Joint Composite Estimating Functions in Spatial and Spatio-Temporal Models

by

Yun Bai

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2011

Doctoral Committee:

Professor Peter Xuekun Song, Chair
Professor Ana V. Diez Roux
Professor Trivellore E. Raghunathan
Assistant Professor Brisa N. Sanchez

To Guangwei, Yubai and my parents

ACKNOWLEDGMENTS

Dissertation research is an intellectual pursuit that builds upon the knowledge and help of many individuals. I would like to express my deepest gratitude to my advisor Dr. Peter X.-K. Song, for his continued support, heart-felt encouragement and enlightening advice throughout my graduate studies. He has always steered me towards better approaches for solving problems and has guided me through many difficult moments in my research. His ideas have helped shape the main framework of my dissertation. I feel very fortunate to have such an approachable and knowledgeable advisor.

I am grateful for Dr. T.E. Raghunanthan who first sparked my interest in spatio-temporal analysis. His insightful comments and suggestion have led me deeper in the methodology investigation.

My sincere thanks also go to Dr. Ana Diez Roux for her generous support of my PhD training. She offered me excellent opportunities to apply statistics to solve real-world problems, which strengthened the scientific backgrounds of my dissertation research.

I am thankful for Dr. Brisa Sanchez for the enlightening discussions I had with her, which have lead to better contents of this dissertation.

Last but not the least, I would like to thank my family for their support and love, which set me worry free in my research endeavor.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

<u>Table</u>

LIST OF APPENDICES

Appendix

ABSTRACT

Joint Composite Estimating Functions in Spatial and Spatio-Temporal Models

by

Yun Bai

Chair: Peter Xuekun Song

Spatial or spatio-temporal data are frequently encountered in many scientific disciplines. One major challenge in modeling these processes is the high dimensionality of such data; that is, the number of observations is usually enormous.

The first part of the dissertation proposes an efficient approach to analyzing spatio-temporal processes. We proposed a new method called joint composite estimating function (JCEF). It reduces the likelihood dimension by utilizing lower-dimensional marginal likelihoods in estimation and inference. This method allows us to account for high-order spatio-temporal dependences through Hansen's generalized method of moments. Simulation experiments show favorable improvement in estimation efficiency over the conventional composite likelihood methods when applied to estimating the spatio-temporal covariance functions. Large sample properties of the proposed JCEF estimator are derived under more realistic settings than what is available in the current literature.

The second part of the thesis presents a much needed review of existing covariance estimation methods parallelly developed for massive spatial data sets. To thoroughly investigate their relative performances in spatio-temporal data analysis, we conduct extensive simulation experiments to compare estimation bias and efficiency among the most popularly used methods, including conventional pairwise composite likelihood, JCEF, Stein's conditional pseudo-likelihood, tapering, weighted least squares, and maximum likelihood, which is served as the golden standard.

The third part of the thesis develops a new modeling and estimating framework for high-dimensional spatial-clustered data, termed as GeoCopula. Marginal distributions are assumed to be the generalized linear models, so that the new method can handle both discrete and continuous outcomes. The within-cluster and between-cluster spatial correlations are modeled by a multivariate Gaussian copula, which results in a fully parametric model for dependent data. This class of models generates population-level regression parameter estimates similar to GEE, while explicitly models the dependence structures separately from the mean model. Estimation and inference are achieved by applying the JCEF method. Through simulation experiments we show efficiency improvement over conventional pairwise composite likelihoods. The proposed model and method are illustrated by an analysis of the Gambia malaria data set.

CHAPTER I

Introduction

Spatial and Spatio-temporal data are frequently encountered in many scientific disciplines, such as environmental sciences, e.g. daily air pollutant records across the country (Paciorek et al., 2009); economics, e.g. real estate transactions over space and time (Gelfand et al., 2003); epidemiology, e.g. infectious disease outbreaks in space and time (Lawson and Song, 2010), among others. Through data analysis, scientists are interested in identifying important factors that associate with the underlying outcome processes and in predicting such processes at locations and time points at which observations are not available.

One of the major challenges in spatio-temporal statistical analysis is the high dimensionality of the data; that is, the number of observations from spatio-temporal processes is usually enormous. For example, the well-known Irish wind speed data (Haslett and Raftery, 1989) includes 12 synoptic meteorological stations during the period of 1961-1978. For every day of the year, daily means of the wind speed of every station are available, leading to an approximately 80,000 observations in total. Studying the dependence structures of these processes poses great computational difficulties in spite of the fast-growing computing capacities. One problem that motivated this dissertation research was the need to model 20-year monthly concentration records of airborne particulate matter with diameter less than 10 microns, or PM10, from 2474 monitors across the United States (Diez Roux et al., 2008). The total number of observations from this PM10 spatio-temporal process is more than 860 thousands, which makes the joint modeling of the spatio-temporal process computationally prohibitive.

On the other hand, in social and health sciences, research studies usually involve subjects that are randomly selected within a large number of clustered geographical units. For example, among the studies of place effects on health, Chaix et al. (2005) investigated individual and contextual factors that determine the health care utilization in France, where 10955 people are randomly surveyed within 4421 municipals in France. To study the association of neighborhood environmental risk factors with cardiovascular diseases, Mujahid et al. (2007) used a sample of 5988 subjects selected from 576 census tracts from three states in USA. Grady (2010) assessed the impact of racial residential segregation on low birth weight from a pool of 10277 cases nested in 1092 census tracts in Michigan. In civil and environmental engineering studies, Sener et al. (2011) analyzed the physical activity participation levels of individuals in a family unit based on data drawn from the 2000 San Francisco Bay Area Household Travel Survey, in which individual and household socio-demographic as well as all activity and travel episodes information were recorded for subjects in 15000 households.

These spatial data examples are just a glimpse of a growing number of research projects that collects data in spatial dimensions, thus necessitate the eminent need to generalize the multilevel data analysis to incorporate the spatial dependences among the clustering units. In classic multilevel models, data from clusters are assumed to be independent, and the focus dwells on appropriately accounting for within-cluster correlations while making statistical inferences. However, when clusters are spatially correlated, such as neighborhoods or census tracts, subjects from clusters are likely to be correlated due to location proximity, hence the between-cluster independence assumption is no longer valid. Statistical analysis ignoring the spatial effect can lead to wrong standard errors of the regression coefficient estimates, which in turn biases hypothesis testing (Anselin and Griffith, 1988). As a result, in order to draw valid statistical inference, it is of critical importance to account for the between-cluster spatial correlation as well as the within-cluster correlation.

The first part of my dissertation develops an efficient yet computationally feasible approach for analyzing high-dimensional spatio-temporal processes. The proposed joint composite estimating function (JCEF) approach aims to reduce the likelihood dimension, to expedite computation, and to minimize the loss of estimation efficiency by utilizing merits of the composite likelihood method and generalized method of moments (GMM Hansen (1982)). The novelty of the proposed method lies in that it first decomposes the high-dimensional process into lower-dimensional components, and then accounts for correlations among the components by using a weight matrix similar to that is used in GMM. As a result, the JCEF approach is expected to significantly improve estimation efficiency over existing composite likelihood methods. Large-sample properties of the JCEF estimator have been derived under a more general setting than those already given in the literature. A comprehensive simulation experiment with varying spatio-temporal dependence structures is carried out to assess the small-sample properties of the proposed JCEF method in terms of covariance function estimation. Moreover, subsampling techniques are reviewed and applied to estimate the weight matrix and parameter standard errors. Effects of different subsample sizes on standard error estimation are also studied via simulations. The proposed method is then used to analyze the dependence structure of the PM10 data for Northeastern United States.

The second part of my thesis comprehensively reviews and compares the existing methods of estimating dependence structures for massive spatio-temporal data. This is a much needed research work in the field of the spatio-temporal data analysis. In spatial statistics, two types of approaches have been developed to facilitate computation. The first approach is based on simplifying covariance structures. Cressie and Jahannesson (2008) proposed fixed rank kriging for very large spatial data sets, where the covariance matrices were specially designed so that the matrix manipulations were of fixed magnitude. A similar idea was exploited in Banerjee et al. (2008). Another approach is based on likelihood approximations, where simplified versions of the full likelihood are considered. For example, Curriero and Lele (1999); Heagerty and Lele (1998); Li and Lin (2006) used pairwise marginal densities to build composite likelihood estimation functions. Also Vecchia (1988) and Stein et al. (2004) suggested approximating the likelihood by a product of conditional densities with truncated conditioning sets. Apart from composite likelihood approaches, Furrer et al. (2006) and Kaufman et al. (2008) used covariance tapering method to shrink small values of covariance entries to zero, so that the sparse matrix algorithm could be used to speed up computation. Fuentes (2007) proposed an approximation by modeling the covariance structures in the spectral domain, which appears to be more involved and hence is of less popularity in application. The performances of some of these estimation methods will be evaluated and compared through extensive simulation experiments. This aims at providing a set of much needed numerical evidences to guide the method selection in practice, which to our knowledge is not currently available.

The third part of the thesis develops a new model and extends the JCEF procedure to spatial-clustered data, where subjects are sampled within clusters which are correlated in space. A unified framework of the GeoCopula regression model is proposed to analyze such data. It not only provides population-level regression parameter estimates similar to GEE, but also models the within-cluster and between-cluster spatial correlation structures separately from the mean regression model. Estimation is carried out using the JCEF approach. Specifically, given this nested data structure, two types of pairs can be identified: pairs within clusters and pairs between clusters. The former group is more informative of the within-cluster variations and the latter is more relevant to between-cluster variations. Thus we can form two group-based composite estimating functions, and then apply the JCEF procedure to improve estimating efficiency. Asymptotic properties of this new JCEF estimators for spatial-clustered data are derived similarly as in the spatio-temporal settings, under the assumption that the spatial dependence decays at an appropriate rate with an increasing domain. In addition to extensive simulation experiments, an analysis of the malaria data set (Diggle et al., 2008) is presented as an illustration.

The dissertation is structured as follows. In Chapter 2, we present a review of composite likelihood methods. We provide all details concerning the development of the JCEF estimator in the spatio-temporal setting in Chapter 3. Extensive comparisons of our proposed method with currently popular approaches are made through simulation experiments in Chapter 4. Chapter 5 is devoted to the development of the GeoCopula model and an estimation approach for spatial-clustered data, followed by an outline of future work in Chapter 6. Appendix lists some technical details of the theoretical proofs.

CHAPTER II

Composite Likelihood Methods

Composite likelihood methodology refers to, in general, a type of pseudo likelihood method that utilizes marginal or conditional distributions of lower-dimensional components from a fully specified parametric distribution. This idea was first proposed by Besag (1974) to make statistical inference and estimation in spatial random fields, and later was termed as ”composite likelihood” by Lindsay (1988). It has received increasing popularity in estimation and inference for parametric and semiparametric models where full likelihood functions are numerically difficult to evaluate due to complex dependence structures of the data. In this chapter, I will begin with an introduction of the two main types of composite likelihood methods frequently used in the literature, and then review their applications to spatial or spatio-temporal models. Emphasis will be given to research work developed to improve the efficiency of composite likelihood estimation. More comprehensive reviews can be found in Varin (2008); Varin et al. (2011) and references therein.

2.1 Definition

Consider an m-dimensional random vector $Y = (y_1, \ldots, y_m)$ with a joint multivariate probability density function $f(Y|\boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is an r-dimension parameter vector of interest. Let $\{\mathcal{A}_1, \ldots, \mathcal{A}_K\}$ be a set of conditional or marginal events with associate likelihoods $L_k(\boldsymbol{\theta}; y) \propto f(y \in \mathcal{A}_k|\boldsymbol{\theta})$. The composite likelihood is constructed by treating these component likelihood functions as independent (Lindsay, 1988).

$$CL(\boldsymbol{\theta}, Y) = \prod_{k=1}^{K} L_k(\boldsymbol{\theta}; y)^{\omega_k},$$

where $\omega_k, k = 1, \ldots, K$ are weights assigned to each component in order to improve estimation efficiency. Usually ω_k is a 0-1 binary variable, indicating whether the k-th component is included in the estimation or not.

There are two main ways to construct composite likelihoods, depending on whether conditional or marginal events are used.

Composite conditional likelihood This type of composite likelihood method was first proposed by Besag (1974) in order to conveniently formulate models for spatial random processes. The idea is to specify the joint probability distribution by conditional probability functions,

$$L_C(\boldsymbol{\theta}; Y) = \prod_{i=1}^{m} f(y_i|y_j \in \partial y_i; \boldsymbol{\theta}),$$

where ∂y_i denotes the set of observations neighboring y_i. This kind of modeling approach is intuitively plausible for correlated data and has been further exploited by many researchers. For example, Vecchia (1988) and Stein et al. (2004) used blocks of observations for conditional events in spatial data analysis, and

Molenberghs and Verbeke (2005) in longitudinal studies, Liang (1987) in stratified case-control studies, and Mardia et al. (2008) in bioinformatics.

Composite marginal likelihood This class of composite likelihood is constructed by compounding lower-dimensional marginal densities. The simplest way is to form pseudo likelihoods under the working independence assumption, the so-called onewise composite likelihood:

$$L_{ind}(\boldsymbol{\theta}; Y) = \prod_{i=1}^{m} f(y_i|\boldsymbol{\theta}),$$

where correlations among the observations are ignored, and estimation involves only marginal parameters.

The most popular form in the current literature is the pairwise composite likelihood based on bivariate marginals:

$$L_P(\boldsymbol{\theta}; Y) = \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} f(y_i, y_j|\boldsymbol{\theta})^{\omega_{ij}},$$

It contains the minimal modeling blocks of marginal and dependence parameters, essential for correlated data analysis.

In some circumstances, one may also consider larger subsets such as triples or quadruples of observations (Varin and Vidoni, 2005). It is also possible to combine L_{ind} and L_P in some optimal way (Reid and Cox, 2004), or one can form a pseudo-likelihood by a mixture of conditional and marginal component density functions.

The method to be proposed in the dissertation is based on the pairwise composite likelihood method, motivated by its substantial computational advantages, its ability to incorporate parameters related to both the mean and the dependence structures and, above all, its relatively high estimation efficiency as demonstrated by many authors.

2.2 Pairwise Composite Likelihood

This section acquaints readers with the basic notations and quantities which are the building blocks for composite estimating equation methods.

Denote the log-likelihood function based on pairwise composite likelihood by l_P. Then

$$l_P(\boldsymbol{\theta}; Y) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \omega_{ij} \log f(y_i, y_j | \boldsymbol{\theta}).$$

The composite score function (CSF) is given by the first order derivative of l_P with respect to $\boldsymbol{\theta}$, specifically,

$$\Psi_P(\boldsymbol{\theta}; Y) = \nabla_{\boldsymbol{\theta}} l_P(\boldsymbol{\theta}; Y) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \omega_{ij} \nabla_{\boldsymbol{\theta}} \log f(y_i, y_j | \boldsymbol{\theta}),$$

which is a linear combination of component score functions, $\log f(y_i, y_j|\boldsymbol{\theta})$. Note that CSF Ψ_P is an unbiased estimating function as long as the two-dimensional marginals are correctly specified. According to Godambe and Heyde (1987), the condition of unbiasedness is essential for estimation consistency.

The maximum composite likelihood estimator $\hat{\boldsymbol{\theta}}_P$ is defined as a solution to the following CSF:

$$\Psi_P(\boldsymbol{\theta}; Y) = 0.$$

The standard theory of estimating equations (Song, 2007, Chap 3) suggests that, the information matrix of the composite likelihood is quantified by

$$G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J^{-1}(\boldsymbol{\theta})H(\boldsymbol{\theta}), \tag{2.1}$$

where $H(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left(-\nabla\Psi_P(\boldsymbol{\theta}; Y)\right)$ is the sensitivity matrix, and $J(\boldsymbol{\theta}) = Var_{\boldsymbol{\theta}}\left(\Psi_P(\boldsymbol{\theta}; Y)\right)$ is the variability matrix. $G(\boldsymbol{\theta})$ is known as the Godambe information or sandwich information matrix (Godambe, 1991). In the classic maximum likelihood setting where the full likelihood $f(y_1, \ldots, y_m)$ is used, the second Bartlett identity holds, i.e., $H(\boldsymbol{\theta}) = J^{-1}(\boldsymbol{\theta})$, and $G(\boldsymbol{\theta})$ becomes the Fisher information matrix. However, composite likelihood may be seen as a likelihood under a misspecified model, where high-order model structures (for example, 3-way dependence) are not assumed. As a result, the second Bartlett identity does not hold, $H(\boldsymbol{\theta}) \neq J^{-1}(\boldsymbol{\theta})$, leading to the loss of efficiency compared to the maximum likelihood estimation (Song, 2007, Chapter 3).

2.2.1 Asymptotic Framework

Given settings to which composite likelihoods are applied, two types of asymptotic scenarios are considered for the derivation of the large-sample properties of $\hat{\boldsymbol{\theta}}_P$. The first scenario corresponds to situations where the increase in sample size is achieved by an increase in the number of independent data replicates. For example, in the longitudinal analysis, subjects are assumed to be independent, and the sample size increase is a result of the inclusion of more subjects. Also for clustered data analysis, observations are usually correlated within clusters, but observations from different clusters are then assumed to be independent. Thus the sample size increases as more clusters are sampled. The other scenario corresponds mostly to spatial or spatio-temporal settings, where oftentimes only one realization of the underlying process is observed. This is analogous to time series data. The increase in sample size is achieved by expanding the process to more observations, and the newly-added data are likely to be correlated with existing observations. Under this sampling scenario, two types of asymptotic arguments have to be distinguished. One is the 'infill asymptotic' (Stein, 1995; Zhang, 2004), where more data are sampled within a fixed spatial domain, resulting in a denser sampling layout. The other is the so-called 'increasing-domain asymptotics'(Mardia and Marshall, 1984), where more data are included in the analysis as a result of expansion in the spatial domain. Refer to Chapter III for details.

I will outline the asymptotic properties of $\hat{\boldsymbol{\theta}}_P$ for the first scenario in the following section 2.2.2, where independent replicates are available, and leave the discussion of the asymptotic derivations in the second spatial/ spatio-temporal scenario to Chapter III.

2.2.2 Asymptotic Properties with Independent Replicates

Now label $Y^{(i)}$ as the realized values of Y for the i-th subject, $i = 1, \dots, n$. The data from subjects are assumed to be mutually independent. Then the pairwise composite score function is of the form:

$$\Psi_P(\boldsymbol{\theta}) = \sum_{k=1}^{n} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \omega_{ij} \nabla_{\boldsymbol{\theta}} \log f(y_i^{(k)}, y_j^{(k)} | \boldsymbol{\theta}).$$

The asymptotic results (e.g. consistency and asymptotic normality) for the composite likelihood estimator can be derived using the same arguments for the classical maximum likelihood estimation. Relevant regularity conditions can be adapted from Lindsay (1988) and Molenberghs and Verbeke (2005).

Thus, the asymptotic normality is postulated as:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_P - \boldsymbol{\theta}) \sim N_r(0, G^{-1}(\boldsymbol{\theta})),$$

where $G(\boldsymbol{\theta})$ is the Godambe information given in (2.1). As Zi (2009) points out that in general, by using Cramer-Rao inequality, the difference between $G^{-1}(\boldsymbol{\theta})$ and, $I^{-1}(\boldsymbol{\theta})$, the inverse of the Fisher information matrix, is positive semi-definite, which means that $\hat{\boldsymbol{\theta}}_P$ is usually more variable than its MLE counterpart. $G(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ if and only if $\Psi_P(\boldsymbol{\theta})$ is a linear function of the score function (Lindsay, 1988). The component-wise ratios of the main diagonal elements of $G(\boldsymbol{\theta})$ over $I(\boldsymbol{\theta})$ describe the asymptotic relative efficiency of the composite likelihood estimator and the MLE.

2.3 Efficiency Improvement

In exceptional cases, pairwise likelihood estimators can achieve full efficiency as the maximum likelihood estimators. For example, Mardia et al. (2007) showed that composite conditional estimators are identical to maximum likelihood estimators in the multivariate normal distribution with arbitrary means and unstructured covariances. Zi (2009) gave the same results based on pairwise likelihoods. Reid and Cox (2004) noted that the likelihood function derived from the quadratic exponential distribution for multivariate binary data (Cox, 1972) is equal to the pairwise likelihood function for binary data generated by a probit link. Pairwise likelihood estimators for two-way contingency table is also equally efficient to the maximum likelihood estimators (Mardia et al., 2009).

The popularity of composite likelihood methods has been boosted by these and other similar findings showing a considerable attainment in efficiency. However, for most other general models, composite likelihood methods are expected to be less efficient than the maximum likelihood methods, because models are misspecified when lower-dimensional components are assumed to be independent or high-order dependences are ignored.

Lindsay (1988) gave some theoretical treatment of this issue. He showed that, for scalar parameters, the Godambe information is proportional to the Fisher information by a factor of squared correlation between the CSF Ψ_P and the score function from the full likelihood Ψ_{MLE}:

$$G(\boldsymbol{\theta}) = I(\boldsymbol{\theta}) corr^2(\Psi_{MLE}(\boldsymbol{\theta}; Y), \Psi_P(\boldsymbol{\theta}; Y)),$$

Thus, the attainment of full information is associated with a linear relationship between Ψ_{MLE} and Ψ_P. The optimal weight w_{ij} of the component composite scores is possible for scaler parameters; however, for vector parameters, the optimal choice of weight is not globally attainable.

Efforts to improve efficiency of composite likelihood methods in the existing literature can be mainly categorized into four types. The first type of work tries to find an optimal combination of composite likelihoods based on different sizes of lower dimensional components, which have been mainly studied in the context of onewise composite likelihood L_{ind} and pairwise composite likelihood L_P. Reid and Cox (2004) showed that it is possible to improve efficiency by constructing estimating functions in an optimal combination of L_{ind} and L_P:

$$l(\boldsymbol{\theta}; Y) = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \omega_{ij} \nabla_{\boldsymbol{\theta}} \log f(y_i, y_j|\boldsymbol{\theta}) - a \sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} \log f(y_i|\boldsymbol{\theta}),$$

where a is a constant to be chosen as a solution subject to an optimality criterion. Zhao and Joe (2005) proposed using L_{ind} for marginal parameters and L_P for association parameters. Kuk (2007) suggested a hybrid method using optimal score functions for the marginal parameters and L_P for association parameters. This method is shown to be better than the alternating logistic regression (Carey et al., 2003).

The second type of approach to improve efficiency is developed in the setting of clustered data, where weights are given to cluster-specific component score functions based on cluster sizes. Joe and Lee (2009) investigate the choice of weights in this setting in detail. Specifically, let n_i denote the cluster size for cluster i, the weight $1/(n_i - 1)$ recommended by Kuk and Nott (2000) and Zhao and Joe (2005) work well when the dependence within cluster is weak. For stronger within-cluster dependence, weight $1/(n_i(n_i - 1))$ is suggested. It is also noted by several authors, that unweighted pairwise likelihood (i.e. $w_{ij} = 1$) can be preferable for inference about the association parameter, while in general weighted versions improve the estimation of marginal mean parameters.

The third type of efficiency enhancement occurs mostly in the context of longitudinal and spatial data analysis, where there is a distance metric associated with observation pairs. The rationale is that pairs further apart usually contain less information than pairs within shorter distances, so should not be included in the estimation for dependence parameters. This simply restricts ω_{ij} to be 0 or 1 depending on pairwise distances. For longitudinal data, Joe and Lee (2009) and Varin and Vidoni (2006) showed that pairwise likelihood constructed from adjacent pairs is preferable to that based on all possible pairs in the sequence. While in spatial analysis, Heagerty and Lele (1998) used pairs that are within a fixed distance lag to build their estimating functions, arguing that using pairs that are further separated is less efficient both computationally and statistically. Recent work by Bevilacqua et al. (2011) proposed to determine the distance lag optimally based on a certain norm of the Godambe information, and showed that estimation based on the optimal distance lag gains efficiency from the unweighted version.

The above outlined efficiency improvement methods are operationally appealing, however, they make no attempt to incorporate the correlations among the pairs (i.e. some form of high-order dependence) into estimation. This is the key source of information loss when using composite likelihood in place of the full likelihood. The fourth type of endeavor is conducted in this direction. Nott and Rydén (1999) proposed a way to account for correlations among paired observations when constructing the composite likelihood estimating equation in image analysis. The idea is to draw a neighborhood (called 'mask') around each observation, and then select pairs containing that observation to form composite score functions. Then these component score functions within the mask are stacked into a column vector. After that, a weight matrix is used to convert the column vector into estimating equations. Note that this way of taking a weighted sum incorporates correlations among pairs, and is different from the usual formulation of composite estimating equations. The weight matrix can be determined optimally by the Godambe information matrix of the column vector. Simulations results, unfortunately, showed little improvement in efficiency of this weighted composite likelihood approach over the ordinary pairwise composite likelihood methods for models considered in their paper.

2.4 Applications in Spatial or Spatio-Temporal Models

Composite likelihood methods have been applied to many practical situations where the full likelihoods are difficult to evaluate. Usually these situations involve correlated data which are either non-Gaussian or of high dimensions and are frequently encountered in spatial and spatio-temporal settings. Curriero and Lele (1999) used the composite likelihood method in spatial variogram estimation. They demonstrated that this method yields consistent estimates and is superior to likelihood-based methods in terms of weaker distributional assumptions and less computational burden. Heagerty and Lele (1998) applied the composite likelihood approach to binary spatial data via the probit model for pairwise observations. Nott and Rydén (1999) developed a version of weighted composite likelihood to incorporate correlations among pairs in image analysis. Pairwise composite likelihood methods have also been applied to model estimation of spatial point processes by Guan (2006). Varin et al. (2005) investigated pairwise likelihood for generalized linear models with spatially-varying random effects, and demonstrated that the proposed method yields estimators with high efficiency. Li and Lin (2006) modeled spatially correlated survival data by Gaussian copulas and bypassed the high-dimensional integration of the likelihood function by again considering the composite likelihood with pairwise observations. More applications of composite likelihood methods in other settings can be found in Varin (2008); Varin et al. (2011).

Applications of composite likelihoods in spatio-temporal analysis have just begun to be embraced by the research community, so are less abundant than related literatures in spatial analysis. Porcu et al. (2007) used pairwise composite likelihood method to estimate spatio-temporal covariance functions. Bevilacqua et al. (2011) proposed the concept of ′optimal distance lag′ based on the Godambe information matrix for the selection of pairs and incorporated it into the pairwise composite likelihood to improve efficiency. The first project of this dissertation will treat it as a bench mark and aims to further improve efficiency of the pairwise composite likelihood methods in the spatio-temporal context. A thorough treatment of the asymptotic properties of the proposed estimators will also be postulated.

CHAPTER III

Joint Composite Estimating Functions in Spatio-Temporal Models

3.1 Introduction

Spatio-temporal data arise from many scientific disciplines such as environmental sciences, climatology, geology, epidemiology, among others. Through data analysis, scientists are interested in understanding important factors that associate with the underlying processes and in predicting such processes at unobserved locations and time points. Both of these tasks require modeling the intrinsic dependency structure of the data, which is usually depicted by the spatio-temporal covariance structure. During past decades, much effort has been made in developing valid yet flexible spatio-temporal covariance models. For example, Cressie and Huang (1999) introduced a class of nonseparable, stationary covariance functions that address space-time interactions. Gneiting (2002) later expanded their work to larger classes of space-time covariance structures that do not depend on closed-form Fourier inversions. Stein (2005) derived space-time covariance functions that are spatially isotropic and not fully symmetric. Porcu et al. (2007) proposed another class of nonseparable space-time covariance structures that are spatially anisotropic, based on which one can formulate temporally asymmetric covariance functions.

Unfortunately, most of these useful covariance models have been seldom applied in practical studies collecting large-scale data sets. This is largely due to the tremendous computational burden in handling high-dimensional covariance matrices for either likelihood-based or Bayesian approaches. To circumvent this difficulty, people usually use simplified approaches to separately model spatial and temporal dependencies (Sahu et al., 2007; Smith and Kolenikov, 2003) or to use a separable spatio-temporal covariance function to ease computation (Zhu et al., 2003). Although these, as well as other similar models, have many desirable properties, they all ignore a crucial model component: the spatio-temporal interaction effect. Paciorek et al. (2009) attempted to capture the spatio-temporal interactions for both PM10 and PM2.5 processes using monthly-varying spatial surfaces. However, they assumed independence across spatial residual surfaces at each time point to reduce computational complexity, which essentially hampers their approach from quantifying the effect of spatio-temporal interaction.

In another area of spatial data research, people have tried to reduce data dimension using composite likelihood (CL) methods (Lindsay, 1988), which is a general class of pseudo-likelihoods based on likelihoods of marginal or conditional events (see Chapter II for details). To name a few, Curriero and Lele (1999) used CL method in spatial variogram estimation. They demonstrated that this method yields consistent estimates and is superior to likelihood-based methods in terms of weaker distributional assumptions and less computational burden. Heagerty and Lele (1998) applied CL approach to binary spatial data via the probit model based on pairwise observations. Li and Lin (2006) modeled spatially correlated survival data by Gaussian copulas and bypassed the high-dimensional integration of the likelihood function by again considering CL with pairwise observations. Varin (2008) and Varin et al. (2011) provide comprehensive reviews of marginal CL methods and their applications.

