Spring 2011

Models and methods for computationally efficient analysis of large spatial and spatio-temporal data

Chengwei Yuan
University of New Hampshire, Durham

MODELS AND METHODS FOR COMPUTATIONALLY EFFICIENT ANALYSIS OF LARGE SPATIAL AND SPATIO-TEMPORAL DATA

BY

CHENGWEI YUAN

B.S., SOUTHEAST UNIVERSITY, CHINA 2004

DISSERTATION

Submitted to the University of New Hampshire

in Partial Fulfillment of

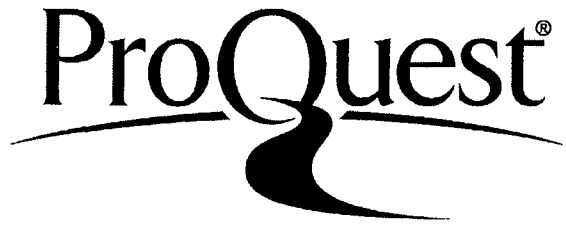
the Requirements for the Degree of

Doctor of Philosophy

in

Mathematics

May, 2011



ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

This thesis has been examined and approved.

Thesis Director, Dr. Ernst Linder

Professor of Mathematics and Statistics

Dr. Linyuan Li

Associate Professor of Mathematics and Statistics

Dr. Philip Ramsey

Instructor of Mathematics and Statistics

Dr. Mark Lyon

Assistant Professor of Mathematics

Dr. Paul Ossenbruggen

Professor Emeritus of Civil Engineering

7/29/2011

Date

DEDICATION

To Zhaoyu and my parents

ACKNOWLEDGMENTS

This dissertation could not have been written without the support and friendship found at the University of New Hampshire and elsewhere. I could not have come this far without the assistance of many individuals and I want to express my deepest appreciation to them.

I would like to give special thanks to my advisor Dr. Ernst Linder for his wisdom and inspiring guidance, encouragement and technical help throughout the research and writing process. I am grateful for my committee members: Drs. Linyuan Li, Philip Ramsey, Mark Lyon and Paul Ossenbruggen, for taking the time to discuss relevant research topics with me, to provide useful suggestions and valuable comments. I would also like to express my gratitude to Dr. Eric Grinberg and Dr. Rita Hibschweiler for their advice throughout my time at the University of New Hampshire and to Jan Jankowski, April Flore, and Ellen O'Keefe for their patience and various forms of support. Many thanks are also due to my fellow graduate students: Shan Yao, Eric LaFlamme, Hyung Kim, Kewei Lu, and Yinbin Pan for their friendship and helpful conversations.

Generous financial support made my doctoral studies and dissertation possible. I would like to thank the UNH Department of Mathematics and Statistics for providing the funding to me in the form of teaching assistantship. I would also like to thank the UNH Complex Systems Research Center for the honor of receiving a

research assistantship.

Finally, this dissertation is dedicated to my greatest blessing, my wife, the most decent and loving person I've ever known. Thanks to all of my family and friends, especially my parents for supporting and encouraging me to pursue what I love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

MODELS AND METHODS FOR COMPUTATIONALLY EFFICIENT ANALYSIS OF LARGE SPATIAL AND SPATIO-TEMPORAL DATA

by

Chengwei Yuan
University of New Hampshire, May 2011
Advisor: Dr. Ernst Linder

With the development of technology, massive amounts of data are often observed at a large number of spatial locations (n). However, statistical analysis is usually not feasible or not computationally efficient for such large dataset. This is the so-called "big n problem".

The goal of this dissertation is to contribute solutions to the "big n problem". The dissertation is devoted to computationally efficient methods and models for large spatial and spatio-temporal data. Several approximation methods to "the big n problem" are reviewed, and an extended autoregressive model, called the EAR model, is proposed as a parsimonious model that accounts for smoothness of a process collected over space. It is an extension of the Pettitt *et al.* as well as Czado and Prokopenko parameterizations of the spatial conditional autoregressive (CAR) model. To complement the computational advantage, a structure removing

orthonormal transformation named "pre-whitening" is described. This transformation is based on a singular value decomposition and results in the removal of spatial structure from the data. Circulant embedding technique further simplifies the calculation of eigenvalues and eigenvectors for the "pre-whitening" procedure.

The EAR model is studied to have connections to the Matérn class covariance structure in geostatistics as well as the integrated nested Laplace approximation (INLA) approach that is based on a stochastic partial differential equation (SPDE) framework. To model geostatistical data, a latent spatial Gaussian Markov random field (GMRF) with an EAR model prior is applied. The GMRF is defined on a fine grid and thus enables the posterior precision matrix to be diagonal through introducing a missing data scheme. This results in parameter estimation and spatial interpolation simultaneously under the Bayesian Markov chain Monte Carlo (MCMC) framework.

The EAR model is naturally extended to spatio-temporal models. In particular, a spatio-temporal model with spatially varying temporal trend parameters is discussed.

INTRODUCTION

In the past few decades, researchers in diverse fields such as climate, ecology and epidemiology, have been facing the task of analyzing data that are both spatially and temporally correlated. In most cases, spatial patterns at locations with short distances from each other are similar, so are the trends over short times. Locations nearby are called "neighbors". Similarities for those neighbors thus can be explained by the correlations in space and time, which can be statistically modeled. Since the milestone work by Cressie (1993), spatial and spatio-temporal statistical models have been investigated that can be used for such complex data. Also, advances in Geographical Information Systems (GIS) and remote sensing (satellites, Lidar, etc.) have enabled accurate geocoding and the collection of large amounts of scientific data. This has also generated considerable interest in statistical modeling for location-referenced spatial data.

Recent developments in Markov chain Monte Carlo (MCMC) procedures such as the Gibbs sampler, Metropolis-Hastings algorithm, or a combination thereof (Gelman *et al.*, 2003) now allow Bayesian analyses of sophisticated multilevel models for complex spatial data. However, the number of locations yielding observations is often too large for fitting desired hierarchical spatial models using MCMC methods, which are iterative and computationally intensive. This computational burden is exacerbated in multivariate settings with several spatially dependent re-

sponse variables as well as when spatial data are collected over time, such as with spatio-temporal data. This is the so-called "big n problem" in spatial statistics that relates to the inversion of the covariance matrix and its determinant calculation.

The computational burden of statistical estimation for large spatio-temporal data is a topic of great current interest. On the one hand, several approximation methods and models have been studied for geo-referenced data, for example, Cholesky decomposition, covariance tapering (Wendland 1998; Furrer *et al.* 2006; Kaufman *et al.* 2008), convolution methods (Higdon 1998, 2002; Higdon *et al.* 2003; Lemos and Sansó, 2009) and spectral domain approximations (Wikle 2002; Paciorek 2007). On the other hand, one of the most popular spatial interaction models for lattice data - data that has been aggregated over fixed areas - is the conditional autoregressive model (CAR) and Markov random fields (MRF) (Besag, 1974; Rue and Held, 2005). Here, the data at one location (area) is modeled conditionally on the data collected at neighboring locations. Lattice analysis is favored from a computational point of view because it directly models the sparse precision matrix $\mathbf{Q}$ which is the matrix inverse of the variance-covariance matrix $\mathbf{\Sigma}$ of the data. In geo-referenced data analysis $\mathbf{Q}$ needs to be calculated from $\mathbf{\Sigma}$, which is computationally taxing. Since $\mathbf{Q}$ is sparse, it helps to achieve fast computation.

Since Gaussian MRF models can serve as computationally efficient alternatives to Gaussian point-referenced, or geostatistical models (GGM), their relationships are of general interest. Rue and Tjelmeland (2002) examined the Gaussian process approximation with MRF. Lindström and Lindgren (2008) and Lindgren, Lindström and Rue (2010) applied finite element method to solving stochastic partial differ-

ential equations to bridge the Gaussian fields and Gaussian MRF. Song *et al.* (2008) conducted an empirical comparison between GGM and GMRF. Rue, Martino, and Chopin (2009) used Integrated Nested Laplace Approximations (INLA package) to numerically integrate out covariance "nuisance" parameters.

In this dissertation, the interest is to modify and extend existing procedures to allow for fast, computationally efficient estimation of parameters and also to provide a better model to represent extremely smooth spatial processes. Results in Linder (2001) and in Hupper (2005) are extended in several ways. First, an extended autoregressive (EAR) model is modified from a previous version along the lines of Czado and Prokopenko (2008) which improves identifiability. Second, the EAR model is investigated in detail and its Markov random field properties are derived. Third, connections are developed between the EAR model and the popular Matérn class of geostatistical models, as well as the new INLA approach. Next, we develop the framework for applying the EAR model for spatially irregular point-referenced, or, geostatistical, data. Here a latent process representation over a large fine grid is proposed combined with a missing data imputation. Finally, we discuss the application of the EAR model for spatio-temporal data.

The dissertation's chapters are arranged as follows. Chapters I-IV are reviews of spatial data, spatial models, approximation methods to "the big n problem", and the Bayesian parameter estimations in a hierarchical paradigm. Chapters V-VII are the main contributions of my dissertation. In Chapter I, a brief introduction to spatial data is discussed. In Chapter II, Gaussian spatial processes as well as both geostatistical models and conditional autoregressive models (CAR) will be

discussed. The Pettitt, Weir, and Hart (2002) and Czado and Prokopenko (2008) parameterizations of the CAR model are also explored here because of their computational efficiency properties. Chapter III introduces "the big n problem" and reviews several approximation methods. Topics covered in Chapter IV will involve Bayesian parameter estimations that rely on Markov chain Monte Carlo methods. An orthogonal data transformation procedure called "pre-whitening" that removes the correlation structure is also examined. In Chapter V, the CAR model is expanded to include a smoothness parameter that is capable to better describe smooth spatial processes. Relationship of this extended autoregressive (EAR) model to ordinary CAR models with higher order neighbor structures will be determined. A circulant embedding technique is discussed. Connections between the EAR model and the Matérn class covariance function as well as the INLA will also be presented here. In Chapter VI, application of the EAR model to geostatistics is studied. Chapter VII will introduce a spatio-temporal hierarchical structure that can model spatially varying temporal trends simultaneously. Conclusions and suggestions for future work are provided in Chapter VIII.

CHAPTER I

Spatial Data

In this chapter, types of spatial data are introduced. In particular, geostatistical data and lattice data are reviewed. In general, a spatial process in d dimensions can be expressed as

$$\{\mathbf{z}(\mathbf{s}) : \mathbf{s} \in D \subset \mathbb{R}^d\}.$$

Here $\mathbf{z}$ denotes the observations, for example, precipitation, ozone values, or the average SAT scores. The location at which $\mathbf{z}$ is observed is $\mathbf{s}$, a $d \times 1$ vector of coordinates. In most of cases, researchers and scientists take much interests in processes in two-dimensional space, $d = 2$, and $\mathbf{s} = (s_x, s_y)'$ are the Cartesian or longitude-latitude coordinates. Spatial data types are characterized by the domain D. In this dissertation, our interests are focused on two most common spatial data types: geostatistical data and lattice data.

1.1 Geostatistical Data

If the domain D is a continuous and fixed set, then we say the data is a "geostatistical data", also called "point-referenced data". The continuity here means $\mathbf{z}(\mathbf{s})$ can be observed at any location $\mathbf{s}$ within domain D. By fixed we mean that the points in D are non-stochastic. Theoretically, $\mathbf{z}(\mathbf{s})$ could be collected at an infinite set

of locations, however, in practice, they cannot be observed exhaustively due to cost or other considerations. For instance, Figure 1 shows 513 indexed measurements of the seasonal (April -August) average of surface ozone data in 1999 in Eastern US. It is impossible for us to detect ozone data at all locations. Therefore, an important task in the analysis of geostatistical data is the reconstruction of the surface of $\mathbf{z}$ over the entire domain. Typically two steps are involved: one is the estimation of unknown parameters, the other is statistical prediction of $\mathbf{z}(\mathbf{s})$ over a fine grid of locations, which is called "kriging" in geostatistics, see Krige (1951).

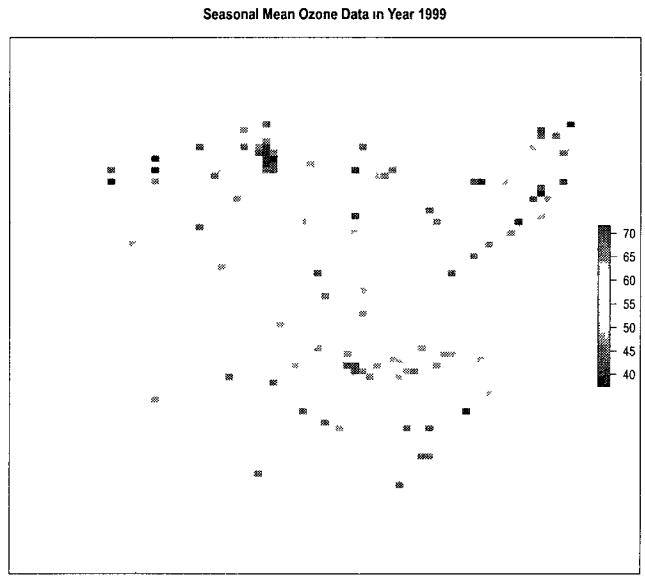


Figure 1: *Example of geostatistical data: seasonal (April-August) average of surface ozone data in 1999 in Eastern US (from UCAR)*

1.2 Lattice Data

Lattice data are spatial data where the domain D is fixed and discrete, and typically in R^2 defined by areas, which means it is not random and it is countable.

Examples include observations collected by town, ZIP code or remote sensing data reported by pixels. Spatial locations with lattice data are often referred to as sites or regions. Two types of lattice data that are usually discussed are *regular* lattice data and *irregular* lattice data, as shown in Figure 2.

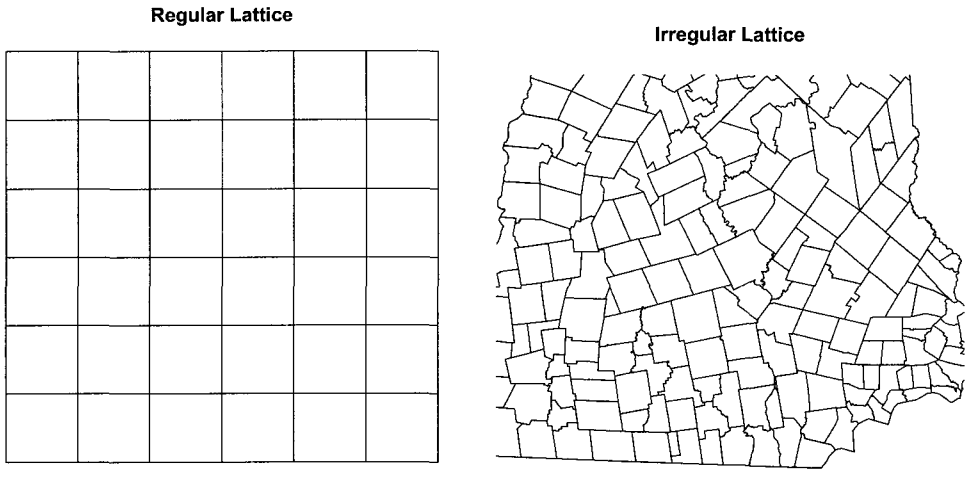


Figure 2: *Left: A generic regular lattice; Right: Irregular lattice for southern New Hampshire towns*

To statistically model the lattice data, we need to spatially index the areas in order to develop measures of spatial dependence. For example, we could utilize the distance between the centroids of any two areas, or we could pose an association between two areas that share a common border. One thing worthy to mention here is: in most of cases, due to the confidentiality and other considerations, for example, individual cancer information, lattice data are spatially aggregated over some areal regions A_i, and thus the notation $\mathbf{z}(A_i)$ is usually used. Due to the discrete nature of space in lattice data analysis, spatial interpolation which is a major goal of geostatistics is not possible. Instead, the goal of lattice analysis is typically to explain uncertainty via a latent smooth process, as well as by assessing

the relationship between observations and other covariates. Examples are land cover classifications (Lunetta and Lyon, 2004), spatial disease mapping (Lawson, 2008), and regional climate model output (Sain, Furrer and Cressie, 2007).

CHAPTER II

Modeling Spatial Data

In this chapter, we review several statistical models for spatial data. Geostatistical models are usually used for point-referenced data and conditional autoregressive (CAR) models are preferred for lattice data. Two parameterizations of modified CAR models are also reviewed.

With the assumption of spatial dependence among responses of $\mathbf{z}(\mathbf{s})$, spatial models are in some way an extension of statistical models for repeated measurement data and longitudinal data. Spatial statistical models are usually formulated as regression models. However, the assumption of independent and identically distributed (i.i.d.) residuals is violated. In the last 30 years, researchers have taken great interest in modeling "correlated" data, the correlation of which are often captured by unknown parameters. In order to estimate parameters efficiently and accurately, one would like to capture the correlation structure with only few parameters.

2.1 Spatial Gaussian Process

The Gaussian assumption is always favorable because of its convenient properties, especially those related to linearity. Following the regression paradigm we model the response variables with a trend (mean) structure and an additive stochastic

structure describing variation and covariation among the responses. Thus a spatial regression model can be written as

$$\mathbf{z}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}(\mathbf{s}),$$

where the trend structure is $\mathbf{X}\boldsymbol{\beta}$ with a $n \times p$ covariate matrix $\mathbf{X}$ and $p \times 1$ vector $\boldsymbol{\beta}$ of regression parameters. $\boldsymbol{\epsilon}(\mathbf{s})$ captures the correlation structure for the response $\mathbf{z}(\mathbf{s})$. Adding the Gaussian assumption we then write

$$\boldsymbol{\epsilon}(\mathbf{s}) \sim N(0, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

Here, $\boldsymbol{\theta}$ is a vector of unknown correlation parameters that will be specified by particular models, and $\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the variance-covariance matrix that determines the dependency among responses. Therefore,

$$\mathbf{z}(\mathbf{s}) \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})).$$

A primary goal is to perform parameter estimation for $\boldsymbol{\beta}, \boldsymbol{\theta}$ and σ^2. The likelihood function of these parameters can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta})\right\}.$$

In the frequentist framework, since the likelihood function is nonlinear in the parameters, the method of choice is maximum likelihood estimation. Here, we will

attempt to find the set of parameter values that minimizes $-2 *$ log-likelihood:

$$-2 \log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = C + n \log(\sigma^2) + \log(|\boldsymbol{\Sigma}(\boldsymbol{\theta})|) + \frac{1}{\sigma^2}(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta})^T \boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta}),$$

where C is a constant term. There are two computationally "expensive" parts: the determinant $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and the inverse $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$. For large data sets (large n), the computation time required can be overwhelming to the typical computer. This issue will be discussed in Chapter III.

2.2 Geostatistical Models

Geostatistical models for point-referenced spatial data have been widely studied in the past few decades. Here, since the popularization of the seminal work of Matheron (Matheron, 1963), the variance-covariance matrix $\sigma^2\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is directly modeled by a covariance function $C(\mathbf{h}) = \sigma^2\rho(\mathbf{h}; \boldsymbol{\theta})$ ($\mathbf{h}$ is the lag-vector) that has only few parameters and is assumed to be second-order stationary. One fact worthy of mentioning here is: for a covariance function $C(\mathbf{h})$ to be valid for a second-order stationary spatial process, C must satisfy the positive-definite condition

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j C(\mathbf{s}_i - \mathbf{s}_j) \geq 0$$

for any set of locations and real numbers a_i, a_j.

Among the most referred to one-parameter correlation functions with range parameters $\theta > 0$ are Exponential, Spherical and Gaussian, where θ is the range parameter. Probably the most important and rich class of two-parameter corre-

Table 1: *Examples of spatial correlation functions that define variance-covariance matrices*

Exponential	$\rho(h) = \exp(-\frac{h}{\theta})$	$h > 0$
Spherical	$\rho(h) = 1 - \frac{3}{2}\frac{h}{\theta} + \frac{1}{2}(\frac{h}{\theta})^3$	$0 < h \leq \theta$
Gaussian	$\rho(h) = \exp(-\frac{h^2}{\theta^2})$	$h > 0$
Matérn	$\rho(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \cdot \left(\frac{2\sqrt{\nu}}{\theta}h\right)^{\nu} \cdot \mathcal{K}_{\nu}\left(\frac{2\sqrt{\nu}}{\theta}h\right)$	$h > 0, \theta > 0, \nu > 0, \mathcal{K}_{\nu}()$:the modified Bessel function of order ν

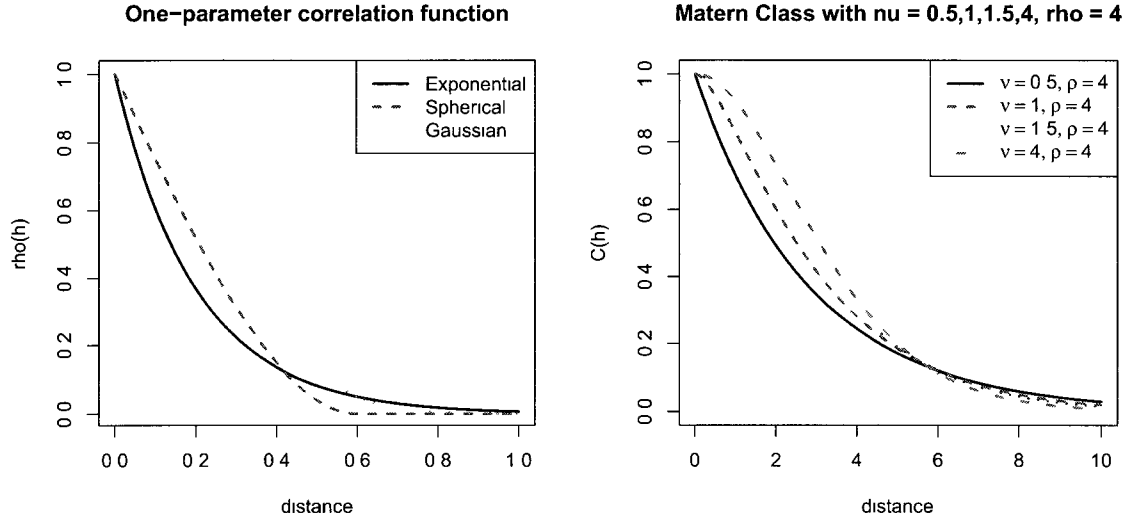


Figure 3: *Left panel: Exponential, Spherical, Gaussian correlation functions with parameters* $\theta = 0.2, 0.6,$ *and* $0.6/\sqrt{3}$*, respectively; Right panel: Matérn class of correlation functions with range parameter* $\theta = 4$ *and smoothness parameter* $\nu = 0.5, 1, 1.5,$ *and* 4.

lation function is the Matérn class. These functions are listed in Table 1, and corresponding correlation graphs are shown in Figure 3.

2.3 Autoregressive Models

Autoregressive models are popular in time series analysis and are denoted as, $AR(p)$, where p is the order. The $AR(p)$ model for a time series $x(t), t = 1, 2, \ldots$ can be written as

$$x_t = c + \sum_{i=1}^{p} \varphi_i x_{t-i} + \varepsilon_t,$$

where $\varphi_1, \ldots, \varphi_p$ are the autoregressive parameters of the model, c is a constant and ε_t is white noise.

In spatial statistics, particularly for lattice data, a spatial neighbor structure in fact introduces a local ordering which then allows us to introduce autoregressive and moving average (*ARMA*) models analogous to similar models in time series analysis. While autoregression ideas are similar in both spatial statistics and time series analysis, they still have a key difference. In regularly spaced time series data, the time index t, since it is 1-dimensional, naturally provides a higher order "neighbor" structure: first order $(t, t-1)$, second order $(t, t-2)$ and so on. With irregular spatial lattices, one rarely considers higher order neighbor stuctures. However a weighting scheme (using ω_{ij} distance based weights, say) would implicitly provide higher order neighbors,

$$\hat{z}_i = \frac{\sum_j \omega_{ij} z_j}{\sum_j \omega_{ij}}.$$

However, the choice of weight functions can be arbitrary and somewhat subjective.

2.3.1 Conditional Autoregressive Model (CAR)

One of the most popular spatial autoregressive models is the *conditional autoregressive or CAR model* (Besag, 1974). Here, "conditional" means: data observed at one location is modeled conditionally on the data collected at neighboring locations. Let $\mathbf{z} = (z_1, \cdots, z_n)^T$ be the observations taken over a spatial lattice at locations $(\mathbf{s}_1, \cdots, \mathbf{s}_n)$. The full conditional distributions of the z_i given all other values are assumed to only depend on the values z_j at the neighboring sites j of i; in which

case we write $j \sim N(i)$. For Gaussian spatial processes, we set

$$z_i|\mathbf{z}_{-i} \sim N\left(\mu_i + \sum_{j\, j\sim N(i)} b_{ij}(z_j - \mu_j), \tau_i^2\right).$$

The condition that z_i given all others z_{-i} only depends on the neighbors of location s_i is specified under the Markov random field (MRF) paradigm (Besag, 1974; Rue and Held, 2005). Through Brooks' Lemma we can obtain the joint distribution from all full conditional distributions, as

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \mathbf{Q}^{-1}),$$

where $\mathbf{Q} = \mathbf{M}^{-1}(\mathbf{I} - \mathbf{B})$, $\mathbf{B} = (b_{ij})$ and $\mathbf{M} = \text{diag}(\tau_i^2), i = 1, \cdots, n$.

Various parameterizations for the b_{ij} have been suggested. The most parsimonious parameterization of the GMRF assumes a variance parameter $\sigma^2/k_i = \frac{1}{\tau_i^2}$, where k_i is the number of neighbors of location s_i, and an interaction parameter ϕ such that $\mathbf{B} = \phi\mathbf{C}$, where $\mathbf{C}$ is the weight matrix defined by one of the weight functions suggested by Pettitt, Weir,and Hart (2002) (ie: linear, uniform, or reciprocal). This will result in the CAR conditional representation

$$z_i|\mathbf{z}_{-i} \sim N\left(\mu_i + \phi \sum_{j\, j\sim N(i)} \frac{c_{ij}}{k_i}(z_j - \mu_j), \frac{\sigma^2}{k_i}\right)$$

and the joint representation $\mathbf{z} \sim N(\mu, \mathbf{Q}^{-1})$, where

$$\mathbf{Q} = \frac{1}{\sigma^2} \begin{pmatrix} k_1 & 0 & \cdots & 0 \\ 0 & k_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & k_n \end{pmatrix} \begin{pmatrix} 1 & -\phi c_{12}/k_1 & \cdots & -\phi c_{1n}/k_1 \\ -\phi c_{21}/k_2 & 1 & \cdots & -\phi c_{2n}/k_2 \\ \vdots & \vdots & \ddots & \vdots \\ -\phi c_{n1}/k_n & -\phi c_{n2}/k_n & \cdots & 1 \end{pmatrix} = \frac{1}{\sigma^2}(\mathbf{K} - \phi \mathbf{C}).$$

Note that the most popular neighbor structure defined by a lattice can be expressed through a neighbor incidence matrix $\mathbf{C}$, the elements of which are determined by

$$c_{ij} = \begin{cases} 1, & \text{if site } i \text{ is a neighbor of site } j \\ 0, & \text{otherwise .} \end{cases}$$

The above $\mathbf{C}$ is the first-order structure. The weight matrix $\mathbf{K}$ is required to ensure that $\mathbf{Q}$ is symmetric and positive definite. Another difficulty with this parameterization is that the parameter space for the spatial interaction term, ϕ, is restricted and its range depends on the eigenvalues of $\mathbf{Q}$ for the same reason (Rue and Held, 2005, Chapter 2).

