ABSTRACT

Title of Dissertation: PARAMETRIC ESTIMATION IN SPATIAL REGRESSION MODELS

Nathan Yu
Doctor of Philosophy, 2022

Dissertation Directed by: Professor Eric Slud
Department of Mathematics

This dissertation addresses the asymptotic theory behind parametric estimation in spatial regression models. In spatial statistics, there are two prominent types of asymptotic frameworks: increasing domain asymptotics and infill asymptotics. The former assumes that spatial data are observed over a region that increases with the sample size, whereas the latter assumes the observations become increasingly dense in a bounded domain. It is well understood that both frameworks lead to drastically different behavior of classical statistical estimators. Under increasing domain asymptotics, we use recently established limit theorems for random fields to prove consistency and asymptotic normality of estimators in a nonlinear regression model. The theory presented here hinges on a crucial assumption that the covariates and error are independent of one another. However, when covariates also exhibit spatial variation, this assumption of independence becomes questionable. This possibility of spatial correlation between the covariates and error is known as spatial confounding. We examine several possible parametric models of spatial confounding and

under increasing domain asymptotics, we determine that the degree of confounding can be estimated with good precision through maximum likelihood methods. Finally, under infill asymptotics, we focus our attention on linear regression models in a Gaussian setting. Existing literature in infill asymptotics tends to ignore estimation of the mean and emphasizes estimation of variance components in the error. For estimation of the mean, the sample path properties of the mean relative to the error play an important role. We show that when the sample paths of the covariates are sufficiently rough, it is possible to obtain consistent, asymptotically normal estimates of regression parameters through maximum likelihood estimation.

PARAMETRIC ESTIMATION IN SPATIAL REGRESSION MODELS

by

Nathan Yu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2022

Advisory Committee:
Professor Eric Slud, Chair/Advisor
Professor Paul Smith
Professor Yu Gu
Professor Maria Cameron
Professor Ingmar Prucha, Dean's Representative

Acknowledgments

The completion of this thesis would not have been possible without the support of my advisor Professor Eric Slud. I am very grateful for his time, patience and commitment throughout the years directing this dissertation. My research topic was not in his area of expertise, yet he still provided invaluable insight, ideas and knowledge. It has been a privilege learning from him and working under his guidance.

Professors Paul Smith, Yu Gu, Maria Cameron and Ingmar Prucha all deserve my deepest gratitude for serving on my committee and providing constructive feedback and comments. A very special thanks to Professor Cameron for agreeing to be on my committee on emergency notice.

I would like to extend another thank you to Professor Paul Smith as well as Professor Benjamin Kedem for serving as graduate directors during my time here. They have done a wonderful job giving guidance to graduate students.

There are members of the math department staff who deserve recognition for all they have done. I am extremely appreciative of Cristina and Sydney in the graduate office for their administrative support. Thanks to Liz for being patient with me every time I locked myself out of my office. And I am indebted to Stephanie and Cristina for their tireless efforts during a hectic defense day.

Last but not least, I wish to thank my family for their undying love, support and encouragement. I dedicate this thesis to them.

Table of Contents

List of Tables

List of Figures

List of Notations

$\mathbb{R}^d$	d-dimensional Euclidean space
$\boldsymbol{v}$	Column vector
$\boldsymbol{v}^T$	Transpose of a vector $\boldsymbol{v}$
$\|\|\boldsymbol{v}\|\|$	Euclidean norm of $\boldsymbol{v}$
$\mathbf{1}$	Vector consisting of all ones
$\|S\|$	Cardinality of a set S
$\boldsymbol{A}^T$	Transpose of a matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	Inverse of a matrix $\boldsymbol{A}$
$\mathrm{tr}(\boldsymbol{A})$	Trace of a matrix $\boldsymbol{A}$
$\|\|\boldsymbol{A}\|\|$	Generic matrix norm of a matrix $\boldsymbol{A}$
$\mathrm{diag}(a_1, \cdots, a_n)$	Diagonal matrix with diagonal entries $\{a_1, ..., a_n\}$
$X_n \xrightarrow{L_2} X$	L_2 convergence of a random sequence X_n to X
$X_n \xrightarrow{L_1} X$	L_1 convergence of a random sequence X_n to X
$X_n \xrightarrow{a.s.} X$	Almost sure convergence of a random sequence X_n to X
$X_n \xrightarrow{P} X$	Convergence in probability of a random sequence X_n to X

$X_n \xrightarrow{D} X$	Convergence in distribution of a random sequence X_n to X
$X_n = o_p(a_n)$	X_n/a_n converges in probability to zero
$f(h) = O(h)$	There exists $C > 0$ such that $\lvert f(h)\rvert \leq Ch$
$\dfrac{\partial f(\boldsymbol{x})}{\partial x_i}$	Partial derivative of f with respect to x_i
$\nabla f(\boldsymbol{x})$	Gradient of $f(\boldsymbol{x})$ with respect to $\boldsymbol{x}$
$\boldsymbol{A}_i(\boldsymbol{\theta})$	Element wise partial derivative of $\boldsymbol{A}$ with respect to θ_i
$\boldsymbol{A}^i(\boldsymbol{\theta})$	Element wise partial derivative of $\boldsymbol{A}^{-1}$ with respect to θ_i
$\boldsymbol{y}\lvert\boldsymbol{x}$	Random vector $\boldsymbol{y}$ conditioned on random vector $\boldsymbol{x}$
$\sum_{\boldsymbol{s}_i} x(\boldsymbol{s}_i)$	Sum of the elements $\{x(\boldsymbol{s}_i)\}_{i=1}^n$
$f(\boldsymbol{x}) \asymp g(\boldsymbol{x})$	There exists c, C such that $cg(\boldsymbol{x}) \leq f(\boldsymbol{x}) \leq Cg(\boldsymbol{x})$ for all $\boldsymbol{x}$
$\mathbf{1}_S(\boldsymbol{x})$	Indicator function of a set S
$\mathbb{P}_1 \equiv \mathbb{P}_2$	The probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ are equivalent
$\mathbb{P}_1 \perp \mathbb{P}_2$	The probability measures $\mathbb{P}_1$ and $\mathbb{P}_2$ are mutually singular
$L_2(D)$	Square integrable functions on $D \subset \mathbb{R}^d$ w.r.t. Lebesgue measure
$W_2^\ell(D)$	$L_2(D)$ functions with derivatives of order ℓ also in $L_2(D)$

Chapter 1 Introduction

Spatial statistics is concerned with modeling data observed over a spatial domain D, typically a subset of d-dimensional Euclidean space $\mathbb{R}^d$. In most applications, the spatial index set D is a subset of $\mathbb{R}^2$, but the dimension can be any integer $d \geq 1$. The case $d = 1$ is the familiar time series and stochastic process setting. Extending the theory from one-dimension to multidimensional index sets presents difficulties because unlike time, space does not have a natural notion of order. However, many concepts and techniques in spatial statistics borrow from the time series literature and prove to be useful. More information regarding the historical applications and development of spatial statistics can be found in Cressie (1993) and Gelfand et al. (2010).

1.1 Spatial regression models

The primary model of interest in this thesis is the spatial regression model, also known as universal kriging in geostatistics (Cressie (1993)),

$$y(\boldsymbol{s}) = m(\boldsymbol{s}) + e(\boldsymbol{s}) \tag{1.1}$$

where $m(\boldsymbol{s})$ represents an unknown trend or mean (large scale variation) and $e(\boldsymbol{s})$ represents a centered error term (small scale variation). When a set of covariates $\{x_i(\boldsymbol{s})\}_{i=1}^m$ is also observed, a common approach is to parametrically represent the trend as a function $f(x_1(\boldsymbol{s}), \cdots, x_m(\boldsymbol{s}); \boldsymbol{\beta})$ where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of real valued regression parameters. A typical choice of trend function is the linear regression model $f(x_1(\boldsymbol{s}), \cdots, x_{p-1}(\boldsymbol{s}); \boldsymbol{\beta}) = \beta_0 + \sum_{i=1}^{p-1} \beta_i x_i(\boldsymbol{s})$. The covariates can either be modelled as deterministic or random. We adopt the latter approach and interpret the trend as a conditional expectation on the observed covariates. In parametric models, we assume that the errors are spatially correlated through a parametric covariance function $C(\boldsymbol{s}, \boldsymbol{t}; \boldsymbol{\theta})$. The overarching theme of this work is the joint estimation of the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We consider the behavior of these estimators under different asymptotic frameworks.

1.2 Spatial confounding

In the regression model (1.1), it is customary to assume independence of the mean and error. However, since the mean and error vary spatially over the same domain, the possibility of correlation arises. This problem has been discussed extensively in the econometrics literature and is referred to as endogeneity (Hayashi (2000)). However, this is a relatively recent topic in the spatial statistics literature. Clayton et al. (1993) are generally cited as the first to discuss this problem in a statistical setting, and were the first to coin the term, spatial confounding. In the case of a Gaussian setting, with a linear mean, one may argue that this is exclusively

a stochastic regression problem, and cannot occur under classical i.i.d. assumptions. As an example, consider the following simple linear regression model,

$$y_i = \alpha + \beta x_i + e_i, \quad e_i \overset{i.i.d}{\sim} N(0, \sigma^2)$$

where $x_i \overset{i.i.d}{\sim} N(0, \sigma_x^2)$ represents an observed covariate and e_i represents an unobserved error. Assume that x_i and e_i are confounded in the sense that $\text{Cov}(x_i, e_i) = \rho\sigma_x\sigma_e$. Then using properties of the Gaussian distribution, $e_i - \dfrac{\rho\sigma_e}{\sigma_x}x_i$ is independent of x_i. Letting $\epsilon_i = e_i - \dfrac{\rho\sigma_e}{\sigma_x}x_i$, we can re-parametrize our original model as,

$$y_i = \alpha + \gamma x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_e^2(1-\rho^2))$$

where $\gamma = \beta + \dfrac{\rho\sigma_e}{\sigma_x}$. Under this new parametrization, we have effectively removed the confounding in our original model, since the regression slope and correlation parameter are not separately identifiable, but γ is. Since Clayton's article, more attention has been brought to this problem and various methods of mitigating the effects have been proposed (Hodges and Reich (2010), Hodges et al. (2006), Hughes and Haran (2013), Hanks (2015)). More recently, Paciorek (2010) and Page et al. (2017) showed that confounding can lead to bias in classical least squares estimators in the linear regression model, assuming a known covariance structure of $e(\boldsymbol{s})$. We generalize the results by these authors and consider spatial confounding with nonlinear trends and unknown covariance parameters of $e(\boldsymbol{s})$.

1.3 Outline of thesis

In Chapter 2, we present some relevant background probability theory. In particular, we highlight concepts pertaining to asymptotic theory in spatial statistics, such as mixing and microergodicity. We review two prominent asymptotic frameworks in spatial statistics: increasing domain asymptotics and infill asymptotics.

Chapter 3 explores nonlinear regression models under increasing domain asymptotics. The novel contribution of this chapter is to bridge a gap between spatial econometrics and statistics by unifying notable results in these respective fields. We use recently developed spatial limit theorems in econometrics to give a thorough theoretical justification of the use of well known estimators in spatial statistics. This is accompanied by a simulation study and real data analysis in Chapter 4.

In Chapter 5, we discuss confounding in nonlinear spatial regression models under increasing domain asymptotics. In this chapter, we give a survey of various possible models of confounding and address identifiability concerns with these models. We then perform a numerical study of these models in Chapter 6, concluding with an analysis on real data.

Chapter 7 discusses linear regression models under infill asymptotics. Current literature tends to emphasize estimation of the error covariance structure and ignores inference on the mean. The main contribution of this chapter are results concerning the identifiability and estimation of the regression parameters. This is followed by a simulation study in Chapter 8.

In Chapter 9, we give a summary of our contributions and discuss possible

directions for future research. This includes a brief introduction to a hybrid asymptotic framework at the intersection of the two frameworks previously mentioned.

All computations and simulations were performed using R software (R Core Team (2020)). In particular, we use the R Studio IDE (RStudio Team (2019)). In addition, datasets used in this thesis can be found in various R libraries, which we refer to in the main text of the chapter they appear in.

Chapter 2 Modeling with spatial random fields

In this thesis, spatial data are assumed to come from a realization of a random field over some domain $D \subset \mathbb{R}^d$. Formally, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $D \subset \mathbb{R}^d$ a spatial index set. In this work, the index set can either be thought of as varying continuously or discretely over $\mathbb{R}^d$. The former is generally called geostatistical data while the latter is called lattice data (Cressie (1993)). Then, a random field is a measurable mapping $y(\boldsymbol{s}, \omega) : D \times \Omega \to \mathbb{R}$. That is, for any location $\boldsymbol{s}_0 \in D$, the quantity $y(\boldsymbol{s}_0, \omega)$ is a random variable. From a spatial profile point of view, for any $\omega_0 \in \Omega$, the function $y(\boldsymbol{s}, \omega_0)$ represents a realization of a sample path. For notational convenience, we omit the dependence on the element $\omega \in \Omega$ and just write $y(\boldsymbol{s})$ to represent the random field. In statistical applications, we take $\{\boldsymbol{s}_1, .., \boldsymbol{s}_n\}$ to be a set of locations in D. Then, $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))^T$ is a random vector that represents the response data at n locations.

2.1 Gaussian random fields

If $y(\boldsymbol{s})$ is a Gaussian random field with mean function $\mu(\boldsymbol{s}) = \mathbb{E}\big[y(\boldsymbol{s})\big]$, and covariance function $C(\boldsymbol{s}, \boldsymbol{t}) = \mathbb{E}\big[(y(\boldsymbol{s}) - \mu(\boldsymbol{s}))(y(\boldsymbol{t}) - \mu(\boldsymbol{t}))\big]$, then the vector of realizations $\boldsymbol{y} = (y(\boldsymbol{s}_1), \cdots, y(\boldsymbol{s}_n))^T$, follows a multivariate normal distribution $\boldsymbol{y} \sim$

$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i = \mathbb{E}\big[\mu(\boldsymbol{s}_i)\big]$ and $\{\boldsymbol{\Sigma}\}_{i,j} = C(\boldsymbol{s}_i, \boldsymbol{s}_j), i, j = 1, ..., n$. In a regression context, suppose $\boldsymbol{y}$ represents the response variable and we have set of observed covariate data $\boldsymbol{x} = (x(\boldsymbol{s}_1), \cdots, x(\boldsymbol{s}_n))^T$. If the response $\boldsymbol{y}$ and covariate $\boldsymbol{x}$ vectors are jointly Gaussian, that is,

$$\begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{x} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_{yx} \\ \boldsymbol{\Sigma}_{xy} & \boldsymbol{\Sigma}_x \end{pmatrix} \right) \tag{2.1}$$

then the conditional distribution of $\boldsymbol{y}|\boldsymbol{x}$ is also Gaussian, specifically,

$$\boldsymbol{y}|\boldsymbol{x} \sim N(\boldsymbol{\mu}_y + \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_x), \boldsymbol{\Sigma}_y - \boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\Sigma}_{xy}) \tag{2.2}$$

This formula is utilized extensively when discussing confounding models in Chapter 5. Particularly, this formula is used to determine the conditional distribution of $\boldsymbol{e}|\boldsymbol{x}$, where $\boldsymbol{e}, \boldsymbol{x}$ are respectively the error and covariate vectors at the n locations. When there is no confounding, their joint distribution in (2.1) has a block diagonal covariance matrix, indicating independence.