The CL approach is known to yield estimators with sound asymptotic properties, however, it is noticeable that most of CLs are based on pairs of observations, resulting in less efficient estimates than the full likelihood. Moreover, the number of pairs can be enormous in cases where the number of correlated observations is large. To deal with the large number of pairs, people usually select subsets of pairs within a certain distance lag (Heagerty and Lele, 1998; Varin et al., 2005). Recently, Bevilacqua et al. (2011) proposed an optimal distance lag in their weighted CL method, where the optimal lag was determined by minimizing a certain norm of the Godambe information (i.e. sandwich asymptotic covariance) matrix. They showed that through this optimal subset selection, a more efficient estimator could be obtained. They also found that estimation based on shorter lags generally yielded more efficient estimates than those based on larger lags.

A shortcoming of their CL estimation method is that correlations among pairs are completely ignored, which could cause some loss of estimation efficiency. This motivates Nott and Rydén (1999) and Kuk and Nott (2000) to formulate optimal composite estimating equations by accounting for correlations among the pairs. However, the method proposed in Nott and Rydén (1999) did not show much efficiency improvement for image data, while Kuk and Nott's method applies only to longitudinal data of moderate lengths or clustered data with moderate cluster sizes.

The objective of this chapter is to develop a more efficient CL estimation method for a joint analysis of spatio-temporal processes. Our approach is proposed to account for correlations among composite pairs in a feasible manner. To do so, we first divide all pairs into spatial, temporal and spatio-temporal cross groups, and then form group-based estimating functions respectively. This often results in over-identified estimating functions. To circumvent this, we construct a joint inference function through a weight matrix, in a similar spirit to the generalized method of moments (GMM) (Hansen, 1982) and the quadratic inference function (QIF) by Qu et al. (2000). The weight matrix is designed to give larger weights to more informative pairs and down weight noisy pairs, hence estimation efficiency can be improved.

The rest of the chapter is structured as follows. In section 2, we present the joint composite estimating function approach for spatio-temporal processes. In section 3, large-sample properties of the proposed estimator are discussed. Simulation studies comparing our method with the conventional composite likelihood method are detailed in section 4. We illustrate an application of our method to study the spatio-temporal dependence structure of PM10 particles in northeastern United States in section 5, followed by a discussion in section 6. Some technical details are listed in the Appendix.

3.2 Methodology

3.2.1 Model

Consider a realization of a spatio-temporal process $\{Y(s,t)$: $s \in \mathcal{S}$, $t \in \mathcal{T}$, $\mathcal{S} \subset \mathbb{R}^2$, $\mathcal{T} \subset \mathbb{R}^+\}$, where $\mathcal{S}$ denotes the set of spatial locations and $\mathcal{T}$ stands for the collection of time points. Assume that $Y(s,t)$ can be decomposed into a deterministic mean function $\mu(s,t)$, and a random component $X(s,t)$ as follows:

$$Y(s,t) = \mu(s,t) + X(s,t), \;\; s \in \mathcal{S}, \; t \in \mathcal{T}.$$

Suppose $X(s,t)$ can be further modeled as

$$X(s,t) = \alpha(s,t) + \epsilon(s,t), \;\; s \in \mathcal{S}, \; t \in \mathcal{T},$$

where the process $\alpha(s,t)$ characterizes the spatio-temporal variations, and $\epsilon(s,t)$ is a normally distributed measurement error with mean zero and variance σ_ϵ^2, independent of each other and independent of $\alpha(s,t)$. Variance σ_ϵ^2 is termed as the nugget effect in Geostatistics. Assume $\{\alpha(s,t)$: $s \in \mathcal{S}$, $t \in \mathcal{T}\}$ follows a multivariate Gaussian process with mean zero and a covariance function C, which is given by, for any two observations at spatio-temporal coordinates (s_1,t_1) and (s_2,t_2)

$$Cov(\alpha(s_1,t_1), \alpha(s_2,t_2)) = C(s_1,s_2,t_1,t_2;\boldsymbol{\theta}^{'}).$$

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}^{'}, \sigma_\epsilon^2)$ be an r-element vector of parameters of interest. We will focus on estimating the covariance structure of $X(s,t)$ in the rest of the paper, provided that the observed process has first been properly de-trended; otherwise, it is relatively straightforward to extend the proposed method by including a mean model for $\mu(s,t)$ (Cressie, 1993).

3.2.2 Composite Estimating Functions

To apply the method of composite likelihoods based on pairs of observations to reduce data dimension, we consider pairwise differences following Curriero and Lele (1999). Let

$$d(k) \equiv d(s_1, t_1, s_2, t_2) = X(s_1, t_1) - X(s_2, t_2), \ \ k \in D_n(p,q), \tag{3.1}$$

where

$$\begin{aligned} D_n(p,q) &= \left\{ (s_1,t_1,s_2,t_2) : \begin{array}{l} s_2 \geq s_1, ||s_1 - s_2|| \leq p, \\ t_2 \geq t_1, |t_1 - t_2| \leq q, \\ t_1 \neq t_2 \text{ if } s_1 = s_2, \\ s_1 \neq s_2 \text{ if } t_1 = t_2 \end{array} \right\} \\ &\subset \mathcal{S} \times \mathcal{T} \times \mathcal{S} \times \mathcal{T} \subseteq \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+. \end{aligned}$$

Here n is the length of the realized process $X(s,t)$, and $|| \bullet ||$ is the Euclidean distance between two points in a d-dimensional space with $d \geq 2$. The ordering of spatial locations is defined as follows: for two locations $s_1 = (a_1, b_1)$, and $s_2 = (a_2, b_2)$, we say $s_1 > s_2$ if $a_1 > a_2$ or if $a_1 = a_2$ and $b_2 > b_1$, where (a,b) are the coordinates for a location. The set $D_n(p,q)$ contains all pairs of observations within p units of space and q units of time lags in the coordinate space $\mathcal{S} \times \mathcal{T}$.

When both p and q equal infinity, the set includes all possible pairs of observations. For simplicity of exposition, we drop the two indices, and write $D_n(p,q)$ as D_n.

The values of p and q may be determined according to different criteria. They can be chosen by practical considerations, such as sample size or boundary limits. They can also be determined by some preliminary evaluations (e.g. empirical variograms) of the spatial and temporal dependence decay rates, and be set to ranges that sustain fairly high level of correlation. Or we may choose such p and q that maximize the Godambe information (Godambe and Heyde, 1987) of the corresponding composite estimating functions, so that the resulting estimator will have minimum variance. Clearly, the latter approach requires the evaluation of the sandwich information matrix for different combinations of these cutoff lags, which is computationally demanding. Many simulation results reported in the literature (e.g. Varin et al., 2005; Bevilacqua et al., 2010; Davis and Yau, 2011), have suggested choosing p and q to include only pairs that are within shorter distances for better estimation efficiency. This means that we can exclude a substantial number of pairs from set D_n that are further apart in either space or time to reduce the computational burden.

It is easy to see that the difference process $d(k)$ in equation (3.1) follows a univariate normal distribution with mean zero and variance given by

$$\begin{aligned} Var\left\{d(k)\right\} &= C(s_1,s_1,t_1,t_1;\boldsymbol{\theta}) + C(s_2,s_2,t_2,t_2;\boldsymbol{\theta}) + 2\sigma_\epsilon^2 - 2C(s_1,s_2,t_1,t_2;\boldsymbol{\theta}) \\ &\equiv 2\gamma_k(\boldsymbol{\theta}). \end{aligned}$$

Denote the composite score function (CSF) for the observed $d(k)$ as $f_k(d(k);\boldsymbol{\theta})$.

Then

$$f_k\left(d(k);\boldsymbol{\theta}\right) = \frac{\dot{\gamma}_k(\boldsymbol{\theta})}{2\gamma_k(\boldsymbol{\theta})}\left\{1 - \frac{d^2(k)}{2\gamma_k(\boldsymbol{\theta})}\right\},$$

where, for any function f, $\dot{f}$ denotes the vector of gradients of function f with respect to the parameter vector $\boldsymbol{\theta}$. It is clear that $f_k\left(d(k);\boldsymbol{\theta}\right)$ is an unbiased estimating function for $\boldsymbol{\theta}$ since it is derived from a valid density function.

According to the CL literature (Reid and Cox, 2004; Varin et al., 2011), a common version of composite estimating functions is,

$$\Psi_{CL}(\boldsymbol{\theta}) = \sum_{k\in D_n} f_k\left(d(k);\boldsymbol{\theta}\right),$$

where $d(k)$ are implicitly treated as being independent.

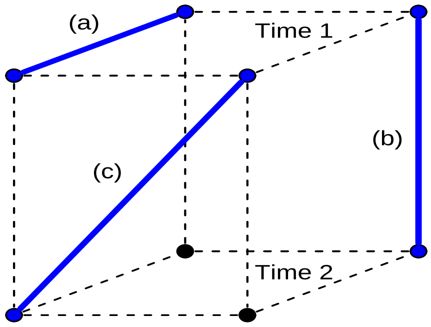
Alternatively, one may stack the individual CSF terms into a column vector $\nu(\boldsymbol{\theta}) = \{f_k\left(d(k);\boldsymbol{\theta}\right)\}_{k\in D_n}$, from which the estimating function is given by

$$\left\{E\left(\dot{\nu}(\boldsymbol{\theta})\right)\right\}^T\left\{Cov(\nu(\boldsymbol{\theta}))\right\}^{-1}\nu(\boldsymbol{\theta}) = 0.$$

As pointed out by Kuk (2007), this version of composite estimating equations effectively accounts for the correlations among the differences. However, the calculation of $Cov(\nu(\boldsymbol{\theta}))$ and/or its inverse is computationally prohibitive when the number of pairs (or differences) is large.

To improve over the existing CL methods and to incorporate correlations among the pairs in the estimation, we propose a new approach. That is, we construct three sets of estimating functions with the utility of spatio-temporal characteristics of the data. Specifically, we first partition D_n into three subsets, namely $D_{S,n}$, with pairs differing only in locations; $D_{T,n}$, with pairs differing only in time; and $D_{C,n}$, with pairs differing both in locations and time. Hence $D_n = D_{S,n} \cup D_{T,n} \cup D_{C,n}$. Figure 3.1 displays such partition with these three types of pairs, (a) for spatial pair, (b) for temporal pair, and (c) for spatio-temporal cross pair, in a typical spatio-temporal setting with four locations observed at two time points.

Figure 3.1: Configurations of spatio-temporal pairs. Upper plane represents four locations observed at time 1, lower plane represents the same four locations observed at time 2. (a) is the spatial pair, (b) the temporal pair, and (c) the spatio-temporal cross pair.



Summing over all pairwise differences of spatial pairs across all time points, we obtain the following spatial composite estimating function (CEF):

$$\Psi_{S,n}(\boldsymbol{\theta}) = \frac{1}{|D_{S,n}|} \sum_{i \in D_{S,n}} f_i\left(d(i); \boldsymbol{\theta}\right),$$

where for any set $\mathcal{A}$, $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$. In a similar fashion, we construct the temporal CEF:

$$\Psi_{T,n}(\boldsymbol{\theta}) = \frac{1}{|D_{T,n}|} \sum_{j \in D_{T,n}} f_j\left(d(j); \boldsymbol{\theta}\right).$$

Likewise, the third CEF is formed by using spatio-temporal cross pairs:

$$\Psi_{C,n}(\boldsymbol{\theta}) = \frac{1}{|D_{C,n}|} \sum_{l \in D_{C,n}} f_l\left(d(l); \boldsymbol{\theta}\right).$$

Note that the resulting estimating functions constructed using the group-specific pairs characterize different profiles of the spatio-temporal process. The spatial piece $\Psi_{S,n}(\boldsymbol{\theta})$ provides paramount information of the spatial dependency; the temporal piece $\Psi_{T,n}(\boldsymbol{\theta})$ contains key information of the temporal dependency; and the spatio-temporal cross piece $\Psi_{C,n}(\boldsymbol{\theta})$ is more relevant to information of the spatio-temporal interaction. The total number of equations, when three pieces are combined as $(\Psi^T_{S,n}(\boldsymbol{\theta}), \Psi^T_{T,n}(\boldsymbol{\theta}), \Psi^T_{C,n}(\boldsymbol{\theta}))^T$, is larger than the number of parameters. As a result, parameters cannot be estimated by directly solving these equations due to the issue of over-identification. We then form a weighted quadratic objective function in a similar spirit to the generalized method of moments (GMM) (Hansen, 1982), so that estimates can be obtained by minimizing the objective function.

Precisely, let W be a positive-definite matrix, and let

$$\Gamma_n(\boldsymbol{\theta}) = (\Psi^T_{S,n}(\boldsymbol{\theta}), \Psi^T_{T,n}(\boldsymbol{\theta}), \Psi^T_{C,n}(\boldsymbol{\theta}))^T.$$

A quadratic inference function takes the following form:

$$Q_n(\boldsymbol{\theta}) = \Gamma^T_n(\boldsymbol{\theta}) W^{-1} \Gamma_n(\boldsymbol{\theta}),$$

and the estimator is given by

$$\hat{\boldsymbol{\theta}}_n = argmin_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_n(\boldsymbol{\theta}). \tag{3.2}$$

We call this $\hat{\boldsymbol{\theta}}_n$ the joint composite estimating function (JCEF) estimator.

The classic GMM theory indicates that the optimal weight matrix is the asymptotic covariance matrix of composite estimating functions, namely $Cov(n\Gamma_n(\boldsymbol{\theta}))$. However, this result cannot be directly applied here, because our objective function $Q_n(\boldsymbol{\theta})$ is special in two aspects. First, the three estimating functions, $\Psi_{S,n}(\boldsymbol{\theta})$, $\Psi_{T,n}(\boldsymbol{\theta})$ and $\Psi_{C,n}(\boldsymbol{\theta})$, are constructed from different sets of observations, while in the standard GMM, different moment conditions are based on the same set of observations. Second, the numbers of terms in the three composite estimating functions are different due to the fact that the numbers of spatial, temporal and cross pairs are different. When one CEF consists of significantly more pairs, it will gain higher weight in the objective function due to its larger stratum size. So it is necessary to adjust for such stratum effect by using a normalized weight matrix, in a similar spirit to stratified sampling.

To proceed, let I_r be the $r \times r$ identity matrix. Write

$$\sqrt{\mathcal{N}} = diag(\sqrt{|D_{S,n}|}, \sqrt{|D_{T,n}|}, \sqrt{|D_{C,n}|}) \otimes I_r,$$

where $\otimes$ denotes Kronecker product of two matrices. This defines a block diagonal matrix with the first r diagonals being $\sqrt{|D_{S,n}|}$, the next r diagonals $\sqrt{|D_{T,n}|}$, and the last r diagonals $\sqrt{|D_{C,n}|}$. The normalized weight matrix is given by

$$W = \sqrt{\mathcal{N}} Cov\left\{\Gamma_n(\boldsymbol{\theta})\right\} \sqrt{\mathcal{N}}.$$

When $|D_{S,n}|$, $|D_{T,n}|$ and $|D_{C,n}|$ are approximately the same, W and $Cov\left\{n\Gamma_n(\boldsymbol{\theta})\right\}$ will play the same role in weighting. However, when one of $|D_{S,n}|$, $|D_{T,n}|$ or $|D_{C,n}|$

is considerably larger than the rest, W will help adjust for the unbalanced stratum sizes, so that the smaller stratum will make comparable contribution to the estimation. Zhao and Joe (2005) used a similar approach to account for different cluster sizes in their CL formulation for familial data. See also Joe and Lee (2009) for more detailed discussion.

3.2.3 Estimation of the Weight Matrix

Although $Cov\left\{\Gamma_n(\boldsymbol{\theta})\right\}$ can be derived analytically using multivariate Gaussian quadrant probabilities, given the large number of possible pairs, computing it based on analytic formulas is not practically feasible. Alternatively, in spatial data analysis, estimation of this covariance matrix is mostly achieved by subsampling techniques as done in Heagerty and Lele (1998); Heagerty and Lumley (2000); Lee and Lahiri (2002); Li and Lin (2006). Specifically, let the sampling region $A_n = \mathcal{S} \times \mathcal{T}$, where $|A_n| = n$. Under the assumption that, asymptotically, $|A_n| E\left\{\Gamma_n(\boldsymbol{\theta})\Gamma_n^T(\boldsymbol{\theta})\right\} \to \Sigma$, we can estimate Σ using sample covariance matrix of statistics computed on subshapes of the sampling region A_n. That is,

$$\hat{\Sigma}_n = k_n^{-1} \sum_{i=1}^{k_n} |A_{l(n)}^i| \left\{\Gamma_n^i(\boldsymbol{\theta}) - \bar{\Gamma}_n(\boldsymbol{\theta})\right\}^2, \tag{3.3}$$

with $\bar{\Gamma}_n(\boldsymbol{\theta}) = \sum_{i=1}^{k_n} \Gamma_n^i(\boldsymbol{\theta})/k_n$, where $\Gamma_n^i(\boldsymbol{\theta})$ is vector $\Gamma_n(\boldsymbol{\theta})$ evaluated in $A_{l(n)}^i, i = 1, \ldots, k_n$, a collection of (non)overlapping subshapes of A_n, and k_n is the number of subshapes.

This subsampling method was first introduced by Carlstein (1987) for strictly stationary time series. Sherman (1996) later showed that it could be used to estimate the moments of a general statistic for random fields on a lattice. Moreover, Kunsch (1989) demonstrated that the use of overlapping replicates led to a more stable variance estimate than non-overlapping replicates. The optimal subsample size was given by Politis and Romano (1994) for a stationary random field on a d-dimensional lattice as $Mn^{d/(d+2)}$, where M is a certain tunning constant. Heagerty and Lumley (2000) studied the effect of different choices of M for regression models. Sherman (1996) pointed out that it was useful to gather some empirical evidence about the range of correlation in determining M. If the correlation decays fast, small subsamples can be used; otherwise a large one should be considered.

We will apply this subsampling technique to estimate our weight matrix and later investigate its performance in the standard error calculation. Other more sophisticated resampling schemes in spatial data analysis can be found in Lele (1991), Lahiri et al. (1999) and Zhu and Morgan (2004), among others. Note that to calculate $Cov\left\{\Gamma_n(\boldsymbol{\theta})\right\}$ for each subsample, parameter values have to be given. We propose to generate some simple consistent estimates by either setting the weight matrix to the identity matrix in the JCEF method or using estimates from the empirical variogram.

Many established numerical optimization methods can be used to obtain parameter estimates that minimize $Q_n(\boldsymbol{\theta})$. However, given the complex nature of the parametric covariance structures $C(\bullet;\boldsymbol{\theta})$, algorithms that do not require calculations of the Hessian matrix are desirable in this case. Quasi-Newton, Nelder-Mead and conjugate-gradient methods are possible choices. These optimization routines are offered by many mathematical and statistical software such as MatLab and R. To ensure that the true minimum of the target function is found, a set of good starting values is very important, which in our case can be found by fitting the corresponding parametric variogram to the empirical variogram (Cressie, 1993). The related detail is illustrated in section 5.

3.3 Large Sample Properties

Asymptotic properties of the estimators based on composite likelihood have been quite established in the context of longitudinal or clustered data settings, where there exist independent replicates of the data, see, e.g., Lindsay (1988); Molenberghs and Verbeke (2005); Zhao and Joe (2005). However, little research work is available in literature to address the asymptotic properties of composite likelihood estimators for spatially and temporally dependent data with no replicates. The well-referenced papers by Heagerty and Lele (1998) and Guan et al. (2004) are among the few sources of asymptotic properties for composite likelihood estimators with spatial data. However, their results are limited either to regular lattice data or strictly stationary random fields, and require the sampling domain to expand in a certain fashion. In the following paragraphs, I will establish the asymptotic properties of the JCEF estimator in more practical settings.

The asymptotic properties of the JCEF estimator defined in equation (3.2) are mainly governed by asymptotic behaviors of $\Gamma_n(\boldsymbol{\theta})$. Once we establish a uniform law of large numbers (ULLN) and a central limit theorem (CLT) for $\Gamma_n(\boldsymbol{\theta})$, the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ will follow from the standard GMM arguments. We derive these large-sample results for fixed spatial and temporal lags p and q under increasing-domain asymptotics. That is, the increase in sample size is achieved by the expansion of the sampling domain in space or time or simultaneously. As a result of fixing p and q, the numbers of pairs in the spatial, temporal and spatio-temporal cross groups are proportional to the total number of data points n for the observed random process, that is $|D_{S,n}|$, $|D_{T,n}|$, and $|D_{C,n}|$ are of the same order $O(n)$. For simplicity, we assume the weight matrix W is known. Otherwise, a $\sqrt{n}$-consistent $\hat{W}$ would be sufficient for us to modify our justifications.

3.3.1 Assumptions

Jenish and Prucha (2009) developed a set of limit theorems for random processes under rather general conditions of nonstationarity, unevenly spaced locations, and general forms of sample regions. We tailor relevant regularity conditions to establish large-sample properties for our JCEF estimator as follows:

Assumption 1 The (possibly unevenly spaced) lattice $D \subset \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+$ is infinitely countable. All elements in D are located at distances of at least $d_0 > 0$ from each other. That is, $\rho(i,j) \geq d_0$, for all $i, j \in D$, where $\rho(i,j)$ is a distance metric for any two points i, $j \in D$. See a detailed definition of the distance metric in the Appendix A.

Assumption 2 $\{D_{\mathcal{A},n} : n \in \mathbb{N}\}$ is a sequence of arbitrary finite subsets of D, satisfying $|D_{\mathcal{A},n}| \to \infty$ as $n \to \infty$, for $\mathcal{A} \in \{S, T, C\}$.

Assumption 3 $(\boldsymbol{\Theta}, v)$ is a totally bounded parameter space with metric v.

Assumption 4 (Uniform $L_{2+\delta}$ integrability) Let $q_k = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||f_k(d(k); \boldsymbol{\theta})||$. Then for some $\delta > 0$, $\lim_{e \to \infty} E q_k^{2+\delta} \mathbf{1}(||q_k|| > e) = 0$, for all $k \in D_n$.

Assumption 5 $E \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||\dot{f}_k(d(k); \boldsymbol{\theta})|| < \infty$, for all $k \in D_n$.

Assumption 1 ensures that the increase of sample size is achieved by an expanding domain, thus it rules out the in-fill asymptotics. Assumption 2 guarantees that sequences of subsets $D_{S,n}$, $D_{T,n}$, and $D_{C,n}$, on which the process is generated, increase in cardinality. Assumption 3 regulates the parameter space. Assumptions 4 and 5 are regularity conditions for score functions. The uniform integrability condition in Assumption 4 is a standard moment assumption postulated in CLTs for one-dimensional processes. A sufficient condition for the uniform $L_{2+\delta}$ integrability of f_k is its uniform L_γ boundedness for some $\gamma > 2 + \delta$. A weaker assumption of L_1 integrability is sufficient for a LLN for f_k. Assumption 5 is a Lipschitz-type condition, implying that the score functions are L_0 stochastically equicontinuous, so that a ULLN can be obtained.

The difference process $d(k)$ is usually not stationary. To regulate its dependence structure, we impose some α-mixing conditions on $d(k)$. Let U and V be two subsets of D_n, and let $\sigma(U) = \sigma\{d(k); k \in U\}$ be the σ-algebra generated by random variables $d(k), k \in U$. Define

$$\alpha(U, V) = \sup\{|P(A \cap B) - P(A)P(B)|; A \in \sigma(U), B \in \sigma(V)\}.$$

Then this α-mixing coefficient for the random field $\{d(k), k \in D_n\}$ is defined as:

$$\alpha(k, l, m) = \sup\{\alpha(U, V), |U| < k, |V| < l, \rho(U, V) \geq m\},$$

with $k, l, m \in \mathbb{N}$ and $\rho(U, V)$ the distance between sets U and V; see the Appendix A for the definition of ρ. In addition, we need the following conditions similar to those stated in Assumption 3 in Jenish and Prucha (2009).

Assumption 6 The process $\{d(k), k \in D_n\}$ satisfies the following mixing conditions in an a-dimensional space:

(a) $\sum_{m=1}^{\infty} m^{a-1}\alpha(1,1,m)^{\delta/(2+\delta)} < \infty$, for some $\delta > 0$,

(b) $\sum_{m=1}^{\infty} m^{a-1}\alpha(k,l,m) < \infty$ for $k + l \leq 4$,

(c) $\alpha(1,\infty,m) = O(m^{-a-\epsilon})$ for some $\epsilon > 0$.

Assumption 6 requires a polynomial decay of the α-mixing coefficient, which can be shown to hold for Gaussian processes, a special case of the Gibbs fields (Winkler, 1995; Doukhan, 1994).

3.3.2 Consistency

Consider a generic case of

$$\Psi_{\mathcal{A},n}(\boldsymbol{\theta}) = \frac{1}{|D_{\mathcal{A},n}|} \sum_{k \in D_{\mathcal{A},n}} f_k(d(k);\boldsymbol{\theta}),$$

where $\mathcal{A} \in \{S, T, C\}$.

Based on Theorems 2 and 3 in Jenish and Prucha (2009), Assumptions 1, 2, 4, and 6 ensure a point-wise LLN for f_k based on sub-series $\{d(k), k \in D_{\mathcal{A},n}\}$; with additional assumption 5 on stochastic equicontinuity of f_k, a uniform version of LLN is warranted. Thus, we have

Lemma 1. *Given Assumptions 1-6,*

$$\sup_{\boldsymbol{\theta} \in \Theta} ||\Psi_{\mathcal{A},n}(\boldsymbol{\theta}) - E\Psi_{\mathcal{A},n}(\boldsymbol{\theta})|| \xrightarrow{p} 0, \;\; \textit{as} \;\; n \to \infty.$$

Lemma 1 holds for $\Psi_{S,n}(\boldsymbol{\theta})$, $\Psi_{T,n}(\boldsymbol{\theta})$, and $\Psi_{C,n}(\boldsymbol{\theta})$, so we can show easily that for any given positive-definite weight matrix W,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - EQ_n(\boldsymbol{\theta})| \xrightarrow{p} 0, \quad \text{as } n \to \infty.$$

Consequently, we establish the consistency of the JCEF estimator in Theorem 1.

Theorem 1. *Under the same conditions stated in Lemma 1, if the true parameter value* $\boldsymbol{\theta}_0$ *is the unique minimizer of* $EQ_n(\boldsymbol{\theta})$*, and* $\hat{\boldsymbol{\theta}}_n$ *minimizes* $Q_n(\boldsymbol{\theta})$*, then* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0, \quad$ *as* $n \to \infty$.

3.3.3 Asymptotic Normality

To derive the asymptotic distribution of the JCEF estimator, the following additional regularity conditions are needed.

Assumption 7 Let $\Sigma_n(\boldsymbol{\theta}) = Var\left\{\Gamma_n(\boldsymbol{\theta})\right\}$, $\lim_{n\to\infty} n\Sigma_n(\boldsymbol{\theta}) = \Sigma(\boldsymbol{\theta})$, where $\Sigma(\boldsymbol{\theta})$ is a positive-definite matrix.

Assumption 8 $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\dot{\Gamma}_n(\boldsymbol{\theta}) - E\dot{\Gamma}_n(\boldsymbol{\theta})|| \xrightarrow{p} 0$. Write $\lim_{n\to\infty} E\dot{\Gamma}_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})$ is a positive-definite matrix.

Assumption 7 assumes that the variance of $\Gamma_n(\boldsymbol{\theta})$ is of order $O(n^{-1})$, which is also a standard assumption for the subsampling estimation of the covariance. Assumption 8 is a ULLN for $\dot{\Gamma}_n(\boldsymbol{\theta})$, which regulates the asymptotic variance of the estimator and can be obtained with the same regularity conditions on $\dot{\Gamma}_n(\boldsymbol{\theta})$ as those in Lemma 1.

Lemma 2. *Given Assumptions 1-4, 6 and 7, we have*

$$\sqrt{n}\,\Gamma_n(\boldsymbol{\theta}) \xrightarrow{d} N(0, \Sigma(\boldsymbol{\theta})), \quad as \quad n \to \infty.$$

A sketch of the proof for Lemma 2 is given in the Appendix B. Then based on the standard GMM arguments (Hansen, 1982), we establish the following theorem:

Theorem 2. *Given Assumptions 1-4, and 6-8, we have*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Omega(\boldsymbol{\theta}_0)\Sigma(\boldsymbol{\theta}_0)\Omega^T(\boldsymbol{\theta}_0)), \quad as \quad n \to \infty,$$

where $\Omega(\boldsymbol{\theta}_0) = -[I^T(\boldsymbol{\theta}_0)W^{-1}I(\boldsymbol{\theta}_0)]^{-1}I^T(\boldsymbol{\theta}_0)W^{-1}$.

Note that the above results are applicable to more general settings than those considered in Heagerty and Lele (1998). Their asymptotic results are based on the theory in Guyon (1995), which required the sample regions to form a strictly increasing sequence on evenly-spaced lattices. In contrast, we do not impose any restrictions on the geometry and growth behavior of the sample regions and allow for unevenly spaced locations, a situation frequently encountered in real data analysis. Moreover, our results accommodate sampling domain expansions both in space and time, while results in Li et al. (2007) only deal with the expansion in time. In fact, our results are even applicable to processes with unbounded moments, which may arise in a wide range of real world applications. For more discussion, refer to Jenish and Prucha (2009). It is also worth pointing out that asymptotic results for infinite spatial or/and temporal lags are slightly different, because convergence rates of $\Psi_{S,n}(\boldsymbol{\theta})$, $\Psi_{T,n}(\boldsymbol{\theta})$ and $\Psi_{C,n}(\boldsymbol{\theta})$ may be of different orders, due to the differences in expansion rates in space and time.

