMs)

In the simpler case, where responses are assumed to be independent and from a Gaussian distribution, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{R})$, the normality of the errors implies MLE and LSE coincide as estimators, i.e. both are optimal (BLUE). The additional assumptions that $\boldsymbol{\phi} \sim N(\mathbf{0}, \boldsymbol{G}_{\phi})$ and $Cov(\boldsymbol{\phi}, \boldsymbol{\varepsilon}) = \mathbf{0}$, are made and it is noted that the joint density of $\boldsymbol{y}$ and $\boldsymbol{\phi}$ may be written as:

$$f(\boldsymbol{y}, \boldsymbol{\phi}) = f(\boldsymbol{y}|\boldsymbol{\phi}) f(\boldsymbol{\phi}) \qquad (3\text{-}2)$$

If the joint log-likelihood of (3-2) is maximized (using partial differentiation) w.r.t $\boldsymbol{\beta}$ and $\boldsymbol{\phi}$ we obtain Henderson's mixed model equations (3-3), which can be seen in the relation below.

$$\begin{pmatrix} \mathbf{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \mathbf{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \mathbf{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \mathbf{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\phi}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \mathbf{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix}$$

As a consequence of the *Gauss-Markov Theorem* we have that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\phi}}$ are Best Linear Unbiased Estimates (BLUE). Derivation of these equations and proof of theorem can be found in Appendix 2A and 3A.

Fitting Generalized Linear Mixed Models (GLMMs)

In the case where response variables come from an exponential family, fitting a GLMM is classically done via maximum likelihood, and involves integrating over the random-effects of the model. However, such integrals are typically high in dimension and a closed form is difficult to obtain. Consequently, numerical integration methods are often more practical in fitting GLMM's. In R, one of the built-in functions used to fit a GLMM is glmm(). This function works by calculating and maximizing the *Monte Carlo likelihood approximation* (MCLA) to find the *Monte Carlo maximum likelihood estimates* (MCMLE's) for both the fixed-effects vector $\boldsymbol{\beta}$, and for the random-effects vector $\boldsymbol{\phi}$. A brief outline of MCMLE theory is provided in Appendix A4 for reference, so that the reader has intuition about how a statistical software package such as R obtains estimates computationally.

We will now wrap up our discussion of generalized linear mixed-effects models by going through an example of fitting data using glmm() in R, and then interpreting the results we obtain. This will help the reader understand how random-effects are incorporated into a mixed model, as well as help gain intuition behind how we build our CAR model for the actual investigation being done in this paper.

§ 3.3 – Fitting a GLMM Salamander Example

Ex. 3.2.1 – We now demonstrate how a glmm is fit using the built in glmm() function. The example we go through is also outlined in the documentation of the glmm package. Likewise, the dataset we use is included in the glmm package, and was collected by the University of Chicago in 1986. It contains information on two types of male and female salamanders, and whether or not they mated. As such, we are looking at a success or failure type of experiment and the family of distributions under consideration is Bernoulli. We load the dataset and check its contents.

Figure 3.3.1. Salamander Data Content

```
> data(salamander)
> names(salamander)
[1] "Mate"   "Cross"  "Female" "Male"
```

The "Mate" variable is a vector of 0's or 1's, where 0 indicates that a pair of salamanders did not mate, and 1 indicates they did. The "Cross" variable indicates the two types of salamander in the pair; i.e. Cross = W/R means a White Side female was paired with a Rough Butt male. Lastly, the "Female" and "Male" variables contain the ID number of each female and male respectively.

As mentioned, we want to fit a Bernoulli mixed model with a logit link function, since our responses are 0 or 1. The fixed-effects component here is the cross type, and consequently we have four parameters ($\beta_{R/R}$, $\beta_{R/W}$, $\beta_{W/R}$, $\beta_{W/W}$). Naturally, some males will be more likely to mate than others, and similarly some females will be more likely to mate than others. This means there's variability among each of the females and males, and for our model to measure these presumably random tendencies, we introduce a random effect for each male (and female) salamander.

Our assumptions for the 2 random effects are $\boldsymbol{u_i} \sim \boldsymbol{iid}$ (0, $\boldsymbol{\sigma^2_m}$) and $\boldsymbol{v_i} \sim \boldsymbol{iid}$ (0, $\boldsymbol{\sigma^2_f}$) as well as that the two random effects are independent of each other. The associated parameters that need estimation are the variances - $\boldsymbol{\sigma^2_m}$ and $\boldsymbol{\sigma^2_f}$.

Shown below is the code used to fit our data, and then display the summary, as well as the summary itself.

Figure 3.3.2. Code for fitting GLMM for salamander data

```
set.seed(1234)
clust <- makeCluster(2)
sal <- glmm(Mate ~ 0 + Cross, random = list(~ 0 + Female,~ 0 + Male),
            varcomps.names = c("F", "M"), data = salamander,
            family.glmm = bernoulli.glmm, m = 10^4,
            debug = TRUE, cluster = clust)

stopCluster(clust)
return(summary(sal))
```

We note that the cluster argument here is used to allow computation of the gradient vector, Hessian matrix and values to be done simultaneously as opposed to sequentially, allowing for a more time-efficient function.