2.2 Stationarity and intrinsic stationarity

If the mean function is constant, $\mu(\boldsymbol{s}) = \mu$, and the covariance function $C(\boldsymbol{s}, \boldsymbol{t})$ depends only on the difference $\boldsymbol{s} - \boldsymbol{t}$, then the random field is called stationary. For a stationary random field, $C(\boldsymbol{h})$ must be a non-negative definite function on $\mathbb{R}^d$. A well known result by Bochner (see for example, p. 20-21 of Gelfand et al. (2010)) states that a real-valued function $C(\boldsymbol{h})$ on $\mathbb{R}^d$ is non-negative definite if and only if

for some symmetric positive measure Λ with distribution function F on $\mathbb{R}^d$,

$$C(\boldsymbol{h}) = \int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T\boldsymbol{h}} dF(\boldsymbol{\omega}) \tag{2.3}$$

Moreover, if F is absolutely continuous with respect to the Lebesgue measure, then F has a density f, called the spectral density. If f is integrable over $\mathbb{R}^d$, then the following inversion formula gives a relationship between f and C,

$$f(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^T\boldsymbol{h}} C(\boldsymbol{h}) d\boldsymbol{h} \tag{2.4}$$

A less restrictive form of stationarity is intrinsic stationarity. For a constant mean random field $y(\boldsymbol{s})$, the variogram is defined as $\gamma(\boldsymbol{s},\boldsymbol{t}) = \frac{1}{2}\mathbb{E}\big[(y(\boldsymbol{s}) - y(\boldsymbol{t}))^2\big]$. If the variogram depends only on the difference between locations $\boldsymbol{s} - \boldsymbol{t}$ then the random field is said to be intrinsically stationary. All stationary random fields are intrinsically stationarity, but the converse is not true (e.g. Brownian motion on $\mathbb{R}$ is intrinsically stationary, but not stationary).

2.2.1 The Matérn class of covariance functions

When the covariance function depends on $\boldsymbol{h}$ only through its Euclidean length $||\boldsymbol{h}||$, it is said to be (weakly) isotropic. A popular choice of isotropic covariance functions comes from the Matérn family,

$$C(\boldsymbol{h}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\theta\sqrt{2\nu}||\boldsymbol{h}||)^\nu K_\nu(\theta\sqrt{2\nu}||\boldsymbol{h}||), \quad \sigma > 0, \theta > 0, \nu > 0 \tag{2.5}$$

where $K_\nu(z)$ is a modified Bessel function of the second kind (Watson (1995)) and $\theta > 0$ and $\nu > 0$ represent scale and smoothness parameters respectively. The corresponding family of spectral densities takes the form,

$$f(\boldsymbol{\omega}) = \sigma^2 \frac{\Gamma(\nu + \frac{d}{2})}{\Gamma(\nu)\pi^{\frac{d}{2}}} \frac{(2\nu\theta^2)^\nu}{(2\nu\theta^2 + ||\boldsymbol{\omega}||^2)^{\nu+\frac{d}{2}}} \tag{2.6}$$

Stein (1999), p.12, strongly advocates the use of the Matérn class when modelling stationary random fields due to its flexibility. The flexibility of this family lies in the choice of smoothness parameter ν. For example, the exponential family $C(\boldsymbol{h}) = \sigma^2 e^{-\theta||\boldsymbol{h}||}$, can be thought of as a member of the Matérn family when $\nu = 1/2$ and the limiting case $\nu \to \infty$ leads to the Gaussian class of covariance functions $C(\boldsymbol{h}) = \sigma^2 e^{-\frac{1}{2}\theta^2||\boldsymbol{h}||^2}$.

2.3 Multivariate random fields

When dealing with several random fields, it will sometimes be necessary to model their dependence through a multivariate covariance function. A multivariate random field takes the form $\boldsymbol{y}(\boldsymbol{s}) = (y_1(\boldsymbol{s}), ..., y_p(\boldsymbol{s}))^T$, where each component, $y_i(\boldsymbol{s}), i = 1, ..., p$ is a scalar random field. Assuming that $\boldsymbol{y}(\boldsymbol{s})$ is mean zero and stationary, the multivariate extension of the covariance function is a $p \times p$ matrix valued function,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} C_{11}(\boldsymbol{h}) & \cdots & C_{1p}(\boldsymbol{h}) \\ \vdots & \ddots & \vdots \\ C_{p1}(\boldsymbol{h}) & \cdots & C_{pp}(\boldsymbol{h}) \end{pmatrix}$$

where $C_{ij}(\boldsymbol{h}) = \mathbb{E}\big[y_i(\boldsymbol{s})y_j(\boldsymbol{s}+\boldsymbol{h})\big]$ is a scalar cross-covariance function between the components $y_i(\boldsymbol{s})$ and $y_j(\boldsymbol{s}), 1 \leq i, j \leq p$. The corresponding spectral density matrix is,

$$\boldsymbol{f}(\boldsymbol{\omega}) = \begin{pmatrix} f_{11}(\boldsymbol{\omega}) & \cdots & f_{1p}(\boldsymbol{\omega}) \\ \vdots & \ddots & \vdots \\ f_{p1}(\boldsymbol{\omega}) & \cdots & f_{pp}(\boldsymbol{\omega}) \end{pmatrix}$$

A result due to Cramér gives a multivariate extension of Bochner's theorem (see for example, Gneiting et al. (2010), p. 1176).

Theorem 2.3.1. *The matrix function $\boldsymbol{C}(\boldsymbol{h})$ is the cross covariance function for a stationary multivariate random random field if and only if the corresponding matrix of cross spectral densities $\boldsymbol{f}(\boldsymbol{\omega})$ is almost everywhere nonnegative definite.*

Gneiting et al. (2010) developed a bivariate version of the Matérn model using this criterion. This was followed by a multivariate extension to any number of components by Apanasovich et al. (2012). When discussing possible confounding models in Chapter 5, we describe the dependence between the error and covariates using a multivariate random field structure.

2.4 Mixing for random fields

Since we are dealing with dependent spatial data, standard laws of large numbers and central limit theorems cannot be applied directly. We use a concept called mixing, which is a way of controlling the asymptotic dependence in a spatial random field. Mixing has been well studied in the time series and random process context,

where there is a notion of order in the index set. Going from $\mathbb{R}$ to $\mathbb{R}^d$ requires a little more care, particularly in both the cardinalities and distances between index sets. Bradley (2005) and Doukhan (1994) give an overview of some results of mixing in random fields. There are different versions of mixing stated by these authors, so we focus our attention on α-mixing random fields. Below are definitions related to α-mixing, one in terms of σ-algebras, and the other in terms of random fields.

Definition 2.4.1. *The α-mixing coefficient between σ-algebras $\mathcal{F}$ and $\mathcal{G}$ is*

$$\alpha(\mathcal{F}, \mathcal{G}) := \sup(|\mathbb{P}(A)\mathbb{P}(B) - \mathbb{P}(A \cap B)| : A \in \mathcal{F}, B \in \mathcal{G}) \tag{2.7}$$

Definition 2.4.2. *For $n \in \mathbb{N}$, let $D_n = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\} \subset D$ denote the sampling domain. For $A \subset D_n$ and $B \subset D_n$, and a random field $y(\boldsymbol{s})$, let $\sigma_n(A) := \sigma(y(\boldsymbol{s}_i), \boldsymbol{s}_i \in A)$, the σ-algebra generated by $\{y(\boldsymbol{s}_i), \boldsymbol{s}_i \in A\}$, and similarly for $\sigma_n(B)$. The α-mixing coefficient for the random field $y(\boldsymbol{s})$ is defined as,*

$$\alpha_n(k, l, h) := \sup(\alpha(\sigma_n(A), \sigma_n(B)) : |A| \leq k, |B| \leq l, d(A, B) \geq h) \tag{2.8}$$

where $d(A, B) = \inf\{||\boldsymbol{a} - \boldsymbol{b}|| : \boldsymbol{a} \in A, \boldsymbol{b} \in B\}$ is the Euclidean distance between the sets A, B.

To account for possibly non-nested sampling domains D_n, define the uniform α-mixing coefficient $\alpha(k, l, h) = \sup_n \alpha_n(k, l, h)$. The following bound (Heyde and Hall (1980), Corollary A.2, p. 278) allows us to control the dependence a random field between two sets of locations.

Lemma 2.4.3. *Suppose A and B are finite sets in D with $|A| = k, |B| = l$ and $d(A,B) = h$. Let X and Y be $\sigma_n(A)$ and $\sigma_n(B)$ measurable respectively. If $\mathbb{E}\big[|X|^p\big]^{1/p} = ||X||_p < \infty$ and $\mathbb{E}\big[|Y|^q\big]^{1/q} = ||Y||_q < \infty$ with $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} = 1$ with $p, q > 1$ and $r > 0$, then,*

$$Cov(X,Y) \leq 8(\alpha(k,l,h))^{1/r}||X||_p||Y||_q \tag{2.9}$$

This is helpful in establishing LLNs and CLTs for random fields. Under specific conditions on the α-mixing coefficient, various LLNs and CLTs have been proved (see for example, Jenish and Prucha (2009), Bolthausen (1982) and Guyon (1995), among others). We will explore these results in more depth in Chapter 3.

2.5 Asymptotic frameworks

2.5.1 Increasing domain asymptotics

Suppose the spatial domain D is unbounded and for each n, there is a $\delta > 0$, independent of n such that $||\boldsymbol{s}_i - \boldsymbol{s}_j|| > \delta$, $\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n$. This precludes the spatial locations from becoming too dense as the number of observations grow. This asymptotic framework is called increasing domain asymptotics. Under this framework, it has been shown that under certain regularity conditions, classical estimators of regression and covariance parameters satisfy some sort of consistency and asymptotic normality result (Mardia and Marshall (1984), Cressie and Lahiri (1996), Crujeiras and van Keilegom (2010)). This asymptotic framework is more compatible with the

concept of mixing since it is natural to assume the dependence between observations decays as the distance between their respective locations grows.

2.5.2 Infill asymptotics

Suppose now that the spatial domain D is compact and the sampling locations become increasingly dense in D with the number of observations. This asymptotic framework is referred to as infill asymptotics. Unlike in increasing domain asymptotics, typical estimators do not behave as expected. In particular, it has been shown (Ying (1991), Chen et al. (2000), Zhang (2004), Du et al. (2009), Wang and Loh (2011), Kaufman and Shaby (2013), Tang et al. (2021)) that for a zero mean Gaussian random field with Matérn covariance, not all covariance parameters can be consistently estimated, but only certain functions of them, called the microergodic parameters (Stein (1999)). Before we define this term, we review the measure theoretic concepts of equivalence and mutual singularity.

Definition 2.5.1. *Let $\mathbb{P}_1$ and $\mathbb{P}_2$ be two probability measures on a measurable space $(\Omega, \mathcal{F})$. Then $\mathbb{P}_1$ is absolutely continuous with respect to $\mathbb{P}_2$ if for all $A \in \mathcal{F}$, $\mathbb{P}_2(A) = 0$ implies $\mathbb{P}_1(A) = 0$. $\mathbb{P}_1$ and $\mathbb{P}_2$ are said to be equivalent, denoted as $\mathbb{P}_1 \equiv \mathbb{P}_2$, if they are absolutely continuous with respect to each other. $\mathbb{P}_1$ and $\mathbb{P}_2$ are said to be mutually singular, denoted as $\mathbb{P}_1 \perp \mathbb{P}_2$, if there is some $A \in \mathcal{F}$ such that $\mathbb{P}_1(A) = 1$ and $\mathbb{P}_2(A) = 0$.*

In particular, for two equivalent measures $\mathbb{P}_1$ and $\mathbb{P}_2$, it would be impossible to determine with certainty which measure is correct based on observing $\omega \in \Omega$. In

general, two probability measures may be neither equivalent nor mutually singular. However, for Gaussian probability measures, there is a dichotomy: they are *either* equivalent or mutually singular (Feldman (1958), Hájek (1958)). We describe the approach by Hájek (1958), who used a form of information theoretic divergence in determining the equivalence or mutual singularity of two Gaussian measures $\mathbb{P}_1, \mathbb{P}_2$. Let $D_n = \{\boldsymbol{s}_1, .., \boldsymbol{s}_n\}$ be a nested sequence of locations whose countable union is dense in D and $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))^T$ be the corresponding realization of the random field $y(\boldsymbol{s})$. Recall from Section 2.1 that $\boldsymbol{y}$ follows a multivariate normal distribution under $\mathbb{P}_1, \mathbb{P}_2$. Denote by p_n, the likelihood ratio of $\mathbb{P}_2$ to $\mathbb{P}_1$ based on the observations $\boldsymbol{y}$.

Definition 2.5.2. *The entropy distance between* $\mathbb{P}_2$ *and* $\mathbb{P}_1$ *is defined as,*

$$J_n = \mathbb{E}_2[\log p_n] - \mathbb{E}_1[\log p_n] \tag{2.10}$$

Here $\mathbb{E}_i, i = 1, 2$ *denotes the expectation of* $\log p_n$ *under* $\mathbb{P}_i, i = 1, 2$.

The term entropy distance has other names in literature, including "symmetrized Kullback-Liebler divergence" and "Jeffreys divergence" (Hájek (1958)). It is known that J_n is monotonically increasing (Stein (1999), p. 116) and thus, J_n either tends to a finite limit or ∞. The condition $\lim_{n\to\infty} J_n = \infty$ is necessary for $\mathbb{P}_1 \perp \mathbb{P}_2$ to hold, even in the non-Gaussian case. For Gaussian measures, the condition $\lim_{n\to\infty} J_n = \infty$ is sufficient for $\mathbb{P}_1 \perp \mathbb{P}_2$ to hold as well. We now formally state this result by Hájek, of which a translated version can be found in Theorem 1, p. 77, of Ibragimov and Rozanov (1978), also in Theorem 4, p. 117, of Stein (1999).

Lemma 2.5.3. *The Gaussian measures* $\mathbb{P}_1$ *and* $\mathbb{P}_2$ *are either equivalent or mutually singular. They are mutually singular if and only if* $\lim_{n\to\infty} J_n = \infty$.

Various results on the equivalence and singularity of Gaussian measures have been known for decades by probabilists (Ibragimov and Rozanov (1978), Skorokhod and Yadrenko (1973)), with Lemma 2.5.3 being an example. It was only until relatively recently that these results began to permeate into the statistics literature (Stein (1988)). A literature review of these results can be found in Chapter 7. We now are ready to define the concept of microergodicity, which in statistical terms, is a form of identifiability.

