STATISTICAL INFERENCE FOR LARGE SPATIAL DATA

A Dissertation

by

FURONG LI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Huiyan Sang
Committee Members,	Michael Longnecker
	Bani Mallick
	Ramalingam Saravanan
Head of Department,	Valen Johnson

May 2017

Major Subject: Statistics

ABSTRACT

The availability of large spatial and spatial-temporal data geocoded at accurate locations has fueled increasing interest in spatial modeling and analysis. In this dissertation, we present one study concerning the inference on properties of a single spatial process, and then turn to multiple processes and provide two modeling approaches exploring the spatially varying relationship between covariates and the response variable of interest.

In the first study, we investigate the inference tool based on quasi-likelihood, composite likelihood (CL) method and propose a new weighting scheme to construct a CL for the inference of spatial Gaussian process models. This weight function approximates the optimal weight derived from the theory of estimating equations. It combines block-diagonal approximation and tapering strategy to facilitate computations. Gains in statistical and computational efficiency over existing CL methods are illustrated through simulation studies.

The second investigation is the development of a new spatial modeling framework to capture the spatial structure, especially clustered structure in the relationship between response variable and explanatory variables. The proposed method, called Spatially Clustered Coefficient(SCC) regression, results in estimators of varying coefficients, which conveys important information about the changing pattern of the relationship. The SCC method works very effectively in estimation for data either with clustered coefficients or smoothly-varying coefficients, based on our simulation results. Thus, it allows the researchers to explore the spatial structure in the regression coefficient without any priori information. We also derive some oracle inequalities, which provides non-asymptotic error bounds on estimators and predictors. An application of the SCC method to temperature and salinity data in the Atlantic basin is provided for illustration.

Motivated by the studies in Geoscience that the influence of turbulent heat flux on sea surface temperature (SST) varies at different spatial scales, we develop a statistical model to quantify the continuous dependence of SST-turbulent heat flux relationship (T-Q relationship) on spatial scales. In particular, we propose a penalized regression model in the spectral domain to estimate the changing relationship with spatial scales. While application to T-Q relationship is the main motivation for this work, it should be emphasized that the penalized spectral regression framework is general and thus is applicable to other phenomena of interest as well.

ACKNOWLEDGMENTS

It is a great pleasure to acknowledge all those who help to make this dissertation possible.

First, I would like to express my sincere gratitude to my Ph.D advisor Huiyan Sang, for her enthusiasm and great efforts. She provided encouragement and excellent research guidance through my Ph.D study. She is also very helpful beyond research, providing me a lot of suggestions about life.

I am very grateful to Dr. Michael Longnecker. He always provided me tremendous help and guidance since the first day I started my statistics study here. He is also one member of my committee, who helped review this dissertation carefully. I would like give thanks to the other members of my committees, Dr. Bani Mallick and Dr. Ramalingam Saravanan.

The Department of Statistics in Texas A&M University has been very supportive. I count myself fortunate to have been student for 5 years in such a friendly environment. Special thanks to many colleagues within the department for their accompanying in the journey of the qualifier, the prelim and the defense.

I wish to thank my family for their supports. A special gratitude is due to my husband, for his endless encouragement and support.

To them I dedicate this dissertation.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Dr. Huiyan Sang, Dr. Michael Longnecker and Dr. Bani Mallick of the Department of Statistics and Dr. Ramalingam Saravanan of the Department of Atmospheric Sciences.

The data analyzed for Section 4 was obtained from ERA-Interim reanalysis dataset, which is available through the online access http://www.ecmwf.int/en/research/climate-reanalysis/era-interim. The analyses depicted in Section 4 were conducted in part by Zhao Jing of the Department of Oceanography.

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by a fellowship from Texas A&M University and dissertation research fellowships from National Science Foundation DMS-1343155, DMS-1622433 and National Institutes of Health grant R01-CA19776101.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Nowadays the popular usage of geographical information systems (GIS) and global positioning systems (GPS) have led to the increased collection of research data geocoded at accurate locations, in the field of geoscience, econometrics and biological science. This has fueled increasing interest in statistical modeling and analysis with the spatial locations of measurements being taken into account. Spatial statistics is now an important field within statistics.

The main feature of spatial data is the dependence of observations since data observed on proximal locations tend to have similar values, possibly resulting from homogeneous physical dynamics or environmental conditions. Ignoring this spatial dependence may result in incorrect estimation of model parameters and inaccurate predictions.

There are three major objectives for spatial analysis. One of the research goals is to make inferences on the properties of the process. That is, to estimate the parameters describing the spatial dependence structure. Another primary goal is to explain the variability in the process of interest using a set of explanatory variables. One useful framework to achieve this goal is spatial regression. Spatial regression coefficients reveal both the effect of covariates, and provide important information on the relationship between two processes. The third goal of spatial analysis is that of prediction. That is, predict an outcome at an unobserved location, given the observed values. This dissertation focuses on the first two research goals.

In investigating the properties of a process, conventional likelihood-based model inference methods, such as maximum likelihood estimation (MLE) and Bayesian inference, are computationally expensive for many spatial models with large data sets. As an alternative inference tool, composite likelihood (CL) methods have gained considerable attention in recent years due to their simplicity and sound asymptotic properties. However, CL estimators often result in substantial loss in statistical efficiency with respect to MLE. To improve statistical efficiency or to reduce the computational burden, recent approaches consider the choice of weights in constructing composite likelihood in the context of spatial process. We follow this path, in the first study, to seek an adaptive weight function for composite likelihood with a good balance between computational complexity and estimation efficiency.

In dealing with multiple spatial processes, spatial regression models such as Gaussian process regression or spatial generalized linear regression models have been widely adopted to address this problem, in which spatial dependence is taken into account by adding a spatial random effect to the (generalized) linear regression model. Regression coefficients in such models are often assumed to be a constant. However, in many problems especially when data are collected across a large region, it is unlikely that a constant regression coefficient can adequately capture the spatially dynamic relationship between response variables and covariates. One example is to have clustered pattern of regression coefficients that abruptly change across the boundary of adjacent clusters but stay relatively homogeneous within clusters. Indeed, it is of great interest to many practitioners to identify such clusters that allow them to explain varying associations between the response of interest and covariates. There is no existing method designed to address the clustered coefficient regression. We develop a spatial modeling approach with the ability to capture the spatial structures in the effect of the explanatory variables.

The relationship between response variables and covariates may not only vary in the physical space but also in the spectral space. The latter is quite common in the applications of geophysics as the dynamics controlling the relationship typically changes with the spatial scales. While a knowledge of the scale-dependent relationship is essential to understand the nature of geophysical system, so far no modelling approaches have been proposed to address these types of problems. In the third study, we extend the varying coefficient regression model developed in the second study to the spectral domain, constructing a penalized spectral regression model to estimate the scale-dependent relationship between response variables and covariates.

The rest of this dissertation is organized as follows. In Section 2, we present the study of weighted composite likelihood, focusing on effective estimation of covariance parameters in spatial Gaussian process. Section 3 details the Spatial Clustered Coefficient (SCC) model and its theoretical properties. An extension of the SCC model in the spectral domain is provided in Section 4 with its application to the relationship between sea surface temperature and turbulent heat flux at the air-sea interface. Section 5 summarizes the studies in this dissertation.

2. ON APPROXIMATING OPTIMAL WEIGHTED COMPOSITE LIKELIHOOD METHOD FOR SPATIAL MODELS

2.1 Introduction

There has been much interest in recent years in a form of likelihood type estimation called composite likelihood (CL). It is a weighted product of a collection of component likelihoods such as low dimensional conditional or marginal densities. Because each component in CL is a valid likelihood object, the corresponding estimating equation obtained from the score function of CL is unbiased under standard regularity conditions. Therefore, the CL inference is known to have well established properties of a likelihood from a misspecified model. Compared to the maximum likelihood (ML) approach, the CL method does not require evaluations of full likelihood functions but only products of low-dimensional marginal or conditional likelihoods, leading to a considerable reduction of computational burden, although a loss of statistical efficiency is generally expected with respect to the ML method.

CL methods have been used in many contexts when it is difficult or computational expensive to evaluate or specify full likelihoods. In particular, various types of CL functions have been introduced in spatial statistics to facilitate computations. For spatial GP models, it is known that the full likelihood function of a GP model involves inversion of an $n \times n$ covariance matrix for a data set of size n, requiring $O(n^3)$ operation and $O(n^2)$ memory. This can be computationally infeasible for large datasets which are becoming increasingly common in geosciences. [1] compared three CL functions for spatial GP models based on the paired marginal distributions, paired conditional distributions and paired differences. Recently, in [2], the authors proposed to use a composite likelihood function defined as a product of the joint densities of pairwise spatial blocks. For spatial generalized linear mixed models, the authors in [3] proposed a composite likelihood approach based on marginal densities of pairwise differences of responses; and the authors in [4] proposed a pairwise composite likelihood approach based on bivariate marginal densities. [5] used the composite likelihoods for the inference of a Gaussian max-stable process for spatial extreme values, in which the closed-form expressions of the corresponding joint likelihoods are intractable.

As outlined in [6], for a given estimation problem, the choice of a suitable CL function should be driven by statistical and computational considerations. However, it is noticeable that many existing methods of CLs are constructed with equally weighted pairs due to its simplicity. To improve statistical efficiency or to further reduce the computational burden associated with large data sets that have enormous number of pairs, several investigations have considered the choice of weights when constructing CL in the context of a spatial process. One popular strategy is to use binary weights to exclude those pairs whose distances are beyond certain taper range [4, 7]. [8] and [9] consider selecting a taper range by maximizing certain criteria derived from the Godambe information matrix of a CL estimator. The CL estimators based on binary weights have improved statistical efficiency over equally weighted CL methods. However, these methods ignore dependence among selected pairs and hence can still lead to considerable loss of statistical efficiency. Thus far, very limited work has been done on designing non-binary weights. [8] investigated weighted composite score for a scalar parameter and constructed a weight by minimizing an upper bound of the asymptotic variance of the estimates. They find that the proposed weighted CL method performs better than both the binary weighted method and the equally weighted CL method for Gaussian random fields. [10] proposed a joint composite estimating function (JCEF) approach through a weight matrix to spatio-temporally clustered data.

Our main contribution in this paper is to construct a new efficient weighted composite likelihood (WCL) method for spatial Gaussian processes. The proposed weight is motivated from the optimal weight derived from the theory of estimating equations. It is known that the computational burden in constructing optimal estimating equations is formidable as it requires the inversion of a large covariance matrix of scores. To circumvent the difficulty in computing optimal weights, we exploit spatial dependence structures among pairs and develop a sparse matrix approximation method based upon block-diagonal structure and tapering. This leads to a weight function with a good balance between computation complexity and estimation efficiency.

Both the cases with a scalar parameter and multiple covariance parameters are investigated. We develop a weighted profile composite likelihood method that iteratively estimates each model parameter using WCL. Our method allows the use of different weights for each individual model parameter to reflect different correlation structures among pairs of score functions.

This section is organized as follows. In Section 2.2, we review some basics for composite likelihoods and spatial Gaussian process relevant to this study, and then introduce our methods for efficient covariance estimation. Section 2.3 illustrates the performance of our method through a number of simulation studies. An application to real data is presented in Section 2.4, using the yearly total precipitation anomalies dataset [11]. Conclusions are summarized in Section 2.5 followed by discussion.

2.2 Methodology

2.2.1 Composite likelihood

Consider a parametric statistical model with probability density function $\{f(\mathbf{z};\boldsymbol{\theta}), z \in \mathcal{Z} \subseteq \mathbb{R}^n\}$, where $\boldsymbol{\theta}$ is a p-dimensional parameter vector to be estimated. Denoting by $\{\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_K\}$ a set of marginal or conditional events, composite likelihood is a weighted product of the likelihood corresponding to each single event [6]

$$\mathbf{C}\mathcal{L}(\boldsymbol{\theta};\mathbf{z}) = \prod_{k=1}^{K} f(\mathbf{z} \in \mathcal{A}_k;\boldsymbol{\theta})^{\omega_k}, \tag{2.1}$$

where $f(\mathbf{z} \in \mathcal{A}_k;\boldsymbol{\theta})$ is the likelihood of event $\mathcal{A}_k$ and $\{\omega_k, k = 1, \dots, K\}$ is a set of non-negative weights to be chosen. The associated weighted composite log-likelihood is

$$c\ell(\boldsymbol{\theta};\mathbf{z}) = \sum_{k=1}^{K} \omega_k \ell_k(\boldsymbol{\theta};\mathbf{z}), \tag{2.2}$$

where $\ell_k(\boldsymbol{\theta};\mathbf{z}) = \log f(\mathbf{z} \in \mathcal{A}_k;\boldsymbol{\theta})$.

CL in (2.2) is a universal expression of the weighted composite log-likelihood allowing for combinations of marginal and conditional densities [12]. For example, a special form of CL is compounded based on pairwise differences between observations

$$c\ell(\boldsymbol{\theta};\mathbf{z}) = \sum_{t=1}^{N} \sum_{i \neq j} w_{ij} \ell_{ij}(\boldsymbol{\theta};\mathbf{z}_i^{(t)} - \mathbf{z}_j^{(t)}), \tag{2.3}$$

where $\mathbf{z}_i^{(t)}$ is the sample of the t-th replicate for $\mathbf{z}_i$.

Consider a spatial random field $Z(\mathbf{s})$ that is modeled as a Gaussian process with mean $\mu(\mathbf{s})$ and a covariance function $\mathcal{C}(\mathbf{s}, \mathbf{s}';\boldsymbol{\theta})$. For example, an exponential covariance function takes the form $\mathcal{C}(\mathbf{s}, \mathbf{s}';\boldsymbol{\theta}) = \sigma^2 \exp(|\mathbf{s} - \mathbf{s}'|/\phi)$, where σ^2 is the variance parameter, ϕ is the spatial dependence range parameter. It is well known that the data likelihood with n observed locations involves the inversion of an $n \times n$ covariance matrix. The computational cost can be very intensive or even prohibitive when n is large.

CL offers an alternative inference approach that only requires low-dimensional likelihood calculation and hence has a clear computational advantage over full likelihood. Indeed, the computational cost for considering all possible pairs is of order $O(n^2)$. In this paper, we focus on constructing a WCL based on pairwise differences for the inference of covariance parameters while assuming $\mu(\mathbf{s})$ is a constant in the rest of the paper. We remark that it is relatively straightforward to extend the proposed method by including a mean model for $\mu(\mathbf{s})$ following similar strategies as in restricted maximum likelihood (REML).

Let $U_{ij,t} = Z(\mathbf{s}_i, t) - Z(\mathbf{s}_j, t), i \neq j, t = 1, \cdots, N$, the differences between any two observations of the t-th replicate. Then, we have $U_{ij,t} \sim \mathcal{N}(0, 2\gamma_{ij}(\boldsymbol{\theta}))$, where $\gamma_{ij}(\boldsymbol{\theta}) =$ var$[Z(\mathbf{s}_i) - Z(\mathbf{s}_j)]$, also known as the variogram in spatial statistics. The composite likelihood of the pairwise difference in (2.3) can be expressed as $c\ell(\boldsymbol{\theta}) = \sum_{t=1}^{N} \sum_{i \neq j} w_{ij} \ell_{ij,t}(\boldsymbol{\theta})$, where $\ell_{ij,t}(\boldsymbol{\theta}) = -\{\log\gamma_{ij}(\boldsymbol{\theta})/2 + [U_{ij,t}]^2/(4\gamma_{ij}(\boldsymbol{\theta}))\}$.

Composite likelihood can be justified within the framework of the theory of estimating functions. Let $\widehat{\boldsymbol{\theta}}_{CL}$ be the maximum composite likelihood estimator of $\boldsymbol{\theta}$. Clearly, $\widehat{\boldsymbol{\theta}}_{CL}$ is also the solution of the following composite score equations,

$$\mathbf{s}(\boldsymbol{\theta}; \mathbf{z}) = \nabla c\ell(\boldsymbol{\theta}; \mathbf{z}) = \sum_{i \neq j} w_{ij} \mathbf{s}_{ij}(\boldsymbol{\theta}; \mathbf{z}) = 0, \tag{2.4}$$

where ∇ denotes the gradient obtained by differentiation with respect to $\boldsymbol{\theta}$. Here $\mathbf{s}_{ij}(\boldsymbol{\theta}; \mathbf{z}) = (\sum_{t=1}^{N} \nabla_{\theta_1} \ell_{ij,t}(\boldsymbol{\theta}; \mathbf{z}), \cdots, \sum_{t=1}^{N} \nabla_{\theta_p} \ell_{ij,t}(\boldsymbol{\theta}; \mathbf{z}))$, representing score contributions from each pair (i, j).

Since $\mathbf{s}(\boldsymbol{\theta}; \mathbf{z})$ is a linear combination of the scores associated with each of the likelihood terms, it is indeed an unbiased estimating equation satisfying $E\{\mathbf{s}(\boldsymbol{\theta}, \mathbf{z})\} = 0$ under standard regularity conditions [6]. Therefore, under the theory of the unbiased estimating function, the maximized composite likelihood estimator $\hat{\boldsymbol{\theta}}_{CL}$ is a consistent and unbiased parameter estimator [4, 6],

$$N^{1/2}(\hat{\boldsymbol{\theta}}_{CL} - \boldsymbol{\theta}) \rightarrow N_q\{0, G^{-1}(\boldsymbol{\theta})\} \tag{2.5}$$

in distribution as $N \to +\infty$, where $G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}H(\boldsymbol{\theta})$ known as Godambe information [13] or sandwich information, $H(\boldsymbol{\theta}) = E\{-\nabla^2 c\ell(\boldsymbol{\theta})\}$ and $J(\boldsymbol{\theta}) = \text{var}\{\nabla c\ell(\boldsymbol{\theta})\}$.

Despite its sound asymptotic properties, the estimation method using CL typically results in loss of statistical efficiency compared with the maximum likelihood estimator counterpart. Indeed, CL can be viewed as a misspecified model, and hence may not attain the Cramér-Rao lower bound [12]. Nevertheless, efficiency gain might be achieved by carefully designing ways to construct composite likelihood while keeping low computational cost [10, 14]. Below, we seek to construct a WCL to provide a good compromise between computation cost and estimation efficiency.