It has been noted that for an underlying smooth process, a Markov random field can incorporate a higher order structure for a regular lattice data, while a distance-based weight function can be assumed for non-regular lattice data. For example, a higher order structure was considered in Rue and Held, 2005, Chapter 5, where the conditional expectation of $E(z_{ij}|z_{-ij})$ is parameterized with multiple

parameters:

$$E(z_{ij}|z_{-ij}) = -\frac{1}{\theta_0}\left(\theta_1 \sum_{i_1 j_1 \in N(ij)} z_{i_1 j_1} + \theta_2 \sum_{i_2 j_2 \in N(ij)} z_{i_2 j_2} + \theta_3 \sum_{i_3 j_3 \in N(ij)} z_{i_3 j_3} + \cdots\right).$$

The $z_{i_1 j_1}$ are the first-order neighbors of z_{ij}, $z_{i_2 j_2}$ the second-order neighbors, and so on. Taking a square lattice as an example, the east-west and north-south neighbors are referred to as first-order neighbors while the four nearest diagonal locations are called second-order neighbors. Figure 4 shows the first three order structures.

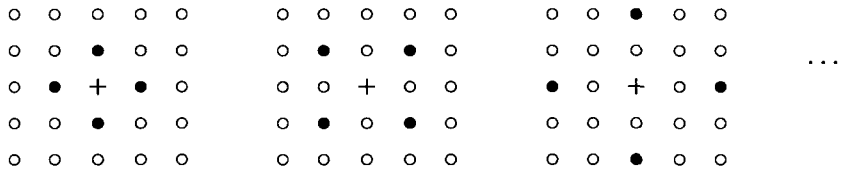


Figure 4: *Higher-order structure for a square lattice. The left graph shows the first-order neighbors for a site labeled as +; the middle graph shows the second-order neighbors; the right graph shows the third-order neighbors.*

The parameters $(\theta_0, \theta_1, \theta_2, \cdots)$ define the higher-order spatial associations. The higher-order CAR model can be treated as an extension to the first-order CAR model. If the underlying process is very smooth, higher-order parameters will be significantly different from zero, while when the process is coarse, most higher-order parameters can be assumed to be equal zero. One difficulty for this higher-order CAR model is the parameter estimation. On the one hand, it is very subjective to determine how many higher-order neighbors we need to include in the model. On the other hand, as more neighbors are included in the model, the accuracy of parameter estimation will be decreased.

As an alternative to a higher order neighbor structure, a distance-based weight function can be assumed for any location $\{s_i : i = 1, 2, \cdots, n\}$ that is surrounded by its neighbors. Sites s_i and s_j are neighbors if and only if they lie within some critical distance $\delta > 0$ of each other. Let d_{ij} denote the Euclidean distance between sites i and j and let $\gamma : [0, \infty] \rightarrow [0, \infty]$ be continuous and non-increasing on $[0, \delta)$ and zero on $[\delta, \infty)$. A $n \times n$ matrix $\boldsymbol{\gamma} = [\gamma_{ij}]$ can be defined by

$$\gamma_{ij} = \begin{cases} \gamma(d_{ij}), & i \neq j \\ 0, & i = j \ . \end{cases}$$

Three distance-based functions for γ_{ij}, uniform, linear, and reciprocal, are suggested by Pettitt, Weir, and Hart (2002). In the CAR model, using different weighting schemes, the **C** matrix will be replaced by a matrix $\boldsymbol{\gamma}$ defined by functions such as those listed in Table 2.

Table 2: *Examples of distance-based weight functions : Uniform, Linear and Reciprocal*

Uniform	$\gamma(d_{ij}) = \begin{cases} 1, & 0 < d_{ij} < \delta \\ 0, & d_{ij} \geq \delta \end{cases}$
Linear	$\gamma(d_{ij}) = \begin{cases} 1 - \frac{d_{ij}}{\delta}, & 0 < d_{ij} < \delta \\ 0, & d_{ij} \geq \delta \end{cases}$
Reciprocal	$\gamma(d_{ij}) = \begin{cases} \frac{\delta}{d_{ij}} - 1, & 0 < d_{ij} < \delta \\ 0, & d_{ij} \geq \delta \end{cases}$

2.3.2 Simultaneous Autoregressive Model (SAR)

For a Gaussian spatial process $\mathbf{z}(\mathbf{s})$, instead of modeling z_i as conditionally dependent on its neighbors $z_j, j \in N(i)$, we model each z_i as a linear combination of all other $z_j, j \neq i$, where the coefficients of the linear combination are denoted by b_{ij} (Note that by definition $b_{ii} = 0$). Then we can write

$$z_i = \sum_j b_{ij} z_j + \epsilon_i, \quad i = 1, 2, \cdots, n,$$

where $\epsilon_i \sim N(0, \sigma_i^2)$. In matrix form, let $\mathbf{z} = \begin{pmatrix} z_1, & z_2, & \cdots, & z_n \end{pmatrix}^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \cdots, \epsilon_n)^T$, $\mathbf{B} = (b_{ij})$, and $\mathbf{D} = \text{diag}\left(\sigma_1^2, \cdots, \sigma_n^2\right)$, then

$$(\mathbf{I} - \mathbf{B})\mathbf{z} = \boldsymbol{\epsilon}.$$

If $\mathbf{I} - \mathbf{B}$ is full rank, we can write $\mathbf{z} = (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\epsilon}$, and thus can obtain

$$\mathbf{z} \sim N\left(\mathbf{0}, (\mathbf{I} - \mathbf{B})^{-1}\mathbf{D}((\mathbf{I} - \mathbf{B})^{-1})'\right).$$

In a regression context, the SAR model is applied to model the residuals $\mathbf{U} = \mathbf{z} - \mathbf{X}\boldsymbol{\beta}$, rather than $\mathbf{z}$ itself. This imitates the first order autoregressive model ($AR(1)$) in time series modeling of the residuals from a linear regression trend. The model now can be written as

$$\begin{cases} \mathbf{U} = \mathbf{z} - \mathbf{X}\boldsymbol{\beta} \\ \mathbf{U} = \mathbf{B}\mathbf{U} + \boldsymbol{\epsilon} \end{cases}.$$

Substituting $\mathbf{U}$ from the first into the second equation, we obtain an attractive form

$$\mathbf{z} = \mathbf{B}\mathbf{z} + (\mathbf{I} - \mathbf{B})\mathbf{X}\beta + \epsilon.$$

The expression above shows that $\mathbf{z}$ can be viewed as a weighted average of its own (neighboring) values and the trend (regression) function. If $\mathbf{B}$ is the zero matrix, we obtain an OLS regression; if $\mathbf{B} = \mathbf{I}$, we obtain a purely spatial model. Note that the SAR model representation will break down for non-Gaussian data, hence the SAR model is not used for generalized linear models (GLM) with say Poisson counts or with binary response data.

One important thing to note here is that SAR models are well suited to maximum likelihood estimation but not at all for MCMC fitting of Bayesian models. That is, the -2log-likelihood function is

$$\sum_i \log(\sigma_i) - \log(|\mathbf{I} - \mathbf{B}|) + (\mathbf{z} - \mathbf{X}\beta)^T(\mathbf{I} - \mathbf{B})\mathbf{D}^{-1}(\mathbf{I} - \mathbf{B})^T(\mathbf{z} - \mathbf{X}\beta).$$

Since there is no matrix inversion required, computing the determinant is relatively quickly. Note that the process is usually accelerated by using diagonally dominant, sparse matrix approximations. Thus, iterative maximization is commonly quite efficient in terms of computer time. However, unlike the CAR random effects that are defined through full conditional distributions, the full conditional distribution for the SAR effects have no convenient form. As a result in Bayesian hierarchical model with large n, the computation of such distributions will be expensive.

2.3.3 Comparison of CAR and SAR Models

Cressie (1993) shows that any SAR model can be represented as a CAR model, but gives a counterexample to prove that the converse is not true. Both CAR and SAR models incorporate spatial dependence parameters, say, ρ_c and ρ_s respectively for CAR and SAR. Both parameters have restrictions which are controlled by the eigenvalues of the lattice neighbor matrices. In addition, Wall (2004) shows as the ρ_c and ρ_s increase from zero to the upper end of the parameter space, the implied correlations between all sites monotonically increase. However, when $\rho_c, \rho_s < 0$, the correlations are not monotone, which gives another reason to avoid negative spatial correlation parameters. Moreover, the ranking of the implied correlations from largest to smallest is not consistent as ρ_s and ρ_c change. For example, under the first-order neighbor structure of the 48 contiguous U.S. states lattice, she models the statewide average SAT verbal scores and finds that when $\rho_c = .49$ the Corr(Alabama; Florida)= .20 and the Corr(Alabama; Georgia)= .16. But, when $\rho_c = .975$ the correlation between Alabama and Georgia is greater than the correlation between Alabama and Florida.

2.3.4 The Pettitt *et al.* Parameterization of the CAR Model

The likelihood for a spatial autoregressive Gaussian process $\mathbf{z} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{Q}(\boldsymbol{\theta})^{-1})$ is

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = \frac{|\mathbf{Q}(\boldsymbol{\theta})|^{1/2}}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta}\right)^T \mathbf{Q}(\boldsymbol{\theta})\left(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta}\right)\right\}.$$

Both CAR and SAR models are computationally more efficient than geostatistical models since no repeated inversions of the variance-covariance matrix are needed when likelihood methods are performed. However, the issue of finding the determinant in the likelihood remains. There are several methods that can be used to attempt this, but most are cumbersome for large data sets. This issue is addressed in a paper by Pettitt, Weir, and Hart (2002), and many other papers.

Pettitt, Weir, and Hart (2002) propose a particular parameterization of the ordinary CAR model that proves to be computationally efficient. In this model, the precision matrix is created in such a way that the determinant is computed easily and in closed form. It also lends itself to the addition of covariates without complicating the model. This computational efficiency is particularly advantageous for large irregular lattices and weighting schemes applied to continuous space data.

Recall that $\mathbf{z}$ is a realization from a conditional autoregressive Gaussian process given by

$$\mathbf{z} \sim N(\mu, \sigma^2(\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}),$$

where $\mathbf{C} = (c_{ij})$ is a matrix that has zeros along its main diagonal and $\mathbf{M} = \text{diag}(m_{11}, m_{22}, \cdots, m_{nn})$ is a diagonal matrix chosen so that the matrix $\mathbf{Q}^{-1} = (\mathbf{I} - \mathbf{C})^{-1}\mathbf{M}$ is symmetric and positive-definite. Pettitt, Weir, and Hart use the matrix γ (the elements of which are defined in Table 2), together with a spatial dependence parameter, ϕ, to construct the precision matrix, that is, the inverse of the variance-covariance matrix (if it is non-singular). The matrices $\mathbf{M}$ and $\mathbf{C}$ are defined so that

the terms of the matrix $\mathbf{C}$ are

$$c_{ij} = \begin{cases} \frac{\phi\gamma_{ij}}{1+|\phi|\sum_{k\in N(i)}\gamma_{ik}}, & j \neq i \\ 0, & j = i \end{cases},$$

and the terms of the matrix $\mathbf{D}$ are

$$m_{ii} = \frac{1}{1+|\phi|\sum_{k\in N(i)}\gamma_{ik}}, \; i = 1, 2, \cdots, n,$$

and thus the matrix $\mathbf{Q} = \mathbf{M}^{-1}(\mathbf{I} - \mathbf{C})$ is symmetric with

$$Q_{ij} = \begin{cases} 1 + |\phi| \sum\limits_{k\in N(i)} \gamma_{ik}, & i = j \\ -\phi\gamma_{ij}, & i \neq j \end{cases},$$

and it is also positive definite since it is a symmetric diagonally dominant matrix for all $-\infty < \phi < \infty$. Therefore, the precision matrix is symmetric and positive-definite, making it a valid precision matrix.

The three conditional moments, the mean, the variance and the covariance can be written as

$$E(z_i|\mathbf{z}_{-i}) = \mu_i + \frac{\phi}{1+|\phi|\sum_{k\in N_i}\gamma_{ik}} \sum_{j\in N(i)} \gamma_{ij}(z_j - \mu_j),$$

$$\mathrm{Var}(z_i|\mathbf{z}_{-i}) = \frac{\sigma^2}{1+|\phi|\sum_{k\in N_i}\gamma_{ik}},$$

$$\mathrm{corr}(z_i, z_j|\mathbf{z}_{-\{ij\}}) = \frac{\phi\gamma_{ij}}{\sqrt{(1+|\phi|\sum_{k\in N(i)}\gamma_{ik})(1+|\phi|\sum_{k\in N(j)}\gamma_{jk})}}.$$

Without the addition of the unit in the denominator of the definitions of c_{ij} and m_{ii} we would have a familiar intrinsic CAR model (Besag and Kooperberg, 1995) with

$\phi > 0$. The parameter ϕ measures the strength of the spatial dependency. There is no spatial dependency, if $\phi = 0$. This corresponds to unstructured random effects. As $|\phi| \to \infty$ while other parameters remain fixed, Pettitt *et al.* state that $|\phi|^{-1}\mathbf{Q}$ tends to an intrinsic CAR model,

$$\lim_{|\phi|\to\infty} |\phi|^{-1}\mathbf{Q} = \begin{cases} \sum_{k\in N(i)} \gamma_{ik}, & i = j \\ -\text{sign}(\phi)\gamma_{ij}, & i \neq j \end{cases}.$$

In contrast to the intrinsic CAR, the joint distribution of $\mathbf{z}$ is a proper distribution, which leads to a proper posterior when this model is used as a prior distribution in a Bayesian analysis.

The determinant of the precision matrix $\mathbf{Q}$ can now be solved through a reasonably efficient numerical technique as follows. Define

$$\mathbf{D} = \text{diag}\left(\sum_{j\in N(i)} \gamma_{ik},\ i = 1, 2, \cdots, n\right)$$

and γ be the matrix with diagonal equal to zero and off-diagonal equal to γ_{ij}, then, $\mathbf{Q}$ can be written in the form

$$\begin{aligned} \mathbf{Q} &= \mathbf{I} + |\phi|\mathbf{D} - \phi\gamma \\ &= \begin{cases} \mathbf{I} - \phi(\gamma - \mathbf{D}), & \phi > 0 \\ \mathbf{I}, & \phi = 0 \\ \mathbf{I} - \phi(\gamma + \mathbf{D}), & \phi < 0 \end{cases}. \end{aligned}$$

If $\{\lambda_i^1 : i = 1,2,\cdots,n\}$ are the eigenvalues of $\gamma - \mathbf{D}$ and $\{\lambda_i^2 : i = 1,2,\cdots,n\}$ are the eigenvalues of $\gamma + \mathbf{D}$, then the eigenvalues $\{\xi_i : i = 1,2,\cdots,n\}$ of $\mathbf{Q}$ can be determined by

$$\xi_i = \begin{cases} 1 - \phi\lambda_i^1, & \phi > 0 \\ 1, & \phi = 0 \\ 1 - \phi\lambda_i^2, & \phi < 0 \end{cases}.$$

The determinant of $\mathbf{Q}$ can now be obtained by taking the product of the appropriate set of eigenvalues,

$$|\mathbf{Q}| = \begin{cases} \prod_i (1 - \phi\lambda_i^1), & \phi > 0 \\ 1, & \phi = 0 \\ \prod_i (1 - \phi\lambda_i^2), & \phi < 0 \end{cases}.$$

Therefore, once the eigenvalues of $\gamma - \mathbf{D}$ and $\gamma + \mathbf{D}$ are known, the determinant of $\mathbf{Q}$ may be computed quickly for any value of ϕ.

In this dissertation, we will restrict ϕ to only have positive values, as discussed in section 2.3.3. Thus, in order to calculate the determinant of $\mathbf{Q}$, we only need to find the eigenvalues of $\gamma - \mathbf{D}$. A singular value decomposition (SVD) of $\gamma - \mathbf{D}$ will result in the desired eigenvalues and eigenvectors. Assume the decomposition

$$\mathbf{F}^T(\gamma - \mathbf{D})\mathbf{F} = \mathbf{\Lambda}$$

exists with $\mathbf{\Lambda} = \text{diag}\{\lambda_i : i = 1,2,\cdots,n\}$, where λ_i is the i-th eigenvalue of $\gamma - \mathbf{D}$. The columns of $\mathbf{F}$ contain the corresponding eigenvectors. Once the eigenvalues

are found, the eigenvalues of $\mathbf{Q}$ can be expressed as

$$\xi_i = 1 - \phi\lambda_i,\ i = 1, 2, \cdots, n$$

and the determinant of $\mathbf{Q}$ can be easily computed as

$$|\mathbf{Q}| = \prod_i (1 - \phi\lambda_i),\ \ \phi > 0.$$

2.3.5 The Czado's Parameterization of CAR Model

It can be noted that when $\phi \to \infty$, the conditional variance $\text{Var}(z_i|\mathbf{z}_{-i}) = \frac{\sigma^2}{1+|\phi|\sum_{k\in N_i}\gamma_{ik}}$ decreases to zero, which is restrictive. Czado and Prokopenko (2008) propose a modified Pettitt's model, where the full conditional distribution for $\mathbf{z}$ is given as follows

$$z_i|\mathbf{z}_{-i} \sim N\left(\mu_i + \frac{\phi}{1+|\phi|\sum_{k\in N(i)}\gamma_{ik}}\sum_{j\in N(i)}\gamma_{ij}(z_j - \mu_j), \frac{(1+|\phi|)\tau^2}{1+|\phi|\sum_{k\in N(i)}\gamma_{ik}}\right).$$

The only difference to Pettitt's model is the conditional variance. In Czado's model, the asymptotic conditional variance

$$\frac{(1+|\phi|)\tau^2}{1+|\phi|\sum_{k\in N(i)}\gamma_{ik}} \to \frac{\tau^2}{\sum_{k\in N(i)}\gamma_{ik}}, \quad \text{as } |\phi| \to \infty.$$

The intrinsic CAR model still arises in the limit, when $|\phi| \to \infty$. This model has the same behavior as Pettitt's CAR parameterization when $|\phi|$ goes to zero (no spatial dependency), and all partial correlations between z_i and z_j given all the other sites

are the same. Further, Czado's CAR model has larger conditional and marginal variance for z_i than the original Pettitt's model for $\phi > 0$, thus allowing for a larger variability.

Czado's CAR model parameterization will result in

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \mathbf{Q}^{-1}),$$

where $\mathbf{Q} = \mathbf{M}^{-1}(\mathbf{I} - \mathbf{C})$. The diagonal elements of $\mathbf{M}$ are

$$m_{ii} = \frac{1 + |\phi|}{1 + |\phi| \sum_{k \in N(i)} \gamma_{ik}}, \quad i = 1, 2, \cdots, n$$

and elements c_{ij} of matrix $\mathbf{C}$ are defined exactly the same as Pettitt's model,

$$c_{ij} = \begin{cases} \frac{\phi \gamma_{ij}}{1+|\phi| \sum_{k \in N(i)} \gamma_{ik}}, & j \neq i \\ 0, & j = i \ . \end{cases}$$

Thus,

$$Q_{ij} = \begin{cases} \frac{1+|\phi| \sum_{k \in N(i)} \gamma_{ik}}{1+|\phi|} = 1 + \frac{|\phi|}{1+|\phi|}(\sum_{k \in N(i)} \gamma_{ik} - 1), & i = j \\ -\frac{\phi}{1+|\phi|}\gamma_{ij}, & i \neq j \ . \end{cases}$$

Now, define

$$\mathbf{D} = \text{diag}\left(\sum_{k \in N(i)} \gamma_{ik} - 1, \ i = 1, 2, \cdots, n \right)$$

and

$$\psi = \frac{\phi}{1 + |\phi|}.$$

Following a similar approach in Pettitt's CAR model, we can get

$$\mathbf{Q} = \mathbf{I} + |\psi|\mathbf{D} - \psi\boldsymbol{\gamma},$$

and if $\{\lambda_i^1 : i = 1, 2, \cdots, n\}$ are the eigenvalues of $\boldsymbol{\gamma} - \mathbf{D}$ and $\{\lambda_i^2 : i = 1, 2, \cdots, n\}$ are the eigenvalues of $\boldsymbol{\gamma} + \mathbf{D}$, then the determinant of $\mathbf{Q}$ is

$$|\mathbf{Q}| = \begin{cases} \prod_i (1 - \psi\lambda_i^1), & \psi > 0 \\ 1, & \psi = 0 \\ \prod_i (1 - \psi\lambda_i^2), & \psi < 0 \end{cases}$$

which can be computed quickly for any value of ψ. Also, we will restrict $\psi > 0$, and in this case, the range of ψ will be $[0, 1]$. $\psi = 0$ indicates no spatial dependency and $\psi = 1$ denotes the intrinsic CAR model.

CHAPTER III

Computational Efficiency: The Big n Problem

3.1 The Big n Problem

As discussed in Chapter II, the parameter estimation procedures require the minimization of -2log-likelihood function for a Gaussian process (n denotes the sample size):

$$-2\log L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2) = C + n\log(\sigma^2) + \log(|\boldsymbol{\Sigma}(\boldsymbol{\theta})|) + \frac{1}{\sigma^2}(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{z}(\mathbf{s}) - \mathbf{X}\boldsymbol{\beta}).$$

Commonly the Gaussian elimination and LU decomposition are used to obtain the determinant and the inverse of a general square matrix. Both algorithms are numerically equivalent to an order $O(n^3)$. Here, the order $O(n^3)$ of the "FLOPS" (**FL**oating point **OP**erations per **S**econd) measures the computational complexity of mathematical operations in relation to the matrix size n. For large n, the evaluation procedure of the likelihood function will be computationally expensive for repetitive evaluations of the determinant $|\boldsymbol{\Sigma}(\boldsymbol{\theta})|$ and the inverse $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$, either under the iterative numerical maximization for maximum likelihood estimation (MLE) or under the iterative Markov chain Monte Carlo evaluation for estimation of a Bayesian posterior distribution. This is called "the big n problem" in spatial statis-

tics. When models are extended to multivariate models, say, with m measurements, at a location or extended to spatio-temporal models with spatial time series at T time points, they will respectively lead to larger matrices: $nm \times nm$ matrix and $nT \times nT$. The problem gets worse as n gets larger. The objective of this Chapter is to review some suggestions and approximations for handling spatial process models in this case.

3.2 Approximation Methods to the Big n Problem

3.2.1 Cholesky Decomposition

Researchers have proposed several approximation methods that relate to the variance-covariance matrices. Since these matrices are symmetric and positive definite, they have a special decomposition, called the *Cholesky decomposition*. A matrix $\mathbf{A}$ can be written as $\mathbf{A} = \mathbf{F}^T\mathbf{F}$, where $\mathbf{F}$ is an upper triangular matrix and is called the "square root" of the matrix $\mathbf{A}$ (if it is real). It can be clearly seen by applying the $\mathbf{LDU}$ decomposition to $\mathbf{A}$. Since $\mathbf{A}$ is symmetric, we can obtain $\mathbf{A} = \mathbf{LDL}^T$, where $\mathbf{L}$ is unit lower triangular and $\mathbf{D}$ is a diagonal matrix with all entries positive. Then, we can write $\mathbf{F}^T = \mathbf{LD}^{1/2}$ to get the Cholesky root.

The Cholesky algorithm can be expressed as follows: for $i = 1, 2, \cdots, n$

$$\begin{cases} f_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} f_{ik}^2\right)^{1/2} \\ f_{ji} = \frac{a_{ij} - \sum_{k=1}^{i-1} f_{jk} f_{ik}}{f_{ii}}, \ j = i+1, \cdots, n \ . \end{cases}$$

While the Cholesky algorithm reduces the operation count to $n^3/6$, however, it is still a $O(n^3)$ algorithm. Hence, when n is large, "the big n problem" still persist.

3.2.2 Covariance Tapering

In geostatistics, correlation functions determine variance-covariance matrices. The typical spatial autocovariance is assumed to be nonzero for any finite distance. Thus, these matrices have nonzero elements everywhere. For "the big n problem" issue, the parameter estimation would be extremely slow and even unfeasible. This represents a major disadvantage for geostatistical methods where distance based models are assumed for the spatial correlation $\boldsymbol{\Sigma}(\boldsymbol{\theta})$.

Sparse representations are sometimes useful to speed up matrix inversion and/or determinant calculation. One idea is to force a variance matrix to be sparse (with many zeros), in order to attain matrix operational efficiency. However, one must maintain positive definiteness of any sparse modification of the variance-covariance matrix. The *Covariance tapering* method was proposed and studied by Wendland (1998), Furrer *et al.* (2006), and Kaufman *et al.* (2008). Let $C(h;\theta)$ be the original covariance function, and suppose the $C_\phi(h)$ is a covariance function that is identically zero outside a particular range described by ϕ. Now consider a tapered covariance that is the elementwise product of $C_\phi(h)$ and $C(h;\theta)$:

$$C_{\text{tap}}(h;\phi,\theta) = C(h;\theta) \circ C_\phi(h).$$

The approximation will be obtained by replacing the covariance matrices $C(h;\theta)$

by those defined by $C_{tap}(h;\phi,\theta)$. The product $C_{tap}(h;\phi,\theta)$ preserves some of the shape of $C(h;\theta)$ but its values are identical to zero outside of a fixed location distance range, controlled by ϕ. Of equal importance, $C_{tap}(h;\phi,\theta)$ is a valid covariance, since the elementwise product of two positive definite matrices is again positive definite (Horn and Johnson, 1994, Theorem 5.2.1). As an example of tapering covariance functions, a spherical covariance and two of the Wendland tapers are considered here. They are all valid covariances in $\mathbb{R}^3$. The functions are plotted in Figure 5 and summarized in Table 3. Based on the theory (Furrer *et al.* 2006, section 2) with respect to the Matérn smoothness parameter, the spherical covariance will be used as a taper for the Matérn covariance with its smoothness parameter $\nu \leq 0.5$, Wendland1 for $\nu \leq 1.5$ and Wendland2 for $\nu \leq 2.5$.

Table 3: *Examples of taper covariance functions: Spherical, Wendland1, and Wendland2* ($x_+ = \max\{0, x\}$)

Spherical	$C(h;\phi) = (1-\frac{h}{\phi})_+^2(1+\frac{h}{2\phi})$	$h > 0$
Wendland1	$C(h;\phi) = (1-\frac{h}{\phi})_+^4(1+4\frac{h}{\phi})$	$h > 0$
Wendland2	$C(h;\phi) = (1-\frac{h}{\phi})_+^6(1+6\frac{h}{\phi}+\frac{35h^2}{3\phi^2})$	$h > 0$

One issue involved in this approach is how to determine the "best" distance maximum for the taper which would optimize estimation accuracy and computational efficiency. Generally the "cut-off" is selected by subjective choice, but this issue needs further investigation.