Definition 2.5.4. *Let* $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ *be a family of measures on a measurable space* $(\Omega, \mathcal{F})$. *Then a function* $h(\boldsymbol{\theta})$ *is microergodic if for any two* $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \Theta$, $h(\boldsymbol{\theta}_1) \neq h(\boldsymbol{\theta}_2)$ *implies that* $\mathbb{P}_{\boldsymbol{\theta}_1} \perp \mathbb{P}_{\boldsymbol{\theta}_2}$. *We say that* $h(\boldsymbol{\theta})$ *is non-microergodic if* $h(\boldsymbol{\theta}_1) \neq h(\boldsymbol{\theta}_2)$ *implies that* $\mathbb{P}_{\boldsymbol{\theta}_1} \equiv \mathbb{P}_{\boldsymbol{\theta}_2}$

Zhang (2004), p. 252, shows that a consistent estimator of $h(\boldsymbol{\theta})$ cannot exist in the non-microergodic case. We state this formally as a theorem below. Note this theorem holds for any probability measures, not necessarily Gaussian.

Theorem 2.5.5. *Let* $\{\mathbb{P}_{\boldsymbol{\theta}}, \boldsymbol{\theta} \in \Theta\}$ *be a family of probability measures. If* $h(\boldsymbol{\theta})$ *is non-microergodic, then* $h(\boldsymbol{\theta})$ *cannot be consistently estimated.*

Appendix: Gaussian sample path behavior

In this appendix, we give a few background results on Gaussian sample path behavior. As we show in Chapter 7, the microergodicity of regression parameters

depends on properties of the covariate sample paths. The behavior of the sample paths is conveniently determined by properties of the covariance function. Adler (1981) (Theorem 3.4.1 and its Corollary, p. 61-62) presents results on the sample path continuity of a Gaussian field over a compact domain $D \subset \mathbb{R}^d$.

Theorem 2.5.6. *Let* $\{e(\boldsymbol{s}), s \in D\}$ *be a zero-mean, Gaussian random field with continuous covariance function. If for some* $0 < C < \infty$ *and some* $\epsilon > 0$,

$$\mathbb{E}\big[(e(\boldsymbol{s}) - e(\boldsymbol{t}))^2\big] \leq \frac{C}{|\log(||\boldsymbol{s} - \boldsymbol{t}||)|^{1+\epsilon}}$$

for all $\boldsymbol{s}, \boldsymbol{t} \in D$, *then* $e(\boldsymbol{s})$ *has continuous sample paths over* D *almost surely.*

Corollary 2.5.7. *Let* $\{e(\boldsymbol{s}), \boldsymbol{s} \in D\}$ *be a zero-mean, stationary Gaussian random field with continuous covariance function* $C(\boldsymbol{s})$. *If for some* $0 < C < \infty$ *and some* $\epsilon > 0$,

$$C(0) - C(\boldsymbol{s}) \leq \frac{C}{|\log(||\boldsymbol{s}||)|^{1+\epsilon}}, \quad \forall \boldsymbol{s} \in D$$

then $e(\boldsymbol{s})$ *has continuous sample paths over* D *almost surely.*

Abrahamsen (1997), p. 20, conjectures that any Gaussian random field with a continuous covariance function satisfies the above inequalities and thus, all such random fields possess continuous sample paths almost surely. For the Matérn case, the following result (see Theorem 3.4.3 of Adler (1981)) implies this conjecture.

Theorem 2.5.8. *Let* $\{e(\boldsymbol{s}), \boldsymbol{s} \in D\}$ *be a zero-mean, stationary Gaussian random field with spectral density* $f(\boldsymbol{\omega})$. *Then* $e(\boldsymbol{s})$ *has continuous sample paths on* D *almost*

surely if for some $\alpha > 0$,

$$\int_{\mathbb{R}^d} [\log(1 + ||\boldsymbol{\omega}||)]^{1+\alpha} f(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty$$

We apply this result to the case where $e(\boldsymbol{s})$ has Matérn spectral density defined in (2.6).

Proposition 2.5.1. *Let* $\{e(\boldsymbol{s}), s \in D\}$ *be a zero-mean, stationary Gaussian random field with Matérn covariance. Then the condition of Theorem 2.5.8 is satisfied and thus,* $e(\boldsymbol{s})$ *has continuous sample paths on* D *almost surely.*

Proof. Due to isotropy and after converting to spherical coordinates, it sufficient to show that,

$$\int_0^\infty \frac{[\log(1+r)]^{1+\alpha}}{(1+r^2)^{\nu+\frac{d}{2}}} r^{d-1} dr < \infty$$

Splitting the integral over $[0, 1]$ and $[1, \infty)$, we have,

$$\int_0^\infty \frac{[\log(1+r)]^{1+\alpha}}{(1+r^2)^{\nu+\frac{d}{2}}} r^{d-1} dr = \int_0^1 \frac{[\log(1+r)]^{1+\alpha}}{(1+r^2)^{\nu+\frac{d}{2}}} r^{d-1} dr + \int_1^\infty \frac{[\log(1+r)]^{1+\alpha}}{(1+r^2)^{\nu+\frac{d}{2}}} r^{d-1} dr$$

The first integral is easily seen to be convergent for $\alpha, \nu > 0$ and $d \geq 1$. For the second,

$$\begin{aligned} \int_1^\infty \frac{[\log(1+r)]^{1+\alpha}}{(1+r^2)^{\nu+\frac{d}{2}}} r^{d-1} dr &= \int_1^\infty \frac{[\log(1+r)]^{1+\alpha}}{r^{2\nu+d}(1+1/r^2)^{\nu+\frac{d}{2}}} r^{d-1} dr \\ &\leq \int_1^\infty \frac{[\log(1+r)]^{1+\alpha}}{r^{2\nu+1}} dr \end{aligned}$$

since $1 + 1/r^2 \geq 1$ on $[1, \infty)$. Since $\lim_{r\to\infty} \frac{[\log(1+r)]^{1+\alpha}}{r^\nu} = 0$ for any $\alpha, \nu > 0$, there exists an $M > 0$ so that,

$$\int_1^\infty \frac{[\log(1+r)]^{1+\alpha}}{r^{2\nu+1}} dr \leq M \int_1^\infty \frac{1}{r^{\nu+1}} dr$$

The above integral converges for $\nu > 0$. □

Since $e(\boldsymbol{s})$ has continuous sample paths on D, the notion of the integral $\int_D e(\boldsymbol{s}) d\boldsymbol{s}$ is well defined and exists almost surely.

Smoothness plays a major factor in the microergodicity of regression parameters. The next few results involve the classes of functions $W_2^\ell(D)$, called Sobolev spaces. Informally, functions in this class are square integrable with square integrable derivatives up to order ℓ. First, let $\alpha = (\alpha_1, ..., \alpha_N)$ be an N-tuple of nonnegative integers and denote $|\alpha| = \sum_{i=1}^N \alpha_i$. Next, let $D^\alpha f(\boldsymbol{x})$ be the set of all $|\alpha|^{th}$ order partial derivatives of f. Finally, let $C^\infty(D)$ denote the space of infinitely differentiable functions with supports contained in D. There are two slightly different definitions of $W_2^\ell(D)$ depending on whether ℓ is an integer or fractional. We first give the definition for integer order ℓ.

Definition 2.5.9. *For $\ell \in \mathbb{N}$, the Sobolev space $W_2^\ell(D)$ of integer order ℓ is defined as the closure of $C^\infty(D)$ relative to the norm,*

$$||f||^2_{W_2^\ell(D)} = \int_D |f(\boldsymbol{x})|^2 d\boldsymbol{x} + \sum_{|\alpha|=\ell} \int_D |D^\alpha f(\boldsymbol{x})|^2 d\boldsymbol{x}$$

Next, we give the definition for fractional order ℓ.

Definition 2.5.10. *For $\ell > 0$, we can write $\ell = k + \gamma$, where $k \in \mathbb{N}$ is the integer part of ℓ and $\gamma \in (0, 1)$ is the fractional part. Then the Sobolev space $W_2^\ell(D)$ of fractional order ℓ is defined as the closure of $C^\infty(D)$ relative to the norm,*

$$||f||^2_{W_2^\ell(D)} = \int_D |f(\boldsymbol{x})|^2 d\boldsymbol{x} + \sum_{|\alpha|=k} \int_D \int_D \frac{|D^\alpha f(\boldsymbol{x}) - D^\alpha f(\boldsymbol{y})|^2}{||\boldsymbol{x} - \boldsymbol{y}||^{2\gamma+d}} d\boldsymbol{x} d\boldsymbol{y}$$

Scheuerer (2010) gives spectral conditions for the sample paths of a stationary Gaussian random field $x(\boldsymbol{s})$ to be in $W_2^\ell(D)$ almost surely. For integer order ℓ, Scheuerer (2010) gives a necessary and sufficient condition (Corollary 1 and Proposition 1).

Theorem 2.5.11. *Let $\{x(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a stationary Gaussian random field with spectral density $f(\boldsymbol{\omega})$. Then the sample paths of $x(\boldsymbol{s})$ are in $W_2^\ell(D)$ almost surely if and only if,*

$$\int_{\mathbb{R}^d} ||\boldsymbol{\omega}||^{2\ell} f(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty \tag{2.11}$$

For fractional order ℓ, Scheuerer (2010) gives a sufficient condition (Theorem 3).

Theorem 2.5.12. *Let $\{x(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a stationary Gaussian random field with spectral density $f(\boldsymbol{\omega})$. If for some $\alpha > 0$ and $\ell \in \mathbb{R}_+ \setminus \mathbb{N}$,*

$$\int_{\mathbb{R}^d} [\log(1 + ||\boldsymbol{\omega}||)]^{1+\alpha} ||\boldsymbol{\omega}||^{2\ell} f(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty \tag{2.12}$$

then the sample paths of $x(\boldsymbol{s})$ are in $W_2^\ell(D)$ almost surely.

Chapter 3 Nonlinear regression under increasing domain asymptotics

We consider the nonlinear regression model with several covariates,

$$y(\boldsymbol{s}) = f(x_1(\boldsymbol{s}), \cdots, x_m(\boldsymbol{s}); \boldsymbol{\beta}) + e(\boldsymbol{s}), \quad \boldsymbol{s} \in D \subset \mathbb{R}^d$$

where $(x_1(\boldsymbol{s}), ..., x_m(\boldsymbol{s}))^T$ is a multivariate random field and $e(\boldsymbol{s})$ is a mean zero random field. We assume that the form of the function f is known and $\boldsymbol{\beta} \in B \subset \mathbb{R}^p$ is a vector of unknown regression parameters. In this chapter, the covariates are assumed to be independent of the error $e(\boldsymbol{s})$. We assume D is a countably infinite, possibly irregularly spaced lattice in $\mathbb{R}^d$, where the minimum euclidean distance between a given pair of spatial locations is bounded below.

Assumption 3.0.1. *There exists a* $\delta > 0$ *such that* $||\boldsymbol{s} - \boldsymbol{t}|| \geq \delta > 0, \ \forall \boldsymbol{s}, \boldsymbol{t} \in D$. *Without loss of generality, we can assume that* $\delta = 1$.

Assume $n > p$ and let $D_n = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\} \subset D$ be a sequence of finite sets, not necessarily nested, converging to D. The sequence D_n represents the spatial locations at which $y(\boldsymbol{s})$ and $x_j(\boldsymbol{s}), j = 1, ..., m$ are observed. In vector notation, we

write the regression model as $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta}) + \boldsymbol{e}$ where,

$$\begin{aligned} \boldsymbol{y} &= (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))^T \\ \boldsymbol{x}_j &= (x_j(\boldsymbol{s}_1), ..., x_j(\boldsymbol{s}_n))^T, \quad j = 1, ..., m \\ \boldsymbol{e} &= (e(\boldsymbol{s}_1), ..., e(\boldsymbol{s}_n))^T \end{aligned}$$

and $\boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta})$ is a $n \times 1$ vector containing $f(x_1(\boldsymbol{s}), ..., x_m(\boldsymbol{s}); \boldsymbol{\beta})$ when evaluated at each spatial location. Let $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ be the $n \times n$ parametric covariance matrix of $\boldsymbol{e}$ where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ are unknown covariance parameters. We investigate the joint estimation of the unknown parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ using the following multistage approach.

1. As a provisional estimator of $\boldsymbol{\beta}$, use ordinary least squares,

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta}))^T (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta})) \tag{3.1}$$

2. Form the residuals $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \hat{\boldsymbol{\beta}}_{OLS})$ as a proxy for the true errors. Then estimate $\boldsymbol{\theta}$ using a least squares variogram approach (see Section 3.2 for a description).

3. Using $\hat{\boldsymbol{\theta}}$ from step 2 as a plug-in estimator, re-estimate $\boldsymbol{\beta}$ using feasible generalized least squares (FGLS),

$$\hat{\boldsymbol{\beta}}_{FGLS} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta}))^T \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}}) (\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta})) \tag{3.2}$$

Note that this procedure is not novel (see for example, Chapter 3 of Gelfand et al. (2010) for a discussion) and is well known to statisticians and econometricians. However, to our knowledge, there is a lack of careful theoretical consideration of the above steps in the spatial statistics literature. Our novel contribution in this chapter is to justify the above steps by unifying the results of Jenish and Prucha (2009), Lahiri et al. (2002) and Crujeiras and van Keilegom (2010). Jenish and Prucha (2009) establish laws of large numbers for random fields, which allow us to prove the consistency of the OLS estimator (3.1) in step 1. Crujeiras and van Keilegom (2010) prove the consistency of least squares variogram estimators in step 2, while Lahiri et al. (2002) establish the asymptotic normality of these estimators. Finally, Crujeiras and van Keilegom (2010) prove the consistency and asymptotic normality of the FGLS estimator in step 3.

This work was inspired by Crujeiras and van Keilegom (2010), who investigate the same estimation procedure. However, we generalize some of their results. First, we relax their assumption that the sampling points $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$ are regularly spaced. Next, these authors only mention in passing, without proof, that the OLS estimator in step 1 is consistent, citing a result by Gallant and Goebel (1976). However, this cited result was given in a time series context with autoregressive errors. The laws of large numbers by Jenish and Prucha (2009) allow us to justify step 1 adequately in a spatial random field setting. Finally, Crujeiras and van Keilegom (2010) use a fairly restrictive assumption that $e(\boldsymbol{s})$ is Gaussian, which essentially gives asymptotic normality of the FGLS estimator for free. We make use of a spatial central limit theorem proved by Jenish and Prucha (2009), which allows us to

bypass this Gaussian assumption. In the results that follow, the proofs are deferred to Section 3.5, unless otherwise stated in the main section. The first step is proving the consistency of the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ in (3.1) as a function of known $\boldsymbol{\theta}$.