2.2.2 Approximate optimal weighted composite likelihood

In this paper, we present a weighting strategy with the goal to improve the efficiency of CL by exploiting the theory of optimal estimating equations [15]. Stack all the individual score vector $\mathbf{s}_{ij}(\boldsymbol{\theta})$ of size p into a column vector $\mathbf{S}(\boldsymbol{\theta})$ of size $pn(n-1)/2$. Now let $\mathbf{W}(\boldsymbol{\theta})$, a $pn(n-1)/2 \times p$ matrix, be the weighting function of $\boldsymbol{\theta}$. Then $\mathbf{Q}(\boldsymbol{\theta}) = \mathbf{W}(\boldsymbol{\theta})^{\mathrm{T}}\mathbf{S}(\boldsymbol{\theta})$ defines a class of valid unbiased estimating functions. For example, the equally weighted CL corresponds to the case where $\mathbf{W}(\boldsymbol{\theta})$ is a binary matrix $\mathbf{1}_{n(n-1)/2} \otimes \mathbf{I}_p$. Let $\boldsymbol{\theta}_w$ be the root of the estimating function $\mathbf{Q}(\boldsymbol{\theta})$ with weight matrix $\mathbf{W}$, the approximate asymptotic covariance matrix of $\boldsymbol{\theta}_w$ is given by

$$E\{\nabla\mathbf{Q}(\boldsymbol{\theta})\}]^{-1}E\{\mathbf{Q}(\theta)\mathbf{Q}(\theta)^{T}\}]^{-1}E\{\nabla\mathbf{Q}(\boldsymbol{\theta})\}]^{-T} \tag{2.6}$$

In the class of estimating functions $\mathbf{Q}(\boldsymbol{\theta})$, the approximate covariance matrix in (2.6) is minimized with respect to the partial ordering of nonnegative definite symmetric matrices (see, [15]) when

$$\mathbf{W} = \mathbf{W}_{Opt} = \text{Cov}\{\mathbf{S}(\theta)\}^{-1}E^{T}\{\nabla\mathbf{S}(\theta)\} \tag{2.7}$$

Although the weight matrix with the above form combines score vectors in an optimal way and leads to efficient estimators, it is rarely used in practice. Indeed, Cov($\mathbf{S}(\theta)$) is a $pn(n-1)/2 \times pn(n-1)/2$ matrix, whose inversion requires $O(n^6)$ of computational complexity, making it computationally prohibitive for large spatial data sets. Below, we seek strategies to approximate $\mathbf{W}_{Opt}$ to circumvent computational difficulties.

We first consider the case in which parameter θ is a scalar. In this case, we stack $\mathbf{s}_{ij}(\boldsymbol{\theta}; \mathbf{z}), j \neq i$ into a vector $\mathbf{S}(\theta)$ of size $n(n-1)/2$. Following (2.7), the optimal weight $\mathbf{W}_{Opt} = -E^T\{\nabla\mathbf{S}(\theta)\}\text{Cov}(\mathbf{S}(\theta))^{-1}$. Under mild regularity conditions, $-E\{\nabla\mathbf{S}_{ij}(\theta)\} = \text{Var}(\mathbf{S}_{ij}(\theta))$, i.e., the diagonal entries of Cov($\mathbf{S}(\theta)$). Rewrite $\text{Cov}\{\mathbf{S}(\theta)\} = D\text{Corr}\{\mathbf{S}(\theta)\}D$, where D is a diagonal matrix proportional to the square roots of the diagonal entries of Cov($\mathbf{S}(\theta)$). The optimal weight $\mathbf{W}_{Opt}$ can be expressed as

$$\mathbf{W}_{Opt} = D^{-1}\text{Corr}(\mathbf{S}(\theta))^{-1}D\mathbf{1}. \tag{2.8}$$

Clearly, equally weighted CL essentially corresponds to the case where correlations among score elements are treated as zeros. However, such assumptions are unrealistic for spatial models in which pairs of scores constructed on spatial differences often show non-negligible dependence.

We seek methods to approximate the optimal weight function that takes into account correlations among spatial pairs while keeping the computation at a low cost. To motivate such an approximation, we investigate the pattern of the weight function $\mathbf{W}_{Opt}(\theta)$ and the covariance of score functions below. First note that $-E^T\{\nabla\mathbf{S}(\theta)\}$ is a vector of size $n(n-1)/2$ with each element to be I_{ij}, the marginal Fisher information of the likelihood for a spatial pair (i, j). The marginal information contribution from each spatial pair is expected to vary as the distance between i and j. In fact, for a spatial Gaussian process model with variogram γ_{ij}, $I_{ij} = N\frac{[\gamma_{ij}^{(1)}(\theta)]^2}{2\gamma_{ij}^2(\theta)}$. Using simple algebra, we can prove that the Godambe information gain by adding a pair (i,j) from a weighted CL is bounded by the marginal information I_{ij}. Therefore, this motivates us to taper pairs if their marginal information is below certain threshold before constructing WCL. For example, when γ_{ij} is an exponential variogram function and the goal is to estimate the range parameter while fixing σ^2, it is easy to show that the marginal information is a monotone decaying function of distance, which indicates that pairs with distances beyond certain taper range, denoted as τ, can be excluded since their contributed information is minimal.

Let $\mathbf{S}(\theta)_{taper}$ denote the score vector stacked from the tapered $s_{ij}(\theta)$. We next examine the pattern of the correlation matrix of the score function $\mathbf{C}_{score,taper} = \mathrm{Corr}\{\mathbf{S}(\theta)_{taper}\}$. Apparently, the dimension of this correlation matrix is greatly reduced thanks to tapering. To further reduce computation, a natural idea is to seek strategies to approximate $\mathrm{Corr}\{\mathbf{S}(\theta)\}$ by only keeping elements with large correlations. Note that for the spatial problem we consider here,

$$\mathrm{Corr}\{\mathbf{S}_{ij}(\theta), \mathbf{S}_{\ell k}(\theta)\} = \frac{\{\gamma_{i\ell}(\theta) - \gamma_{j\ell}(\theta) + \gamma_{jk}(\theta) - \gamma_{ik}(\theta)\}^2}{4\gamma_{ij}(\theta)\gamma_{\ell k}(\theta)}. \tag{2.9}$$

It can be proved that for a given pair (i,j),

$$\mathrm{Corr}\{\mathbf{S}_{ij}(\theta), \mathbf{S}_{\ell k}(\theta)\} \le \max\{\mathrm{Corr}(\mathbf{S}_{ij}(\theta), \mathbf{S}_{\ell_1 \ell_2}(\theta))\} \tag{2.10}$$

holds for any $\ell_1 \in \{i \neq j\}$ and $\ell_2 \notin \{\ell \neq k\}$. This inequality suggests that two pairs achieve the largest correlation when two vertices from each pair coincides with each other. Motivated by this finding, for a score function corresponding to a given pair (i,j), we propose to keep its correlation with $\{(i,\ell)\}$, for all $\{\ell : d_{i\ell} < \tau\}$ and set correlations for pairs without shared vertex to be 0. It clearly has an advantage over equal-weight CL which completely ignores correlations among pairs. By using a more accurate approximation of the optimal weights, $\mathbf{W}_{BT}(\phi)$ is expected to achieve greater statistical efficiency compared with other WCL methods.

We acknowledge that this approximation ignores correlations among pairs without shared vertices. We explain below why it is necessary to do so for the sake of computation efficiency. Under a proper ordering of pairs, the approximation method described above results in a block diagonal matrix approximation to the correlation matrix $\mathbf{C}_{score,taper}$, denoted as $\mathbf{C}_{score,BT}$. Let M denote the number of blocks, which equals the number of unique vertices from all remaining pairs after tapering. For a given order of these vertices from 1 to M, the m-th block is the correlation matrix of a set of two pairs $\{(m,j),(m,\ell)\}$, for $\{j \neq \ell > m, d_{mj} < \tau, d_{m\ell} < \tau\}$. Compared to the original full correlation matrix of the score vector for all pairs, the computational cost associated with this approximated correlation matrix is greatly reduced for two reasons: first the dimension of the matrix is substantially reduced from the total number of pairs to the number of close pairs only; and the approximated correlation of the close pairs has block diagonal structures, whose computation can be handled efficiently and in parallel.

To illustrate the above idea, below we plot the pattern of the covariance matrix of the score function $\mathbf{C}_{score,taper}$ through a simulation. We generate 100 independent replicates of realizations from a spatial Gaussian process at 100 randomly selected locations from $[0,100] \times [0,100]$. An exponential covariance function is used with the range parameter $\phi = 30$ and the variance $\sigma^2 = 1$ (no nugget effect). We first consider estimating ϕ assuming σ^2 is known. Figure 2.1(a) shows the averaged score correlation matrix $\mathbf{C}_{score,taper}$ for the re-ordered remaining pairs after tapering. A notable feature for $\mathbf{C}_{score,taper}$ is that the value in its block diagonals are generally significantly larger than the non-block diagonals, which justifies our approximation strategy by only considering block diagonal correlations that capture dependence for two pairs that share a vertex.

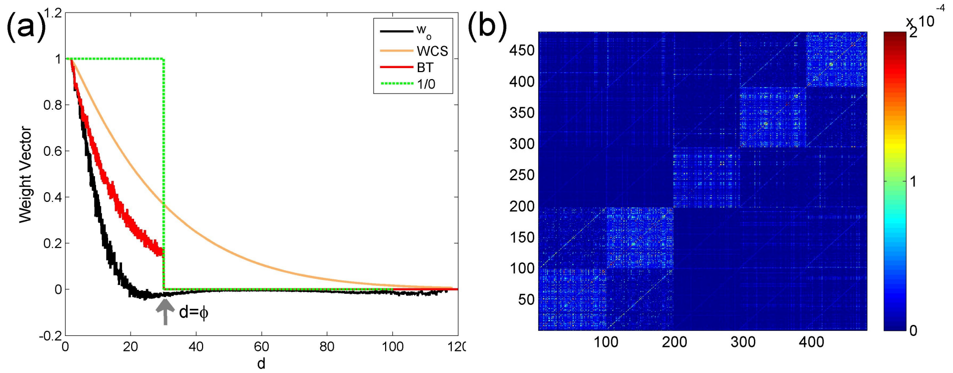

Figure 2.1: Weights for composite likelihood. (a) Various weights as a function of distance of pairs. (b) The covariance matrix of scores Cov($\mathbf{S}(\boldsymbol{\theta})$) based on exponential covariance with $\sigma^2 = 1$ and $\phi = 30$.

With the use of the approximated score correlation matrix $\mathbf{C}_{score,BT}$, we propose a new weighting function termed as block-tapering (BT) weight

$$\mathbf{W}_{BT}(\theta) = -E^T\{\nabla\mathbf{S}(\theta)_{taper}\}\mathbf{C}_{score,BT}(\theta)^{-1}, \tag{2.11}$$

for all pairs with $d_{ij} < \tau$, and 0 otherwise.

Using the same simulated dataset as above, we evaluate the proposed weight function $\mathbf{W}_{BT}$ (referred to as the BT-WCL method) and compare it with various other weight functions, including a binary 1/0 weight by tapering distant pairs (denoted as $\mathbf{W}_{1/0}(\phi)$, referred to as 1/0-WCL), an adaptive weight $\mathbf{W}_{WCS}(\theta) = \text{diag}\{-E[\ell_{ij}^{(2)}(\theta)]\}$ (referred to as the WCS method) proposed by [8], and the optimal weight $\mathbf{W}_{Opt}(\phi)$ as a benchmark. Clearly $\mathbf{W}_{Opt}(\phi)$ appears to have a strong decreasing trend as distance d as shown in Figure 2.1(b), which is consistent with the findings in previous studies that distant pairs are nearly uncorrelated and hence contribute little information in terms of estimating range parameter. For the examples considered here with the exponential covariance, indeed the optimal weight $\mathbf{W}_{Opt}(\phi)$ decreases from 1 ($d = 0$) to 0.05 at $d \geq \phi$, suggesting the use of ϕ as a threshold to guarantee little loss of efficiency. It is also noticeable that the weight $\mathbf{W}_{BT}(\phi)$ is in greater agreement with the curve $\mathbf{W}_{Opt}(\phi)$ than $\mathbf{W}_{WCS}(\phi)$ or $\mathbf{W}_{1/0}(\phi)$. This finding is not surprising considering that $\mathbf{W}_{1/0}(\phi)$ is essentially equivalent to the weight by approximating Corr$(\mathbf{S}(\phi))_{taper}$ as an identity matrix, and $\mathbf{W}_{WCS}(\phi)$ is essentially equivalent to the weight by approximating Cov$(\mathbf{S}(\phi))$ as an identify matrix. For $d \geq \phi$, $\mathbf{W}_{Opt}(\phi)$,$\mathbf{W}_{BT}(\phi)$ and $\mathbf{W}_{1/0}(\phi)$ are nearly zero while $\mathbf{W}_{WCS}(\phi)$ is still significantly greater than zero. For $d \leq \phi$, $\mathbf{W}_{Opt}(\phi)$,$\mathbf{W}_{BT}(\phi)$ and $\mathbf{W}_{WCS}(\phi)$ all decrease fairly smoothly with spatial lag d. But the decay rate of $\mathbf{W}_{BT}(\phi)$ is much closer to that of the $\mathbf{W}_{Opt}(\phi)$ than that of $\mathbf{W}_{WCS}(\phi)$. These findings imply that the our proposed method uses a more accurate approximation to the optimal weight and hence is expected to improve statistical efficiency compared to the 1/0 weight and WCS weight methods. We will further demonstrate the utility our method in Section 2.3 through numerical simulations.

We now consider the case with multiple parameters. The optimal weight in this case involves the inversion of Cov$(\mathbf{S}(\theta))$, a $pn(n-1)/2 \times 2n(n-1)/2$ matrix. In view of the computational expense of the joint optimal weight in (2.8), we propose to iteratively estimate model parameters following a similar spirit as in the profile likelihood method. That is, given current values of $(\widehat{\sigma}^2, \hat{\phi})$, we calculate a weighting function for each individual parameter and then estimate the parameter by maximizing the weighted profile composite likelihood. For example, assume both the range parameter ϕ and the variance σ^2 are now unknown, we estimate model parameters according to the following procedures:

(1) Start from some initial values of $(\widehat{\sigma}^2, \hat{\phi})$. One way is to obtain preliminary estimates of $(\widehat{\sigma}^2, \hat{\phi})$ using fast estimation methods such as the tapered equally-weighted CL.

(2) Given $\widehat{\sigma}^2$, update $\hat{\phi}$ using the BT-WCL method, and

(3) Given $\hat{\phi}$, update $\widehat{\sigma}^2$ using the CL estimation with equal weights.

The entire procedure repeats steps (2) and (3) until convergence. Of course, there is no guarantee that convergence to a fixed pint of the iterative process will occur, especially when parameters are not orthogonal [16]. However, our simulation studies show that in general convergence to a fixed point is rapid.

2.3 Simulation studies

We design a number of simulation studies to investigate the use of the BT-WCL method for the inference of spatial Gaussian process regression models.

2.3.1 Simulation 1: only range parameter is unknown

We simulate N process realizations from a Gaussian process with exponential covariance function at n spatial locations. We set the true value of the variance parameter σ^2 to be 1 and experiment with a range of true values of the range parameter ϕ. We first consider the situation where only the range parameter is unknown while the other parameters are fixed. We compare the estimators under different CL methods via Monte Carlo simulation results by setting $N = 1000$ replicates. We set $n = 100$ spatial locations to make it computationally feasible to obtain the results of the optimal weight CL estimator $\mathbf{W}_{Opt}(\phi)$ for our comparison analysis. To avoid numerical singularities, locations of observations are generated following a sampling approach similar to the one in [11]. Specifically, a two-dimensional regular grid is first generated with increments of 2 over the domain $[0, \sqrt{100N}] \times [0, \sqrt{100N}]$. Then each grid point is perturbed by adding a random noise, uniformly distributed on $[-0.5, 0.5]$ to each coordinate. In this case, each perturbed gridpoint is at least 1 unit away from any of its neighbors. Finally, n locations are randomly chosen from the perturbed grid points without replacement.

Figure 2.2 presents the boxplots of the estimates using the five CL methods (Eq-WCL, WCS, $1/0$-WCL, BT-WCL, Opt-WCL) and the MLE method for various values of ϕ. No strong biases are observed across all the CL estimators. In contrast, evident differences

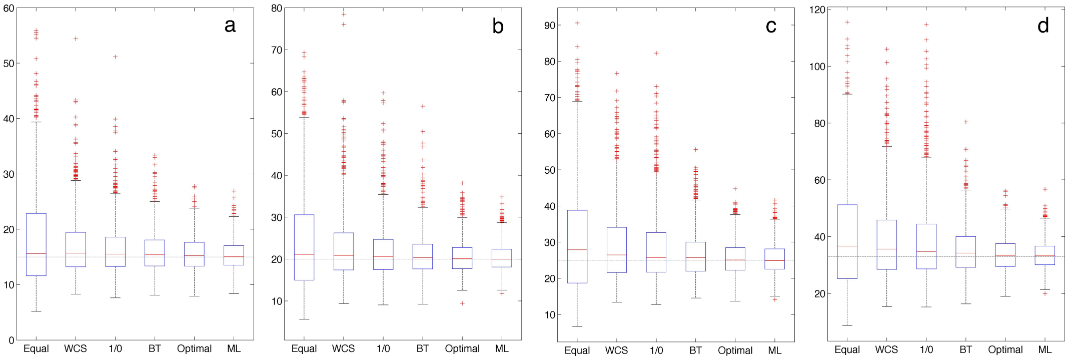

Figure 2.2: The boxplots of estimates for $\hat{\phi}$ using various methods. Those include Eq-WCL, WCS, 1/0-WCL, BT-WCL, Opt-WCL and MLE when (a) $\phi = 15$, (b) $\phi = 20$, (c) $\phi = 25$ and (d) $\phi = 30$ with observation number $N = 100$. Dot-dashed horizontal lines represent the true value of ϕ. The variance is known as $\sigma^2 = 1$.

in standard error for the 6 estimators are observed from the boxplots. Generally, the standard error for all estimators becomes larger as the range parameter ϕ increases. We also calculate the relative efficiency (RE) between each of the 5 versions of composite likelihood ($\mathbf{W}_{BT}(\phi)$,$\mathbf{W}_{equal}(\phi)$, $\mathbf{W}_{1/0}(\phi)$ and $\mathbf{W}_{WCS}(\phi)$, $\mathbf{W}_{Opt}(\phi)$) and the MLE, denoted as $MSE(\cdot)/MSE(MLE)$, and report the results in Figure 2.3 (a). As expected, the highest RE among the five CL methods is achieved by using the Opt-WCL method. Moreover, all of the three adaptive weight CL approaches (BT-WCL, WCS, 1/0-WCL) show better performance than the equally weighted CL. Among them, the estimates obtained from the proposed BT-WCL estimator yield a value of RE closest to that of the Opt-WCL estimates, which is consistent with the findings in Section 2.2.2. In this study, the RE of the 1/0-WCL and the WCS are generally $30\% - 65\%$ lower than that of the BT-WCL. We also examine the performance of these CL methods for two larger numbers of locations $n = 400$ and $n = 1000$, respectively. The results of the $\mathbf{W}_{Opt}(\phi)$ are not shown since the computation becomes formidable for large n. Overall, the results in Figure 2.3(a) and (b) indicate that the use of the BT weight achieves significant efficiency gains over the other CL methods in parameter estimations. Efficiency gain also seems to be stronger as spatial dependence range becomes shorter.