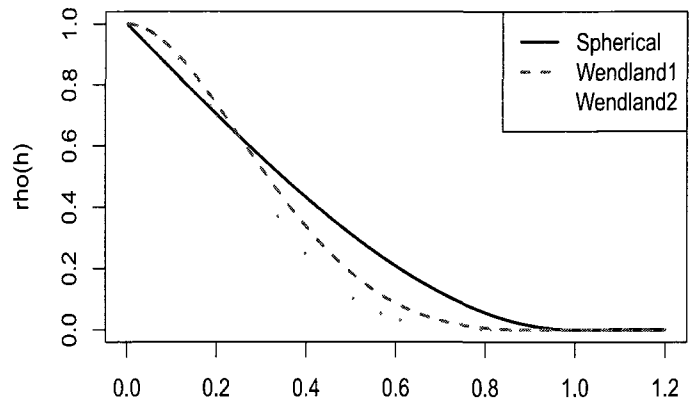


Figure 5: *Spherical, Wendland1, Wendland2 taper covariance functions with taper length 1*

3.2.3 Dimension Reduction

The dimension reduction approach (Higdon 1998, 2002 ; Higdon *et al.* 2003; Lemos and Sanso, 2009) is another strategy for "the big n problem". This is also known as the kernel convolution method with a latent process. The kernel convolution method has been widely and successfully applied in density estimation and regression modeling. An attractive way of using kernel convolution in spatial statistics is to reduce the dimension of variance-covariance matrices, and also to introduce a more general nonstationary spatial process while retaining clear interpretation and permitting analytical calculations. Suppose the process $\mathbf{z}(\mathbf{s}) = (z(\mathbf{s}_1), \cdots, z(\mathbf{s}_n))^T$ can be represented by

$$\mathbf{z}(\mathbf{s}) = \int k(\mathbf{s} - \mathbf{s}')w(\mathbf{s}')d\mathbf{s}'.$$

The corresponding finite approximation will be

$$\mathbf{z}(\mathbf{s}) = \sum_{j=1}^{m} k(\mathbf{s} - \mathbf{s}_j^*)w(\mathbf{s}_j^*),$$

where $w(\mathbf{s})$ is a stationary latent spatial process, k is a kernel function such as, for instance, the popular bivariate Gaussian kernel in the form of

$$k(\mathbf{s}-\mathbf{s}') = \exp\{-\frac{1}{2}(\mathbf{s}-\mathbf{s}')^T\Sigma(\mathbf{s}-\mathbf{s}')\}.$$

One natural choice of Σ would be a diagonal but allowing for componentwise scaling to the separation vector $\mathbf{s}-\mathbf{s}'$. Figure 6 shows the ideas of the latent process

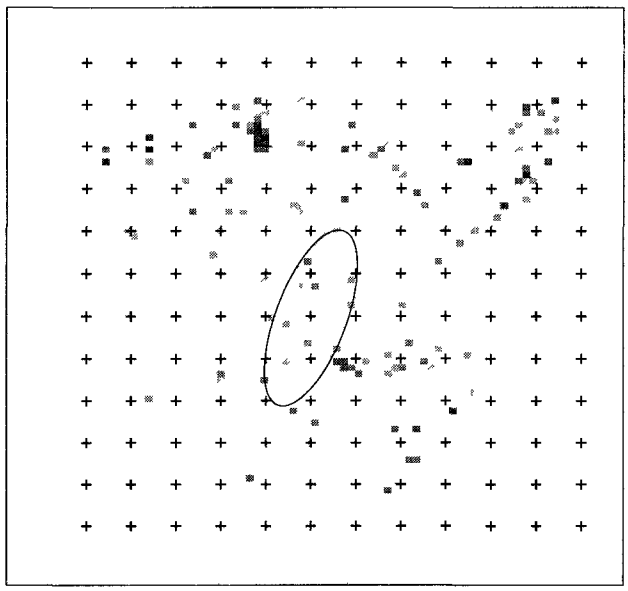


Figure 6: *Convolution method with latent process applied to ozone data. The "+" signs denote spatial locations of the underlying grid process $w(\mathbf{s})$. The ellipse shows the kernel function.*

approximation method. Note that, the kernel function $k(\cdot)$ might be parametric, say with parameters that determine the smoothness of the process, or might be spatially or temporally varying, which allows to capture the local anisotropy and lends itself to specifying models with non-stationary dependence structure (Higdon, 1998).

The finite approximation shows that given the kernel $k(\cdot)$, the process $\{\mathbf{z}(\mathbf{s}_i), i = 1, 2, \cdots, n\}$ in the region can be expressed as a linear combination of the set $\{w(\mathbf{s}_j), j = 1, \cdots, m\}$. Therefore, no matter how large n is, working with the latent process $w(\mathbf{s})$, we only need to consider a $m \times m$ matrix calculations, where $m \ll n$. This will make the computation more efficient.

However, two issues have been raised with regards to this method. First, the computational efficiency depends on the size of the grid of the underlying process. For example, the Gaussian kernel allows for a rather coarse representation of the underlying grid process without any appreciable bias, and thus its computations will be fast. However, specifying $k(\cdot)$ to have the form of a Gaussian density dictates the smoothness of $\mathbf{z}(\mathbf{s})$. As Higdon (1998) points out any choice of kernel k that allows less smooth realizations of $\mathbf{z}(\mathbf{s})$ will generally require a finer grid for the latent process, but this will hinder the computation efficiency. So, how to determine the number of the $\{\mathbf{s}_j^*\}$? The second issue is how sensitive the inference will be to the choice of $\{\mathbf{s}_j^*\}$? These two issues are still under discussion as shown in Lemos and Sansó (2009).

3.2.4 Spectral Basis Representation

Similar to Higdon's convolution methods, Wikle (2002) and Paciorek (2007) suggest using a Fourier basis function to spectrally represent a stationary Gaussian process. However, rather than to specify a coarse (at least not very fine) grid in the convolution method, their model requires a fine grid but the computation of matrix inverses is made more efficient by use of the Fast Fourier Transform (FFT).

Suppose we have an isotropic Gaussian process $\mathbf{z}(\mathbf{s}) = (z(\mathbf{s}_1), \cdots, z(\mathbf{s}_n))^T$, then it can be represented by

$$\mathbf{z}(\mathbf{s}) = \mathbf{K}w(\mathbf{s}^*) + \epsilon(\mathbf{s}).$$

The key idea is to approximate $w(\mathbf{s}^*)$ on a grid $\mathbf{s}^*$, of size $M = M_1 \times M_2$, where M_1 and M_2 are powers of two. The $\mathbf{K}$ is an incidence matrix, which maps each observation location to the nearest grid location in Euclidean space. Evaluated at the grid points, the vector of $w(\mathbf{s}^*)$ can be written as

$$w(\mathbf{s}^*) = \mathbf{\Psi u},$$

where $\mathbf{\Psi}$ is a matrix of orthogonal spectral basis functions, and $\mathbf{u}$ is a vector of complex-valued basis coefficient, $u_m = a_m + b_m i, m = 1, \cdots, M$. To approximate the mean zero stationary isotropic Gaussian process, the basis coefficients have the prior distribution,

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} \sim N(\mathbf{0}, \mathbf{\Sigma}_\theta),$$

where $\mathbf{\Sigma}_\theta$ is a diagonal matrix, parameterized by θ. The conditional variance of $\mathbf{u}$ given the observed data will then be

$$\text{Var}(\mathbf{u}|\mathbf{z}, ...) = \left(\zeta \mathbf{\Psi}^T \mathbf{K}^T \mathbf{K} \mathbf{\Psi} + \mathbf{\Sigma}_\theta^{-1}\right)^{-1}.$$

The sampling scheme requires calculation of $\mathbf{\Psi}^T \mathbf{K}^T \mathbf{K} \mathbf{\Psi}$, which is not feasible for large number of grid points. Wikle and Paciorek's idea is to assume no more than

one observation per grid cell, so that $\mathbf{K} = \mathbf{I}$ can be achieved using a missing data scheme (see Appendix A.2, Paciorek, 2007). Since $\mathbf{\Psi}$ is an orthogonal matrix, then the conditional variance will become

$$\text{Var}(\mathbf{u}|\mathbf{z}, ...) = \left(\zeta\mathbf{I} + \mathbf{\Sigma}_\theta^{-1}\right)^{-1},$$

which is a diagonal matrix. This will result in a computationally efficient approximation to a Gaussian process.

CHAPTER IV

Bayesian Parameter Estimation Using MCMC

This chapter reviews Bayesian approaches for parameter estimations under the hierarchical Markov chain Monte Carlo paradigm. Bayesian methods, which have been largely applied in parameter estimation and statistical inference, serve as an alternative approach to the maximum likelihood estimation method. By modeling both the observed data and any unknown parameters as random variables, it provides a cohesive framework for combining complex data models and external knowledge or expert opinion. In this approach, in addition to specifying the distribution model, let it be $f(\mathbf{y}|\theta)$ for the observed data $\mathbf{y} = (y_1, \cdots, y_n)$ given a vector of unknown parameters $\theta = (\theta_1, \cdots, \theta_k)$, we suppose that θ is a random vector from a *prior* distribution $\pi(\theta, \lambda)$, where λ is a vector of hyperparameters. If λ are known, inference concerning θ is based on the *posterior* distribution,

$$p(\theta|\mathbf{y}, \lambda) = \frac{p(\mathbf{y}, \theta|\lambda)}{p(\mathbf{y}|\lambda)} = \frac{f(\mathbf{y}|\theta)\pi(\theta|\lambda)}{\int f(\mathbf{y}|\theta)\pi(\theta|\lambda)d\theta} \propto f(\mathbf{y}|\theta)\pi(\theta|\lambda).$$

Notice the contribution of both the likelihood of data and the external knowledge to the posterior. In practice, λ will not be known, a second stage distribution (called

hyperprior) $\pi(\lambda)$ will often be required. Thus, we have the posterior

$$p(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta|\lambda)\pi(\lambda).$$

A computational challenge in applying Bayesian methods is that for most realistic problems, the integrations required to do inference in $p(\theta|\mathbf{y}, \lambda)$ are generally not tractable in closed form, and thus must be approximated numerically. In some cases, the posterior distribution can be expressed as a closed form solution, such as when *conjugate priors* are assumed for unknown parameters. However, due to the presence of unknown quantities, some intractable integrations remain. Markov chain Monte Carlo (MCMC) integration methods, thus, have been developed and serve as the most popular tools in Bayesian practice. In this dissertation, we will introduce the two most popular MCMC algorithms, the Gibbs sampler and the Metropolis-Hastings algorithm.

4.1 Gibbs Sampler

Suppose our model contains k parameters, $\theta = (\theta_1, \cdots, \theta_k)'$. To implement the Gibbs sampler, we must assume that samples can be generated from each of the full conditional distributions

$$p(\theta_i|\theta_{j,j\neq i}, \mathbf{y}), \quad i = 1, 2, \cdots, k.$$

Such samples might be available directly, say, the posterior distributions are normal or gamma; or indirectly, say using rejection sampling approach. In this latter case,

two popular alternatives are the adaptive rejection sampling, and the Metropolis algorithm described in the next section. In both cases, the collection of full conditional distributions uniquely determine the joint posterior distribution, $p(\boldsymbol{\theta}|\mathbf{y})$.

Given an arbitrary set of starting value $\{\theta_2^{(0)}, \theta_3^{(0)}, \cdots, \theta_k^{(0)}\}$, the algorithm proceeds as follows:

Gibbs Sampler: For $t \in 1:T$, repeat:

- **Step 1:** Sample $\theta_1^{(t)}$ from $p(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)}, \cdots, \theta_k^{(t-1)}, \mathbf{y})$
- **Step 2:** Sample $\theta_2^{(t)}$ from $p(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)}, \cdots, \theta_k^{(t-1)}, \mathbf{y})$
- $\vdots$
- **Step k:** Sample $\theta_k^{(t)}$ from $p(\theta_k|\theta_2^{(t)}, \theta_3^{(t)}, \cdots, \theta_{k-1}^{(t)}, \mathbf{y})$.

Notice that for any sample $\{\theta_i^t, i = 1, \cdots, k, t = 1, \cdots, T\}$, its conditional distribution always uses the most updated parameters. The parameters obtained at iteration t, $(\theta_1^{(t)}, \cdots, \theta_k^{(t)})$, converge in distribution to a draw from the true joint posterior distribution $p(\theta_1, \cdots, \theta_k|\mathbf{y})$. This means that for t sufficiently large, say $t > T_0$, $\{\boldsymbol{\theta}^{(t)}, t = T_0 + 1, \cdots, T\}$ is a sample from the true posterior, from which any posterior quantities of interest may be estimated. For example, we may use a sample mean to estimate the posterior mean, i.e.,

$$E(\hat{\theta}_i|\mathbf{y}) = \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} \theta_i^{(t)},$$

and use an empirical sample 95% interval as a credible interval for any $\theta_i, i = 1, \cdots, k$, etc. The time in range $t \in \{0, 1, \cdots, T_0\}$ is commonly known as the *burn-in*

period. This is used to ensure the convergence, and thus minimize the bias of posterior inferences. In practice, we may actually run m (instead of 1, say $m = 3$) parallel Gibbs sampling chains with m different initial values. This technique is applied to assess sampler convergence, and can be produced with no extra time on multiprocessor computer. In this case, we would again discard all samples from the burn-in period, and obtain the posterior mean estimate,

$$E(\hat{\theta}_i|\mathbf{y}) = \frac{1}{m(T-T_0)} \sum_{j=1}^{m} \sum_{t=T_0+1}^{T} \theta_{i,j}^{(t)} \ .$$

The above Gibbs sampler draws samples of k scalar parameters one by one. Block schemes, which allow for updating an entire subvector of $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \cdots, \theta_k^{(t)})$ are also possible. As a footnote, recall that the CAR model defines the joint distribution of all data in terms of its full conditional distributions, and thus the Gibbs sampler arises as a natural scheme for simulation based inference.

4.2 The Metropolis-Hastings Algorithm

The Gibbs sampler is easy to understand and implement, but requires the full conditional distributions to be known. Unfortunately, when the prior distribution $p(\boldsymbol{\theta})$ and the likelihood $f(\mathbf{y}|\boldsymbol{\theta})$ are not a conjugate pair, one or more of these full conditionals may not be available in closed form. Even in this setting, however, $p(\theta_i|\boldsymbol{\theta}_{j,j\neq i}, \mathbf{y})$ will be available up to a proportionality constant, since it is proportional to the portion of $f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ that involves θ_i.

The Metropolis or Metropolis-Hastings algorithm is a rejection algorithm that deals precisely with this problem, since it requires only a function proportional

to the distribution to be sampled, at the cost of requiring a rejection step from a particular proposal density. Due to its flexibility and easy implementation, the Metropolis-Hastings (Hastings, 1970) algorithm has become the most commonly used MCMC techniques for finding posterior distributions. Suppose we wish to generate samples from a joint posterior distribution $p(\theta|\mathbf{y}) \propto g(\theta) = f(\mathbf{y}|\theta)p(\theta)$, we begin by specifying a proposal density $J(\theta^*|\theta^{(t-1)})$ that is a valid density function for every possible value of the conditioning variable $\theta^{(t-1)}$, and satisfies

$$J(\theta^*|\theta^{(t-1)}) = J(\theta^{(t-1)}|\theta^*) ,$$

which denotes that J is symmetric. Given a starting value $\theta^{(0)}$, the algorithm proceeds as follows.

Metropolis Algorithm: For $t \in 1 : T$, repeat:

- **Step 1:** Sample θ^* from $J(\theta^*|\theta^{(t-1)})$
- **Step 2:** Compute the ratio $r = \frac{g(\theta^*)}{g(\theta^{(t-1)})} = \exp[\log g(\theta^*) - \log g(\theta^{(t-1)})]$
- **Step 3:** If $r \geq 1$, set $\theta^{(t)} = \theta^*$;

$$\text{If } r < 1, \text{ set } \theta^{(t)} = \begin{cases} \theta^*, & \text{with probability } r \\ \theta^{(t-1)}, & \text{with probability } 1 - r \end{cases} .$$

The Metropolis algorithm offers substantial flexibility to choose the proposal density J. Theoretically, we can choose an ideal "good" density which will result in adequate proposed θ^* to be accepted. An usual selection would be normal

distribution

$$J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}) = N(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)}, \boldsymbol{\Sigma}),$$

since it obviously satisfies the symmetry property. The acceptance ratio will then depend on $\boldsymbol{\Sigma}$. Different choices of $\boldsymbol{\Sigma}$ could result in very high acceptance ratio (say, 1) or very low ratio (say, 0.01). On the one hand, an overly narrow proposal density proposes values around the parameter space with small steps, leading to high acceptance ratio, and high autocorrelation in the sampled chain; on the other hand, an overly wide proposal density will propose values far away from the majority of the posterior's support, leading to high rejection, and also, high autocorrelation. Gelman *et al.* (2003) proposed that an acceptance ratio between 25% and 40% is optimal, but also varies with the dimension and true posterior correlation structure $\boldsymbol{\theta}$. In this sense, acceptance ratio is always tuned by $\boldsymbol{\Sigma}$, which is called *tuning parameter*.

In practice, the Metropolis algorithm often serves as a substep in a larger Gibbs sampling algorithm framework, in which not all parameters posterior distributions have closed-form solutions, or some of them are awkward full conditionals. This is called "Metropolis within Gibbs" or "Metropolis substeps".

Sometimes, we may encounter situations with restrictions to parameters, for instance, $\boldsymbol{\theta} > 0$. In this case, Gaussian proposals will not be appropriate. The Metropolis-Hastings algorithm (Hastings, 1970) was proposed to resolve this issue. It does not require the symmetry property for proposal density. There is only a small difference to the Metropolis algorithm in Step 2.

Metropolis-Hastings Algorithm : Replace r in Step 2 in the Metropolis algorithm by

$$r = \frac{g(\boldsymbol{\theta}^*)J(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^*)}{g(\boldsymbol{\theta}^{(t-1)})J(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{(t-1)})} \ .$$

4.3 Bayesian Hierarchical Models

Bayesian techniques assume parameters in the model to be random variables, and assign prior distributions to them assuming prior information. Combining the likelihood of parameters and the prior distributions, posterior probability densities for each parameter can be determined. However, when more than one level of priors and parameters are needed, a hierarchical model can be applied.

Assume that $\mathbf{z(s)}$ is an underlying spatial process that follows the traditional CAR model. Let $\mathbf{y(s)}$ be one realization of this process. Then, one can express the data vector $\mathbf{y(s)}$ in terms of the process $\mathbf{z(s)}$ as

$$\mathbf{y(s)} = \mathbf{z(s)} + \boldsymbol{\epsilon}\mathbf{(s)},$$

where the vector $\boldsymbol{\epsilon}\mathbf{(s)}$ consists of identically independent normally distributed white noise components with mean zero and variance σ_y^2, and is assumed independent of $\mathbf{z(s)}$. Thus, the joint distribution of the data conditional given the process can be written as

$$\mathbf{y(s)}|\mathbf{z(s)}, \sigma_y^2 \sim N(\mathbf{z(s)}, \sigma_y^2\mathbf{I}).$$

To fully describe the distribution of the data, one needs to specify the distributions of both the spatial process and the variability in the data. As is common

practice in Bayesian Gaussian linear models, the variance, σ_y^2 , is assumed to follow an inverse gamma prior distribution with parameters α_y and β_y, which is a conjugate prior. The mean of the gamma distribution is $\frac{\alpha_y}{\beta_y}$. Since the spatial process is assumed to follow a Gaussian CAR model, its distribution is

$$\mathbf{z(s)}|\boldsymbol{\beta}, \sigma_z^2, \phi \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_z^2\mathbf{Q}^{-1}),$$

where $\mathbf{Q}$ is the rescaled precision matrix as defined for the computer efficient CAR, $\mathbf{X}\boldsymbol{\beta}$ allows for a linear trend over some set of independent explanatory variables, and ϕ is the spatial interaction (dependence) parameter.

The introduction of additional parameters in the prior distributions, so-called hyperparameters, requires another level of priors. Here, the distribution of $\boldsymbol{\beta}$, σ_z^2 and ϕ need to be specified. A convenient prior distribution for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0}).$$

In most cases, $\boldsymbol{\beta}_0$ is just the zero vector as is tested against in regression analysis, although it can be chosen to be some other $p \times 1$ vector. The variance covariance matrix $\boldsymbol{\Sigma}_{\beta_0}$ can be simplified to be a diagonal matrix $\sigma_\beta^2\mathbf{I}$ where the constant σ_β^2 is usually chosen based on past experience or some other knowledge, or is noninformative as a sufficiently large value.

As with the variance constant for the data, the variance for the spatial process, σ_z^2 , is assumed to follow an inverse gamma distribution with parameters α_z and

β_z. In both cases, if no strong prior information is available, the parameters for the inverse gamma distribution are chosen constants that result in relatively flat priors, which is achieved by selecting very small values for both α_z and β_z. The log normal distribution is chosen for the spatial interaction parameter ϕ . This implies that the values of ϕ are always positive, since to have negative spatial association implies that measurements at locations close together have opposite signs, something that tends not to occur in applications. Thus,

$$\pi = \log(\phi) \sim N(\mu_\phi, \sigma_\phi^2).$$

The hierarchical model structure is listed in Table 4.

Table 4: *Hierarchical model structure for a spatial process*

(1) Data process	$\mathbf{y(s)}\|\mathbf{z(s)}, \sigma_y^2 \sim N\left(\mathbf{z(s)}, \sigma_y^2\mathbf{I}\right)$	$\mathbf{y(s)}$ is observed data $\mathbf{z(s)}$ is the latent process
(2) Latent process	$\mathbf{z(s)}\|\boldsymbol{\beta}, \sigma_z^2, \phi \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma_z^2\mathbf{Q}^{-1}\right)$	$\mathbf{X}\boldsymbol{\beta}$ is the spatial trend $\mathbf{Q}^{-1}$ is the variance-covariance matrix defined by CAR model
(3) Priors	$\boldsymbol{\beta}\|\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0})$	$\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0}, \alpha_y, \beta_y, \alpha_z, \beta_z, \mu_\phi, \sigma_\phi^2$ are constants
	$\sigma_y^2 \sim InverseGamma(\alpha_y, \beta_y)$	$\boldsymbol{\Sigma}_{\beta_0}$ is generally chosen to be diagonal
	$\sigma_z^2 \sim InverseGamma(\alpha_z, \beta_z)$	
	$\pi = \log(\phi) \sim N(\mu_\phi, \sigma_\phi^2)$	

4.4 Precision Matrix Diagonalization

It can be noted that the precision matrix $\mathbf{Q}$ in the latent process in Table 4 can be expressed in different ways, as discussed in section 2.3. The log likelihood function

for the latent process is:

$$\text{loglik}(\mathbf{z}, \phi, \sigma_z^2) = -\frac{n}{2}\log(2\pi\sigma_z^2) + \frac{1}{2}\log|\mathbf{Q}| - \frac{1}{2\sigma_z^2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{Q}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta}),$$

where $|\mathbf{Q}|$ denotes the determinant. Despite the fact that the parameters in the CAR model proposed by Pettitt, Weir, and Hart (2002) were chosen in such a way as to make parameter estimation more computationally efficient, the data values are still correlated over space which can be computationally demanding for $|\mathbf{Q}|$ when many repeated calculations are required such as in maximum likelihood and MCMC based estimation. Rue (2001) recommends reordering the sites so that $\mathbf{Q}$ becomes a sparse band matrix and subsequent application of the Cholesky factorization.

In this section, we use a precision matrix diagonalization approach that results in a process that is uncorrelated over space. Due to the particular parameterization we find a diagonalization of $\mathbf{Q}$ that is free of the parameters, hence it needs to be performed only once and can be done prior to estimation. This approach is thus also called a "pre-whitening" method. Thus, the process variance contains no covariance component, resulting in full conditional posterior distributions that are easier to calculate and have a simpler form. After transformation calculation of determinant and matrix inversion are simple arithmetic operations with diagonal matrices and do not pose any computational challenge even for large data.

An alternative and an enhancement to this data transformation for gridded data is to do circulant embedding. That is, by enclosing the original lattice from which

the data is collected in a grid that is wrapped around a torus, all observed locations would have the same number of neighbors (4 neighbors). This creates a weight matrix that allows for easy computing of eigenvalues and eigenvectors, the most computationally taxing part of parameter estimation.

In the following, to clarify our diagonalization approach, we will assume the spatial interaction parameter to be $0 < \psi = \frac{\phi}{1+\phi} < 1$ based on Czado's CAR model, which is a realistic assumption since negative interactions are unlikely. When $\psi > 0$, $\mathbf{Q}$ can be written as

$$\mathbf{Q} = \mathbf{I} - \psi(\boldsymbol{\gamma} - \mathbf{D}).$$

As discussed in section 3.2, the eigenvalues of $\mathbf{Q}$ are $\eta_i = 1 - \psi\lambda_i$ where λ_i are the eigenvalues of $\boldsymbol{\gamma} - \mathbf{D}$. Note that $\boldsymbol{\gamma} - \mathbf{D}$ is completely determined by the given lattice, and does not depend on the model parameters. This can be utilized for calculation of $|\mathbf{Q}|$. The eigenvalue calculation of λ_i needs to be done only once, and $|\mathbf{Q}| = \prod_{i=1}^{n}(1 - \psi\lambda_i)$ for an updated value of ψ requires only a simple calculation. We can now write the eigenvector based diagonalization as follows,

$$\mathbf{F}^T(\boldsymbol{\gamma} - \mathbf{D})\mathbf{F} = \boldsymbol{\Lambda}\,,$$

where $\boldsymbol{\Lambda} = \text{diag}(\lambda_i)$ and $\mathbf{F}$ is the orthonormal matrix consisting of the unit length eigenvectors of $\boldsymbol{\gamma} - \mathbf{D}$ as columns. Note that by the properties of orthonormal vectors, $\mathbf{F}^T\mathbf{F} = \mathbf{F}\mathbf{F}^T = \mathbf{I}$. Thus, $\mathbf{F}^T = \mathbf{F}^{-1}$, and it can be shown that $\mathbf{Q}$ and $\mathbf{Q}^{-1}$ can be

expressed as

$$\mathbf{Q} = \mathbf{F}\text{diag}(1 - \psi\lambda_i)\mathbf{F}^T$$

and

$$\mathbf{Q}^{-1} = \mathbf{F}\text{diag}(\frac{1}{1 - \psi\lambda_i})\mathbf{F}^T .$$

By transforming the data and the process by way of the eigenvectors, the terms in the resulting hierarchical model are related to the original model such that

$$\mathbf{y}^* = \mathbf{F}^T\mathbf{y} ,$$

$$\mathbf{z}^* = \mathbf{F}^T\mathbf{z} ,$$

$$\mathbf{X}^* = \mathbf{F}^T\mathbf{X} .$$

Hence, the hierarchical model for the transformed data is identical to the structure in the original model. The only difference is that the $\mathbf{z}^*$ are uncorrelated with new variance matrix

$$\sigma_z^2\text{diag}(\frac{1}{1 - \psi\lambda_i}) .$$

Parameter estimation is done using these transformed values and the back transformation, $\mathbf{z} = \mathbf{F}\mathbf{z}^*$, is used after the parameter estimation is complete to obtain the original process estimates.

4.5 Posterior Distributions for Unknown Parameters

With the transformed hierarchical model structure listed in Table 5, we can now ease the computation for parameter estimations in MCMC procedure. Full conditional distribution for some parameters can be found using Bayesian methods,

Table 5: *Hierarchical model structure for a transformed spatial process, the precision matrix of which has been diagonalized.*

(1) Data process	$\mathbf{y}^*(\mathbf{s})\|\mathbf{z}^*(\mathbf{s}), \sigma_y^2 \sim N\left(\mathbf{z}^*(\mathbf{s}), \sigma_y^2\mathbf{I}\right)$	$\mathbf{y}^*(\mathbf{s}) = \mathbf{F}^T\mathbf{y}(\mathbf{s})$ is the transformed data $\mathbf{z}^*(\mathbf{s}) = \mathbf{F}^T\mathbf{z}(\mathbf{s})$ is the transformed latent process
(2) Latent process	$\mathbf{z}^*(\mathbf{s})\|\boldsymbol{\beta}, \sigma_z^2, \psi$ $\sim N\left(\mathbf{X}^*\boldsymbol{\beta}, \sigma_z^2\text{diag}(\frac{1}{1-\psi\lambda_i})\right)$	$\mathbf{X}^*\boldsymbol{\beta} = \mathbf{F}^T\mathbf{X}\boldsymbol{\beta}$ is the transformed spatial trend
(3) Priors	$\boldsymbol{\beta}\|\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0})$	$\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0}, \alpha_y, \beta_y, \alpha_z, \beta_z, a_\psi, b_\psi$ are constants
	$\sigma_y^2 \sim InverseGamma(\alpha_y, \beta_y)$	$\boldsymbol{\Sigma}_{\beta_0}$ is generally chosen to be diagonal
	$\sigma_z^2 \sim InverseGamma(\alpha_z, \beta_z)$	
	$\psi = \frac{\phi}{1+\phi} \sim Unif(0,1)$	
	or $\psi = \frac{\phi}{1+\phi} \sim Beta(a_\psi, b_\psi)$	

while others have no closed-form representation, and thus need a Metropolis step to update those parameters. Below we will calculate and list the available full conditional distributions.