3.1 Consistency of ordinary least squares

3.1.1 Linear trend

First, consider the linear regression model,

$$\boldsymbol{f}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_m; \boldsymbol{\beta}) = \beta_* \mathbf{1} + \sum_{j=1}^{p-1} \beta_j \boldsymbol{x}_j = \boldsymbol{X}\boldsymbol{\beta}, \quad \boldsymbol{X} = \begin{bmatrix} \mathbf{1} & \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{p-1} \end{bmatrix}_{n \times p} \tag{3.3}$$

where the OLS estimator has a closed form, $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Letting $\boldsymbol{\beta}_0$ denote the true regression parameters, we can write,

$$\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 = \left(\frac{1}{n} \sum_{\boldsymbol{s}_i} \boldsymbol{X}_i \boldsymbol{X}_i^T \right)^{-1} \frac{1}{n} \sum_{\boldsymbol{s}_i} \boldsymbol{X}_i e_i \tag{3.4}$$

where $\boldsymbol{X}_i = (1, x_1(\boldsymbol{s}_i), \cdots, x_{p-1}(\boldsymbol{s}_i))$ is the i^{th} row of the design matrix $\boldsymbol{X}$, and $e_i = e(\boldsymbol{s}_i)$, $i = 1, ..., n$. The following are L_1 and L_2 laws of large numbers which can be applied directly to $\hat{\boldsymbol{\beta}}_{OLS}$. The L_1 LLN can be found in Theorem 3 of Jenish and Prucha (2009), and so we state it without proof. The L_2 LLN is not directly stated by these authors, but is proved in Section 3.5 as an extension using their assumptions and arguments. Naturally, the L_2 law requires stronger moment and mixing conditions. Recall the definition of α-mixing coefficient in (2.8).

Theorem 3.1.1. *Suppose that* $\{z(\boldsymbol{s}_i), \boldsymbol{s}_i \in D_n\}$ *is a realization of an* α*-mixing random field with mixing coefficient* $\alpha(k,l,m)$ *and that the limit* $\frac{1}{n}\sum_{\boldsymbol{s}_i} \mathbb{E}\big[z(\boldsymbol{s}_i)\big]$ *exists.*

1. *(*L_1 *LLN) If for some* $\eta > 0$,

$$\sup_{\boldsymbol{s}_i \in D_n} \mathbb{E}\big[|z(\boldsymbol{s}_i)|^{1+\eta}\big] < \infty \quad \textit{and} \quad \sum_{m=1}^{\infty} m^{d-1}\alpha(1,1,m) < \infty \tag{3.5}$$

then $\frac{1}{n}\sum_{\boldsymbol{s}_i}(z(\boldsymbol{s}_i) - \mathbb{E}\big[z(\boldsymbol{s}_i)\big]) \xrightarrow{L_1} 0$.

2. *(*L_2 *LLN) If for some* $\eta > 0$,

$$\sup_{\boldsymbol{s}_i \in D_n} \mathbb{E}\big[|z(\boldsymbol{s}_i)|^{2+\eta}\big] < \infty \quad \textit{and} \quad \sum_{m=1}^{\infty} m^{d-1}(\alpha(1,1,m))^{\frac{\eta}{2+\eta}} < \infty \tag{3.6}$$

then $\frac{1}{n}\sum_{\boldsymbol{s}_i}(z(\boldsymbol{s}_i) - \mathbb{E}\big[z(\boldsymbol{s}_i)\big]) \xrightarrow{L_2} 0$. *In fact, the variance of* $\frac{1}{n}\sum_{\boldsymbol{s}_i}(z(\boldsymbol{s}_i) - \mathbb{E}\big[z(\boldsymbol{s}_i)\big])$ *is* $O(\frac{1}{n})$.

Since both L_1 and L_2 convergence imply convergence in probability, we can use either to show that $\hat{\boldsymbol{\beta}}_{OLS}$ is consistent. In the following result, we establish the L_2 consistency of $\hat{\boldsymbol{\beta}}_{OLS}$ and the proof can be found in Section 3.5.

Proposition 3.1.1. *Assume that the* $\mathbb{R}^p$*-valued random field* $(x_1(\boldsymbol{s}), ..., x_{p-1}(\boldsymbol{s}), e(\boldsymbol{s}))$ *is jointly* α*-mixing with mixing coefficient* $\alpha(k,l,m)$ *satisfying (3.6) in Theorem 3.1.1. Moreover, assume that the squared covariates* $\{x_j^2(\boldsymbol{s})\}_{j=1}^{p-1}$ *and mean zero error* $e(\boldsymbol{s})$ *satisfy the uniform integrability condition in (3.6) of Theorem 3.1.1. Let* $\boldsymbol{A} = \lim_{n\to\infty} \frac{1}{n}\sum_{\boldsymbol{s}_i} \mathbb{E}\big[\boldsymbol{X}_i \boldsymbol{X}_i^T\big]$ *be the limiting expectation of the matrix defined in (3.4).*

If $\boldsymbol{A}$ *exists and is non-singular, then the OLS estimator in the linear regression model (3.3) satisfies* $\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 \xrightarrow{L_2} 0$.

Note in Proposition 3.1.1, we do not require intrinsic stationarity of $e(\boldsymbol{s})$.

3.1.2 Nonlinear trend

For a more general nonlinear trend $\boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta})$, we look for an estimator solving the following least squares criterion,

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} q(\boldsymbol{s}_i; \boldsymbol{\beta}) \tag{3.7}$$

where $q(\boldsymbol{s}_i; \boldsymbol{\beta}) = (y(\boldsymbol{s}_i) - f(x_1(\boldsymbol{s}_i), \cdots, x_m(\boldsymbol{s}_i); \boldsymbol{\beta}))^2$. Generally, for M-estimation problems such as this, we require a uniform law of large numbers. The following result on consistency of M-estimators can be found in Theorem 5.7 from van der Vaart (1998).

Lemma 3.1.2. *Let* M_n *be random functions and let* M *be a non-random function of* $\boldsymbol{\beta}$ *so that,*

1. $\sup_{\boldsymbol{\beta} \in B} |M_n(\boldsymbol{\beta}) - M(\boldsymbol{\beta})| \xrightarrow{P} 0$
2. $\inf_{||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \geq \epsilon} M(\boldsymbol{\beta}) > M(\boldsymbol{\beta}_0), \quad \forall \epsilon > 0$

Then, any sequence of minimizers, $\hat{\boldsymbol{\beta}}_n$*, such that* $M_n(\hat{\boldsymbol{\beta}}_n) \leq M_n(\boldsymbol{\beta}_0) + o_p(1)$ *converges in probability to* $\boldsymbol{\beta}_0$.

The second condition is an identifiability condition that ensures the limiting function M has a unique minimum at $\boldsymbol{\beta}_0$. The first condition is the one that needs to be

proved in these types of problems. Our random function in this case is the sequence of averages $M_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} q(\boldsymbol{s}_i;\boldsymbol{\beta})$. Jenish and Prucha (2009) give a result that establishes the uniform convergence of a sequence of spatial random averages like $M_n(\boldsymbol{\beta})$. First, we need to define the concept of stochastic equicontinuity.

Definition 3.1.3. *Let $\{g(\boldsymbol{s}_i;\boldsymbol{\beta}), \boldsymbol{s}_i \in D_n, \boldsymbol{\beta} \in B\}$ be a sequence of random functions. Then $\{g(\boldsymbol{s}_i;\boldsymbol{\beta})\}$ is stochastically equicontinuous on B if for every $\epsilon > 0$,*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\frac{1}{n}\sum_{\boldsymbol{s}_i}\mathbb{P}\left(\sup_{\boldsymbol{\beta}'\in B}\sup_{||\boldsymbol{\beta}-\boldsymbol{\beta}'||<\delta}|g(\boldsymbol{s}_i;\boldsymbol{\beta})-g(\boldsymbol{s}_i;\boldsymbol{\beta}')|>\epsilon\right)=0$$

We also need the following uniform integrability condition on $\{g(\boldsymbol{s}_i;\boldsymbol{\beta})\}$,

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{\boldsymbol{s}_i}\mathbb{E}\big[d(\boldsymbol{s}_i)\mathbb{1}(d(\boldsymbol{s}_i)>k)\big]=0, \quad \text{as} \;\; k\to\infty \tag{3.8}$$

where $d(\boldsymbol{s}_i) = \sup_{\boldsymbol{\beta}\in B}|g(\boldsymbol{s}_i;\boldsymbol{\beta})|$. We are now ready to state, without proof, the ULLN result found in Theorem 2 of Jenish and Prucha (2009).

Theorem 3.1.4. *Let B be a bounded subset of Euclidean space. Suppose that the sequence of random function $\{g(\boldsymbol{s}_i;\boldsymbol{\beta})\}$ are stochastically equicontinuous as in Definition 3.1.3 and satisfy the condition in (3.8). If for any $\boldsymbol{\beta} \in B$, the $\{g(\boldsymbol{s}_i;\boldsymbol{\beta})\}$ satisfy a pointwise law of large numbers in the sense that,*

$$\frac{1}{n}\sum_{\boldsymbol{s}_i}(g(\boldsymbol{s}_i;\boldsymbol{\beta})-E\big[g(\boldsymbol{s}_i;\boldsymbol{\beta})\big]) \xrightarrow{P} 0 \tag{3.9}$$

then, it follows that a uniform law of large numbers holds,

$$\sup_{\boldsymbol{\beta}\in B}\left|\frac{1}{n}\sum_{\boldsymbol{s}_i}(g(\boldsymbol{s}_i;\boldsymbol{\beta})-E\big[g(\boldsymbol{s}_i;\boldsymbol{\beta})\big])\right|\xrightarrow{P}0$$

We should remark that this type of generic ULLN is not recent and has been discussed historically in non-spatial settings (Pötscher and Prucha (1994), Newey (1991)). The concept of stochastic equicontinuity even dates back to at least Billingsley (1968), whose book covered the convergence of stochastic processes. In fact, the only spatial aspect of Theorem 3.1.4 is the pointwise LLN in (3.9). If one can prove that (3.9) holds, then the ULLN holds. Theorem 3.1.1 can be used to establish a pointwise LLN for $\{q(\boldsymbol{s}_i;\boldsymbol{\beta}),\boldsymbol{s}_i\in D_n\}$. Indeed, assume that for each $\boldsymbol{\beta}\in B$, the $\{q(\boldsymbol{s}_i;\boldsymbol{\beta}),\boldsymbol{s}_i\in D_n\}$ have uniformly bounded moments as in (3.6) of Theorem 3.1.1. Since we assume that $(x_1(\boldsymbol{s}),....,x_m(\boldsymbol{s}),e(\boldsymbol{s}))$ is jointly α-mixing with mixing coefficient $\alpha(k,l,m)$, the mixing conditions in (3.6) of Theorem 3.1.1 are preserved for $\{q(\boldsymbol{s}_i;\boldsymbol{\beta}),\boldsymbol{s}_i\in D_n\}$ since it is a measurable transformation of the covariates and error. So by Theorem 3.1.1, the pointwise LLN in (3.9) holds for $\{q(\boldsymbol{s}_i;\boldsymbol{\beta}),\boldsymbol{s}_i\in D_n\}$. These arguments establish the following uniform LLN for our nonlinear estimator.

Proposition 3.1.2. *Let* $M_n(\boldsymbol{\beta})=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}q(\boldsymbol{s}_i;\boldsymbol{\beta})$ *denote the sequence of random functions given in (3.7). If the* $\{q(\boldsymbol{s}_i;\boldsymbol{\beta}),\boldsymbol{s}_i\in D_n\}$ *satisfy the mixing and moment conditions of Theorems 3.1.1, and the the stochastic equicontinuity and domination conditions of 3.1.4, then,*

$$\sup_{\boldsymbol{\beta}\in B}|M_n(\boldsymbol{\beta})-E\big[M_n(\boldsymbol{\beta})\big]|\xrightarrow{P}0$$

Therefore, the first condition of Lemma 3.1.2 is satisfied with the non-random function being $M(\boldsymbol{\beta}) = \lim_{n\to\infty} \frac{1}{n} \sum_{\boldsymbol{s}_i} E\big[q(\boldsymbol{s}_i; \boldsymbol{\beta})\big]$. Together with the identifiability condition of Lemma 3.1.2, this in turn, implies the consistency of the OLS estimator.

Corollary 3.1.5. *The OLS estimator* $\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} M_n(\boldsymbol{\beta})$ *is consistent.*

This finishes our discussion of step 1 in the FGLS estimation procedure. The next step is finding a consistent estimator of the covariance parameters $\boldsymbol{\theta}$.

3.2 Consistency and asymptotic normality of least squares variogram estimation

In this section, we show under some general conditions that the least squares variogram estimators in step 2 are consistent and asymptotically normal. In the previous section, we did not require the error to be intrinsically stationary, but we impose that condition here. These results stem from the work of Lahiri et al. (2002), later discussed by Crujeiras and van Keilegom (2010) for nonlinear trends. Recall from Section 2.2 the definition of the variogram of an intrinsically stationary random field. A classical non-parametric estimator of the variogram is the empirical variogram,

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2|N(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in N(\boldsymbol{h})} (\hat{\epsilon}(\boldsymbol{s}_i) - \hat{\epsilon}(\boldsymbol{s}_j))^2 \tag{3.10}$$

where $\hat{\epsilon}(\boldsymbol{s}_i), i = 1, ..., n$ are elements of the residual vector $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_p; \hat{\boldsymbol{\beta}}_{OLS})$ fitted from OLS and $N(\boldsymbol{h}) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j = \boldsymbol{h}\}$. Since our data are not assumed

to be regularly spaced, this estimator needs a slight modification,

$$\hat{\gamma}(\boldsymbol{h}) = \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in T(\boldsymbol{h})} (\hat{\epsilon}(\boldsymbol{s}_i) - \hat{\epsilon}(\boldsymbol{s}_j))^2 \tag{3.11}$$

where $T(\boldsymbol{h})$ is a neighborhood of $\boldsymbol{h}$. Cressie (1993) refers to this as a tolerance region. For instance, we can take our tolerance region to be

$$T(\boldsymbol{h}) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_i - \boldsymbol{s}_j \in B(\boldsymbol{h}, \delta_n), \text{ for some } \delta_n > 0\}, \quad \delta_n \to 0 \tag{3.12}$$

The empirical estimator $\hat{\gamma}(\boldsymbol{h})$ generally does not satisfy the non-positive definiteness property needed. To avoid this issue, it is common to assume a parametric form, $\gamma(\boldsymbol{h}; \boldsymbol{\theta})$ and estimate $\boldsymbol{\theta}$ using a least squares approach. Let $\boldsymbol{h}_i, i = 1, ..., K$ be a set of K lag-vectors and $\gamma(\boldsymbol{h}; \boldsymbol{\theta})$ a valid (non-positive definite) parametric family of variogram functions. Then, the least squares estimator is,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \boldsymbol{\ell}(\boldsymbol{\theta})^T \boldsymbol{V}(\boldsymbol{\theta}) \boldsymbol{\ell}(\boldsymbol{\theta}) \tag{3.13}$$

where $\boldsymbol{\ell}(\boldsymbol{\theta}) = (\hat{\gamma}(\boldsymbol{h}_1) - \gamma(\boldsymbol{h}_1; \boldsymbol{\theta}), ..., \hat{\gamma}(\boldsymbol{h}_K) - \gamma(\boldsymbol{h}_K; \boldsymbol{\theta}))^T$ and $\boldsymbol{V}(\boldsymbol{\theta})$ is a positive definite weight matrix. The choice of $\boldsymbol{V}(\boldsymbol{\theta})$ determines the asymptotic efficiency of this class of estimators. There are three common choices for the weight matrix, each having its own benefits and drawbacks:

1. Ordinary least squares (OLS) where, $\boldsymbol{V}(\boldsymbol{\theta})$ is the $K \times K$ identity matrix.

2. Weighted least squares (WLS) where, $\boldsymbol{V}(\boldsymbol{\theta}) = \text{diag}(w_1(\boldsymbol{\theta}), \cdots, w_K(\boldsymbol{\theta}))$ is a

diagonal matrix containing non-negative weights.