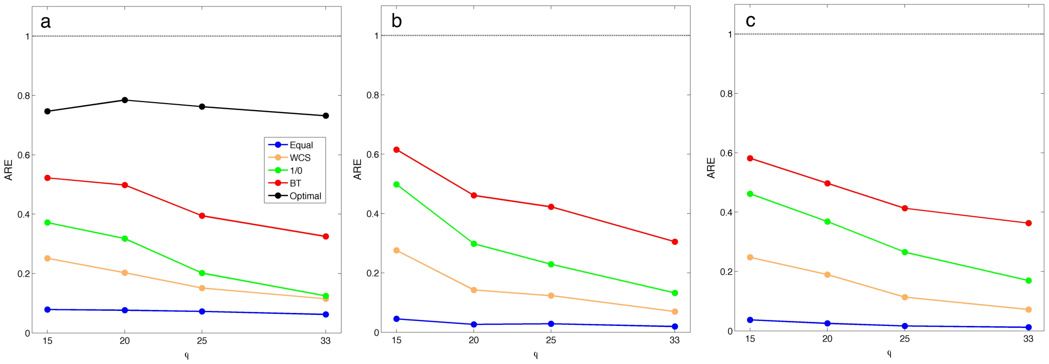

Figure 2.3: The relative efficiency (RE) of CL estimates for ϕ using different weighting schemes. RE are calculated in case of (a)$N = 100$, (B)$N = 400$ and (c)$N = 1000$. The variance is known as $\sigma^2 = 1$. Dot-dashed horizontal lines represent $RE = 1$.

2.3.2 Simulation 2: all parameters unknown

To examine the performance of the iterative CL method for the case with multiple parameters described in 2.2.2, in this study we consider a similar simulation design as the one in 2.3.1 but assume both ϕ and σ^2 are unknown. For each of the 1000 simulation replicates, we generate data at $N = 1000$ sampling locations. We compare the results of the estimates of ϕ and σ^2 using the (iterative) BT-WCL, Eq-WCL, 1/0-WCL, and WCS methods. In addition, we also include the results of the estimates of $c = \sigma^2/\phi$, whose MLE has been shown to be a consistent estimator under the fixed domain asymptotics [17]. Figure 2.4 shows the RE for each estimator. Overall, we observe similar results as in the case where σ^2 is known. The (iterative) BT-WCL approach outperforms the Eq-WCL, 1/0-WCL, and WCS method. For ϕ and c, all the adaptive weighted CL methods have substantial improvement in RE over the equally weighted CL method, especially for spatial GP with smaller scale spatial dependence structures. But for σ^2, the relative efficiencies (RE) of the estimates from various CL methods are comparable with each other.

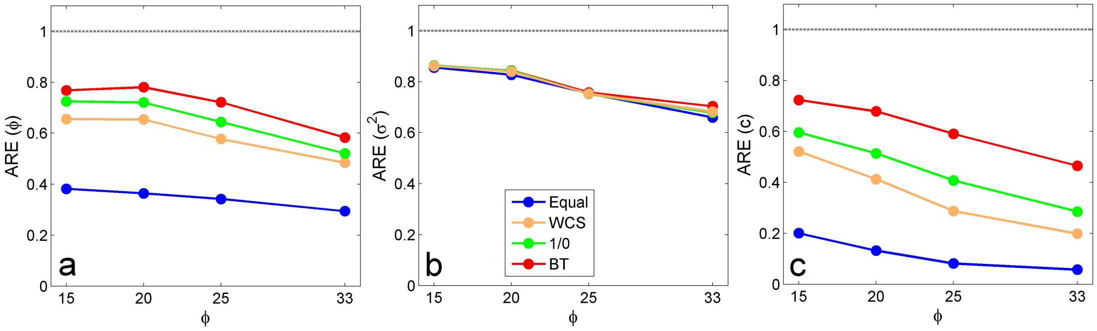

Figure 2.4: The relative efficiency (RE) of CL estimates for different parameters. Estimates for (a)ϕ, (b)σ^2 and (c)c are calculated using different weighting schemes in case of $N = 1000$. Dot-dashed horizontal lines represent $RE = 1$.

2.3.3 Simulation 3: computational efficiency

We have shown that the BT-WCL method achieves significant statistical efficiency gain in the above simulations. We now focus on examining the performance of the BT-WCL in terms of its computational efficiency. We use the simulation designed as study 1 but varying the number of locations K from 100 to 6400 to compare the computation time associated with each different methods. All computations are carried out on a 2.3GHz four-core processor with 16GB of memory.

Figure 2.5(a) shows the computation times required for a single evaluation of the full likelihood function and the composite likelihood function associated with each of the WCL methods. These are calculated by averaging over 100 repetitions of evaluation. Note that

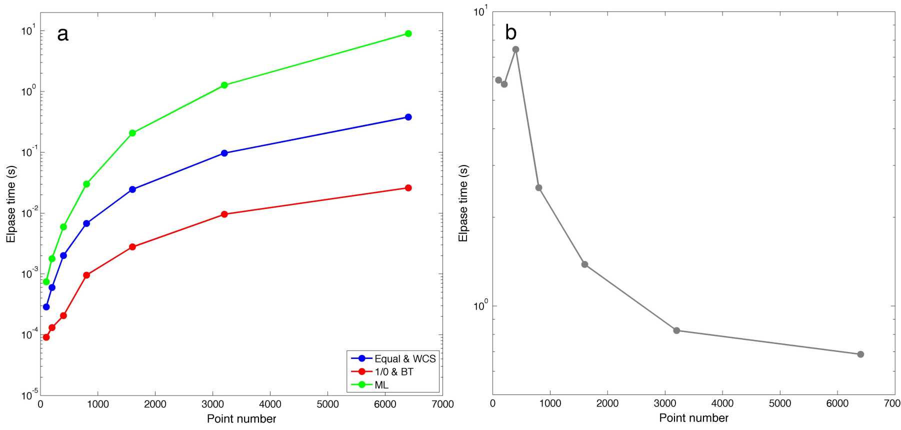

Figure 2.5: Computational burden for different weights. (a) The computation time for a single evaluation of full likelihood function and CL function with different weighting schemes. (b) The ratio of computation time for computing BT weight to that for evaluating composite likelihood.

the calculation of composite likelihood is of the same order for the BT-WCL and the 1/0-WCL method if the same taper range is used. This is also the case for the WCS and the Eq-WCL approach in which all pairs of observations are included. All the CL estimators result in a reduction of computational time compared to the MLE as expected. The computational gains of the BT-WCL are substantial: with 6400 observations, the computation time for the BT-WCL is only 0.3% of that for the MLE. It also outperforms BT-WCL the equal-weighted CL method and WCS method thanks to the exclusion of distant pairs: the computational time for the BT-WCL is only 7% of that for the WCS and equal-weighted CL method. It is also noticeable that the computational gain becomes more pronounced when increasing the number of observations, making it desirable for large spatial data sets.

We remark that the computation of the BT-WCL estimator requires precomputing the BT weight matrix to be used for the evaluation of the composite likelihood function. Therefore, we also examine the extra computational cost associated with the computation of the BT weight matrix. Figure 2.5(b) shows the ratio of the computation times between calculating weights and composite likelihood functions given weights. The result indicates that the computation cost is mainly attributed to CL function evaluation as n becomes large. The computation of the inversion of the sparse block matrix Cov($\mathbf{S}(\theta)$) used in BT-WCL does not cause a substantial increase in computational burden.

2.4 Application to precipitation data

We illustrate the BT-WCT method using the yearly total precipitation anomalies at 7,352 weather stations from the year 1962 in United States. This large, irregularly spaced spatial data was used by [11]. The yearly totals precipitation anomalies is yearly totals standardized by the long-term mean and standard deviation for each station. [11] mentioned it shows no obvious nonstationarity and anisotropy. Therefore, we fit to the data Gaussian process model with an exponential covariance function (without nugget effect), which is stationary and isotropic.

Table 2.1: Estimates of ϕ, σ^2 and c using MLE and CL method with different weighting schemes. The bottom row presents the computation time required for a single evaluation of full likelihood function and composite likelihood.

Parameter	MLE	BT-WCL	Eq-WCL	WCS	1/0-WCL
ϕ(km)	65.9	101.4	203.8	162.8	145.8
σ^2	0.72	0.77	0.78	0.77	0.77
c	0.011	0.008	0.004	0.005	0.005
Times(s)	12.64	0.03	0.47	0.47	0.03

The estimates for ϕ, σ^2, and c using MLE and CL method with different type of weights are provided in Table 2.1. The MLE for ϕ and σ^2 are 65.9 and 0.722, respectively, while the CL estimators are 101.4 and 0.765, larger than MLE. Among various weighted CL estimates, BT-WCL estimate is the closet to MLE. It is not surprising that the equally weighted CL estimate is furthest from the MLE. As to the computation cost, the CL estimators lead to a substantial reduction of computational time compared to the MLE (Table 2.1). The MLE requires 12.64 seconds to evaluate a single likelihood, which is about 400 times of the cost of BT-WCL.

2.5 Conclusions and discussion

The section addressed the problem of estimating covariance parameters of spatial Gaussian processes when the dataset is large and irregularly spaced. We have proposed a new adaptively weighted CL method, i.e., the BT-WCL method. The BT weight is an approximation to the optimal weight derived from the theory of optimal estimation equation. It is calculated with the strategy of combining block-diagonal feature and tapering. This weighting scheme leads to a considerable reduction of computational burden and retains sound estimation efficiency.

We have shown the utility of our method through simulations and data examples. In this section, we only investigate the use of a BT weight for the estimation of a spatial Gaussian process with exponential covariance function. It is possible to extend the techniques to more generalized covariance function with decay correlation with distance. Indeed, we also find that the block-diagonal feature of the covariance of scores exists for the model with power exponential covariance in an unreported simulation study. A challenge in some of these cases will be to analytically evaluate the covariance of scores. However, sampling or subsampling based methods might be adopted to estimate them as done in [18] and [10]. Finally, our method of estimating the optimal weight in WCL by blocking and tapering also has great potential to be applied for non-Gaussian spatial data in the context of copula models or spatial generalized linear models. These topics will be investigated in future work.

3. SPATIAL CLUSTERED COEFFICIENT REGRESSION MODEL

3.1 Introduction

Numerous problems in environmental, earth, and biological sciences nowadays involve large amounts of spatial data, obtained from remote ground sensors, satellite images, scientific climate computer models, geographic information systems, public health and spatial genetics, etc. In many such applications, a main problem of interest is to explain the variability in a response variable observed over the region of interest using a set of explanatory variables, considering spatial dependence of observations.

Spatial regression models such as Gaussian process regression or spatial generalized linear regression models have been widely adopted to address this problem, in which spatial dependence is accounted for by adding a spatial random effect to the (generalized) linear regression models. Regression coefficients in such models are often assumed to be a constant. However, in many problems especially when data are collected across a large region, it is unlikely that a constant regression coefficient can adequately capture the spatially dynamic relationship between response variables and covariates. One example is to have clustered pattern of regression coefficients that abruptly change across the boundary of adjacent clusters but stay relatively homogeneous within clusters. Indeed, it is of great interest to many practitioners to identify such clusters that allow them to explain varying associations between responses of interest and covariates.

Our specific motivation problem is from an important scientific question in oceanography. Geophysical fluids (i.e., air and sea water) consist of distinct fluids masses [19]. Within each fluids mass, the physical and chemical properties are relatively homogeneous. But they change rapidly across the narrow boundary between adjacent fluids masses (termed as fronts in geoscience). Such a phenomenon is formed as a result of nonlinear nature of geophysical fluids dynamics and ubiquitous in the atmosphere and ocean [20]. When exploring the relationship between features of fluids, it is likely that the relationship will change abruptly across the fronts. One notable instance is the relationship between temperature and salinity of sea water (referred to as the T-S relationship henceforth). In oceanography, temperature and salinity are two important features of water masses and strongly affect the ocean currents [19]. Knowledge of spatial distribution of T-S relationship in the ocean provides important information on the movement and extent of individual water masses. Such information can be further used to monitor the pathway and strength of meridional overturning circulation (MOC) which plays a key role in the global climate system. It is desirable to built a model with the ability to capture such spatial structures in the effect of the explanatory variables.

However, to the best of our knowledge, there are very limited work on spatially clustered coefficient regression models. Thus far, literature on spatially varying coefficient models mainly focus on smoothly varying coefficient models. Geographically Weighted Regression (GWR) [21] and Spatially-Varying Coefficients (SVC) model [22] are two popular models of this type. GWR is an ensemble of spatial local regression models fitted separately. That is, a linear regression model is fitted at each spatial point, giving greater weights to the closer points. In SVC model, the spatially-varying coefficient surface is modeled as a multivariate spatial process with stationary specification. The SVC adopts a Bayesian approach with posterior inference for all attainable model parameters and thus offers a richer inferential framework. [23] and [24] compare these two methods and suggest that SVC generally produces more accurate inferences on the regression coefficients especially in the presence of strong colinearity among explanatory variables. However, the superiority of SVC in estimation accuracy is at the cost of much higher computational burden due to the requirement of Metropolis MCMC. For a moderate data size of 1000 points and three spatially varying coefficients, it may take several days to collect sufficient

MCMC samples [23]. This largely limits its application to large spatial datasets which become more and more common with advances in observational techniques. Moreover, both GWR and SVC implicitly or explicitly assume stationarity in the spatially-varying coefficient surface over the space as neither the weighting function in GWR or covariance function of the coefficient process in SVC depends on the location. However, such a simple assumption would not be consistent with the complicated spatial structure of regression coefficients in certain applications. The coefficients could vary more rapidly in some parts of the domain than others.

In this section, our main contribution is to propose a new spatial modeling approach to estimate regression coefficients with the presence of a spatial pattern, especially a clustered pattern, influencing the effect of the explanatory variables without assuming stationarity in effect. The proposed method, called Spatially Clustered Coefficient (SCC) regression, employs penalized least square by penalizing the pairwise difference of regression coefficients between two locations that are connected by an edge in the graph. Edge selection in our problem is challenging since there is no clear ordering of spatial points. Previous studies such as 2d (gridded) fused lasso [25] considers all the edges in a lattice graph, which can have costly computational complexity for large spatial data sets. We implemented a spatial graph based on minimum spanning trees (MST). Two coefficients with locations connected by a MST tends to be similar. Therefore, the SCC model can capture the spatial structures in coefficients by encouraging the homogeneity of the coefficient on proximate locations. Even for the smoothly varying coefficients, we will see in the simulations that SCC model has strong local adaptivity and can also accurately detect a spatially highly variable pattern. With this property, the SCC model allows researchers to explore the spatial structures in regression coefficients either clustered or smoothly varying. In practice, the proposed method is computationally highly efficient, thanks to the use of MST and transformation that reduces the problem to a usual lasso-type optimization with $n - 1$

penalty terms for a spatial data set of size n.

The rest of the section is organized as follows. Section 3.2 details the Spatial Clustered Coefficient (SCC) model and discusses theoretical properties of SCC. In Section 3.3, a series of simulation studies are designed to illustrate the performance of SCC. An application of the method is presented in Section 3.4 using the aforementioned temperature and salinity data in the Atlantic basin. Section 3.5 summarizes the major conclusions of this study followed by discussion. Related proof are provided in the Appendix A.

3.2 Methodology

3.2.1 Spatially clustered coefficient model

Suppose we observe spatial data $\{(\mathbf{x}(s_i), Y(s_i)), i = 1, \ldots, n\}$ at locations $s_1, \ldots, s_n \in \mathbb{R}^2$, where the response variable $Y(s_i)$ is assumed to be spatially correlated and $\mathbf{x}(s_i) = (x_1(s_i), \ldots, x_p(s_i))^{\mathsf{T}}$ is the p-dimensional vector of explanatory variables for the observation located at s_i. The intercept can be included by defining $x_p(s_i) = 1$ for $i = 1, \ldots, n$. Consider the standard linear regression, $Y(s_i) = \mathbf{x}(s_i)^{\mathsf{T}}\boldsymbol{\beta} + \epsilon(s_i)$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathsf{T}}$ is the vector of regression coefficients and $\epsilon(s_i)$'s are independently identically distributed random noises with mean 0 and variance σ^2. Without loss of generality, we assume that the explanatory variables are standardized to have mean 0 and unit variance.

The extension of the linear regression model to allow spatially varying regression coefficients is straightforward,

$$Y(s_i) = \mathbf{x}(s_i)^{\mathsf{T}}\boldsymbol{\beta}(s_i) + \epsilon(s_i). \tag{3.1}$$

However, in many spatial datasets, it is very common to observe only one or a limited number of replicates at each location, making the above model ill-posed if without any assumptions on $\boldsymbol{\beta}(s_i)$. Indeed, strong spatial patterns of $\boldsymbol{\beta}(s_i)$ do exist since association between response variables and explanatory variables at nearby locations are expected to be highly homogeneous. This motivates us to assign a regularization function for $\boldsymbol{\beta}(\boldsymbol{s})$ utilizing their spatial homogeneity patterns.

Specifically, we propose to estimate $\boldsymbol{\beta}$ by minimizing the following objective function

$$\sum_{i=1}^{n} \{Y(\boldsymbol{s}_i) - \sum_{k=1}^{p} \mathbf{x}_k(\boldsymbol{s}_i)\beta_k(\boldsymbol{s}_i)\}^2 + \sum_{k=1}^{p} \sum_{(i,j)\in\mathbb{E}} P_\lambda(|\beta_k(\boldsymbol{s}_i) - \beta_k(\boldsymbol{s}_j)|), \tag{3.2}$$

where $\mathbb{E}$ is a set of coordinate pairs.