3.4 Simulation Experiments

To assess the performance of the proposed JCEF method, we conduct a simulation study, in which we compare the proposed method with the weighted composite likelihood (WCL) method given in Bevilacqua et al. (2011). Following Bevilacqua et al. (2011), we form our composite estimating functions based on neighboring pairs for both cases of WCL and JCEF. It is noted that tuning the distance lag according to a certain optimality criterion (e.g. minimizing the trace of the inverse of the Godambe information) for each specific case can result in better efficiency. However, using a common distance lag in the simulation study serves the purpose of comparison, while the related computational burden appears manageable.

The spatio-temporal covariance function used in the data generation is a non-separable spatio-temporal covariance structure proposed in Cressie and Huang (1999):

$$C(h,u|\boldsymbol{\theta}) = \begin{cases} \frac{\sigma^2(2\beta)}{(a^2u^2+1)^{\nu}(a^2u^2+\beta)\Gamma(\nu)} \left\{ \frac{b}{2} \left(\frac{a^2u^2+1}{a^2u^2+\beta} \right)^{\frac{1}{2}} h \right\}^{\nu} K_{\nu} \left(b \left(\frac{a^2u^2+1}{a^2u^2+\beta} \right)^{\frac{1}{2}} h \right), & \text{if } h > 0, \\ \frac{\sigma^2(2\beta)}{(a^2u^2+1)^{\nu}(a^2u^2+\beta)}, & \text{if } h = 0, \end{cases} \tag{3.4}$$

where $u = |t_1 - t_2|$ is the time lag and $h = ||s_1 - s_2||$ the Euclidean distance between two locations. K_{ν} is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun (1972), p.374), where $\nu > 0$ is a smoothness parameter characterizing the behavior of the correlation function near the origin. If $u = 0$, $C(h, 0|\boldsymbol{\theta})$ degenerates into a purely spatial covariance, which is the popular Matèrn class used in spatial statistics. When $\nu = 0.5$, this spatial correlation model is an exponential function of h, when $\nu \to \infty$, the Gaussian correlation function. Let

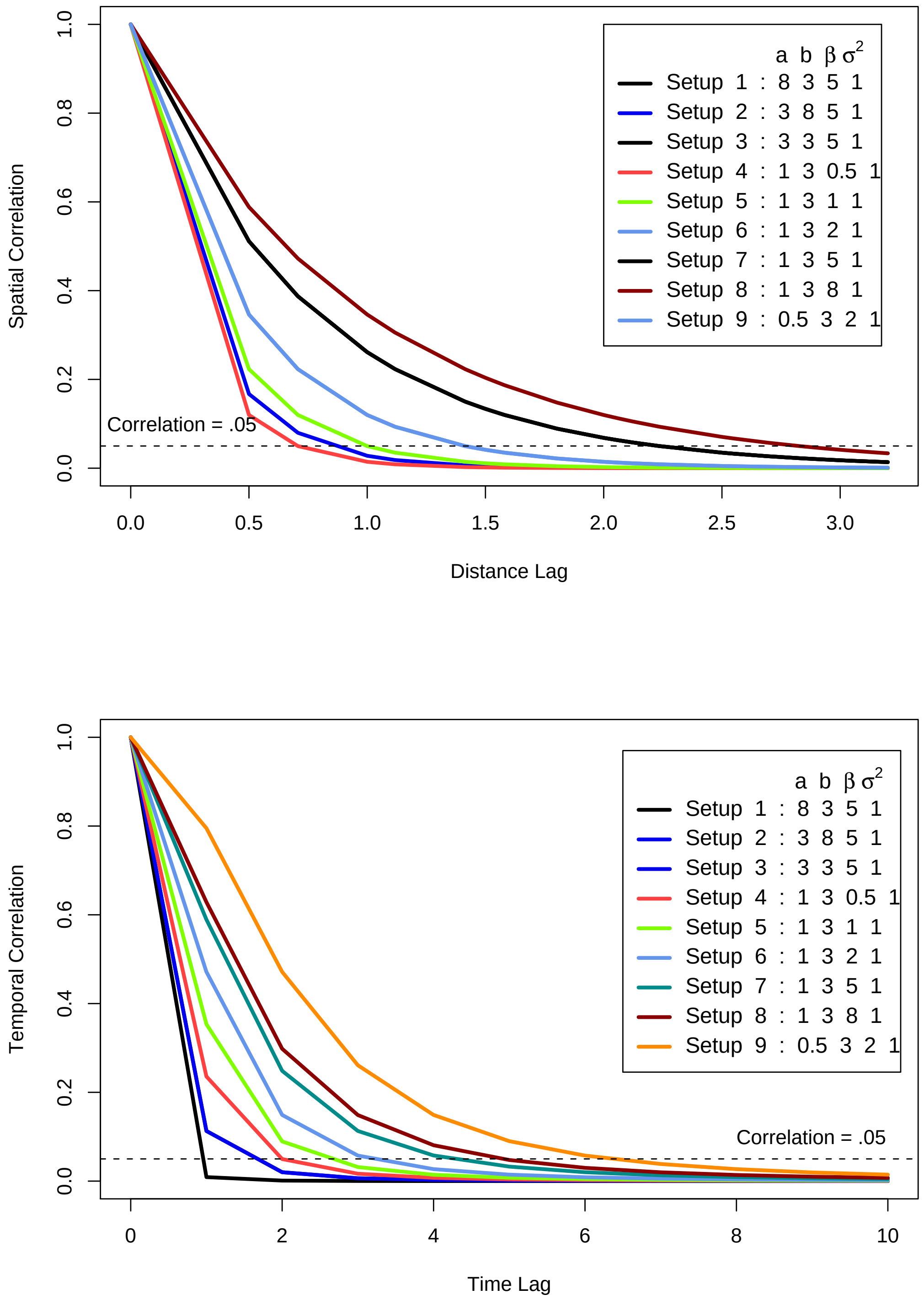
$\boldsymbol{\theta} \equiv (a, b, \beta, \nu, \sigma^2)$. $a \geqslant 0$ is the scaling parameter of time, $b \geqslant 0$ is the scaling parameter of space, $\beta > 0$ is a space-time interaction parameter, and $\sigma^2 = C(0,0) > 0$, the variance at the origin. Note that a separable covariance function is obtained when $\beta = 1$.

We generate $X(s,t)$ on a regular grid of $7 \times 7 \times 30$ space-time points, with the spatial coordinates being set at $(1, 1.5, \ldots, 4) \times (1, 1.5, \ldots, 4)$ and $\mathcal{T} = (1, 2, \ldots, 30)$. Table 3.1 includes nine simulation setups, whose marginal spatial/temporal correlation patterns are displayed in Figure 3.2, respectively. To create a variety of dependence structures in the simulation study, we start in setup 1 with short-range spatial and temporal dependences, and next vary the spatial and temporal scaling parameters a and b and then vary the spatio-temporal interaction parameter β. It is clear that the decaying rate of correlation slows down from setup 1 to setup 9 in both directions of space and time.

Also, for this type of covariance structure, the spatial/temporal scaling parameter, together with the interaction parameter, determine the marginal dependence patterns in space or time. For example, in setup 2 and setup 3, the same a and β values lead to identical marginal temporal dependence patterns. The same phenomenon occurs in setup 1 and setup 3 as well as setup 6 and setup 9 for marginal spatial correlations. Note that the interpretation of the temporal and spatial lags on the horizontal axes should be based on the specific units used to specify coordinates.

Parameter ν is fixed at 0.5 in the simulation. In practice, ν is difficult to estimate, because it requires dense space data and may run into identifiability problem (Stein, 1999). Also as pointed out by Huang et al. (2007), the estimation of σ_ϵ^2 may cause numerical instability, hence is fixed in the simulation for the convenience

Figure 3.2: Spatial and temporal correlation patterns for simulation setups 1-9. Parameter ν is fixed at 0.5.



of comparison. We will outline a profile quadratic inference function approach to estimating ν in the discussion.

Table 3.1: Mean squared errors (MSE) of parameter estimates. Results are from 200 simulations based on covariance structure in equation (3.4). Total MSE is the sum of MSEs for four parameters. RE is the relative efficiency defined as the total MSE of WCL over that of JCEF.

Simulation Scenarios					Method	Mean Squared Errors				Total MSE	RE
						a	b	β	σ^2		
	a	b	β	σ^2							
Setup 1	8	3	5	1	WCL	3.6035	0.1034	0.3400	0.0060	4.05	
					JCEF	3.3502	0.0943	0.3420	0.0058	3.79	1.07
Setup 2	3	8	5	1	WCL	0.6664	0.9276	0.9997	0.0017	2.60	
					JCEF	0.1656	0.6594	0.4080	0.0018	1.23	2.10
Setup 3	3	3	5	1	WCL	0.0869	0.0794	0.4294	0.0052	0.60	
					JCEF	0.0777	0.0888	0.2266	0.0053	0.40	1.51
Setup 4	1	3	0.5	1	WCL	0.0047	0.1046	0.0030	0.0018	0.11	
					JCEF	0.0037	0.0703	0.0023	0.0019	0.08	1.46
Setup 5	1	3	1	1	WCL	0.0029	0.0599	0.0073	0.0028	0.07	
					JCEF	0.0022	0.0527	0.0037	0.0027	0.06	1.19
Setup 6	1	3	2	1	WCL	0.0025	0.0637	0.0183	0.0026	0.09	
					JCEF	0.0020	0.0467	0.0085	0.0026	0.06	1.46
Setup 7	1	3	5	1	WCL	0.0031	0.1157	0.2439	0.0073	0.37	
					JCEF	0.0025	0.1080	0.1288	0.0072	0.25	1.50
Setup 8	1	3	8	1	WCL	0.0038	0.1950	1.3044	0.0111	1.51	
					JCEF	0.0028	0.1967	0.5698	0.0113	0.78	1.94
Setup 9	0.5	3	2	1	WCL	0.0007	0.0822	0.0321	0.0064	0.12	
					JCEF	0.0005	0.0670	0.0172	0.0073	0.09	1.32

Estimation of the weight matrix is achieved by sub-group sampling on overlapping sub-blocks of size $4 \times 4 \times 15$, determined according to the rule given in Politis and Romano (1994). We use the estimates from WCL to evaluate the individual score functions in each sub-blocks. A total of 200 simulated datasets are generated for each setup. We compare the JCEF and WCL in terms of mean squared errors (MSE).

The results, summarized in Table 3.1, show that the JCEF method clearly outperforms WCL in all nine simulation setups in terms of the total MSE, the sum of individual MSEs. The relative efficiency (RE), defined as the ratio of the total MSE of WCL over that of JCEF, shows that, for all scenarios, JCEF clearly gains efficiency compared to WCL and in some cases such gain is substantial. Parameter-specific MSEs indicate that most of the improvement occurs in the estimation of the interaction parameter β. On average, the MSE reduction is 40.52% for β, followed by 26.1% for a, the temporal scaling parameter, and then 13.52% for b, the spatial scaling parameter. The estimates for the variance parameter σ^2 are comparable for the two methods. The significant efficiency improvement for β, a and b are very desirable, since these are important parameters pertaining to the dependence structure. Especially for the interaction parameter β, valid parameter and standard error estimates will help researchers make infernece about whether a simpler, separable spatio-temporal covariance is supported by the data.

3.4.1 Standard Error Estimation

The key to obtaining valid standard error estimates is to create proper replicates of the data. As done in the step of weight matrix estimation, we invoke the subsampling method to calculate standard errors. A similar formula to that in equation (3.3) is used with $\boldsymbol{\theta}_n^i$ replacing $\Gamma_n^i(\boldsymbol{\theta})$. The subsample size is determined by $Mn^{b/(b+2)}$, where b equals 3 in the spatio-temporal setting. Following Heagerty and Lumley (2000), we vary the tuning parameter M from 2 to 4 to assess the effects of different subsample sizes on the standard error estimation, resulting in three subsampling schemes: $3 \times 3 \times 15$, $4 \times 4 \times 15$, and $4 \times 4 \times 20$, respectively. The same weight matrix used in the previous JCEF estimation is used for each subsample evaluation.

Another popular approach to creating data replicates is through parametric bootstrap. That is, after obtaining JCEF estimates, we generate data based on the estimated model, and the square root of the sample variance of the JCEF estimates across replicates is obtained as the estimate. This method involves more computation, but is less prone to bias than subsampling, which is likely to introduce extra bias with artificially created subsamples in finite samples. Bevilacqua et al. (2010) adopts the parametric bootstrap approach for constructing tests of separability of space-time covariance functions. We consider a comparison of subsampling and parametric bootstrap with bootstrap sample size 200. Given the importance of the spatio-temporal interaction parameter β, we devote our attention to β in the simulation.

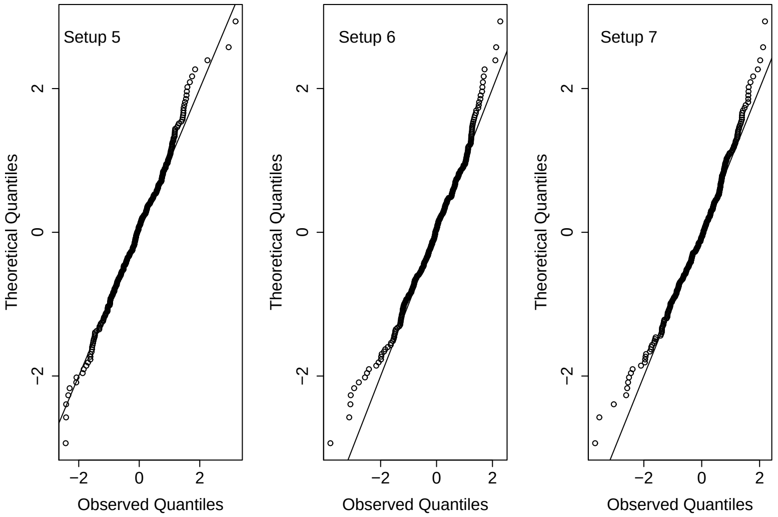
Table 3.2 lists results from 300 rounds of simulation for setups 5-7 with β equals 1, 2 and 5, respectively. We can see that different subsample sizes do have an impact on standard error estimation. Smaller subsamples yield standard error estimates closer to the empirical standard deviations, while larger subsamples tend to underestimate the variations. The reason may be that we use all overlapping sub-blocks and larger sub-blocks share more common observations, leading to less variations among blocks. However, truncation bias can occur if subsamples are too small, since it may fail to account for correlations at longer distances. Parametric bootstrapped standard error estimates perform very well in all three settings with estimates very close to the empirical standard deviations. This is because with consistent parameter estimates, the bootstrap procedure would yield a standard error estimate similar to the empirical one. In summary, if parametric bootstrap is feasible computationally, it is recommended; otherwise, subsampling is a way to do it. Obviously, some further investigation is needed to determine the tuning parameter M.

From the QQ-plot in Figure 3.3, we can see that the estimated β values follow the normal distribution closely, which means that using the normal approximation and the bootstrapped standard error estimates, valid inference on β is insured. This is confirmed by computing the 95% coverage probabilities across replicates for the three setups. In Table 3.2, parametric bootstrap and subsampling with $3 \times 3 \times 15$ partition scheme yield covarage probabilities close to the nominal 95%, while the other two subsampling schemes have smaller coverage probabilities due to underestimated standard errors. As a byproduct of this simulation, WCL estimates as inputs for the weight estimation are also recorded. The calculated MSE in Table 2 again shows that the JCEF method considerably lowers MSE leading to efficiency gain. The reduction in MSE is mainly due to the reduction in standard deviations, that is, both methods produce consistent estimates, but those from JCEF have smaller variances, which again corroborates the theory.

Table 3.2: Standard errors of parameter estimates for β. Results are from 300 simulations based on covariance structure in equation (3.4). SE is the standard error. CP is the 95% coverage probability. MSE is the mean squared error. Subsampling (Sub) and parametric bootstrap(Par.boot) are used to calculate SE of β. SE_e is the empirical standard deviation of $\hat{\beta}$.

	JCEF							WCL		
	Method		SE	CP	SE_e	Mean	MSE	SE_e	Mean	MSE
Setup 5	Sub	$4 \times 4 \times 20$	0.0728	87.67 %	0.0937	0.9998	0.0088	0.2126	1.0282	0.0458
$\beta = 1$		$4 \times 4 \times 15$	0.0814	93.33 %						
		$3 \times 3 \times 15$	0.0929	95.67 %						
	Par.boot		0.0997	96.67 %						
Setup 6	Sub	$4 \times 4 \times 20$	0.1273	86.33 %	0.1804	1.9944	0.0325	0.3879	2.0393	0.1515
$\beta = 2$		$4 \times 4 \times 15$	0.1423	89.67 %						
		$3 \times 3 \times 15$	0.1920	96.00 %						
	Par.boot		0.1744	94.00 %						
Setup 7	Sub	$4 \times 4 \times 20$	0.4034	79.67 %	0.6125	5.0254	0.3746	1.1050	5.2141	1.2627
$\beta = 5$		$4 \times 4 \times 15$	0.4696	86.00 %						
		$3 \times 3 \times 15$	0.6053	94.00 %						
	Par.boot		0.6221	94.67 %						

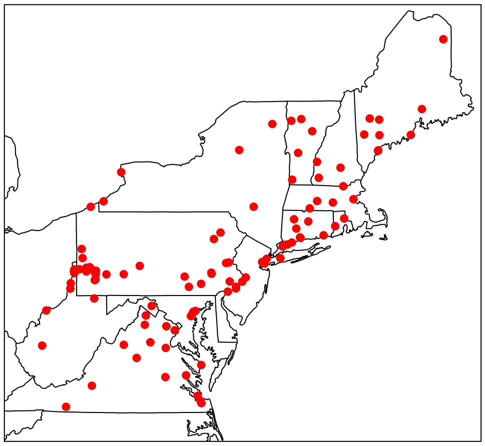
Figure 3.3: Normal QQ-plots of the standardized estimates of $\hat{\beta}$ by JCEF, fixing other parameters. Observed quantiles are ordered $(\hat{\beta} - \beta)/SE(\hat{\beta})$, based on standard error estimates from parametric bootstrap.



3.5 Analysis of Particulate Matter Data

We analyze 20-year PM10 data across the northeastern United States from August 1982 to August 2002. The goal is to study the spatio-temporal dependence structure of air pollutant PM10 so that predictions can be made at specific locations and time points. Monthly mean PM10 measures are obtained by averaging all available readings for a given month and are log transformed. Because not all monitors are observed all the time, we use 108 monitors with consecutive monthly records between January 2000 and August 2002 (32 months) in this area for an illustration of the JCEF method. A layout of the monitor locations is displayed in Figure 3.4. The distance between two monitor locations ranges from 0.45 to 956 miles.

Figure 3.4: Layout of PM10 monitor stations for Northeastern United States from January 2000 to August 2002.



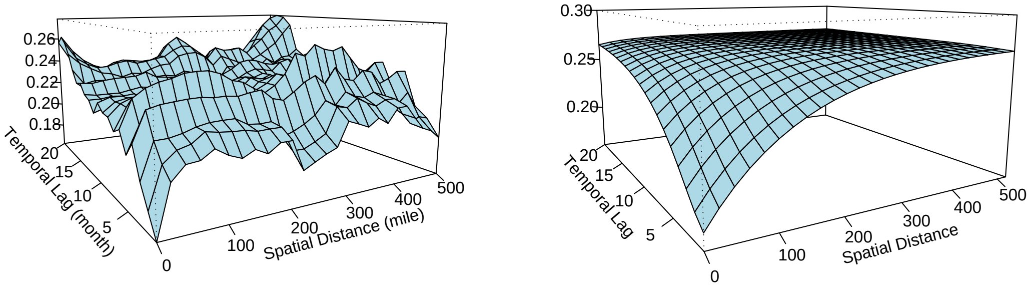
Northeastern United States

We first remove month and location effects by an ANOVA model treating each month and location as class variables (Diez Roux et al., 2008), and then use the resulting residuals to estimate the spatio-temporal dependence structure. To visualize the spatio-temporal pattern, we plot the estimated spatio-temporal empirical variogram in Figure 3.5(left). Observation pairs are grouped by distance lags of 20 to 500 miles with unit increase of 20 miles and temporal lags of one to 20 months, with unit increase of one month.

We fit the nonseparable covariance structure in equation (3.4) with a nugget effect of variance σ_ϵ^2 to the data. A set of initial parameter values is obtained using WLS by minimizing the weighted difference of the empirical variogram to the parametric variogram at pre-specified lags.

As pointed out previously, subsampling may not be appropriate for the irreg-

Figure 3.5: Empirical and fitted spatio-temporal variogram for PM10 residuals. Observation pairs are grouped by distance lags of 20 to 500 miles with unit increase of 20 miles and temporal lags of one to 20 months, with unit increase of one month.



ular spatial monitor grid, we use parametric bootstrap to create sample replicates for the subsequent determination of the optimal distance lag, the weight matrix estimation and standard error estimation.

To determine the optimal distance lags, it is computationally prohibitive in practice to compute the Godambe information for all possible combinations of spatial and temporal lags. We use the grid search method to find the optimal lags from a pool of spatial and temporal lags with time ranging from 1 to 6 months with one month increment and spatial distances ranging from 20 to 260 miles with 20 miles increment. The optimal combination is 6 months in time and 100 miles in space, which means we will include pairs that are within $p = 100$ miles in distance and $q = 6$ months in time to specify our composite estimating functions. Then WCL is carried out for estimation and its estimates are used for weight matrix calculation. Finally, the JCEF method is applied to estimate the model parameters.

Table 3.3: Parameter and standard error estimates of the spatio-temporal covariance structure in equation (3.4) fitted to the PM10 data set. Standard error estimates are obtained by parametric bootstrap. CI denotes for confidence interval.

	WCL			JCEF		
Parameter	Estimate	95% CI		Estimate	95% CI	
a	1.0112	0.6213	1.5048	1.1636	0.7833	1.7285
b	0.0382	0.0148	0.0981	0.0403	0.0173	0.0939
β	4.1129	0.9423	20.7327	6.4341	1.7373	23.8292
σ^2	0.0219	0.0173	0.0265	0.0224	0.0180	0.0277
σ_ϵ^2	0.0194	0.0167	0.0231	0.0199	0.0173	0.0229

Parameter estimates, standard errors and 95% confidence intervals from JCEF and WCL are listed in Table 3.3. Point estimates from the two methods are similar, but JCEF yields smaller standard error estimates, especially for the interaction parameter β, which is consistent with the simulation results. For the JCEF method, given $\hat{\beta} = 6.4341$, $\hat{a} = 1.1636$ means that the marginal temporal correlation decays by around 40% with one month increase in time, and $\hat{b} = 0.0403$ indicates that the marginal spatial correlation decays by approximately 15% with 10 miles increase in space. $\hat{\beta} = 6.4341$ indicates that the temporal correlation decays approximately 1.5% faster with 10 miles further away in space, while the spatial correlation decays about 2.5% faster with 1 month further apart in time. The confidence interval for $\hat{\beta}$ does not cover one, indicating that there is a significant spatio-temporal interaction effect. As a result, a separable covariance strucutre is not applicable to this data set if the covariance function in equation (3.4) is used to model the dependence structure.

We also compare the sums of squared differences (SSD) between the fitted parametric variograms obtained by JCEF (Figure 3.5, right) and WCL to the empirical variogram at grid points at which the empirical variogram is computed. The SSD ratio of JCEF over WCL is 0.67, indicating that the fitted surface by JCEF is 33% closer to the empirical one than that by WCL. In summary, the proposed JCEF outperforms WCL in point estimates, standard error estimates, as well as the goodness-of-fit. Some additional analysis of the data may be carried out. For example, one could use tests proposed in Li et al. (2007) to test the symmetry and isotropy of the data dependence and fit the corresponding parametric covariance function to the data to improve overall model fit. It is worth further investigation.

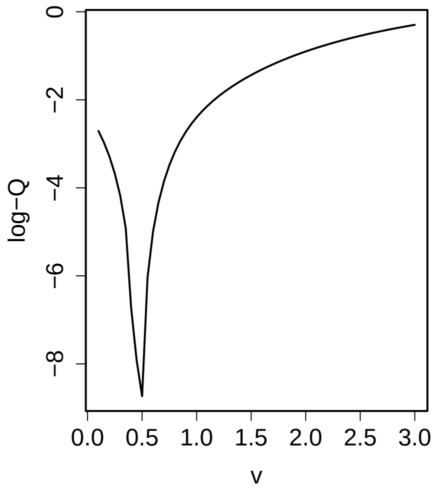
3.6 Discussion

In this chapter, we have proposed a statistically efficient and computationally feasible approach to estimating spatio-temporal covariance models for massive data sets. The proposed JCEF method constructs separate composite likelihoods based on spatial, temporal and spatio-temporal cross pairs, and then join them into a quadratic inference function. Through such GMM formulation, our method accounts for correlations among the pairs via the weight matrix and allocates higher weights to groups of pairs with more information, hence substantially improves the estimation efficiency over existing WCL methods. The JCEF estimator has also proven to be consistent and asymptotically Gaussian under the increasing-domain asymptotics. Comprehensive simulation studies have shown that JCEF recovers significant amount of efficiency from the conventional weighted composite likelihood approach.

Another advantage of JCEF is the possibility of deriving a goodness-of-fit statistic to test the mean-zero model assumption, $H_0 : E\{\Gamma_n(\boldsymbol{\theta})\} = 0$. This can be used for testing the separability structure of the covariance matrix. Since $\hat{\boldsymbol{\theta}}_n$ is obtained by an over-identified estimating function $\Gamma_n(\boldsymbol{\theta})$, $Q_n(\hat{\boldsymbol{\theta}}_n)$ falls in the 'over-identifying restriction' test by Hansen (1982). He proved that the asymptotic distribution of $Q_n(\hat{\boldsymbol{\theta}}_n)$ is χ^2 with degrees of freedom equal to the number of estimating functions minus the number of parameters, which in our case is $2r$. However, many researchers have pointed out that the first-order asymptotic theory often provides inadequate approximations to the distributions of the test statistics obtained from GMM estimators; see, for example, a special issue of the Journal of Business & Economics Statistics (July 1996). To improve inference, a number of alternative estimators have been suggested. These include empirical likelihood (Qin and Lawless, 1994; Owen, 1988; Imbens, 1997), modified bootstrap procedures (Hall and Horowitz, 1996), and the continuous updating estimator (Hansen et al., 1996). Qu et al. (2000) used the latter approach to construct the QIF and showed that the finite-sample distribution of the objective function agrees well with the asymptotic counterpart. Performances of these goodness-of-fit methods under the JCEF framework for spatio-temporal data is worth further exploration.

As noted previously, the smoothness parameter ν is usually difficult to estimate. However, the quadratic objective function in JCEF provides an analogy to the profiled likelihood for estimating ν. Specifically, given a range of ν values, we perform the JCEF estimation procedure for each ν, and record the parameter values and the target function value. Then ν is estimated to be the one with the smallest target function value, and the corresponding parameter estimates are used as the final estimates. We plot log-$Q(\boldsymbol{\theta})$ and ν in Figure 3.6 for simulation setup 7 considered in Table 3.1. The true value for ν is 0.5. We can see that this profile approach provides an accurate estimate of ν, and may be a very promising method. Further detailed work is needed for this development.

Figure 3.6: Estimated log-$Q(\hat{\boldsymbol{\theta}})$ against smoothness parameter ν, evaluated in setup 7 in Table 3.1. The true value for ν is 0.5.



We have considered covariance estimation from a detrended process. As known, detrending may introduce artificial correlation into the residuals, which may distort the intrinsic correlations of the data. In fact, this is a common concern when a two-stage procedure is used to estimate covariance structures. A simple solution would be to jointly estimate mean and covariance parameters. From a large-sample point of view, as long as mean parameters are consistently estimated, covariance estimates can be consistent under some mild conditions. On the other hand, in actual applications, finite sample performances matter more. Our experiments in both theory and computation suggested that two factors are crucial to ensure similar performances between the two-stage approach and the joint estimation method: (i) the strength of the intrinsic spatio-temporal dependency and (ii) the sample size. For large data sets, these two factors are usually in favor of the two-stage procedure.

It is worth noting that variance estimates of the JCEF estimator do not account for the uncertainty in the weight matrix estimation. This pertains to the uncertainty resulted from the plugged-in parameter estimates in the evaluation of the weight matrix. According to Windmeijer (2005), such variation is known to be of order $O(n^{-1})$, which is a lower-order term than $O(n^{-1/2})$ and thus may be ignorable when n is large. In addition, this issue concerning the finite-sample performance of the JCEF estimator has been well studied in the GMM literature. Several methods have been proposed to correct for the downward bias in parameter standard error estimates when the sample size is inadequate. This includes adding a variance correction term (Windmeijer, 2005), or using a parametric bootstrap procedure (Hall and Horowitz, 1996) to account for the uncertainty in the weight matrix estimation.