Recall the fact for the Gaussian conditional distributions. Suppose the joint Gaussian distribution for $(\mathbf{X}_1, \mathbf{X}_2)'$ is

$$\begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right).$$

Then, the conditional distribution of $\mathbf{X}_2|\mathbf{X}_1$ is:

$$\mathbf{X}_2|\mathbf{X}_1 \sim N\left(\mu_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{X}_1 - \mu_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right).$$

4.5.1 Posterior Distribution for Latent Spatial Process

Let

$$\eta_i = 1 - \psi\lambda_i \,,$$

where, $\{\lambda_i,\ i = 1, \cdots, n\}$ are the eigenvalues of matrix $\gamma - \mathbf{D}$, here, γ is defined by a distance based weight matrix by Pettitt *et al.*(2002), $\mathbf{D}$ is a diagonal matrix with the ith element corresponding to the $i-$th row summation of γ minus 1, and

$$\omega_i = \frac{\sigma_z^2}{\sigma_z^2 + \sigma_y^2\eta_i} \,.$$

The joint distribution of $\mathbf{y}^*$ and $\mathbf{z}^*$ is determined by

$$\begin{pmatrix} \mathbf{y}^* \\ \mathbf{z}^* \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{X}^*\boldsymbol{\beta} \\ \mathbf{X}^*\boldsymbol{\beta} \end{pmatrix}, \begin{pmatrix} \sigma_z^2\mathrm{diag}(\frac{1}{\eta_i}) + \sigma_y^2\mathbf{I} & \sigma_z^2\mathrm{diag}(\frac{1}{\eta_i}) \\ \sigma_z^2\mathrm{diag}(\frac{1}{\eta_i}) & \sigma_z^2\mathrm{diag}(\frac{1}{\eta_i}) \end{pmatrix} \right),$$

where the variance of $\mathbf{y}^*$ is:

$$\mathrm{Var}(\mathbf{y}^*) = \mathrm{Var}(\mathbf{z}^* + \boldsymbol{\epsilon}) = \sigma_z^2\mathrm{diag}(\frac{1}{\eta_i}) + \sigma_y^2\mathbf{I}$$

and the covariance of $\mathbf{y}$ and $\mathbf{z}$ is:

$$\mathrm{Cov}(\mathbf{y}^*, \mathbf{z}^*) = \mathrm{Cov}(\mathbf{z}^* + \boldsymbol{\epsilon}, \mathbf{z}^*) = \mathrm{Var}(\mathbf{z}^*) = \sigma_z^2\mathrm{diag}(\frac{1}{\eta_i}) \,.$$

Therefore, the conditional distribution of $p(\mathbf{z}^*|\mathbf{y}^*, \sigma_z^2, \sigma_y^2, \boldsymbol{\beta}, \psi)$ is:

$$\mathbf{z}^*|\mathbf{y}^*, \sigma_z^2, \sigma_y^2, \boldsymbol{\beta}, \psi \sim N(\boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}}, \Sigma_{\mathbf{z}|\mathbf{y}}) ,$$

where,

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{z}|\mathbf{y}} &= \mathbf{X}^*\boldsymbol{\beta} + \sigma_z^2 \text{diag}(\frac{1}{\eta_i}) \left(\sigma_z^2 \text{diag}(\frac{1}{\eta_i}) + \sigma_y^2 \mathbf{I} \right)^{-1} (\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) \\ &= \mathbf{X}^*\boldsymbol{\beta} + \text{diag}(\omega_i)(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{z}|\mathbf{y}} &= \sigma_z^2 \text{diag}(\frac{1}{\eta_i}) - \sigma_z^2 \text{diag}(\frac{1}{\eta_i}) \left(\sigma_z^2 \text{diag}(\frac{1}{\eta_i}) + \sigma_y^2 \mathbf{I} \right)^{-1} \sigma_z^2 \text{diag}(\frac{1}{\eta_i}) \\ &= \sigma_z^2 \text{diag}\left(\frac{1 - \omega_i}{\eta_i} \right) . \end{aligned}$$

4.5.2 Posterior Distribution for Variance Parameters

The posterior full distribution $p(\sigma_y^2|\mathbf{y}^*, \mathbf{z}^*, \sigma_z^2, \boldsymbol{\beta}, \psi)$ for data variance will involve data likelihood $p(\mathbf{y}^*|\mathbf{z}^*, \sigma_z^2, \sigma_y^2, \boldsymbol{\beta}, \psi)$ and its prior $p(\sigma_y^2)$. In our case, the prior distribution is inverse gamma with known shape parameter α_y and scale parameter β_y, which is

$$p(\sigma_y^2) \propto (\sigma_y^2)^{-\alpha_y - 1} \exp\left(-\beta_y / \sigma_y^2 \right) .$$

The data likelihood $p(\mathbf{y}^*|\mathbf{z}^*, \sigma_z^2, \sigma_y^2, \boldsymbol{\beta}, \psi)$ is proportional to

$$p(\mathbf{y}^*|\mathbf{z}^*, \sigma_z^2, \sigma_y^2, \boldsymbol{\beta}, \psi) \propto \frac{1}{(\sigma_y^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^{n}(y_i^* - z_i^*)^2\right\} .$$

Therefore, the posterior distribution of σ_y^2 given all other parameters will be:

$$\begin{aligned} p(\sigma_y^2|\mathbf{y}^*, \mathbf{z}^*, \sigma_z^2, \boldsymbol{\beta}, \psi) &\propto \frac{1}{(\sigma_y^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_y^2}\sum_{i=1}^{n}(y_i^* - z_i^*)^2\right\} \cdot (\sigma_y^2)^{-\alpha_y - 1} \exp\left(-\beta_y/\sigma_y^2\right) \\ &= IG(\alpha_{y(post)}, \beta_{y(post)}) , \end{aligned}$$

where,

$$\alpha_{y(post)} = \alpha_y + \frac{n}{2}$$

and

$$\beta_{y(post)} = \beta_y + \frac{1}{2}\sum_{i=1}^{n}(y_i^* - z_i^*)^2 .$$

Now, let us consider the posterior distribution for latent process variance σ_z^2. The posterior full distribution $p(\sigma_z^2|\mathbf{y}^*, \mathbf{z}^*, \sigma_y^2, \boldsymbol{\beta}, \psi)$ will involve the product of three items: data likelihood $p(\mathbf{y}^*|\mathbf{z}^*, \sigma_y^2, \sigma_z^2, \boldsymbol{\beta}, \psi)$, the latent process likelihood $p(\mathbf{z}^*|\sigma_y^2, \sigma_z^2, \boldsymbol{\beta}, \psi)$ and its prior $p(\sigma_z^2)$. However, notice that the data process only involves the latent process with parameters $\boldsymbol{\beta}$ and σ_y^2, it has no contribution to the posterior distributions of σ_z^2, which therefore only depends on its prior and latent process likelihood. In our case, the prior distribution is inverse gamma with known shape parameter α_z and

scale parameter β_z, which is

$$p(\sigma_z^2) \propto (\sigma_z^2)^{-\alpha_z-1} \exp\left(-\beta_z/\sigma_z^2\right) .$$

The latent process likelihood for $\mathbf{z}^*$ is:

$$p(\mathbf{z}^*|\sigma_z^2, \sigma_y^2, \boldsymbol{\beta}, \psi) \propto \frac{1}{(\sigma_z^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_z^2} \sum_{i=1}^{n} \eta_i (\mathbf{z}_i^* - \mathbf{X}_i^* \boldsymbol{\beta})^2\right\} .$$

Therefore, the posterior distribution of σ_z^2 given all other parameters will be:

$$\begin{aligned} p(\sigma_z^2|\mathbf{y}^*, \mathbf{z}^*, \sigma_z^2, \boldsymbol{\beta}, \psi) &\propto \frac{1}{(\sigma_z^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma_z^2} \sum_{i=1}^{n} \eta_i (\mathbf{z}_i^* - \mathbf{X}_i^* \boldsymbol{\beta})^2\right\} \cdot (\sigma_z^2)^{-\alpha_z-1} \exp\left(-\beta_z/\sigma_z^2\right) \\ &= IG(\alpha_{z(post)}, \beta_{z(post)}) , \end{aligned}$$

where,

$$\alpha_{z(post)} = \alpha_z + \frac{n}{2}$$

and

$$\begin{aligned} \beta_{z(post)} &= \beta_z + \frac{1}{2} \sum_{i=1}^{n} \eta_i (\mathbf{z}_i^* - \mathbf{X}_i^* \boldsymbol{\beta})^2 \\ &= \beta_z + \frac{1}{2} \sum_{i=1}^{n} \eta_i e_i^2, \quad \text{where, } e_i = \mathbf{z}_i^* - \mathbf{X}_i^* \boldsymbol{\beta} \text{ is the residual.} \end{aligned}$$

4.5.3 Posterior Distribution for Trend Parameters

The prior distribution for $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}) = N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_{\beta_0}) .$$

Then, the joint distribution of $(\boldsymbol{\beta}, \mathbf{z}^*)$ is:

$$\begin{pmatrix} \mathbf{z}^* \\ \boldsymbol{\beta} \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{X}^*\boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_0 \end{pmatrix}, \begin{pmatrix} \sigma_z^2 \text{diag}(\frac{1}{\eta_i}) + \mathbf{X}\boldsymbol{\Sigma}_{\beta_0}\mathbf{X}^T & \mathbf{X}\boldsymbol{\Sigma}_{\beta_0} \\ \boldsymbol{\Sigma}_{\beta_0}\mathbf{X}^T & \boldsymbol{\Sigma}_{\beta_0} \end{pmatrix} \right) .$$

Therefore, the conditional distribution $p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{z}^*, \sigma_z^2, \sigma_z^2, \psi)$ is

$$p(\boldsymbol{\beta}|\mathbf{y}^*, \mathbf{z}^*, \sigma_z^2, \sigma_z^2, \psi) = N(\boldsymbol{\mu}_{\beta(post)}, \boldsymbol{\Sigma}_{\beta(post)}) ,$$

where,

$$\boldsymbol{\mu}_{\beta(post)} = \boldsymbol{\beta}_0 + \boldsymbol{\Sigma}_{\beta_0}\mathbf{X}^T \left(\sigma_z^2 \text{diag}(\frac{1}{\eta_i}) + \mathbf{X}\boldsymbol{\Sigma}_{\beta_0}\mathbf{X}^T \right)^{-1} (\mathbf{z}^* - \mathbf{X}^*\boldsymbol{\beta}_0)$$

and

$$\boldsymbol{\Sigma}_{\beta(post)} = \boldsymbol{\Sigma}_{\beta_0} - \boldsymbol{\Sigma}_{\beta_0}\mathbf{X}^T \left(\sigma_z^2 \text{diag}(\frac{1}{\eta_i}) + \mathbf{X}\boldsymbol{\Sigma}_{\beta_0}\mathbf{X}^T \right)^{-1} \mathbf{X}\boldsymbol{\Sigma}_{\beta_0} .$$

4.5.4 Non-closed-form Posterior Distributions

Some parameters do not have closed-form posterior full conditional distributions, and thus cannot use Gibbs sampler to update them in MCMC. For example, in our case here, parameter $\psi \in (0, 1)$ cannot be written in an explicit solution. What

we will do is to provide a proposal distribution for ψ, say, uniform distribution between 0 and 1, and then use "Metropolis within Gibbs" to update parameter ψ. A similar situation will also happen later for our proposed new model, and thus we will use a similar estimation scheme.

CHAPTER V

An Extended Spatial Autoregressive Model

Geostatistical models and spatial autoregressive models as well as Gaussian Markov random fields (GMRF) are discussed in the previous chapters. The conditional autoregressive (CAR) model is also expanded to Pettitt's or Czado's parameterizations which are computer efficient models that allow for direct and computationally fast calculation of the precision matrix.

However, CAR models are somehow too limited in practice to be suitable for a underlying smooth latent process. CAR models with a defined low order neighbor structure are not capable of modeling an underlying smooth random field. It has been noted that a CAR model assumes a single interaction parameter between first order neighbors which produces rough spatial surfaces. Griffith *et al.* (1996) show heuristically that the CAR model corresponds approximately to an exponentially decaying correlation structure over a large lattice ignoring the subtleties of the edge effects. It is well known that random fields with an exponential correlation structure are not differentiable hence they are not smooth. High order neighbor structures may be specified to capture the smoothness of an underlying process. Rue and Tjelmeland (2002), and Rue and Held (2005, Chapter 5), provide a more general correspondence whereby the degree of smoothness is increased by increas-

ing the order of the neighborhood which they apply to Gaussian Markov random fields (GMRF) over regular lattices. They recommend up to at least five orders of neighbors which requires at least 6 interaction parameters in the isotropic case. While this approximation is excellent it may be difficult to implement such higher order neighbor structures on irregular lattices. Even on regular lattices, the required large number of parameters may be a burden for model fitting. Further, Rue and Held (2005) note that the estimated parameter values typically have alternating signs and are not particularly insightful with respect to the underlying model structure.

In this chapter, we will propose a parsimonious model with two parameters for the spatial dependence structure that is suitable for estimation where the underlying spatial random field can have any degree of smoothness. Our model is an extension of the one-parameter Czado's CAR model (or modified Pettitt's model).

5.1 Model Extension: The EAR Model

We now utilize the diagonalization (section 4.4) to define a new extended model by introducing a parameter $\theta > 0$ that describes the smoothness of the underlying random field. We call this the extended autoregression model, or abbreviated: the "EAR" model.

Recall that the computer efficient Czado's CAR model for a spatial process $\mathbf{z}$ can be written as

$$\mathbf{z} \sim N(\boldsymbol{\mu}, \sigma^2 \mathbf{Q}^{-1}) ,$$

where, for $0 < \psi < 1$ and with eigenvalues $\{\lambda_i, i = 1, \cdots, n\}$ of matrix $\gamma - \mathbf{D}$ in section 2.3.5

$$\mathbf{Q} = \mathbf{F}\text{diag}(1 - \psi\lambda_i)\mathbf{F}^T .$$

Now, let us define and specify our EAR model as follows.

Definition 5.1 (EAR Model) *A spatial process* $\mathbf{z}$ *defined over a lattice with neighbor index matrix* γ *follows the EAR model with parameters* $(\mu, \sigma^2, \psi, \theta)$ *if*

$$\mathbf{z} \sim N(\mu, \sigma^2\mathbf{Q}^{-1})$$

with $\mathbf{Q}$ *defined by*

$$\mathbf{Q} = \mathbf{F}diag(1 - \psi\lambda_i)^\theta\mathbf{F}^T ,$$

where $\mathbf{F}$ *is the eigenvector matrix of* $\gamma - \mathbf{D}$ *and* $\mathbf{D}$ *is the diagonal matrix of the row sums of* γ *minus 1, and* $\{\lambda_i, i = 1, \cdots, n\}$ *are eigenvalues of* $\gamma - \mathbf{D}$.

Alternatively we can write:

A spatial random field $\mathbf{z}$ is EAR$(\mu, \sigma^2, \psi, \theta)$ if its transformed process $\mathbf{z}^*$ follows

$$\mathbf{z}^* = \mathbf{F}^T\mathbf{z} \sim N\left(\mathbf{F}^T\mu, \sigma^2\text{diag}\left(\frac{1}{1 - \psi\lambda_i}\right)^\theta\right) .$$

Notice that like the Czado *et al.* parameterization, the spatial interaction parameter ψ takes any value between $(0, 1)$. When $\psi = 0$, the process is independent, and the smooth parameter will take no effect; while when $\psi = 1$, spatial realizations are highly correlated. The parameter θ governs the smoothness of the random field,

and is specified to be strictly larger than zero.

As an illustration, a 30×30 grid has been created representing a spatial random field following the extended model. The values for each grid cell are generated from the EAR model with mean $\boldsymbol{\mu} = \mathbf{0}$, $\sigma^2 = 1$ and different smoothness parameters $\theta = 1, 2, 3, 4, 6, 9$, and fixed interaction parameter $\psi = 0.75$. The simulation procedure follows Chapter 2 in Rue and Held (2005) as listed in Table 6.

Table 6: *An algorithm to simulate a Gaussian random spatial process from EAR model*

Algorithm : Sampling $\mathbf{z} \sim EAR(\boldsymbol{\mu}, \sigma^2, \psi, \theta) = N(\mu, \sigma^2 \mathbf{Q}^{-\theta}), \mathbf{Q} = \mathbf{I} - \psi(\gamma - \mathbf{D})$

1. Compute the eigendecomposition of matrix $\gamma - \mathbf{D}$, $\gamma - \mathbf{D} = \mathbf{F}\Lambda\mathbf{F}^T$, where, $\Lambda = \text{diag}(\lambda_i, i = 1, \cdots, n)$
2. Calculate $\boldsymbol{\Sigma}^{\frac{1}{2}} = \mathbf{F}\text{diag}\left(\frac{1}{1-\psi\lambda_i}\right)^{\theta/2}\mathbf{F}^T$
3. Sample $\mathbf{x} \sim N(\mathbf{0}, \mathbf{I})$
4. Compute $\mathbf{y} = \boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{x}$
5. Compute $\mathbf{z} = \mathbf{y} + \boldsymbol{\mu}$
6. **Return**

Figure 7 shows simulations of $\mathbf{z}$ for various values of θ, illustrating how it is related to smoothness. Note that we used the same $\mathbf{x}$ for each realization. From the graphs, we can obviously detect the smoothness pattern as θ varies. The larger θ is, the smoother the random field. This thus allows for the modeling of a spatial random field with any level of smoothness. Notice that when $\theta = 1$, our model reduces to the CAR model.

5.2 Circulant Embedding

It has been noted that GMRF are not stationary even on a regular lattice because of the differing neighbor structure along the boundaries of the lattice. This is the

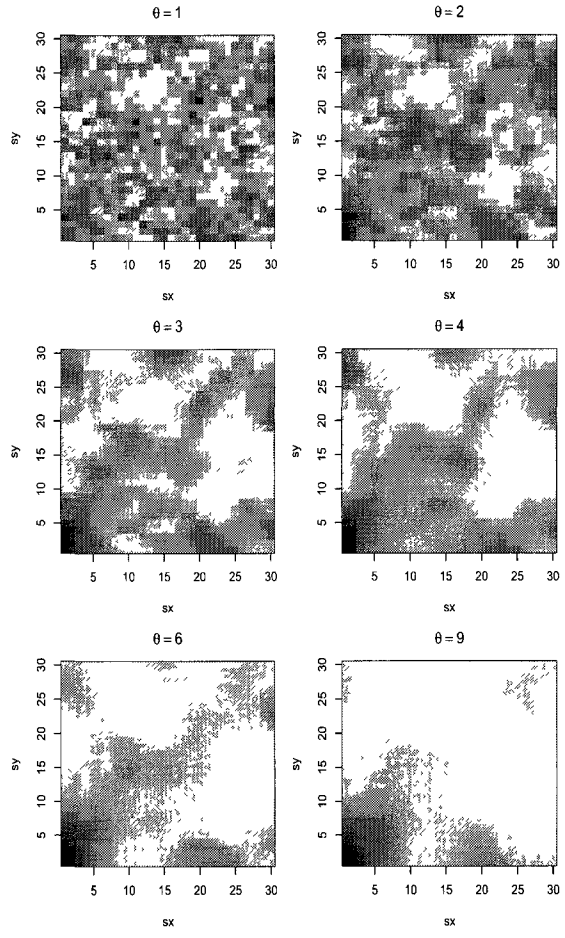


Figure 7: *Simulated random fields with mean 0 of the EAR model with* $\psi = 0.75$ *for* $\theta = 1, 2, 3, 4, 6,$ *and* 9, *respectively, and* $\sigma^2 = 1$. *Note the same* $\mathbf{x}$ *was used for each realization.*

well known edge effect problem. This problem is resolved if one assumes that the lattice is wrapped around onto a torus, which in effect removes any edges and provides for a uniform neighbor structure throughout the spatial domain.

What's more, estimation with Gaussian spatial data has been limited to moderately sized data because of the computationally demanding operations of matrix inversion and determinant calculation, as discussed in the "big n problem". The number of floating point operations on a computer for these operations is typically of order $O(n^3)$. For regular lattices one can take advantage of the regular structure

in **Q** and thus accelerate the computation. It has been noted that a further gain in computation can be achieved if the spatial domain is a regular torus.

A torus can be viewed as a regular lattice with cyclical boundary conditions. Figure 8 illustrates the form of a torus (or donut). More precisely if the x and y coordinates of a regular lattice are ordered from 0 to $n_1 - 1$, and from 0 to $n_2 - 1$, respectively, then the torus implies a cyclical extension of the numbering where in the x direction sites are numbered mod n_1 and in the y direction sites are numbered mod n_2. For example site $(2, 5)$ equals site $(n_1 - 2, 5)$ and site $(4, n_2)$ equals site $(4, 0)$, and so on. Such a cyclical extension removes any boundaries and the assumed neighbor structure is identical at any point on the torus.

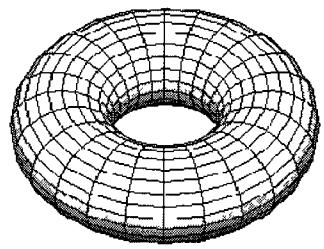


Figure 8: *Torus - an illustration of a two-dimensional lattice with cyclic boundary conditions*

A GMRF defined on a torus will then result in a circulant precision matrix **Q**. Circulant matrices have the property that their eigenvalues and eigenvectors are related to the discrete Fourier transform. This allows for fast algorithms for common matrix operation such as obtaining inverse, determinant, and so on. Below we will provide some details about circulant matrices.

Definition 5.2 (Circulant matrix) *An $n \times n$ matrix $\mathbf{C}$ is circulant if and only if it has the form*

$$\mathbf{C} = \begin{pmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \cdots & c_{n-2} \\ c_{n-2} & c_{n-1} & c_0 & \cdots & c_{n-3} \\ \vdots & \vdots & \vdots & & \vdots \\ c_1 & c_2 & c_3 & \cdots & c_0 \end{pmatrix} = (c_{j-i \bmod n})$$

for some vector $\mathbf{c} = (c_0, c_1, \cdots, c_{n-1})^T$. The vector $\mathbf{c}$ is called the base of $\mathbf{C}$.

A circulant matrix is fully specified by only one column or one row. Let $\omega = \sqrt{-1}$, then the jth eigenvalues can be found by

$$\lambda_j = \sum_{i=0}^{n-1} c_i \exp(-2\pi\omega ij/n) ,$$

and the jth eigenvector is

$$\mathbf{e}_j = \frac{1}{\sqrt{n}} \begin{pmatrix} 1 \\ \exp(-2\pi\omega j/n) \\ \exp(-2\pi\omega j2/n) \\ \vdots \\ \exp(-2\pi\omega j(n-1)/n) \end{pmatrix} .$$

Now, the eigenvector matrix can be defined by,

$$\begin{aligned}\mathbf{F} &= (\mathbf{e}_0|\mathbf{e}_1|\cdots|\mathbf{e}_{n-1}) \\ &= \frac{1}{\sqrt{n}}\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \rho^1 & \rho^2 & \cdots & \rho^{n-1} \\ 1 & \rho^2 & \rho^4 & \cdots & \rho^{2(n-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \rho^{n-1} & \rho^{2(n-1)} & \cdots & \rho^{(n-1)(n-1)} \end{pmatrix},\end{aligned}$$

where $\rho = \exp(-2\pi\omega/n)$. Note that $\mathbf{F}$ does not depend on $\mathbf{c}$.

A natural generalization of circulant matrices are block-circulant matrices. They share the same properties as circulant matrices. The block-circulant matrix can be defined as,

Definition 5.3 (Block-circulant matrix) *An $Nn \times Nn$ matrix $\mathbf{C}$ is block circulant if and only if it has the form*

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_0 & \mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_{N-1} \\ \mathbf{C}_{N-1} & \mathbf{C}_0 & \mathbf{C}_1 & \cdots & \mathbf{C}_{N-2} \\ \mathbf{C}_{N-2} & \mathbf{C}_{N-1} & \mathbf{C}_0 & \cdots & \mathbf{C}_{N-3} \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{C}_1 & \mathbf{C}_2 & \mathbf{C}_3 & \cdots & \mathbf{C}_0 \end{pmatrix} = (\mathbf{C}_{j-i \bmod N})$$

where, $\mathbf{C}_i$ is a circulant $n \times n$ matrix with base $\mathbf{c}_i$.

When considering data collected on a rectangular lattice, either regular in boundary shape or not, it can be wrapped onto a torus directly. Another approach is to use a slightly larger regular rectangular lattice to enclose the original irregular lattice, and then wrap it onto a torus. Thus, this embedding scheme becomes circulant in nature. The fact that the region is expanded indicates that the original lattice on which the observations are taken is minimally affected by the "additional" neighbors. Using this type of embedding creates a γ matrix and thus a $\mathbf{Q}$ matrix that is not only sparse but also symmetric block circulant, which looks like

$$\mathbf{Q} = \begin{pmatrix} q_0 & q_1 & q_2 & \cdots & q_{n-1} \\ q_{n-1} & q_0 & q_1 & \cdots & q_{n-2} \\ q_{n-2} & q_{n-1} & q_1 & \cdots & q_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ q_1 & q_2 & q_3 & \cdots & q_0 \end{pmatrix} = (q_{j-i \bmod n}) .$$

Taking a 3×3 regular lattice wrapped onto a torus as an example, the first-order neighbor indicator matrix is

$$\gamma = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix},$$

and γ is sparse and has four neighbors for each site.

It is well known that the eigenvalues of $\mathbf{Q}$ are the discrete Fourier transform of any row of $\mathbf{Q}$, whereas the eigenvectors are the corresponding Fourier bases of size n (Brockwell and Davis, 1987). Note that the eigenvectors are constants hence they do not depend on $\mathbf{Q}$ except for its size n. These calculations are of order $O(n \log n)$ and thus are possible for large n. The eigenvectors and eigenvalues are used in the obvious way for calculating the determinant of $\mathbf{Q}$ and for generating draws from a model with precision $\mathbf{Q}$. Note that for a non-isotropic model that has different interaction parameters in the x and the y directions, $\mathbf{Q}$ is a block circulant matrix, and the eigenvector / eigenvalue calculations involve the two-dimensional discrete Fourier transform (see Rue and Held, 2005, for details). For our model when applied on a torus we will apply the Fourier transform to one row of the matrix $\gamma - \mathbf{D}$.

In order for this type of embedding to be valid for use in CAR and EAR models, the precision matrix, $\mathbf{Q} = I - \psi(\gamma - \mathbf{D})$, needs to be symmetric positive definite. For now, consider only the uniform weight system described by

$$\gamma_{ij} = \begin{cases} 1, & j \in N(i) \\ 0, & j \notin N(i) . \end{cases}$$

For any order of neighbors considered, this results in a sparse symmetric matrix. In the circulant embedding scheme, matrix γ is circulant, and the elements of $\mathbf{D}$ represent the number of neighbors for each spatial location minus 1. This number is constant and thus can be written as $\mathbf{D} = d\mathbf{I}$. Since γ is symmetric and both $\mathbf{D}$ and $\mathbf{I}$ are diagonal, $\mathbf{Q}$ is symmetric. In addition, the $\mathbf{Q}$ matrix can be expressed as

$$\mathbf{Q}_{ij} = \begin{cases} 1 + d\psi, & i = j \\ -\psi\gamma_{ij}, & i \neq j . \end{cases}$$

Recall that a matrix $\mathbf{A} = [a_{ij}]$ is said to be diagonally dominant if $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$ for all i. Since uniform weights are assumed, there are $(d + 1)$ off diagonal elements that are equal to one and the rest are zero. Becasue $0 \leq \psi \leq 1$, we get

$$|Q_{ii}| = 1 + d\psi \geq \sum_{j \neq i} |Q_{ij}| = d\psi + \psi, \quad \forall\, i .$$

Therefore, $\mathbf{Q}$ is diagonally dominant. From Theorem 12.2.16 in Graybill (1983), $\mathbf{Q}$ is positive definite. Therefore, the precision matrix with uniform weights is symmetric positive definite. The proof follows for other weight systems. That is,

for any valid $\mathbf{Q}$, circulant embedding creates another valid $\mathbf{Q}$.