The term P_λ is a penalty function to encourage homogeneity between two regression coefficients if their corresponding locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are connected by an edge in $\mathbb{E}$. Here λ is a regularization parameter determining the strength of penalization. The selection of the penalty functions P_λ, the edge set $\mathbb{E}$ and tuning parameters λ are three key ingredients in the model in (3.2). Below, we discuss strategies to address these problems.

3.2.1.1 Selection of the penalty function P_λ

There are various forms of penalty functions encouraging sparsity in the literature of variable selection. The simplest and perhaps the most widely adopted one is the Lasso [26] that employs an L_1-penalty of the form

$$P_\lambda(t) = \lambda|t|. \tag{3.3}$$

As the penalty (3.3) is a convex function, efficient convex optimization algorithms can be readily applied. In this case, the L_1 penalty enforces sparsity of the difference in two edge-connected coefficients. This method allows the estimation of regression coefficients with a spatially piece-wise constant if edge sets are selected appropriately to incorporate spatial information. The non-zero elements of the estimated $|\beta_k(\boldsymbol{s}_i) - \beta_k(\boldsymbol{s}_j)|$ correspond to boundary points, whereas any two edge-connected coefficient with zero difference be-long to the same cluster. Naturally, spatial cluster for each explanatory variable can be automatically detected. However, as Lasso assigns large penalties to large values of t, it tends to underestimate t, in our case, the difference in two regression coefficients, when its true value is large. To remedy this flaw, various penalty functions have been proposed, including adaptive Lasso [27], smoothly clipped absolute deviation (SCAD) [28], minimax concave penalty (MCP) [29], and reciprocal L_1-regularization (rLasso) [30]. Adaptive Lasso assigns larger weights to the terms with small values in the L_1 penalty. SCAD and MCP adopt some concave functions that converge to constants as the penalized term becomes large. The rLasso uses a class of penalty functions that are decreasing in $(0, \infty)$ with a discontinuity at 0 and converging to infinity when the penalized term approaches zero. These two forms have smaller estimation errors compared to Lasso, which, however, is at the expense of considerable increase in computational cost. Therefore, in practice, penalty functions are often selected by weighing a trade-off between statistical efficiency and computational complexity for specific problems. It should be noted that the reduction of computational burden is critically important in spatial analysis as large datasets have become common in diverse fields such as geoscience, ecology, and econometrics. In this section, we mainly focus on the Lasso penalty to demonstrate the power of SCC method for computational simplicity and conservative comparison. We remark that using more advanced penalty forms may further improve the performance of SCC method.

3.2.1.2 Edge selections based on minimum spanning tree

Here we pay special attention to the selection of an edge set $\mathbb{E}$. In spatial problems, each location is a node in a graph. If we believe that the physical properties of geographical locations with close distance tend to be homogeneous, it is not unreasonable to anticipate similar coefficients for proximate locations. Therefore, $\mathbb{E}$ should only include the coordinate pairs that are close to each other. A simple choice of $\mathbb{E}$ satisfying this criterion is the set consisting of neighboring coordinate pairs, i.e., $\mathbb{E} = \{(s_i, s_j) : i = 1, \ldots, n; s_j \in \mathbb{N}_{s_i}\}$ where $\mathbb{N}_{s_i}$ is the set representing the neighbors of s_i.

However, it is known that, unlike temporal data, spatial data do not have a natural ordering, making it challenging to construct $\mathbb{E}$.

The neighboring set $\mathbb{N}_{s_i}$ can be defined using either the nearest neighbor of s_i or four neighbors with common borders for grid data such as in 2D fused lasso. Although such selections of $\mathbb{E}$ appear to be natural, they suffer from two evident deficiencies. First, $\mathbb{E}$ defined above does not necessarily connect all the data points together. In this case, (3.2) does not reduce to a constant regression coefficients model, when $\lambda \to \infty$. Second, the set $\mathbb{E}$ may include redundant coordinate pairs, imposing great computational challenges [31, 32]. For example, consider a regular grid consisting of n points. The number of penalty terms is $2n - 2\sqrt{n}$, many of which are redundant.

The aforementioned analysis suggests that an appropriate choice for $\mathbb{E}$ should only include coordinate pairs close to each other, lead to connectivity of all data points, and have no redundant pairs. One choice of $\mathbb{E}$ satisfying all the three criteria is the minimum spanning tree (MST). For spatial data, we can construct an undirected graph, a set of vertices connected pairwise by edges, $G = (\mathbb{V}, \mathbb{E})$ consisting of the set $\mathbb{V}$ of vertices and the set $\mathbb{E}$ of edges. In particular, the vertices correspond to the spatial locations and weights of edges correspond to the distance between two locations. In graph theory, a minimum spanning tree (MST) is a sub-graph that connects all the vertices together without any cycles and with the minimum total edge weight. The edge set $\mathbb{E}$ constructed from the MST automatically satisfies the three criteria by definition and is thus an appropriate choice. We choose $\mathbb{E}$ as the edge set of the minimum spanning tree in this study and the corresponding penalized least square (3.2) with the Lasso penalty becomes

$$\sum_{i=1}^{n}\{Y(s_i)-\sum_{k=1}^{p}x_k(s_i)\beta_k(s_i)\}^2+\lambda\sum_{k=1}^{p}\|\mathbf{T}\boldsymbol{\beta}_k\|_1, \tag{3.4}$$

where $\boldsymbol{\beta}_k=(\beta_k(s_1),\dots,\beta_k(s_n))^{\mathsf{T}}$ and $\mathbf{T}$ is a $(n-1)\times n$ matrix constructed from the edge set $\mathbb{E}$ in the MST. $\mathbf{T}$ has full rank and each row vector of $\mathbf{T}$ only contains two non-zero elements, 1 and -1, by this construction.

3.2.1.3 Selection of tuning parameter λ

When $\lambda\to\infty$, the model (3.2) yields a constant regression coefficient; when $\lambda=0$, it reduces to the ordinary least square with all different coefficients across the region. With an appropriate λ, the penalized least square model (3.2) produces clustered regression coefficients. In practice, the optimal λ can be determined via some data-dependent model selection criteria, such as generalized cross-validation (GCV) [33], Bayesian information criterion (BIC) [34] and extended Bayesian information criterion (EBIC) [35,36]. In this section, we use BIC instead of EBIC to choose λ since the latter tends to produce over-sparsity in the penalized term (the difference of regression coefficients in our model) according to previous studies [30].

3.2.2 Computation

The SCC model (3.2) is an optimization problem. It is easy to implement as it can be transformed into a Lasso, or Lasso type problem after suitable reparameterization. We first consider the transformed parameters $\boldsymbol{\theta}_k=(\theta_k(s_1),\dots,\theta_k(s_n))^{\mathsf{T}}$, $k=1,\dots,p$ defined as

$$\boldsymbol{\theta}_k=\begin{pmatrix}\mathbf{T}\\ \frac{1}{n}\mathbf{1}^{\mathsf{T}}\end{pmatrix}\boldsymbol{\beta}_k=\widetilde{\mathbf{T}}\boldsymbol{\beta}_k. \tag{3.5}$$

Note that the $\widetilde{\mathbf{T}}$ is a $n\times n$ invertible matrix since $\mathbf{T}$ has full rank and the row vectors of

$\mathbf{T}$ are orthogonal to the unit vector $\mathbf{1}$. Thus, there is a one-to one transformation between $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$. Define a new design matrix as

$$\widetilde{\mathbf{X}} = [diag(\mathbf{x}_1), \dots, diag(\mathbf{x}_p)]\widetilde{\mathbf{T}}^{-1}, \tag{3.6}$$

where $diag(\mathbf{x}_k)$ is a diagonal matrix with the diagonal entries $\mathbf{x}_k = (x_k(s_1), \dots, x_k(s_n))^{\mathsf{T}}$. Then the SCC model (3.4) can be rewritten as

$$\|\mathbf{Y} - \widetilde{\boldsymbol{X}}\boldsymbol{\theta}\|_2^2 + \lambda \sum_{t \in B} |\theta_t|, \tag{3.7}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathsf{T}}, \dots, \boldsymbol{\theta}_p^{\mathsf{T}})^{\mathsf{T}}$ and B represents the set $B = \{t : \mathrm{mod}(t, n) \neq 0\}$. Henceforth, we will denote $\sum_{t \in B} |\theta_t|$ as $\|\theta_B\|_1$ for neatness.

Therefore, the solution to the SCC model (3.2) with Lasso penalty can be obtained by solving the Lasso problem (3.7) with respect to the parameters $\boldsymbol{\theta}$. Estimators for $\boldsymbol{\beta}$ are given by $\widehat{\boldsymbol{\beta}_k} = \widetilde{\mathbf{T}}^{-1}\widehat{\boldsymbol{\theta}_k}$. Many efficient algorithms for the Lasso, such as LARS algorithms [37] and coordinate decent algorithm [38] can be readily applied for the SCC model. In this study, we implement the SCC method using the package *glmnet* (both in R and Matlab) which is based on the coordinate descent algorithm.

3.2.3 Theoretical properties

In this subsection, we establish the oracle inequalities for the SCC estimators. As there is a one-to one transformation between $\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$, we present theorem in terms of $\boldsymbol{\theta}$.

Assumptions 1. *(a) There is a positive constant C_1 so that $n^{-1} \sum_{i=1}^{n} \widetilde{X}_{i,t}^2 \le C_1$ for any $n > 0$ and $t \in \{1, ..., n \cdot p\}$. (b) There is a positive constant Φ so that for any vector $\mathbf{u} \in \mathbb{R}^{n \cdot p}$ satisfying $\sum_{t=1}^{n \cdot p} |u_t| \le 4\sqrt{|A|}\sqrt{\sum_{t \in A} u_t^2}$ where $A = \{t : \theta_t \neq 0\} \cup B^C$ and $|A|$ denotes its* *cardinality, we have*

$$\frac{1}{n}\mathbf{u}^{\mathsf{T}}(\widetilde{\mathbf{X}}^{\mathsf{T}}\widetilde{\mathbf{X}})\mathbf{u} \geq \Phi \sum_{t \in A} u_t^2. \tag{3.8}$$

Assumption 1 is widely adopted in previous literature [39–42]. Assumption 1(a) requires the random variables $V_t = n^{-1} \sum_{i=1}^{n} \widetilde{X}_{i,t}\varepsilon_i$ to be sub-Gaussian for any $t \in \{1, ..., n \cdot p\}$. A sufficient condition for assumption 1(b) to be satisfied is that the restriction of Gram matrix $\widetilde{\mathbf{X}}^{\mathsf{T}}\widetilde{\mathbf{X}}$ to column A is positive definite.

Theorem 1. *Suppose that Assumption 1 holds. If* $\lambda\sqrt{n}/log(n) \geq 4\sqrt{(1 + C_2)2C_1^2\sigma^2}$ *where* C_2 *is a positive constant for any* $n > 0$*, we have the following inequalities with probability tending to unity as* $n \to \infty$

$$\frac{1}{n}||\widetilde{\boldsymbol{X}}\boldsymbol{\theta} - \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\theta}}||_2^2 \leq \frac{4\lambda_n^2|A|}{\Phi}, \tag{3.9}$$

$$||\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}||_1 \leq \frac{8\lambda_n|A|}{\Phi}. \tag{3.10}$$

The detailed proof for (3.9) and (3.10) is provide in Appendix A. We note that for the case of infilling domain, $|A| \sim O(\sqrt{n})$ as $n \to \infty$. Accordingly, the right hand sides of (3.9) and (3.10) decrease asymptotically to zero as $n \to \infty$.

3.3 Simulation studies

In this section, we present two simulation studies to illustrate the performance of the SCC method under two different scenarios: the true regression coefficients having clustered pattern and smoothly varying pattern respectively.

In both studies, we use 2000 spatial locations that are randomly selected from the square domain $[-0.5, 0.5] \times [-0.5, 0.5]$. The responses at each location are generated using the linear regression model with two predictors and an intercept:

$$Y(s_i) = \beta_1(s_i)x_1(s_i) + \beta_2(s_i)x_2(s_i) + \beta_3(s_i) + \epsilon(s_i), \tag{3.11}$$

where $\epsilon(s_i) \overset{iid}{\sim} N(0, \sigma^2)$. We set σ to be 0.1 in the following simulations.

$x_1(s)$ and $x_2(s)$ are then generated by linearly transforming two independent realizations of a spatial Gaussian process with mean zero and covariance matrix defined from an anisotropic exponential function:

$$\mathrm{Cov}\{x_k(s_i), x_k(s_j)\} = \exp\left(-\sqrt{\frac{(s_{h,i} - s_{h,j})^2}{\phi_{h,k}^2} + \frac{(s_{v,i} - s_{v,j})^2}{\phi_{v,k}^2}} \right), k = 1, 2, \tag{3.12}$$

where $(s_{h,i}, s_{v,i})$ is the coordinate in the horizontal and vertical direction and $(\phi_{h,k}, \phi_{v,k})$ is the anisotropic range parameter. Specifically, suppose $x_{1,0}(s)$ and $x_{2,0}(s)$ are the two independent realizations. We let $x_{1,0}(s) = x_1(s)$ and $x_{2,0}(s) = rx_{1,0}(s_i) + \sqrt{1 - r^2}x_{2,0}(s_i)$, allowing for the colinearity between the two spatially varying predictors. In the following analysis, we let $r = 0.75$, corresponding to moderate colinearity.

We remark that numerical data analyses in previous works often generate the value of predictors from a white noise process, corresponding to the special case $(\phi_{h,k}, \phi_{v,k}) = (0, 0)$ [23, 24]. However, such a design of independent variables is far from the reality for spatial data. For example, most of the variables used in geoscience, such as temperature, precipitation, wind speed, ocean primary productivity, and dissolved oxygen in the sea water, have evident spatial structures [19]. When serving as predictors of a regression model in numerical studies, they should be generated from spatially-correlated processes. In the following simulation studies, we reveal that the spatial correlation of predictors will have a profound influence on the efficiency of estimation and prediction.

We use Gaussian processes to produce predictors, considering various extent of spatial correlation. Three combinations of range parameters $(\phi_{h,k}, \phi_{v,k})$, corresponding to weak, moderate and strong spatial correlation, are provided in the Table 3.1. To avoid loss of generosity, we allow distinct spatial structures for different predictors by assigning different values of $(\phi_{h,k}, \phi_{v,k})$ for different k. For each value of $(\phi_{h,k}, \phi_{v,k})$, we simulate 100 datasets. In each dataset, we randomly select 1000 data points out of the 2000 data points for estimation with the remaining data points held-out for prediction.

Table 3.1: Spatial range parameters for predictors used in simulation studies.

Range parameters	Weak	Moderate	Strong
$(\phi_{h,1}, \phi_{v,1})$	(0.3,0.1)	(1,0.3)	(3,1)
$(\phi_{h,2}, \phi_{v,2})$	(0.1,0.3)	(0.3,1)	(1,3)

In both studies, we compare the results with those of GWR. For the GWR method, the regression coefficients at location s_i are estimated by $\beta(s_i) = (\mathbf{X}^T\mathbf{W}(s_i)\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}(s_i)\mathbf{Y}$ where $\mathbf{X}$ is a $n \times p$ matrix with $\mathbf{x}(s_i)$ as the i-th row and $\mathbf{W}(s_i)$ is a diagonal matrix determined from a chosen spatial kernel function. Here we employ an exponential spatial kernel function with the optimal range parameter estimated through cross-validation. We use the packages *glmnet* to implement SCC method and the package *gwr* to implement GWR (both packages are available in R and Matlab). The results of the SVC model are not reported here since it is computationally too expensive for the size of the data considered here. Moreover, previous studies suggest that SVC typically produces comparable results as GWR [23].

3.3.1 Study 1: clustered coefficients

In this study, the true regression coefficients are set to be spatially clustered. Specifically, the coefficients are constant within each cluster with an abrupt change of value across the boundaries of clusters. Three different cluster patterns are assigned to each of the coefficients $\beta_1(s_i)$, $\beta_2(s_i)$, and $\beta_3(s_i)$ to reveal the ability of SCC for detecting various clusters as shown in the top panel of Figure 3.1.

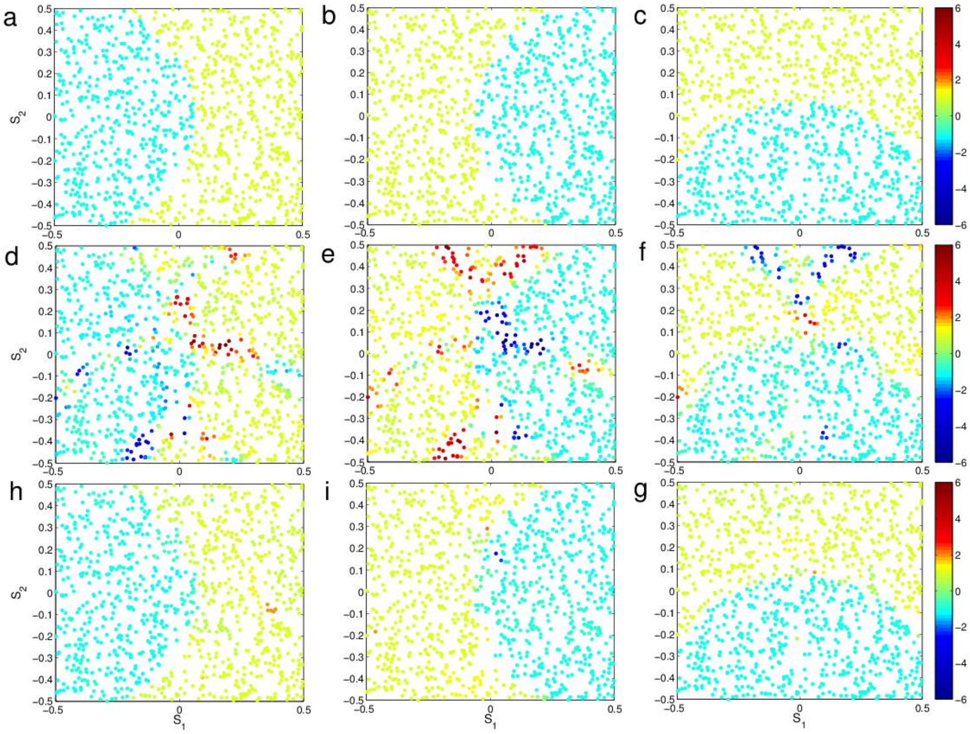

Figure 3.1: Spatial structure for clustered coefficients. Panel (a-c) corresponds to true coefficient β_1, β_2 and β_0. The estimated coefficient surface from (d-f) GWR method and (h-g) SCC method in one simulation with spatial range parameters $(\phi_{h,1}, \phi_{v,1}) = (1, 0.3)$ and $(\phi_{h,2}, \phi_{v,2}) = (0.3, 1)$ for predictors.