We focus our attention on evaluating the efficiency gain of JCEF over the existing methods in terms of covariance estimation in this paper. Kriging, one of the popular approaches used for prediction in Geostatistics, relies heavily on covariance functions, as the kriging predictor is the best linear unbiased estimator based on the covariance model specified for the process. It may also be interesting to study whether more efficient covariance estimators will yield more efficient predictors.

CHAPTER IV

Estimation Methods for High-Dimensional Space-Time Covariances

4.1 Introduction

In chapter III, a new estimating procedure called Joint Composite Estimating Function (JCEF) has been proposed to estimate high-dimensional spatio-temporal covariance functions. It significantly improves estimating efficiency over conventional weighted pairwise marginal composite likelihood methods, and has desirable large-sample statistical properties. In this chapter, we investigate the relative performance of JCEF in a wider context, comparing it to many other popular estimating approaches proposed in the literature.

Thanks to the technology advancements in all fields of modern sciences, e.g. remote satellite sensing in environmental sciences, functional-MRI in medical studies, and the next-generation sequencing in bioinformatics etc., data has never been more accessible to researchers than what can be available today. Statistics has been greatly challenged by the need to explore such massive data sets. In recent decades, there has been a spawn of statistical research in the development of computationally feasible methods for analyzing large-scale data sets. The focus of this chapter will be dwelt on the spatio-temporal data arising frequently in environmental sciences, where millions of observations can be instantaneously collected over a large number of spatial locations.

The difficulty in estimating dependence structures for massive data has long been recognized in spatial statistics. Two types of approaches have been developed to facilitate computations. The first approach is based on simplifying covariance structures. For stationary spatial processes on regular grids, Zimmerman (1989) showed that covariance structures of those processes possess patterned structures that could be utilized to reduce the computational burden. Cressie and Jahannesson (2008) proposed fixed rank kriging for very large spatial data sets, where the covariance matrices were specially designed so that the matrix manipulations were of a fixed magnitude. A similar idea was exploited in Banerjee et al. (2008). However, these approaches either require the spatial processes to be stationary, or impose over-simplified structures for the covariance matrices, hence may not be well generalized to real data analysis.

Another approach is based on likelihood approximations, where simplified versions of the full likelihood are considered. For example, composite likelihood (CL) methods (Lindsay, 1988) have been proposed to model spatial data. As a general class of pseudo-likelihoods, composite likelihood is based on valid marginal or conditional likelihood functions. Curriero and Lele (1999); Heagerty and Lele (1998); Li and Lin (2006) all used pairwise marginal densities to build composite likelihood estimation functions, while Vecchia (1988) and Stein et al. (2004) suggested approximating the likelihood by a product of conditional densities with truncated conditioning sets. Apart from composite likelihood approaches, Furrer et al. (2006) and Kaufman et al. (2008) used covariance tapering method to shrink small values of covariance entries to zero, so that the sparse matrix algorithm could be used to speed up computation. Fuentes (2007) proposed an approximation by modeling the covariance structures in the spectral domain, which appears to be more involved and hence is of less popularity in application.

Additional challenges arise in spatio-temporal settings. With the addition of a time domain, data scale is much larger. Also, the distinct yet intricately involved nature of the space and time further complicates the data analysis. To simplify covariance structures, people usually separately model spatial and temporal dependencies (Sahu et al., 2007; Smith and Kolenikov, 2003) or to apply a separable spatio-temporal covariance function for the ease of computation (Haas, 1995; Genton, 2007). Although these and other similar approaches have many desirable properties, they all ignore a crucial model component: the spatio-temporal interaction effect.

The objective of this chapter is to provide numerical evidences on why JCEF proposed in chapter III is more appropriate in spatio-temporal data analysis than other available methods. We use pairwise marginal densities as the building blocks of the estimating function due to the following considerations:

(i) Pairwise CL is both analytically and numerically simple to work with.

(ii) It only requires the correct specification of bivariate densities, hence the resulting estimation and inference are robust to the misspecification of high-dimensional moment structures (Varin et al., 2011). In contrast, conditional CL approaches (e.g. Vecchia, 1988; Stein et al., 2004) are usually vulnerable to model misspecification, as they require the formulation of higher-dimensional distributions. In addition, it is easier to check assumptions on the bivariate distribution than on the high-dimensional multivariate distribution.

(iii) The pairwise CL approach does not require a distance metric accommodating both space and time, while a unified distance norm is needed by the tapering approach (Kaufman et al., 2008). As we will see in the simulation experiments (see section 4.3.2), tapering expedites computing time only when the number of non-zero covariance elements is small.

Thus, the pairwise composite likelihood seems appealing in modeling large-scale spatio-temporal data, for its simplicity, flexibility and feasibility in statistical inference and numerical computation. The rest of chapter is organized as follows. A review of some of the popular approaches in covariance estimation is given in section 4.2. Extensive simulation experiments are carried out to compare the relative performances of the methods in section 4.3, followed by a discussion section.

4.2 A Review of Covariance Estimation Methods

4.2.1 Fixed Rank Kriging

To reduce the computational intensity of handling large matrices, Cressie and Jahannesson (2008) proposed the fixed rank kriging technique based on a reduced rank approximation of the underlying process. The prediction is based on a family of covariance functions constructed using a set of basis functions that is fixed in number. Essentially, the covariance of the spatial process is assumed to be generated from a hidden process of a fixed dimension, so that one only needs to invert the kernel covariance matrix of the hidden process. Banerjee et al. (2008) proposed a similar method in the Bayesian framework, and termed it as the predictive process. I briefly introduce the key idea in the spatial context.

Let $Y(s), s = 1, \ldots, n$ be a pure spatial process in $\mathcal{S}$, write

$$Y(s) = \mathbf{S}(s)'\boldsymbol{\eta} + \epsilon(s), \tag{4.1}$$

where $\epsilon(s)$ is the independent error term, and

$$\mathbf{S}(s) = (S_1(s), \ldots, S_r(s))',$$

is a set of r basis functions. $\boldsymbol{\eta}$ is an r-dimensional vector with $Var(\boldsymbol{\eta}) = \mathbf{K}$. Both $\mathbf{S}(s)$ and $\mathbf{K}$ are assumed to be known. The model in equation (4.1) may be regarded as a spatial random-effects model. When a deterministic mean model is included, equation (4.1) becomes a mixed effects model. Now let $\mathbf{S}$ be an $n \times r$ matrix whose (i, l) element is $S_l(s_i)$, then the covariance of $\mathbf{Y} = (Y(1), \ldots, Y(n))$ is

$$\Sigma = \mathbf{SKS}' + \sigma^2 \mathbf{V}_n, \tag{4.2}$$

Where $\mathbf{V}_n$ is an $n \times n$ diagonal matrix with entries given by the measurement error variance, and are assumed known. Kriging requires the inversion of a generic covariance of $\mathbf{Y}$, whose computational complexity is of order $O(n^3)$. Given the specific expression of (4.2), the complexity is reduced to $O(r^3)$, with r being a fixed constant given in equation (4.1). A fast computation is achieved by applying the following Sherman-Morrison-Woodbury formulae (Henderson and Searle, 1981) to equation (4.2).

$$(\mathbf{I}+\mathbf{PAP}')^{-1} = \mathbf{I} - \mathbf{P}(\mathbf{A}^{-1}+\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}',$$

where **P** and **A** are matrices of appropriate dimensions.

Many research studies have successfully used the fixed rank kriging to capture the large scale structure of spatial processes. However, it is usually inadequate in capturing the local and small-scale dependence structure (Stein, 2008; Finley et al., 2009). To solve this problem, Sang and Huang (2011) proposed a full-scale approximation method, in which the residual covariance that is not accounted for by the fixed-rank process is modeled and estimated through tapering (Furrer et al., 2006; Kaufman et al., 2008).

4.2.2 Spectral Methods

Another approach to reduce computation of the full likelihood is through the spectral decomposition of the covariances. For observations on a regular complete lattice, the resulting spectral density can be conveniently estimated by periodogram, and Whittle's approximation can be used to compute the likelihood (Whittle, 1954). Guyon (1982) and Stein (1995, 1999) applied the spectral methods to study stationary spatial process on regular grids without missing data. Fuentes (2002) extended the spectral methods to nonstationary spatial processes, and Fuentes (2007) adapted the procedure to irregularly spaced spatial data. Following notations in Fuentes (2007), I introduce the spectral decomposition of the covariance functions and the likelihood approximation method as follows.

For two observations from a stationary spatial process, the spatial covariance function C depends on the relative configuration of two locations $\mathbf{s}_1$ and $\mathbf{s}_2$ in $\mathcal{R}^2$.

$$C(\mathbf{s}_1 - \mathbf{s}_2) = cov(Y(\mathbf{s}_1), Y(\mathbf{s}_2))$$

The spectral density function f is the Fourier transform of the covariance function:

$$f(\omega) = \frac{1}{(2\pi)^2} \int_{\mathcal{R}^2} \exp(-i\mathbf{s}^T)C(\mathbf{s})d\mathbf{s},$$

where i is the complex unit.

Furthermore, if Y is observed only at n uniformly spaced spatial locations Δ units apart, then the spectrum of observations of the sample sequence $Y(\Delta\mathbf{s})$, for $\mathbf{s} \in \mathcal{Z}^2$, is concentrated within the finite-frequency band $-\pi/\Delta \geq \omega \leq \pi/\Delta$. The spectral density of f_Δ of the process on the lattice is written as

$$f_\Delta(\omega) = \sum_{Q \in \mathcal{Z}^2} f\left(\omega + \frac{2\pi Q}{\Delta}\right),$$

where $\omega \in [-\pi/\Delta, \pi/\Delta]^2$.

The spectral density of a lattice process, observed on an $n_1 \times n_2$ grid, where $n = n_1 n_2$ can be estimated by the periodogram,

$$I_n(\omega) = (2\pi)^{-2}(n_1 n_2)^{-1} \left| \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} Y(\mathbf{s}_{ij}) \exp\left\{-i\mathbf{s}_{ij}^T \omega\right\} \right|^2,$$

where $| \bullet |$ is the norm for a complex number.

The above equation is evaluated in the set of Fourier frequencies $\mathbf{\Omega} = 2\pi(f_1/n_1, f_2/n_2)$, where

$$(f_1, f_2) = \left\{ \left\lfloor -\frac{n_1 - 1}{2} \right\rfloor, \ldots, n_1 - \left\lfloor \frac{n_1}{2} \right\rfloor \right\} \times \left\{ \left\lfloor -\frac{n_2 - 1}{2} \right\rfloor, \ldots, n_2 - \left\lfloor \frac{n_2}{2} \right\rfloor \right\}.$$

Moreover, according to Whittle (1954), if $Y(\mathbf{s})$ follows a Gaussian process with mean zero, the negative log-likelihood can be approximated as:

$$\frac{n}{(2\pi)^2} \sum_{\omega \in \mathbf{\Omega}} \log f(\omega) + I_n(\omega) f(\omega)^{-1},$$

where f is the spectral density of the lattice process. The approximated likelihood can be calculated very efficiently using the fast Fourier transformation, which requires only $O(n \log_2 n)$ operations.

Spectral methods are less intuitively appealing than other methods. It is more involved to generalize the method to nonstationary and irregular spatial data. How this type of approach can be used to facilitate estimation of spatio-temporal covariance functions has not yet been addressed in the literature.

4.2.3 Pseudo-Conditional Likelihoods

Besides the marginal composite likelihood methods, another form of compositing the likelihood is through conditional density functions (see Chapter II for details). Vecchia (1988) proposed to approximate the full likelihood through products of conditional density functions, and truncate the conditioning sets to avoid large matrix manipulations. This method was later extended by Stein et al. (2004) to include distant observations in the conditioning sets.

Specifically, partition a random vector $\mathbf{Y}$ into subvectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_b$ of possibly different lengths, and denote $\mathbf{Y}_{(j)} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_j), j \geq 2$. Then, the likelihood can be written as

$$f(\mathbf{Y}; \boldsymbol{\theta}) = f(\mathbf{Y}_1; \boldsymbol{\theta}) \prod_{j=2}^{b} f(\mathbf{Y}_j | \mathbf{Y}_{(j-1)}; \boldsymbol{\theta}).$$

To reduce the dimension of the conditioning sets, let $\mathbf{S}_{(j)}$ be some subvector of $\mathbf{Y}_{(j)}$, then

$$f(\mathbf{Y}; \boldsymbol{\theta}) \approx f(\mathbf{Y}_1; \boldsymbol{\theta}) \prod_{j=2}^{b} f(\mathbf{Y}_j | \mathbf{S}_{(j-1)}; \boldsymbol{\theta}),$$

where $\mathbf{Y}_j$ is called the prediction set and $\mathbf{S}_j$ the conditioning set.

Choosing $\mathbf{S}_j$ to be of smaller sizes reduces computations, however two problems are associated with this decomposition. First, a multivariate distribution is required, which is not readily available for non-Gaussian data. Also this approach is more susceptible to model miss-specification, since it requires the correct specification of high-order distribution structures. It is also a difficult task to check the model assumption for multivariate distributions of more than two dimensions. In these regards, pairwise composite likelihood has clear advantages. Second, the choice of the conditioning sets is tricky and usually requires an overall distance metric of space and time.

Nevertheless, Stein's method is representative of the conditional formulation within the composite likelihood framework and is closely connected to Gaussian Markov random fields. We also include it in our simulation experiments.

4.2.4 Covariance Tapering

Tapering (Furrer et al., 2006; Kaufman et al., 2008) is becoming increasingly popular in spatial statistics due to its simplicity both in concept and in implementation. The idea is to set certain elements of the covariance matrix to 0, such that the resulting matrix is sparse and positive definite, and retains the original properties for proximate locations. Specifically, let $C(h; \boldsymbol{\theta})$ be the covariance function for two observations with distance h in space, and $K_{taper}(h; \eta)$ be the tapering function that is identically 0 whenever $h \geq \eta$, where η is a pre-specified cutoff. Then the tapered covariance function is given by:

$$C_{taper}(h; \boldsymbol{\theta}) = C(h; \boldsymbol{\theta}) K_{taper}(h; \eta).$$

In our spatio-temporal setting, applying the tapering technique requires the specification of a joint distance metric accommodating both space and time coordinates. This is generally difficult, as space and time are distinct with respect to distance. Nevertheless, for the pure simulation purpose, we use the Euclidean distance norm on standardized spatial and temporal coordinates. Note that MLE is a special case of the tapering method when the taper range η is set at infinity.

Besides the requirement of a join spatio-temporal distance metric, tapering has some other limitations that have not yet been addressed. For example, the class of tapering functions depends only on distances between observations, hence is stationary in nature. It may not be appropriate to taper a nonstationary covariance matrix with a stationary tapering function. Constructing more flexible tapering functions that accommodate non-stationarity, anisotropy and other potential spatial characteristics can better preserve the original covariance structures and greatly enhance the performance of the tapering method.

4.3 Simulation Experiments

To assess the performance of the JCEF method, we conduct simulation experiments to compare it to some of the available methods in the literature, including (i) weighted composite likelihood (WCL), the current available CL approach, (ii) the tapering method (taper) based on covariance regularization (Kaufman et al., 2008), (iii) conditional pseudo-likelihood methods (Stein) proposed in Stein et al. (2004) and Vecchia (1988), (iv) weighted least square approach (WLS), the one used most often by practitioners in spatial statistics (Cressie, 1993), and (v) maximum likelihood estimates (MLE), the golden standard.

We compare their performances in terms of mean squared errors (MSE) of parameter estimates. We also scale parameter-specific MSEs by their corresponding parameter values and sum them together to obtain an overall efficiency measure, called *total scaled MSE*. This scaling balances different scales of parameter values, so that a fair comparison can be made. Relative efficiency (RE) is then computed as the ratio of the total scaled MSEs between two methods under comparison. All simulations are coded in R 2.11.1 (R Development Core Team, 2010) and executed on a Linux cluster with Intel Xeon X5680 processors (3.33 GHz CPU and 1.5G memory for each of 16 nodes).

The spatio-temporal covariance function used in the data generation is a non-separable spatio-temporal covariance structure proposed in Cressie and Huang

(1999):

$$C(h,u;\boldsymbol{\theta}) = \begin{cases} \frac{\sigma^2(2\beta)}{(a^2u^2+1)^{\nu}(a^2u^2+\beta)\Gamma(\nu)} \left\{ \frac{b}{2} \left(\frac{a^2u^2+1}{a^2u^2+\beta}\right)^{\frac{1}{2}} h \right\}^{\nu} K_{\nu}\left(b \left(\frac{a^2u^2+1}{a^2u^2+\beta}\right)^{\frac{1}{2}} h \right), & \text{if } h > 0, \\ \frac{\sigma^2(2\beta)}{(a^2u^2+1)^{\nu}(a^2u^2+\beta)}, & \text{if } h = 0, \end{cases} \tag{4.3}$$

where $u = |t_1 - t_2|$ is the time lag and $h = ||s_1 - s_2||$ is the Euclidean distance between two locations. K_ν is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun (1972), p.374), where $\nu > 0$ is a smoothness parameter characterizing the behavior of the correlation function near the origin. If $u = 0$, $C(h, 0; \boldsymbol{\theta})$ degenerates into a purely spatial covariance, which is the popular Matèrn class used in spatial statistics. When $\nu = 0.5$, this spatial correlation is an exponential function of h, when $\nu \to \infty$, the Gaussian correlation function. In practice, ν is difficult to estimate, because it requires dense space data and may run into identifiability problem (Stein, 1999). We will discuss a profile quadratic inference function approach for estimating ν in the discussion section.

For the rest of the parameters, $a \geqslant 0$ is the scaling parameter of time, $b \geqslant 0$ is the scaling parameter of space, $\beta > 0$ is a space-time interaction parameter, and $\sigma^2 = C(0,0) > 0$, the variance at the origin. Note that a separable covariance function is obtained when $\beta = 1$. We also study the presence of a nugget effect in our simulation comparison, and denote its variance as σ^2_ϵ. As a result, the parameter vector of interests is $\boldsymbol{\theta} \equiv (a, b, \beta, \sigma^2, \sigma^2_\epsilon)$.

4.3.1 Comparison to Weighted Composite Likelihood

We first compare JCEF with WCL. We form our composite estimating functions based on neighboring pairs for both WCL and JCEF, following the suggestion given in Bevilacqua et al. (2010). It is noted that tuning the distance lag according to a certain optimality criterion (e.g. minimizing the trace of the inverse of the Godambe information) for each specific case can result in better efficiency. However, using a common distance lag in the simulation study serves the purpose of comparison, and the related computational burden appears manageable.

We generate $X(s,t)$ on a regular grid of $7 \times 7 \times 30$ space-time points, with spatial coordinates being set at $(1, 1.5, \ldots, 4) \times (1, 1.5, \ldots, 4)$ and $\mathcal{T} = (1, 2, \ldots, 30)$. Table 4.1 includes three simulation setups. We vary β values from 0.5, 1 to 5, corresponding to negative, none and positive spatio-temporal interaction effect, respectively. Each column-wise plot in Figure 4.1 shows the marginal spatial and temporal correlation patterns, respectively. It is clear that the decay rate of spatial or temporal correlation given different temporal or spatial lags changes with different β values. Parameter ν is fixed at 0.5 in the simulation.

Estimation of the weight matrix is achieved by sub-group sampling on overlapping sub-blocks of size $4 \times 4 \times 15$, following the rule suggested in Politis and Romano (1994). We use estimates from WCL to evaluate individual score functions in each sub-blocks. A total of 200 simulated datasets are generated for each setup.

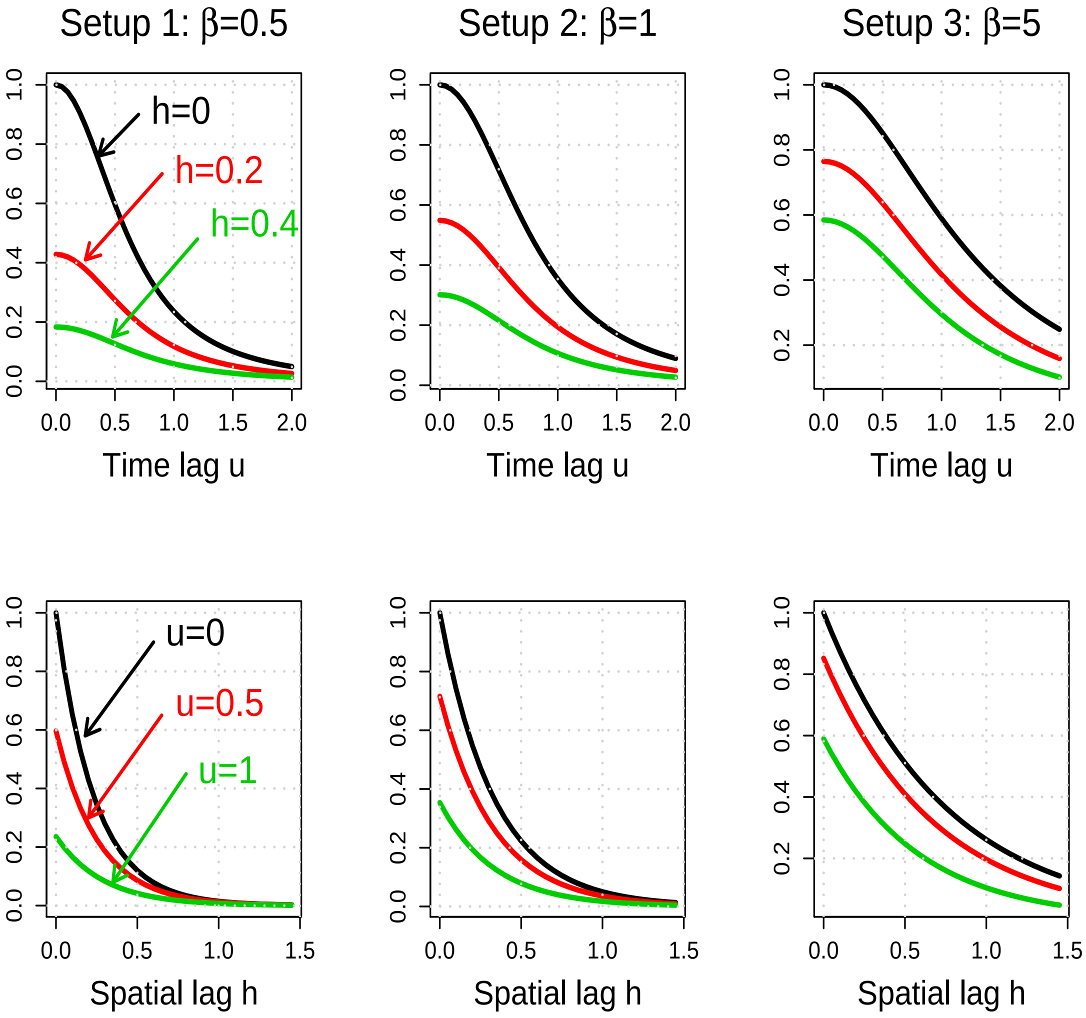
We first compare JCEF and WCL in the presence of a nugget effect σ_ϵ^2. The results, summarized in Table 4.1, show that the JCEF method clearly outperforms WCL in all three simulation setups in terms of the total scaled MSE. The resulting REs show that, for all three scenarios, JCEF clearly reaches 25% or higher efficiency improvement compared to WCL. Unscaled parameter-specific MSEs indicate that on average, approximate 10% reduction in MSE is achieved for parameter a, b, β and σ_ϵ^2.

Table 4.1: (With Nugget) Mean squared errors (MSE) of parameter estimates. Results are from 200 rounds of simulations based on the covariance structure in equation (3.4) and a nugget effect σ_ϵ^2. Total scaled MSE is the sum of MSEs for four parameters scaled by parameter means. RE is the relative efficiency defined as the total scaled MSE of WCL over that of JCEF.

Scenarios	Method	Mean Squared Errors					Total Scaled MSE	RE
		a	b	β	σ^2	σ_ϵ^2		
Setup 1		1	3	0.5	1	0.5		
	WCL	0.0122	0.2915	0.0060	0.0060	0.0039	0.0901	
	JCEF	0.0122	0.2651	0.0039	0.0051	0.0025	0.0724	1.25
Setup 2		1	3	1	1	0.5		
	WCL	0.0086	0.1492	0.0186	0.0047	0.0027	0.0594	
	JCEF	0.0070	0.1376	0.0107	0.0051	0.0016	0.0447	1.33
Setup 3		1	3	5	1	0.5		
	WCL	0.0078	0.1593	0.3855	0.0133	0.0014	0.0599	
	JCEF	0.0074	0.1065	0.2576	0.0125	0.0013	0.0471	1.27
Average MSE Reduction		12.24%	9.09%	11.04%	3.76%	10.59%		

We then compare the two methods without the nugget effect in the covariance structures. Similar summary statistics are listed in Table 4.2, part of which is reported in Table 3.1. It appears that in this case, JCEF gains even more efficiency for a, b and β. On average, the MSE reduction is 40.5% for β, followed by 26.1% for a (the temporal scaling parameter), and then 13.5% for b (the spatial scaling parameter). The estimates for the variance parameter σ^2 are comparable between the two

Figure 4.1: Plot of $C(h, u)$ in equation (3.4). Each column of the plot corresponds to spatial and temporal correlation patterns for simulation setups 1-3 considered in Table 4.2. Parameter ν is fixed at 0.5.



methods. The significant efficiency improvement for β, a and b are very desirable, since these are important parameters pertaining to the dependence structure. In addition, for the interaction parameter β, valid parameter and standard error estimates will help researchers make infernece about whether a simpler, separable spatio-temporal covariance is supported by data.

Table 4.2: (Without Nugget) Mean squared errors (MSE) of parameter estimates. Results are from 200 rounds of simulations based on the covariance structure in equation (3.4). Total scaled MSE is the sum of MSEs for four parameters scaled by parameter means. RE is the relative efficiency defined as the total scaled MSE of WCL over that of JCEF.

Scenarios	Method	Mean Squared Errors				Total Scaled MSE	RE
		a	b	β	σ^2		
Setup 1		1	3	0.5	1		
	WCL	0.0047	0.1046	0.0030	0.0018	0.11	
	JCEF	0.0037	0.0703	0.0023	0.0019	0.08	1.46
Setup 2		1	3	1	1		
	WCL	0.0029	0.0599	0.0073	0.0028	0.07	
	JCEF	0.0022	0.0527	0.0037	0.0027	0.06	1.19
Setup 3		1	3	5	1		
	WCL	0.0031	0.1157	0.2439	0.0073	0.37	
	JCEF	0.0025	0.1080	0.1288	0.0072	0.25	1.50
Average MSE Reduction		22.38%	17.17%	40.18%	-1.08%		

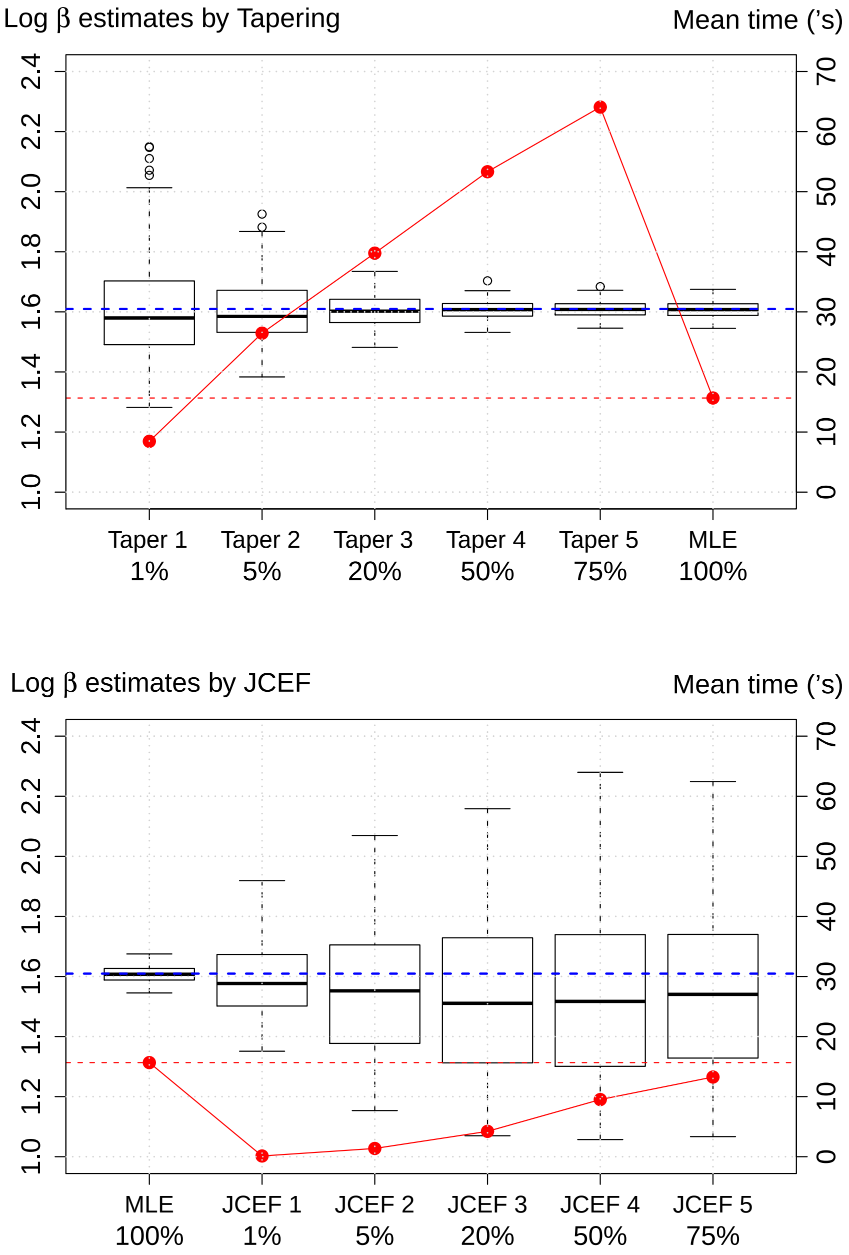
4.3.2 Comparison to Tapering

We compare the tapering method and JCEF with varying distance lags and taper ranges. For each combination of the spatial and temporal lags (p,q) used for JCEF, we select an appropriate taper range, so that the same pairs of observations are included in the latter method. Given that each pair of observations corresponds to two entries in the full covariance matrix, we quantify the shared amount of information by both methods in terms of the percent of covariance elements utilized in estimation for each set of spatial and temporal lags (p, q) and the respective taper range η. These percentages are marked below the horizontal axis label in Figure 4.2, where boxplots of estimates of $log(\beta)$ and the averaged computing times (red-dotted line for MLE, and solid line for the specific method) are presented. Data are generated based on setup 3 considered in Table 4.2.