Figure 3.3.3. Summary for GLMM salamander model

```
Link is: "logit (log odds)"

Fixed Effects:
        Estimate Std. Error z value Pr(>|z|)
CrossR/R   1.2328     0.3048   4.045 5.24e-05 ***
CrossR/W   0.3199     0.2670   1.198  0.23077
CrossW/R  -1.9986     0.3308  -6.042 1.52e-09 ***
CrossW/W   0.9159     0.2970   3.084  0.00204 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Variance Components for Random Effects (P-values are one-tailed):
  Estimate Std. Error z value Pr(>|z|)/2
F   1.4588     0.3107   4.695   1.33e-06 ***
M   1.6436     0.3342   4.918   4.36e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the summary, we see coefficient estimates for the fixed-effects, along with their associated standard errors, Wald statistics, and p-values indicating significance. We also have estimates for the variance components, as well as their Wald test statistics and their p-values. Of the 4 fixed-effects covariates, we see 3 are significant, meaning that the R/R and W/W crosses positively influence mating in salamander pairs, whereas the W/R cross has a negative influence in mating between male and female salamanders (it's coefficient is negative). Further, we see that both variance components cannot be discarded from the model. This is an implication that there exists an unknown factor which causes differences between the likelihoods of mating, across the 4 different breeds of salamander.

§ 4 – Spatial Conditional Autoregressive Models

§ 4.1 Overview of Spatial Autoregressive Models

Although there are various spatial models used for geostatistical analysis, the family of *Conditional Autoregressive* (CAR) models are typically used to model areal data that exhibit some sort of spatial pattern or dependence. Since it is the focus of this investigation to model spatially correlated data, we narrow our focus to CAR models. Loosely speaking, spatial dependency is defined as the co-variation of data across a geographic space. This spatial dependency leads to spatial autocorrelation among observations, and as mentioned in the Preface, this poses a problem in regression analysis since the assumption of independence between observations is no longer valid. This may yield unstable parameter estimates and consequently render unreliable significance tests, and incorrect interpretation of a fitted model. To see how spatial regression models like a CAR model address spatial dependence, and capture information from spatial relationships, one must understand the structure of the model.

Def 4.1. A *conditional autoregressive* (CAR) model is a particular case of a *Markov Random Field* (MRF). Formally, a random process, or collection of variables $\{\boldsymbol{Y}(s) : s \in D\}$ is defined as a Markov Random Field, if

(1) It is described by an undirected graph (or spatial lattice), D

(2) It satisfies the memoryless property (independence of the past)

Where the spatial lattice $D := \{s : s = 1, \ldots, n\}$ is a countable collection of *n* geographic locations, and $\boldsymbol{Y}(s)$ is one observation of our data at location *s*.

In the view of a GLMM framework, a slight change of notation is required,

$$g(\mu_i) = \boldsymbol{X_i^T \beta} + \boldsymbol{Z_i^T \phi} \qquad (4\text{-}1)$$

where, in accordance with Def 4.1, $i=1,\ldots,n$ denotes the i[th] geographic location, Y_i is one observation at of our data, at location i, and the rest of the model components are defined as in the definition of a GLMM given by (3-1).

What makes CAR models distinct from non-spatial GLMM models is the variance-covariance structure of their random-effects component. Consider the model in (4-1). Here, we further assume that the random effects component is multivariate Gaussian, specified as

$$\boldsymbol{\phi} \sim N(0,\ \tau(\boldsymbol{D} - \alpha \boldsymbol{W})^{-1}) \qquad (4\text{-}2)$$

where τ is a measure of overall variance, accounted for by the lattice structure of the model,

$\mathbf{W}_{nxn} = \{w_{ij}\}$, and each entry $w_{ij} = \begin{cases} 1, & if\ locations\ i \neq j\ are\ neighbors \\ 0, & otherwise \end{cases}$

$\mathbf{D}_{nxn} = \{d_{ii}\}$ where $d_{ii} = N_i$ where N_i is the number of neighbours of location i

α , $0 < \alpha < 1$ is a measure of the strength of the spatial dependency component

$\sum_{\phi} = \tau(\boldsymbol{D} - \alpha \boldsymbol{W})^{-1}$ is the positive semi-definite covariance matrix of the random-effects component

Def 4.2. We note that the matrix given by

$\boldsymbol{W}_{nxn} = \{w_{ij}\}$, and each entry $w_{ij} = \begin{cases} 1, & if\ locations\ i \neq j\ are\ neighbors \\ 0, & otherwise \end{cases}$

is called the spatial weights matrix or spatial adjacency matrix, where an entry $w_{ij} = 1$ if communities (i,j) are neighbours, and $w_{ij} = 0$ otherwise.

By convention, each entry on the main diagonal w_{ii} is 0. There are several known methods for characterizing two locations (*i*,*j*) as neighbours, most of which are based on a distance metric, or on an undirected graph as in the picture below. It is through the spatial weights matrix that the autocorrelation is accounted for.

It should also be noted that we say "*Conditional*" Autoregressive model since, given a set of observations corresponding to *n* different areal regions, the random spatial relationship between observation *i* and *j* can be modelled conditionally as

$$\phi_i \,|\phi_j \sim N\left(\overline{\phi}_i, \frac{\tau}{N_i}\right) \qquad (4\text{-}4)$$

where N_i is the number of neighbours for location *i*, and the mean $\overline{\phi}_i$ is given by

$$\bar{\phi}_i = \sum_{j \in N_i} \frac{\phi_j}{N_i} \qquad (4\text{-}5)$$

The model given by (4-1) whose has random-effects component has the form in (4-2) is called the *intrinsic*-CAR model whenever the parameter $\alpha = 1$. Note that CAR models are distinct from SAR (simultaneous autoregressive) models.

§ 4.2 Fitting Poisson CAR Model

Obtaining the Variance – Covariance Structure

Let our model take the form $\mu_i = e^{\boldsymbol{X_i^T \beta} + \boldsymbol{Z_i^T \phi}}$

where $\boldsymbol{\phi}$ has a spatial CAR structure as in (4-2), our response data comes from an exponential family of distributions, and our error terms are at least independently distributed. We recall that in the general case, $\boldsymbol{\phi} \sim N(0, \tau(\boldsymbol{D} - \alpha \boldsymbol{W})^{-1})$ where $\boldsymbol{G_\phi} = \tau(\boldsymbol{D} - \alpha \boldsymbol{W})^{-1}$ is the variance-covariance matrix to be estimated, with parameters τ and α being unknown and **W** being the spatial weights matrix.

To obtain estimates for the parameters, we must maximize the joint log-likelihood function of $\boldsymbol{y}$ and $\boldsymbol{\phi}$ which is given by the following expression**:**

$$l(G_{\boldsymbol{\phi}}, \boldsymbol{\phi}, \boldsymbol{\beta}) = log \prod_{i=1}^{n} \left(\int f_{y_i|\boldsymbol{\phi}}(y_i \mid \boldsymbol{\phi}, \boldsymbol{\beta}, \boldsymbol{G_\phi}) f_{\boldsymbol{\phi}}(\boldsymbol{\phi} | \boldsymbol{G_\phi}) \right)$$

$$= log \prod_{i=1}^{n} \int \frac{e^{-exp[\boldsymbol{X_i^T \beta}]} \; exp[\boldsymbol{X_i^T \beta}]}{y_i\,!} \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{G_\phi}|^{\frac{1}{2}}} \; exp\left(-\frac{1}{2} (\boldsymbol{\phi}^{(i)})^T \boldsymbol{G_\phi^{-1}} (\boldsymbol{\phi}^{(i)}) \right)$$

$$= \int \sum_{i=1}^{n} \frac{e^{-exp[\boldsymbol{X_i^T \beta}]} \; exp[\boldsymbol{X_i^T \beta}]}{y_i\,!} \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{G_\phi}|^{\frac{1}{2}}} \; exp\left(-\frac{1}{2} (\boldsymbol{\phi}^{(i)})^T \boldsymbol{G_\phi^{-1}} (\boldsymbol{\phi}^{(i)}) \right)$$

(4-6)

Where $\boldsymbol{\phi}^{(i)}$ is an entry of the random vector $\boldsymbol{\phi}$, *p* is its dimension, and *n* is the size of the response vector. Typically in real world applications where data is of high dimension, as discussed in section 3.2, a *Markov Chain Monte Carlo* (MCMC) "Pseudo-likelihood"

maximization technique is used for obtaining estimates. This is done to optimize run-time efficiency of our code when obtaining solutions of high-dimensional integrals, which in theory are computable, but have closed form solutions that are hard to obtain. When using the R software package *rstan* which fits our data, this estimation is done automatically by the code. The procedure is outlined in brief detail in Appendix A.

Overall measure of correlation

As in the case of non-spatial data, where Pearson's correlation coefficient *p* is used to measure overall correlation between two points, traditional spatial autocorrelation statistics such as *Moran's I*, *Geary's C,* are tools that estimate the overall autocorrelation within our geographic space. Geary's *C* and Moran's *I* are defined respectively as

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}\left(y_i - y_j\right)^2}{2W\sum_{i=1}^{n}(y_i - \overline{y})^2} \qquad I = \frac{n}{W}\frac{\sum_i \sum_j w_{ij}(y_i - \overline{y})\left(y_j - \overline{y}\right)}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

where W is the sum of all w_{ij}, *n* is the number of spatial units indexed by *i* and *j*, *y* is the response variate of interest, and $\overline{y}$ is its sample mean, and w_{ij}, are the entries of the spatial weights matrix for the CAR model, as in Def 4.2. We have that both $-1 \leq C \leq 1$ and $-1 \leq I \leq 1$. A positive correlation suggests that the observations clustered together in the same geographic area somehow related, while an overall negative correlation suggests dissimilarity among neighbouring values.