5.3 Is the EAR Model a Markov Random Field?

Through the definition, it is not clear whether the EAR model is in general a Markov random field. That is, the conditional distributions of one realization from the process given all other locations, $p(z_i|\mathbf{z}_{-i})$ cannot be determined using only the neighbors of that location. In our case, $p(z_i|\mathbf{z}_{-i})$ also depends upon the smoothing parameter θ. Recall that this is important because in order to use the Gibbs sampler for estimation, it is necessary that when values are simulated from the joint distribution, that this distribution is both stationary and unique. It will be shown in the following text that under some conditions, the EAR model is equivalent to a higher order CAR model which is a Markov random field. Rue and Tjemeland (2002) have shown a similar correspondence between geostatistical models and Markov random fields.

In this section we show that for a regular square lattice embedded onto a torus the (isotropic) EAR model is for integer values of θ a Markov random field with higher order neighbor structure, and that for other values of θ it can be approximated by such a model.

The following are some obvious facts of neighbor indicator matrices defined over a torus. We denote the neighbor indicator matrices by γ, or γ_j. For instance, when $j = 1$, γ_1 denotes the first-order neighbor indicator matrix; when $j = 2$, γ_2 denotes the second-order neighbor indicator matrix, etc. They are assumed to be circulant. For non-isotropic models similar results (but more complex though) can

be derived; but here the γ matrices are block circulant. We use $\mathbf{D}$ or $\mathbf{D}_j$ to denote the row sum diagonal matrices of the form diag $\left\{\sum_k \gamma_{ik} - 1\right\}$. Matrices γ_j and $\mathbf{D}_j$ can be treated as a pair for each j-th order neighbor structure. In this section we assume throughout that $\sigma = 1$. The generalization to $\sigma \neq 1$ is trivial.

Fact 1: Let λ_i be eigenvalues of $(\gamma - \mathbf{D}) = (\gamma - d\mathbf{I})$, where d denotes the number of neighbors minus 1 of any site on the torus. Then $\lambda_i = \delta_i - d$, where δ_i are the eigenvalues of γ.

Fact 2: To consider a higher order CAR model, it is necessary to separate γ to distinguish the order of the neighbors being considered. Then, let $\gamma_1, \gamma_2, \cdots, \gamma_k$ be a set of indicator neighbor matrices where the subscript i indicates the order of the neighbors,

$$\gamma = \gamma_1 + \gamma_2 + \cdots + \gamma_k$$

where

$$\text{Elements of } \gamma_j = \begin{cases} 1, & \text{if } s_i \text{ is a } j\text{-th order neighbor} \\ 0, & \text{otherwise .} \end{cases}$$

Using the Czado *et al.* parameterization where $\psi > 0$, $\mathbf{Q} = \mathbf{I} - \psi(\gamma - \mathbf{D})$. In the case where one is only interested in the first order neighbors, $\gamma = \gamma_1$, and so

$$\mathbf{Q} = \mathbf{I} - \psi_1(\gamma_1 - \mathbf{D}_1) ,$$

where ψ_1 is the spatial interaction using first order neighbors. Similarly, considering both first and second order neighbors results in $\gamma = \gamma_1 + \gamma_2$, and

$$\mathbf{Q} = \mathbf{I} - \psi_1(\gamma_1 - \mathbf{D}_1) - \psi_2(\gamma_2 - \mathbf{D}_2) .$$

Then for the combined higher order neighbor structure $\gamma_1 + \gamma_2 + \cdots + \gamma_k$, the following is a k-parameter CAR precision matrix:

$$\mathbf{Q} = \mathbf{I} - \psi_1(\gamma_1 - \mathbf{D}_1) - \psi_2(\gamma_2 - \mathbf{D}_2) - \cdots - \psi_k(\gamma_k - \mathbf{D}_k) .$$

It is easily established that $\mathbf{Q}$ is symmetric and positive definite for any values of the ψ (positive) parameters, as shown in section 5.2.

Fact 3: Since in the above representation all γ_j are circulant and all $\mathbf{D}_j$ are $\mathbf{D}_j = d_j\mathbf{I}$, the eigenvalues of $\mathbf{Q}$ are given by:

$$1 - \psi_1(\delta_{1,i} - d_1) - \psi_2(\delta_{2,i} - d_2) - \cdots - \psi_k(\delta_{k,i} - d_k)$$

for $i = 1, \cdots, n$, and where $\delta_{j,i}$ are the eigenvalues of γ_j.

Fact 4: Since all γ_j have the same eigenvector matrix $\mathbf{F}$ (defined with columns as eigenvectors) we can represent the precision matrix $\mathbf{Q}$ as follows:

$$\mathbf{Q} = \mathbf{F}\text{diag}\left(1 - \psi_1(\delta_{1,i} - d_1) - \psi_2(\delta_{2,i} - d_2) - \cdots - \psi_k(\delta_{k,i} - d_k)\right)\mathbf{F}^T .$$

Since the $\mathbf{Q}$ in Fact 4 is the precision matrix of a higher order GMRF we now attempt to show when this $\mathbf{Q}$ is equal or a close approximation to an EAR model. To demonstrate the connections between the EAR model and a higher order Markov random field, consider the first order EAR model with precision matrix

$$\mathbf{Q} = \mathbf{F}\text{diag}(1 - \psi\lambda_i)^{\theta}\mathbf{F}^T .$$

Assuming a normal distribution,

$$\mathbf{z} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma} = \mathbf{Q}^{-1})$$

and

$$\mathbf{F}^T\mathbf{z} \sim MVN\left(\mathbf{F}^T\boldsymbol{\mu}, \text{diag}\left(\frac{1}{1 - \psi\lambda_i}\right)^{\theta}\right) .$$

We use the diagonal matrix of the representation in Fact 4 to match the diagonal matrix of the EAR model by pointwise Taylor series expansion. Recall that any function can be approximated by a Taylor series expansion as follows,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)\frac{(x - x_0)^2}{2!} + f'''(x_0)\frac{(x - x_0)^3}{3!} + \cdots ,$$

where x_0 is any constant. Thus, with $x_0 = 0$, via a Taylor series approximation, each element $(1 - \psi\lambda_i)^{\theta}$ can be written as,

$$(1 - \psi\lambda_i)^{\theta} = h(\lambda_i) = h(0) + h'(0)\lambda_i + h''(0)\frac{\lambda_i^2}{2!} + h'''(0)\frac{\lambda_i^3}{3!} + \cdots ,$$

where

$$h'(0) = \theta(1-\psi\lambda_i)^{\theta-1}(-\psi)\mid_{\lambda_i=0} = -\psi\theta$$

$$h''(0) = \theta(\theta-1)(1-\psi\lambda_i)^{\theta-2}\psi^2\mid_{\lambda_i=0} = \psi^2\theta(\theta-1)$$

$$h'''(0) = \theta(\theta-1)(\theta-2)(1-\psi\lambda_i)^{\theta-3}(-\psi^3)\mid_{\lambda_i=0} = -\psi^3\theta(\theta-1)(\theta-2)$$

$$h^{(m)}(0) = (-1)^m\psi^m\theta(\theta-1)\cdots(\theta-m+1)\,,$$

and so on. It is obvious that if θ is an integer, say $\theta = p > 0$, then $h^{(p+q)}(0) = 0$ for $q \geq 1$. In this case we have an exact expansion as follows:

$$(1-\psi\lambda_i)^p = 1 - p\psi\lambda_i + p(p-1)\frac{\psi^2\lambda_i^2}{2!} + \cdots + (-1)^p p!\frac{\psi^p\lambda_i^p}{p!} = \sum_{j=0}^{p}(-1)^j\begin{pmatrix} p \\ j \end{pmatrix}\psi^j\lambda_i^j\,.$$

Fact 5: An EAR model over a regular torus (isotropic) that has an integer valued smoothness parameter $\theta = p$ has the following representation of the precision matrix. Let $\mathbf{\Psi} = \text{diag}(1-\psi\lambda_i)$, and let $\mathbf{Q}_j$ denote the precision matrix of an EAR model with $\theta = j$ for $j = 1, \cdots, p$. Note that $\mathbf{Q}_1$ denotes the first-order precision matrix.

$$\mathbf{Q}_p = \mathbf{F}\text{diag}(1-\psi\lambda_i)^p\mathbf{F}^T = \mathbf{F}\mathbf{\Psi}^p\mathbf{F}^T = \mathbf{F}\mathbf{\Psi}\mathbf{F}^T\mathbf{F}\mathbf{\Psi}\mathbf{F}^T\cdots\mathbf{F}\mathbf{\Psi}\mathbf{F}^T = \mathbf{Q}_1^p\,.$$

The expansion of Fact 4 can now directly be performed on the $\mathbf{Q}$ matrix which provides a mechanism to translate the parameters of an EAR model to those of a higher order CAR model. In order to simplify the matrix expansion we define the order of regular lattice neighbors in a slightly non-standard way. Define the

following neighbor incidence matrices ordered by distance:

$$\gamma\{1\}, \gamma\{1,1\}, \gamma\{2\}, \gamma\{1,2\}, \gamma\{2,1\}, \gamma\{2,2\}, \gamma\{3\}, \gamma\{1,3\}, \gamma\{3,1\}, \gamma\{2,3\}, \gamma\{3,2\}, \gamma\{4\},$$
$$\gamma\{1,4\}, \gamma\{4,1\}, \gamma\{3,3\}, \gamma\{2,4\}, \gamma\{4,2\}, \gamma\{5\}, \gamma\{4,3\}, \gamma\{3,4\}, \cdots$$

Here $\gamma\{i\}$ denotes the neighbor incidence matrix for a set of neighbors in the primary directions (east, north, west, and south) at distance i. $\gamma\{k,l\}$ denotes the indicator matrix of a set of neighbors in the "diagonal" directions obtained by moving k nodes in a forward direction and l nodes turning left starting in any of a all four primary directions. Obviously each $\gamma\{k,l\}$ corresponds to a set of 4 neighbors at each node of the torus embedded lattice. Further the distance associated with these neighbors is $d\{k,l\} = (k^2 + l^2)^{1/2}$. Note that in this notation we can also define $\gamma\{0\} = \gamma\{0,0\} = \mathbf{I}$ (the identity matrix) and $\gamma\{i,0\} = \gamma\{0,i\} = \gamma\{i\}$.

Fact 6: The powers of a first order neighbor incidence matrix $\gamma\{1\}$ over a torus are given as follows:

$$\gamma\{1\}^2 = \gamma\{2\} + 2\gamma\{1,1\} + 2^2\gamma\{0\}$$

$$\gamma\{1\}^3 = \gamma\{3\} + 3\gamma\{2,1\} + 3\gamma\{1,2\} + 3^2\gamma\{1\}$$

$$\gamma\{1\}^4 = \gamma\{4\} + 4\gamma\{3,1\} + 6\gamma\{2,2\} + 4\gamma\{1,3\} + 4[4\gamma\{2\} + 6\gamma\{1,1\}] + 6^2\gamma\{0\}$$

$$\gamma\{1\}^5 = \gamma\{5\} + 5\gamma\{4,1\} + 10\gamma\{3,2\} + 10\gamma\{2,3\} + 5\gamma\{1,4\}$$
$$+ 5[5\gamma\{3\} + 10\gamma\{2,1\} + 10\gamma\{1,2\}] + 10^2\gamma\{1\}$$

$$\gamma\{1\}^6 = \gamma\{6\} + 6\gamma\{5,1\} + 15\gamma\{4,2\} + 20\gamma\{3,3\} + 15\gamma\{2,4\} + 6\gamma\{1,5\}$$
$$+ 6[6\gamma\{4\} + 15\gamma\{3,1\} + 20\gamma\{2,2\}$$
$$+ 15\gamma\{1,3\}] + 15[15\gamma\{2\} + 20\gamma\{1,1\}] + 20^2\gamma\{0\} .$$

The pattern of these multiplications is more clearly illustrated by the graphs given in Figure 9.

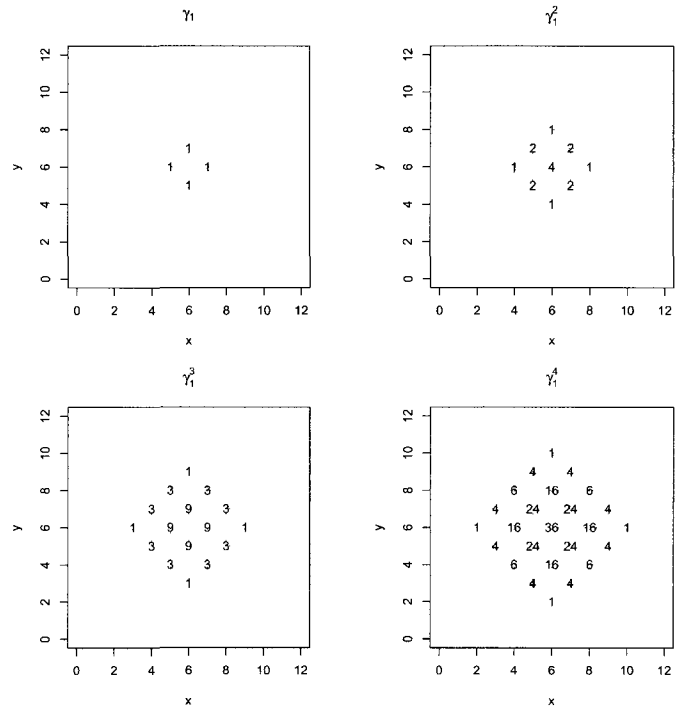


Figure 9: *Patterns of neighbor weights corresponding to powers of the first order incidence matrix* γ_1.

These equations can be programmed using the following:

$$\begin{aligned}\gamma^k = \gamma\{1\}^k &= \sum_{j=0}^{k-1} \gamma\{k-j, j\} + \binom{k}{1} \sum_{j=1}^{k-2} \binom{k}{j} \gamma\{k-j-1, j-1\} \\ &\quad + \binom{k}{2} \sum_{j=2}^{k-3} \binom{k}{j} \gamma\{k-j-2, j-2\} + \cdots \\ &\quad + \binom{k}{[m]} \binom{k}{[m]} \gamma\{k-2[m], k-2[m]\} \\ &= \sum_{i=0}^{[m]} \left[\binom{k}{i} \sum_{j=i}^{\max\{k-1-i, i\}} \binom{k}{j} \gamma\{k-j-i, j-i\} \right],\end{aligned}$$

where $[m]$ is the largest integer $\leq k/2$.

Fact 7: Consider a CAR model on a torus. Then in our representation:

Let $\Lambda = \text{diag}(\lambda_i) = \text{diag}(\text{eigenvalues of}(\gamma_1 - d_1\mathbf{I})) = \mathbf{F}^T(\gamma_1 - d_1\mathbf{I})\mathbf{F} = \mathbf{F}^T\gamma_1\mathbf{F} - d_1\mathbf{I} = \Delta - d_1\mathbf{I}$; where $\Delta = \text{diag}(\delta_i) = \text{diag}(\text{eigenvalues}(\gamma_1))$.

Fact 7a: For the first-order precision matrix $\mathbf{Q}_1 = I - \psi(\gamma_1 - \mathbf{D}_1)$, $\mathbf{D}_1 = d_1\mathbf{I} = (4-1)\mathbf{I} = 3\mathbf{I}$. Using the letter κ instead of ψ, $\mathbf{Q}_1$ turns out to be

$$\mathbf{Q}_1 = (3\psi + 1)\mathbf{I} - \psi\gamma_1 = w\mathbf{I} - w\kappa\gamma_1 = \frac{1}{1-3\kappa}(\mathbf{I} - \kappa\gamma_1),$$

where, $\kappa = \frac{\psi}{1+3\psi}$ and $w = \frac{1}{1-3\kappa}$. Notice that as the range of ψ varies from $0 < \psi < 1$ the range of κ is $0 < \kappa < \frac{1}{4}$. Further, it follows that

$$\mathbf{Q}_1 = \mathbf{F}(\mathbf{I} - \psi\mathbf{\Lambda})\mathbf{F}^T = \mathbf{F}(\mathbf{I} - \psi(\mathbf{\Delta} - 3\mathbf{I}))\mathbf{F}^T = \mathbf{F}\left(\frac{1}{1-3\kappa}(\mathbf{I} - \kappa\mathbf{\Delta})\right)\mathbf{F}^T .$$

Therefore, using Fact 5: $\mathbf{Q}_p = \mathbf{F}\text{diag}(1 - \psi\lambda_i)^p\mathbf{F} = \mathbf{Q}_1^p$ for some positive integer p; and hence:

$$\mathbf{Q}_1^p = \left(\frac{1}{1-3\kappa}\right)^p (\mathbf{I} - \kappa\gamma_1)^p = \left(\frac{1}{1-3\kappa}\right)^p \sum_{i=0}^{p} \begin{pmatrix} p \\ i \end{pmatrix} (-\kappa)^i \gamma_1^i .$$

The equation above indicates that, on a regular torus with inter-node distance= 1, the precision matrix of the isotropic EAR model with an integer smoothness parameter $\theta = p > 0$ is a linear combination of neighbor incidence matrices (including the identity matrix) of neighbors up to including a distance of p. Such a linear combination defines a higher order CAR model.

When the smoothness parameter θ in the EAR model is not an integer, the explicit connections between EAR and higher-order CAR model is not available. However, θ still governs the smoothness of the spatial process, and can be interpreted as a smoothness parameter in the EAR model. Through the matrix logarithm and exponential, we will demonstrate that when θ is not an integer, the $\mathbf{Q}$ matrix in the EAR is still valid.

Recall that the exponential of a $n \times n$ matrix $\mathbf{X}$ can be written as

$$e^{\mathbf{X}} = \mathbf{I} + \frac{1}{1!}\mathbf{X} + \frac{1}{2!}\mathbf{X}^2 + \frac{1}{3!}\mathbf{X}^3 + \cdots ,$$

and the logarithm of $\mathbf{I} + \mathbf{X}$ is

$$\ln(\mathbf{I} + \mathbf{X}) = \mathbf{X} - \frac{\mathbf{X}^2}{2} + \frac{\mathbf{X}^3}{3} - \frac{\mathbf{X}^4}{4} + \cdots .$$

The precision matrix $\mathbf{Q}$ in the EAR model can be expressed as

$$\mathbf{Q} = \mathbf{Q}_1^{\theta} = \left[\mathbf{I} - \psi(\gamma_1 - \mathbf{D}_1)\right]^{\theta} .$$

Let $\mathbf{F}$ and Λ be matrices of eigenvectors and eigenvalues of $\gamma_1 - \mathbf{D}_1$, then,

$$\begin{aligned}
\mathbf{F}^T\mathbf{Q}\mathbf{F} &= \mathbf{F}^T\left[\mathbf{I} - \psi(\gamma_1 - \mathbf{D}_1)\right]^{\theta}\mathbf{F} \\
&= \mathbf{F}^T \cdot \exp\left\{\theta \log(\mathbf{I} - \psi(\gamma_1 - \mathbf{D}_1))\right\} \cdot \mathbf{F} \\
&= \mathbf{F}^T \cdot \exp\left\{\theta\left(-\psi(\gamma_1 - \mathbf{D}_1) - \frac{\psi^2}{2}(\gamma_1 - \mathbf{D}_1)^2 - \frac{\psi^3}{3}(\gamma_1 - \mathbf{D}_1)^3 + \cdots\right)\right\} \cdot \mathbf{F} \\
&= \mathbf{F}^T \cdot \left[e^{-\theta\psi(\gamma_1 - \mathbf{D}_1)} \times e^{-\theta\frac{\psi^2}{2}(\gamma_1 - \mathbf{D}_1)^2} \times e^{-\theta\frac{\psi^3}{3}(\gamma_1 - \mathbf{D}_1)^3} \times \cdots\right] \cdot \mathbf{F} .
\end{aligned}$$

Note that for any square, positive definite matrices A and B, we have

$$\mathbf{F}^T \cdot [\mathbf{A}\mathbf{B}] \cdot \mathbf{F} = \mathbf{F}^T\mathbf{A}\mathbf{F} \cdot \mathbf{F}^T\mathbf{B}\mathbf{F} .$$

So,

$$\mathbf{F}^T\mathbf{Q}\mathbf{F} = \mathbf{F}^T e^{-\theta\psi(\gamma_1-\mathbf{D}_1)}\mathbf{F}\cdot\mathbf{F}^T e^{-\theta\frac{\psi^2}{2}(\gamma_1-\mathbf{D}_1)^2}\mathbf{F}\cdot\mathbf{F}^T e^{-\theta\frac{\psi^3}{3}(\gamma_1-\mathbf{D}_1)^3}\mathbf{F}\cdot\cdots \ .$$

For any order k, the exponential of $e^{c(\gamma_1-\mathbf{D}_1)^k}$ (c is a constant) can be written as,

$$e^{c(\gamma_1-\mathbf{D}_1)^k} = \mathbf{I} + \frac{1}{1!}[c(\gamma_1-\mathbf{D}_1)^k] + \frac{1}{2!}[c(\gamma_1-\mathbf{D}_1)^k]^2 + \frac{1}{3!}[c(\gamma_1-\mathbf{D}_1)^k]^3 + \cdots \ .$$

Thus,

$$\begin{aligned}\mathbf{F}^T\cdot e^{c(\gamma_1-\mathbf{D}_1)^k}\cdot\mathbf{F} &= \mathbf{I} + \frac{1}{1!}[c(\Lambda)^k] + \frac{1}{2!}[c(\Lambda)^k]^2 + \frac{1}{3!}[c(\Lambda)^k]^3 + \cdots \\ &= e^{c\Lambda^k} \ .\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbf{F}^T\mathbf{Q}\mathbf{F} &= e^{-\theta\psi\Lambda}\cdot e^{-\theta\frac{\psi^2}{2}\Lambda^2}\cdot e^{-\theta\frac{\psi^3}{3}\Lambda^3}\cdot\cdots \\ &= \exp\left\{-\theta\left(\psi\Lambda + \frac{\psi^2}{2}\Lambda^2 + \frac{\psi^3}{3}\Lambda^3 + \cdots\right)\right\} \\ &= \exp\{\theta\log(\mathbf{I}-\psi\Lambda)\} \\ &= (\mathbf{I}-\psi\Lambda)^\theta \ .\end{aligned}$$

Note that when θ is not an integer, the precision matrix can be treated as an exponential and logarithm of the weighted combination of infinite neighbor matrices coefficients. In fact, as the neighbor distance d gets larger, the corresponding

contribution of those neighbors with distance d to the conditional mean quickly approaches zero, since the weight

$$\frac{[\theta(\psi^m/m)]^k}{k!} \to 0, \quad \text{as } m \to \infty, k \to \infty,$$

where, m and k correspond to the $m-$th and $k-$th expansion of logarithm and exponential, respectively.

5.4 Connections to the Matérn Class of Covariance Matrices

One of the most popular covariance structures in spatial statistics is the Matérn class which provides a family of covariance functions with two parameters,

$$C(d;\rho,\nu) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}d}{\rho}\right)^{\nu} \mathcal{K}_{\nu}\left(\frac{2\sqrt{\nu}d}{\rho}\right),$$

where d is distance, σ^2 is the variance of the process, ρ is the range parameter, and $\mathcal{K}_{\nu}(\cdot)$ is the modified Bessel function of the second kind, whose order is the differentiability parameter, $\nu > 0$ (see Stein, 1999). This covariance function has the desirable property that sample functions of Gaussian processes parameterized with the covariance are $\lceil \nu - 1 \rceil$ times differentiable, where $\lceil \cdot \rceil$ is the ceiling function. When $\nu = 0.5$, the Matérn reduces to the exponential covariance function; when $\nu = 1.5$, the Matérn class is the same as Whittle's covariance function, and when $\nu \to \infty$, it has the form of Gaussian covariance structure.

The spectral density of the Matérn covariance, evaluated at spectral frequency, ω, is

$$f(\boldsymbol{\omega};\rho,\nu) = \sigma^2 \frac{\Gamma(\nu+\frac{D}{2})(4\nu)^\nu}{\pi^{\frac{D}{2}}\Gamma(\nu)(\pi\rho)^{2\nu}} \cdot \left(\frac{4\nu}{(\pi\rho)^2} + \boldsymbol{\omega}^T\boldsymbol{\omega}\right)^{-(\nu+\frac{D}{2})},$$

where D is the dimension of the space. Generally, $D = 2$ for two-dimensional space.

In this section we provide a motivation for the particular form of the EAR model given in Definition 1. The connection is illustrated by the fact that the spectral density function of the Matérn class has the same general form as the EAR Fourier transformed covariance matrix, and that the smoothness parameter in the Matérn class has the same role as the exponent parameter θ in the EAR model.

To build the connections between the Matérn spectral density function and the EAR Fourier transformed covariance matrix, we use one of the Matérn class parameterization leading to the following spectral density function (SPD) (see eg. Schabenberger & Gotway, 2005):

$$f(\boldsymbol{\omega};a,\nu^*) = \tau^* \left(\frac{1}{a^2 + \|\boldsymbol{\omega}\|^2}\right)^{\nu^*+\frac{D}{2}}.$$

Here a plays the role of a decay parameter of the covariance function, or equivalently, $1/a$ represents a range parameter and D denotes the space dimension, usually $D = 2$. Through the comparison between $f(\boldsymbol{\omega};\rho,\nu)$ and $f(\boldsymbol{\omega};a,\nu^*)$, we note that ν and ν^* are exactly the same, and

$$a = \frac{2\sqrt{\nu}}{\pi\rho},$$

and τ^* is a multiple of the variance parameter as follows:

$$\tau^* = \sigma^2 a^{\nu} \frac{\Gamma(\nu + \frac{D}{2})}{\Gamma(\nu)\pi^{\frac{D}{2}}} .$$

Comparing this to the EAR Fourier transformed covariance matrix

$$\sigma^2 \text{diag}\left(\frac{1}{1 - \psi\lambda_i}\right)^{\theta} ,$$

we can connect these two models via

$$\sigma^2 \text{diag}\left(\frac{1}{1 - \psi\lambda_i}\right)^{\theta} = \sigma^2 \text{diag}\left(\frac{1/\psi}{1/\psi + (-\lambda_i)}\right)^{\theta} = \frac{\sigma^2}{\psi^{\theta}} \text{diag}\left(\frac{1}{1/\psi + (-\lambda_i)}\right)^{\theta} .$$

Hence the interaction parameter ψ in the EAR model corresponds to $(1/a^2)$ in the Matérn representation, and $\theta = \nu^* + 1$ for spatial data in a two-dimensional space.