In terms of parameter estimates, boxplots in the bottom panel of Figure 4.2 show that for JCEF, increasing spatial and temporal lags does not improve the estimates, consistent to findings in the current literature (e.g. Varin et al., 2005; Davis and Yau, 2011). This is because, pairs further apart are less likely to be correlated, hence contain little information about the dependence. Including them in the estimation will add more noise in the estimation of covariance structures.

On the contrary, boxplots in the top panel of Figure 4.2 show that, increasing the taper range from nearing neighbors (1%) to the maximum distance (100%), leads to improved estimation. This is because tapering works on the covariance matrix. Extending the taper range increases the non-zero covariance elements used in estimation, which in turn brings in high-order correlations among the included covariance elements, leading to efficiency gain. This explanation does not apply to pairwise CL methods, since no high-order correlations are contained in pairs. However, this weakness is overcome, to some extent, by JCEF in which the weight matrix effectively accounts for some of the correlations beyond pariwise dependences. Obviously, WCL does not incorporate such high-order correlation information, so it is less efficient than JCEF, as already shown in chapter II.

Figure 4.2: Boxplots of $log(\beta)$ estimates from tapering (top panel) and JCEF (bottom panel) for setup 3 considered in Table 4.2 with a spatial grid of 7 by 7 and 30 time points. Five sets of spatial and temporal lag combination (p, q) (with increasing values) are considered for JCEF, corresponding to JCEF 1 - JCEF 5. The percentages of information utilized by each (p, q) are marked below the horizontal axis label, ranging from 1% to 75%. Five taper ranges η are chosen respective to each (p, q), and are labeled as Taper 1 - Taper 5. The same percentages are marked for tapering accordingly. MLE is the special case when $p = q = \infty$ for JCEF and $\eta = \infty$ for tapering. Red dotted lines are the corresponding mean computing time (in seconds). Red dashed line indicates the mean time used by MLE. Blue dashed line indicates the true $log(\beta)$ value.



In terms of computing time needed for the optimization to converge, the red-solid line in the top panel of Figure 4.2 shows that, tapering requires much longer time than MLE (red-dotted line) as the taper range increases. Tapering is faster when only 1% of the covariance elements (nearing neighbors) are used in estimation. Note that we use the R code posted on http://www.image.ucar.edu/Data /precip_tapering/ for executing the tapering method (with minor changes), which is the same code used by Kaufman et al. (2008). So the comparison of computing time is based on the same sparse matrix algorithm. What makes tapering run slow may due to the time spent in indexing and retrieving non-zero entries, which can be a substantial workload for a larger taper range. Figure 4.2 clearly indicates that tapering is only competitive when the taper range is small. However, in this case, JCEF is superior to tapering in both estimation and computational efficiency.

4.3.3 Comparison to WLS and MLE

We now include WLS and MLE in the comparison. WLS is probably the most commonly used method in spatial data analysis. It estimates dependence parameters by fitting a parametric covariance function to the computed empirical spatio-temporal variogram. As already shown in Lele and Taper (2002), WLS is less efficient than WCL, which as we have shown is less efficient than JCEF.

We use setup 3 considered in Table 4.2 for comparing the five methods. Table 4.3 lists results of β estimates from three increasing grids. In particular, we choose the taper range so that it is computationally competitive to MLE. Then the spatial and temporal lags in JCEF are set at values comparable to the tapering method.

Table 4.3: A comparison of MSEs and computing time for MLE, JCEF, WCL, Tapering and WLS for setup 3 considered in Table 4.2, based on 200 rounds of simulations. Data are generated from three increasing grids of $5 \times 5 \times 15$, $6 \times 6 \times 20$, and $7 \times 7 \times 30$.

	$5 \times 5 \times 15$		$6 \times 6 \times 20$		$7 \times 7 \times 30$	
	MSE	Time	MSE	Time	MSE	Time
MLE	0.09	2.61	0.03	4.11	0.02	15.66
JCEF	1.62	0.03	0.85	0.06	0.37	0.12
WCL	4.47	0.03	2.31	0.05	1.08	0.10
Taper	5.13	3.40	2.11	3.94	1.13	8.45
WLS	8.91	0.00	6.81	0.00	4.73	0.01

As the golden method, MLE has the smallest MSEs for the price of the most computing time. WLS is the fastest for the cost of being least accurate. It is clear that JCEF well balances between time and MSE, and is the best among all methods in this simulation setup.

4.3.4 Comparison to Conditional Pseudo-Likelihood

Alternative to marginal bivariate distributions used in JCEF, estimation based on conditional density functions is also extensively considered in the literature. See Vecchia (1988) and Stein et al. (2004), among others.

Given innumerable ways to construct the conditioning sets, in the simulation study, we follow Stein (2005) to select half of the conditioning set from the nearest neighbors, and the other half from observations further apart. We vary the number of conditioning observations from 1, 2, 4, 6, and 8, and term them as Stein 1 to Stein 8, respectively. Results shown in Figure 4.3 are obtained based on setup 3 considered in Table 4.2, the same setting used in Table 4.3 and Figure 4.2. Figure 4.3 displays boxplots of $log(\beta)$ estimates and mean computing time for 5 versions of Stein's method and for our JCEF based on neighboring pairs. Results of the MLE are included as the golden standard. From Figure 4, we learn:

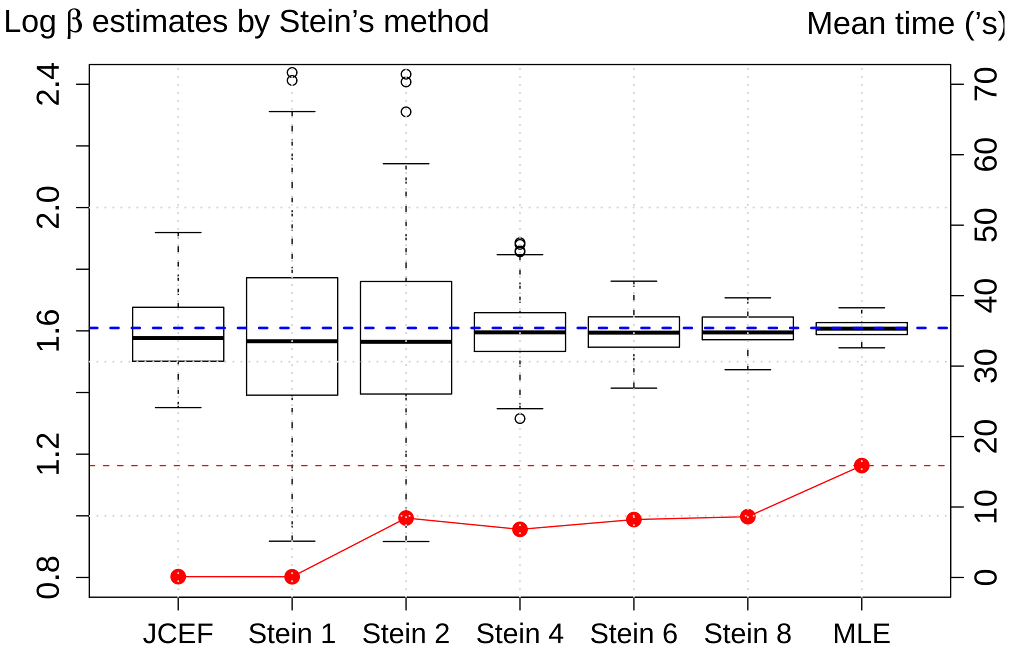
(i) As the size of conditioning sets increases, Stein's method yields improved efficiency, as a result of including the high-order conditional dependence.

(ii) When the size of the conditioning set is 1, Stein 1 uses bivariate density functions, and hence similar pairs are used in both Stein 1 and JCEF. Clearly, JCEF performs much better in terms of estimation efficiency. Interestingly, JCEF has shown to be comparable to Stein's method up to four conditioning observations. This suggests that the weight matrix used in JCEF incorporates additional amount of information beyond pairwise correlation, comparable to Stein 4.

(iii) Although Stein's method is always faster than MLE, it is clearly slower than JCEF. Thus, as far as computing time is concerned, JCEF will be advantageous for large-size data problems as well as running analysis on ordinary PCs.

In summary, we conclude that compared to Stein's method, JCEF is a desirable compromise between estimation and computational efficiency. In addition, unlike Stein's method, JCEF does not require an explicit specification and evaluation of the full multivariate density functions. This can be a considerable challenge when Stein's method is to be generalized to non-normal data, such as binary and Poisson data.

Figure 4.3: Boxplots for estimates of $log(\beta)$ by Stein's method with varying sizes of conditioning sets. Stein 1 refers to Stein's method with 1 conditioning observation and so forth. Estimates by JCEF based on neighboring pairs and by MLE are also plotted for comparison. Red solid line is the mean computing time (in seconds) for both JCEF and Stein's methods. Red dashed line is the mean time of MLE. Blue dashed line indicates the true $log(\beta)$ value.



4.4 Discussion

In this chapter, I have reviewed most of the methods proposed in the literature to estimate covariance structures for massive data sets. These parallel methodology developments have been mostly focused on spatial data sets, and have not yet been applied in spatio-temporal settings. Even within the spatial context, how these methods compare is still an open question.

The simulation experiments have compared the conventional WCL, JCEF, tapering, Stein's method, WLS and MLE. Results show that JCEF is advantageous over these methods in terms of balancing estimation and computational efficiency for large data sets. JCEF is faster and has smaller MSE than tapering when the taper range is short. It is a better method than Stein's method with one observation in the conditioning set, a situation where bivariate distributions are used by both methods. It also recovers a significant amount of efficiency from WCL. Unlike tapering, it does not require a joint distance metric in space and time, and is more robust to model misspecification than Stein's methods which require parametric forms of higher-order distributions.

Spectral methods and the fixed ranking kriging are not chosen because it is hard to set up the simulation for fair comparison. The spectral method requires a set of Fourier frequencies to approximate the likelihood, which cannot be easily related to a scenario in the composite likelihood or tapering setting. While the fixed rank kriging needs the specification of a set of basis functions, which again cannot be formulated to resemble any of the pseudo-likelihood methods.

However, it is possible to formulate a composite likelihood version comparable to the predictive process method. Specifically, if the predictive process is built on a coarsened spatial grid of the data on a regular lattice, and the dimension of the process is m. Then we can formulate composite likelihoods based on m-dimensional marginal densities of observations on the grids of the same dimension and shape. The resulting composite estimation will utilize much more information from the higher dependence structure of the data and should be more efficient than the pairwise version. Since different m-dimensional observations can be directly incorporated into the estimation. This enlarged composite estimation approach may be more efficient than the predictive process, where observations off the grid need to be predicted.

Sang and Huang (2011) proposed a full-scale approximation which aimed to retain the merit of the predictive process in capturing large-scale variations and that of the tapering in capturing short-range variations. A similar setup can be constructed within the composite likelihood framework. Namely, one part of the composite likelihoods can utilize observations of a larger dimension, in analogy to the predictive process. The other part can still be based on pairwise observations, similar to the tapering approach to capture a short range of dependences. Then, this mixture of composite likelihoods may provide an improved version of the current pairwise composite likelihoods to deal with anisotropic spatial dependences.

It is worth mentioning that by including more observations in the conditioning set, stein's method yields more efficient estimates at a contained computing cost, as seen in Figure 4.3. This shows that incorporating high-order dependence among observations increases estimation efficiency. A similar gain in efficiency can be expected for the marginal composite likelihood. I have conducted a small scale simulation to assess whether using tri-variate marginal composite likelihoods improves efficiency. Parameter values in setup 3 of Table 4.3 are set as the true values. Data are generated on a $5 \times 5 \times 5$ regular grid.

Figure 4.4: Boxplots of $\log \beta$ estimates from various methods, labeled along the horizontal axis. Stein 1 - Stein 3 correspond to conditioning sets of 1-3 observations. JCEF and WCL are based on bivariate densities. TCL denotes tri-variate marginal composite likelihood estimation. Computing times are marked in red above the horizontal axis. Parameter values in setup 3 Table 4.3 are set as the true values. Data are generated on a $5 \times 5 \times 5$ regular grid.

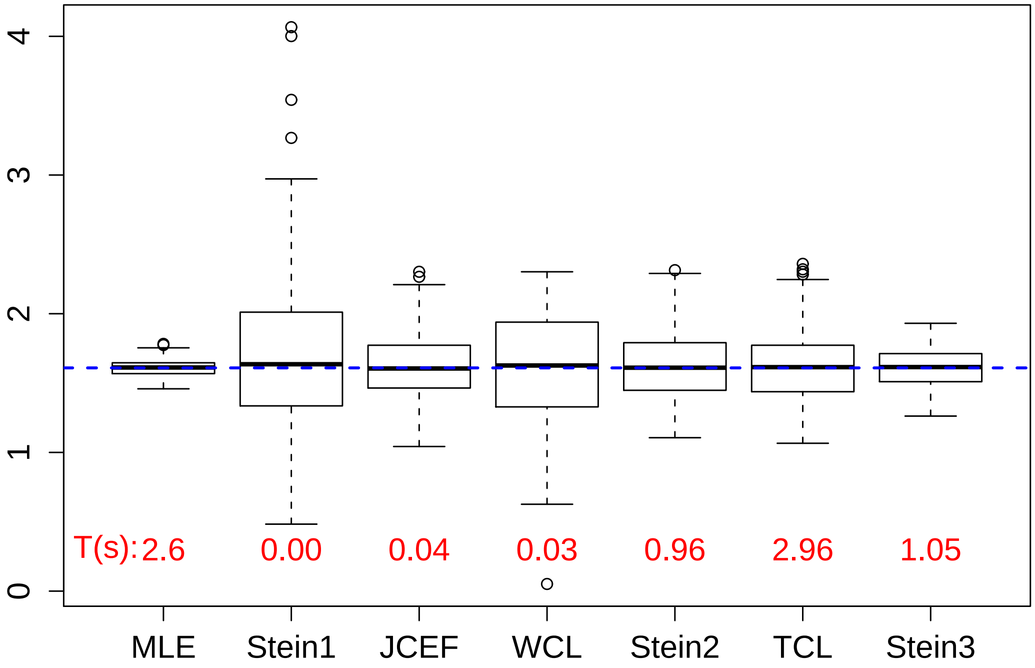


Figure 4.4 displays the boxplots of $\log \beta$ estimates from various methods, labeled along the horizontal axis. Stein1 - Stein3 correspond to the conditioning set of 1-3 observations. JCEF and WCL are based on bivariate densities. TCL denotes tri-variate marginal composite likelihood estimation. Computing times are

marked in red above the horizontal axis. It is clear that TCL is more efficient than WCL and is comparable to Stein2. This indicates that triplets of observations yield more efficient estimates than pairs of observations in spatio-temporal setting.

CHAPTER V

GeoCopula Regression Models for Spatial-Clustered Data

5.1 Introduction

In social and health sciences, research studies usually involve subjects that are randomly selected within a large number of geographical units. For example, among the studies of place effects on health, Chaix et al. (2005) investigated individual and contextual factors that determine the health care utilization in France, where 10955 people are randomly surveyed within 4421 municipals in France. To study the association of neighborhood environmental risk factors with cardiovascular diseases, Mujahid et al. (2007) used a sample of 5988 subjects selected from 576 census tracts from three states in USA. Grady (2010) assessed the impact of racial residential segregation on low birth weight from a pool of 10277 cases nested in 1092 census tracts in Michigan. In civil and environmental engineering studies, Sener et al. (2011) analyzed the physical activity participation levels of individuals in a family unit based on data drawn from the 2000 San Francisco Bay Area Household Travel Survey, in which individual and household socio-demographic as well as all activity and travel episodes information were recorded for subjects in 15000 households.

These examples are just a glimpse of a growing number of research projects that collect data in spatial dimensions, thus necessitate the eminent need to generalize the multilevel data analysis to incorporate the spatial dependences among the clustering units. In classic multilevel models, data from clusters are assumed to be independent, and the focus dwells on appropriately accounting for within-cluster correlations while making statistical inferences. However, when clusters are spatially correlated, such as neighborhoods or census tracts, subjects from clusters are likely to be correlated due to location proximity, hence the between-cluster independence assumption is no longer valid. Statistical analysis ignoring the spatial effect can lead to wrong standard errors of the regression coefficient estimates, which in turn biases hypothesis testing (Anselin and Griffith, 1988). As a result, in order to draw valid statistical inference, it is of critical importance to account for the between-cluster spatial correlation as well as the within-cluster correlation.

In the current literature, there are two popular modeling frameworks for analyzing spatially correlated data. One approach is based on random effects models, where mean models are specified conditional on cluster-specific random effects (e.g. Diggle et al., 2008). The spatial structures are accounted for by allowing random effects to distribute as a spatial stochastic process. For non-Gaussian data, regression parameters in such hierarchical specification only have conditional or cluster-specific interpretations, which may not be desirable when population characteristics are of interest. The other approach is the generalized estimating equation (GEE, Liang and Zeger 1986), which specifies the mean model and covariance separately. In the covariance model, the spatial dependence is incorporated via a spatially structured working correlation matrix (e.g. Albert and McShane, 1995; Gotway and Stroup, 1997). GEE is suitable when the mean model is of central interest, since it treats spatial dependences as nuisance components. As a result, GEE is not appropriate for spatial interpolation, which however is an important task in many practical studies, such as disease mapping (Diggle et al., 2008).

In this chapter, we propose a new and flexible modeling framework that models both mean and covariance structures of spatial-clustered data, termed as GeoCopula regression model. In this model, univariate margins are specified by generalized linear models, while the spatial and cluster dependences are modeled through the multivariate Gaussian copula. The proposed framework allows us to analyze a large variety of multivariate discrete and continuous spatial-clustered data, including normal, binary and count data as special cases. Since the mean and the dependence structure are separately formulated, regression parameters have marginal interpretations, and at the same time, spatial dependence is explicitly modeled by the copula and is not constrained by the mean model.

It is worth mentioning that Bárdossy (2006) and Bárdossy and Li (2008) proposed to use bivariate copulas as an alternative to variograms and covariance functions to describe spatial variability. They showed that copula-based approach is more flexible in accounting for asymmetrical dependence and is superior in terms of prediction when the normality assumption is violated. Moreover, Kazianka and Pilz (2010) proposed a similar regression model in which exponential dispersion distribution family (Jorgensen, 1997) is used as the marginal distributions and a multivariate copula is applied to model the spatial dependence. Our work in this chapter extends Kazianka and Pilz's model to analyze more complex spatial-clustered data, and attempts to provide a richer statistical presentation (e.g. large sample properties) of the multivariate copula regression model. Most importantly, our joint composite estimating function is new and computationally efficient in complex data structures.

A key obstacle of preventing the wide spread of spatial analysis in contextual research is mostly due to computational issues. Almost all existing models require either high-dimensional matrix manipulations such as in GEE, or high-dimensional integrations, such as in random effects models. Numerical calculations quickly become intractable for data sets with a large number of spatial units, as in the previous examples. Similar computational problems are faced by Bayesian approaches as well.

The need to reduce computational burden is eminent in many practical situations. For spatial data, people have tried to use composite likelihood (CL) methods (Lindsay, 1988), which is a general class of pseudo-likelihoods based on likelihoods of marginal or conditional events. Among many others Curriero and Lele (1999) used CL in spatial variogram estimation, and demonstrated that the CL approach provides consistent estimates and is superior to likelihood-based methods in terms of weaker distributional assumptions and much less computational burden. Heagerty and Lele (1998) applied CL approach to binary spatial data, which is modeled via a probit model of pairwise observations using an exponential decaying covariance structure. Li and Lin (2006) modeled spatially correlated survival data by a Gaussian copula and avoided the high-dimensional integration of the likelihood function by again considering pairwise observations. Varin et al. (2005) used pairwise CL to estimate GLMM for Poisson data and show that CL can considerably reduce computing burden and retain adequate efficiency. Bárdossy (2006) and Bárdossy and Li (2008) developed bivariate spatial copulas to model ground water quality parameters. While Kazianka and Pilz (2010) proposed the same CL approach for estimating spatial copula models.

Though the pairwise CL approach, currently the most popular version of CL, is computationally appealing and yields estimators with sound asymptotic properties, it implicitly treats observation pairs as independent, resulting in some loss of efficiency in comparison to the full likelihood method. Recently, Bai et al. (2011) proposed a joint composite estimating function (JCEF) approach to accounting for correlations among the pairs in the analysis of spatio-temporal data. In the context of spatial-clustered data, we aim at the development of a new JCEF method that yields better efficiency than the existing pairwise CL methods. Recognizing the differences between within-cluster correlation and between-cluster spatial correlation, we group pairs into between-cluster pairs that come from different clusters, and within-cluster pairs that lie within a cluster, (see Figure 5.1 in section 5.3). The former group is expected to provide more information about cluster-level covariate effects (e.g. environmental factors) and between-cluster spatial dependence, while the latter may be more informative about within-cluster covariate effects (e.g. subject-level characteristics) and within-cluster correlations. Then we combine the two sets of composite estimating functions into a quadratic objective function, in a similar way suggested by Qu et al. (2000). Then the estimation is carried out by minimizing the objective function. In this way efficiency can be improved.

In this chapter, GeoCopula model is constructed from the multivariate Gaussian copula for three reasons: (i) When margins are normal linear models, the proposed GeoCopula model becomes the multivariate Gaussian distribution, the most widely used model for spatial continuous data. When the probit link is used for binary data, the GeoCopula model results in a multivariate probit model, another very popular model for spatial binary data. (ii) The dependence structure is conveniently depicted by a correlation matrix in the Gaussian copula, which can be straightforwardly utilized in the spatial interpolation, such as kriging. (iii) Other copulas such as Archimedean copulas cannot accommodate as rich and flexible spatial dependences as the Gaussian copula. We will revisit this point in a later section.

The rest of the chapter is structured as follows. In section 5.2, the GeoCopula model is proposed and detailed for multivariate Gaussian and binary data. Section 5.3 proposes a joint composite estimating function approach to estimating parameters in the GeoCopula model. Large-sample properties of the proposed estimator is investigated in section 5.4. Simulation experiments are conducted in section 5.5. A real data example is illustrated in section 5.6, followed by some discussions in section 5.7.

5.2 Model

Let Y_{si} denote the outcome of the ith subject nested in geographic cluster s, where $i \in \mathcal{I}_s$, the index set of subjects in cluster s, and $s \in \mathcal{S} \subset \mathcal{R}^2$, with S being a collection of spatial clusters under study. Denote the number of subjects in cluster s as n_s, and the total number of subjects is $n = n_1 + \cdots + n_S$. Suppose that each outcome Y_{si} follows a generalized linear model (McCullagh and Nelder, 1989), whose mean (or systematic component) μ_{si} is specified as a function of p covariates, $\mathbf{x}_{si} = (x_1^{si}, \ldots, x_p^{si})^T$ via a known link function h; that is,

$$h(\mu_{si}) = \eta(\mathbf{x}_{si}) = \mathbf{x}_{si}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_1^{si} + \cdots + \beta_p x_p^{si},$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is a vector of regression coefficients.

The cumulative distribution function (CDF) of Y_{si} is given by $F_{si}(y_{si}; \mu_{si}, \varphi_{si})$, where φ_{si} is the dispersion parameter. For simplicity, write the univariate CDF by $F_{si}(y_{si})$, and the corresponding density function by $f_{si}(y_{si})$.

5.2.1 Copula Dependence Model

To specify a fully parametric model for all Y_{si}'s, we invoke a copula dependence model to characterize both spatial and within-cluster correlations. In short, an n-dimensional copula function is a multivariate parametric distribution with univariate uniform margins. That is, copula $C(\mathbf{u})$ is a CDF in the n dimensional cube with uniformly distributed marginals $\mathbf{u} = (u_1, \ldots, u_n)$. A copula can be easily constructed from a given multivariate distribution. Let $W = (W_1, \ldots, W_n)^T \sim G$ where G is an n dimensional CDF with margins $G_1, \ldots, G_n$. Then the resulting copula takes the form

$$C_G(\mathbf{u}) = G\left\{G_1^{-1}(u_1), \ldots, G_n^{-1}(u_n)\right\}, \tag{5.1}$$

where $\mathbf{u} = (u_1, \ldots, u_n)^T \in (0,1)^n$, provided the existence of all marginal inverse CDFs (i.e. quantile functions) G_i^{-1} of G_i. Note that the dependence structure in the original G is transfered into copula C_G. Equivalently, by a change of variables,

$$C_G\left(G_1(w_1), \ldots G_n(w_n)\right) = G\left(w_1, \ldots, w_n\right).$$

By supplementing the copula C_G in equation (5.1) with any given margins, say $F_{11}, \ldots, F_{Sn_S}$, a new multivariate distribution can be obtained as

$$F(\mathbf{y}) = C_G\{F_{11}(y_{11}), \ldots, F_{Sn_S}(y_{Sn_S})\}, \tag{5.2}$$

where $\mathbf{y} = (y_{11}, \ldots, y_{Sn_S})$.

When marginal outcomes are all continuous, the first order derivative of CDF (5.2) leads to the density function of $\mathbf{y}$ as follows:

$$f(\mathbf{y}) = c_G\{F_{11}(y_{11}), \ldots, F_{Sn_S}(y_{Sn_S})\} \prod_{s \in \mathcal{S}, i \in \mathcal{I}_S} f(y_{si}), \tag{5.3}$$

where c_G is the density function corresponding to C_G.

When marginal outcomes are all discrete, a multivariate probability mass function is obtained by taking the Radon-Nikodym derivative of CDF $F(\mathbf{y})$ in equation (5.2), and given as follows:

$$\begin{aligned} f(\mathbf{y}) &= P\left(Y_{11} = y_{11}, \ldots, Y_{Sn_S} = y_{Sn_S}\right) \\ &= \sum_{j_{11}=1}^{2} \cdots \sum_{j_{Sn_S}=1}^{2} (-1)^{j_{11}+\cdots+j_{Sn_S}} C_G(u_{11}^{j_{11}}, \ldots, u_{Sn_S}^{j_{Sn_S}}), \end{aligned} \tag{5.4}$$

where $u_{si}^1 = F_{si}(y_{si})$, and $u_{si}^2 = F_{si}(y_{si}-)$. Here $F_{si}(y_{si}-)$ is the left-hand limit of F_{si}, which is equal to $F_{si}(y_{si} - 1)$ when the support of F_{si} is an integer set, such as the case for Poisson or binomial outcomes.

Different choices of copula models result in different multivariate parametric distributions. The following examples indicate the richness and flexibility of the copula framework.

Example 1 (Multivariate Gaussian Copula). When G is the n dimensional multivariate Gaussian distribution with zero means and a correlation matrix Σ, the resulting copula is known as the multivariate Gaussian copula, with CDF given by:

$$C_\Phi(\mathbf{u}|\Sigma) = \Phi_n \left\{ \Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_n)|\Sigma \right\}, \tag{5.5}$$

Here Φ_n and Φ are CDFs for n-variate normal $N_n(0, \Sigma)$ and the standard univariate normal $N(0, 1)$ marginals, respectively. Note that $(u_1, \ldots, u_n)$ are independent when Σ is the identity matrix.

Example 2 (Multivariate *t* Copula). When G is the n dimensional multivariate t distribution with zero means, degrees of freedom κ and a correlation matrix Σ, the resulting copula is called the multivariate t copula with CDF given by,

$$C_T(\mathbf{u}|\Sigma, \kappa) = T_n \left\{ T^{-1}(u_1|\kappa), \ldots, T^{-1}(u_n|\kappa)|\Sigma, \kappa \right\},$$

where $T_n(\bullet|\Sigma, \kappa)$ is the n-variate t CDF with d.f. κ and $T(\bullet|\kappa)$ is the corresponding univariate t CDF with d.f. κ. It is known that the t copula captures some tail dependences as opposed to the Gaussian copula that yields independence in the lower or upper tails. Hence t copula is often applied to model continuous outcomes of extreme values. When d.f. $\kappa \to \infty$, the t copula approaches the Gaussian copula. However, for the t copula, an identity correlation matrix does not imply independence among margins. Also, in practice it is very subtle to handle tail dependences for discrete outcomes, such as binary data.