§ 4.3 Common Spatial CAR Models and Inference Techniques

Spatial Neighbourhood Structure

There are several practical ways to characterize spatial relationships in a spatial model. We will summarize the two approaches most relevant to our investigation. The first, and more common approach is to have the spatial weights matrix categorize geographic regions that are adjacent to one another on a map as neighbours. Adjacency is typically measured with a distance metric such as the *Absolute Value Norm* or the *Euclidean Norm*. The second approach, which is more susceptible to human errors, is to build the spatial weights matrix based on locations having common traits. That is, if data agglomerations for locations *i* and *j* are similar, then deem them as neighbours. How one defines similarity is tentative, and is typically a source of location fallacy.

Bayesian Inference of Model Parameters

Depending on the specific model that is employed, spatial dependency can enter the regression model either through the relationship between responses and fixed-predictors, relationship between responses and random-effects, or through the residual terms. Loosely speaking, in a regression setting, Bayesian inference fits a conditional probabilistic model to the unknown regression parameters. It allows one to account for spatial dependency regardless of how it is included in the model. We can then use the probabilistic model to summarize statistical properties of the estimated parameters (i.e. build confidence intervals, prediction intervals, etc.), and ultimately draw conclusions from the data.

§ 4.4 Sources of Error and Assessment of Model Quality

Locational Fallacy

A common error in fitting aerial data with spatial autoregressive models such as the Poisson ICAR model is locational fallacy. This refers to the error that arises from the particular spatial neighbourhood structure that is chosen for the model, often at the discretion of the statistician. The spatial characterization could either be too simple, in the sense that it does not account for key details, or it could be wrong. In turn, this would yield an incorrect analysis, misinterpretation of data and ultimately misinformation.

Consider for example, building a spatial model which analyzes the degree of gun ownership and violent crime in the United States. As one would assume, different states will have varying laws regarding gun ownership. Intuitively then, the spatial characterization for the model will be designed to account for gun-laws across states. For privacy reasons however, the exact addresses of people may be inaccessible, and some people who have registered guns in a state with less restrictions on guns may actually have an address in another state with stricter gun laws. The resulting model would then inaccurately relate gun violence.

Limitation of Mathematical Tools and Computational Power

In general, spatial data analysis is still a relatively new and growing field. Techniques which minimize potential error are still being developed. Technology also plays a huge role. For example consider trying to statistically model coastlines and marine-life. Due to the fractal

nature of coastlines, a correct measure of its length and surface area may not be possible. Consequently, approximate GIS data is used in the statistical model, yielding potentially inaccurate results.

§ 5 – Spatial Analysis of Homicide Count Data

§ 5.1 - Procedure

Credibility of Data

The crime data we have for this investigation was obtained from the Chicago Police Dept. data portal, and the socioeconomic data was obtained from the City of Chicago's website, from a census conducted in 2008. The homicide count data has multiple components, including longitude and latitude coordinates, exact time and date of when the homicide took place, street address were the homicide occurred, etc. For the purpose of this investigation we will narrow our focus only to the Community Area component of the homicide count data, as it plays a key role in determining our spatial weights matrix later on.

Adjusting the Level of Violence to account for Community Population

Using population data, the murder levels for each community area was computed by taking the homicide count for that area, diving it by the area's population. Incorporating the populations of each district and pairing it with each count observation is preferred since it is an adjusted measure of the level of violence for a region, independent of the effect that a population may have on the level of violence. Namely, it accounts for the fact that 2 murders in a community with a population of 5,000 is much worse than 3 homicides in an area with a population of 15,000. A log offset term is introduced into the model so that we retain numerical stability when computing estimates. This is done because the population sizes of each community is in the 10,000's raising that number as an exponent would generate computational errors.

Justification for using Bayesian Inference

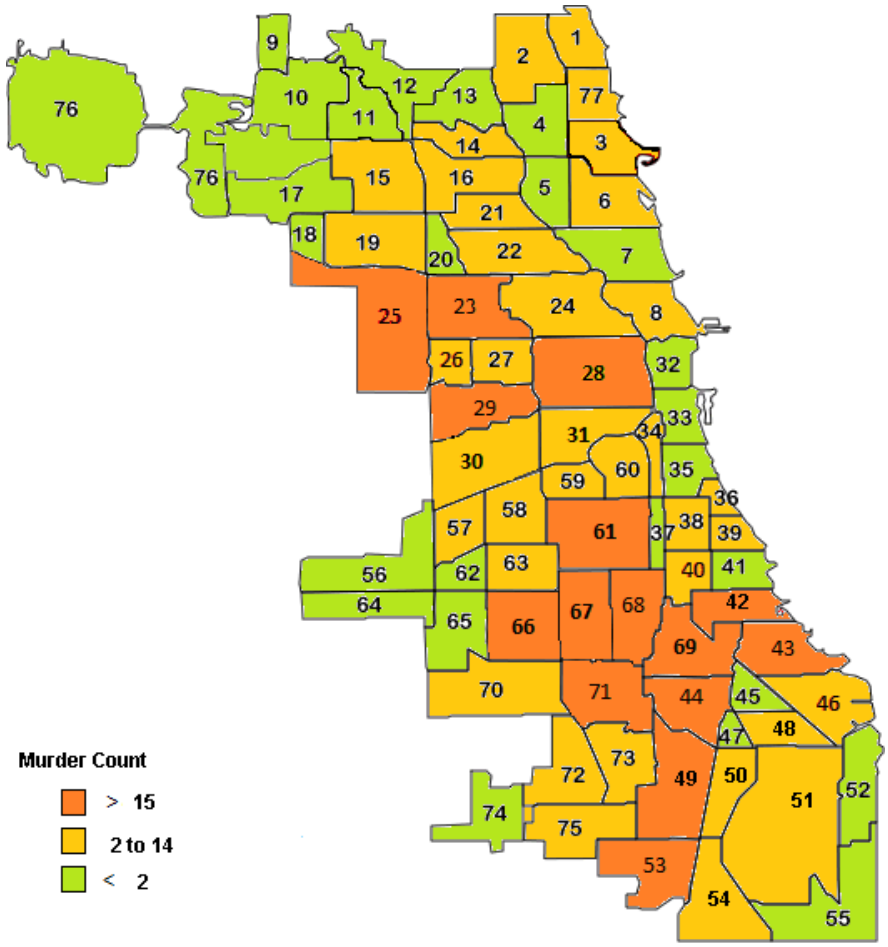
Using Bayesian hierarchical modeling in conjunction with Markov chain Monte Carlo (MCMC) in the context of modelling homicide counts across Chicago's districts is justified for 3 reasons.

(1) This approach allowed for imposing a correlation structure among random effects through their prior distributions, without affecting our ability to estimate the impact of the non-random socioeconomic factors on homicides.
(2) With the use of MCMC simulation, and sampling from a prior distribution, we overcome issues that arise in computation of marginal log-likelihood functions in multidimensional settings. This method of obtaining estimates is preferred over fitting the full likelihood model which has a less efficient runtime, and does not render more accurate results.
(3) It gives us a way of specifying a complicated non-Gaussian CAR model, which is then easily interpreted and used to make inference on homicide counts in spatially related regions of Chicago.

Choice of Neighbourhood / Spatial Weights Structure

Based on its murder counts, each community area is assigned to one of 3 groups (or neighbourhoods - in the spatial statistics sense). If the murder count for a community was between 0 and 1, it was assigned to the low group, if it was between 2 and 15, it was assigned to a moderate group, and if it was greater than 15 then it was assigned to an extreme group. A visualization of this can be seen in the heat map (Figure 5.1.1) on the next page. These specific numbers were chosen after some trial and error, in order to provide a balanced representation of crime among districts.

Figure 5.1.1. Heat map of Chicago based on murder counts per community



Using this visualization, we create our spatial adjacency matrix W. If two community areas, say 3 and 22, are grouped into the same neighbourhood, i.e. they are both yellow, then the entries (3,22) and (22,3) in the matrix are 1. Anywhere where entries w_{ij} are 1, communities i and j are neighbours. Keeping in mind that our goal is to verify if communities with similar levels of violence are correlated, the adjacency (weights) matrix is essential in achieving this as it adds the hypothesized correlation structure to our regression model. It can be seen that group 1 has 24 communities, group 2 has 38 communities, and group 3 has 15 communities

Setting up our CAR model

Once a preferred neighbourhood structure was devised our theoretical model was ready and we used it to generate 20,000 sample estimates. We extracted the sampled values and used their posterior distributions to analyze their statistical properties. Using the estimated intensity parameter, we were able to test if our hypothesized correlation structure was accurate, and measure the overall quality of our model.

As discussed, we've chosen our CAR model to be of the form:

$$\ln(\mu_i) = ln(offset) + \boldsymbol{X_i^t\beta} + \boldsymbol{Z_i^t\phi} \qquad (5\text{-}1)$$

For the fixed component of the model, the 77 x 5 design matrix $\boldsymbol{X}$ is composed of five covariates, $\boldsymbol{X_1, X_2,\ldots,X_5} \in \mathbb{R}^{77}$. These 5 covariates are socioeconomic factors, which are hypothesized to influence the murder rate in the city.

- $\boldsymbol{X_1}$ is a vector of the percent of people aged 16+ who are unemployed for each of the 77 communities
- $\boldsymbol{X_2}$ is a vector of the percent of households below the poverty line
- $\boldsymbol{X_3}$ is a vector of the percent of persons aged 25+ without a highschool diploma
- $\boldsymbol{X_4}$ is a vector containing the per capita income for each community area
- $\boldsymbol{X_5}$ is a vector containing the hardship index for every community area

For the random component of the model, the design matrix **Z** is assumed to be a 77x77 identity matrix, and the random vector of coefficients, $\boldsymbol{\phi}$, is assumed to follow a multivariate normal (p=77) distribution as outlined in chapter 4.

$$\boldsymbol{\phi} \sim N(0, \tau(\boldsymbol{D} - \alpha\boldsymbol{W})^{-1}) \qquad (5\text{-}2)$$

Note also here, the diagonal matrix **D** has entries $d_{ii} = N_i$ on its main diagonal that correspond to the number of neighbours that each community, *i*, has.

Code to fit Spatial CAR Model (courtesy of Prof. Jason D. Nielsen)