The estimated coefficients obtained from the GWR and the SCC methods are plotted in Figure 3.1(d-i). It is noted that spatial patterns of the coefficients derived from the SCC method shown in the bottom panel of Figure 3.1 are highly consistent with the true regression coefficients shown in the top panel. It successfully captures the clustered structure in coefficients and detects the abrupt changes across the boundaries of adjacent clusters. In contrast, the estimated coefficients obtained from the GWR method do not exhibit clear cluster structure. Specifically, the GWR method produces poor estimations of regression coefficients near the boundary of clusters, mainly due to its nature of smoothing regression coefficients.

We further examine the performance of the SCC in terms of parameter estimation and prediction at hold-out locations. Specifically, we consider the mean squared error of coefficient estimation (MSE_β) and mean squared error of prediction (MSE_p), defined as:

$$MSE_\beta = \frac{1}{|I_e|}\sum_{i\in I_e}\sum_{k=1}^{p}(\beta_k(s_i)-\widehat{\beta_k(s_i)})^2,$$

$$MSE_p = \frac{1}{|I_p|}\sum_{i\in I_p}(Y(s_i)-\widehat{Y(s_i)})^2,$$

where I_e (I_p) is the subset of data points used for estimation (prediction) with $|I_e|$ ($|I_p|$) denoting its cardinality.

The comparison of MSE_β and MSE_p for GWR and SCC, under 3 different settings of spatial dependence for covariates, is reported in Figure 3.2 and Table 3.2. For coefficient estimation, the SCC method clearly outperforms the GWR method with considerably smaller values of MSE_β in all the 3 settings. As the spatial correlation in covariates becomes stronger, the SCC estimators remain relatively more stable, whereas the performance of the GWR degrades substantially. For instance, the MSE_β for GWR estimator is 0.60 when spatial correlations in covariates is weak ($(\phi_{h,1}, \phi_{v,1}) = (0.3, 0.1)$ and $(\phi_{h,1}, \phi_{v,1}) = (0.1, 0.3)$) but increases to 6.82 in the case of strong spatial correlation

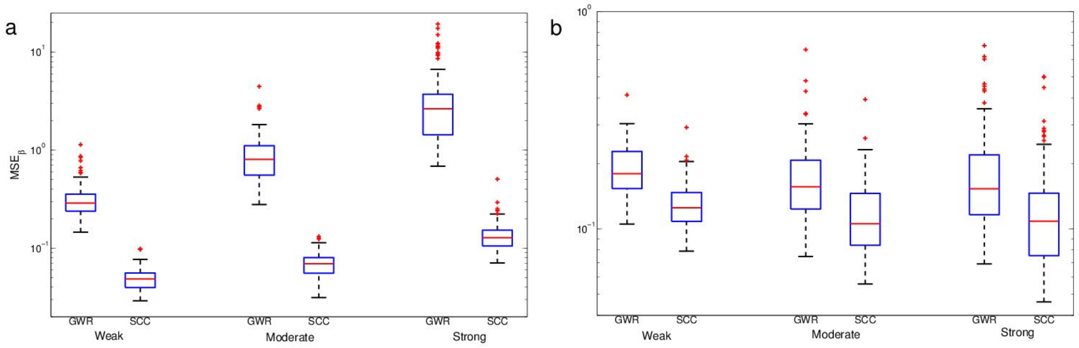

Figure 3.2: Boxplot of errors in Study 1. The (a)MSE_β and (b) MSE_p for GWR and SCC method, under 3 setting of spatial correlation (weark, moderate and strong) for predictors, based on 100 simulated datasets.

($(\phi_{h,1}, \phi_{v,1}) = (3, 1)$ and $(\phi_{h,1}, \phi_{v,1}) = (1, 3)$). In contrast, MSE_β associated with the SCC estimator changes less by a factor of 3. Therefore, the SCC method provides more robust inference on the regression coefficients especially in presence of strong spatial correlation in covariates.

Table 3.2: Summary of Study 1 (clustered coefficients). The MSE_β and MSE_p for the SCC and GWR method, under various spatial correlation for predictors.

Spatial correlation	MSE_β		MSE_p	
	GWR	SCC	GWR	SCC
Weak	0.60	0.10	0.30	0.21
Moderate	1.88	0.16	0.28	0.18
Strong	6.36	0.28	0.27	0.17

For prediction, the MSE_p for the SCC method is also systematically smaller than that of the GWR method. The improvement in prediction of the SCC method over the GWR is substantial but less striking compared with the improvement in parameter estimations. We note that although the MSE_{β} derived from the GWR method increases rapidly as the spatial correlation in covariates become stronger, its MSE_{p} is almost unchanged. This might be partly due to the canceling effect when the estimation errors in regression coefficients for individual covariates are carried to calculate the prediction of the response variable.

3.3.2 Study 2: smoothly varying coefficients

This study has a similar design as in Study 1, except that the regression coefficients are independently generated from a Gaussian spatial process and hence are smoothly varying. Here we use a zero mean and an anisotropic exponential covariance function as in (3.12). The variance parameter is fixed at 4 and the anisotropic range parameter is set to be $(3, 1)$ for $\beta_1(s_i)$, $(1, 3)$ for $\beta_2(s_i)$, and $(2, 2)$ for $\beta_3(s_i)$, respectively.

Figure 3.3 displays the true coefficient and the estimated values using GWR and SCC methods, respectively. The spatial pattern of coefficients derived from the SCC method agrees reasonably well with that of the true model. The estimates from the GWR method are, however, quite noisy with artificially large coefficient values in some parts of the domain. This is partly due the fact that the isotropic kernel function used in the GWR is too restrictive when fitting an anisotropic spatial field. In contrast, an advantage of SCC is its strong local adaptivity. The use of a local pairwise penalty function allows the fitting of a spatial field that is constant in one area but highly changing in another. Comparisons of the MSE_{β} further confirm the superiority of the SCC method in estimating the smoothly-varying coefficients (Figure 3.4). For instance, in the presence of strong spatial correlation in predictors, the MSE_{β} for SCC estimator is only 1/8 of that for GWR estimator (Table 3.3).

3.3.3 Summary of simulation results

In this section, we evaluated the performance of SCC method for estimating the clustered and smoothly-varying coefficients, respectively. In both cases, the SCC method is

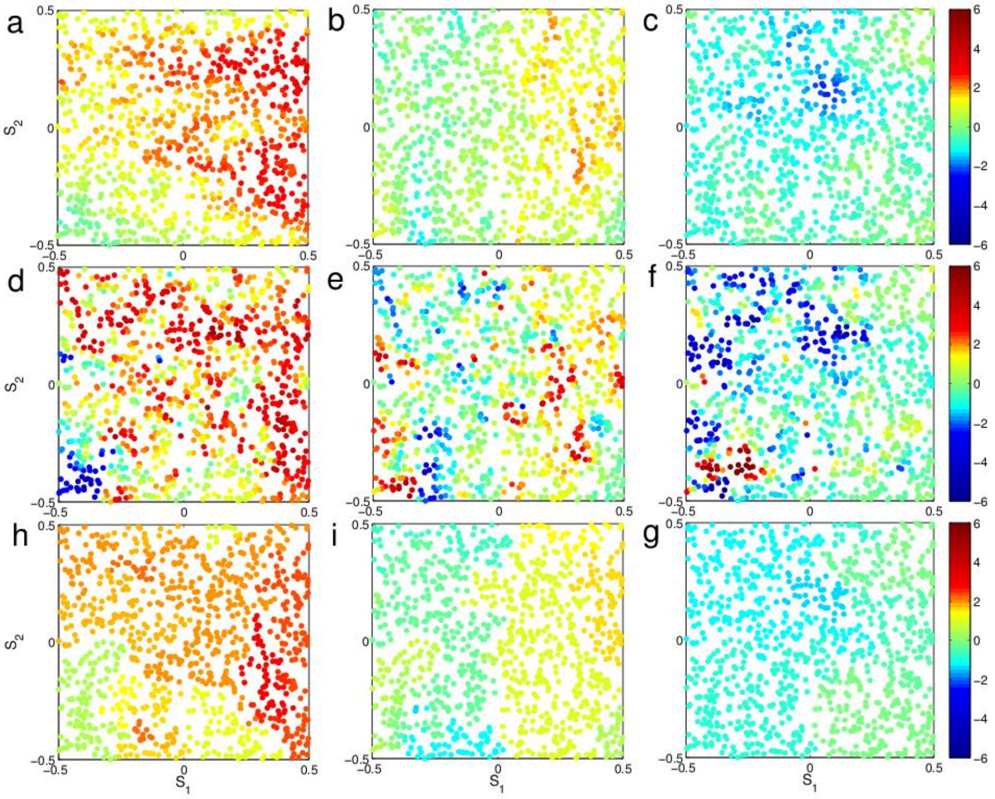

Figure 3.3: Spatial structure for smoothly varying coefficients. Panel (a-c)correspond to true coefficient β_1, β_2 and β_0. The estimated coefficient surface from (d-f) GWR method and (h-g) SCC method in one simulation with spatial range parameters $(\phi_{h,1}, \phi_{v,1}) = (1, 0.3)$ and $(\phi_{h,2}, \phi_{v,2}) = (0.3, 1)$ for predictors.

Table 3.3: Summary of Study 2 (smoothly varying coefficients). The MSE_β and MSE_p for the SCC and GWR method, under various spatial correlation for predictors.

Spatial correlation	MSE_β		MSE_p	
	GWR	SCC	GWR	SCC
Weak	0.29	0.17	0.15	0.25
Moderate	0.83	0.23	0.13	0.22
Strong	2.62	0.37	0.13	0.20

capable of capturing the spatial pattern in coefficients and outperforms the GWR method with considerably smaller MSE_β. It should be noted that Study 1 and Study 2 can be

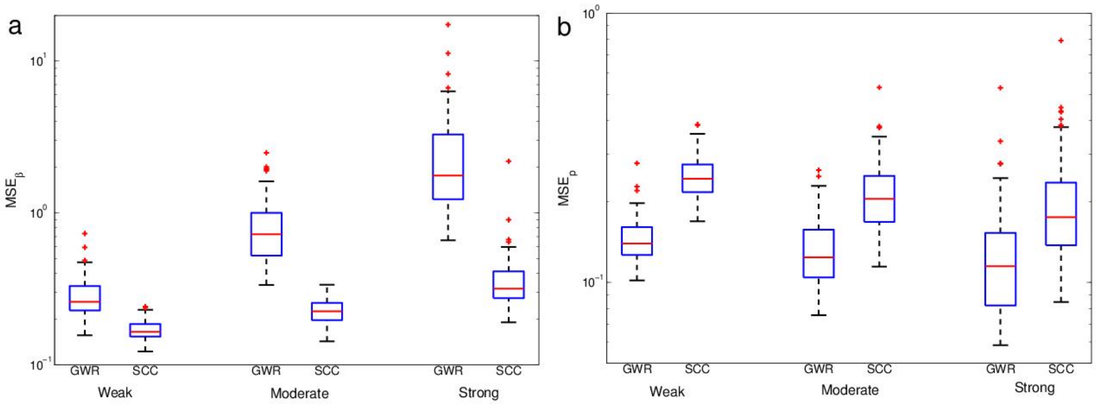

Figure 3.4: Boxplot of errors in Study 2. The (a)MSE_β and (b) MSE_p for GWR and SCC method, under 3 setting of spatial correlation (weark, moderate and strong) for predictors, based on 100 simulated datasets.

treated as the two extreme cases for the spatial pattern of coefficients. The former is consistent with the assumption underpinning the SCC method while the latter favors the setting (i.e., a global spatial kernel function independent from location) of the GWR method. Most of the applications probably lie in between these two extreme cases. Therefore, we expect that the SCC method should have a superior performance to the GWR method for estimating the spatially varying coefficients in real examples, which will be demonstrated in the following section using the hydrographic data in the ocean.

3.4 Real data analysis

3.4.1 Dataset

An application of the SCC method is performed by analyzing the temperature-salinity relationship in the Atlantic Ocean, along with a comparison analysis with the GWR and OLS methods. The temperature and salinity records are obtained from the World Atlas 2013 version 2 (WOA 13 V2) archived at National Oceanographic Data Center (available through https://www.nodc.noaa.gov/OC5/woa13/). WOA is a dataset of objectively analyzed climatological fields of temperature, salinity, dissolved oxygen, Apparent Oxygen Utilization (AOU), percent oxygen saturation, phosphate, silicate, and nitrate derived by combining all the available observations. All the variables are provided on regular grids. The grid size in the zonal and meridional directions is 1 ° while the vertical grid size increases from 10 m near the sea surface to 200 m just above the sea floor. The non-uniform vertical grid size is typical in oceanic datasets and is rationalized by the fact that oceanic variables like temperature and salinity change much more rapidly in the upper ocean than in the abyss.

To facilitate analysis, we take a meridional segment of temperature and salinity in the Atlantic basin along 25°W between 60°S-60°N (Figure 3.5), leading to 11166 grid points in total. This is a standard segment widely used in oceanographic studies as it is well representative of spatial variations of oceanic variables. Figure 3.5a and 3.5b display the spatial distributions of temperature and salinity along the 25°W segment, respectively.

There are three notable features. First, the temperature and salinity are not randomly distributed but have well organized spatial structures. Specifically, the temperature is generally higher at the lower latitudes and in the upper ocean as a result of solar radiation. The spatial distribution of salinity is somewhat more complicated. Near the sea surface, the salinity values peak around 30°S and 30°N due to the low precipitation rates at these latitudes. In addition, there is a pronounced low-salinity tongue originating from the sea surface around 50°S-60°S and extending northward and downward to the equatorial region at 1000 m or so. This low-salinity tongue corresponds to the Antarctic Intermediate Water (AAIW). The encounter of AAIW with high-salinity water mass centered at 30°S leads to a strong salinity front.

The second notable feature is that the temperature and salinity are highly anisotropic. The temperature and salinity gradients in the vertical directions are several orders of magnitude larger than those in the horizontal direction. The anisotropy is essentially a result

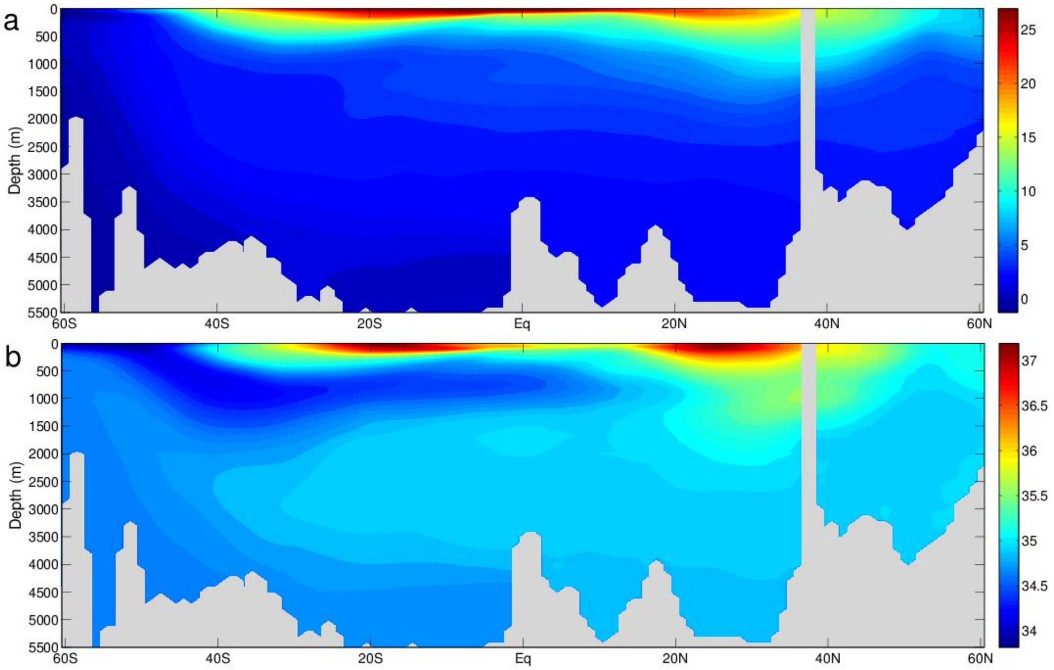

Figure 3.5: Spatial distribution of (a) temperature in °C and (b) salinity in PSU along the meridional segment 25°W.

of ocean geometry. It has a width of around 20000 km but a thickness about 4 km. To account for the geometry of ocean, oceanic studies typically adopt non-dimensional coordinates $(s'_h, s'_v) = (s_h/L, s_v/H)$ where s'_h (s'_v) is the non-dimensional horizontal (vertical) coordinate and L (H) is the horizontal (vertical) length of ocean, respectively [20]. In the non-dimensional coordinates, the magnitudes of horizontal and vertical gradients are on the same order of magnitude, largely eliminating the anisotropy. In this study, we will adopt this scaling technique following the convention of oceanic studies.

Finally, the distributions of temperature and salinity appear to be non-stationary. As mentioned above, there is strong salinity gradient around the front formed by the AAIW and high-salinity water mass centered at 30°S. In addition, there is also a strong temperature gradient near the sea surface around 40°S and 40°N. These temperature fronts are maintained by the energetic eastward ocean currents through the thermal wind relation. Furthermore, both the gradients of temperature and salinity are generally stronger in the upper ocean than in the abyss. This is because the turbulent mixing, a process homogenizing the fluids properties, dominates the evolution of temperature and salinity in the abyss [20].

3.4.2 Analysis results

To explore the spatial structure of the T-S relationship, We adopt the proposed SCC model (3.2) with Lasso penalty function (3.3) and the regression model below:

$$S(s_h{}', s_v{}') = \beta_1(s_h{}', s_v{}')T(s_h{}', s_v{}') + \beta_0(s_h{}', s_v{}'),$$

where the response variable $S(s_h{}', s_v{}')$ denotes salinity, $T(s_h{}', s_v{}')$ denotes temperature, regression coefficient $\beta_1(s_h{}', s_v{}')$ measures the T-S relationship of interest, and $\beta_0(s_h{}', s_v{}')$ is the intercept.