Example 3 (Multivariate Archimedean copula). A copula is called Archimedean if it admits the following representation

$$C(\mathbf{u}) = \psi \left(\psi^{-1}(u_1) + \ldots + \psi^{-1}(u_n) \right),$$

where ψ is the copula generator. Some important generators include Clayton, Frank and Gumbel copula. Unlike the elliptical copulas such as Gaussian and t, Archimedean copulas usually have explicit CDF formulas, and hence are easier to evaluate. Although, they can deal with arbitrary n dimensional outcomes, there is only one dependence parameter specified to measure the strength of the association among n components. For example, the Clayton generator function is $(1+t)^{-1/\theta}$, where θ describes the dependence among all univariate margins. For complex data structures, like the spatial-clustered data considered in this chapter, multi-level dependences are present. So a simple dependence parameter is not enough to describe the dependence structure of the data.

Example 4 (Vine copulas). To formulate more flexible multivariate copulas that deal with complex dependence structures and that are computationally tractable, a pair-copula decomposition of a multivariate copula density function was proposed in Joe (1996); Bedford and Cooke (2001). Specifically, an n-variate copula density c can be expressed as a product of $n(n-1)/2$ bivariate conditional copula densities, in a sequential manner,

$$c(\mathbf{u}) = \prod_{j=1}^{n-1} \prod_{j=1}^{n-j} c_{i,i+j|i+1,\ldots,i+j-1} \left\{ u_{i|i+1,\ldots,i+j-1}, u_{i+j|i+1,\ldots,i+j-1} \right\}, \tag{5.6}$$

where

$$u_{i|i+1,\ldots,i+j-1} = F(y_i | y_{i+1}, \ldots, y_{i+j-1}),$$

and

$$u_{i+j|i+1,\ldots,i+j-1} = F(y_{i+j} | y_{i+1}, \ldots, y_{i+j-1})$$

are the conditional CDFs, based on given margins $F(y_i), i = 1, \ldots, n$ and bivariate copula dependence measures, $c_{i,i+j|i+1,\ldots,i+j-1}$. For example, when $n = 3$, expression (5.6) becomes

$$c(u_1, u_2, u_3) = c_{12}(u_1, u_2) * c_{23}(u_2, u_3) * c_{13|2}(u_{1|2}, u_{3|2}). \tag{5.7}$$

The decomposition of the multivariate copula density is not unique. Bedford and Cooke (2002) used graphical models to organize possible decompositions in a systematic way, and called them ”vines”. The formulation adopted in equation (5.6) is the D-vine structure. In parallel, there is a C-vine structure.

One major advantage of the vine copula construction is that different copula functions can be specified for different pairs to more flexibly capture complex dependences of real data. In addition, via this decomposition the evaluation of the full likelihood only involves calculation of bivariate densities. For example, in equation (5.7), c_{12}, c_{23} and $c_{13|2}$ can be specified by different two-dimensional copulas. This form of pair-copula construction has been applied to study serial dependences for longitudinal data (Smith et al., 2010), and cross-sectional dependence for multiple time series (Min and Czado, 2010). The current practice is only limited to $n \leq 50$ or so, due to the escalating computational burden.

The presence of spatial dependence structures requires the multivariate copula to fulfill the following three properties as pointed out by Bárdossy (2006); Kazianka and Pilz (2010): (i) exchangeability, that is the dependence between location s_1 and s_2 is the same as the dependence between s_2 and s_1; (ii) arbitrarily strong or weak dependence can be modeled, that is, given a set of very close spatial locations, there should be a parameterization of the copula to achieve full dependence. On the contrary, when spatial locations are far away, the parameterization of the copula should lead to independence; (iii) The geometric position of the corresponding locations can be incorporated into the copula parameterization. In addition, we believe that another three features are also desirable: (iv) high-dimensional and flexible dependence structures can be incorporated; (v) continuous and discrete outcomes can be modeled; (vi) the copula regression model should encompass existing popular models as special cases.

Given these constraints, not all the popular multivariate copulas are appropriate in modeling spatial-clustered data. For example, in Example 3, the Archimedean copulas have only one dependence parameter, hence the strength of dependence does not vary with spatial distances. As for the multivariate t copula in Example 2, even if the correlation matrix can be formulated as a function of spatial distances, zero correlations do not imply statistical independence among margins (Kotz and Nadarajah, 2004), and the tail dependence is hard to interpret for binary outcomes. Furthermore, although the pair-copula construction in Example 4 allows flexible dependences to be specified in each bivariate copulas, such flexibility hampers the use of a unified dependence measure to describe a certain overall dependence pattern over the spatial domain (e.g. different bivariate copulas have different dependence ranges and interpretations), which is a crucial component needed in spatial interpolation, such as kriging (Cressie, 1993, Chapter 3). Also, given a set of pairwise dependence parameters, it is difficult to model, estimate ($n(n-1)/2$ dependence parameters) and interpret related dependence patterns. For example, in equation (5.7) c_{12}, c_{23} and $c_{13|2}$ have their own pair-specific dependence parameters, which appear to be too generic to model overall spatial features using the vine copula model.

5.2.2 GeoCopula Regression Models

In this chapter, we choose the multivariate Gaussian copula described in Example 1 as the dependence model to build the GeoCopula regression model. The advantages include (i) the multivariate Gaussian copula is both analytically and theoretically well studied; (ii) the correlation matrix enables us to model a dependence map across the entire spatial region under study, which can accommodate full dependence with correlations approaching 1, and full independence with zero correlation coefficients, as well as positive and negative correlations; (iii) the correlation pattern can be easily formulated as functions of spatial coordinates and covariates, which can be conveniently estimated and applied for spatial interpolation.

For example, if we assume a compound symmetry (i.e. exchangeable) structure for within-cluster correlation, then the within-cluster correlation matrix for cluster i is

$$\Sigma_{ii} = (1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i}, \quad i = 1, \ldots, S \tag{5.8}$$

where ρ is the correlation among individuals within the same cluster, and $\mathbf{I}_{n_i}$ is an $n_i \times n_i$ identity matrix, and $\mathbf{J}_{n_i}$ an $n_i \times n_i$ matrix with all entries being 1.

Furthermore, if we assume the spatial correlation to be the Matérn class across clusters, the spatial correlation matrix between observations in clusters s and t is

$$\Sigma_{st} = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu} d_{st}}{\alpha} \right) K_\nu \left(\frac{2\sqrt{\nu} d_{st}}{\alpha} \right) * \mathbf{J}_{n_s \times n_t}, \tag{5.9}$$

where d_{st} is the distance between cluster s and t, and $\mathbf{J}_{n_s \times n_t}$ is an $n_s \times n_t$ matrix with all entries being 1. That is, subjects in cluster s are equally correlated with subjects in cluster t. The strength of the correlation is a function of the distance between two clusters.

It follows that the overall correlation matrix Σ is an $S \times S$ block matrix of the form

$$\Sigma = [\Sigma_{ij}]_{S \times S}, \quad i, j = 1, \ldots, S, \tag{5.10}$$

where the block-diagonal Σ_{ii} is given in (5.8) and the off block-diagonal Σ_{ij} is given in (5.9).

Moreover, various spatial correlation patterns like, for example, nonstationarity and anisotropy, can be easily formulated for between-cluster correlation patterns as functions of spatial ordinates and covariates. In summary, almost all existing spatial correlation structures studied in the literature can be straightforwardly adapted into the multivariate Gaussian copula model with little effort, through blocks of the between-cluster correlation matrices.

Another important advantage of using Gaussian copula is that by specifying different marginal GLMs for $F_{si}(y_{si})$, the resulting GeoCopula, given by

$$F(\mathbf{y}) = \Phi_n \left\{ \Phi^{-1}(F_{11}(y_{11})), \ldots, \Phi^{-1}(F_{Sn_S}(y_{Sn_S})) | \Sigma \right\}, \tag{5.11}$$

encompasses a wide range of useful models in practice, including the Gaussian spatial model (when $F_{si}(y_{si})$ is normal), and the multivariate probit model (when $F_{si}(y_{si})$ is binomial with a probit link) as special cases.

Example 5 (GeoCopula Special Case I: Multivariate Gaussian Regression Model). Assume marginally $Y_{si} \sim N(\mathbf{x}_{si}^T \boldsymbol{\beta}, \sigma_{si}^2)$, and denote the uniform random variable

$u_{si} = \Phi\left(\frac{y_{si} - \mathbf{x}_{si}^T \boldsymbol{\beta}}{\sigma_{si}}\right)$. Plug u_{si} into equation (5.5), we obtain

$$F(\mathbf{y}) = \Phi\left(\frac{y_{11} - \mathbf{x}_{11}^T \boldsymbol{\beta}}{\sigma_{11}}, \ldots, \frac{y_{Sn_S} - \mathbf{x}_{Sn_S}^T \boldsymbol{\beta}}{\sigma_{Sn_S}} | \Sigma\right).$$

That is

$$\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, D\Sigma D\right), \tag{5.12}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{11}^T \\ \ldots \\ \mathbf{x}_{Sn_S}^T \end{pmatrix}, \quad D = diag\{\sigma_{11}, \ldots, \sigma_{Sn_S}\}.$$

Example 6 (GeoCopula Special Case II: Multivariate Probit Model). Assume marginally $Y_{si} \sim Bernoulli(p_{si})$. Then the CDF of Y_{si} is

$$F_{si}(y_{si}) = \begin{cases} 0, & y_{si} < 0 \\ 1 - p_{si}, & 0 \leq y_{si} < 1 \\ 1, & y_{si} \geq 1. \end{cases}$$

Consider a probit regression model $p_{si} = \Phi(\mathbf{x}_{si}^T \boldsymbol{\beta})$. Plug $F_{si}(y_{si})$ into equation (5.11), we obtain a multivariate distribution for n-variate binary data, which is shown in Song (2000) to have the same probability mass function as that generated by the following multivariate probit model.

Specifically, let the latent normal variable

$$Z_{si} = \mathbf{x}_{si}^T \boldsymbol{\beta} + \epsilon_{si}, \quad \text{and} \quad \boldsymbol{\epsilon} = (\epsilon_{11}, \ldots, \epsilon_{Sn_S})^T \sim N(0, \Sigma).$$

Define a dichotomous procedure as follows:

$$Y_{si} = I(Z_{si} > 0),$$

where $I(\bullet)$ is the indicator function. Then $(Y_{11}, \ldots, Y_{Sn_S})^T$ defined by this threshold model has the same probability mass function as the random vector constructed in Example 6. That is, the multivariate probit model is a special case of the proposed GeoCopula regression model.

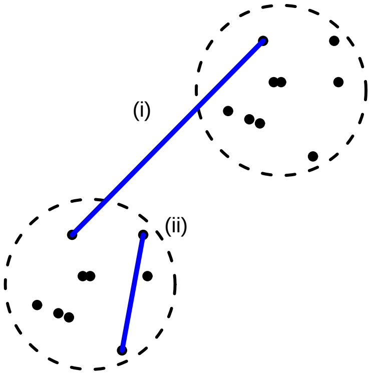
5.3 Estimation

5.3.1 General Theory

For a large-scale data set, computing the distribution function of the GeoCopula models in equation (5.11) either requires high-dimensional integration or large matrix inversion, hence is not numerically feasible. Following Besag (1974), we consider a pseudo-likelihood approach to perform parameter estimation and inference for the GeoCopula models. Estimation functions are formulated from pairwise marginal composite likelihoods (Lindsay, 1988; Varin et al., 2011). Bai et al. (2011) proposed the joint composite estimating function (JCEF) approach to further improve the estimation efficiency by forming a quadratic objective function from different types of pairwise estimating functions. We develop an analog of the JCEF approach in this new class of models. Each type of estimating functions is constructed by grouping pairs of outcome variables according to characteristics of the underlying spatial process. In our spatial-clustered data, a natural grouping scheme is to partition pairs into within-cluster and between-cluster groups (e.g.

villages), as shown in Figure 5.1. The former contains pairs of observations from a common cluster, which are more relevant to subject-level effects and within-cluster correlations. While the latter consists pairs of observations from different clusters, which capture more information relevant to cluster-level covariate effects and between-cluster spatial correlations.

Figure 5.1: Configurations of spatial-clustered data with two clusters. (i) between-cluster pair, (ii) within-cluster pair.



To develop JCEF, the first step is to marginalize the high-dimensional CDF function in equation (5.11) into 2-dimensional marginals. Let the vector of parameters of interest be $\boldsymbol{\theta}$, which includes the mean regression coefficients $\boldsymbol{\beta}$, the dispersion parameter ψ, and the variance and covariance parameters ρ and α in Σ, provided that a Matérn class spatial correlation function is assumed. Assume the length of $\boldsymbol{\theta}$ is p. By the property of marginal closure of the Gaussian copulas, a 2-dimensional marginal CDF is given by,

$$F(y_{si}, y_{tj}; \boldsymbol{\theta}) = \Phi_2\{\Phi^{-1}(F_{si}(y_{si}; \boldsymbol{\beta}, \psi)), \Phi^{-1}(F_{tj}(y_{tj}; \boldsymbol{\beta}, \psi)) | \Sigma_{si,tj}(\rho, \alpha)\}, \qquad (5.13)$$

where $\Sigma_{si,tj}$ is the 2×2 corresponding sub correlation matrix.

Let $f(y_{si}, y_{tj}; \boldsymbol{\theta})$ be the density with respect to $F(y_{si}, y_{tj}; \boldsymbol{\theta})$, whose explicit expression form is given by equation (5.3) for continuous outcomes and in equation (5.4) for discrete outcomes. Let $U(y_{si}, y_{tj}; \boldsymbol{\theta})$ be the marginal score function associated with $f(y_{si}, y_{tj}; \boldsymbol{\theta})$, which is called the component score function (CSF). According to Varin et al. (2011), the conventional composite likelihood estimating functions is formed by summing all such CSFs within a certain distance lag d:

$$S(\boldsymbol{\theta}, d) = \sum_{||s-t||<d} U(y_{si}, y_{tj}; \boldsymbol{\theta}, d),$$

where $|| \bullet ||$ is the Euclidean distance in R^2. We call this method weighted composite likelihood (WCL) approach (Bevilacqua et al., 2011). The weight is 0/1, depending on the distance d between two clusters.

The optimal d can be determined by evaluating the Godambe information matrix (i.e. asymptotic covariance of the estimates) of the corresponding estimating equations (Bevilacqua et al., 2011). A value of d leading to the most informative set of estimating equations is then used. However, when calculating the information matrix is computationally costly, empirical guidelines can be used. For example, from the empirical spatial variogram, one can learn the spatial dependence patterns, and choose a value for d within which pairwise correlations are fairly high. Numerous simulation experiments are reviewed in (Varin et al., 2011) and studied in Davis and Yau (2011) and Bai et al. (2011). All have shown that including pairs within shorter distances usually results in more efficiency than including pairs further apart. This is desirable, since a substantial number of pairs can be eliminated from estimation, which greatly facilitates the computational feasibility and speed.

To develop the JCEF approach, we first partition pairs into between-cluster and within-cluster groups. Label the two sets as $D_{W,n}$ and $D_{B,n}$, respectively. They are given by

$$D_{W,n} = \{(s,i,t,j) : s = t \in \mathcal{S}, \text{ and } i \neq j,\, i,j \in \mathcal{I}_s\},$$

$$D_{B,n} = \{(s,i,t,j) : 0 < ||s-t|| \leq d;\, s,t \in \mathcal{S}, \text{ and } i \in \mathcal{I}_s, j \in I_t\}.$$

Then $D_n = D_{W,n} \cup D_{B,n} \in \mathcal{S} \times \mathcal{I}_s \times \mathcal{S} \times \mathcal{I}_s \subset \mathcal{R}^2 \times \mathcal{R}^+ \times \mathcal{R}^2 \times \mathcal{R}^+$ is the set containing all pairs used in estimation.

The between-cluster CSF is constructed as

$$\Psi_{B,n}(\boldsymbol{\theta}, d) = \frac{1}{|D_{B,n}|} \sum_{(s,i,t,j) \in D_{B,n}} U(y_{si}, y_{tj}; \boldsymbol{\theta}, d).$$

And, the within-cluster CSF is constructed as

$$\Psi_{W,n}(\boldsymbol{\theta}) = \frac{1}{|D_{W,n}|} \sum_{(r,l,r,m) \in D_{W,n}} U(y_{rl}, y_{rm}; \boldsymbol{\theta}),$$

where $|A|$ is the cadinality of set A.

Instead of taking the sum of the two groups of CSFs, we stack them into an extended CSF:

$$\boldsymbol{\Gamma}_n(\boldsymbol{\theta}, d) = \left(\Psi^T_{B,n}(\boldsymbol{\theta}, d), \Psi^T_{W,n}(\boldsymbol{\theta})\right)^T.$$

Note that the dimension of $\boldsymbol{\Gamma}_n$ is larger than that of $\boldsymbol{\theta}$, leading to a so-called over-identification scenario (Hansen, 1982). To obtain an estimate of $\boldsymbol{\theta}$, following Hansen (1982) and Qu et al. (2000), we form a quadratic objective function of the following form:

$$Q_n(\boldsymbol{\theta}, d) = \boldsymbol{\Gamma}^T_n(\boldsymbol{\theta}, d) W^{-1} \boldsymbol{\Gamma}_n(\boldsymbol{\theta}, d),$$

where W is a $2p \times 2p$ positive-definite weight matrix. Consequently, a JCEF estimator is defined as

$$\hat{\boldsymbol{\theta}}_n(d) = argmin_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} Q_n(\boldsymbol{\theta}, d). \tag{5.14}$$

According to Hansen (1982), the optimal weight matrix is $W = Cov\left(\boldsymbol{\Gamma}_n(\boldsymbol{\theta}, d)\right)$, in the sense that the resulting estimator has the maximum efficiency.

5.3.2 JCEF in Multivariate Probit Model

It is relatively easy to derive JCEF in the multivariate Gaussian model by following the general procedure outlined in the above session. Here, we illustrate the derivation of JCEF in the GeoCopula regression model for binary data. We refer to Heagerty and Lele (1998) that considered multivariate probit model for spatial binary data.

First, the bivariate probability mass function for Y_{si} and Y_{tj} in the general canonical form is given by:

$$\log P\left(Y_{si} = y_{si}, Y_{tj} = y_{tj}\right) = \alpha_0(si, tj) + \alpha_1(si, tj) y_{si} + \alpha_2(si, tj) y_{tj} + \alpha_3(si, tj) y_{si} y_{tj}.$$

Second, according to Zhao and Prentice (1990), the score function may be expressed as follows:

$$U_{si,tj}(\boldsymbol{\theta}) = D_{si,tj}^T V_{si,tj}^{-1} R_{si,tj},$$

with

$$D_{si,tj} = \frac{\partial}{\partial \boldsymbol{\theta}} \begin{pmatrix} \mu_{si} \\ \mu_{tj} \\ \sigma_{si,tj} \end{pmatrix} \text{ and } R_{si,tj}(\boldsymbol{\theta}) = \begin{pmatrix} y_{si} - \mu_{si} \\ y_{tj} - \mu_{tj} \\ (y_{si} - \mu_{si})(y_{tj} - \mu_{tj}) - \sigma_{si,tj} \end{pmatrix}.$$

The detailed expression of $V_{si,tj} = var\left(R_{si,tj}\right)$ can be found in Appendix A in Heagerty and Lele (1998).

Third, based on the GeoCopula model, we have

$$\mu_{si} = E(y_{si}) = \Phi(\mathbf{x}_{si}^T \boldsymbol{\beta}), \quad \mu_{tj} = E(y_{tj}) = \Phi(\mathbf{x}_{tj}^T \boldsymbol{\beta}),$$

$$\sigma_{si,tj}^2 \quad = \quad \Phi_2\left(\mathbf{x}_{si}^T \boldsymbol{\beta}, \mathbf{x}_{tj}^T \boldsymbol{\beta} | \Sigma_{si,tj}\right) - \Phi(\mathbf{x}_{si}^T \boldsymbol{\beta}) \Phi(\mathbf{x}_{tj}^T \boldsymbol{\beta}).$$

Finally, the group-based composite score functions are

$$\Psi_{B,n}(\boldsymbol{\theta}, d) = \frac{1}{|D_{B,n}|} \sum_{(s,i,t,j) \in D_{B,n}} U_{si,tj}(\boldsymbol{\theta}, d),$$

and

$$\Psi_{W,n}(\boldsymbol{\theta}) = \frac{1}{|D_{W,n}|} \sum_{(r,l,r,m) \in D_{W,n}} U_{rl,rm}(\boldsymbol{\theta}).$$

5.3.3 Estimation of the Weight Matrix

Although the optimal weight matrix $Cov(\Gamma_n(\boldsymbol{\theta}))$ can be derived analytically using multivariate Gaussian quadrant probabilities, given the large number of possible pairs, computing it based on the analytic formula is not practically feasible. Alternatively, in spatial data analysis, estimation of this covariance matrix is mostly achieved by subsampling techniques as suggested in Heagerty and Lele (1998); Heagerty and Lumley (2000); Lee and Lahiri (2002); Li and Lin (2006). Specifically, let the sampling region be $A_n = \mathcal{S} \times \mathcal{T}$, where $|A_n| = n$. Under the assumption that, asymptotically, $|A_n| E(\Gamma_n(\boldsymbol{\theta})\Gamma_n^T(\boldsymbol{\theta})) \to \Lambda$, we can estimate Λ using sample covariance matrix of statistics computed on either overlapping or non-overlapping subshapes of the sampling region A_n. That is:

$$\hat{\Lambda} = k_n^{-1} \sum_{i=1}^{k_n} |A^i_{l(n)}| \left\{ \Gamma^i_n(\boldsymbol{\theta}) - \bar{\Gamma}(\boldsymbol{\theta}) \right\}^2, \tag{5.15}$$

with $\bar{\Gamma}(\boldsymbol{\theta}) = \sum_{i=1}^{k_n} \Gamma^i_n(\boldsymbol{\theta})/k_n$, where $\Gamma^i_n(\boldsymbol{\theta})$ is vector $\Gamma(\boldsymbol{\theta})$ evaluated in $A^i_{l(n)}, i = 1, \ldots, k_n$, a collection of (non)overlapping subshapes of A_n and k_n denotes the number of subshapes.

We will apply this subsampling technique for our weight matrix estimation as well as later the standard error estimation. We follow Politis and Romano (1994) to choose the optimal subsample size proportional to $Cn^{a/a+2}$, where a is the dimensional of the spatial domain and C is a tuning constant.

5.4 Large Sample Properties

In the spatial-clustered setting considered in this chapter, the increase of the sample size can be achieved by either increasing the number of subjects within each cluster, or by increasing the number of spatial clusters. For the latter case, two scenarios are possible: (i) more sample locations are added within a fixed spatial domain, known as the in-fill asymptotics (Zhang, 2004); (ii) More locations are included by expanding the spatial domain, corresponding to the increasing-domain asymptotics (Mardia and Marshall, 1984). Sampling more people within clusters can be regarded as a special case of the in-fill asymptotic scenario, where more observations are collected at the same locations. Since these extra data are likely to be highly correlated, for some parameters, consistent estimates may not exist under the in-fill asymptotics (Zhang, 2004).

In this chapter, we establish large-sample properties of the JCEF estimator under the increasing domain context. Because we form the estimating functions based on pairwise observations, the asymptotic properties of $\hat{\boldsymbol{\theta}}_n$ will be based on properties of the extended pairwise random process

$$\mathbf{y}(k) \equiv \left(y_{si}, y_{tj}\right)^T, \tag{5.16}$$

where $k = (s, i, t, j) \in D_n$. Under appropriate conditions of the correlation decay rate for the process $\mathbf{y}(k)$, usually postulated by certain mixing conditions (Guyon, 1995), we expect to have "pseudo-independent" pairs when they are beyond a certain distance. In such cases, we can derive laws of large numbers (LLN) and central limit theorems for $\Psi_{B,n}(\boldsymbol{\theta}, d)$ and $\Psi_{W,n}(\boldsymbol{\theta})$ respectively, and then for $\boldsymbol{\Gamma}_n(\boldsymbol{\theta}, d) = (\Psi_{B,n}(\boldsymbol{\theta}, d), \Psi_{W,n}(\boldsymbol{\theta}))^T$. Moreover, by using standard GMM arguments (Hansen, 1982), we can show the consistency and asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ defined in equation (5.14).

Jenish and Prucha (2009) developed a set of limit theorems for random processes under rather general conditions of nonstationarity, unevenly spaced locations, and general forms of sample regions. We exploit those results to sketch large-sample properties for our JCEF estimator as follows.

5.4.1 Consistency

Consider a generic case of the composite estimating function

$$\Psi_{\mathcal{A},n}(\boldsymbol{\theta}) = \frac{1}{|D_{\mathcal{A},n}|} \sum_{k \in D_{\mathcal{A},n}} U_k(\mathbf{y}(k); \boldsymbol{\theta}),$$

where $\mathcal{A} \in \{B, W\}$.

We assume the following assumptions for the component score functions.

Assumption 1 The (possibly unevenly spaced) lattice $D \subset \mathbb{R}^2 \times \mathbb{Z}^+ \times \mathbb{R}^2 \times \mathbb{Z}^+$ is infinitely countable. All elements in D are located at distances of at least $d_0 > 0$ from each other. That is, $\rho(i,j) \geq d_0$, for all $i, j \in D$, where $\rho(i,j)$ is a distance metric for any two points i, $j \in D$. See a detailed definition of the distance metric in the Appendix C.

Assumption 2 $\{D_{\mathcal{A},n} : n \in \mathbb{N}\}$ is a sequence of arbitrary finite subsets of D, satisfying $|D_{\mathcal{A},n}| \to \infty$ as $n \to \infty$, for $\mathcal{A} \in \{B, W\}$.

Assumption 3 (Uniform $L_{2+\delta}$ integrability) Let $q_k = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||U_k(\mathbf{y}(k); \boldsymbol{\theta})||$. For some $\delta > 0$, $\lim_{e \to \infty} E q_k^{2+\delta} \mathbf{1}(||q_k|| > e) = 0$, for all $k \in D_n$.

Assumption 4 $E \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} ||\dot{U}(\mathbf{y}(k); \boldsymbol{\theta})|| < \infty$, for all $k \in D_n$.

Assumption 1 ensures that the increase of sample size is achieved by an expanding domain, thus it rules out the in-fill asymptotics. Assumption 2 guarantees that sequences of subsets $D_{B,n}$ and $D_{W,n}$ on which the process is generated, increase in cardinality. Assumptions 3 and 4 are regularity conditions for score functions. The uniform integrability condition in Assumption 3 is a standard moment assumption postulated in CLTs for one-dimensional processes. A sufficient condition for the uniform $L_{2+\delta}$ integrability of U_k is its uniform L_γ boundedness for some $\gamma > 2 + \delta$. A weaker assumption of L_1 integrability is sufficient for a LLN for U_k. Assumption 4 is a Lipschitz-type condition, implying that the score functions are L_0 stochastically equicontinuous, so that a ULLN can be obtained.

Lemma 3. *When the sample size increases with the increasing spatial domain, given assumptions 1 - 4, and certain appropriate mixing conditions for* $\mathbf{y}(k)$ *given in the Appendix D,*

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\Psi_{\mathcal{A},n}(\boldsymbol{\theta}) - E\Psi_{\mathcal{A},n}(\boldsymbol{\theta})|| \xrightarrow{p} 0, \ \ \textit{as} \ \ n \to \infty.$$

As shown in Jenish and Prucha (2009), a polynomial decay of the mixing coefficient for the process is enough for their results to hold, which is satisfied by Gaussian random processes considered in our chapter (Guyon, 1995).

Lemma 3 holds for $\Psi_{B,n}(\boldsymbol{\theta}, d)$ and $\Psi_{W,n}(\boldsymbol{\theta})$ respectively, so we can show easily that for any given positive-definite weight matrix W,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |Q_n(\boldsymbol{\theta}) - EQ_n(\boldsymbol{\theta})| \xrightarrow{p} 0, \quad \text{as } n \to \infty.$$

Consequently, we establish the consistency of the JCEF estimator in Theorem 1.

Theorem 3. *Under the same regularity conditions stated in Lemma 3, if the true parameter value* $\boldsymbol{\theta}_0$ *is the unique minimizer of* $EQ_n(\boldsymbol{\theta})$*, and* $\hat{\boldsymbol{\theta}}_n$ *minimizes* $Q_n(\boldsymbol{\theta})$*, then* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0, \quad \textit{as } n \to \infty.$

5.4.2 Asymptotic Normality

To derive the asymptotic distribution of the JCEF estimator, the following additional regularity conditions are needed.

Assumption 5 Let $\Lambda_n(\boldsymbol{\theta}) = Var\left\{\Gamma_n(\boldsymbol{\theta})\right\}$, $\lim_{n\to\infty} n\Lambda_n(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta})$, where $\Lambda(\boldsymbol{\theta})$ is a positive-definite matrix.

Assumption 6 $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} ||\dot{\Gamma}_n(\boldsymbol{\theta}) - E\dot{\Gamma}_n(\boldsymbol{\theta})|| \xrightarrow{p} 0$. Write $\lim_{n\to\infty} E\dot{\Gamma}_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})$ is a positive-definite information matrix.