3. Generalized least squares (GLS) where, $\boldsymbol{V}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix of $(\hat{\gamma}(\boldsymbol{h}_1), ..., \hat{\gamma}(\boldsymbol{h}_K))^T$.

Under certain conditions, Lahiri et al. (2002) show that the GLS estimator is the optimal choice in the sense that limiting variance is the smallest among all estimators in the class (3.13). However, this optimality comes at the cost of inverting $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which often does not even have an exact expression. When the number of lag vectors K equals the dimension of the parameter space, the authors show that all three have the same asymptotic efficiency. Regardless of the choice of $\boldsymbol{V}(\boldsymbol{\theta})$, it is shown that the least squares estimator is consistent and asymptotically normal. Lahiri et al. (2002) give the result for a linear regression model, while Crujeiras and van Keilegom (2010) extends this to a nonlinear regression model with Gaussian errors. Both papers assume that the regressors are deterministic functions. Here, we assume stochastic regressors and remove the Gaussian assumption, imposing mixing conditions and bounded moments as before. For simplicity, we consider the ordinary least squares estimator,

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} (\hat{\gamma}(\boldsymbol{h}_k) - \gamma(\boldsymbol{h}_k; \boldsymbol{\theta}))^2 = \arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}) \tag{3.14}$$

since it is computationally the easiest to work with and still has desirable asymptotic properties. Note that this can be considered as another M-estimation type problem like the one in Lemma 3.1.2. First, the following conditions are imposed.

(C1) (*Identifiability* of $\boldsymbol{\theta}$) For all $\epsilon > 0$, there exists a $\mu > 0$ such that,

$$\inf_{||\boldsymbol{\theta}-\boldsymbol{\theta_0}||\geq\boldsymbol{\epsilon}} \sum_{k=1}^{K} (\gamma(\boldsymbol{h}_k;\boldsymbol{\theta}) - \gamma(\boldsymbol{h}_k;\boldsymbol{\theta}_0))^2 > \mu$$

(C2) (*Mixing and boundedness of moments of* $e(\boldsymbol{s})$) The error random field $e(\boldsymbol{s})$ is intrinsically stationary, has mean 0 and for some $\eta > 0$,

(a) $\sup_{\boldsymbol{s}_i \in D_n} \mathbb{E}\left[|e^2(\boldsymbol{s}_i)|^{2+\eta}\right] < \infty$

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

(c) $\sum_{m=1}^{\infty} m^{d-1}(\alpha(1,1,m))^{\frac{\eta}{2+\eta}} < \infty$

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

(C3) (*Regularity and boundedness of the trend function*) Let the parameter space B for $\boldsymbol{\beta}$ be compact in $\mathbb{R}^m$. The gradient of the trend function with respect to $\boldsymbol{\beta}$ exists and,

$$\mathbb{E}\left[\max_{j=1,..,m} \sup_{\boldsymbol{\beta}\in B} \left|\frac{\partial}{\partial\beta_j} f(x_1(\boldsymbol{s}),...,x_p(\boldsymbol{s}),\boldsymbol{\beta})\right|\right] < \infty$$

Moreover, the trend satisfies the following Lipschitz condition,

$$|f(x_1(\boldsymbol{s}),..,x_p(\boldsymbol{s});\boldsymbol{\beta}_1) - f(x_1(\boldsymbol{s}),..,x_p(\boldsymbol{s});\boldsymbol{\beta}_2)| \leq C(\boldsymbol{s})||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2||$$

where $C(\boldsymbol{s})$ satisfies $\sup_{\boldsymbol{s}} E\big[|C(\boldsymbol{s})|^2\big] < \infty$ and is the same for all $\boldsymbol{\beta}$.

(C4) (*Regularity of the variogram*) The variogram $\gamma(\boldsymbol{h};\boldsymbol{\theta})$ and its partial derivatives

$\frac{\partial \gamma(\boldsymbol{h};\boldsymbol{\theta})}{\partial h_k}, k = 1, ..., K$ are uniformly bounded in $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ and $\boldsymbol{h} \in \mathbb{R}^d$.

(C5) (*Rate of convergence of* $\hat{\boldsymbol{\beta}}_{OLS}$) $||\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0||^2 = o_p(\frac{1}{\sqrt{n}})$.

(C6) (*Size of tolerance region* $T(\boldsymbol{h})$) For every n, $|T(\boldsymbol{h})| > 0$ and $|T(\boldsymbol{h})| = o(n)$.

Condition (C1) ensures that a unique minimum exists in a neighborhood of the true $\boldsymbol{\theta}_0$. Condition (C2) is borrowed from Jenish and Prucha (2009). Under these conditions, they state the following CLT (given in Theorem 1 of their paper).

Theorem 3.2.1. *Let D_n be a sequence of sampling domains satisfying the increasing domain property. Let $\{z(\boldsymbol{s}_i), \boldsymbol{s}_i \in D_n\}$ be a realization of a zero mean random field, with mixing coefficient $\alpha(k, l, m)$ satisfying (C2). If $\lim_{n\to\infty} n^{-1}\sigma_n^2 > 0$ where $\sigma_n^2 = \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n} \text{Cov}(z(\boldsymbol{s}_i), z(\boldsymbol{s_j}))$, then,*

$$\frac{1}{\sqrt{n}} \sum_{\boldsymbol{s}_i \in D_n} z(\boldsymbol{s}_i) \xrightarrow{D} N(0, \lim_{n\to\infty} n^{-1}\sigma_n^2)$$

Condition (C3) is a regularity condition that is used in the proof of the asymptotic normality of the LS variogram estimator. It is an analog of condition (C.6) of Lahiri et al. (2002), who assume a deterministic linear trend. Condition (C4) is a smoothness condition on the variogram that can be directly verified for the choice of variogram model (e.g. exponential variogram). If the variogram and its gradient are continuous in $\boldsymbol{\theta}$ and Θ is a compact neighborhood of $\boldsymbol{\theta}_0$, then this condition holds. In the linear regression model, assumption (C5) is satisfied as seen in the proof of Proposition 3.1.1. In fact, we show in the proof that $\mathbb{E}[||\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0||^2] = O(\frac{1}{n})$, which implies (C5).

Before stating the consistency result of the variogram estimator, we require another lemma. First, define the quantity,

$$\gamma^*(\boldsymbol{h}) = \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in T(\boldsymbol{h})} (e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2 \tag{3.15}$$

This expression differs from the estimator $\hat{\gamma}(\boldsymbol{h})$ where the residuals $\hat{\epsilon}(\boldsymbol{s}_i)$ are replaced with the true errors $e(\boldsymbol{s}_i)$. This is an analog of Matheron method of moments variogram estimator (Cressie (1993)). However, since the errors are not observed, this is not exactly an estimator, but rather a theoretical tool.

Lemma 3.2.2. *Under conditions (C2), (C4) and (C6),* $\gamma^*(\boldsymbol{h}) \xrightarrow{P} \gamma^*(\boldsymbol{h}; \boldsymbol{\theta}_0)$.

The proof can be found in Section 3.5. Additional cardinality arguments needed for this lemma are supplied in the Appendix. We note that a result like this was proved in Davis and Borgman (1982), but for stationary, m-dependent Gaussian random fields on a regular lattice. With the above lemma on hand, the following result on consistency can be found in Proposition 3.1 of Crujeiras and van Keilegom (2010). The proof, given in Section 3.5, requires a bit of modification since the authors assume Gaussian errors as a way of bounding the moments. To relax the Gaussian assumption, we apply the moment conditions in (C2).

Proposition 3.2.1. *Denote* $\boldsymbol{\theta}_0$ *as the true covariance parameter. Under conditions (C1) - (C6),* $\hat{\boldsymbol{\theta}} \xrightarrow{P} \boldsymbol{\theta}_0$.

Lahiri et al. (2002) also give asymptotic normality results for $\hat{\boldsymbol{\theta}}$. The following result can be found in Theorem 3.3 of the aforementioned paper. The proof given

in Section 3.5 is modified to take into account the nonlinear trend and somewhat different assumptions used in (C1) - (C6). In addition, we use the spatial CLT in Theorem 3.2.1, which is slightly different than the CLT given in Lemma A.1 of Lahiri et al. (2002). The latter assumes stationarity and different mixing conditions.

Proposition 3.2.2. *Assume conditions (C1) - (C6) hold. Then for any set of K lag vectors, $\boldsymbol{h}_k, k = 1, .., K$,*

$$\sqrt{n}(\hat{\gamma}(\boldsymbol{h}_1) - \gamma(\boldsymbol{h}_1; \boldsymbol{\theta}_0), \cdots, \hat{\gamma}(\boldsymbol{h}_K) - \gamma(\boldsymbol{h}_K; \boldsymbol{\theta}_0))^T \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega}(\boldsymbol{\theta}_0))$$

where $\boldsymbol{\Omega}_{k\ell}(\boldsymbol{\theta}_0)$ *is* $\lim_{n\to\infty} \frac{1}{4n} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n} \mathrm{Cov}_{\boldsymbol{\theta}_0}((e(\boldsymbol{s}_i) - e(\boldsymbol{s}_i + \boldsymbol{h}_k))^2, (e(\boldsymbol{s}_j) - e(\boldsymbol{s}_j + \boldsymbol{h}_\ell))^2)$.

From Proposition 3.2.2, asymptotic normality of $\hat{\boldsymbol{\theta}}$ is then established in Corollary 3.1 of Lahiri et al. (2002) with a first order Taylor expansion argument, in the same vein as typical M-estimators. We state this result without proof, because it requires no modification in our context.

Corollary 3.2.3. *Assume conditions (C1) - (C6) hold. Then,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{B}(\boldsymbol{\theta}_0)^{-1} \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)^T \boldsymbol{\Omega}(\boldsymbol{\theta}_0) \boldsymbol{\Gamma}(\boldsymbol{\theta}_0) \boldsymbol{B}(\boldsymbol{\theta}_0)^{-1}) \tag{3.16}$$

where,

1. $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$ *is the* $K \times q$ *Jacobian matrix of the vector* $(\gamma(\boldsymbol{h}_1; \boldsymbol{\theta}), \cdots, \gamma(\boldsymbol{h}_K; \boldsymbol{\theta}))$ *evaluated at* $\boldsymbol{\theta} = \boldsymbol{\theta}_0$

2. $\boldsymbol{B}(\boldsymbol{\theta}_0) = \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)^T \boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$

3. $\boldsymbol{\Omega}(\boldsymbol{\theta}_0)$ *is given in Proposition 3.2.2.*

Finally, we re-estimate $\boldsymbol{\beta}$ using the FGLS estimator (3.2) in step 3, with $\hat{\boldsymbol{\theta}}$ serving as plug-in estimates for the covariance parameters.

3.3 Consistency and asymptotic normality of the FGLS estimator

For notational convenience, define $\boldsymbol{f}(\boldsymbol{\beta}) = \boldsymbol{f}(\boldsymbol{x}_1, ..., \boldsymbol{x}_m; \boldsymbol{\beta})$, but we stress that $\boldsymbol{f}$ contains the covariate vectors. Note that the estimator can be seen as the solution to another M-estimation problem,

$$\hat{\boldsymbol{\beta}}_{FGLS} = \arg\min_{\boldsymbol{\beta}} \frac{1}{n}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta}))^T \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})^{-1}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta})) := \arg\min_{\boldsymbol{\beta}} U_n(\boldsymbol{\beta}) \tag{3.17}$$

Define $\nabla f(\boldsymbol{\beta})$ to be the gradient of the trend function $f(x_1(\boldsymbol{s}), ..., x_m(\boldsymbol{s}); \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Let $\boldsymbol{J}(\boldsymbol{\beta}) = \dfrac{\partial \boldsymbol{f}}{\partial \boldsymbol{\beta}^T}$ be the $n \times m$ Jacobian matrix of the vector $\boldsymbol{f}(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. In words, the i^{th} row of $\boldsymbol{J}(\boldsymbol{\beta})$ is $\nabla f(\boldsymbol{\beta})^T$ evaluated at the location $\boldsymbol{s}_i, i = 1, ..., n$. In addition to (C1) - (C6), we impose the following conditions,

(C7) (*Identifiability of* $\boldsymbol{\beta}$) For all $\epsilon > 0$, there exists a $\delta > 0$ such that, $\inf_{||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| \geq \epsilon} R(\boldsymbol{\beta}) > \delta$ where,

$$R(\boldsymbol{\beta}) = \lim_{n \to \infty} \frac{1}{n}(\boldsymbol{f}(\boldsymbol{\beta}) - \boldsymbol{f}(\boldsymbol{\beta}_0))^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)(\boldsymbol{f}(\boldsymbol{\beta}) - \boldsymbol{f}(\boldsymbol{\beta}_0))$$

and the limit exists in probability as a non-stochastic quantity.

(C8) (*Regularity of the inverse covariance matrix*, $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$) The limit,

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}} \left\| \frac{\partial}{\partial \theta_k} \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \right\| < \infty, \quad k = 1, ..., q$$

where the norm $||\boldsymbol{A}||$ can either be the maximum absolute column sum or row sum of the matrix $\boldsymbol{A}$.

(C9) (*Existence and invertibility of asymptotic covariance*) The limiting asymptotic covariance matrix $\lim_{n\to\infty} \frac{1}{n} \boldsymbol{J}(\boldsymbol{\beta}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{J}(\boldsymbol{\beta}_0)$ exists in probability and is invertible.