To illustrate our connections between the Matérn class covariance function under geostatistical modeling and the EAR representation under Gaussian Markov random field, and in addition, to explore the linear or nonlinear relations between parameters in both structures, we will generate data on a 30×30 regular lattice, as shown in Figure 10. Note that the four corners only have two first-order neighbors (distance = 1), and edge locations have three neighbors. They have fewer neighbors than these sites inside, which is the famous "edge problem" that we mentioned in the very beginning of the dissertation. Define the first-order precision matrix

$$\mathbf{Q}_1 = \mathbf{I} - \psi(\gamma_1 - \mathbf{D}_1) ,$$

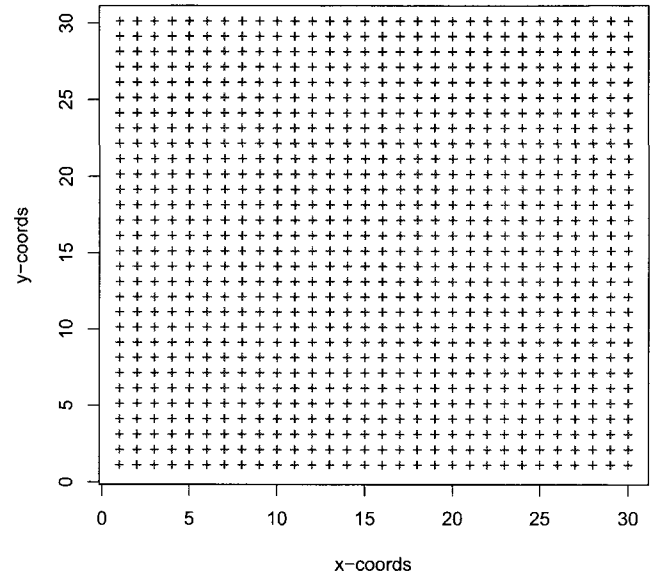


Figure 10: *An illustration of regular* 30×30 *grids*

where, γ_1 is the first-order neighbor matrix using the uniform function proposed by Pettitt *et al.* (2002), and $\mathbf{D}_1 = \text{diag}\left(\sum_j \gamma_{ij} - 1\right)$.

The matrix diagonalization procedure will be applied to $\gamma_1 - \mathbf{D}_1$ to obtain its eigenvalues $\{\lambda_i, i = 1, \cdots, n\}$ and eigenvectors $\mathbf{F}$. Recall that the EAR model is specified via the precision matrix,

$$\mathbf{Q} = \frac{1}{\sigma^2}\mathbf{F}\text{diag}(1 - \psi\lambda_i)^{\theta}\mathbf{F}^T ,$$

or via its covariance matrix,

$$\mathbf{\Sigma} = \sigma^2\mathbf{F}\text{diag}\left(\frac{1}{1 - \psi\lambda_i}\right)^{\theta}\mathbf{F}^T .$$

In our illustration, σ^2 is fixed to be 1, and various values of ψ and θ will be assigned. Therefore, we can obtain the covariance matrix $\mathbf{\Sigma}$ easily. Each row of the covariance matrix will then be divided by its diagonal element to obtain the correlation matrix $\mathbf{R}$. The connections shown above indicate that this correlation matrix is, in some ways, connected to a Matérn correlation function with parameters ρ and ν. Our strategy here is to estimate ρ and ν in Matérn from the correlation matrix $\mathbf{R}$ in EAR, and once ρ and ν are estimated, relations between (ρ, ν) and (ψ, θ) can be further examined. We will calculate all correlation values in the matrix $\mathbf{R}$ for all locations at the regular grid distances $0, 1, \sqrt{2}, 2, \cdots$ up to a maximal distance $d_{\max}$, here we choose $d_{\max} = 15$. Collecting all correlation values of $\mathbf{R}$ into a column vector $\mathbf{y}$ and letting $g(\mathbf{d}; \rho, \nu)$ be a column consisting of correlation values from Matérn class with distances $\mathbf{d}$, we will minimize the sum of square errors,

$$\arg\min_{\rho,\nu} \left\{ (g(\mathbf{d}; \rho, \nu) - \mathbf{y})^T (g(\mathbf{d}; \rho, \nu) - \mathbf{y}) \right\} ,$$

and obtain parameter estimates for ρ and ν using *nlminb* of the software R.

Figure 11 presents an example of Matérn fitting based on EAR correlation specification. In the EAR model, parameters are fixed to $\psi = 0.8$, and $\theta = 3$. Notice that, vertical grey points are correlation values corresponding to distances $0, 1, \sqrt{2}, 2, \cdots$ uo to 15 from the matrix $\mathbf{R}$ in the EAR model, the black line connects all average correlations in each group of distances. The red line is the fitted Matérn correlation function using the grey points. Estimated parameters for Matérn class are $\rho = 5.60$ and $\nu = 1.73$, respectively. It can be clearly identified that black and red lines are

almost surely overlapped. This shows that the Matérn correlation function can almost perfectly be fitted to these correlation values from the EAR model. In other words, the Matérn correlation structure and the EAR model specification are well connected. Smoothness parameters for both θ and ν have the same function and interpretation: they represent the smoothness of the spatial process. The parameters ρ and ϕ, although defined in different ways, they both quantify the association between locations within certain distances.

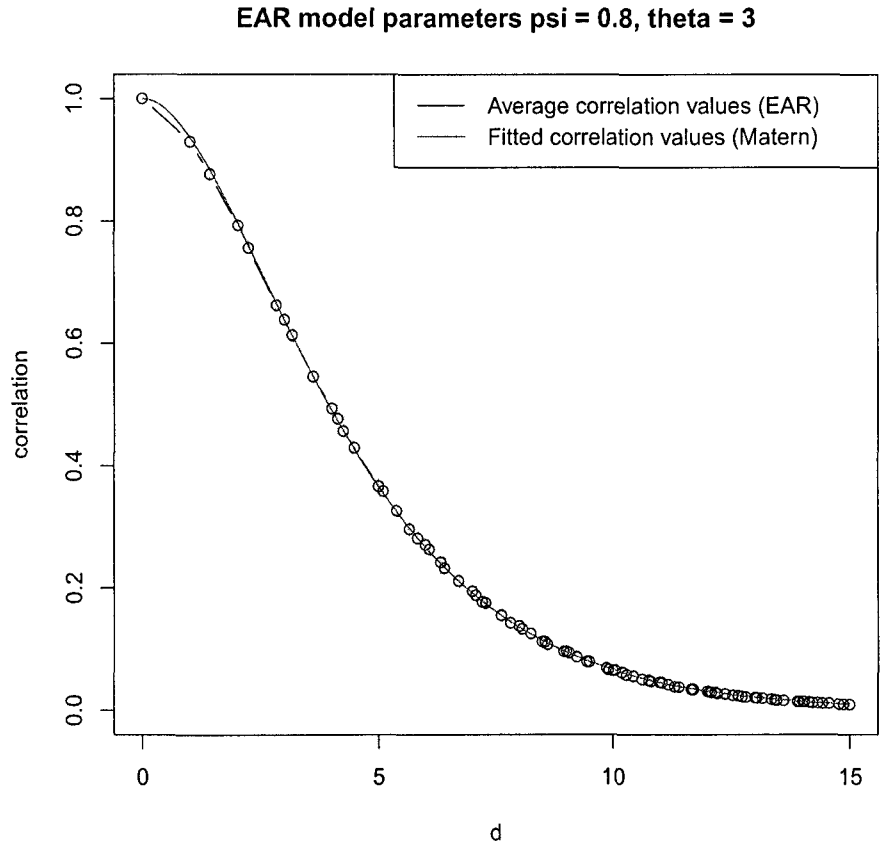


Figure 11: *Fitted Matérn correlation function resulting estimated parameters* $\rho = 5.60$ *and* $\nu = 1.73$. *The grey points are correlation values from correlation matrix in EAR*$(\psi = 0.8, \theta = 3)$ *model; the black line connects the average correlation values in each distance group; the red line is the fitted Matérn correlation function.*

To explore the possible relationships between parameters in the EAR model and in the Matérn correlation function, we estimate and fit parameters ρ and ν given different values of ψ and θ. Figure 12 shows the Matérn class fit with EAR model parameters $\psi = 0.6$ fixed and various integer values of $\theta = 1, 2, \cdots, 20$. The Matérn correlation functions provide nearly perfect fits. Notice that as θ goes larger, the Matérn class also goes smoother.

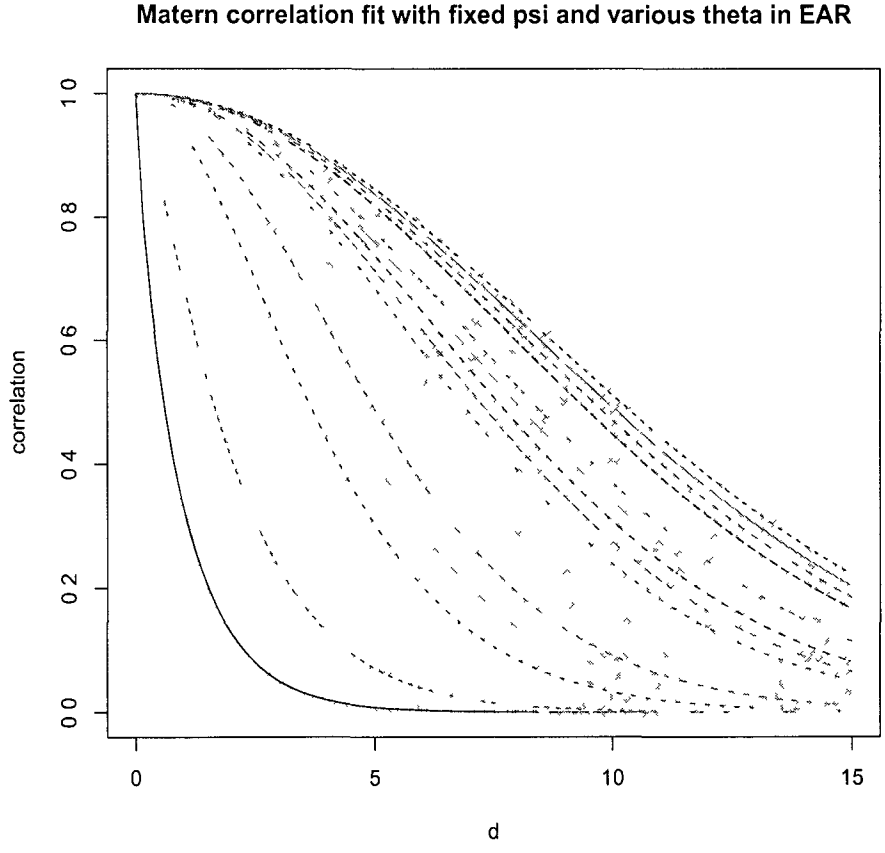


Figure 12: *Matérn class fit for various smoothness parameters θ and fixed correlation parameter $\psi = 0.6$ in the EAR model. Grey points denote the average correlation values at each distance. Lines with different colors are fitted Matérn functions.*

Table 7 presents the fitted values for parameters ν and ρ in Matérn function. Figure 13 shows the visualization of fitted ρ and ν versus θ. An exponential pattern

for ν versus θ can be clearly detected, and so do the square relationship between ρ and θ. We thus take an exponential fit for ν and θ as

$$E(\log(\nu)) = \beta_0 + \beta_1 \theta$$

and take a square fit for ρ and θ as

$$E(\rho) = \beta_0 + \beta_1 \sqrt{\theta}$$

where $E(\cdot)$ is the expectation, β_0 and β_1 are regression parameters. Table 8 presents the estimated coefficients for each fitting. Note that as the smoothness parameter θ in the EAR increases (especially > 15), the smoothness parameter ν in the Matérn class increases exponentially. One possible reason is due to the restrictions of grids. The lattice here is a 30×30 regular grid, which has been fixed. As θ increases, more and more neighbors will be included to serve as conditional weights, which may cover all grid points, and thus results in a process that is too smooth. We do expect that if our grids is wrapped onto a torus, or the grids is dynamic (infinite), the relationship between ν and θ could be linear. Moreover, note that the range parameter ρ also increases when θ increases. This in another way reflects the famous estimation issue for the Matérn class, in which not all parameters can be estimated consistently, but one property can (Zhang, 2004).

Another interest is to explore the relations between spatial correlation parameter ψ in the EAR model and the parameters ρ and ν in the Matérn class when the

Table 8: *Parameter estimates of exponential fitting for ν and θ, and square fitting for ρ and θ.*

Model fitting	β_0	β_1
Exponential	-0.508	0.272
Square	-1.400	3.087

where $E(\cdot)$ is the expectation, β_0 and β_1 are regression parameters. Table 8 presents the estimated coefficients for each fitting. Note that as the smoothness parameter θ in the EAR increases (especially > 15), the smoothness parameter ν in the Matérn class increases exponentially. One possible reason is due to the restrictions of grids. The lattice here is a 30×30 regular grid, which has been fixed. As θ increases, more and more neighbors will be included to serve as conditional weights, which may cover all grid points, and thus results in a process that is too smooth. We do expect that if our grids is wrapped onto a torus, or the grids is dynamic (infinite), the relationship between ν and θ could be linear. Moreover, note that the range parameter ρ also increases when θ increases. This in another way reflects the famous estimation issue for the Matérn class, in which not all parameters can be estimated consistently, but one property can (Zhang, 2004).

Another interest is to explore the relations between spatial correlation parameter ψ in the EAR model and the parameters ρ and ν in the Matérn class when the smoothness parameter θ is fixed. Figure 14 shows the Matérn class fit for fixed $\theta = 3$ and twenty various ψ ranging from 0.1 to 0.99. The fit still is excellent. However, we notice in Figure 15, when ψ is near the boundaries 0 or 1, the fitted ρ appears unstable. It decreases as ψ changes from 0.01 to 0.3, and increases with ψ from 0.3 to 0.94, and then decreases again. We can detect a nearly perfect pattern

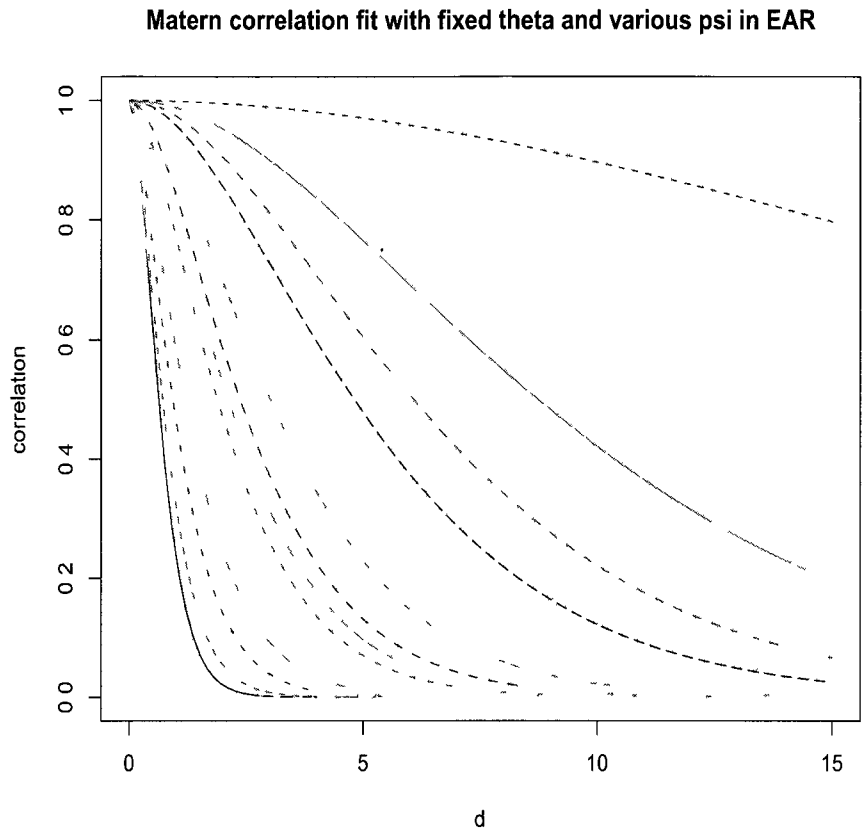


Figure 14: *Matérn class fit for various correlation parameters ψ and fixed smoothness parameter $\theta = 3$ in the EAR model. Grey points denote the average correlation values at each distance. Lines with different colors are fitted Matérn functions.*

only for a limited range of ψ values; probably it is due to the fixed grid problem we discussed above.

As discussed in section 5.3, when the smoothness parameter θ is an integer value p, the precision matrix $\mathbf{Q}_1^p$ can be expressed as a linear combination of different orders of incidence matrices,

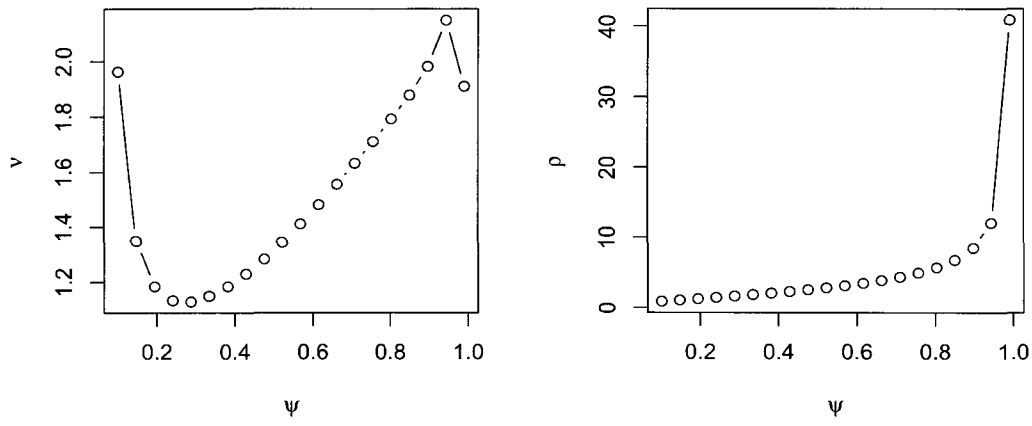


Figure 15: *Fitted values of ν and ρ in the Matérn class versus ψ in the EAR model (θ is fixed).*

$$\mathbf{Q}_1^p = \left(\frac{1}{1-3\kappa}\right)^p (\mathbf{I} - \kappa\gamma_1)^p = \left(\frac{1}{1-3\kappa}\right)^p \sum_{i=0}^{p} \binom{p}{i} (-\kappa)^i \gamma_1^i \ .$$

Note that the signs of the coefficients are fixed and are the same for all values of p where they are overlapping. The relative values of these coefficients corresponding to $Q^m_{1\,ij}/Q^m_{1\,ii}$ are shown in Figure 16. There is an interesting pattern emerging that is related to the Matérn class correspondence. The envelope constructed for the absolute values of the weights corresponds roughly to a Matérn correlation function for the corresponding value of θ . Surprisingly this envelope function is relatively stable for changing values of ψ , which may be explained by the fact that we used relative weights. Nevertheless the correspondence between the EAR weights and the Matérn correlation function is not exact. In particular for small values of θ and ψ , there is a slight difference. However for practical purposes these differences

may be irrelevant. Note that for Figure 16 once can find by trial that a range parameter of $a^{-1} = 0.36$ in the Matérn correlation function produces the closest correspondence between the Matérn class and the EAR weights. We suspect that this particular value arises due to the particular parameterization chosen for the Matérn class.

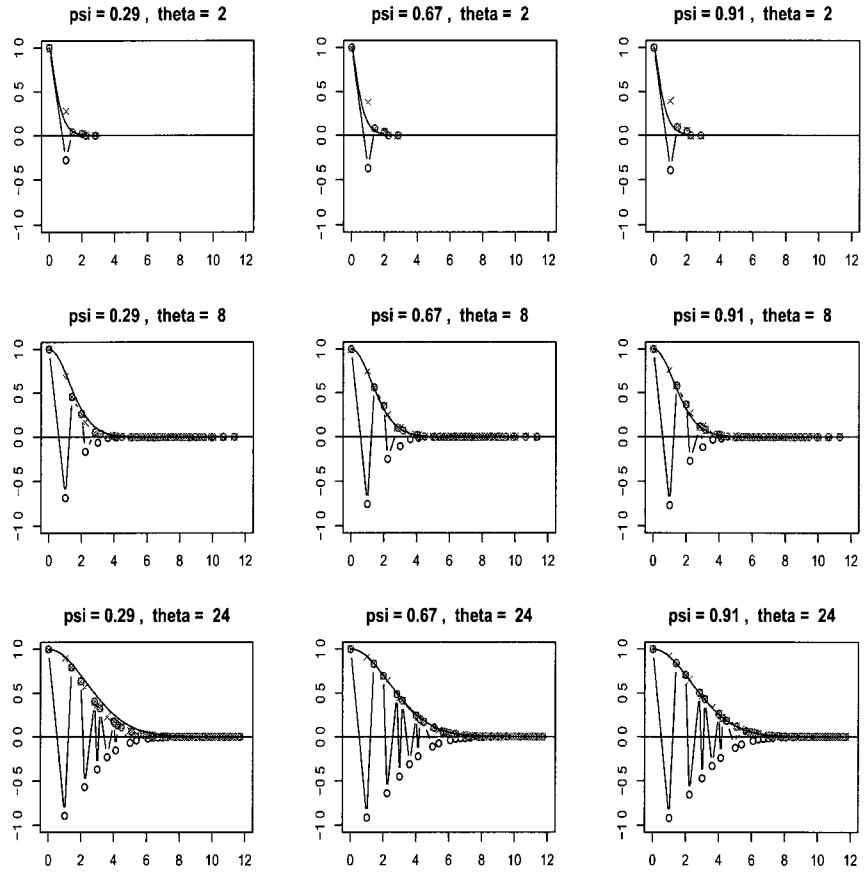


Figure 16: *Calculations of the higher order Markov random field coefficients (weights) as a function of neighbor distance that correspond to an EAR model with θ, for various values of θ (theta) and ψ (psi). Note we plot the relative weights: Q_{1j}/Q_{11} . Also drawn as a smooth line is the Matérn correlation function with $\nu^* = \theta - 1$and range parameter $a^{-1} = 0.36$.*

5.5 Connections to the INLA

A recent increasingly popular stochastic method to geostatistical modeling is the integrated nested Laplace approximation (INLA) that was developed using a stochastic partial differential equations (SPDE) approach (Rue, Martino and Chopin, 2009; Lindgren, Lindström and Rue, 2010). It provides an explicit bridge between Gaussian fields and Gaussian Markov random fields. To briefly explain their method, we use the parameterization of the Matérn function as follows,

$$\text{Cov}(\mathbf{d}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\kappa \|\mathbf{d}\|)^{\nu} K_{\nu}(\kappa \|\mathbf{d}\|) ,$$

where $\kappa > 0$ is the scale parameter and $\nu > 0$ is the smoothness/shape parameter.

It is known that on infinite lattices, fields with Matérn covariances are solutions to an SPDE(Whittle, 1954) based on the Laplacian, $\Delta = \nabla^T \nabla$,

$$(\kappa^2 - \Delta)^{\alpha/2} \mathbf{x}(\mathbf{s}) = \boldsymbol{\epsilon}(\mathbf{s}), \quad \alpha = \nu + D/2 ,$$

where $\boldsymbol{\epsilon}(\mathbf{s})$ is spatial Gaussian white noise, D is the dimension.

A finite element method is used to represent

$$\mathbf{x}(\mathbf{u}) = \sum_{k=1}^{N} \psi_k(\mathbf{u}) w_k$$

for basis functions $\{\psi_k\}$ and Gaussian weights $\{w_k\}$. Note that a stochastic weak formulation of the SPDE states that

$$\langle \phi_k, (\kappa^2 - \Delta)^{\alpha/2}\mathbf{x}\rangle = \langle \phi_k, \boldsymbol{\epsilon}\rangle, \quad k = 1, 2, \cdots$$

for all test functions $\{\phi_k\}$, where $\langle f, g\rangle$ is defined by $\int f(\mathbf{s})g(\mathbf{s})d\mathbf{s}$. Lindgren, Lindström and Rue (2010) show that when $\alpha = 1$, then

$$\phi_k = (\kappa^2 - \Delta)^{1/2}\psi_k \,,$$

and when $\alpha = 2$, then

$$\phi_k = \psi_k \,.$$

They then construct the precision matrices $\mathbf{Q}$ for integers of $\alpha = 1, 2, \cdots$, with $\mathbf{Q}$ specified by,

$$\begin{aligned}
\mathbf{Q}_{1,\kappa} &= \kappa^2\mathbf{C} + \mathbf{G} \\
\mathbf{Q}_{2,\kappa} &= \mathbf{K}\mathbf{C}^{-1}\mathbf{K} \\
&\vdots \\
\mathbf{Q}_{\alpha,\kappa} &= \mathbf{K}\mathbf{C}^{-1}\mathbf{Q}_{\alpha-2,\kappa}\mathbf{C}^{-1}\mathbf{K} \,,
\end{aligned}$$

where,

$$\begin{aligned}
C_{ij} &= \langle \phi_i, \phi_j\rangle, \quad i \neq j \\
C_{ii} &= \langle \phi_i, 1\rangle \\
G_{ij} &= \langle \nabla\phi_i, \nabla\phi_j\rangle \\
\mathbf{K} &= \kappa^2\mathbf{C} + \mathbf{G} \,.
\end{aligned}$$

Taking a regular lattice as an example, the explicit solution to $\mathbf{Q}$ is expressed in the following format, when $\alpha = 2$(or $\nu = 1$), see Rue (2009),

$$\mathbf{Q} = \frac{\kappa^{-2}}{4\pi} \times \begin{bmatrix} & & 1 & & \\ & 2 & -2(4+\kappa^2) & 2 & \\ 1 & -2(4+\kappa^2) & 4+(4+\kappa^2)^2 & -2(4+\kappa^2) & 1 \\ & 2 & -2(4+\kappa^2) & 2 & \\ & & 1 & & \end{bmatrix}.$$

Recall that the EAR model is also shown to be connected to the Matérn class. Notice that in the INLA, the smoothness parameter $\alpha = \nu + D/2$ functions similarly to the parameter θ in the EAR. Below we will show that when the EAR model is wrapped onto a torus, both precision matrices $\mathbf{Q}$ exhibits similar patterns.

Consider a two dimensional regular lattice ($D = 2$) with $\theta = \nu + 1 = 2$. When the EAR model is embedded onto a torus, we know the first-order precision matrix is

$$\mathbf{Q}_1 = (3\psi + 1)\mathbf{I} - \psi\gamma_1 .$$

Let $\eta = 3 + \frac{1}{\psi}$, then

$$\begin{aligned} \mathbf{Q} = \mathbf{Q}_1^2 &= \psi^2 \left(\eta\mathbf{I} - \gamma_1\right)^2 \\ &= \psi^2 \left(\eta^2\mathbf{I} - 2\eta\gamma_1 + \gamma_1^2\right) . \end{aligned}$$

From the powers of first-order incidence matrices in Figure 9, we can clearly see the weights for γ_1 and γ_1^2, which result in

$$\mathbf{Q} = \psi^2 \times \begin{bmatrix} & & 1 & & \\ & 2 & -2(3+\frac{1}{\psi}) & 2 & \\ 1 & -2(3+\frac{1}{\psi}) & 4+(3+\frac{1}{\psi})^2 & -2(3+\frac{1}{\psi}) & 1 \\ & 2 & -2(3+\frac{1}{\psi}) & 2 & \\ & & 1 & & \end{bmatrix} .$$

The INLA approach and the EAR model do have identical pattern for the precision matrix with the relation

$$4 + \kappa^2 = 3 + \frac{1}{\psi} \Longrightarrow 1 + \kappa^2 = \frac{1}{\psi} .$$

When applied to geostatistical data, both of the methods can be applied to do kriging with Gaussian Markov random fields paradigm. For instance, suppose the observation model is

$$\mathbf{y} \sim N(\mathbf{Kx}, \boldsymbol{\Sigma}) ,$$

and the prior distribution model of $\mathbf{x}$ is

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \mathbf{Q}^{-1}) .$$

The posterior density $p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ is also a GMRF, with precision

$$\mathbf{Q}_{\mathbf{x}|\mathbf{y}} = \mathbf{Q} + \mathbf{K}^T\boldsymbol{\Sigma}\mathbf{K}\,,$$

and expectation

$$\boldsymbol{\mu}_{\mathbf{x}|\mathbf{y}} = \mathbf{Q}_{\mathbf{x}|\mathbf{y}}^{-1}\left(\mathbf{Q}\boldsymbol{\mu} + \mathbf{K}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}\right)\,.$$

Notice that in the INLA method, a lattice is constructed via triangulations to construct the precision matrix; while in the EAR model, a regular grid is created and the first-order neighbor incidence matrix is used to generate a precision matrix. For the EAR model, a dense grid will be created to conduct kriging, the idea of which is based on the Paciorek's method (2007) as discussed in section 3.2.4. Application of the EAR model in geostatistics will be later discussed in Chapter 6.