```
#Load Data
load("Chicago_CAR.Rdata")

# Define MCMC parameters
niter<-1e4
nchains<-4

# Define full data list to be used in stan()
W<-data$W
X<-scale(data$X)
n<-length(data$count)
X<-cbind(1,X)
offset<-c(scale(log(data$popl)))

full_d <- list(n = nrow(X),        # number of observations
          p = ncol(X),              # number of coefficients
          X = X,                  # design matrix
          y = data$count,         # observed number of murders
          log_offset = offset,    # log(expected) num. cases
          W = W)                  # spatial weights matrix

# Run the Poisson CAR model with Stan
fit <- stan("chicago_car.stan", data = full_d,
            iter = niter, chains = nchains, verbose = FALSE)

# Print Summary  and provide density plots
print(fit, pars = c('beta', 'tau', 'alpha', 'phi','lp__'))
to_plot <- c('beta', 'tau', 'alpha', 'lp__')
traceplot(fit, pars = to_plot)
stan_plot(fit, pars = "alpha", point_est = "mean", show_density = TRUE)

# Extract samples
fit_ss<-extract(fit)
```

§ 5.2 - Results

Running the code using R gives us the mean of each estimated unknown parameter, and is summarized below.

Figure 5.2.1. Summary of estimates given by spatial CAR model fit.

Parameter	Mean	Standard Error	Std. Dev
$\widehat{\beta_1}$	0.96	0.03	0.43
$\widehat{\beta_2}$	-0.54	0	0.24
$\widehat{\beta_3}$	0.05	0	0.18
$\widehat{\beta_4}$	0.35	0	0.19
$\widehat{\beta_5}$	-0.16	0	0.16
$\widehat{\beta_6}$	0.67	0.01	0.44
$\hat{\tau}$	0.31	0	0.15
$\hat{\alpha}$	0.98	0	0.03

Figure 5.2.2. Fit Summary as printed by R

```
Inference for Stan model: chicago_car.
4 chains, each with iter=10000; warmup=5000; thin=1;
post-warmup draws per chain=5000, total post-warmup draws=20000.