Once again, condition (C7) is an identifiability assumption that ensures the limiting function is non-stochastic and has a unique minimum. Condition (C8) is a regularity condition on the inverse matrix that is needed to control the error in the first order Taylor expansion of $\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})$. The limit in condition (C9) is the inverse of the asymptotic variance of $\hat{\boldsymbol{\beta}}_{FGLS}$. Assuming mixing conditions and moment bounds on the covariates as in Theorem 3.1.1, this condition will hold. For example, in the linear regression case, this matrix has the form $\frac{1}{n} \boldsymbol{X}^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0) \boldsymbol{X}$. This can be seen as a weighted version of the law of large numbers involving the covariates. The extension to the nonlinear case is straightforward since the Jacobian $\boldsymbol{J}(\boldsymbol{\beta}_0)$ is just a measurable function of the covariates.

The following results on the consistency and asymptotic normality of $\hat{\boldsymbol{\beta}}_{FGLS}$ can be found in Crujeiras and van Keilegom (2010), in Proposition 3.2 and Theorem 3.3 respectively. Once again, we note that these authors use the Gaussian assump-

tion on the errors to establish these results. We avoid imposing this distributional assumption by using the spatial LLN and CLT of Theorem 3.2.1.

Proposition 3.3.1. *Under conditions (C1) - (C8),* $\hat{\boldsymbol{\beta}}_{FGLS} \xrightarrow{P} \boldsymbol{\beta}_0$.

Proposition 3.3.2. *Under conditions (C1) - (C9),* $\sqrt{n}(\hat{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{V}^{-1})$ *where* $\boldsymbol{V} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{J}(\boldsymbol{\beta}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\beta}_0)$ *and the limit exists in probability.*

The modified proofs can be found in Section 3.5. Before we proceed to the proofs, we briefly discuss spatial maximum likelihood estimation as an alternative to FGLS.

3.4 A note on spatial maximum likelihood estimation

If one does assume that the error is a Gaussian random field, we may employ the method of maximum likelihood to jointly estimate $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\theta} \in \mathbb{R}^q$. Sweeting (1980) established general conditions for the consistency and asymptotic normality of maximum likelihood estimators of a dependent sample of jointly Gaussian data. Previous papers did consider maximum likelihood estimation for dependent observations, but these were mainly in a time series context, where various techniques such as the martingale CLT could be employed (see for example, Bhat (1974) and Crowder (1976)). Mardia and Marshall (1984) later applied the results of Sweeting in a spatial linear regression model. Cressie and Lahiri (1993, 1996) similarly applied the results of Sweeting for REML estimation in a spatial linear regression model. For a nonlinear regression model $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}_1, .., \boldsymbol{x}_m; \boldsymbol{\beta}) + \boldsymbol{e}$, we may still use the results given in these papers. Due to the independence of $\boldsymbol{e}$ and $\boldsymbol{x}_1, ..., \boldsymbol{x}_m$ we

have,

$$\boldsymbol{y}|\boldsymbol{x}_1, ..., \boldsymbol{x}_p \sim N(\boldsymbol{f}(\boldsymbol{x}_1, .., \boldsymbol{x}_m; \boldsymbol{\beta}), \boldsymbol{\Sigma}(\boldsymbol{\theta}))$$

The second order partial derivatives (Hessian) of the likelihood can be calculated in closed form. First, define the following matrices,

$$\boldsymbol{\Sigma}_i(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i},\ \boldsymbol{\Sigma}^i(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})}{\partial \theta_i},\ \boldsymbol{\Sigma}_{ij}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{\Sigma}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j},\ \boldsymbol{\Sigma}^{ij}(\boldsymbol{\theta}) = \frac{\partial^2 \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}$$

Next, define $t_{ij} = \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}_i(\boldsymbol{\theta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}_j(\boldsymbol{\theta}))$ for $i, j = 1, .., q$ and $||\boldsymbol{A}|| = \text{tr}(\boldsymbol{A}^T\boldsymbol{A})$ as the Euclidean norm of a matrix $\boldsymbol{A}$. Next, denote by $\lambda_1 \leq \cdots \leq \lambda_n$ as the eigenvalues of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$. Finally, let $|\lambda_1^i| \leq \cdots \leq |\lambda_n^i|$ and $|\lambda_1^{ij}| \leq \cdots \leq |\lambda_n^{ij}|$ be the eigenvalues of $\boldsymbol{\Sigma}_i(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{ij}(\boldsymbol{\theta})$ respectively for $i, j = 1, ..., q$. After taking the expectation of the Hessian, the Fisher information can be calculated as the following block diagonal matrix,

$$\boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{I}_{\beta\beta} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{\theta\theta} \end{bmatrix}_{(p+q)\times(p+q)} \tag{3.18}$$

where $\boldsymbol{I}_{\beta\beta} = \boldsymbol{J}^T(\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{J}(\boldsymbol{\beta})$ and the ij^{th} element of $\boldsymbol{I}_{\theta\theta}$ is t_{ij}. We state the result here without proof, as it can be found in Theorem 2 of Mardia and Marshall (1984).

Theorem 3.4.1. *Suppose that as* $n \to \infty$,

(i) $\lim_{n\to\infty} \lambda_n = C < \infty$, $\lim_{n\to\infty} |\lambda_n^i| = C_i < \infty$, $\lim_{n\to\infty} |\lambda_n^{ij}| = C^{ij} < \infty$

(ii) $||\boldsymbol{\Sigma}_i(\boldsymbol{\theta})||^{-2} = O(n^{-1/2-\delta})$ *for some* $\delta > 0$ *for* $i = 1, ..., q$

(iii) $\lim_{n\to\infty} \frac{t_{ij}}{\sqrt{t_{ii}t_{jj}}} = a_{ij}$ *exists for all* $i, j = 1, ..., q$ *and the resulting matrix with* ij^{th} *entry* a_{ij} *is nonsingular*

(iv) $(\boldsymbol{J}^T(\boldsymbol{\beta})\boldsymbol{J}(\boldsymbol{\beta}))^{-1} \xrightarrow{P} \mathbf{0}$

Then the maximum likelihood estimator $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)^T$ *is consistent and* $\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{V}^{-1})$*, where* $\boldsymbol{V} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{I}(\boldsymbol{\beta}_0, \boldsymbol{\theta}_0)$ *and* $\boldsymbol{\phi}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\theta}_0^T)^T$ *are the true parameters.*

We note that the asymptotic variance of $\hat{\boldsymbol{\beta}}$ obtained from both FGLS estimation and maximum likelihood estimation is the same. This is expected since in FGLS estimation, we are minimizing the quadratic form (3.17) which resembles the quadratic form found in the Gaussian likelihood function; the only difference being that we are using the consistent least squares estimator $\hat{\boldsymbol{\theta}}$ in place of the true $\boldsymbol{\theta}$. Regarding estimation of $\boldsymbol{\theta}$, to our knowledge, there is no result in spatial statistics literature that definitively compares the asymptotic variances of the LS estimator and MLE. Zimmerman and Zimmerman (1991) determined through numerical simulations that the least squares method performs just as well computationally as the MLE. Their simulations were done with a constant mean Gaussian random field, with two different variogram models. We perform a similar study with a fitted trend in Chapter 4. We conjecture that the MLE in general has smaller variances. However, from a practical standpoint, least squares estimation may be preferred because the inversion of covariance matrices in maximum likelihood can be prohibitive for large sample sizes. If the only goal is estimation of the regression parameters, then FGLS is preferred for the computational advantage and no assumption of Gaussianity.

3.5 Proofs of results

Proof of Theorem 3.1.1 (L_2 part)

Proof. For $i = 1, ..., n$, let $y(\boldsymbol{s}_i) = z(\boldsymbol{s}_i) - \mathbb{E}\big[z(\boldsymbol{s}_i)\big]$. It suffices to prove that $\text{Var}\left(\frac{1}{n}\sum_{\boldsymbol{s}_i \in D_n} y(\boldsymbol{s}_i)\right)$ tends to 0. We have,

$$\begin{aligned}
\text{Var}\left(\frac{1}{n}\sum_{\boldsymbol{s}_i \in D_n} y(\boldsymbol{s}_i)\right) &= \frac{1}{n^2}\sum_{\boldsymbol{s}_i \in D_n} \text{Var}(y(\boldsymbol{s}_i)) + \frac{1}{n^2}\sum_{\boldsymbol{s}_i \neq \boldsymbol{s}_j \in D_n} \text{Cov}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_j)) \\
&\leq \frac{C}{n} + \frac{K_1}{n^2}\sum_{\boldsymbol{s}_i \neq \boldsymbol{s}_j \in D_n} (\alpha(1,1,d(\boldsymbol{s}_i, \boldsymbol{s}_j)))^{\frac{\eta}{2+\eta}} \\
&\leq \frac{C}{n} + \frac{K_2}{n^2}\sum_{\boldsymbol{s}_i \in D_n}\sum_{m=1}^{\infty} m^{d-1}(\alpha(1,1,m))^{\frac{\eta}{2+\eta}} \\
&= \frac{C}{n} + \frac{K_2}{n}\sum_{m=1}^{\infty} m^{d-1}(\alpha(1,1,m))^{\frac{\eta}{2+\eta}}
\end{aligned}$$

The constant C comes from the bounded moment condition in (3.6). The constant K_1 comes from the covariance inequality of Lemma 2.4.3 with $p = q = 2 + \eta$. The constant K_2 comes from Lemma A.1 (iii) of Appendix A in Jenish and Prucha (2009), which says that the number of points $\boldsymbol{s}_i$ that are within a distance m of $\boldsymbol{s}_j$ is $O(m^{d-1})$. Since the sum is bounded by assumption (3.6), the variance is $O(\frac{1}{n})$. □

Proof of Proposition 3.1.1

Proof. First, notice that the matrix, $\frac{1}{n}\sum_{\boldsymbol{s}_i} \boldsymbol{X}_i \boldsymbol{X}_i^T$ contains the upper left entry 1 because of the regression intercept. The remaining terms are of the form,

$$\frac{1}{n}\sum_{\boldsymbol{s}_i} x_j(\boldsymbol{s}_i), \qquad \frac{1}{n}\sum_{\boldsymbol{s}_i} x_j^2(\boldsymbol{s}_i), \qquad \frac{1}{n}\sum_{\boldsymbol{s}_i} x_j(\boldsymbol{s}_i)x_k(\boldsymbol{s}_i), \quad 1 \leq j,k \leq p-1$$

Next, the vector $\frac{1}{n}\sum_{\boldsymbol{s}_i} \boldsymbol{x}_i e_i$ contains terms of the form,

$$\frac{1}{n}\sum_{\boldsymbol{s}_i} e(\boldsymbol{s}_i), \qquad \frac{1}{n}\sum_{\boldsymbol{s}_i} e(\boldsymbol{s}_i)x_j(\boldsymbol{s}_i), \quad j = 1,..,p-1$$

All terms inside the sums above are measurable functions of the multivariate random field $(x_1(\boldsymbol{s}_i),...,x_{p-1}(\boldsymbol{s}_i),e(\boldsymbol{s}_i))^T, \boldsymbol{s}_i \in D_n$. So the σ-algebra generated by the above terms is contained within the σ-algebra generated by $(x_1(\boldsymbol{s}_i),...,x_{p-1}(\boldsymbol{s}_i),e(\boldsymbol{s}_i))^T$ for $\boldsymbol{s}_i \in D_n$. So, the mixing condition in (3.6) of Theorem 3.1.1 is preserved for each of these quantities. Then an application of Theorem 3.1.1 gives,

$$\begin{aligned} \frac{1}{n}\sum_{\boldsymbol{s}_i} x_j(\boldsymbol{s}_i) &\xrightarrow{L_2} \frac{1}{n}\sum_{\boldsymbol{s}_i} \mathbb{E}\big[x_j(\boldsymbol{s}_i)\big] \\ \lim_{n\to\infty}\frac{1}{n}\sum_{\boldsymbol{s}_i} x_j(\boldsymbol{s}_i)x_k(\boldsymbol{s}_i) &\xrightarrow{L_2} \lim_{n\to\infty}\frac{1}{n}\sum_{\boldsymbol{s}_i} \mathbb{E}\big[x_j(\boldsymbol{s}_i)x_k(\boldsymbol{s}_i)\big] \end{aligned}$$

By assumption, the limiting $p \times p$ matrix $\lim_{n\to\infty}\frac{1}{n}\sum_{\boldsymbol{s}_i} \mathbb{E}\big[\boldsymbol{X}_i\boldsymbol{X}_i^T\big]$ exists and is invertible. Next, since $e(\boldsymbol{s})$ is centered at 0, Theorem 3.1.1 gives,

$$\frac{1}{n}\sum_{\boldsymbol{s}_i} e(\boldsymbol{s}_i) \xrightarrow{L_2} 0, \qquad \frac{1}{n}\sum_{\boldsymbol{s}_i} x_j(\boldsymbol{s}_i)e(\boldsymbol{s}_i) \xrightarrow{L_2} 0, \quad j = 1,...,p$$

and so the vector $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{e}$ converges to the vector $\mathbf{0}$ in L_2. So by the continuous mapping and Slutsky's theorems, $\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 \xrightarrow{P} \mathbf{0}$. □

Proof of Lemma 3.2.2

Proof. First, recall the definition of the tolerance region $T(\boldsymbol{h}) = \{\{\boldsymbol{s}_i, \boldsymbol{s}_j\} \subset D_n : \boldsymbol{s}_i - \boldsymbol{s}_j \in B(\boldsymbol{h}, \delta_n)\}$, where δ_n is a sequence of decreasing positive numbers converging to 0. Note that this estimator is not exactly unbiased, unlike in the regularly spaced lattice case since,

$$\mathbb{E}\big[\gamma^*(\boldsymbol{h})\big] = \frac{1}{2|T(\boldsymbol{h})|} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in T(\boldsymbol{h})} \mathbb{E}\big[(e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2\big] = \frac{1}{|T(\boldsymbol{h})|} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in T(\boldsymbol{h})} \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j, \boldsymbol{\theta}_0)$$

and the difference $\boldsymbol{s}_i - \boldsymbol{s}_j$ is not necessarily equal to $\boldsymbol{h}$. However, by assumption (C4), the variogram is differentiable with bounded gradient on $\mathbb{R}^n$. Then an application of the mean value theorem gives $\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j, \boldsymbol{\theta}_0) = \gamma(\boldsymbol{h}, \boldsymbol{\theta}_0) + \nabla\gamma(\boldsymbol{u}, \boldsymbol{\theta}_0)^T(\boldsymbol{s}_i - \boldsymbol{s}_j - \boldsymbol{h})$, where $\boldsymbol{u}$ lies on the line segment between $\boldsymbol{s}_i - \boldsymbol{s}_j$ and $\boldsymbol{h}$. The second term converges to 0 since $\boldsymbol{s}_i - \boldsymbol{s}_j \in B(\boldsymbol{h}, \delta_n)$. Therefore, it is asymptotically unbiased and so by Chebyshev's inequality, it suffices to show that the variance,

$$\frac{1}{4|T(\boldsymbol{h})|^2} \sum_{i,j \in T(\boldsymbol{h})} \sum_{k,l \in T(\boldsymbol{h})} \mathrm{Cov}((e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2, (e(\boldsymbol{s}_k) - e(\boldsymbol{s}_l))^2) \tag{3.19}$$

goes to 0 in the limit.