Assumption 5 assumes that the variance of $\Gamma_n(\boldsymbol{\theta})$ is of order $O(n^{-1})$, which is also a standard assumption for the subsampling estimation of the covariance. Assumption 6 is a ULLN for the Hessian matrix $\dot{\Gamma}_n(\boldsymbol{\theta})$, which regulates the asymptotic variance of the estimator and can be obtained with the same regularity conditions on $\dot{\Gamma}_n(\boldsymbol{\theta})$ as those in Lemma 3.

Lemma 4. *Under the increasing domain framework, given Assumptions 1-6, we have*

$$\sqrt{n}\,\Gamma_n(\boldsymbol{\theta}) \xrightarrow{d} N(0, \Lambda(\boldsymbol{\theta})), \quad as \quad n \to \infty.$$

A sketch of the proof for Lemma 4 is given in the Appendix E. Then using the standard GMM arguments (Hansen, 1982), we establish the following theorem:

Theorem 4. *Under the increasing domain framework, given assumptions 1-6 and proper mixing conditions for* $\mathbf{y}(k)$*, we have*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(0, \Omega(\boldsymbol{\theta}_0)\Lambda(\boldsymbol{\theta}_0)\Omega^T(\boldsymbol{\theta}_0)), \quad as \quad n \to \infty,$$

where $\Omega(\boldsymbol{\theta}_0) = -[I^T(\boldsymbol{\theta}_0)W^{-1}I(\boldsymbol{\theta}_0)]^{-1}I^T(\boldsymbol{\theta}_0)W^{-1}$.

5.5 Simulation Experiments

5.5.1 Set Up

To assess the performance of the JCEF estimator developed for spatial-clustered data, two simulation experiments are conducted, one based on clustered Gaussian (Example 5) and the other based on multivariate probit model for clustered binary data (Example 6). We compare the estimation efficiency of the JCEF, weighted composite likelihood (WCL) and classic MLE. For convenience, the number of subjects within a cluster is fixed at 4 across all clusters. There are 100 clusters located on a 10×10 spatial grid with two coordinates spanning from 1 to 10, and in total there are 400 observations for a simulation data set.

For both Gaussian and binary simulation experiments, the marginal mean model is the same and specified with two covariates:

$$h(\mu_{si}) = \beta_1 x_s^1 + \beta_2 x_{si}^2, \tag{5.17}$$

where x_s^1 is a cluster-level covariate, and x_{si}^2 is a subject-level covariate, both generated from the uniform distribution in $(0, 1)$. The correlation matrix Σ consists of diagonal blocks, Σ_w, and off-diagonal blocks, Σ_{st}, given below:

$$\Sigma_w = (1-\rho) * \mathbf{I}_4 + \rho * \mathbf{J}_4,$$

$$\Sigma_{st} = \rho \exp(-\alpha||s-t||) * \mathbf{J}_4, \tag{5.18}$$

where $\rho \in (-1, 1)$ is the within-cluster correlation, and $\alpha \in (0, \infty)$ is a spatial scaling parameter for between-cluster spatial correlation. Σ_{st} represents an exponential decay correlation function. The vector of parameters of interest is $\boldsymbol{\theta} = (\beta_1, \beta_2, \rho, \alpha)$.

5.5.2 Spatial-Clustered Gaussian Data

For Gaussian data, four estimation methods are compared, namely, the maximum likelihood estimation (MLE), the weighted composite likelihood estimation (WCL), the JCEF approach with the weight estimated by parametric bootstrap (JCEF.p), and the JCEF approach with the weight estimated by subgroup sampling (JCEF.s). The subgroups in subsampling are chosen as overlapping subregions of 3×3 clusters.

In Table 5.1, two scenarios with different rates of spatial correlation decay are considered. The spatial scaling parameter α are set at 1 and 3, respectively; the larger the value, the faster the decay. Averaged parameter estimates across 200 rounds of simulation are reported as a summary measure for point estimation. As shown in Table 5.1, MLE, JCEF.p and JCEF.s are comparable in terms of bias, while WCL tends to have slightly larger biases, especially for the estimates of β_2, ρ and α. We also compare the root mean squared errors (RMSE) of parameter estimates across simulation replicates. Each RMSE is scaled by the corresponding parameter value, and is then summed together to obtain a measure of overall efficiency, termed as Total Scaled RMSE in Table 5.1. It is shown that, in general, MLE has the smallest RMSE, followed by JCEF.p and JCEF.s. WCL has the largest RMSE. This confirms that when the model assumption is satisfied, MLE achieves the highest efficiency, and for JCEF the weight estimated by the parametric bootstrap is more accurate than that estimated by subsampling. The WCL appears to be the least efficient among the four methods.

It is interesting to note that for the scaling parameter α, MLE yields larger RMSE as compared to JCEF.p in all three settings. One possible explanation is that for the spatial process, only one realization is observed, MLE may not achieve it asymptotic efficiency for some parameters in small samples. Similar phenomena have also been reported in Bai et al. (2011) and Zi (2009) . Generally speaking, results from this simulation experiment show that both JCEF approaches improve estimation efficiency over WCL and parametric JCEF sometimes even outperforms MLE for some parameters.

Table 5.1: Parameter estimates and root mean squared errors for spatial-clustered binary data. Total scaled RMSE is the summation of RMSEs scaled by the corresponding parameter values. Results are from 200 rounds of simulation

		Mean				RMSE				Total Scaled RMSE
		β_1	β_2	ρ	α	β_1	β_2	ρ	α	
Scenario 1	TRUE	1	1	0.8	1					
	MLE	0.9956	0.9971	0.7992	1.0500	0.0457	0.0530	0.0120	0.1657	0.28
	JCEF.p	0.9996	1.0039	0.7972	1.0012	0.0531	0.0655	0.0134	0.0913	0.23
	JCEF.s	1.0006	1.0084	0.7971	0.9513	0.0598	0.0957	0.0124	0.1773	0.35
	WCL	0.9995	0.9784	0.7898	0.9514	0.0539	0.2257	0.0132	0.2083	0.50
Scenario 2	TRUE	1	1	0.8	3					
	MLE	0.9946	0.9927	0.8000	3.0894	0.0309	0.0653	0.0107	0.3909	0.24
	JCEF.p	0.9934	0.9856	0.7903	2.8254	0.0383	0.0840	0.0111	0.2403	0.22
	JCEF.s	0.9915	0.9855	0.7997	3.0625	0.0553	0.1421	0.0160	0.4545	0.36
	WCL	0.9928	1.0206	0.7798	2.6564	0.0478	0.3429	0.0227	0.6108	0.62

5.5.3 Spatial-Clustered Binary Data

Now we compare the same four methods in the multivariate probit model for spatial-clustered binary data. Following Chan and Kuk (1997), we implement MLE using an EM algorithm, where we treat latent continuous variables z_{si} as missing data and apply Gibbs sampler to generate Monto Carlo samples from truncated multivariate normal distributions.

We consider two simulation scenarios and related results are summarized in Table 5.2. The spatial scaling parameter α is set to 1 and 3. As shown in Table 5.2, these four methods yield similar point estimates, and the estimation bias seems to decrease as the spatial correlation decays from scenario 1 to scenario 2. In general, both JCEF approaches have lower RMSE compared to WCL, as shown by the total scaled RMSE. Once again, JCEF.p appears to outperform the JCEF.s in terms of RMSE, similar to results for the Gaussian data in section 5.5.2.

It is interesting to observe from Tables 5.1 and 5.2 that, the JCEF gains more efficiency when the spatial dependence weakens, as indicated by the total scaled RMSE. The reason may be that when the spatial correlation diminishes, the data across different clusters become more variable, so that both within-cluster and between-cluster pairs become more distinct and informative. As a result, the weight matrix enables us to better account for such variations in estimation, leading to efficiency gains.

Table 5.2: Parameter estimates and root mean squared errors for spatial-clustered binary data. Total scaled RMSE is the summation of RMSEs scaled by the corresponding parameter values. Results are from 200 rounds of simulation

		Mean				RMSE				Total Scaled RMSE
		β_1	β_2	ρ	α	β_1	β_2	ρ	α	
Scenario 1	TRUE	1	-1	0.8	1					
	MLE	1.0502	-1.0585	0.7903	1.1541	0.1964	0.1971	0.0502	0.2772	0.73
	JCEF.p	1.0825	-1.0458	0.7995	1.2176	0.1999	0.2204	0.0480	0.3274	0.81
	JCEF.s	1.0624	-1.0669	0.7948	1.1710	0.2002	0.2158	0.0526	0.3382	0.82
	WCL	1.1066	-1.0314	0.7924	1.1980	0.2203	0.2532	0.0560	0.3459	0.88
Scenario 2	TRUE	1	-1	0.8	3					
	MLE	1.0030	-1.0143	0.7970	2.7650	0.2146	0.1493	0.0424	0.8641	0.70
	JCEF.p	1.0202	-1.0076	0.8056	2.9736	0.2230	0.1548	0.0440	0.9153	0.74
	JCEF.s	1.0290	-1.0006	0.8054	2.9578	0.2446	0.1762	0.0446	0.9184	0.78
	WCL	1.0343	-0.9929	0.8013	2.9201	0.3080	0.1761	0.0562	0.9924	0.89

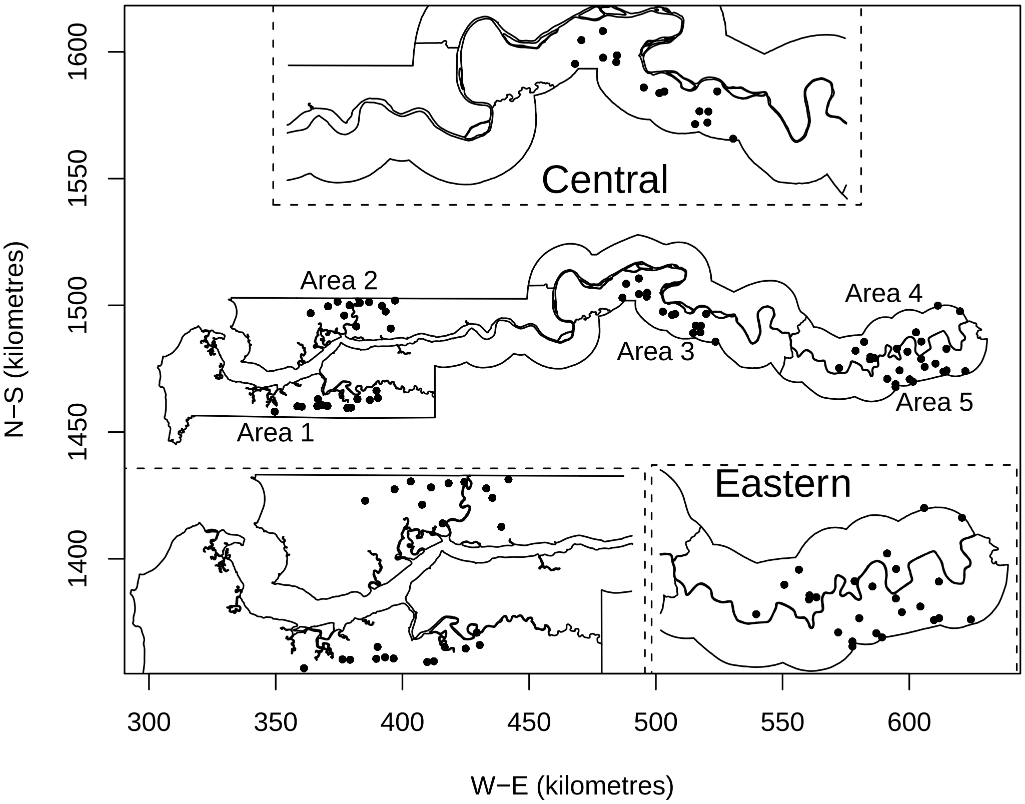
In summary, JCEF proposed for spatial-clustered data analysis improves the estimation efficiency over the existing WCL for both Gaussian and binary data. The extent of the improvement depends on the ability of the weight matrix to account for high-order dependences and how accurate the weight matrix can be estimated. It appears that the weight matrix plays a more significant role for Gaussian data then for binary data in efficiency improvement. The efficiency gain also increases as the spatial dependence among clusters diminishes.

5.6 Data Example

In this section, we illustrate an application of the JCEF to a real-world data. Spatial-clustered data are frequently encountered in spatial epidemiology. One of the key interests is to identify environmental risk factors associated with disease prevalences. Diggle et al. (2008) investigated the spatial variation in the prevalence of malaria among village resident children in Gambia. They developed a spatial generalized linear mixed model to account for the spatial correlation among the residuals at the village level and implemented it in a Bayesian MCMC framework. Thomson et al. (1999) used GEE to obtain regression estimates and accounted for the extra-binomial variation by a working correlation matrix with an exponential spatial correlation function. We now re-analyze this binary malaria incidence data using the proposed GeoCopula model.

Two thousand and thirty five children were randomly sampled from 65 villages along the Gambia river. A graphical representation of the spatial configuration of the sampled villages is given in Figure 5.2. Villages scatter into four distinct regions on the map and are labeled from Area 1 - Area 5. The pairwise distances between two villages range from 0.95 km to 273.3 km.

Figure 5.2: Spatial Configuration of the Sampled Villages



The response from each child is a binary indicator of the presence of malarial parasites in a blood sample. Covariates include child level variables: age, bed net use (NetUse) and whether the bed net is treated (Treated); and the village level variables: inclusion or exclusion from the primary health care (PHC) system and greenness of surrounding vegetation as derived from satellite information (Green). In the final model suggested by Diggle et al. (2008), the five-level area dummy variables (Area) are also included to adjust for the regional effects, however, information about the partition of area 4 and area 5 is not available in the data given in R package Gambia (R Development Core Team, 2010), nor can it inferred from the map. As a result, we had to combine Area 4 and Area 5 into one region in our analysis.

For child i in village s, let the binary random variable Y_{si} denote the presence of malaria (1 for yes; 0 for no). Let $p_{si} = E(Y_{si})$ be the probability of the malaria infection. Then the probit model is:

$$\begin{aligned} p_{si} &= \Phi(\mathbf{x}_{si}^T \boldsymbol{\beta}) \\ &= \Phi(\beta_0 + \beta_1 Age + \beta_2 NetUse + \beta_3 Treated + \beta_4 Green + \beta_5 PHC + \beta_6 Area), \end{aligned}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_6)^T$ is the vector of corresponding regression coefficients.

The correlation matrix Σ is specified similar to equation (5.10). That is, the within-village correlation is specified as compound symmetry, and the between-village correlation is given by an exponential decay function of distance between two villages, as in equation (5.18). Note that exponential correlation decay is a special case of the Matérn class with the smoothness parameter $\nu = 0.5$.

The choice of the distance lag d is based on the level of the empirical spatial correlation. Results from Diggle et al. (2008) show that the spatial dependence decays at a fairly fast rate, so pairs of villages within 3 miles are used to construct the pairwise composite score function.

To create subsamples for the weight matrix estimation and the subsequent standard error estimation, overlapping subregions within radius of 10 km is used as the sub-blocks. Given the fact that the villages scatter into four major regions, the subsampling is carried out in each region and then combined to form the overall subsample. In this way, spatial dependence patterns from different regions are all represented in the subsample.

Results are summarized in Table 5.3, including JCEF estimates and their corresponding 95% confidence intervals. JCEF found that Age (in years) is positvely associated with malaria prevalance, and the bed net use and the treatment of the bed net tend to reduce the risk, although marginally significant. Prevalance in the eastern region is significatly higher than the rest of regions. The 95% confidence interval for ρ is $(0.4318, 0.6447)$. The confidence interval for the spatial scaling parameter α is $(0.3488, 0.4750)$, corresponding to an approximately 65% decrease in dependence with one kilometer increase in distance. This means that the spatial variation operates on a relatively small scale. On the other hand, the WCL approach yield larger confidence intervals than JCEF for most of the variables, hence fail to identify some significant covariate effects (e.g. Age).

The findings based on our JCEF approach are consistent with those in Diggle et al. (2008). For example, in the final model proposed by Diggle et al., age and being in area 5 are positively associated with the risk of malaria. However, it is important to note that results in this chapter are not directly comparable to results in Diggle et al. (2008). The GeoCopula model provides population-level effect estimates, while the spatial linear mixed model used in Diggle et al. (2008) is a cluster-specific model. In addition, we do not have the specific boundary information for creating the same five regions as done in their analysis. Also for identification, correlation ρ instead of the variance is estimated in GeoCopula model. Our spatial scaling parameter α corresponds to the inverse of their scaling parameter as well.

Table 5.3: Parameter estimates and 95% confidence intervals for the malaria data, estimated from JCEF, WCL and Diggle's final spatial GLMM model. Note (*): Age from Diggle's model is in days and α from Diggle's model corresponds to the inverse of α from JCEF and WCL. We combine Area 4 & 5 because of insufficient location information.

	JCEF			WCL				Diggle		
	Estimate	95% C.I.		Estimate	95% C.I.			Estimate	95% C.I.	
Int	-0.0221	-0.2590	0.2148	-0.1138	-0.8224	0.5949		-0.1312	-2.9665	2.6243
Age (year)	0.3723	0.0750	0.6696	0.5079	-0.1488	1.1646	Age (Day)	0.0007*	0.0005*	0.0009*
NetUse	-0.2156	-0.5042	0.0729	-0.1994	-0.7192	0.3205		-0.3579	-0.6731	-0.0420
Treated	-0.1690	-0.3967	0.0587	-0.0450	-0.2679	0.1778		-0.3295	-0.7539	0.0884
Green	-0.0557	-0.2672	0.1558	0.0037	-0.2882	0.2956		-0.0201	-0.0857	0.0479
PHC	-0.2727	-0.7106	0.1651	-0.5945	-0.8438	-0.3452		-0.3448	-0.7879	0.1299
Area 2	-0.1948	-0.4206	0.0310	0.1087	-0.1385	0.3559		-0.3247	-1.1442	0.5102
Area 3	-0.4320	-0.6328	-0.2312	-0.4325	-0.7108	-0.1541		-0.5321	-1.4086	0.5586
Area 4 & 5	0.5145	0.2467	0.7823	0.4057	0.0499	0.7614	Area 4	1.0494*	-0.1095*	2.4253*
							Area 5	1.3096	0.1648	2.6064
σ^2	0.5401	0.4318	0.6447	0.5444	0.4288	0.6554		0.5856	0.3118	1.0502
α	0.4070	0.3488	0.4750	0.2567	0.2068	0.3186	$1/\alpha$	1.0841*	0.0795*	2.7846*

5.7 Discussion

In this chapter, we have developed an innovative statistical modeling and estimation methodology for high-dimensional spatial-clustered data. The proposed GeoCopula model provides population-level regression parameter estimates and allows the modeling of flexible within-cluster and between-cluster spatial dependence structures. The newly proposed JCEF procedure provides more efficient parameter estimates than conventional pairwise composite likelihood. The novelty lies in the incorporation of the correlation among different pairs of composite scores, and then the integration of them into a quadratic objective function for estimation. This strategy can better utilize the information from the two groups than just summing them together as currently used in the WCL. As shown in various simulation studies, the JCEF method gains a significant higher amount of efficiency over the WCL approach for both Gaussian and binary spatial data, and is very comparable to MLE for Gaussian data.

The GeoCopula model is built upon the multivariate Gaussian copula dependence model. Bárdossy (2006); Kazianka and Pilz (2010) discussed some of the disadvantages of the Gaussian copula, such as the tail independence and symmetrical correlation structures in the lower and upper tails and provides some modifications based on the Gaussian copula. They propose some more flexible copulas constructed from Gaussian copulas to address these issues. It is not clear if these modifications are necessary to reflect spatial dependences.

The JCEF method is a general methodology that can be applied to many established modeling framework for spatial-clustered data. The current JCEF method is built upon a general class of multivariate exponential dispersion distributions, which separates the marginal and covariance specifications, hence provides a great flexibility in model formulation. However, a similar JCEF approach can be adapted to the generalized linear mixed models, which offers a cluster-specific parameter interpretations as opposed to GeoCopula models, see e.g. Varin et al. (2005).

Till now, Bayesian methods are predominantly used in the analysis of spatial/temporal data, owing to the numerical limitations of the traditional likelihood methods. The method proposed in this chapter offers a competitive alternative for analyzing high-dimensional data from a frequentist perspective.

The construction of the JCEF requires specifying a distance lag for pair inclusions. This can be achieved by setting the lag to a pre-determined value, based on geographic boundaries, sample size considerations, or substantial research. It can also be set, from a statistical point of view, to maximize the information of the corresponding estimating functions. As proposed in Bevilacqua et al. (2011), the optimal distance lag can be chosen to maximize a certain matrix norm of the Godambe information of the selected composite score functions. When there is no data replicate, the evaluation of the Godambe information can be carried out using subsampling. Usually a grid search is adopted to locate the optimal distance lag value from a pool of potential candidate. This procedure can be easily incorporated into the current JCEF framework to achieve better efficiency if the related computational burden is manageable.

As pointed out in Bai et al. (2011), the quadratic objective function also provides a way for a goodness-of-fit test of the mean-zero model assumption. H_0 : $E\left\{\Gamma_n(\boldsymbol{\theta})\right\} = 0$. Since $\hat{\boldsymbol{\theta}}$ is obtained by an over-identified $\Gamma_n(\boldsymbol{\theta})$, $Q_n(\hat{\boldsymbol{\theta}})$ falls in the 'over-identifying restriction' test by Hansen (1982), who proved that the asymptotic distribution of $Q_n(\hat{\boldsymbol{\theta}})$ is χ^2 with degrees of freedom equal to the number of estimating functions minus the number of parameters, in our case equal to p.

CHAPTER VI

Future Work

This dissertation has focused on the development of composite likelihood methodology for high-dimensional spatio-temporal and spatial-clustered data. Due to the complex dependence structures and large scales of such data, innovative statistical modeling and estimating procedures that are both statistically and computationally efficient are in eminent need.

For spatio-temporal data, a new joint composite estimating function (JCEF) approach has been proposed in Chapter III to estimate high-dimensional spatio-temporal covariance structures. JCEF builds upon the popular pairwise marginal composite score functions, and further improves the estimating efficiency by incorporating the high-order correlations among the pairwise scores through a weight matrix. The resulting target function takes a similar form as the generalized method of moments (Hansen, 1982), hence preserves some desirable properties of GMM. For example, the JCEF estimator is consistent and asymptotically normally distributed. Also the over-identified set of estimating functions offers a means for hypothesis testing. This is potentially useful since different types of the spatio-temporal covariance structures can be selected based on the test statistic.

Specifically, a goodness-of-fit statistic can be derived to test the mean-zero model assumption, $H_0 : E\{\Gamma_n(\boldsymbol{\theta})\} = 0$. This can be used for testing, for example, the separability structure of the covariance matrix. Since $\hat{\boldsymbol{\theta}}_n$ is obtained by an over-identified estimating function $\Gamma_n(\boldsymbol{\theta})$, $Q_n(\hat{\boldsymbol{\theta}}_n)$ falls in the 'over-identifying restriction' test by Hansen (1982), who proved that the asymptotic distribution of $Q_n(\hat{\boldsymbol{\theta}}_n)$ is χ^2 with degrees of freedom equal to the number of estimating functions minus the number of parameters. However, many researchers have pointed out that the first-order asymptotic theory often provides inadequate approximations to the distributions of the test statistics obtained from GMM estimators; see, for example, a special issue of the Journal of Business & Economics Statistics (July 1996). To improve inference, a number of alternative estimators have been suggested. These include empirical likelihood (Qin and Lawless, 1994; Owen, 1988; Imbens, 1997), modified bootstrap procedures (Hall and Horowitz, 1996), and the continuous updating estimator (Hansen et al., 1996). Qu et al. (2000) used the latter approach to construct the QIF and showed that the finite-sample distribution of the objective function agrees well with the asymptotic counterpart. Performances of these goodness-of-fit methods under the JCEF framework for spatio-temporal data is worth further exploration.

In the discussion section of Chapter IV, it was briefly mentioned that composite likelihoods built from triplets of observations have higher efficiency than those based on pairs. This indicates that in some situations, high-order marginal density functions can work better in estimation. However, given a large number of observations, the number of all possible triplets is enormous, in the order of $O(n^3)$ in comparison to that of $O(n^2)$ for pairwise CL. This leads to the problem of selecting informative lower-dimensional likelihoods (pairs/triplets) to reduce the number of terms in CL. In the current composite likelihood literature, pairs of observations are usually selected by a fixed distance lag: pairs within the lag are included. Bevilacqua et al. (2011) has proposed to select an optimal lag using a criterion based on the Godambe information matrix. Their method appears computationally burdensome in spatial data analysis, where there are usually no replicates. There is a clear need of developing new ways, which are both statistically sound and computationally fast, of determining which components to include and which to discard in the estimation.

High-dimensional spatial-clustered data is studied in Chapter V. A novel modeling strategy termed as the GeoCopula regression model is proposed. The GeoCopula model yields population-level regression parameters and explicitly models the spatial and within-cluster correlations. Two directions of research are worth pursuing within the GeoCopula framework. The first interesting problem is to develop more flexible copula models based on the Gaussian copula but allows for tail dependence and asymmetric correlation structures. Retaining the Gaussian copula preserves many nice dependence properties required in spatial data analysis. Work may be focused on how to generalize the Gaussian copula constructions. Bárdossy (2006); Bárdossy and Li (2008) has done some work in this regard. They build χ^2 copulas from the Gaussian copula which circumvent the tail independence and symmetric correlation patterns of the latter. It will be interesting to see how their proposed copulas perform in spatial data analysis. We note in passing that when different copulas are available, the goodness-of-fit test suggested previously can be a potential method for model selection.

The second interesting problem is to extend the estimation of the GeoCopula regression model by increasing the dimension of the composite likelihoods. In

Chapter V, only bivariate density functions are utilized in estimation. Efficiency may be further recovered by using higher-dimensional densities. In this case, the vine-type pair copula construction can be useful, since copula densities of any dimensions can be decomposed into bivariate densities. R software packages are readily available for the D-vine and C-vine decomposition for relatively low dimensional likelihood calculations.

One of the reasons for advocating the use of composite likelihood is its robustness against model misspecification, since only lower dimensional distributions need to be correctly specified for valid statistical inference. With the construction of a quadratic objective function, JCEF is even robust to outliers (Qu and Song, 2004). It will be worth investigating how JCEF performs with misspecified high-order distribution structures and/or in the presence of outliers.

APPENDICES

APPENDIX A

Definition of the distance metric ρ in spatio-temporal setting

The distance between two pairwise differences $d(k_1)$ and $d(k_2)$ defined in equation (3.1) depends on configurations of four points in the spatio-temporal domain $\mathbb{R}^2 \times \mathbb{R}^+$. Denote the coordinates of one point by (s, t). The distance between two points $p_1 = (s_1, t_1)$ and $p_2 = (s_2, t_2)$ in $\mathbb{R}^2 \times \mathbb{R}^+$ is defined as $\tau(p_1, p_2) = \max\{||s_1 - s_2||, |t_1 - t_2|\}$. Let $k_1 = (p_1, p_1^{'})$ and $k_2 = (p_2, p_2^{'})$. Then the distance between two points in $D \subset \mathbb{R}^2 \times \mathbb{R}^+ \times \mathbb{R}^2 \times \mathbb{R}^+$ is defined as $\rho(k_1, k_2) = \min\{\tau(p_1, p_2), \tau(p_1, p_2^{'}), \tau(p_1^{'}, p_2), \tau(p_1^{'}, p_2^{'})\}$, i.e., the minimum distance of two points in sets $(p_1, p_1^{'})$ and $(p_2, p_2^{'})$. The distance between any subsets $U, V \subset D$ is defined as $\rho(U, V) = \min\{\rho(i, j) : i \in U, j \in V\}$.

APPENDIX B

Proof of Lemma 1

Lemma 1 states a CLT for $\Gamma_n(\boldsymbol{\theta})$, which is comprised of three estimating functions based on different groups of pairwise differences with varying numbers of terms. The three groups of pairwise differences are subseries of $d(k)$, hence satisfy the same mixing conditions in Assumption 6 imposed on $d(k)$. In addition, $|D_{S,n}|$, $|D_{T,n}|$, and $|D_{C,n}|$ are of the same order $O(n)$, making it possible to use a common scaling factor to unify the convergence rates.

We prove the asymptotic normality of $\Gamma_n(\boldsymbol{\theta})$ through the Cramer-Wold device. For ease of argument, we work on sums of component score functions instead of means. Define $\Gamma_n^*(\boldsymbol{\theta}) = (\Psi_{S,n}^{*T}(\boldsymbol{\theta}), \Psi_{T,n}^{*T}(\boldsymbol{\theta}), \Psi_{C,n}^{*T}(\boldsymbol{\theta}))^T$, where $\Psi_{\mathcal{A},n}^*(\boldsymbol{\theta}) = |D_{\mathcal{A},n}|\Psi_{\mathcal{A},n}(\boldsymbol{\theta})$ for $\mathcal{A} \in \{S, T, C\}$. The aim is to prove that for arbitrary constants c_1, c_2, and c_3, the linear combination

$$c_1\Psi_{S,n}^*(\boldsymbol{\theta}) + c_2\Psi_{T,n}^*(\boldsymbol{\theta}) + c_3\Psi_{C,n}^*(\boldsymbol{\theta})$$

is asymptotically Gaussian. Define

$$G_n(\boldsymbol{\theta}) \equiv c_1\Psi^*_{S,n}(\boldsymbol{\theta}) + c_2\Psi^*_{T,n}(\boldsymbol{\theta}) + c_3\Psi^*_{C,n}(\boldsymbol{\theta}) = \mathbf{c}^T\Gamma^*_n(\boldsymbol{\theta}),$$

where $\mathbf{c} = (\mathbf{c}_1; \mathbf{c}_2; \mathbf{c}_3)^T$, a $3r$ by r matrix with $\mathbf{c}_i = c_i I_r$, $i = 1, 2, 3$, and I_r is the r by r identity matrix. Let $Var(\Gamma^*_n(\boldsymbol{\theta})) = \Sigma^*_n(\boldsymbol{\theta})$, $\Sigma_{G,n}(\boldsymbol{\theta}) \equiv Var(G_n(\boldsymbol{\theta})) = \mathbf{c}^T\Sigma^*_n(\boldsymbol{\theta})\mathbf{c}$.