          mean se_mean    sd   2.5%    25%    50%    75%  97.5% n_eff Rhat
beta[1]   0.96    0.03  0.43  -0.03   0.73   0.99   1.22   1.82   280 1.01
beta[2]  -0.54    0.00  0.24  -1.01  -0.69  -0.53  -0.38  -0.07  2587 1.00
beta[3]   0.05    0.00  0.18  -0.31  -0.07   0.05   0.17   0.39  3616 1.00
beta[4]   0.35    0.00  0.19  -0.02   0.22   0.35   0.48   0.73  3301 1.00
beta[5]  -0.16    0.00  0.16  -0.49  -0.27  -0.16  -0.06   0.15  5230 1.00
beta[6]   0.67    0.01  0.44  -0.18   0.37   0.66   0.96   1.57  2348 1.00
tau       0.31    0.00  0.15   0.13   0.21   0.28   0.36   0.69  1541 1.00
alpha     0.98    0.00  0.03   0.91   0.97   0.99   0.99   1.00  2297 1.00
lp__    762.13    0.31 12.56 737.69 753.88 761.82 770.14 787.86  1650 1.00
```

Interpretation of Beta Estimates

Figure 5.2.3. A plot of the posterior densities of all 6 regression coefficient estimates

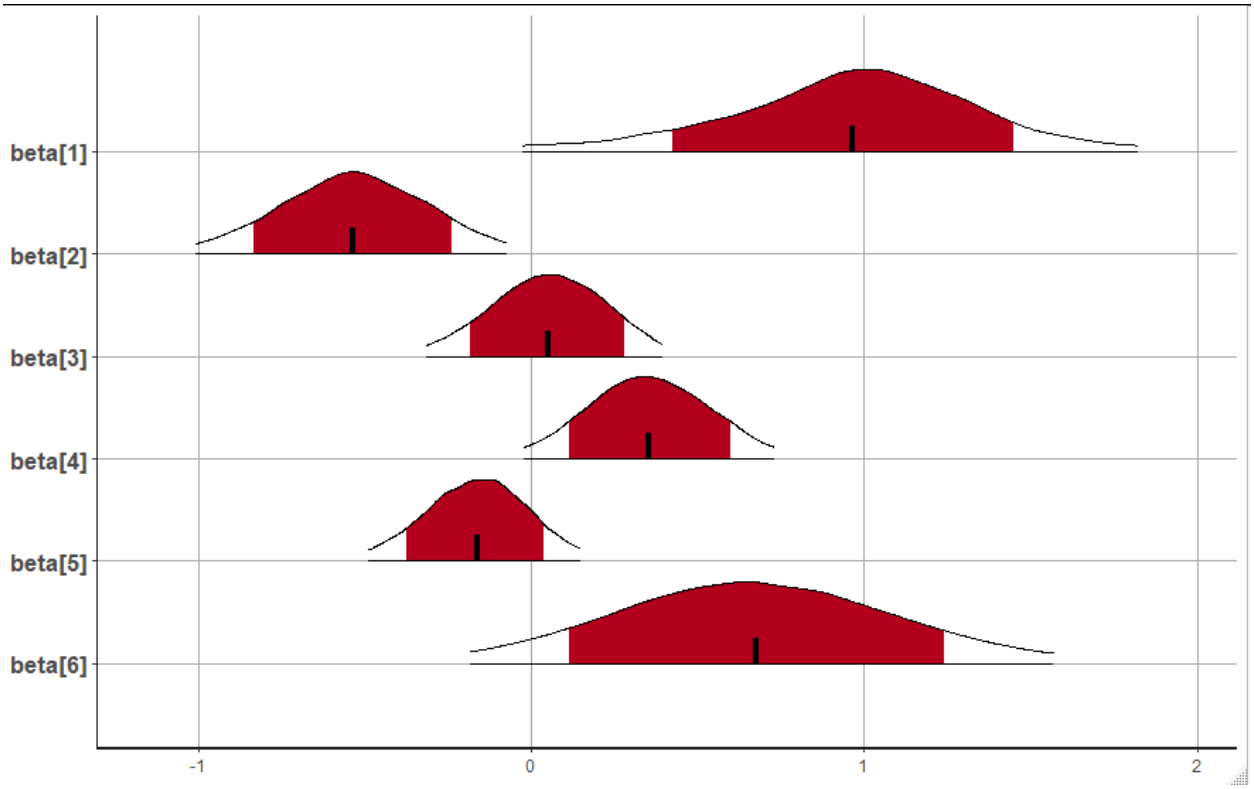
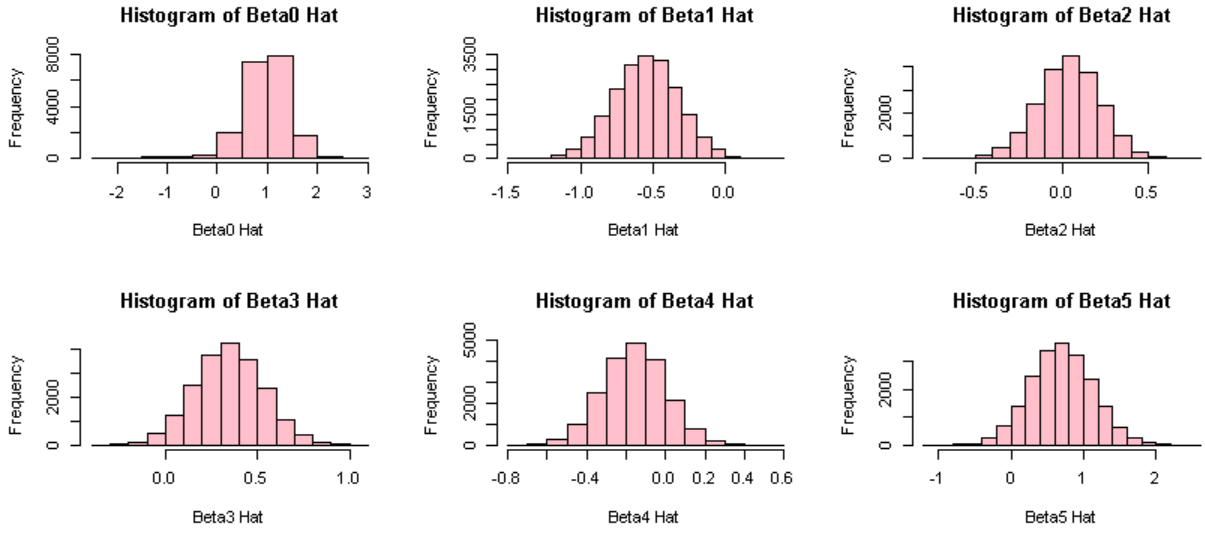


***Figure 5.2.4*:** The histograms corresponding to each sampled estimate for fixed-components



Looking at the plots we notice that the fixed-effects coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5,$ have sample means close to 0, and, under a normality assumption on the sampling distribution of the $\hat{\beta}_j$'s, if we create 95% confidence interval for each, 0 will be included in the interval.

Using the formula for an approximate 100(1-α) % confidence interval given by

$$\widehat{\beta_J} \pm z_{\alpha/2}\sqrt{Var(\widehat{\beta_J})}$$

where each standard error is the standard deviation of $\hat{\beta}_j$ given in the output above, we compute the respective 95% confidence intervals to be:

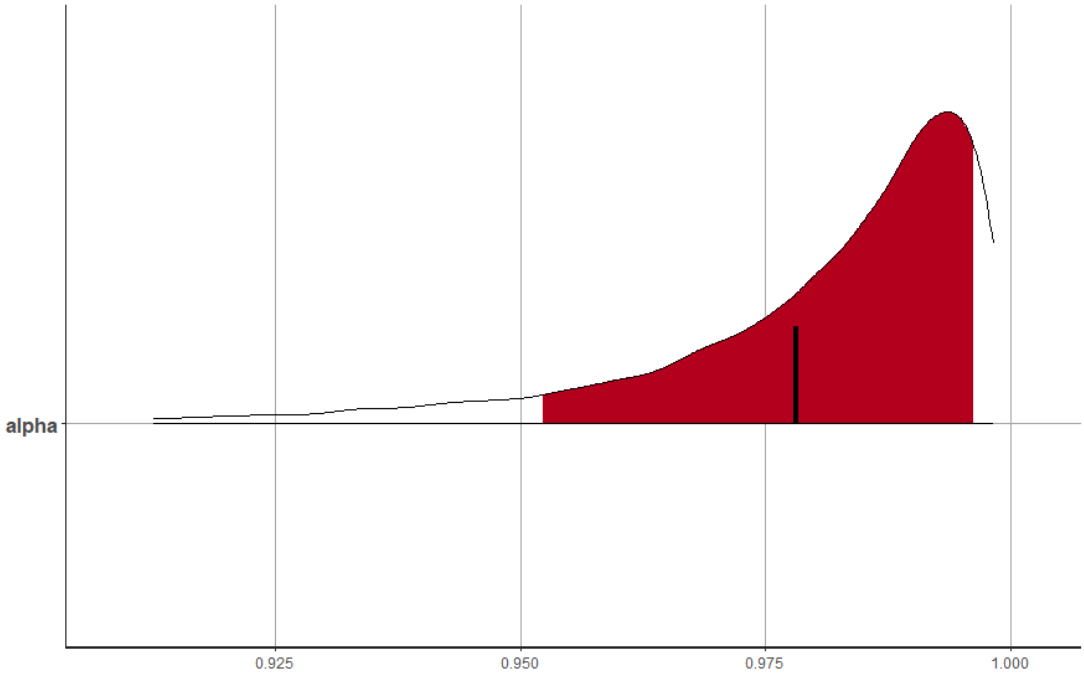
$$\hat{\beta}_0: [0.1172, 1.983],\ \hat{\beta}_1: [-0.0104, -0.0696],\ \hat{\beta}_2: [-0.303, 0.403]$$

$$\hat{\beta}_3: [-0.022, 0.722],\ \hat{\beta}_4: [-0.474, 0.154],\ \hat{\beta}_5: [-0.1924, 1.5324]$$

We notice that each interval, besides the ones corresponding to $\hat{\beta}_0$ and $\hat{\beta}_1$, contain 0. This gives us reasonable belief that the null hypothesis in the test $H_o: \beta_j = 0\ vs. H_\alpha: \beta_j \neq 0\ ,\ j = 2,3,4,5$ should *not* be rejected. In simpler terms, the covariates $\boldsymbol{X_2, X_3, X_4, X_5}$ that correspond to each socioeconomic factor, respectively, have no significant relationship with homicide counts across the city, while the covariate $\boldsymbol{X_1}$ that corresponds to percent of people aged 25+ without a high school diploma does indeed have somewhat of an impact on the homicide levels in Chicago. This is telling us that the correlation structure in the random effects component explains the majority of the data, implying that as long as we know the location of a murder, we can completely model homicide counts, and almost completely discount socioeconomic factors such as hardship index, or per capita income. This is obviously hard to believe, and instead, we should interpret this is as a potential flaw in the design of our spatial weights structure. This also gives us more incentive on defining the spatial weights matrix according to distance and adjacency of neighbourhoods.

Interpretation of alpha parameter

Figure 5.2.5. Posterior distribution of 20000 sampled alpha estimates



We observe that the estimated intensity parameter alpha has a mean value of 0.98, which is nearly 1. This implies that the hypothesized spatial correlation does indeed play significant role in the model, and that the spatial weights matrix (which was designed under the assumption that communities exhibit spatial correlation) remains in our model, and is 98% identical to the originally constructed spatial weights matrix. This isn't necessarily a good thing, since there should be a balance between how much of the data is being explained by the fixed and random components of our model. In terms of this model, this seems reasonable, since we just saw that the fixed-effects regression weights had nearly no significance on homicides across the city.

Random Effects Coefficients

Figure 5.2.6: Means for estimates of the 77 random effects coefficients, given by R

	mean	se_mean	sd
phi[1]	0.40	0.05	0.59
phi[2]	0.23	0.05	0.59
phi[3]	0.44	0.05	0.60
phi[4]	-0.97	0.05	0.66
phi[5]	-0.94	0.05	0.67
phi[6]	0.19	0.05	0.61
phi[7]	-1.00	0.05	0.66
phi[8]	0.58	0.05	0.62
phi[9]	-0.89	0.05	0.68
phi[10]	-0.85	0.05	0.66
phi[11]	-0.96	0.05	0.66
phi[12]	-0.91	0.05	0.67
phi[13]	-0.94	0.05	0.67
phi[14]	0.47	0.05	0.60
phi[15]	0.19	0.05	0.60
phi[16]	0.24	0.05	0.61
phi[17]	-0.89	0.05	0.65
phi[18]	-0.92	0.05	0.67
phi[19]	0.19	0.05	0.59
phi[20]	-0.83	0.05	0.66
phi[21]	0.32	0.05	0.61
phi[22]	0.12	0.05	0.61
phi[23]	0.58	0.05	0.59
phi[24]	0.51	0.05	0.60
phi[25]	0.19	0.05	0.57
phi[26]	0.51	0.05	0.62
phi[27]	0.39	0.05	0.60
phi[28]	1.31	0.05	0.61
phi[29]	0.59	0.05	0.60
phi[30]	0.50	0.06	0.62
phi[31]	0.60	0.06	0.59
phi[32]	-0.93	0.05	0.67
phi[33]	-0.76	0.05	0.68
phi[34]	0.42	0.05	0.61
phi[35]	-1.01	0.05	0.66
phi[36]	0.40	0.05	0.61
phi[37]	-0.80	0.05	0.66
phi[38]	0.51	0.05	0.59
phi[39]	0.57	0.06	0.62
phi[40]	1.05	0.05	0.62
phi[38]	0.51	0.05	0.59
phi[39]	0.57	0.06	0.62
phi[40]	1.05	0.05	0.62
phi[41]	-0.95	0.05	0.67
phi[42]	1.21	0.05	0.59
phi[43]	0.46	0.05	0.58
phi[44]	0.59	0.05	0.59
phi[45]	-0.95	0.05	0.67