To quantitatively evaluate the performance of the SCC, GWR and OLS methods, we compare the prediction error measured by $MSEP$ for the three methods. First, N_t points are randomly selected from the 11166 observations for estimation with the remaining points for prediction. Next, let $N_t = 1000$, 2000 and 4000 to partition the data into training and testing sets to investigate the influence of sample size on prediction efficiency. Finally, for each value of N_t, we repeat the random partition of data sets for 100 times and calculate the MSE_P for each dataset.

Figures 3.6a and 3.6b display the estimated coefficient surface for $\beta_1(s_h{}', s_v{}')$ from the SCC and GWR methods, respectively. Although the true value of β_1 is not available in this case, the performance of SCC and GWR methods can still be inferred from their estimated surface of β_1. The $\beta_1(s_h{}', s_v{}')$ estimated from the SCC method has a well organized spatial structure. Its value is positive throughout the segment except in the AAIW.

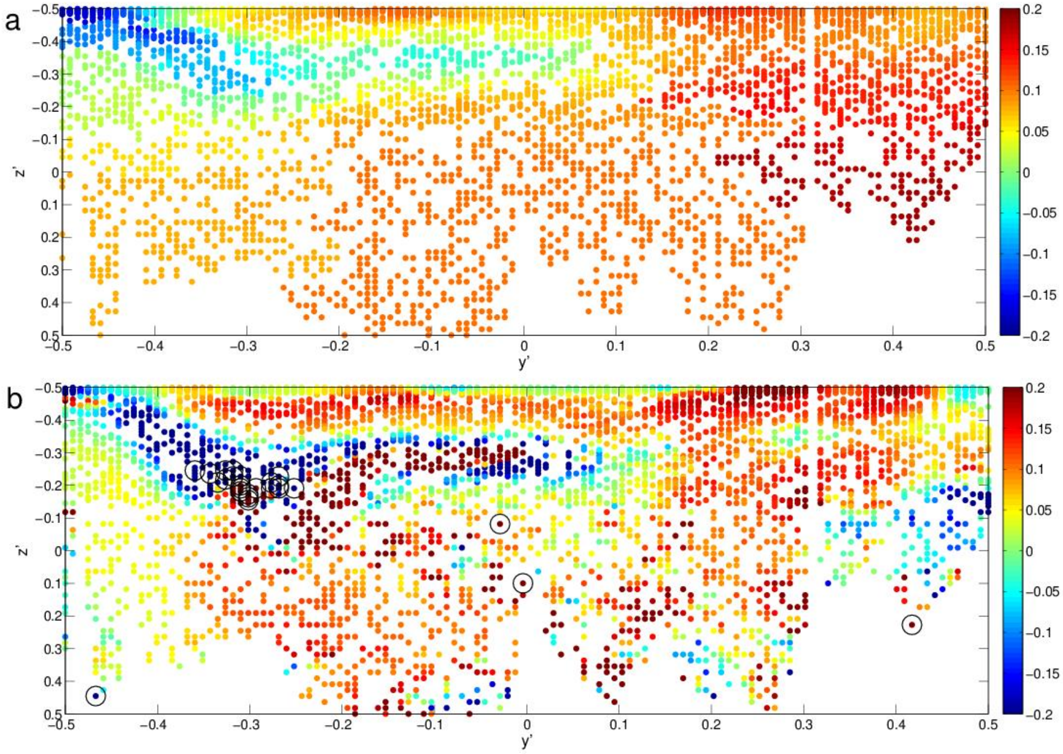

Figure 3.6: The T-S relationship β_1 estimated from the (a) SCC method and (b) GWR method. Computation are implemented based on 4000 data points randomly selected from the 11166 data points. Note that the colorbar in (b) is saturated. The largest positive and negative values of the GWR estimator are 11.99 and -3.61, respectively. Data points with $|\beta_1| > 1$ are marked by grey circles.

This leads to a rapid change of $\beta_1(s_h', s_v')$ in the frontal region between the AAIW and high-salinity water mass centered around 30°S. Such a rapid change is not unreasonable as these two water masses are formed through different dynamics and characterized by distinct water properties [19]. Furthermore, the $\beta_1(s_h', s_v')$ value estimated from the SCC method is generally more variable in the upper ocean than in the abyss. This feature is also supported by ocean dynamics: in the upper ocean, there are many active dynamical processes affecting the T-S relationship, such as air-sea interaction, turbulent mixing, and advection [20]; different dynamical processes may dominate in different parts of the upper ocean, contributing to a spatially varying T-S relationship; whereas the abyssal ocean is much less active and thus a relatively uniform T-S relationship is expected [20].

The estimator from the GWR, however, is quite noisy with occasional outliers. We note that the noisy result similarly occurs in the simulation studies (see Section 3.3), which strongly implies that such noise in estimation is perhaps not an actual feature of this data but probably due to the deficiency of the GWR method in the case with spatially correlated explanatory variables. Indeed, the noise, especially in the abyss, is not consistent with the oceanic dynamics since there is no dynamical process that can lead to changes of the T-S relationship at such a short distance in the abyssal ocean [19]. According to the above heuristic interpretation of the results with regard to ocean dynamics, the SCC method tends to produce a more reasonable estimate for the T-S relationship than the GWR method.

We further examine the performance of SCC in terms of prediction. A summary of the prediction error (MSE_p) for the three method, SCC, GWR and OLS, is provided in Figure 3.7 and Table (3.4). The MSE_p for SCC method are consistently smaller than that for GWR and OLS. The boxplot of MSE_p for GWR spreads much more widely and exhibits more unrealistic predictive outliers, compared with that for SCC and OLS. This inferiority of GWR is highly notable when the sample size N_t is small. These findings suggest that the predictive performance of GWR method can be unstable and hence unreliable especially when data size is small. The large variability of MSE_p for the GWR method is partially attributed to the severely biased estimator of coefficients as we observed in Figure 3.6b.

Concerning that the mean value of MSE_p for the GWR method may be dramatically affected by a few outliers, we also compute the median value of MSE_p for various methods (Table 3.4). The SCC method again outweighs the GWR method in terms of both the mean and median of MSE_p. For example, the mean MSE_p for GWR method is 36.7×10^{-3}, 12.7×10^{-3}, and 45.2×10^{-3} for $N_t = 1000, 2000$, and 4000, respectively. But the corresponding values for SCC method is 0.9×10^{-3}, 0.8×10^{-3}, and 0.8×10^{-3}.

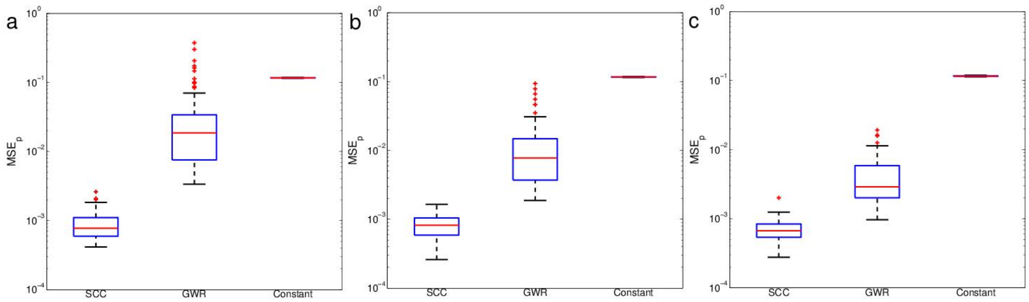

Figure 3.7: Boxplots of MSE_p for temperature and salinity dataset using SCC, GWR and OLS method. N_t = (a) 1000, (b) 2000, and (c) 4000 data points are used for estimation.

Table 3.4: The mean and median of MSE_β and MSE_p for the SCC, GWR and OLS method. The unit here is PSU^2.

Sample size	Mean MSE_β			Median MSE_p		
	GWR	SCC	OLS	GWR	SCC	OLS
$N = 1000$	0.0367	0.0009	0.1165	0.0186	0.0008	0.1164
$N = 2000$	0.0127	0.0008	0.1166	0.0077	0.0008	0.1165
$N = 2000$	0.0452	0.0007	0.1161	0.0029	0.0007	0.1158

Both SCC and GWR methods are superior to the OLS model with constant regression coefficients, as evidenced by their much smaller mean and median MSE_p. This highlights the necessity of allowing regression coefficients to vary over space. Moreover, we note that there is a gain in prediction efficiency for all three methods as the sample size N_t increases. These comparisons demonstrate that the SCC method is more robust than the GWR method.

3.5 Conclusions and discussion

When exploring complicated phenomena in ecology, econometric and environmetrics, it is desirable to built a spatial model with the ability to capture the spatial structure in the relationship between a response variable and explanatory variables. This section described a new spatial regression approach, called spatially clustered coefficient (SCC) method, to address the problems with the presence of spatial pattern, especially clustered pattern in the effect of covariates. Specifically, the SCC method accommodates spatial dependence through structured coefficients and employs penalized least square for estimation, where the penalty function encourages similarity in coefficients with locations connected by minimum spanning tree. Although the SCC method is designed to detect the cluster structure coefficients, it also works reasonably well in capturing a wide range of spatial patterns, as illustrated in the simulation studies. Thus, it allows researchers to explore the spatial pattern in the regression coefficient without any priori information. We establish the oracle inequalities for SCC estimator and demonstrate that its mean square error asymptotically approaches zero as the number of spatial points increases to infinity, under the infilling domain asymptotics. The SCC method is easy to implement as it can be transformed into the a Lasso (or Lasso-type) problem, so that a lot of efficient algorithms and some well-established packages in R and Matlab can be readily applied.

Both simulation studies and real data analysis indicate that the SCC method outperforms the GWR method in terms of estimation and prediction. This superior performance becomes more evident when the explanatory variables have spatial correlation. In the analysis of temperature and salinity data in the Atlantic Ocean, the SCC method produces a more reasonable and robust estimate for the T-S relationship than the GWR method and OLS. The coefficients surface estimated from the SCC method are consistent with the inference from the oceanic dynamics. In contrast, the GWR estimator is quite noisy with unrealistic large values occurring in some regions.

Through simulation studies we show that the SCC method is flexible in terms of local adaptation even for spatially smooth varying coefficients. However, a large number of piece-wise constants need to be fitted for such situations. Ideally, a penalty function that simultaneously detects the boundaries of a cluster while encouraging smoothly varying coefficients within a cluster is desired. We leave the extension of SCC to this setting to future work.

4. QUANTIFICATION OF CONTINUOUS DEPENDENCE OF SST-TURBULENT HEAT FLUX RELATIONSHIP THROUGH PENALIZED SPECTRAL REGRESSION*

4.1 Introduction

Turbulent heat flux at the air-sea interface is an important quantity in the atmosphere-ocean interactions and is closely related to the sea surface temperature (SST). An in-depth knowledge of the relationship between SST and turbulent heat flux (referred to as T-Q relationship henceforth) is essential to understand the nature of coupled atmosphere-ocean system and to improve the model representation of oceans meridional overturning circulation [43–46].

The T-Q relationship is complicated by the dual role of turbulent heat flux in the dynamics of SST [47]. At large scales ($>$1000 km), the turbulent heat flux largely contributes to the generation of SST anomalies, and the generated SST anomalies in turn can modulate the turbulent heat flux. The former leads to a negative T-Q relationship (The heat flux is defined positive upward throughout this study) while the latter tends to produce a positive T-Q relationship (also known as the negative air-sea feedback) [48]. At mesoscales (100-1000 km) [19], the SST anomalies are mainly generated through the baroclinic instability of major oceanic fronts and strongly damped by the turbulent heat flux [49]. This results in a strong positive T-Q relationship.

Despite the differed influence of turbulent heat flux on SST at different spatial scales, a continuous dependence of T-Q relationship on spatial scales has not been quantified mainly due to the lack of appropriate statistical tools. Previous studies [48, 50, 51] usually

*Reprinted from "Quantify the continuous dependence of SST-turbulent heat flux relationship on spatial scales" by Furong Li, Huiyan Sang, and Zhao Jing, 2017. *Geophysical Research Letters*, vol. 44, doi:10.1002/2017GL073695, Copyright [2017] by American Geophysical Union.

evaluated the scale dependence by spatially averaging the data to several arbitrarily chosen coarser grids and then estimating the relationship from coarse-grained data. However, such a method can only provide a crude estimate on the dependence of T-Q relationship on spatial scales. In particular, it evaluates the T-Q relationship at several pre-chosen scale ranges rather than a continuous dependence on scales.

Another deficiency of previous methods is the requirement of sufficiently long time series to get a reliable estimate of the T-Q relationship [48, 52]. This ignores the strong temporal variability in the T-Q relationship and may lead to biased estimates. Therefore, it is desirable to have a statistical model that can provide reasonable estimates even only one time record is available. Such a statistical model can not only reduce biases in the estimates but also allow analyzing the temporal variability of T-Q relationship. We note that the variability of T-Q relationship at interannual and longer time scales has not been reported in the previous literature probably due to the limitation in the existing methods.

In this study, we propose a novel statistical modeling approach, penalized spectral regression, based on state-of-the-art statistical methodologies to explore the scale dependence of the T-Q relationship and its temporal variability at seasonal and longer time periods. As demonstrated below, the penalized spectral regression (PSR) is able to capture continuous dependence of T-Q relationship on spatial scales. Furthermore, it is applicable to the case where only one time record is available. The section is organized as follows. The data and methodology are provided in Section 4.2. Section 4.3 presents the analysis of the T-Q relationship in the Kuroshio extension region by applying the PSR method to the ERA-Interim reanalysis [53] dataset. Conclusions and discussion are included in Section 4.4.

4.2 Data and methodology

4.2.1 Data

The SST and turbulent heat flux data used in this study are obtained from the ERA-Interim reanalysis [53] dataset. The data are monthly outputs spanning from January 1979 to October 2016. To analyze the scale dependence of T-Q relationship, we use the data in the Kuroshio extension region (148.5°E-171.5°W, 17.5°N-47.5°N), one of the regions with the strongest atmosphere-ocean coupling. The spatial resolution of the dataset is $0.75^\circ \times 0.75^\circ$ which is fine enough to resolve a large portion of variability at mesoscales (A recent high-resolution (0.1°) coupled modeling study by [49] reveals that most of the mesoscale variability in the Kuroshio extension region occurs around 600 km). The box size chosen here is small enough to avoid strong spatial non-stationarity in the T-Q relationship [51] but is sufficiently large to include the interested spatial scales.

4.2.2 The PSR method

A regression model capturing the continuous dependence of T-Q relationship on spatial scales is

$$\widetilde{Q}^t_{k,\ell} = (\alpha_{k,\ell} + i\beta_{k,\ell})\widetilde{T}^t_{k,\ell} + (c_{k,\ell} + id_{k,\ell}) + \epsilon^t_{k,\ell}, \tag{4.1}$$

where $\widetilde{Q}^t_{k,\ell}$ and $\widetilde{T}^t_{k,\ell}$ are the 2-D Fourier transform of turbulent heat flux and SST anomalies, k and ℓ are the wavenumber components in the zonal and meridional directions, t is the index for time series, $\alpha_{k,\ell} + i\beta_{k,\ell}$ and $c_{k,\ell} + id_{k,\ell}$ are the complex regression coefficients and intercepts to be estimated, and $\epsilon^t_{k,\ell}$ is the complex independently identically distributed random noise. The real part of regression coefficients $\alpha_{k,\ell}$ characterizes the continuous dependence of T-Q relationship on wavenumbers.

Canonically, $\alpha_{k,\ell} + i\beta_{k,\ell}$ and $c_{k,\ell} + id_{k,\ell}$ can be estimated by solving the ordinary least square problem:

$$min\{\sum_t \sum_{k,\ell} |\widetilde{Q}^t_{k,\ell} - (\alpha_{k,\ell} + i\beta_{k,\ell})\widetilde{T}^t_{k,\ell} - (c_{k,\ell} + id_{k,\ell})|\}. \tag{4.2}$$

In Eq. (4.2), $\alpha_{k,\ell} + i\beta_{k,\ell}$ and $c_{k,\ell} + id_{k,\ell}$ are estimated pointwisely in the wavenumber space. It requires a relatively large number of samples (a sufficiently long time series in our case) to get a reliable estimate. In particular, $\alpha_{k,\ell}+i\beta_{k,\ell}$ and $c_{k,\ell}+id_{k,\ell}$ are undetermined with only one time record available. To overcome this shortage, we propose a penalized spectral regression (PSR) method that allows borrowing information across wavenumbers to estimate $\alpha_{k,\ell} + i\beta_{k,\ell}$ and $c_{k,\ell} + id_{k,\ell}$. Specifically, $\alpha_{k,\ell} + i\beta_{k,\ell}$ and $c_{k,\ell} + id_{k,\ell}$ are estimated by minimizing the following objective function,

$$\begin{aligned} &\sum_t \sum_{k,\ell} |\widetilde{Q}^t_{k,\ell} - (\alpha_{k,\ell} + i\beta_{k,\ell})\widetilde{T}^t_{k,\ell} - (c_{k,\ell} + id_{k,\ell})| \\ &+ \lambda \sum_{k,\ell} P(\alpha_{k,\ell}) + P(\beta_{k,\ell}) + P(c_{k,\ell}) + P(d_{k,\ell}), \end{aligned} \tag{4.3}$$

where λ is a tuning parameter determining the amount of penalization, the penalty function $P(\theta), \theta = \alpha, \beta, c, d$ is defined as

$$P(\theta_{k,\ell}) = |\theta_{k,\ell} - \frac{1}{4}(\theta_{k+\Delta k,\ell} + \theta_{k-\Delta k,\ell} + \theta_{k,\ell+\Delta\ell} + \theta_{k,\ell-\Delta\ell})|. \tag{4.4}$$

Eq. 4.4 belongs to generalized Lasso penalty functions [25]. It penalizes the difference of $\alpha_{k,\ell}+i\beta_{k,\ell}$ ($c_{k,\ell}+id_{k,\ell}$) at some wavenumber from its values at neighboring wavenumbers and thus encourages the similarity of $\alpha_{k,\ell} + i\beta_{k,\ell}$ ($c_{k,\ell} + id_{k,\ell}$) between adjacent wavenumbers. When λ approaches infinity, Eq. 4.3 will result in constant T-Q relationship in the wavenumber space. For $\lambda = 0$, Eq. 4.3 reduces to Eq. 4.2. In practice, the optimal value of λ can be determined from the Bayesian information criterion (BIC) or cross validation.