Note that the expression for the variance of $\gamma^*(\boldsymbol{h})$ contains terms of the form, $\mathrm{Cov}((e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2, (e(\boldsymbol{s}_k) - e(\boldsymbol{s}_l))^2)$. Since all the terms inside the covariance are measurable functions of $\{e(\boldsymbol{s}_i), \boldsymbol{s}_i \in D_n\}$, the α-mixing conditions are preserved for these quantities. Moreover, in light of condition (C2) and the inequality $|a - b|^p \leq$

$2^{p-1}(|a|^2 + |b|^2)$ for $a, b \in \mathbb{R}, p \geq 1$, we have,

$$\sup_{\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n} \mathbb{E}\big[|(e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2|^{2+\eta}| \leq C \left[\sup_{\boldsymbol{s}_i \in D_n} \mathbb{E}\big[|e(\boldsymbol{s}_i)^2|^{2+\eta}\big] + \sup_{\boldsymbol{s}_j \in D_n} \mathbb{E}\big[|e(\boldsymbol{s}_j)^2|^{2+\eta}\big] \right]$$

where $C = 2^{3+2\eta}$. Thus, the squared differences $(e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2$ will satisfy the covariance inequality in Lemma 2.4.3 with $p = q = 2 + \eta$ and $r = \dfrac{2+\eta}{\eta}$. Next, partition $T(\boldsymbol{h})$ into two sets. Define A_n be the set of pairs of points $\{\boldsymbol{s}_i, \boldsymbol{s}_j\}$ and $\{\boldsymbol{s}_k, \boldsymbol{s}_l\}$ in $T(\boldsymbol{h})$ with minimum distance between them greater than some arbitrary $M \in \mathbb{N}$, that is,

$$A_n = \{\{\boldsymbol{s}_i, \boldsymbol{s}_j\}, \{\boldsymbol{s}_k, \boldsymbol{s}_l\} \subset T(\boldsymbol{h}) : d(\{\boldsymbol{s}_i, \boldsymbol{s}_j\}, \{\boldsymbol{s}_k, \boldsymbol{s}_l\}) > M\}$$

Then, the variance expression in (3.19) becomes,

$$\text{Var}(\gamma^*(\boldsymbol{h})) = \frac{1}{4|T_n(\boldsymbol{h})|^2} G_n + \frac{1}{4|T_n(\boldsymbol{h})|^2} H_n \tag{3.20}$$

where, $G_n = \sum_{A_n} \text{Cov}((e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2, (e(\boldsymbol{s}_k) - e(\boldsymbol{s}_l))^2)$ and H_n is defined similarly on A_n^c. For H_n, after applying the Cauchy-Schwarz inequality, condition (C2) implies that the covariance terms are uniformly bounded. So there exists a constant C such that,

$$H_n \leq C|A_n^c| = C|\{\{\boldsymbol{s}_i, \boldsymbol{s}_j\}, \{\boldsymbol{s}_k, \boldsymbol{s}_l\} \subset T(\boldsymbol{h}) : d(\{\boldsymbol{s}_i, \boldsymbol{s}_j\}, \{\boldsymbol{s}_k, \boldsymbol{s}_l\}) \leq M\}|$$

For a fixed pair, $\{\boldsymbol{s}_i, \boldsymbol{s}_j\}$, the number of pairs $\{\boldsymbol{s}_k, \boldsymbol{s}_l\}$ within a distance of M of

$\{\boldsymbol{s}_i, \boldsymbol{s}_j\}$ is $O(M^d)$ by Lemma 3.5.2 of the Appendix. Since there are $|T(\boldsymbol{h})|$ such fixed pairs, we have the upper bound, $H_n \leq C|A_n^c| \leq KM^d|T(\boldsymbol{h})|$ for some constant K. So the second term of (3.20) above goes to 0. Now, consider the first term, G_n. In Lemma 2.4.3, take $A = \{\boldsymbol{s}_i, \boldsymbol{s}_j\}$ and $B = \{\boldsymbol{s}_k, \boldsymbol{s}_l\}$. Then, together with condition (C2), the covariance terms on the set A_n are bounded by,

$$\text{Cov}((e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2, (e(\boldsymbol{s}_k) - e(\boldsymbol{s}_l))^2) \leq C(\alpha(2, 2, d(A, B)))^{\frac{\eta}{2+\eta}}$$

for some C independent of n. It follows that,

$$\begin{aligned} G_n &= \sum_{A_n} \text{Cov}((e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2, (e(\boldsymbol{s}_k) - e(\boldsymbol{s}_l))^2) \\ &\leq C \sum_{A_n} (\alpha(2, 2, d(A, B)))^{\frac{\eta}{2+\eta}} \\ &\leq C \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in T(\boldsymbol{h})} \sum_{m=M}^{\infty} \sum_{d(A,B) \in (m, m+1]} (\alpha(2, 2, m))^{\frac{\eta}{2+\eta}} \\ &\leq K|T(\boldsymbol{h})| \sum_{m=M}^{\infty} m^{d-1} (\alpha(2, 2, m))^{\frac{\eta}{2+\eta}} \end{aligned}$$

where K is independent of n. From the second to third line, we used the fact that

$$A_n \subset \bigcup_{m=M}^{\infty} \{\{\boldsymbol{s}_i, \boldsymbol{s}_j\}, \{\boldsymbol{s}_k, \boldsymbol{s}_l\} \in T(\boldsymbol{h}) : d(A, B) \in (m, m + 1]\}$$

and that $\alpha(2, 2, d(A, B)) \leq \alpha(2, 2, m)$ for $d(A, B) \in [m, m + 1)$. For the third to fourth line, note that for a fixed pair $\{\boldsymbol{s}_i, \boldsymbol{s}_j\}$, the number of pairs $\{\boldsymbol{s}_k, \boldsymbol{s}_l\}$ that satisfy $d(\{\boldsymbol{s}_i, \boldsymbol{s}_j\}, \{\boldsymbol{s}_k, \boldsymbol{s}_l\}) \in (m, m + 1]$ is $O(m^{d-1})$ by Lemma 3.5.3 of the Appendix. By conditions (C2), (C6), the first term on the right hand side of (3.20) goes to 0. □

Proof of Proposition 3.2.1

Proof. Define the following quantities,

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{K} (\gamma(\boldsymbol{h}_i; \boldsymbol{\theta}_0) - \gamma(\boldsymbol{h}_i; \boldsymbol{\theta}))^2, \quad \hat{Q}_n(\boldsymbol{\theta}) = \sum_{i=1}^{K} (\hat{\gamma}(\boldsymbol{h}_i) - \gamma(\boldsymbol{h}_i; \boldsymbol{\theta}))^2$$

To prove the proposition, we apply Lemma 3.1.2 to $M_n = \hat{Q}_n$ and $M = Q$. The second condition of the lemma is satisfied by (C1). It remains to prove that $\sup_{\boldsymbol{\theta} \in \Theta} |\hat{Q}_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \xrightarrow{P} 0$. By the triangle inequality,

$$|\hat{Q}_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \leq |\hat{Q}_n(\boldsymbol{\theta}) - Q_n^*(\boldsymbol{\theta})| + |Q_n^*(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \tag{3.21}$$

where $Q_n^*(\boldsymbol{\theta}) = \sum_{i=1}^{K} (\gamma^*(\boldsymbol{h}_i) - \gamma(\boldsymbol{h}_i; \boldsymbol{\theta}))^2$ and $\gamma^*(\boldsymbol{h}) = \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in T(\boldsymbol{h})} (e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2$. By Lemma 3.2.2, $\gamma^*(\boldsymbol{h}) \xrightarrow{P} \gamma(\boldsymbol{h}; \boldsymbol{\theta}_0)$ and thus, $\sup_{\boldsymbol{\theta} \in \Theta} |Q_n^*(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \xrightarrow{P} 0$, since the variogram is uniformly bounded (C4). We now show that the first term in (3.21) converges to 0 in probability uniformly in $\boldsymbol{\theta}$. By an identical algebraic manipulation as that of Crujeiras and van Keilegom (2010), we have,

$$\hat{Q}_n(\boldsymbol{\theta}) = Q_n^*(\boldsymbol{\theta}) + \sum_{i=1}^{K} (A_n(\boldsymbol{h}_i) + B_n(\boldsymbol{h}_i))^2 - 2 \sum_{i=1}^{K} (\gamma(\boldsymbol{h}_i; \boldsymbol{\theta}) - \gamma^*(\boldsymbol{h}_i))(A_n(\boldsymbol{h}_i) + B_n(\boldsymbol{h}_i))$$

where, after defining $g(\boldsymbol{s}; \boldsymbol{\beta}) = f(\boldsymbol{s}; \boldsymbol{\beta}_0) - f(\boldsymbol{s}; \boldsymbol{\beta})$,

$$A_n(\boldsymbol{h}) = \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_j) \in T(\boldsymbol{h})} [g(\boldsymbol{s}_i; \boldsymbol{\beta}) - g(\boldsymbol{s}_j; \boldsymbol{\beta})]^2$$

$$B_n(\boldsymbol{h}) = \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in T(\boldsymbol{h})} [e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j)]\,[g(\boldsymbol{s}_i;\boldsymbol{\beta}) - g(\boldsymbol{s}_j;\boldsymbol{\beta})]$$

Therefore,

$$\begin{aligned}\sup_{\boldsymbol{\theta}\in\Theta} |\hat{Q}_n(\boldsymbol{\theta}) - Q_n^*(\boldsymbol{\theta})| &\leq \sum_{i=1}^{K} (A_n(\boldsymbol{h}_i) + B_n(\boldsymbol{h}_i))^2 \\ &+ 2\sum_{i=1}^{K} \sup_{\boldsymbol{\theta}\in\Theta} |(\gamma(\boldsymbol{h}_i;\boldsymbol{\theta}) - \gamma^*(\boldsymbol{h}_i))(A_n(\boldsymbol{h}_i) + B_n(\boldsymbol{h}_i))|\end{aligned}$$

In order to show that this term converges to 0 in probability, it suffices to show that $A_n(\boldsymbol{h}_i)$ and $B_n(\boldsymbol{h}_i)$ both converge to 0 in probability for each $i = 1, .., K$. First, by the mean value theorem applied to $g(\boldsymbol{s}_i;\boldsymbol{\beta}) - g(\boldsymbol{s}_j;\boldsymbol{\beta})$, we have $g(\boldsymbol{s}_i;\boldsymbol{\beta}) - g(\boldsymbol{s}_j;\boldsymbol{\beta}) = \nabla G^T(\bar{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0)$, where

$$\nabla G^T(\bar{\boldsymbol{\beta}}) = \left(\frac{\partial g(\boldsymbol{s}_i;\bar{\boldsymbol{\beta}})}{\partial\beta_1} - \frac{\partial g(\boldsymbol{s}_j;\bar{\boldsymbol{\beta}})}{\partial\beta_1}, \cdots, \frac{\partial g(\boldsymbol{s}_i;\bar{\boldsymbol{\beta}})}{\partial\beta_m} - \frac{\partial g(\boldsymbol{s}_j;\bar{\boldsymbol{\beta}})}{\partial\beta_m}\right)$$

and $\bar{\boldsymbol{\beta}}$ lies on the line segment between $\hat{\boldsymbol{\beta}}_{OLS}$ and $\boldsymbol{\beta}_0$. Then for the $A_n(\boldsymbol{h}_i)$ terms,

$$\begin{aligned}A_n(\boldsymbol{h}) &= \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in T(\boldsymbol{h})} \left[\nabla G^T(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{OLS})\right]^2 \\ &= (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{OLS})^T \left[\frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in T(\boldsymbol{h})} \nabla G^T(\bar{\boldsymbol{\beta}})\nabla G^T(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{OLS})\right]\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta}_0$ and condition (C3) on the uniform boundedness of the trend, the above converges to 0 in probability for any $\boldsymbol{h}_i, i = 1, ..., K$. For the $B_n(\boldsymbol{h}_i)$ terms,

$$
\begin{aligned}
B_n(\boldsymbol{h}) &= \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in T(\boldsymbol{h})} (e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))\nabla G^T(\bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{OLS}) \\
&= (\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}_{OLS})^T \frac{1}{2|T(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i,\boldsymbol{s}_j)\in T(\boldsymbol{h})} (e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))\nabla G(\bar{\boldsymbol{\beta}})
\end{aligned}
$$

The fact that $\hat{\boldsymbol{\beta}}_{OLS} \xrightarrow{P} \boldsymbol{\beta}_0$, together with conditions (C2, C3) imply that the above converges to 0 in probability for any $\boldsymbol{h}_i, i = 1, ..., K$. □

Proof of Proposition 3.2.2

Proof. By the Cramér-Wold device, it suffices to prove that for any $\boldsymbol{a} = (a_1, ..., a_K)^T \in \mathbb{R}^K$, the quantity,

$$
\begin{aligned}
T_n &= \sqrt{n}\sum_{k=1}^{K} a_k(\hat{\gamma}(\boldsymbol{h}_k) - \gamma(\boldsymbol{h}_k;\boldsymbol{\theta}_0)) \\
&= \sqrt{n}\sum_{k=1}^{K} a_k \left\{ \frac{1}{2|T_n(\boldsymbol{h}_k)|} \sum_{\boldsymbol{s}_i,\boldsymbol{s}_j} (\hat{\epsilon}(\boldsymbol{s}_i) - \hat{\epsilon}(\boldsymbol{s}_j))^2 - \gamma(\boldsymbol{h}_k;\boldsymbol{\theta}_0)) \right\}
\end{aligned}
$$

converges in distribution to $N(0, \boldsymbol{a}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\boldsymbol{a})$. Define the quantities,

$$
T_{1n} = \sqrt{n}\sum_{k=1}^{K} a_k \left\{ \frac{1}{2|T_n(\boldsymbol{h}_k)|} \sum_{T_n(\boldsymbol{h}_k)} (e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2 - \gamma(\boldsymbol{h}_k;\boldsymbol{\theta}_0)) \right\}
$$

$$
T_{2n} = \sqrt{n}\sum_{k=1}^{K} a_k \left\{ \frac{1}{2n} \sum_{\boldsymbol{s}_i} (e(\boldsymbol{s}_i) - e(\boldsymbol{s}_i + \boldsymbol{h}_k))^2 - \gamma(\boldsymbol{h}_k;\boldsymbol{\theta}_0)) \right\}
$$