5.6 Identifiability Issue: the Intrinsic EAR Model

Both Czado's and Pettitt's versions of the CAR model have the intrinsic CAR in the limit, when $\psi \to 1$(or $\phi \to \infty$), and the conditional variance of $z_i|\mathbf{z}_{-i}$ decreases to σ^2/N_j. When ψ goes to zero (no spatial dependency), all partial correlations between z_j and z_i given all the other sites are the same. In the EAR model, the spatial interaction parameter ψ and the smoothness parameter θ are both included and need to be estimated. It has been shown that the parameters ψ and θ are somehow corresponding to the range and smoothness parameters in the Matérn class covariance functions, given the variance terms σ^2 are equal to 1. Zhang (2004) showed non-consistency in parameter estimation for the Matérn class of geosta-

tistical models indicating an identifiability problem. Similar parameter estimation non-consistency and identifiability issues also exisit in the EAR model. As an illustration, Figure 17 shows the simulated Gaussian random fields with mean 0 and EAR specification with parameters $\psi = 0.1, 0.6$ and 0.9, and $\theta = 1, 4$ and 10. The σ^2 is assumed to be 1. It can be noted that random fields in the upper right corner and in the lower left corner appear to have similar smoothness and patterns. Their corresponding parameters are $\psi = .1, \theta = 10$ and $\psi = 0.9, \theta = 1$, respectively. This, in some sense, points to the identifiability issue of the EAR model: The situation of a high value of the smoothness parameter θ with a low value of the interaction parameter ψ can not clearly be distinguished by data from the situation of a low θ value with a high ψ value. The distance between grid points in a regular lattice represents a maximal resolution, and intuitively it is understandable that strong spatial interaction cannot objectively be distinguished from smoothness.

Figure 18 shows the image plot for the -2loglikelihood for various parameter values of ψ and θ. The data is simulated on a regular 30×30 lattice from the EAR model with values of parameters $\psi = 0.5$, $\theta = 4$ and $\sigma^2 = 1$. The same values of the likelihood are clearly noticeable in the dark blue area. The likelihood for the small ψ and large θ (say, $\psi = 0.2, \theta = 8$) is quite close to that for the large value of ψ and small value of θ (say, $\psi = 0.8, \theta = 3$) . The likelihood appears to be roughly constant along curves from top left to bottom right.

The remedy of the identifiability issues can be proposed by using an intrinsic version of the EAR model for spatial data. When the EAR model is applied to a spatio-temporal process, repeated measurements for each location at different times

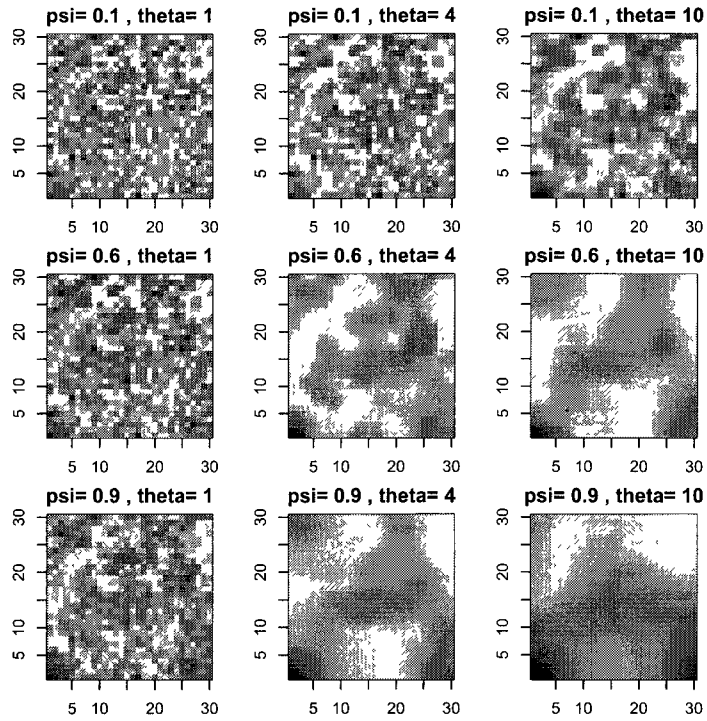


Figure 17: *Simulated random fields with mean 0 of the EAR model with* $\psi = 0.1, 0.6$ *and* 0.9 *for* $\theta = 1, 4$ *and* 10*, and* $\sigma^2 = 1$*. The same* $\mathbf{x}$ *was used for each realization.*

provide sufficient information and thus can resolve the identifiability issue. Recall that the intrinsic CAR model has been widely used in application, for instance, disease mapping, image analysis etc, since the work by Besag, York, and Mollie (1991). Intrinsic CAR models are rank deficient versions of the CAR model that are invariant with respect to linear contrasts. They are not proper models but suitable as prior models where linear contrasts are required. The popular rank $n-1$ intrinsic CAR model is used widely as a spatial random effects prior i.e. a spatial a random field $\mathbf{z}$, where the defining contrast is $\sum_i \mathbf{z}_i = 0$. Specification of such an intrinsic CAR model is equivalent to specifying that the conditional means are averages of

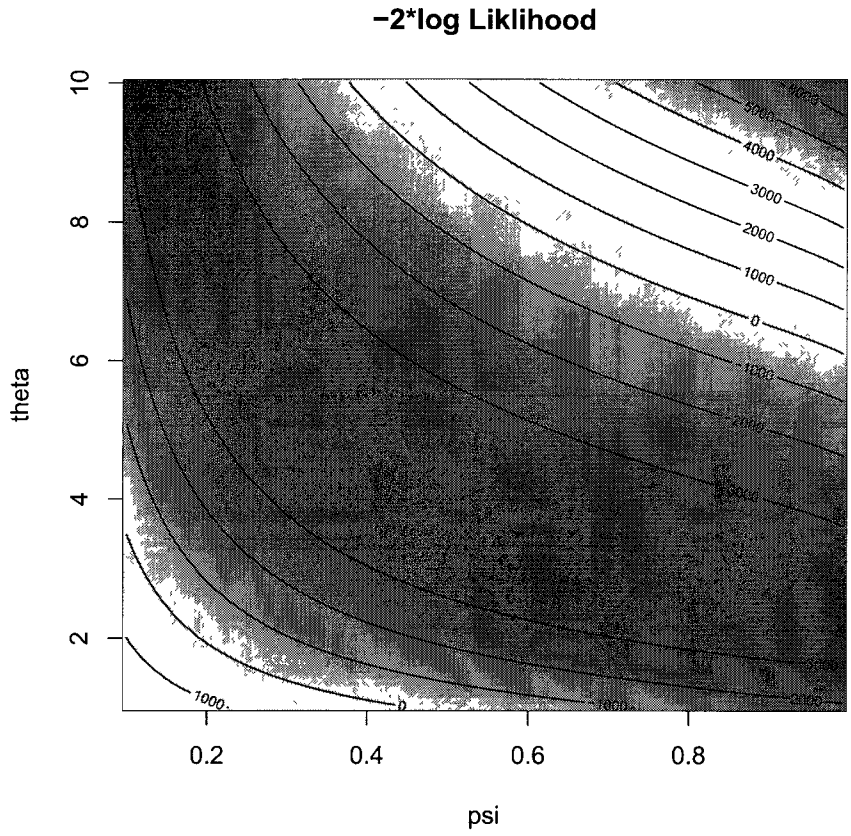


Figure 18: *The profile log likelihood of EAR model to the simulated data in a regular* 30×30 *grids with parameters* $\psi = 0.5, \theta = 4$ *and* $\sigma^2 = 1$.

the neighboring values. More specifically if we assume that

$$z_i|\mathbf{z}_i \sim N\left(\mu_i + \frac{1}{N_i}\sum_{j:j\in N(i)}\left(z_j - \mu_j\right), \frac{\sigma^2}{N_i}\right),$$

then $\mathbf{z}$ is said to follow an intrinsic CAR model with corresponding precision matrix

$$\mathbf{Q} = \frac{1}{\sigma^2}\begin{pmatrix} N_1 & 0 & \cdots & 0 \\ 0 & N_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & N_n \end{pmatrix}\begin{pmatrix} 1 & -\gamma_{12}/N_1 & \cdots & -\gamma_{1n}/N_1 \\ -\gamma_{21}/N_2 & 1 & & -\gamma_{2n}/N_2 \\ \vdots & & \ddots & \vdots \\ -\gamma_{n1}/N_n & -\gamma_{n2}/N_n & \cdots & 1 \end{pmatrix} = \frac{1}{\sigma^2}(\mathbf{D} - \gamma) .$$

Here, γ_{ij} is defined as $\gamma_{ij} = 1$ if j is a neighboring site of i, and equals 0 otherwise. Recall that in the Czado's CAR model, when $\psi \to 1$, the conditional variance reduces to σ^2/N_i, resulting in the intrinsic CAR model. Therefore, we can define an intrinsic EAR model as follows,

Definition 5.4 (Intrinsic EAR Model) *A spatial process* $\mathbf{z}$ *defined over a lattice with neighbor index matrix* γ *follows the intrinsic EAR model with parameters* (μ, σ^2, θ) *if*

$$\mathbf{z} \sim N(\mu, \sigma^2\mathbf{Q}^{-1})$$

with $\mathbf{Q}$ *defined by*

$$\mathbf{Q} = \mathbf{F}diag(1 - \lambda_i)^{\theta}\mathbf{F}^T$$

where $\mathbf{F}$ *is the eigenvector matrix of* $\gamma - \mathbf{D}$ *and* $\mathbf{D}$ *is the diagonal matrix of the row sums of* γ *minus 1, and* $\{\lambda_i, i = 1, \cdots, n\}$ *are eigenvalues of* $\gamma - \mathbf{D}$.

Equivalently, the precision matrix can be expressed as:

$$\mathbf{Q} = \mathbf{F}\text{diag}(\xi_i)^{\theta}\mathbf{F}^T ,$$

where $\mathbf{F}$ is the eigenvector matrix of $\mathbf{D} - \gamma$ and $\mathbf{D}$ is the diagonal matrix of the row sums of γ, and $\{\xi_i, i = 1, \cdots, n\}$ are eigenvalues of $\mathbf{D} - \gamma$.

The intrinsic EAR model will be used in parameter estimation and spatial interpolation in Chapter VI.

CHAPTER VI

EAR Model in Geostatistics

In this chapter we develop the framework for applying the EAR model for possibly irregular point-referenced data, i.e. geostatistical data. We adopt a fine grid latent process representation similar to Paciorek (2007) and provide the full conditional distribution required for MCMC estimation. We conclude with a simulation example.

As discussed in Chapter II, the most important task in analyzing geostatistical data is the spatial prediction or the interpolation. One of the most widely used methods for interpolation of spatial data is "kriging", named after Krige (1951) and popularised when Matheron (1963) applied linear interpolation in a geostatistical context. The kriging predictor is a linear combination of observations; and thus suitable for Gaussian data, or data that is Gaussian after appropriate transformation (Box & Cox, 1964). The kriging weights in the linear combination depend on the estimated mean and covariance structure of the data.

Diggle *et al.* (1998) formalized the idea of generalized geostatistical models, with a latent Gaussian spatial process, as the natural extension of kriging models to an exponential family of responses. They used Bayesian estimation, suggesting a Metropolis-Hastings implementation, with the spatial function sampled sequen-

tially at each observation location at each MCMC iteration. However, as we mentioned in Chapter III, this implementation is slow to converge and mix, as well as being computationally inefficient. Paciorek (2007) focuses on a spectral representation via a particular parameterized prior structure that approximates stationary Gaussian processes on a regular grid. In his approach, the latent grids are specified to be fine enough so that the process at the observation locations can be calculated through an incidence matrix, which maps each observation location to only one nearest latent grid location in Euclidean space. Rue and Tjelmeland (2002) provide a link between GMRFs and kriging by showing that a GMRF on a rectangular grid in $\mathbb{R}^2$ can be used to approximate fields with a wide class of covariance functions. As pointed out in their paper, a problem with defining the field on a rectangular grid is that observations seldom fall on the grid points. However, this can be remedied either by assigning each observation to the closest grid point or by letting values at the observations points to be some linear interpolation of the values at nearby grid points.

In this Chapter, we will apply the EAR model as a latent process in rectangular grids to approximate geostatistical data. The grids will be defined fine enough to ensure that each observation only be associated with its most closest grid point. A natural Bayesian model will be considered and specified later. Parameter estimation and interpolation are performed using a MCMC approach.

Suppose the observed geostatistical data are $\mathbf{y}(\mathbf{s}), \mathbf{s} = (s_1, \cdots, s_n)^T$. For simplicity, we assume $E(\mathbf{y}) = 0$. This can be easily extended to the general case of $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ with regression terms. The data is modeled by a latent process $\mathbf{z}(\mathbf{w}), \mathbf{w} =$

$(w_1, \cdots, w_N)^T$ through an incidence matrix $\mathbf{K}$. In our approach, the grids are defined fine enough which indicates that $n < N$.

The observation model can be written as

$$\mathbf{y}_{n\times 1} = \mathbf{K}_{n\times N}\mathbf{z}_{N\times 1} + \boldsymbol{\epsilon}_{n\times 1} ,$$

with,

$$\boldsymbol{\epsilon}_{n\times 1} \sim N(0, \tau_y^{-1}\mathbf{I}) ,$$

where τ_y is the precision parameter for data $\mathbf{y}$. With the assumption of $\mathbf{y}$ following the Gaussian process, we can write

$$\mathbf{y} \sim N(\mathbf{Kz}, \tau_y^{-1}\mathbf{I}) .$$

The prior distribution model for $\mathbf{z}$ is,

$$\mathbf{z} \sim N(0, \mathbf{Q}^{-1}), \text{ which is an EAR model.}$$

Let $\mathbf{F}$ and $\{\lambda_i, i = 1, \cdots, N\}$ be the eigenvectors and eigenvalues, respectively, of the matrix $\gamma - \mathbf{D}$ (see Section 5.1). Then through the pre-whitening procedure, we define $\mathbf{z}^* = \mathbf{F}^T\mathbf{z}$, and obtain

$$\mathbf{z}^* \sim N\left(0, \tau_z^{-1}\text{diag}\left(\frac{1}{1+\psi\lambda_i}\right)^{\theta}\right) ,$$

where τ_z is the precision parameter for the latent process, $0 < \psi < 1$ is the spatial interaction parameter, and θ is the smoothness parameter.

Rewrite $\mathbf{z} = \mathbf{F}\mathbf{z}^*$, the observation model can be written as,

$$\mathbf{y} \sim N(\mathbf{K}\mathbf{F}\mathbf{z}^*, \tau_y^{-1}\mathbf{I}) = N\left(\mathbf{K}^*\mathbf{z}^*, \tau_y^{-1}\mathbf{I}\right) ,$$

where $\mathbf{K}^* = \mathbf{K}\mathbf{F}$.

Given the prior distributions for τ_y, τ_z, ψ and θ, we can get the closed-form conditional posterior distributions for $\tau_y|..., \tau_z|...$ and $\mathbf{z}^*|...$, but not for ψ and θ. Typically gamma priors will be assigned for precision parameters, and thus

$$\pi(\tau_y) \sim \Gamma(a_y, b_y)$$
$$\pi(\tau_z) \sim \Gamma(a_z, b_z) .$$

Since ψ is in the range of $(0, 1)$, a uniform prior or more generally a beta prior can be used. The beta distribution is preferred here since we can tune its parameters to achieve a desired acceptance ratio in the Metropolis posterior sampling. The parameter θ is greater than zero, and thus a log-normal prior will be suitable,

$$v = \log(\theta) \sim N(\mu_\theta, \sigma_\theta^2) .$$

Therefore, the posterior distribution of $\pi(\mathbf{z}^*, \tau_y, \tau_z, \psi, \theta|\mathbf{y})$ is:

$$\begin{aligned}
\pi(\mathbf{z}^*, \tau_y, \tau_z, \psi, \theta|\mathbf{y}) &= \pi(\mathbf{y}|\mathbf{z}^*, \tau_y) \cdot \pi(\mathbf{z}^*|\tau_z, \psi, \theta) \cdot \pi(\tau_y) \cdot \pi(\tau_z) \cdot \pi(\psi) \cdot \pi(\theta) \\
&\propto \tau_y^{n/2} \exp\left\{-\frac{1}{2}\tau_y(\mathbf{y} - K^*\mathbf{z}^*)^T(\mathbf{y} - K^*\mathbf{z}^*)\right\} \\
&\quad \times \tau_z^{N/2} \prod_{i=1}^{N}(1 + \psi\lambda_i)^{\theta/2} \exp\left\{-\frac{1}{2}\tau_z \cdot \mathbf{z}^{*T}\text{diag}(1 + \psi\lambda_i)^{\theta}\mathbf{z}^*\right\} \\
&\quad \times \tau_y^{a_y - 1} \exp\{-b_y\tau_y\} \cdot \tau_z^{a_z - 1} \exp\{-b_z\tau_z\} \cdot \pi(\psi) \cdot \pi(\theta) \,.
\end{aligned}$$

Recall that in Rue and Held (2005), a Gaussian Markov random field $\mathbf{x}$ with expectation $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}$ can be defined via the density

$$\pi(\mathbf{x}) = \frac{|\mathbf{Q}|^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T\mathbf{Q}(\mathbf{x} - \boldsymbol{\mu})\right)$$

and its corresponding canonical form is

$$\pi(\mathbf{x}) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{b}^T\mathbf{x}\right) ,$$

where the mean $\boldsymbol{\mu}$ can be expressed as $\boldsymbol{\mu} = \mathbf{Q}^{-1}\mathbf{b}$.

Let $\Delta = \text{diag}(1 + \psi\lambda_i)^{\theta}$, then the full conditional distributions for $\mathbf{z}^*, \tau_y, \tau_z$ are:

$$\begin{aligned}
&\mathbf{z}^*|... \sim N\left((\tau_y\mathbf{K}^{*T}\mathbf{K}^* + \tau_z\Delta)^{-1}\tau_y\mathbf{K}^{*T}y, (\tau_y\mathbf{K}^{*T}\mathbf{K}^* + \tau_z\Delta)^{-1}\right) \\
&\tau_y|... \sim \Gamma(a_y + \tfrac{n}{2}, b_y + 0.5(\mathbf{y} - \mathbf{K}^*\mathbf{z}^*)^T(\mathbf{y} - \mathbf{K}^*\mathbf{z}^*)) \\
&\tau_z|... \sim \Gamma(a_z + \tfrac{N}{2}, b_z + 0.5\mathbf{z}^{*T}\Delta\mathbf{z}^*) \,.
\end{aligned}$$

Notice that this sampling scheme requires calculation of the inverse of the posterior conditional variance matrix $(\tau_y \mathbf{K}^{*T}\mathbf{K}^* + \tau_z\Delta)^{-1}$, which will not be feasible for large number of grids points. However, since $\mathbf{F}$ is an orthogonal matrix, if $\mathbf{K}^T\mathbf{K}$ were the identity matrix with dimension N, then,

$$\mathbf{K}^{*T}\mathbf{K}^* = (\mathbf{KF})^T\mathbf{KF} = \mathbf{F}^T\mathbf{K}^T\mathbf{KF} = \mathbf{I}\,.$$

This simplifies the variance matrix,

$$(\tau_y \mathbf{K}^{*T}\mathbf{K}^* + \tau_z\Delta)^{-1} = (\tau_y \mathbf{I} + \tau_z\Delta)^{-1}\,,$$

which is a diagonal matrix, and would be easy to calculate. As discussed in Paciorek (2007), assuming no more than one observation per grid cell, $\mathbf{K}^T\mathbf{K} = \mathbf{I}$ can be achieved using a missing data scheme by introducing latent pseudo-observations for all grid cells without any associated data. To illustrate this idea, we simply assume 3 observations $\mathbf{y} = (y_1, y_2, y_3)^T$ with 6 latent grid cells $\mathbf{z} = (z_1, z_2, \cdots, z_6)^T$, and suppose y_1 is associated with z_2, y_2 with z_5, and y_3 with z_1, then this association can be denoted as

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{pmatrix}.$$

Through introducing pseudo-observations $\tilde{y}_4, \tilde{y}_5, \tilde{y}_6$, which are associated to z_3, z_4, z_6, we get

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \tilde{y}_4 \\ \tilde{y}_5 \\ \tilde{y}_6 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \\ z_5 \\ z_6 \end{pmatrix} = \mathbf{K}\mathbf{z} \, .$$

It is obvious that $\mathbf{K}^T\mathbf{K} = \mathbf{I}$.

Collecting pseudo-observations into a vector, $\tilde{\mathbf{y}}$, they can be sampled within the MCMC using a Gibbs step as

$$\tilde{\mathbf{y}} \sim MVN_{N-n}(\tilde{\mathbf{K}}\mathbf{F}\mathbf{z}^*, \tau_y^{-1}\mathbf{I}) \, ,$$

where the matrix $\tilde{\mathbf{K}}$ functions as a bridge to connect grid cells with no associated data to pseudo-observations. For instance, in our example above,

$$\tilde{\mathbf{K}} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} .$$

Now, the "observation data" $\mathbf{y}$ is augmented on the full latent grids, $\mathbf{y} = (\mathbf{y}_{\text{obs}}, \tilde{\mathbf{y}})$, which combines the actual observations with pseudo-observations. Using the MCMC approach, the posterior samples of $\mathbf{z}^*, \tau_y, \tau_z$ can be drawn via the Gibbs sampler, and those of ψ and θ can be drawn by the Metropolis algorithm.

As an example of applying the EAR model as a latent process for geostatistical data, we simulated a Matérn process with mean 0 and true variance, smoothness and range parameters $\sigma^2 = 1$, $\nu = 2$ and $\rho = 3$, respectively. The observations are then generated with added random Gaussian noise with mean 0 and standard deviation 0.5. The total number of 238 geostatistical locations are uniformly distributed in a $[1,30] \times [1,30]$ square panel. The intrinsic EAR model ($\psi = 1$) is then applied to serve as latent process in a 60×60 lattice, where spatial interpolations are carried out. Figure 19 shows the simulated observations (left panel) and spatial interpolations (right panel). With regard to the parameter estimation consistency and accuracy, Zhang (2004) found that parameters in the Matérn class cannot be consistently estimated, and we also have had difficulty in achieving reasonable mixing for two variance components σ_z^2 and σ_y^2 as well as the smoothness parameter θ in the EAR model, as shown in Figure 20. It is in fact that the signal to noise ratio is confounded with the process smoothness. This leads to the slow mixing for posterior sampling.

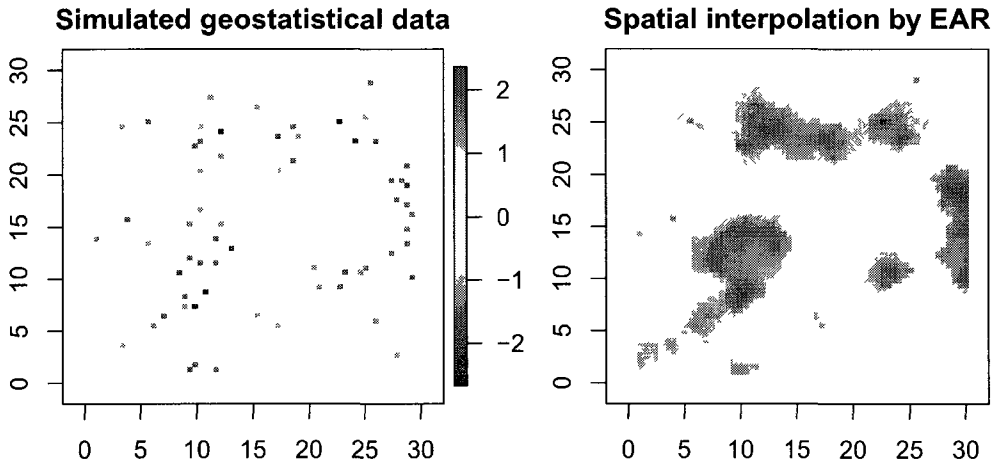


Figure 19: *An illustration of the EAR latent process to geostatistical data. Left: simulated observations from Matérn process with* $\sigma^2 = 1$, $\nu = 2$ *(smoothness) and* $\rho = 3$ *(range) added Gaussian noise with mean 0 and variance 0.5. Right: Spatial interpolation on a* 60×60 *grid*

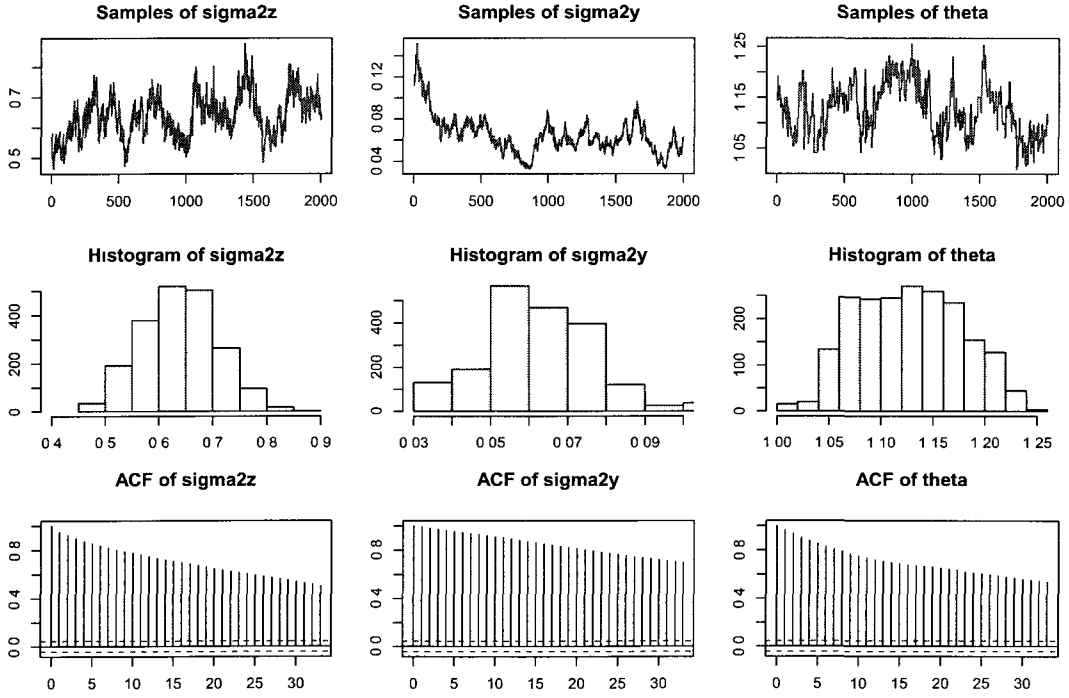


Figure 20: *Posterior samples for the data variance parameter* σ_y^2 *(sigma2y), latent EAR process variance parameter* σ_z^2 *(sigma2z) and its smoothness parameter* θ *(theta)*

CHAPTER VII

Spatio-Temporal Model

In this chapter, we formalize the EAR model representation for spatio-temporal data. First we give a very brief review of autoregressive time series models, and then provide calculations for separable space-time processes as well as for spatially varying parameter models. We give details of the steps required for estimation.