Finally, we remark that more advanced penalty functions, such as the smoothly clipped absolute deviation (SCAD) penalty [28] and minimax concave penalty (MCP) [29] are also applicable but at the expense of increased computational burden.

The rationale of adopting the penalty functions Eq. 4.4 is rooted in the belief that the T-Q relationship should have a well-organized structure rather than fluctuate randomly in the wavenumber space. From a dynamical point of view, there are no foreseen reasons why such a belief is unreasonable. As Eq. 4.2 fails to utilize such prior information, it requires a sufficiently long time series to get a reliable estimate and thus is an inefficient method to estimate the T-Q relationship. In contrast, the PSR method borrows neighboring information to accommodate structured T-Q relationship in the wavenumber space. As illustrated in Section 4.3, it can provide reasonable estimates even when only one time record is available.

4.2.3 Computation of the PSR

The monthly data are temporally averaged to construct seasonal time series of SST and turbulent heat flux. To compute the SST and turbulent heat flux anomalies, we remove a linear fit along each longitude from the SST and heat flux map, followed by subtracting a linear fit along each latitude (See Figure 4.1a) for an instance of resultant SST and turbulent heat flux anomalies). Given the domain size used in this study, such defined anomalous fields are isolated from the basin-scale background and at the same time largely preserve the variability from the smallest resolved spatial scale (150 km) to 8000 km, making it suitable to analyze the T-Q relationship on a wide range of interested spatial scales. Finally, a 2-D fast Fourier transform (FFT) is applied to the SST and heat flux anomalies to evaluate $\widetilde{Q}^t_{k,\ell}$ and $\widetilde{T}^t_{k,\ell}$. As the linear fits make the Fourier components at the zero wavenumber, i.e., $\widetilde{Q}^t_{0,0}$ and $\widetilde{T}^t_{0,0}$, become zero, $\alpha_{0,0} + i\beta_{0,0}$ is poorly constrained. To avoid unrealistic large values for $\alpha_{0,0} + i\beta_{0,0}$ and its contamination on neighboring wavenumbers, we replace the penalty terms $\lambda(P(\alpha_{0,0})+P(\beta_{0,0}))$ in Eq. 4.3 by $\lambda|\alpha_{0,0}+\beta_{0,0}|$ to constrain the value of $\alpha_{0,0}+i\beta_{0,0}$. Such a penalty term is also reasonable in a dynamical sense as previous studies suggest that the T-Q relationship should be much weaker at basin scales than at mesoscales [48, 50].

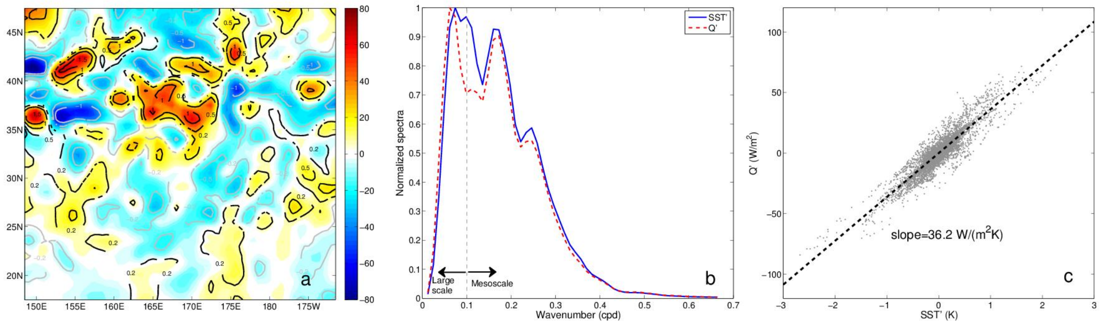

Figure 4.1: A A case study of SST and turbulent heat flux anomalies. (a) The spatial distribution of SST anomalies in K (contours) and turbulent heat flux anomalies in W/m^2 (color) in DJF between 2011 and 2012. (b) The wavenumber spectra of SST anomalies (blue solid) and turbulent heat flux anomalies (red dashed). Each spectrum has been normalized by its maximum. (c) The scatter plot of SST anomalies v.s. turbulent heat flux anomalies. The black dashed line denotes the pattern regression.

With the values of $\widetilde{Q}^t_{k,\ell}$ and $\widetilde{T}^t_{k,\ell}$ available, Eq. 4.3 can be solved for any length of time series using the coordinate descent algorithm [38]. To fully assess the temporal variability of the T-Q relationship and also to illustrate the power of the PSR method, we will solve Eq. 4.3 for each season of individual years, in which case there is only one time record available for each estimation.

4.3 Results

4.3.1 The continuous dependence of T-Q relationship on spatial scales

We first perform a case study in the winter season (DJF) between 2011 and 2012 to illustrate the performance of PSR method. Figure 4.1a and b display the SST and turbulent heat flux anomalies and their wavenumber spectra, respectively. Both the SST and turbulent heat flux anomalies exhibit a full spectrum of variability with dominant variability occurring within 4°-30° (1° corresponds to about 100 km in the studied domain). There is a positive pattern correlation between SST and turbulent heat flux anomalies (Figure 4.1c). Their pattern regression coefficient is 36.2 W/(m^{2}K).

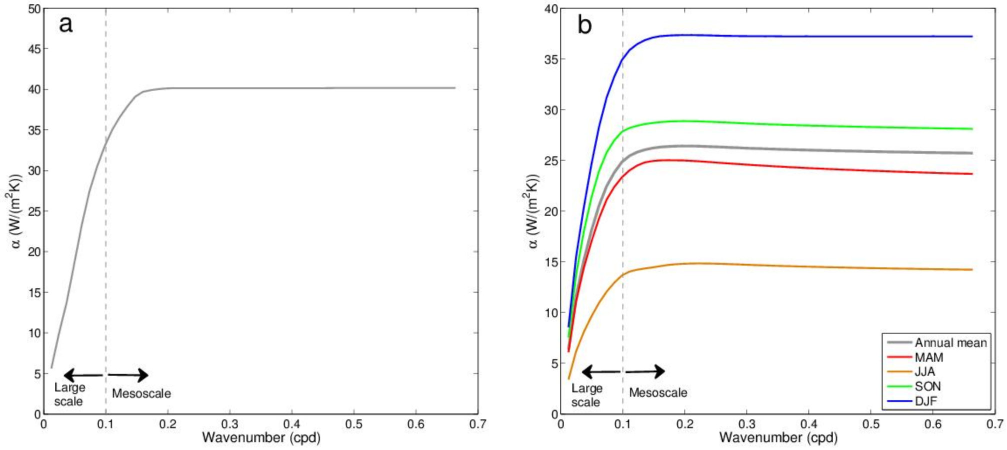

Figure 4.2: The T-Q relationship. (a) The continuous dependence of T-Q relationship on spatial scales in the case study of DJF between 2011 and 2012. (b) Same as (a) but for the climatological T-Q relationship in each season.

Figure 4.2a displays the continuous dependence of T-Q relationship on the wavenumber ($k_H = \sqrt{k^2 + l^2}$) derived from the PSR method. It provides much richer information than a single regression coefficient derived from the pattern regression. The value of α averaged over all the wavenumbers is 36.3 W/(m^{2}K), compatible to 36.2 W/(m^{2}K) derived from the pattern regression. However, as revealed by Figure 4.2a, the T-Q relationship varies significantly with wavenumber. The value of α stays relatively stable at mesoscales. It varies by less than 20% for k_H ranging from 0.1 cycle per degree (cpd) to the largest resolved wavenumber (0.67 cpd). The mean value of α within 0.1-0.67 cpd is 39.9 W/(m^{2}K), larger than that derived from the pattern regression which essentially merges the T-Q relationship at all the spatial scales. Then the value of α decreases rapidly and monotonically as the scales become larger. It ends at 5.6 W/(m^2K)) at the smallest resolved wavenumber (0.0125 cpd). Such a scale-dependence of α is supported by the SST dynamics in the Kuroshio extension region. At mesoscales, the SST anomalies are mainly generated by the baroclinic instability of Kuroshio extension jet and are strongly damped by the turbulent heat flux once generated [49]. Such a scenario is consistent with a large positive value for α at mesoscales. At large scales, the turbulent heat flux plays a dual role in the dynamics of SST anomalies. On one hand, it largely contributes to the generation of SST anomalies, contributing to a negative value of α. On the other hand, the SST anomalies are subjected to the damping by the turbulent heat flux after their generation but this damping effect is expected to be weaker than that at mesoscales [54, 55]. Combining these two effects will result in a rapid decrease of magnitude of α as the spatial scales migrate from mesoscales to large scales.

The scale-dependence of the SST-heat flux relationship averaged over 38 years is qualitatively similar to the relationship obtained in the case study (Figure 4.2b). The climatological mean value of α varies by less than 4% at mesoscales and then decreases rapidly as the spatial scales become larger. However, the climatological mean value of α at mesoscales is only about 26.0 W/(m^{2}K)), significantly smaller than 39.9 W/(m^{2}K) in the case study. As demonstrated in the next subsection, such a difference is mainly due to the pronounced seasonality of α at mesoscales.

4.3.2 Low-frequency variability of scale-dependent T-Q relationship

In this subsection, we examine the variability of scale-dependent T-Q relationship at seasonal and longer time periods. Figure 4.2b displays the value of α in different seasons averaged over 38 years. In all the seasons, the dependence of α on spatial scales is qualitatively similar to that in the case study above, i.e., the value of α is relatively stable at mesoscales and then decreases rapidly as the spatial scales further increase. Despite these similarities, α exhibits an evident seasonal cycle. Its phase is coherent at all the spatial scales but its amplitude is more pronounced at mesoscales than at large scales. In winter (DJF), the mean value of α at mesoscales reaches up to 37.2 W/(m^{2}K), more than twice the value (14.5 W/(m^{2}K)) in summer (JJA). The strong damping of mesoscale SST anomalies in winter is mainly due to the cold dry air masses coming from the Asian continent and high wind speed associated with the energetic winter storms, which significantly enhances the air-sea coupling [56].

The seasonality of the T-Q relationship in the Kuroshio extension region has also been studied in previous literature but without taking the scale dependence into consideration [48, 57]. These studies reported that the damping of SST anomalies by turbulent heat flux in the Kuroshio extension region is stronger in winter and autumn than in spring and summer, which is consistent with our results derived from the PSR method. However, the results in this study clearly show that the phase of seasonality of T-Q relationship is coherent at all the spatial scales.

Thus far the interannual and decadal variabilities of the T-Q relationship in the Kuroshio extension region remain poorly understood. However, such variabilities could play an important role in modulating the strength of Kuroshio extension jet at corresponding time scales. As demonstrated in a recent study by [49], the Kuroshio extension jet becomes weaker and wider when the damping of mesoscale SST anomalies by turbulent heat flux is suppressed. In this subsection, we analyze the interannual and decadal variabilities of the T-Q relationship at mesoscales derived from the PSR method. As the value of α does not change significantly at mesoscales (Figure 4.2), we use its mean value averaged between 0.1-0.67 cpd (denoted as α_{meso} hereinafter) to characterize the T-Q relationship at mesoscales to facilitate analysis.

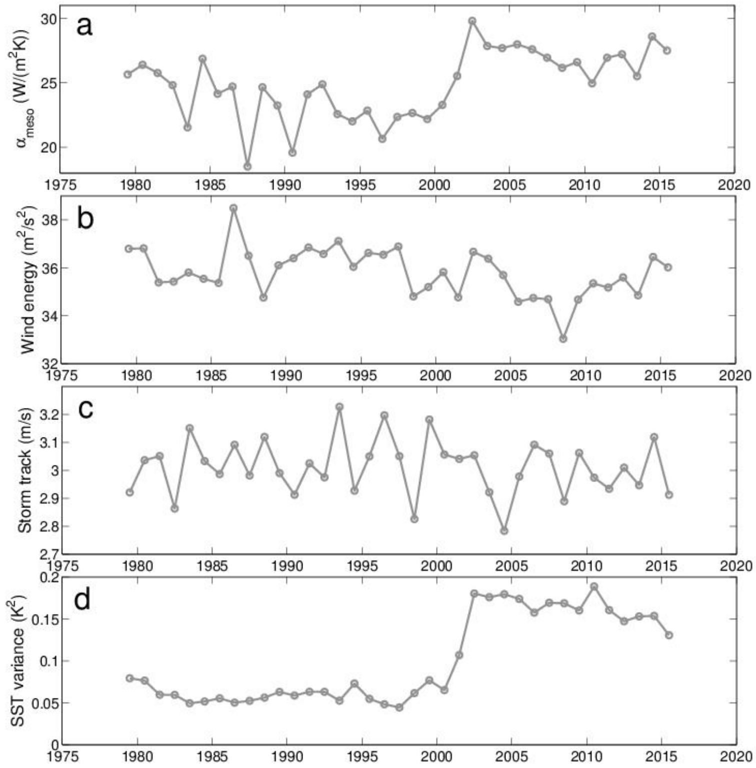

Figure 4.3: Time series of annual mean for (a) α_{meso} , (b) wind energy, (c) storm track intensity, and (d) variance of mesoscale SST anomalies. Here the wind energy is computed as $(u_{10}^2 + v_{10}^2)/2$ where u_{10} and v_{10} are daily mean 10-m wind velocity derived from the ERA-Interim reanalysis. The storm track intensity is evaluated as the standard deviation of 2-8 day band-pass filtered v_{10}.

Figure 4.3a displays the annual mean time series of α_{meso} from 1979 to 2016. The most notable feature is that the value of α_{meso} during 1979-2001 is systematically smaller than that during 2002-2016. This difference is due to an abrupt increase of α_{meso} from 2001 to 2002. Despite the key role of wind speed in the T-Q relationship [56], there is no evidence for the enhancement of wind energy or storm track intensity at the same time period (Figure 4.3b and c). Instead, the abrupt increase of α_{meso} from 2001 to 2002 coincides remarkably well with the abrupt increase of variance of mesoscale SST anomalies (Figure 4.3d). As such, the correlation coefficient between α_{meso} and wind energy (storm track intensity) is only about -0.25 (-0.17) (not statistically significant at 5% significance level) but increases to 0.73 (statistically significant at 5% significance level) between α_{meso} and mesoscale SST anomaly variance. It should be noted that the abruptly elevated mesoscale SST anomaly variance is not a realistic feature but simply results from the change of prescribed lower boundary conditions of ERA-Interim simulation from low-resolution (1°) SST data products to higher-resolution ($< 0.5^\circ$) products [53]. Therefore, we suspect the abrupt increase of α_{meso} from 2001 to 2002 may be an artifact of numerical model configurations. Furthermore, it also suggests that the T-Q relationship at mesoscales might not be well represented when low-resolution SST is used to force the atmosphere general circulation models. The climatological mean α_{meso} derived from the ERA-Interim reanalysis during 1979-2016 may bias low.

4.4 Conclusion and discussion

In this section, we developed a novel statistical model, penalized spectral regression (PSR), to evaluate the continuous dependence of SST- turbulent heat flux relationship (T-Q relationship) on spatial scales. By penalizing the difference of T-Q relationship at adjacent wavenumbers to reflect the belief that the T-Q relationship should be well organized in the wavenumber space rather than fluctuate randomly, the PSR model is able to provide reasonable estimates even when only one time record is available. Application of PSR model to the ERA-Interim reanalysis in the Kuroshio extension region reveals pronounced variation of T-Q relationship with spatial scales. The regression coefficient α stays stable at mesoscales (< 1000 km) with a climatological mean value of 26 W/(m^2K). Then its value decreases rapidly as the spatial scales further increase but is always positive. There is a pronounced seasonal cycle in α with its phase coherent at all the resolved spatial scales (150-8000 km). The largest and smallest values occur in winter and summer, respectively. In addition, the value of α during 1979-2001 is systemically smaller than that during 2002-2016 due to an abrupt increase of its value from 2001 to 2002. However, we suspect that the abrupt increase is not a realistic feature but probably due to the use of high-resolution SST datasets as lower boundary conditions in ERA-Interim reanalysis since January 2002.

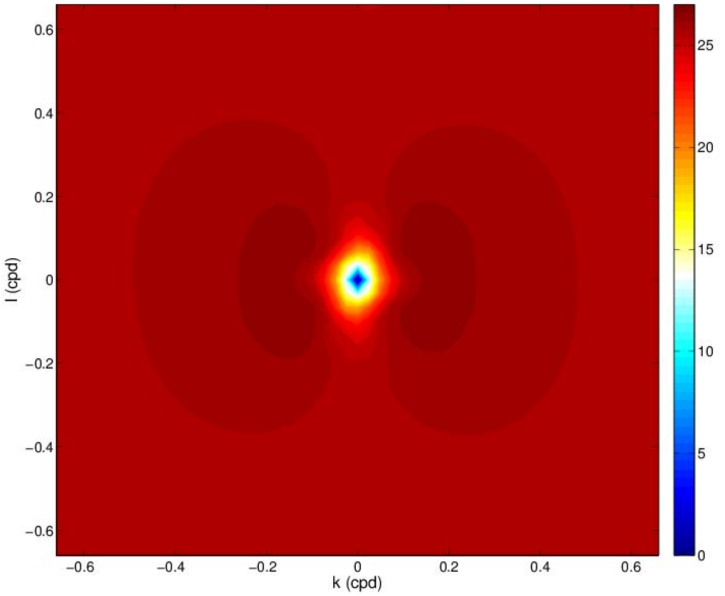

Figure 4.4: The climatological mean α in W/(m^{2}K) as a function of k and ℓ.

The PSR method is not only able to evaluate the continuous dependence of T-Q relationship on spatial scales but also its continuous dependence on azimuth. Figure 4.4

displays the climatological mean value of α as a function of k and ℓ. It seems that the T-Q relationship is basically isotropic although the value of α tends to be slightly stronger for east-west directed SST gradient than north-south directed SST gradient at spatial scales between 250-1000 km. Whether such a slight difference ($< 5\%$) has a clear dynamical interpretation remains unknown but deserves further investigation in future studies. Furthermore, the PSR method can be readily extended to analyze the continuous dependence of T-Q relationship on geographic location and time. These extensions can be done by modifying the penalty functions to penalize the difference of regression coefficients at adjacent locations and time. We conclude that the PSR model provides a feasible tool to analyze the relationship between various quantities (e.g., SST v.s. sea surface height, SST v.s. precipitation, and SST v.s. wind speed) in geophysics.