phi[46]	0.80	0.05	0.59
phi[47]	-0.76	0.05	0.67
phi[48]	0.78	0.05	0.61
phi[49]	0.70	0.05	0.58
phi[50]	0.69	0.05	0.62
phi[51]	0.98	0.06	0.61
phi[52]	-0.97	0.05	0.67
phi[53]	0.94	0.05	0.59
phi[54]	0.32	0.05	0.62
phi[55]	-0.76	0.05	0.67
phi[56]	-0.86	0.05	0.66
phi[57]	0.51	0.05	0.62
phi[58]	0.49	0.05	0.60
phi[59]	0.61	0.05	0.62
phi[60]	0.47	0.05	0.60
phi[61]	1.01	0.05	0.59
phi[62]	-0.82	0.05	0.66
phi[63]	0.51	0.05	0.60
phi[64]	-0.94	0.05	0.67
phi[65]	-0.85	0.05	0.66
phi[66]	0.75	0.05	0.58
phi[67]	0.07	0.05	0.63
phi[68]	0.28	0.05	0.61
phi[69]	0.78	0.05	0.58
phi[70]	0.60	0.05	0.59
phi[71]	-0.12	0.05	0.61
phi[72]	0.51	0.05	0.62
phi[73]	0.32	0.05	0.59
phi[74]	-0.94	0.05	0.67
phi[75]	0.71	0.06	0.61
phi[76]	-0.77	0.05	0.67
phi[77]	0.40	0.05	0.60

The overall mean for each of these coefficients was computed to be 0.04793, which is very close to 0, as one would expect based on the theoretical model.

Variance Parameter

Next, τ is the overall estimated variance parameter which denotes the variability of the data that remains unaccounted for by the fixed-components of the model. We see its estimated mean value is 0.31. This is telling us that once the neighbourhood structure has been devised, and the number of 'neighbours' for each grouped homicide observation has been accounted for in the matrix **D**, the overall variation of homicides across members of each of the 3 spatial neighbourhoods (or clusters) is 0.31. Intuitively this makes sense, as data points were grouped together based on their similarity should not be too different. Again, this points to the conclusion that almost all of the information of our homicide data for each city district is explained by the spatial structure of our *CAR* model.

Overall Measure of Correlation

The statistic Moran's I was computed to be 0.8187924 indicating an overall positive spatial autoc orrelation level among data agglomerations across the city.

§ 5.3 – Conclusion

We saw that our model had flaws in its ability to find significance in 4 of the 5 socioeconomic factors that were used as predictor variables in the fixed-components of the Poisson CAR model. As discussed, this could likely be due to *locational fallacy*, i.e. the choice of spatial characterization of neighbourhoods, done via the spatial weights matrix. A proposed solution would be to re-design the spatial weights matrix, and base our neighbour characterization on adjacency of communities that are close in proximity to each other. This is a more natural characterization of spatial relationships. We could have also characterized two communities i,j as neighbours based on similarity of socioeconomic factors. Another potential reason for why the covariates were deemed insignificant is how their significance was evaluated. Typically, one can also use the *deviance residuals* and compare across multiple proposed models which include varying combinations of the five socioeconomic factors, choosing the proposed model that yields the best deviance.

An interesting extension to this project would be to perform a time-series analysis of the homicides, and try to detect seasonal trend in crimes. Typically, more gun violence occurs when there are more people outside, and there is usually more people outdoors during the summer. Using a time series model to fit our data would then also allow for predicting future crime levels.

It should also be noted that the model used for fitting homicide counts serves as a template, and is not limited to only modelling homicides. We have data of multiple types of crimes, so a similar analysis could be performed to each and every one.

Appendix A

A1: Estimation of Poisson GLM parameters via iterative weighted least squares (IRWLS) technique.

Let **Y** be an nx1 vector, coming from a Poisson distribution, such that for each Y_i

$$P(Y_i = y_i \mid X_i, \beta) = \frac{\mu_i^{y_i} \exp[X_i^T \boldsymbol{\beta}]}{y_i !}$$

Where $\mu_i = \exp[X_i^T \beta]$ and our log link function is given by $g(\mu_i) = \ln \mu_i = \boldsymbol{X_i^T \beta}$ and

The likelihood function for a Poisson regression model where the observations have dimension *n* is given by:

$$L(\beta \mid Y_i, X_i) = \prod_{i=1}^{n} P(Y_i = y_i \mid \boldsymbol{X_i}, \boldsymbol{\beta}) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} \exp[\boldsymbol{X_i^T \beta}]}{y_i !}$$

Taking the log of the likelihood function, we obtain

$$l(\beta) = \sum_{i=1}^{n} y_i X_i \boldsymbol{\beta} - \sum_{i=1}^{n} \exp[\boldsymbol{X_i^T \beta}] - \sum_{i=1}^{n} \log(y_i !)$$

Differentiating $l(\beta)$ with respect to β_k (k = 1,2,…,p), we obtain the likelihood equations:

$$\frac{\partial}{\partial \beta_k} l(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i X_{i,k} - \sum_{i=1}^{n} \exp\left[\boldsymbol{X_i^T \beta}\right] X_{i,k} = \sum_{i=1}^{n} (y_i - \mu_i) X_{i,k} = 0$$

For a Poisson model we have $\sigma_i^2 = \mu_i = \exp[\boldsymbol{X_i^T \beta}]$ and so the above equations become

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) X_{i,k} \mu_i}{\mu_i} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\sigma_i^2} \frac{\partial \mu_i}{\partial \beta_k} = 0 \qquad (1)$$

Since the last equation above has no analytical form, we obtain estimates recursively:

1. Begin with an initial guess for unknown σ_i, (e.g. for simplicity take $\sigma_i = 1$)
2. Substitute guessed value into (1) to obtain initial value $\widehat{\boldsymbol{\beta}^{(0)}}$
3. Use $\widehat{\boldsymbol{\beta}^{(0)}}$ (plug back in to (1)) to obtain estimated variance $\widehat{\sigma_i^2}$ Use this value and re-do weighted least squares with these initial estimates to obtain $\widehat{\boldsymbol{\beta}^{(1)}}$.
4. Iterate until convergence

To elaborate on the convergence part, suppose $\widehat{\boldsymbol{\beta}^{(k)}} \to \widehat{\boldsymbol{\beta}}$ as $k \to \infty$, it follows from (1) that

$$\sum_{i=1}^{n} \left[\frac{y_i - \mu_i(\widehat{\boldsymbol{\beta}^{(k+1)}})}{\sigma_i^2(\widehat{\boldsymbol{\beta}^{(k)}})} \right] \frac{\partial \mu_i(\widehat{\boldsymbol{\beta}^{(k+1)}})}{\partial \widehat{\boldsymbol{\beta}^{(k+1)}}} = 0$$

Then we see that $\hat{\beta}$ must be a root of the equation

$$\sum_{i=1}^{n} \left[\frac{y_i - \mu_i(\widehat{\boldsymbol{\beta}})}{\sigma_i^2(\widehat{\boldsymbol{\beta}})} \right] \frac{\partial \mu_i(\widehat{\boldsymbol{\beta}})}{\partial \hat{\beta}} = 0$$

A2: Gauss-Markov Theorem on optimality of LSE (MLE)

Given a linear regression model: $\boldsymbol{y} = \boldsymbol{X\beta} + \varepsilon$ where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I\sigma^2})$, recall that the Least-Squares Estimator for $\boldsymbol{\beta}$ is given by

$$\widehat{\beta} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'Y}$$

We claim the LSE (MLE) is the estimator with the smallest variance among the class of all unbiased estimators of β.

Proof: To see, let $\beta^* = \mathbf{C}\boldsymbol{y}$ be another unbiased linear estimator of β with $\mathbf{C} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'} + \mathbf{D}$ where D is a (k x n) non-zero matrix. We first show that β^* unbiased iff $\boldsymbol{DX} = 0$.

$$\begin{aligned}
\mathrm{E}(\beta^*) &= \mathrm{E}(\mathrm{C}y) \\
&= \mathrm{E}[((\boldsymbol{X'X})^{-1}\boldsymbol{X'} + \mathbf{D})(\mathbf{X\beta} + \varepsilon)] \\
&= ((\boldsymbol{X'X})^{-1}\boldsymbol{X'} + \mathbf{D})\mathbf{X\beta} + ((\boldsymbol{X'X})^{-1}\boldsymbol{X'} + \mathbf{D})\mathrm{E}(\varepsilon) \\
&= \left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' + \mathbf{D}\right)\mathbf{X\beta} \\
&= \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\mathbf{X\beta} + \mathrm{D}\mathbf{X\beta} \\
&= (\mathbf{I_k} + \mathbf{DX})\boldsymbol{\beta}
\end{aligned}$$

So, it must be that $\boldsymbol{DX} = 0$. Consider now the variance of this new estimator.

$$\begin{aligned}
\mathrm{Var}(\beta^*) = \mathrm{Var}(\mathbf{C}\mathrm{y}) = \mathbf{C}\,\mathrm{Var}(\mathbf{y})\mathbf{C}' &= \boldsymbol{\sigma^2}\mathbf{C}\mathbf{C}' \\
&= \boldsymbol{\sigma^2}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' + \mathbf{D}\right)\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}' + \mathbf{D}\right)' \\
&= \boldsymbol{\sigma^2}\left(\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{D}' + \boldsymbol{DX}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \boldsymbol{DD}'\right) \\
&= \boldsymbol{\sigma^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \boldsymbol{\sigma^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}(\boldsymbol{DX})' + \boldsymbol{\sigma^2}\boldsymbol{DX}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \boldsymbol{\sigma^2}\boldsymbol{DD}' = \\
&= \boldsymbol{\sigma^2}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} + \ \boldsymbol{\sigma^2}\boldsymbol{DD}' = \mathrm{Var}\left(\widehat{\beta}\right) + \ \boldsymbol{\sigma^2}\boldsymbol{DD}'
\end{aligned}$$

Hence, it follows that $\mathrm{Var}(\widehat{\beta}) < \mathrm{Var}(\beta^*)$ as claimed. □

A3: Henderson's Mixed Model Equations and Mixed Model Parameter Estimation

Consider the following mixed linear regression mode: $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \varepsilon$**,** where with the usual normality assumptions, we have

$$y|u \sim N(\mathbf{X}\beta + \mathbf{Zu}, \mathbf{R})$$

$$\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G})$$

Where X and Z are design matrices for the fixed and random effects respectively, and $\boldsymbol{\beta}$ and $\boldsymbol{u}$ are the fixed and random effects respectively.

The log likelihood is then:

$$-2\log L(\boldsymbol{\beta}, \mathbf{u}) = \log|\mathbf{R}| + (\mathbf{y} - \mathbf{X\beta} - \mathbf{Zu})'\boldsymbol{R^{-1}}(\mathbf{y} - \mathbf{X\beta} - \mathbf{Zu}) + \log|\mathbf{G}| + \mathbf{u}'\boldsymbol{G^{-1}}\mathbf{u}$$

Differentiating w.r.t both random and fixed effects and setting both to 0:

$$\frac{\partial \log L}{\partial \boldsymbol{u}} = \mathbf{Z}'\boldsymbol{R^{-1}}(\mathbf{y} - \mathbf{X\beta} - \mathbf{Zu}) - \boldsymbol{G^{-1}}\mathbf{u} = 0$$

$$\frac{\partial \log L}{\partial \boldsymbol{\beta}} = \mathbf{X}'R^{-1}(\mathbf{y} - \mathbf{X\beta} - \mathbf{Zu}) = 0$$

After some rearranging we obtain Henderson's mixed model equations

$$\mathbf{Z}'\boldsymbol{R^{-1}}\boldsymbol{y} = \mathbf{Z}'\boldsymbol{R^{-1}}\mathbf{X\beta} + \mathbf{u}(\mathbf{Z}'\boldsymbol{R^{-1}}\boldsymbol{Z} + \boldsymbol{G^{-1}})$$

$$\mathbf{X}'\boldsymbol{R^{-1}}\boldsymbol{y} = \mathbf{X}'\boldsymbol{R^{-1}}\mathbf{X\beta} + \mathbf{u}\mathbf{X}'\boldsymbol{R^{-1}}\boldsymbol{Z}$$