In addition to spatial-only models, it is often of great interest to incorporate temporal trends into spatial models, hence, the spatio-temporal model. This type of model arises when repeated measurements are collected over time as well as across space. For instance, total yearly precipitation observed at various weather stations over the African Sahel from year 1982 to 1996 (Lindström J., and Lindgren, F., 2008). In this case, the data analysis has to take account of spatial dependence among the stations, but also that the observations at each station typically are not independent but form a time series. In other words, one must take account of temporal correlations as well as spatial correlations. Therefore, the linear trend parameters, the spatial interaction parameter and the temporal interaction parameter, are all incorporated in the distribution of the data to represent linear trend, spatial interaction, and temporal autoregressive behavior, respectively.

To model data collected over both space and time, the computer efficient CAR model of Pettitt, Weir, and Hart (2002) as well as Czado and Prokopenko (2008) described for analyzing spatial data can be modified to incorporate not only spatial interaction, but temporal dependencies as well, as in the space-time hierarchical model of Wikle, Berliner, and Cressie (1998). In this model, it is assumed that the data at each location come from a normal distribution with errors that could contain spatial or temporal correlations. If the data do not come from a normal distribution, an appropriate transformation can be made so that the transformed data is Gaussian. The purpose here is to estimate simultaneously linear regression trend as well as spatial and temporal structure in the residuals.

7.1 Autoregression in Time Series

Consider a time series data $\{x_t, t = 1, 2, \cdots, T\}$ collected over time. Without loss of generality, we assume $E(x_t) = \mu$ is a constant. One approach in modeling this type of data is to use an autoregressive model of order p, or an $AR(p)$ model. That is,

$$x_t - \mu = a_1(x_{t-1} - \mu) + a_2(x_{t-2} - \mu) + \cdots + a_p(x_{t-p} - \mu) + \epsilon_t \,,$$

where a_j, $j = 1, \cdots, p$ are the autocorrelation parameters and are related to so-called partial autocorrelations:

$$a_j = \text{corr}(x_t, x_{t+j}|\text{others}) \,.$$

The ranges of a_j are restricted so that roots of the associated polynomial lie outside the unit circle. Typically the ranges are contained in $[-1, 1]$ (see Shumway and

Stoffer, Chapter 3, 2006). ϵ_t is assumed to be the Gaussian error term, that is, $\epsilon_t \sim N(0, \sigma^2)$.

In spatial-temporal models, since in most cases, the process is only related to what happened in the previous time, $AR(1)$ structure between consecutive times for the field is typically assumed, which is

$$x_t - \mu = a(x_{t-1} - \mu) + \epsilon_t \,,$$

or it can be written as

$$(1 - aB)(x_t - \mu) = \epsilon_t \,,$$

where B is the backshift operator such that $Bx_t = x_{t-1}$. Then, x_t can be solven as

$$x_t = \mu + \epsilon_t + a\epsilon_{t-1} + a^2\epsilon_{t-2} + \cdots \,.$$

Therefore,

$$E(x_t) = \mu$$

and

$$\mathrm{Var}(x_t) = \sigma^2(1 + a^2 + a^4 + \cdots) = \frac{\sigma^2}{1 - a^2}$$

and

$$\mathrm{Cov}(x_t, x_{t-j}) = \sigma^2 \frac{a^j}{1 - a^2}, \quad j = 1, \cdots, (T-1) \,.$$

Let $\mathbf{X} = (x_1, x_2, \cdots, x_T)'$ be the $T \times 1$ vector of all the data, the joint distribution can be written as

$$\mathbf{X} \sim N(\mathbf{0}, \sigma^2 \mathbf{Q}_T^{-1}),$$

where the variance-covariance matrix $\mathbf{Q}_T^{-1}$ is

$$\mathbf{Q}_T^{-1} = \frac{1}{1-a^2} \begin{pmatrix} 1 & a & a^2 & a^3 & \cdots & a^{T-1} \\ a & 1 & a & a^2 & \cdots & a^{T-2} \\ a^2 & a & 1 & a & \cdots & a^{T-3} \\ \vdots & \vdots & \vdots & \vdots & \ddots & a \\ a^{T-1} & a^{T-2} & \cdots & a^2 & a & 1 \end{pmatrix}.$$

Thus, the precision matrix $\mathbf{Q}_T$ is a tri-diagonal matrix,

$$\mathbf{Q}_T = \begin{pmatrix} 1 & -a & 0 & & & \\ -a & 1+a^2 & -a & 0 & & \\ & \ddots & \ddots & \ddots & & \\ & 0 & -a & 1+a^2 & -a & 0 \\ & & & \ddots & \ddots & \ddots \\ & & & 0 & -a & 1+a^2 & -a \\ & & & & 0 & -a & 1 \end{pmatrix}.$$

From the appendix of Lindström J., and Lindgren, F. (2008), we know that the determinant $|\mathbf{Q}_T| = 1 - a^2$. Refer to Shumway and Stoffer (2006) for more about time series analysis.

7.2 Separable Spatio-temporal Model

Recall that the theory of Gaussian Markov random fields is used in the spatial model setting. Letting $\mathbf{z}_t, (t = 1, \cdots, T)$ denote the n-by-1 column vector representing the GMRF at each time point, then the spatio-temporal field can be represented as

$$\mathbf{z} = [\mathbf{z}_1', \cdots, \mathbf{z}_T']' .$$

From the section 7.1, assuming an $AR(1)$ structure between consecutive times and a mean field, $\mu(\mathbf{s})$, that is constant in time, the field $\mathbf{z}$, can be modeled as

$$(\mathbf{z}_t - \mu) = a(\mathbf{z}_{t-1} - \mu) + \epsilon_t ,$$

where ϵ_t are independent in time but spatially correlated,

$$\epsilon_t \sim N(\mathbf{0}, \mathbf{Q}_S^{-1}) ,$$

where, $\mathbf{Q}_S$ denotes the precision matrix for the spatial dependence. The term separability refers to the covariance matrix. Naturally it implies that both spatial dependence does not change over time, and temporal dependence does not change in space. Further we take $\mathbf{y}_1 \sim N(\mu, \frac{1}{1-a^2}\mathbf{Q}_S^{-1})$, and thus can get the distribution for the spatio-temporal GMRF,

$$\mathbf{y} \sim N\left(\mathbf{1} \otimes \mu, (\mathbf{Q}_T \otimes \mathbf{Q}_S)^{-1}\right) ,$$

where, $\mathbf{Q}_T$ is give in section 7.1, $\mathbf{1}$ denotes unity (column) vectors and $\otimes$ is the Kronecker product. Note that the separable space-time covariance matrix $(\mathbf{Q}_T \otimes \mathbf{Q}_S)^{-1}$ is convenient in terms of computational efficiency. $\mathbf{Q}_S$ can be as defined for the computer efficient spatial CAR model, or be simply a diagonal matrix indicating the lack of spatial structure in the residuals. $\mathbf{Q}_T$ here has been set to $AR(1)$ model for computational efficiency, or be a diagonal matrix indicating no temporal structure in the residuals. In addition, $|\mathbf{Q}_T \otimes \mathbf{Q}_S| = |\mathbf{Q}_T|^n |\mathbf{Q}_S|^T = (1-a^2)^n |\mathbf{Q}_S|^T$, since $|\mathbf{Q}_T| = 1-a^2$.

In a hierarchical modeling, Gaussian observations are assumed as noisy versions of an underlying latent GMRF. Stacking all the observations in a nT-by-1 vector $\mathbf{y}$, it now can be written as a sum of the unknown GMRF $\mathbf{z}$ with additive, independent Gaussian errors $\epsilon \sim N(\mathbf{0}, \sigma_y^2 \mathbf{I})$,

$$\mathbf{y} = \mathbf{z} + \boldsymbol{\epsilon} \, .$$

The distribution of the data given the random fields is,

$$\mathbf{y}|\mathbf{z} \sim N(\mathbf{z}, \sigma_y^2 \mathbf{I}_{nT}) \, .$$

A general class prior distribution of the underlying GMRF $\mathbf{z}$ is

$$\mathbf{z}|\sigma_z^2, \psi, a \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma_z^2 (\mathbf{Q}_T \otimes \mathbf{Q}_S)^{-1}\right) \, ,$$

where, $\mathbf{X}$ is a known matrix of regression basis vectors and $\boldsymbol{\beta}$ contains the unknown regression parameters. ψ is a spatial interaction parameter, and a is a temporal correlation parameter.

If we assume an EAR model structure for $\mathbf{Q}_S$, then an additional smoothness parameter θ will be incorporated,

$$\mathbf{z}|\sigma_z^2, \psi, \theta, a \sim N\left(\mathbf{X}\boldsymbol{\beta}, \sigma_z^2(\mathbf{Q}_T \otimes \mathbf{Q}_S)^{-1}\right) .$$

To complete the hierarchical model, the prior distributions need to be specified for all parameters $\sigma_y^2, \sigma_z^2, \psi, \theta$ and a. As in the spatial hierarchical model, both variance parameters are conventionally assumed to have an inverse Gamma distribution. So, the prior distribution for the measurement error is

$$\sigma_y^2 \sim InvGamma(\alpha_y, \beta_y) ,$$

and the prior distribution for the error in the process is

$$\sigma_z^2 \sim InvGamma(\alpha_z, \beta_z) .$$

As in the spatial hierarchical model, the inverse gamma parameters are chosen constants.

Recall that a is the $AR(1)$ parameter representing temporal structure. It is assumed that $-1 \leq a \leq 1$ in temporal autoregressive processes. Since it is unknown which values of a are more likely, the prior distribution for a is assumed to be uniform$(-1, 1)$. As in the spatial hierarchical setup, ψ is the spatial interaction parameter and assumed to be $0 < \psi < 1$, a uniform$(0, 1)$ prior or beta distribution

can be assigned, that is,

$$\psi \sim Beta(\alpha_\psi, \beta_\psi) ,$$

where α_ψ and β_ψ are chosen constants. The distribution of $\theta > 0$ is chosen to be lognormal, that is

$$\nu = \log(\theta) \sim N(\mu_\theta, \sigma_\theta^2) ,$$

where μ_θ and σ_θ^2 are chosen constants. The priors for regression parameters $\boldsymbol{\beta}$ are conventionally normal distributions,

$$\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_\beta) .$$

In many cases, the mean vector $\boldsymbol{\beta}_0$ is chosen to be the zero vector.

7.3 Spatio-temporal Model with Spatially Varying Parameters

As discussed in section 7.2, the EAR model can be extended in a straightforward manner for spatio-temporal data. Of one particular interest are the spatio-temporal models with several parameters that are spatially varying. Typical choices for spatially varying parameters are the mean and the temporal trend. That is, for different spatial locations, temporal trends have various intercepts and slopes. In the general formulation we assume q spatially varying parameters. Now the data model is,

$$\mathbf{y}|\mathbf{z}, \sigma_y^2 \sim N(\mathbf{z}, \sigma_y^2 \mathbf{I}_{nT}) ,$$

and the underlying process $\mathbf{z}$ is as follows,

$$\mathbf{z}|\boldsymbol{\beta}, \sigma_z^2, \psi, \theta, a \sim N\left(\mathbf{X}\boldsymbol{\beta} + \sum_{k=1}^{q} \mathbf{U}_k \boldsymbol{\alpha}_k, \sigma_z^2(\mathbf{Q}_T \otimes \mathbf{I}_n)^{-1}\right).$$

We denote the spatially varying parameters by $\boldsymbol{\alpha}_k$ and assume they have an EAR process prior, and thus

$$\boldsymbol{\alpha}_k \sim N(\boldsymbol{\mu}_{\alpha,k}, \sigma_{\alpha,k}^2 \mathbf{Q}_k^{-1}),$$

where

$$\mathbf{Q}_k = [\mathbf{I} - \psi_k(\gamma - \mathbf{D})]^{\theta_k}.$$

As in the computer efficient CAR model, γ is the neighbor weight matrix and $\mathbf{D}$ is the diagonal matrix containing the row sums of γ minus 1.

Recall that the spatial structure can be removed using a "pre-whitening" method involving the singular value decomposition (SVD) of the $\gamma - \mathbf{D}$ within the precision matrix. If $\mathbf{F}$ is the matrix containing the eigenvectors of $\gamma - \mathbf{D}$ and Λ is the diagonal matrix containing the eigenvalues of $\gamma - \mathbf{D}$, then the spatial precision matrix $\mathbf{Q}_\alpha$ can be rewritten as

$$\mathbf{Q}_\alpha = \mathbf{F}\text{diag}(1 - \psi\lambda_i)^\theta \mathbf{F}',$$

and the process variance-covariance structure is

$$\sigma_\alpha^2 \mathbf{Q}_\alpha^{-1} = \sigma_\alpha^2 \mathbf{F}\text{diag}\left(\frac{1}{1 - \psi\lambda_i}\right)^\theta \mathbf{F}'.$$

To apply the transformation to the vector of observations, $\mathbf{y}$,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_T \end{pmatrix},$$

we apply the trasnformation to each component of the data vector, and obtain

$$\mathbf{y}^* = \begin{pmatrix} \mathbf{F}'\mathbf{y}_1 \\ \mathbf{F}'\mathbf{y}_2 \\ \vdots \\ \mathbf{F}'\mathbf{y}_T \end{pmatrix} = (\mathbf{I}_T \otimes \mathbf{F}')\mathbf{y} .$$

Similarly, we obtain,

$$\mathbf{z}^* = \begin{pmatrix} \mathbf{F}'\mathbf{z}_1 \\ \mathbf{F}'\mathbf{z}_2 \\ \vdots \\ \mathbf{F}'\mathbf{z}_T \end{pmatrix} = (\mathbf{I}_T \otimes \mathbf{F}')\mathbf{z} ,$$

and also apply the transformation to the fixed effects $\mathbf{X}$,

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{F}'\mathbf{X}_1 \\ \mathbf{F}'\mathbf{X}_2 \\ \vdots \\ \mathbf{F}'\mathbf{X}_T \end{pmatrix} = (\mathbf{I}_T \otimes \mathbf{F}')\mathbf{X} .$$

By applying the transformation to the space-varying random effects and their corresponding regressor matrices, the result is

$$\sum_{k=1}^{q} (\mathbf{I}_T \otimes \mathbf{F}')\mathbf{U}_k \boldsymbol{\alpha}_k \, .$$

The constant matrices $\mathbf{U}_k$ take on special forms for spatially varying mean and spatially varying temporal trend that are particularly easy to work with. For the mean, $\mathbf{U}_k = (\mathbf{1}_T \otimes \mathbf{I}_n)$, and its pre-whitening is as follows:

$$(\mathbf{I}_T \otimes \mathbf{F}')\mathbf{U}_k \boldsymbol{\alpha}_k = (\mathbf{I}_T \otimes \mathbf{F}')(\mathbf{1}_T \otimes \mathbf{I}_n)\boldsymbol{\alpha}_k = (\mathbf{1}_T \otimes \mathbf{F}')\boldsymbol{\alpha}_k = (\mathbf{1}_T \otimes \mathbf{I}_n)\mathbf{F}'\boldsymbol{\alpha}_k = \mathbf{U}_k\mathbf{F}'\boldsymbol{\alpha}_k \, .$$

For the temporal trend, $\mathbf{U}_k = (\mathbf{v}_{time} \otimes \mathbf{I}_n)$, and its pre-whiteing is,

$$(\mathbf{I}_T \otimes \mathbf{F}')\mathbf{U}_k \boldsymbol{\alpha}_k = (\mathbf{I}_T \otimes \mathbf{F}')(\mathbf{v}_{time} \otimes \mathbf{I}_n)\boldsymbol{\alpha}_k = (\mathbf{v}_{time} \otimes \mathbf{F}')\boldsymbol{\alpha}_k = (\mathbf{v}_{time} \otimes \mathbf{I}_n)\mathbf{F}'\boldsymbol{\alpha}_k = \mathbf{U}_k\mathbf{F}'\boldsymbol{\alpha}_k \, .$$

Thus, for the random effects, the regressors remain unchanged while the effects are transformed as

$$\begin{aligned} \boldsymbol{\alpha}_k^* &= \mathbf{F}'\boldsymbol{\alpha}_k \\ &\sim N\left(\mathbf{F}'\boldsymbol{\mu}_{\alpha,k}, \sigma_{\alpha,k}^2 \mathrm{diag}\left(\frac{1}{1-\psi_k\lambda_i}\right)^{\theta_k}\right) . \end{aligned}$$

Applying the transformation to the process precision matrix $\mathbf{Q}_T \otimes \mathbf{Q}_S$ results in

$$\begin{aligned}(\mathbf{I}_T \otimes \mathbf{F}')(\mathbf{Q}_T \otimes \mathbf{Q}_S)(\mathbf{I}_T \otimes \mathbf{F}) &= (\mathbf{I}_T\mathbf{Q}_T \otimes \mathbf{F}'\mathbf{Q}_S)(\mathbf{I}_T \otimes \mathbf{F}) \\ &= \mathbf{I}_T\mathbf{Q}_T\mathbf{I}_T \otimes \mathbf{F}'\mathbf{Q}_S\mathbf{F} \\ &= \mathbf{Q}_T \otimes \mathbf{F}'\mathbf{Q}_S\mathbf{F} .\end{aligned}$$

In the case of $\mathbf{Q}_S = \mathbf{I}_n$, the transformed space-time precision matrix becomes

$$\mathbf{Q}_T \otimes \mathbf{I}_n .$$

Spatio-temporal full conditional distributions with spatially varying EAR processes can then be specified given priors as discussed in section 7.2.

CHAPTER VIII

Conclusions and Future Work

8.1 Conclusions

In this dissertation, several approximation methods to "the big n problem" are reviewed, and methods for improving computational efficiency in estimating spatial parameters of a large dataset are proposed. In particular, the Pettitt *et al.* as well as Czado and Prokopenko parameterizations for the CAR model are discussed. Both parameterizations result in a sparse symmetric neighbor weight matrix that is relatively easy to work with, but still uses a considerable amount of computation time when working with very large data. To complement the computationally advantageous parameterization, a structure removing orthonormal transformation named "pre-whitening" is described. This transformation is based on a singular value decomposition and results in the removal of spatial structure from the data. Iterative computations can then be performed much faster in transformed space.

The circulant embedding technique is also discussed as a method to decrease computation time for very large data sets. Here, a smaller regular lattice structure is embedded within a larger rectangular grid and wrapped around onto a torus. On the torus, each location has exactly the same number of neighbors. This results

in a block circulant neighbor-weight matrix for which the calculation of eigenvalues and eigenvectors for the "pre-whitening" procedure are much faster, of order $O(n\log(n))$ as opposed to the typical $O(n^3)$.

The EAR model is proposed as a parsimonious extension to the autoregressive model that accounts for smoothness of a spatial process. In particular, it is an extension of the Czado and Prokopenko parameterization of the CAR model on a regular lattice or a torus, when the smoothness parameter θ takes integer values, the EAR model is shown to be equivalent to higher order CAR models when uniform weights are used; while when θ is not an integer, it can be treated as an exponential and logarithm of the weighted combination of infinite higher order neighbor matrices. However, as the neighbor distance d increases, the corresponding contribution of those neighbors with distance d to the conditional mean approaches zero at a fast rate. Thus, to model extremely smooth processes in space, use of the EAR model provides a more efficient analysis and accurate parameer estimation since it reduces the number of parameters needed in the model.

The EAR model structure is shown to have connections with the Matérn class in geostatistics. The smoothness parameter θ in the EAR and ν in the Matérn correlation function behave similarly, while they have theoretically relationship of $\theta = \nu+1$ on a torus. A simulation study shows a deviation of this lienar relationship between θ and ν towards an exponential relationship for large θ, possibly due to the edge effect in finite lattices. The study also shows a quadratic pattern between θ and ρ, where ρ is the range parameter in the Matérn class. Our model is also connected to INLA, which uses a finite method for solving SPDE. If wrapped onto

a torus, both models result in identical patterns of precision matrices, that only differ in their parameterizations.

In addition to applying the EAR model for lattice data, a latent GMRF with EAR model prior is proposed to model geostatistical data. A latent fine grid is created to ensure no more than one observation per grid cell. A missing data scheme by introducing latent pseudo-observations for all grid cells without any associated data is also used. This thus enables the posterior precision matrix to be diagonal, which does not require time-consuming inversion and determinant calculation. Parameter estimation and spatial kriging can be done simultaneously under MCMC iterations. An intrinsic EAR model is also proposed due to the identifiability issue between the smoothness parameter and the spatial interaction parameter.

Finally, the EAR model is used as the prior for spatio-temporal models. Of particular interest is the non-separable spatio-temporal model with spatially varying parameters.

8.2 Future Work

Two major areas of future work may involve: (1) the weighting scheme of the EAR model and (2) the application of the EAR model in geostatistics. First, in the EAR model in my dissertation, only uniform weights are assumed and used. Pettitt *at el.* (2002) proposed two other weighting functions: reciprocal and linear. The difficulty in using the linear or reciprocal weighting scheme is to determine the best cut-off distance $r_{\max}$ beyond which the interaction between two locations is zero.

The incorporation of the smoothness parameter θ in the EAR model exacerbates this problem, since the smoothness parameter and the $r_{\max}$ are confounded, that is, the larger the $r_{\max}$, the larger the θ. Second, when applying the EAR model in geostatistics, we might consider using other association matrices $\mathbf{K}$ that connect observations with a latent process. In this dissertation, we use an incidence matrix $\mathbf{K}$ to associate an observation to its closest grid point. We may define other $\mathbf{K}$ such that values at the observations points are some linear interpolation of the values at nearby latent grids. However, computation efficiency still needs to be achieved.

REFERENCES

[1] Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data.* New York: Chapman & Hall/CRC Press.

[2] Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B,* **36**, 192-236.

[3] Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika,* **82**, 4, 733-46.

[4] Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics,* **43**, 1-59.

[5] Box, George E.P. and Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B,* **26(2)**, 211-252.

[6] Brockwell, P.J. and Davis, R.A. (2003). *Introduction to Time Series and Forecasting: Second Edition.* New York: Springer Verlag.

[7] Cliff, A.D. and Ord, J.K. (1981). *Spatial Pocesses: Models and Applications.* London: Pion Limited.

[8] Cressie, N. (1991). *Statistics for Spatial Data.* New York: Wiley.

[9] Czado C. and Prokopenko S. (2008). Modelling transport mode decisions using hierarchical logistic regression models with spatial and cluster effects. *Statistical Modelling,* **8**, 4, 315-345.

[10] Diggle, P.J., Tawn, J.A., and Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics,* **47**, 299-350.

[11] Furrer, R., Genton, M. G. and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics,* **15(3)**, 502-523.

[12] Gaudard, M., Karson, M., Linder, E. and Sinha, D. (1999). Bayesian spatial prediction. *Environmental and Ecological Statistics,* **6**, 147-182.

[13] Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2003). *Bayesian Data Analysis: Second Edition* New York: Chapman & Hall / CRC Press.

[14] Graybill, F.A. (1983). *Matrices with Applications in Statistics.* Belmont, CA: Wadsworth Publishing.

[15] Griffith, D. A., Layne, L.J., and Doyle, P.G. (1996). Further explorations of relationships between semi-variogram and spatial autoregressive models. General Technical Report RM-GTR-277, United States Department of Agriculture.

[16] Handcock, M.S. and Wallis, J. R. (1994). An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, **89**, 426, 368-378.

[17] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57** (1): 97-109.

[18] Higdon D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, **5**, 173-190.

[19] Higdon D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, eds. C. Anderson, V. Barnett, P.C. Chatwin, and A.H. El-Shaarawi. London: Springer-Verlag, pp. 37-56.

[20] Higdon D., Lee, H. and Holloman, C. (2003). Markov chain Monte Carlo-based approaches for inference in computationally intensive inverse problems (with discussion). *Bayesian Statistics 7. Proceedings of the Seventh Valencia International Meeting*, 181-197.

[21] Hjort, N.L. and Omre, H. (1994). Topics in spatial statistics. *Scandinavian Journal of Statistics*, **21**, 289-357.

[22] Horn, R. A. and Johnson, C. R. (1994). *Topics in Matrix Analysis*. Cambridge University Press, Cambridge.

[23] Hrafnkelsson, B. and Cressie, N. (2003). Hierarchical modeling of count data with application to nuclear fall-out. *Environmental Ecological Statistics*, **10**, 197-200.

[24] Hupper, V.P. (2005). Contributions to modeling and computer efficient estimation for Gaussian space-time processes. PhD Dissertation.

[25] Kaufman, C., Schervish, M., and Nychka, D. (2008). Covariance tapering for likelihood based estimation in large spatial datasets. *Journal of the American Statistical Association*, **103**, 1545-1555.

[26] Krige, D.G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 6, 119C139.

[27] Lawson A. (2008). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*. New York: Chapman & Hall / CRC Press.

[28] Lemos, R.T., Sansó, B. (2009). A spatio-temporal model for mean, anomaly and trend fields of North Atlantic sea surface temperature (with discussion). *Journal of the American Statistical Association*, **104**, 5-18.

[29] Linder, E. (2001). Computer-efficient spatial estimation and interpolation based on conditional Gaussian autoregressive models. Proceedings: Joint Statistical Meetings. 2001. American Statistical Association. Alexandria, VA.

[30] Lindgren, F., Lindström, J., and Rue, H. (2010). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach. Technical report.

[31] Lindström J., and Lindgren, F. (2008). A Gaussian Markov random field model for total yearly precipitation over the Afircan Sahel. *Preprints in Mathematical Sciences*, 2008:8.

[32] Lunetta S.R., Lyon, G.J. (2004). *Remote Sensing and GIS Accuracy Assessment.* New York: Chapman & Hall / CRC Press.

[33] Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, **58**, 1246-1266.

[34] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1092.

[35] Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, **70**, 120-126.

[36] Pace, R.K. and Barry, R. (1996). Sparse spatial autoregressions. *Statistics and Probability Letters*, 2158.

[37] Paciorek, J.C. (2007a). Computational techniques for spatial logistic regression with large data set. *Computational Statistics and Data Analysis*, **51**, 3631-3653.

[38] Paciorek, J.C. (2007b). Bayesian smoothing with Gaussian processes using Fourier basis functions in the spectralGP package. *Journal of Statistical Software*, volumn 19, issue 2.

[39] Pettitt, A.N., Weir, I.S., and Hart, A.G. (2002). A conditional autoregressive Gaussian process for irregularly spaced multivariate data with application to modeling large sets of binary Data. *Statistics and Computing*, **12**, 353-367.

[40] Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society*, **63**, part 2, 325-338.

[41] Rue, H. (2009). Spatial modeling and inference using SPDEs . SAMSI Opening Workshop, September 13-16.

[42] Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications.* New York: Chapman & Hall / CRC Press.

[43] Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 319-392.

[44] Rue, H. and Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, **29**, 31 - 49.

[45] Sain, S.R., Furrer, R., Cressie, N. (2007). Combining regional climate model output via a multivariate Markov random field model. In: 56th Session of the International Statistical Institute, Lisbon, Portugal.

[46] Schabenberger, O. and Gotway C.A. (2005). *Statistical Methods for Spatial Data Analysis.* Chapman & Hall/CRC.

[47] Shumway, R. and Stoffer, D. (2006). *Time Series Analysis and Its Applications with R Examples.* New York: Springer Verlag.

[48] Song H.R., Fuentes,M., and Ghosh S. (2008). A comparative study of Gaussian geostatistical models and Gaussian Markov random field models. *Journal of Multivariate Analysis*, **99**, 1681-1697.

[49] Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer-Verlag, New York.

[50] Wall, M.M. (2004). A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, **121**, 311-324.

[51] Wendland, H. (1998). Error estimates for interpolation by compactly supported radial basis functions of minimal degree. *Advances in Computational Mathematics*, **93**, 258-272.

[52] Wikle, C.K., (2002). Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains. In *Spatial Cluster Modelling*, A. Lawson and D. Denison, eds. Chapman and Hall, 199-209.

[53] Wikle, C.K., Berliner, L.M., and Cressie, N. (1998). Hierarchical Bayesian space-time models. *Environmental and Ecological Statistics*, **5**, 117-154.

[54] Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, **41**, 434-449.

[55] Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250-261.