Write

$$\begin{aligned} G_n(\boldsymbol{\theta}) &= \sum_{i\in D_{S,n}} c_1 f_i(d(i);\boldsymbol{\theta}) + \sum_{j\in D_{T,n}} c_2 f_j(d(j);\boldsymbol{\theta}) + \sum_{l\in D_{C,n}} c_3 f_l(d(l);\boldsymbol{\theta}) \quad \text{(B.1)} \\ &\equiv \sum_{k\in D_n} h_k(d(k);\boldsymbol{\theta}), \end{aligned}$$

where

$$h_k(d(k);\boldsymbol{\theta}) = \begin{cases} c_1 f_k(d(k);\boldsymbol{\theta}), & \text{if} \quad k \in D_{S,n}; \\ c_2 f_k(d(k);\boldsymbol{\theta}), & \text{if} \quad k \in D_{T,n}; \\ c_3 f_k(d(k);\boldsymbol{\theta}), & \text{if} \quad k \in D_{C,n}. \end{cases}$$

Equation (B.1) simply multiplies each set of estimating functions by a constant and sums them together. Then given Assumptions 1-4 and 6-7, according to Theorem 1 in Jenish and Prucha (2009),

$$\Sigma^{-1/2}_{G,n}(\boldsymbol{\theta})G_n(\boldsymbol{\theta}) \sim N(0, I_r), \quad \text{as} \quad n \to \infty.$$

Note that Assumption 4 is imposed on f_k, which also applies to h_k, since h_k differs from f_k by a multiplicative constant. Assumption 7 implies the convergence of $n^{-1}\Sigma^*_n(\boldsymbol{\theta})$ to a positive-definite constant matrix, provided that $|D_{S,n}|$, $|D_{T,n}|$, and $|D_{C,n}|$ are of order $O(n)$.

Since c_1, c_2, and c_3 are arbitrary constants, by Cramer-Wold device, we obtain,

$$(\Sigma_n^*(\boldsymbol{\theta}))^{-1/2}\Gamma_n^*(\boldsymbol{\theta}) \sim N(0, I_{3r}), \text{ as } n \to \infty.$$

Let $B = diag\{\frac{1}{|D_{S,n}|}I_r, \frac{1}{|D_{T,n}|}I_r, \frac{1}{|D_{C,n}|}I_r\}$. Then $\Gamma_n(\boldsymbol{\theta}) = B\Gamma_n^*(\boldsymbol{\theta})$, whose asymptotic normality follows immediately.

APPENDIX C

Definition of the Distance Metric ρ for spatial-clustered data

The distance between two extended pairs $\mathbf{y}(k_1)$ and $\mathbf{y}(k_2)$ defined in equation (5.16) depends on configurations of four points in the spatial-clustered domain $\mathbb{R}^2 \times \mathbb{Z}$. Denote the coordinates of one point by $(\mathbf{s}, i)$, where $\mathbf{s}$ is the vector of spatial coordinates, and i is the index within a cluster. The distance between two points $p_1 = (\mathbf{s}, i)$ and $p_2 = (\mathbf{t}, j)$ in $\mathbb{R}^2 \times \mathbb{Z}$ is defined as $\tau(p_1, p_2) = ||\mathbf{s} - \mathbf{t}|| + I(i - j \neq 0)d_0$, where $|| \bullet ||$ is the Euclidean distance in $\mathcal{R}^2$. Defined in this way, the distance between any two different observations consists of two parts. The first part is the spatial distance between two clusters they reside in, and the second part is d_0 if they have different indices within clusters. This ensures that different observations are at least d_0 distance away. Let $k_1 = (p_1, p_2)$ and $k_2 = (p_1', p_2')$. Then the distance between two points in $D \subset \mathbb{R}^2 \times \mathbb{Z} \times \mathbb{R}^2 \times \mathbb{Z}$ is defined as

$$\rho(k_1, k_2) = \min\{\tau(p_1, p_2), \tau(p_1, p_2'), \tau(p_1', p_2), \tau(p_1', p_2')\},$$

i.e., the minimum distance of two points in sets (p_1, p_2) and (p_1', p_2'). The distance between any subsets $U, V \subset D$ is defined as $\rho(U, V) = \min\{\rho(i, j) : i \in U, j \in V\}$.

APPENDIX D

Mixing Conditions for $\mathbf{y}(k)$

To regulate the dependence structure of $\mathbf{y}(k)$ defined in equation (5.16), we impose some α-mixing conditions on $\mathbf{y}(k)$. Let U and V be two subsets of D_n, and let $\sigma(U) = \sigma\{\mathbf{y}(k); k \in U\}$ be the σ-algebra generated by random variables $\mathbf{y}(k), k \in U$. Define

$$\alpha(U,V) = \sup\{|P(A \cap B) - P(A)P(B)|; A \in \sigma(U), B \in \sigma(V)\}.$$

Then this α-mixing coefficient for the random field $\{\mathbf{y}(k), k \in D_n\}$ is defined as:

$$\alpha(k,l,m) = \sup\{\alpha(U,V), |U| < k, |V| < l, \rho(U,V) \geq m\},$$

with $k, l, m \in \mathbb{N}$ and $\rho(U,V)$ the distance between sets U and V, defined in Appendix C. We need the following conditions similar to those stated in Assumption 3 (Jenish and Prucha, 2009).

Mixing Conditions The process $\{\mathbf{y}(k), k \in D_n\}$ satisfies the following mixing conditions in an a-dimensional space:

(a) $\sum_{m=1}^{\infty} m^{a-1}\alpha(1,1,m)^{\delta/(2+\delta)} < \infty$, for some $\delta > 0$,

(b) $\sum_{m=1}^{\infty} m^{a-1}\alpha(k,l,m) < \infty$ for $k + l \leq 4$,

(c) $\alpha(1,\infty,m) = O(m^{-a-\epsilon})$ for some $\epsilon > 0$.

This requires a polynomial decay of the α-mixing coefficient, which can be shown to hold for Gaussian processes, a special case of the Gibbs fields (Winkler, 1995; Doukhan, 1994).

APPENDIX E

Proof of Lemma 3

Lemma 3 states a CLT for $\Gamma_n(\boldsymbol{\theta})$, which is comprised of two estimating functions based on different groups of pairwise differences with varying numbers of terms. The two groups of pairwise differences are subseries of $\mathbf{y}(k)$, hence satisfy the same mixing conditions in in Appendix C. In addition, $|D_{B,n}|$ and $|D_{W,n}|$ are of the same order, making it possible to use a common scaling factor to unify the convergence rates.

We prove the asymptotic normality of $\Gamma_n(\boldsymbol{\theta})$ through the Cramer-Wold device. For ease of argument, we work on sums of component score functions instead of means. Define $\Gamma_n^*(\boldsymbol{\theta}) = (\Psi_{B,n}^{*T}(\boldsymbol{\theta}), \Psi_{W,n}^{*T}(\boldsymbol{\theta}))^T$, where $\Psi_{\mathcal{A},n}^*(\boldsymbol{\theta}) = |D_{\mathcal{A},n}|\Psi_{\mathcal{A},n}(\boldsymbol{\theta})$ for $\mathcal{A} \in \{B, W\}$.

The aim is to prove that for arbitrary constants c_1 and c_2, the linear combination

$$c_1\Psi_{B,n}^*(\boldsymbol{\theta}) + c_2\Psi_{W,n}^*(\boldsymbol{\theta})$$

is asymptotically Gaussian. Define

$$G_n(\boldsymbol{\theta}) \equiv c_1 \Psi^*_{B,n}(\boldsymbol{\theta}) + c_2 \Psi^*_{W,n}(\boldsymbol{\theta}) = \mathbf{c}^T \Gamma^*_n(\boldsymbol{\theta}),$$

where $\mathbf{c} = (\mathbf{c}_1; \mathbf{c}_2)^T$, a $2p$ by p matrix with $\mathbf{c}_i = c_i I_p$, $i = 1, 2$, and I_p is the p by p identity matrix. Let $Var(\Gamma^*_n(\boldsymbol{\theta})) = \Lambda^*_n(\boldsymbol{\theta})$, $\Lambda_{G,n}(\boldsymbol{\theta}) \equiv Var(G_n(\boldsymbol{\theta})) = \mathbf{c}^T \Lambda^*_n(\boldsymbol{\theta})\mathbf{c}$.

Write

$$\begin{aligned} G_n(\boldsymbol{\theta}) &= \sum_{i \in D_{B,n}} c_1 U_i(\mathbf{y}(i); \boldsymbol{\theta}) + \sum_{j \in D_{W,n}} c_2 U_j(\mathbf{y}(j); \boldsymbol{\theta}) \\ &\equiv \sum_{k \in D_n} h_k(\mathbf{y}(k); \boldsymbol{\theta}), \end{aligned} \tag{E.1}$$

where

$$h_k(\mathbf{y}(k); \boldsymbol{\theta}) = \begin{cases} c_1 U_k(\mathbf{y}(k); \boldsymbol{\theta}), & \text{if} \quad k \in D_{B,n}; \\ c_2 U_k(\mathbf{y}(k); \boldsymbol{\theta}), & \text{if} \quad k \in D_{W,n}. \end{cases}$$

Equation (E.1) simply multiplies each set of estimating functions by a constant and sums them together. Then given Assumptions 1- 6 and the mixing conditions stated in Appendix D, according to Theorem 1 in Jenish and Prucha (2009),

$$\Lambda^{-1/2}_{G,n}(\boldsymbol{\theta}) G_n(\boldsymbol{\theta}) \sim N(0, I_p), \quad \text{as} \quad n \to \infty.$$

Note that Assumption 3 and 4 are imposed on U_k, which also apply to h_k, since h_k differs from U_k by a multiplicative constant. Assumption 6 implies the convergence of $n^{-1}\Lambda^*_n(\boldsymbol{\theta})$ to a positive-definite constant matrix, provided that $|D_{B,n}|$ and $|D_{W,n}|$ are of the same order.

Since c_1 and c_2 are arbitrary constants, by Cramer-Wold device, we obtain,

$$(\Lambda_n^*(\boldsymbol{\theta}))^{-1/2}\,\Gamma_n^*(\boldsymbol{\theta}) \sim N(0, I_{2p}), \text{ as } n \to \infty.$$

Let $A = diag\{\frac{1}{|D_{B,n}|}I_p, \frac{1}{|D_{W,n}|}I_p\}$. Then $\Gamma_n(\boldsymbol{\theta}) = A\Gamma_n^*(\boldsymbol{\theta})$, whose asymptotic normality follows immediately.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abramowitz, M. and Stegun, I. (1972), *Handbook of mathematical functions*, Dover Publications.

Albert, P. and McShane, L. (1995), "A generalized estimating equations approach for spatially correlated binary data: with an application to the analysis of neuroimaging data," *Biometrics*, 51, 627–638.

Anselin, L. and Griffith, D. A. (1988), "Do spatial effects really matter in regression analysis," *Papers, Regional Science Association*, 65, 11–34.

Bai, Y., Song, P. X.-K., and Raghunathan, T. (2011), "Joint composite estimating functions in spatio-temporal models," *Journal of the Royal Statistical Society, Series B*, In revision.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008), "Gaussian predictive process models for large spatial data sets," *Journal of the Royal Statistical Society: Series B*, 70, 825–848.

Bárdossy, A. (2006), "Copula-based geostatistical models for groundwater," *Water Resources Research*, 42.

Bárdossy, A. and Li, J. (2008), "Geostatistical interpolation using copulas," *Water Resources Research*, 44.

Bedford, T. and Cooke, R. (2001), "Probability density decomposition for conditionally dependent random variables modeled by vines," *Annals of Mathematics and Artificial Intelligence*, 32, 245–268.

Bedford, T. and Cooke, R. (2002), "Vines - a new graphical model for dependent random variables," *Annals of Statistics*, 30, 1031–1068.

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

Bevilacqua, M., Mateu, J., Porcu, E., Zhang, H., and Zini, A. (2010), "Weighted composite likelihood-based tests for space-time separability of covariance functions," *Statistics and Computing*, 20, 283–293.

Bevilacqua, M., Gaetan, C., Mateu, J., and Porcu, E. (2011), "Estimating space and space-time covariance functions: a weighted composite likelihood approach," *Journal of the American Statistical Association (to appear)*.

Carey, V., Zeger, S., and Diggle, P. (2003), "Modelling multivariate binary data with alternating logistic regressions," *Biometrika*, 80, 517–526.

Carlstein, E. (1987), "The use of subseries values for estimating the variance of a general statistic from a stationary sequence," *The Annals of Statistics*, 14, 1171–1179.

Chaix, B., Merlo, J., and Chauvin, P. (2005), "Comparison of a spatial approach with the multilevel approach for investigating place effects on health: the example of healthcare utilisation in France," *Journal of Epidemiology and Community Health*, 59, 517–526.

Chan, J. S. and Kuk, A. Y. (1997), "Maximum likelihood estimation for probit-linear mixed models with correlated random effects," *Biometrics*, 53, 86–97.

Cox, D. (1972), "The analysis of multivariate binary data," *Applied Statistics*, 21, 113–120.

Cressie, N. and Huang, H.-C. (1999), “Classes of nonseparable, spatio-temporal stationary covariance functions,” *Journal of the American Statistical Association*, 94, 1330–1340.

Cressie, N. and Jahannesson, G. (2008), “Fixed rank kriging for very large spatial data sets,” *Journal of the Royal Statistical Society: Series B*, 70, 209–226.

Cressie, N. A. (1993), *Statistics for Spatial Data*, Wiley, revised edn.

Curriero, F. C. and Lele, S. (1999), “A composite likelihood approach to semivariogram estimation,” *Journal of Agricultural, Biological, and Environmental Statistics*, 4, 9–28.

Davis, R. A. and Yau, C.-Y. (2011), “Comments on Pairwise Likelihood in Time Series Models,” *Statistica Sinica*, 21, 255–277.

Diez Roux, A. V., Auchincloss, A. H., Franklin, T. G., Raghunathan, T., Barr, R. G., Kaufman, J., Astor, B., and Keeler, J. (2008), “Long-term exposure to ambient particulate matter and prevalence of subclinical atherosclerosis in the multi-ethinic study of atherosclerosis,” *American Journal of Epidemiology*, 167, 667–675.

Diggle, P., Moyeed, R., Rowlingson, B., and Thomson, M. (2008), “Childhood malaria in the Gambia: a case-study in model-based geostatistics,” *Applied Statistics*, 51, 493–506.

Doukhan, P. (1994), *Mixing: Properties and Examples*, Springer.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009), “Improving the performance of predictive process modeling for large data sets,” *Computational Statistics and Data Analysis*, 53.

Fuentes, M. (2002), “Spectral methods for nonstationary spatial processes,” *Biometrika*, 89, 197–210.

Fuentes, M. (2007), “Approximate likelihood for large irregularly spaced spatial data," *Journal of the American Statistical Association*, 102, 321–331.

Furrer, R., Genton, M. G., and Nychka, D. (2006), "Covariance Tapering for Interpolation of Large Spatial Datasets," *Journal of Computational and Graphical Statistics*, 15, 502–523.

Gelfand, A. E., KIM, H.-J., Sirmans, C. F., and Banerjee, S. (2003), "Spatial Modeling With Spatially Varying Coef. cient Processes," *Journal of the American Statistical Association*, 98, 387–396.

Genton, M. G. (2007), "Separable approximations of space-time covariance matrices," *Environmetrics*, 18, 681–695.

Gneiting, T. (2002), "Nonseparable, stationary covariance functions for space-time data," *Journal of the American Statistical Association*, 97, 590–600.

Godambe, V. (1991), *Estimating Functions*, Oxford University Press, New York.

Godambe, V. P. and Heyde, C. (1987), "Quasi-Likelihood and Optimal Estimation," *International Statistical Review*, 55, 231–244.

Gotway, C. and Stroup, W. (1997), "A generalized linear model approach to spatial data analysis," *Journal of Agricultural, Biological, and Environmental Statistics*, 2, 157–178.

Grady, S. C. (2010), "Racial residential segregation impacts on low birth weight using improved neighborhood boundary definitions," *Spatial and Spatio-Temporal Epidemiology*, 1, 239–249.

Guan, Y. (2006), "A composite likelihood approach in fitting spatial point process models," *Journal of the American Statistical Association*, 101, 1502–1512.

Guan, Y., Sherman, M., and Calvin, J. A. (2004), "Nonparametric Test for Spatial Isotropy Using Subsampling," *Journal of the American Statistical Association*, 99, 810–821.

Guyon, X. (1982), "Parameter estimation for a stationary process on a *d*-dimensional lattice," *Biometrika*, 69, 95–105.

Guyon, X. (1995), *Random fields on a network: modeling, statistics and applications*, Springer-Verlag.

Haas, T. C. (1995), "Local prediction of a spatio-temporal process with an application to wet sulfate deposition," *Journal of the American Statistical Association*, 90, 1189–1199.

Hall, P. and Horowitz, J. L. (1996), "Bootstrap critical values for tests based on generalized-method-of-moments estimators," *Econometrika*, 64, 891–916.

Hansen, L. P. (1982), "Large sample properties of generalized method of moments estimators," *Econometrica*, 50, 1029–1054.

Hansen, L. P., Heaton, P., and Yaron, A. (1996), "Finite-sample properties of some alternative GMM estimators," *Journal of Business and Economic Statistics*, 14, 262–280.

Haslett, J. and Raftery, A. (1989), "Space-time modeling with long-memory dependence: assessing Ireland's wind power resource," *Applied Statistics*, 38, 1–50.

Heagerty, P. J. and Lele, S. R. (1998), "A composite likelihood approach to binary spatial data," *Journal of the American Statistical Association*, 93, 1099–1111.

Heagerty, P. J. and Lumley, T. (2000), "Window subsampling of estimating functions with application to regression models," *Journal of the American Statistical Association*, 95, 197–211.

Henderson, H. V. and Searle, S. R. (1981), "On deriving the inverse of a sum of matrices," *SIAM Review*, 23.

Huang, H.-C., Martinez, F., Mateu, J., and Montes, F. (2007), "Model comparison and selection for stationary space-time models," *Computational Statistics & Data*

Analysis, 51, 4577–4596.

Imbens, G. (1997), "One-step estimators for over-identified generalized method of moments models," *Review of Economic Studies*, 64, 359–383.

Jenish, N. and Prucha, I. R. (2009), "Central limit theorems and uniform laws of large numbers for arrays of random fields," *Journal of Econometrics*, 150, 86–98.

Joe, H. (1996), "Families of $m-$variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters," in *Distributions with Fixed Marginals and Related Topics*, eds. L. R uschendorf, B. Schweizer, and M. D. Taylor, pp. 120–141, Hayward, CA: Institute of Mathematical Statistics.

Joe, H. and Lee, Y. (2009), "On weighting of bivariate margins in pairwise likelihood," *Journal of Multivariate Analysis*, 100, 670–685.

Jorgensen, B. (1997), *The Theory of Dispersion Models*, Chapman & Hall.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008), "Covariance tapering for likelihood-based estimation in large spatial data sets," *Journal of the American Statistical Association*, 103, 1545–1555.

Kazianka, H. and Pilz, J. (2010), "Copula-based geostatistical modeling of continuous and discrete data including covariates," *Stochastic Environmental Research and Risk Assessment*, 24, 661–673.

Kotz, S. and Nadarajah, S. (2004), *Multivariate t Distributions and Their Applications*, Cambridge University Press.

Kuk, A. and Nott, D. (2000), "A pairwise likelihood approach to analyzing correlated binary data," *Statistics and Probability Letters*, 47, 329–335.

Kuk, A. Y. (2007), "A hybrid pairwise likelihood method," *Biometrika*, 94, 939–952.

Kunsch, H. (1989), "The jackknife and bootstrap for general stationary observations," *Annals of Statistics*, 17, 1217–1241.

Lahiri, S., Kaiser, M., Cressie, N., and Hsu, N. (1999), “Prediction of spatial cumulative distribution functions using subsampling,” *Journal of the American Statistical Association*, 94, 86–97.

Lawson, A. B. and Song, H. R. (2010), “Bayesian hierarchical modeling of the dynamics of spatio-temporal influenza season outbreaks,” *Spatial and Spatio-temporal Epidemiology*, 1, 187–195.

Lee, Y. D. and Lahiri, S. N. (2002), “Least squares variogram fitting by spatial subsampling,” *Journal of the Royal Statistical Society: Series B*, 64, 837–854, Part 4.

Lele, S. (1991), “Jackknifing linear estimating equations: asymptotic theory and applications in stochastic processes,” *Journal of the Royal Statistical Society: Series B*, 53, 253–267.

Lele, S. and Taper, M. L. (2002), “A composite likelihood approach to (co)variance components estimation,” *Journal of Statistical Planning and Inference*, 103, 117–135.

Li, B., Genton, M. G., and Sherman, M. (2007), “A nonparametric assessment of properties of space-time covariance functions,” *Journal of the American Statistical Association*, 102, 736–744.

Li, Y. and Lin, X. (2006), “Semiparametric Normal Transformation Models for Spatially Correlated Survival Data,” *Journal of the American Statistical Association*, 101, 591–603.

Liang, K.-Y. (1987), “Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models,” *Biometrics*, 43, 289–299.

Liang, K.-Y. and Zeger, S. (1986), “Longitudinal data analysis using generalized linear models,” *Biometrika*, 73, 13–22.

Lindsay, B. G. (1988), “Composite Likelihood Methods,” *Contemporary Mathematics*, 80, 221–239.

Mardia, K. and Marshall, R. (1984), "Maximum likelihood estimation of models for residual covariance in spatial regression," *Biometrika*, 71, 135–146.

Mardia, K. V., Hughes, G., and Taylor, C. C. (2007), "Efficiency of the pseudolikelihood for multivariate normal and von mises distributions," Tech. rep., Statistics Department, University of Leeds.

Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H. (2008), "A multivariate von Mises distribution with applications to bioinformatics," *Canadian Journal of Statistics*, 36, 99–109.

Mardia, K. V., Kent, J., Hughes, G., Taylor, C. C., and Singh, H. (2009), "Maximum likelihood estimation using composite likelihoods for closed exponential families," *Biometrika*, 96, 975–982.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, CRC Press, 2 edn.

Min, A. and Czado, C. (2010), "Bayesian inference for multivariate copulas using pair-copula constructions," *Journal of Financial Econometrics*, 8, 511–546.

Molenberghs, G. and Verbeke, G. (2005), *Models for Discrete Longitudinal Data*, Springer,New York.

Mujahid, M. S., Diez Roux, A. V., Morenoff, J. D., and Raghunathan, T. (2007), "Assessing the measurement properties of neighborhood scales: from psychometrics to ecometrics," *American Journal of Epidemiology*, 165, 858–867.

Nott, D. and Rydén, T. (1999), "Pairwise likelihood methods for inference in image models," *Biometrika*, 86, 661–676.

Owen, A. (1988), "Empirical likelihood ratio confidence intervals for a single functional," *Biometrika*, 75, 234–249.

Paciorek, C. J., Yanosky, J. D., and Puett, R. C. (2009), "Practical large-scale spatio-temporal modeling of particulate matter concentrations," *The Annals of Applied*

Statistics, 3, 370–397.

Politis, D. and Romano, J. (1994), “Large sample confidence regions based on subsamples under minimal assumptions,” *The Annals of Statistics*, 22, 2031–2050.

Porcu, E., Mateu, J., and Bevilacqua, M. (2007), “Covariance functions that are stationary or nonstationary in space and stationary in time,” *Statistica Neerlandica*, 61, 358–382.

Qin, J. and Lawless, J. (1994), “Empirical likelihood and general estimating equations,” *Annals of Statistics*, 22, 300–325.

Qu, A. and Song, P. X.-K. (2004), “Assessing robustness of generalized estimating equations and quadratic inference functions,” *Biometrika*, 91, 447–459.

Qu, A., Lindsay, B., and Li, B. (2000), “Improving generalized estimating equations using quadratic inference functions,” *Biometrika*, 87, 823–836.

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Reid, N. and Cox, D. (2004), “A note on pseudolikelihood constructed from marginal densities,” *Biometrika*, 91, 729–737.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2007), “High-resolution space-time ozone modeling for assessing trends,” *Journal of the American Statistical Association*, 102, 1221–1234.

Sang, H. and Huang, J. (2011), “A full-scale approximation of covariance functions for large spatial data sets,” *Journal of the Royal Statistical Society: Series B*, in press.

Sener, I., Pendyala, R., and Bhat, C. (2011), “Accommodating spatial correlation across choice alternatives in discrete choice models: an application to modeling residential location choice behavior,” *Journal of Transport Geography*, 19, 294–303.

Sherman, M. (1996), "Variance estimation for statistics computed from spatial lattice data," *Journal of the Royal Statistical Society: Series B*, 58, 509–523.

Smith, M., Min, A., Almeida, C., and Czado, C. (2010), "Modeling longitudinal data using a pair-copula decomposition of serial dependence," *Journal of the American Statistical Association*, 105, 1467–1479.

Smith, R. L. and Kolenikov, S. (2003), "Spatiotemporal modeling of $PM_{2.5}$ data with missing values," *Journal of Geophysical Research*, 108, 4–27.

Song, P. X.-K. (2000), "Multivariate dispersion models generated from Gaussian copula," *Scandinavian Journal of Statistics*, 27, 305–320.

Song, P. X.-K. (2007), *Correlated Data Analysis*, Springer.

Stein, M. L. (1995), "Fixed domain asymptotics for spatial periodograms," *Journal of the American Statistical Association*, 90, 1277–1288.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory of Kriging*, Springer-Verlag.

Stein, M. L. (2005), "Space-time covariance functions," *Journal of the American Statistical Association*, 100, 310–321.

Stein, M. L. (2008), "A modeling approach for large spatial data sets," *Journal of the Korean Statistical Society*, 37.

Stein, M. L., Chi, Z., and Welty, L. (2004), "Approximating likelihoods for large spatial data sets," *Journal of the Royal Statistical Society: Series B*, 66, 275–296.

Thomson, M., Connor, S., D'Alessandro, U., Rowlingson, B., Diggle, P., Creswell, M., and Greenwood, B. (1999), "Predicting malaria infection in Gambian children from satellite data and bed net use surveys: the importance of spatial correlation in the interpretation of results," *The American Journal of Tropical Medicine and Hygiene*, 61.

Varin, C. (2008), "On composite marginal likelihoods," *Advances in Statistical Analysis*, 92, 1–28.

Varin, C. and Vidoni, P. (2005), "A note on composite likelihood inference and model selection," *Biometrika*, 92, 519–528.

Varin, C. and Vidoni, P. (2006), "Pairwise likelihood inference for ordinal categorical time series," *Computational Statistics and Data Analysis*, 51, 2365–2373.

Varin, C., Høst, G., and Skare, Ø. (2005), "Pairwise likelihood inference in spatial generalized linear mixed models," *Computational Statistics & Data Analysis*, 49, 1173–1191.

Varin, C., Reid, N., and Firth, D. (2011), "An Overview of Composite Likelihood Methods," *Statistica Sinica*, pp. 5–42.

Vecchia, A. (1988), "Estimation and model identification for continuous spatial processes," *Journal of the Royal Statistical Society: Series B*, 50, 297–312.

Whittle, P. (1954), "On stationary processes in the plane," *Biometrika*, 41, 434–449.

Windmeijer, F. (2005), "A finite sample correction for the variance of linear efficient two-step GMM estimators," *Journal of Econometrics*, 126, 25–51.

Winkler, G. (1995), *Image analysis, Random fields and Dynamic monte carlo methods: a mathematical introduction*, Springer-Verlag, New York.

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261.

Zhao, L. P. and Prentice, R. L. (1990), "Correlated binary regression using a quadratic exponential model," *Biometrika*, 77, 642–648.

Zhao, Y. and Joe, H. (2005), "Composite likelihood estimation in multivariate analysis," *The Canadian Journal of Statistics*, 33, 335–356.

Zhu, J. and Morgan, G. (2004), “Comparison of spatial variables over subregions using a block bootstrap,” *Journal of Agricultural, Biological, and Environmental Statistics*, 9, 91–104.

Zhu, L., Carlin, B. P., and Gelfand, A. E. (2003), “Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta,” *Environmetrics*, 14, 537–557.

Zi, J. (2009), “On some aspects of composite likelihood,” Ph.D. thesis, University of Toronto.

Zimmerman, D. L. (1989), “Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models,” *Journal of Statistical Computation and Simulation*, 32, 1–15.