5. SUMMARY

Statistical Inference for large spatial data is studied in this dissertation from three aspects. We first address the problem of estimating covariance parameters for large and irregularly spaced dataset. Then a new modeling approach, called spatially clustered coefficient (SCC) regression, is proposed to explore the spatially-varying relationship between covariates and the response variable of interest. Finally, a penalized spectral regression (PSR) model is constructed by extending the idea of SCC model to the spectral domain. The major progresses of this dissertation are summarized as follows:

In Section 2, a new weighting scheme is proposed to construct a composite likelihood (CL) for the inference of spatial Gaussian process models. This weight function is an approximation to the optimal weight derived from the theory of optimal estimating equations. It is calculated with the strategy of combing block-diagonal approximation and tapering. Gains in statistical and computational efficiency over existing CL methods are illustrated through simulation studies and applications to the rainfall data.

In Section 3, we propose a new modelling approach, called spatially clustered coefficient (SCC) regression, to capture the spatial structure, especially clustered structure in the relationship between response variable and explanatory variables. It is demonstrated based on simulation studies that the SCC method works very effectively in estimation for data either with clustered coefficients or smoothly-varying coefficients. Thus, it is a feasible and power tool to explore the spatial structure in the regression coefficient without any priori information. Some oracle inequalities are derived, providing non-asymptotic error bounds on estimators and predictors. Finally, the SCC method is applied to analyzing the temperature-salinity relationship in the Atlantic basin and shows good performance.

In Section 4, we extend the idea of SSC method to the spectral domain and construct a penalized spectral regression (PSR) model. The PSR method is applied to quantifying the wavenumber-dependent relationship between sea surface temperature (SST) and turbulent heat flux (T-Q relationship). The T-Q relationship derived from the PSR method is consistent with geophysical dynamics, lending support to its good performance. Moreover, some new features in the T-Q relationship are disclosed by the PSR method and are likely to raise broad interest in the geoscience community.

REFERENCES

[1] M. Bevilacqua and C. Gaetan, "Comparing composite likelihood methods based on pairs for spatial Gaussian random fields," *Statistics and Computing*, vol. 25, no. 5, pp. 877–892, 2015.

[2] J. Eidsvik, B. A. Shaby, B. J. Reich, M. Wheeler, and J. Niemi, "Estimation and prediction in spatial models with block composite likelihoods," *Journal of Computational and Graphical Statistics*, vol. 23, no. 2, pp. 295–315, 2014.

[3] F. C. Curriero and S. Lele, "A composite likelihood approach to semivariogram estimation," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 4, no. 1, pp. 9–28, 1999.

[4] C. Varin, G. Høst, and Ø. Skare, "Pairwise likelihood inference in spatial generalized linear mixed models," *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 1173–1191, 2005.

[5] S. A. Padoan, M. Ribatet, and S. A. Sisson, "Likelihood-based inference for max-stable processes," *Journal of the American Statistical Association*, vol. 105, no. 489, pp. 263–277, 2010.

[6] B. G. Lindsay, "Composite likelihood methods," *Contemporary Mathematics*, vol. 80, no. 1, pp. 221–39, 1988.

[7] T. V. Apanasovich, D. Ruppert, J. R. Lupton, N. Popovic, N. D. Turner, R. S. Chapkin, and R. J. Carroll, "Aberrant crypt foci and semiparametric modeling of correlated binary data," *Biometrics*, vol. 64, no. 2, pp. 490–500, 2008.

[8] M. Bevilacqua, C. Gaetan, J. Mateu, and E. Porcu, "Estimating space and space-time covariance functions for large data sets: a weighted composite likelihood approach,"

Journal of the American Statistical Association, vol. 107, no. 497, pp. 268–280, 2012.

[9] H. Sang and M. G. Genton, "Tapered composite likelihood for spatial max-stable models," *Spatial Statistics*, vol. 8, pp. 86–103, 2014.

[10] Y. Bai, P. X.-K. Song, and T. Raghunathan, "Joint composite estimating functions in spatiotemporal models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 74, no. 5, pp. 799–824, 2012.

[11] C. G. Kaufman, M. J. Schervish, and D. W. Nychka, "Covariance tapering for likelihood-based estimation in large spatial data sets," *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1545–1555, 2008.

[12] D. R. Cox and N. Reid, "A note on pseudolikelihood constructed from marginal densities," *Biometrika*, vol. 91, no. 3, pp. 729–737, 2004.

[13] V. P. Godambe, "An optimum property of regular maximum likelihood estimation," *The Annals of Mathematical Statistics*, vol. 31, no. 4, pp. 1208–1211, 1960.

[14] B. G. Lindsay, G. Y. Yi, and J. Sun, "Issues and strategies in the selection of composite likelihoods," *Statistica Sinica*, vol. 21, no. 1, pp. 71–105, 2011.

[15] C. C. Heyde, *Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter estimation*. New York: Springer Science & Business Media, 2008.

[16] D. R. Cox and N. Reid, "Parameter orthogonality and approximate conditional inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 49, no. 1, pp. 1–39, 1987.

[17] H. Zhang, "Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics," *Journal of the American Statistical Association*, vol. 99, no. 465, pp. 250–261, 2004.

[18] Y. D. Lee and S. N. Lahiri, "Least squares variogram fitting by spatial subsampling," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 64, no. 4, pp. 837–854, 2002.

[19] L. D. Talley, *Descriptive Physical Oceanography: An Introduction*. London: Academic Press, 2011.

[20] G. K. Vallis, *Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-scale Circulation*. Cambridge: Cambridge University Press, 2006.

[21] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression*. Chichester: John Wiley & Sons, Ltd, 2003.

[22] A. E. Gelfand, H.-J. Kim, C. Sirmans, and S. Banerjee, "Spatial modeling with spatially varying coefficient processes," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 387–396, 2003.

[23] A. O. Finley, "Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence," *Methods in Ecology and Evolution*, vol. 2, no. 2, pp. 143–154, 2011.

[24] D. C. Wheeler and C. A. Calder, "An assessment of coefficient accuracy in linear regression models with spatially varying coefficients," *Journal of Geographical Systems*, vol. 9, no. 2, pp. 145–166, 2007.

[25] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 1, pp. 91–108, 2005.

[26] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[27] H. Zou, “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[28] J. Fan and R. Li, “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[29] C.-H. Zhang, “Nearly unbiased variable selection under minimax concave penalty,” *The Annals of Statistics*, vol. 38, no. 2, pp. 894–942, 2010.

[30] Q. Song and F. Liang, “High-dimensional variable selection with reciprocal L_1-regularization,” *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1607–1620, 2015.

[31] X. Shen and H.-C. Huang, “Grouping pursuit through a regularization solution surface,” *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 727–739, 2012.

[32] Z. T. Ke, J. Fan, and Y. Wu, “Homogeneity pursuit,” *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 175–194, 2015.

[33] G. H. Golub, M. Heath, and G. Wahba, “Generalized cross-validation as a method for choosing a good ridge parameter,” *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979.

[34] G. Schwarz *et al.*, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[35] J. Chen and Z. Chen, “Extended Bayesian information criteria for model selection with large model spaces,” *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[36] J. Chen and Z. Chen, “Extended BIC for small-n-large-p sparse GLM,” *Statistica Sinica*, vol. 22, no. 2, pp. 555–574, 2012.

[37] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, *et al.*, “Least angle regression,” *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[38] J. Friedman, T. Hastie, and R. Tibshirani, “Regularization paths for generalized linear models via coordinate descent,” *Journal of Statistical Software*, vol. 33, no. 1, p. 1, 2010.

[39] Y. Kim, H. Choi, and H.-S. Oh, “Smoothly clipped absolute deviation on high dimensions,” *Journal of the American Statistical Association*, vol. 103, no. 484, pp. 1665–1673, 2008.

[40] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, “Simultaneous analysis of lasso and Dantzig selector,” *The Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.

[41] S. A. Van De Geer, P. Bühlmann, *et al.*, “On the conditions used to prove oracle results for the lasso,” *Electronic Journal of Statistics*, vol. 3, pp. 1360–1392, 2009.

[42] J. Guo, J. Hu, B.-Y. Jing, and Z. Zhang, “Spline-lasso in high-dimensional linear regression,” *Journal of the American Statistical Association*, vol. 111, no. 513, pp. 288–297, 2016.

[43] S. Zhang, R. J. Greatbatch, and C. A. Lin, “A reexamination of the polar halocline catastrophe and implications for coupled ocean-atmosphere modeling,” *Journal of Physical Oceanography*, vol. 23, no. 2, pp. 287–299, 1993.

[44] F. Yin and E. Sarachik, “Interdecadal thermohaline oscillations in a sector ocean general circulation model: advective and convective processes,” *Journal of Physical Oceanography*, vol. 25, no. 11, pp. 2465–2484, 1995.

[45] F. Chen and M. Ghil, “Interdecadal variability of the thermohaline circulation and high-latitude surface fluxes,” *Journal of Physical Oceanography*, vol. 25, no. 11, pp. 2547–2568, 1995.

[46] S. Rahmstorf and J. Willebrand, “The role of temperature feedback in stabilizing the thermohaline circulation,” *Journal of Physical Oceanography*, vol. 25, no. 5, pp. 787–805, 1995.

[47] C. Frankignoul, A. Czaja, and B. LHeveder, “Air–sea feedback in the North Atlantic and surface boundary conditions for ocean models,” *Journal of Climate*, vol. 11, no. 9, pp. 2310–2324, 1998.

[48] C. Frankignoul and E. Kestenare, “The surface heat flux feedback. Part i: Estimates from observations in the Atlantic and the North Pacific,” *Climate Dynamics*, vol. 19, no. 8, pp. 633–647, 2002.

[49] X. Ma, Z. Jing, P. Chang, X. Liu, R. Montuoro, R. J. Small, F. O. Bryan, R. J. Greatbatch, P. Brandt, D. Wu, *et al.*, “Western boundary currents regulated by interaction between ocean eddies and the atmosphere,” *Nature*, vol. 535, no. 7613, pp. 533–537, 2016.

[50] U. Hausmann, A. Czaja, and J. Marshall, “Estimates of air–sea feedbacks on sea surface temperature anomalies in the Southern Ocean,” *Journal of Climate*, vol. 29, no. 2, pp. 439–454, 2016.

[51] U. Hausmann, A. Czaja, and J. Marshall, “Mechanisms controlling the SST air-sea heat flux feedback and its dependence on spatial scale,” *Climate Dynamics*, vol. 48, no. 3, pp. 1297–1307, 2017.

[52] R. Wu, B. P. Kirtman, and K. Pegion, “Surface latent heat flux and its relationship with sea surface temperature in the National Centers for Environmental Prediction Climate Forecast System simulations and retrospective forecasts,” *Geophysical Research Letters*, vol. 34, no. 17, 2007.

[53] D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, *et al.*, “The ERA-Interim reanalysis: configuration and performance of the data assimilation system,” *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 656, pp. 553–597, 2011.

[54] F. P. Bretherton, “Ocean climate modeling,” *Progress in Oceanography*, vol. 11, no. 2, pp. 93–129, 1982.

[55] C. Frankignoul, “Sea surface temperature anomalies, planetary waves, and air-sea feedback in the middle latitudes,” *Reviews of Geophysics*, vol. 23, no. 4, pp. 357–390, 1985.

[56] C. W. Fairall, E. F. Bradley, D. P. Rogers, J. B. Edson, and G. S. Young, “Bulk parameterization of air-sea fluxes for tropical ocean-global atmosphere coupled-ocean atmosphere response experiment,” *Journal of Geophysical Research: Oceans*, vol. 101, no. C2, pp. 3747–3764, 1996.

[57] S. Park, C. Deser, and M. A. Alexander, “Estimation of the surface heat flux response to sea surface temperature anomalies over the global oceans,” *Journal of Climate*, vol. 18, no. 21, pp. 4582–4599, 2005.

APPENDIX A

PROOF OF THEOREM 1

To prove Theorem 1, we first derive the following two lemmas, and then prove the oracle inequalities using the lemmas.

Lemma 1. *Define* $\Lambda_n = \{\max_{t=1,\ldots np} |V_t| \leqslant \lambda_n/4\}$ *where* $V_t = n^{-1}\sum_{i=1}^{n} \widetilde{X}_{i,t}\varepsilon_i$. *Then*

$$P(\Lambda_n) \geqslant 1 - 2p \cdot n^{-C_2}. \tag{A.1}$$

Lemma 2. *On the event* Λ_n*, we have*

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 + \frac{\lambda_n}{2}\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1 \leq r_n\|\boldsymbol{\theta}_A - \widehat{\boldsymbol{\theta}}_A\|_2, \tag{A.2}$$

where $r_n = 2\lambda_n\sqrt{|A|}$.

Proof. According to Assumption 1a, V_t is a sub-Gaussian random variable with a zero mean and a sub-Gaussian parameter $C_1\sigma/\sqrt{n}$. Using the upper and lower deviation inequalities, we have:

$$P(|V_t| \leq \lambda_n/4) \geq 1 - 2\exp(-\frac{\lambda_n^2/16}{2C_1^2\sigma^2/n}) \geq 1 - 2n^{-(1+C_2)},$$

and

$$P(\max_{t=1,\ldots np} |V_t| \leq \lambda_n/4) \geq (1 - 2n^{-(1+C_2)})^{np} \geq 1 - 2p \cdot n^{-C_2}.$$

This proves Lemma 1. □

Proof. As the estimator is the minimizer of penalized least square, we have

$$\frac{1}{n}\|\mathbf{Y}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2+\lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1 \leq \frac{1}{n}\|\mathbf{Y}-\widetilde{\mathbf{X}}\boldsymbol{\theta}\|_2^2+\lambda_n\|\boldsymbol{\theta}_B\|_1.$$

After some manipulations, we have:

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \leq \lambda_n\|\boldsymbol{\theta}_B\|_1-\lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1+\frac{2}{n}\boldsymbol{\epsilon}^{\mathsf{T}}\widetilde{\mathbf{X}}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}).$$

Then on the event Λ_n, we have

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \leq \lambda_n\|\boldsymbol{\theta}_B\|_1-\lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1+\frac{\lambda_n}{2}\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_1.$$

Therefore, we have

$$\begin{aligned}\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2+\frac{\lambda_n}{2}\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_1 &\leq \lambda_n\|\boldsymbol{\theta}_B\|_1-\lambda_n\|\widehat{\boldsymbol{\theta}}_B\|_1+\lambda_n\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_1\\ &\leq 2\lambda_n\|\boldsymbol{\theta}_A-\widehat{\boldsymbol{\theta}}_A\|_1 \qquad\qquad (A.3)\\ &\leq 2\lambda_n\sqrt{|A|}\|\boldsymbol{\theta}_A-\widehat{\boldsymbol{\theta}}_A\|_2.\end{aligned}$$

This proves Lemma 2. □

Now we prove the oracle inequalities. According to Lemma 2, we have:

$$\|\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}\|_1 \leq 4\sqrt{|A|}\|\boldsymbol{\theta}_A-\widehat{\boldsymbol{\theta}}_A\|_2. \qquad (A.4)$$

Then $\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}$ satisfies Assumption 1b. Therefore, we have

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta}-\widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \geq \Phi\|\boldsymbol{\theta}_A-\widehat{\boldsymbol{\theta}}_A\|_2^2. \qquad (A.5)$$

Combining (A.3) and (A.5) yields

$$\frac{1}{n}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2^2 \leq \frac{r_n}{\sqrt{n\Phi}}\|\widetilde{\mathbf{X}}\boldsymbol{\theta} - \widetilde{\mathbf{X}}\widehat{\boldsymbol{\theta}}\|_2. \tag{A.6}$$

This directs leads to inequality (3.9). Based on inequality (3.9) and (A.5), we have

$$\Phi\|\boldsymbol{\theta}_A - \widehat{\boldsymbol{\theta}}_A\|_2^2 \leq \frac{4\lambda_n^2|A|}{\Phi}. \tag{A.7}$$

Combining (A.3) and (A.7) directly leads to inequality (3.10).

Finally, we derive the oracle inequality for $\boldsymbol{\beta}$. According to inequality (3.10), we have

$$\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2 \leq \|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_1 \leq \frac{8\lambda_n|A|}{\Phi}. \tag{A.8}$$

Let r_{min} denote the smallest eigenvalue of $\widetilde{\boldsymbol{T}}^{\mathsf{T}}\widetilde{\boldsymbol{T}}$. Then r_{min} is positive as $\widetilde{\boldsymbol{T}}^{\mathsf{T}}\widetilde{\boldsymbol{T}}$ is positive definite. Therefore, we have

$$\|\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}\|_2^2 \geq r_{min}\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2^2. \tag{A.9}$$

Combining (A.8) and (A.9) yields

$$\|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|_2 \leq \frac{8\lambda_n|A|}{\Phi\sqrt{r_{min}}}.$$

APPENDIX B

THE DIAGRAM FOR MINIMUM SPANNING TREE

The following schematic diagram illustrates the rationale of minimum spanning tree in the SCC method .

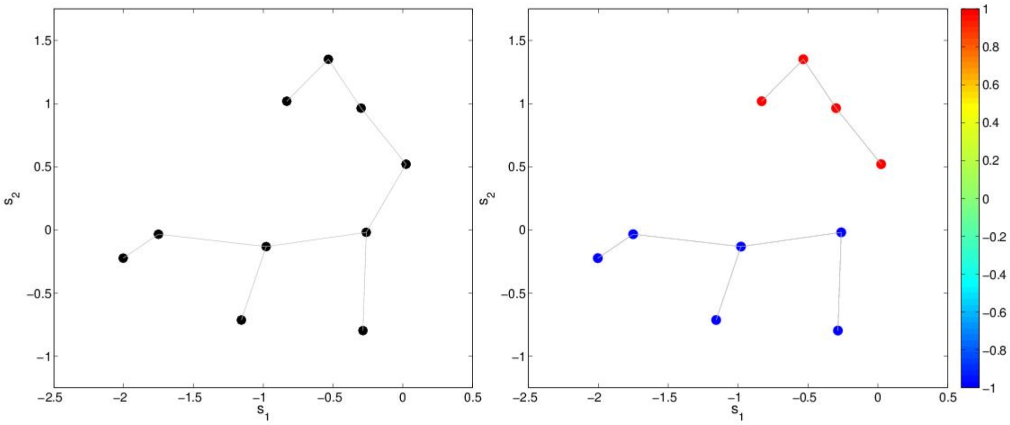

Figure B.1: A schematic diagram for minimum spanning tree (MST). The left panel depicts the MST for 10 spatial location. The right panel displays the estimated coefficient after constructing the MST.