In matrix notation:

$$\begin{pmatrix} \mathbf{X}'\boldsymbol{R^{-1}}\boldsymbol{X} & \mathbf{X}'\boldsymbol{R^{-1}}\boldsymbol{Z} \\ \mathbf{Z}'\boldsymbol{R^{-1}}\boldsymbol{X} & \mathbf{Z}'\boldsymbol{R^{-1}}\boldsymbol{Z} + \boldsymbol{G^{-1}} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\boldsymbol{R^{-1}}y \\ \mathbf{Z}'\boldsymbol{R^{-1}}y \end{pmatrix} \qquad (I)$$

The equations in (I) are MLE equations, and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MLE}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{u}} \end{pmatrix}$. Since residuals are assumed to be normal, LSE and MLE coincide and so by the Gauss-Markov Theorem, $\widehat{\boldsymbol{\theta}}_{\boldsymbol{MLE}}$ is optimal.

A4: Estimation of high-dimensional GLMM models via Markov Chain MLEs

In the case of trying to obtain Maximum Likelihood estimates from multi-dimensional models, it is often hard to obtain an analytical (closed form) solution, since we must calculate a multi-dimensional integral.

Consider a GLMM regression model of the form

$$g(\mu_i) = X_i^T \boldsymbol{\beta} + Z_i^T \boldsymbol{u}$$

Where $\boldsymbol{y}|\boldsymbol{u}$ comes from a distribution with pmf or pdf given by $f_{y_i|u}(y_i|\mathrm{u}, \beta, \phi)$ and lastly assume that the random effects component is multivariate normal, $\boldsymbol{u} \sim N(\mathbf{0}, \boldsymbol{G_u})$.

Suppose interest lies in obtaining estimates for this model's parameters. First note that by definition, the likelihood function is

$$L(\boldsymbol{\beta}, \phi, \mathbf{G}) = \int \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} f_{y_i|u}(y_i|\boldsymbol{u}, \boldsymbol{\beta}, \phi) f_u(\boldsymbol{u}|\boldsymbol{G})$$

$$= \int f_{\mathbf{y}|\boldsymbol{u}}(\mathbf{y}|\boldsymbol{u}, \boldsymbol{\beta}, \phi) f_u(\boldsymbol{u}|\boldsymbol{G})\ d\boldsymbol{u}$$

$$= \int \frac{f_{\mathbf{y}|\boldsymbol{u}}(\mathbf{y}|\boldsymbol{u}, \boldsymbol{\beta}, \phi) f_u(\boldsymbol{u}|\boldsymbol{G})}{\boldsymbol{h_u}(\boldsymbol{u})} \boldsymbol{h_u}(\boldsymbol{u}) d\boldsymbol{u}$$

$$\doteq \frac{1}{N} \sum_{k=1}^{\boldsymbol{N}} \frac{f_{\mathbf{y}|\boldsymbol{u}}\left(\mathbf{y}\middle|\boldsymbol{u}^{(k)}, \boldsymbol{\beta}, \phi\right) f_u\left(\boldsymbol{u}^{(k)}\middle|\boldsymbol{G}\right)}{\boldsymbol{h_u}(\boldsymbol{u}^{(k)})}$$

Then, then using a numerical optimization scheme, the simulated likelihood given in the last term is maximized, and so we obtain our MCMLE. As with any iterative procedure, we iterate until we see convergence of simulated estimates.

We note that here, N is the number of iterations, h($\boldsymbol{u}$) is the *importance sampling* distribution, corresponding to the distribution from which values of $\boldsymbol{u}$ are drawn. The idea behind importance sampling is that we select only values of $\boldsymbol{u}$ which have significant impact on the parameter being estimated, thereby reducing the variance of the estimator. As such, we look for a distribution (h($\boldsymbol{u}$)) which generates more of these important values than other sampling distributions – hence its name. The simulated likelihood in (3-4) is numerically optimized after every iteration, thus producing our MCMLE for our generalized linear mixed-effects model.

In the case of the Poisson CAR model, we obtain the log-likelihood function by integrating over the 77-dimensional Gaussian random-effects vector, and R carries out the recursive optimization procedure outlined above.

Bibliography

Bacastow, T. (2014). *Understanding Spatial Fallacies.* Penn State Department of Geography.

Cameron, A. C. (2013). *Regression Analysis of Data* (2nd ed.). New York: Cambridge Press.

Cameron, A., & Trivedi, P. K. (1998). *Regression Analysis of Count Data.* New York: Cambridge Press.

Dupont, W. (2002). *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data.* New York: Cambridge Press.

Gelfand, A. E., & Vounatsou, P. (2003). Proper multivariate conditional autoregressivemodels for spatial data analysis. *Biostatistics*, 11-25.

Li, H. (2006). *Bayesian Hierarchical Models for Spatial Count Data withApplication to Fire Frequency in British Columbia.* Victoria, BC: University of Victoria.

Long, J. (1997). *Regression Models for Categorical and Limited Dependent Variables.* Thousand Oaks, CA: Sage Publications.

Long, J., & Freese, J. (2006). *Regression Models for Categorical Dependent Variables Using Stata, Second Edition.* College Station, TX: Stata Press.

Mardia, K. (1989). Maximum Likelihood Estimation for Spatial Models. In D. A. Griffith, *Spatial Statistics: Past, Present, and Future* (pp. 203-253). Ann Arbor, MI: Institute of Mathematical Geography.

Morris, M. (2019). *Spatial Models in Stan: Intrinsic Auto-Regressive Models for Areal Data.* Roswell, GA: ScienceDirect.

Wang, Y., & Kockelman, K. M. (2013). A POISSON-LOGNORMAL CONDITIONAL-AUTOREGRESSIVE MODEL FOR MULTIVARIATE SPATIAL ANALYSIS OF PEDESTRIAN CRASH COUNTS ACROSS NEIGHBORHOODS. *Accident Analyis and Prevention*, 71-84.

Warnes, J., & Ripley, B. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 640-642.

Zhou, X., & Lin, H. (2008). Spatial Weights Matrix. In X. H. Shekhar S., *Encyclopedia of GIS.* Boston, MA: Springer.