The main idea of the proof in Lahiri et al. (2002) is that T_n is well approximated by T_{2n} for large n and that T_{2n} has the desired asymptotic normality. First we

show that $|T_n - T_{1n}| \xrightarrow{P} 0$ and $|T_{1n} - T_{2n}| \xrightarrow{P} 0$ and then conclude by showing that $T_{2n} \xrightarrow{D} N(0, \boldsymbol{a}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\boldsymbol{a})$. Letting $g(\boldsymbol{s}; \boldsymbol{\beta}) = f(\boldsymbol{s}; \boldsymbol{\beta}_0) - f(\boldsymbol{s}; \boldsymbol{\beta})$, we have,

$$\begin{aligned}
|T_n - T_{1n}| &\leq \sqrt{n} \sum_{k=1}^{K} \frac{|a_k|}{2|T_n(\boldsymbol{h}_k)|} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j} |(\hat{\epsilon}(\boldsymbol{s}_i) - \hat{\epsilon}(\boldsymbol{s}_j))^2 - (e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))^2| \\
&= \sqrt{n} \sum_{k=1}^{K} \frac{|a_k|}{2|T_n(\boldsymbol{h}_k)|} \Bigg\{ \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j} |(g(\boldsymbol{s}_i; \hat{\boldsymbol{\beta}}_{OLS}) - g(\boldsymbol{s}_j; \hat{\boldsymbol{\beta}}_{OLS}))^2 \\
&\quad + \; 2(g(\boldsymbol{s}_i; \hat{\boldsymbol{\beta}}_{OLS}) - g(\boldsymbol{s}_j; \hat{\boldsymbol{\beta}}_{OLS}))(e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))| \Big\} \\
&\leq \sqrt{n} \sum_{k=1}^{K} |a_k| \left\{ A_{kn} ||\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0||^2 + B_{kn} ||\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0|| \right\}
\end{aligned}$$

where we defined A_{kn}, B_{kn} as (recalling the Lipschitz property of the trend in (C3)),

$$\begin{aligned}
A_{kn} &= \frac{1}{2|T_n(\boldsymbol{h}_k)|} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j} (C^2(\boldsymbol{s}_i) + C^2(\boldsymbol{s}_j)) \\
B_{kn} &= \frac{1}{|T_n(\boldsymbol{h}_k)|} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j} (C(\boldsymbol{s}_i) + C(\boldsymbol{s}_j))(e(\boldsymbol{s}_i) - e(\boldsymbol{s}_j))
\end{aligned}$$

Since $||\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0||^2 = o_p(\frac{1}{\sqrt{n}})$, it suffices to show that A_{kn} and B_{kn} are bounded in probability for $k = 1, ..., K$. This follows from the uniform boundedness of the errors in condition (C2) and of the Lipschitz random variables $C(\boldsymbol{s})$ in (C3). Thus, $|T_n - T_{1n}| \xrightarrow{P} 0$.

Next, we heuristically reason why $|T_{1n} - T_{2n}| \xrightarrow{P} 0$. We note that the indices of the two sums in T_{1n} and T_{2n} differ slightly. For T_{1n}, we are summing over $T_n(\boldsymbol{h}_k) = \{(\boldsymbol{s}_i, \boldsymbol{s}_j) : \boldsymbol{s}_j \in B_{\delta_n}(\boldsymbol{s}_i + \boldsymbol{h}_k)\}$, which is the tolerance region as defined before. For T_{2n}, we are summing over all pairs of the form $\{(\boldsymbol{s}_i, \boldsymbol{s}_i + \boldsymbol{h}_k) : i = 1, ..., n\}$. Since $\delta_n \to 0$ in the tolerance region, the difference between $T_n(\boldsymbol{h}_k)$ and $\{(\boldsymbol{s}_i, \boldsymbol{s}_i + \boldsymbol{h}_k) : i = 1, ..., n\}$

becomes negligible for large n and thus, $|T_{1n} - T_{2n}| \xrightarrow{P} 0$. For more details in this step of the proof, we refer to p. 82 of Lahiri et al. (2002).

Finally, we show that T_{2n} is asymptotically normal. We can apply the CLT result in Theorem 3.2.1. Write T_{2n} as,

$$T_{2n} = \sqrt{n}\sum_{k=1}^{K} a_k \left\{ \frac{1}{2n}\sum_{\boldsymbol{s}_i}(e(\boldsymbol{s}_i) - e(\boldsymbol{s}_i + \boldsymbol{h}_k))^2 - \gamma(\boldsymbol{h}_k, \boldsymbol{\theta}_0)) \right\} = \frac{1}{\sqrt{n}}\sum_{\boldsymbol{s}_i} z(\boldsymbol{s}_i)$$

where $z(\boldsymbol{s}_i) = \sum_{k=1}^{K} a_k \left\{ \frac{(e(\boldsymbol{s}_i) - e(\boldsymbol{s}_i + \boldsymbol{h}_k))^2}{2} - \gamma(\boldsymbol{h}_k, \boldsymbol{\theta}_0) \right\}$. Then, by Theorem 3.2.1,

$$\frac{1}{\sqrt{n}}\sum_{\boldsymbol{s}_i} z(\boldsymbol{s}_i) \xrightarrow{D} N\left(0, \lim_{n\to\infty} n^{-1} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n} \mathrm{Cov}(z(\boldsymbol{s}_i), z(\boldsymbol{s}_j))\right)$$

The covariance term can be calculated as,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n} \mathrm{Cov}(z(\boldsymbol{s}_i), z(\boldsymbol{s}_j)) &= \sum_{i=1}^{K}\sum_{j=1}^{K} a_k a_\ell \boldsymbol{\Sigma}_{k\ell}(\boldsymbol{\theta}_0) \\ &= \boldsymbol{a}^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\boldsymbol{a} \end{aligned}$$

where $\boldsymbol{\Sigma}_{k\ell}(\boldsymbol{\theta}_0) = \lim_{n\to\infty} \frac{1}{4n} \sum_{\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n} \mathrm{Cov}((e(\boldsymbol{s}_i) - e(\boldsymbol{s}_i + \boldsymbol{h}_k))^2, (e(\boldsymbol{s}_j) - e(\boldsymbol{s}_j + \boldsymbol{h}_\ell))^2)$. So asymptotic normality of T_{2n} holds. □

Proof of Proposition 3.3.1

Proof. By Lemma 3.1.2, consistency follows if we prove that,

$$\forall \epsilon > 0, \exists \nu > 0, \text{s.t.} \inf_{||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| > \epsilon} |U(\boldsymbol{\beta}) - U(\boldsymbol{\beta}_0)| > \nu$$

$$\sup_{\boldsymbol{\beta}} |U_n(\boldsymbol{\beta}) - U(\boldsymbol{\beta})| \xrightarrow{P} 0$$

where $U(\boldsymbol{\beta}) = 1 + R(\boldsymbol{\beta})$. The first part follows from condition (C7) and the fact that $R(\boldsymbol{\beta})$ is uniquely minimized at $\boldsymbol{\beta}_0$. It remains to prove the second part. By an identical decomposition as that of Crujeiras and van Keilegom (2010), we have, $U_n(\boldsymbol{\beta}) = U_{n1} + U_{n2}(\boldsymbol{\beta}) + U_{n3}(\boldsymbol{\beta})$, where,

$$\begin{aligned} U_{n1} &= \frac{1}{n}\boldsymbol{e}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{e} \\ U_{n2}(\boldsymbol{\beta}) &= \frac{2}{n}\boldsymbol{e}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}(\boldsymbol{f}(\boldsymbol{\beta}_0) - \boldsymbol{f}(\boldsymbol{\beta})) \\ U_{n3}(\boldsymbol{\beta}) &= \frac{1}{n}(\boldsymbol{f}(\boldsymbol{\beta}_0) - \boldsymbol{f}(\boldsymbol{\beta}))^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}(\boldsymbol{f}(\boldsymbol{\beta}_0) - \boldsymbol{f}(\boldsymbol{\beta})) \end{aligned}$$

So it follows that

$$\sup_{\boldsymbol{\beta}} |U_n(\boldsymbol{\beta}) - U(\boldsymbol{\beta})| \leq |U_{n1} - 1| + \sup_{\boldsymbol{\beta}} |U_{n2}(\boldsymbol{\beta})| + \sup_{\boldsymbol{\beta}} |U_{n3}(\boldsymbol{\beta}) - R(\boldsymbol{\beta})| \qquad (3.22)$$

and we show that these three terms converge to 0 in probability. By the Mean Value Theorem, we can expand U_{n1} as,

$$U_{n1} = \frac{1}{n}\boldsymbol{e}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{e} + \frac{1}{n}\sum_{i=1}^{q}(\hat{\theta}_i - \theta_{0i})\boldsymbol{e}^T\frac{\partial\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\theta}})}{\partial\theta_i}\boldsymbol{e} \qquad (3.23)$$

where $\bar{\boldsymbol{\theta}}$ lies on the line segment between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$. Since $\hat{\theta}_i \xrightarrow{P} \theta_{0i}$, the second term of (3.23) converges to 0 in probability if we show that $\frac{1}{n}\boldsymbol{e}^T\frac{\partial\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\theta}})}{\partial\theta_i}\boldsymbol{e}$ is bounded in

probability. Denote $c_{jk}(\bar{\boldsymbol{\theta}})$ as the jk^{th} element of $\dfrac{\partial \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\theta}})}{\partial \theta_i}$. By condition (C2),

$$\mathbb{E}\left[\left|\frac{1}{n}\boldsymbol{e}^T\frac{\partial \boldsymbol{\Sigma}(\bar{\boldsymbol{\theta}})}{\partial \theta_i}\boldsymbol{e}\right|\right] = \mathbb{E}\left[\frac{1}{n}\sum_{\boldsymbol{s}_j}\sum_{\boldsymbol{s}_k}|c_{jk}(\bar{\boldsymbol{\theta}})e(\boldsymbol{s}_j)e(\boldsymbol{s}_k)|\right] \leq C\sup_{\boldsymbol{\theta}}\left\|\frac{\partial}{\partial \theta_k}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\right\|$$

for some constant $C > 0$. Thus, $\dfrac{1}{n}\boldsymbol{e}^T\dfrac{\partial \boldsymbol{\Sigma}(\bar{\boldsymbol{\theta}})}{\partial \theta_i}\boldsymbol{e}$ is bounded in probability by (C8). For the first term of (3.23), write,

$$\frac{1}{n}\boldsymbol{e}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{e} = \frac{1}{n}\boldsymbol{z}^T\boldsymbol{z} = \frac{1}{n}\sum_{\boldsymbol{s}_i} z^2(\boldsymbol{s}_i)$$

where $\boldsymbol{z} = \boldsymbol{L}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{e}$ and $\boldsymbol{L}$ comes from the Cholesky factorization of $\boldsymbol{\Sigma}$. If we assume that $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}_0)$, then $\{y(\boldsymbol{s}_i), \boldsymbol{s}_i \in D_n\}$ are i.i.d. $N(0,1)$ and a standard LLN can be applied. Otherwise for non-Gaussian errors, note that $\{y(\boldsymbol{s}_i), \boldsymbol{s}_i \in D_n\}$ is a linear transformation of the errors $\{e(\boldsymbol{s}_i), \boldsymbol{s}_i \in D_n\}$. Thus the uniform boundedness and mixing properties of the error are preserved. Since $E\big[y(\boldsymbol{s}_i)\big] = 1$ for each $i = 1, ..., n$, Theorem 3.1.1 implies that,

$$\frac{1}{n}\boldsymbol{e}^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{e} = \frac{1}{n}\sum_{\boldsymbol{s}_i} y^2(\boldsymbol{s}_i) \xrightarrow{P} 1$$

which shows that $|U_{n1} - 1| \xrightarrow{P} 0$ and so the first term of (3.22) goes to 0 in probability.

Next, we consider the term $U_{n2}(\boldsymbol{\beta})$ in (3.22). By the Mean Value Theorem, we can expand $U_{n2}(\boldsymbol{\beta})$ as,

$$U_{n2}(\boldsymbol{\beta}) = \frac{2}{n}(\boldsymbol{f}(\boldsymbol{\beta}_0) - \boldsymbol{f}(\boldsymbol{\beta}))^T\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{e} + \frac{2}{n}\sum_{i=1}^{q}(\hat{\theta}_i - \theta_{0i})(\boldsymbol{f}(\boldsymbol{\beta}_0) - \boldsymbol{f}(\boldsymbol{\beta}))^T\frac{\partial \boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{\theta}})}{\partial \theta_i}\boldsymbol{e}$$

where $\bar{\boldsymbol{\theta}}$ lies on the line segment between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. By similar arguments as in the preceding paragraph for U_{n1}, both terms above converge to zero in probability. Thus, $|U_{n2}(\boldsymbol{\beta})| \xrightarrow{P} 0$ pointwise for any $\boldsymbol{\beta}$. Since the parameter space B is compact and trend is smooth, convergence is uniform. Thus, the second term of (3.22) converges to zero in probability.

Finally, for $U_{n3}(\boldsymbol{\beta})$, it suffices again to consider the term,

$$\frac{1}{n}(\boldsymbol{f}(\boldsymbol{\beta}_0) - \boldsymbol{f}(\boldsymbol{\beta}))^T \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)^{-1}(\boldsymbol{f}(\boldsymbol{\beta}_0) - \boldsymbol{f}(\boldsymbol{\beta}))$$

where we replaced $\hat{\boldsymbol{\theta}}$ with the true parameter $\boldsymbol{\theta}_0$ since it can be shown by a Taylor expansion argument that these terms are asymptotically equivalent. By definition of $R(\boldsymbol{\beta})$, the third term of (3.22) goes to 0 in probability. □

Proof of Proposition 3.3.2

Proof. After taking the gradient of $U_n(\boldsymbol{\beta}) = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta}))^T \boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})(\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{\beta}))$ with respect to $\boldsymbol{\beta}$, and applying the Mean Value Theorem, Crujeiras and van Keilegom (2010) arrive at the following expression,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}_0) = \left(\frac{1}{n}\boldsymbol{J}^T(\hat{\boldsymbol{\beta}}_{FGLS})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{J}(\bar{\boldsymbol{\beta}})\right)^{-1} \frac{1}{\sqrt{n}}\boldsymbol{J}^T(\hat{\boldsymbol{\beta}}_{FGLS})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{e} \quad (3.24)$$

where $\bar{\boldsymbol{\beta}}$ lies on the line segment between $\hat{\boldsymbol{\beta}}_{FGLS}$ and $\boldsymbol{\beta}_0$. By assumption (C9), the limit $\boldsymbol{V} = \lim_{n\to\infty} \frac{1}{n}\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\beta}_0)$ exists in probability. Since $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\beta}}_{FGLS}$ are consistent, the limit inside the parentheses above equals $\boldsymbol{V}$ by the Continuous Map-

ping Theorem. For the same reason, the limiting behavior of $\frac{1}{\sqrt{n}}\boldsymbol{J}^T(\hat{\boldsymbol{\beta}}_{FGLS})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{e}$ and $\frac{1}{\sqrt{n}}\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{e}$ are the same. We now show that conditional on the covariates, the limiting distribution of $\frac{1}{\sqrt{n}}\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{e}$ is $N(\mathbf{0}, \boldsymbol{V})$. By the Cramér Wold device, it suffices to prove that,

$$\frac{1}{\sqrt{n}}\boldsymbol{a}^T\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{e} \xrightarrow{D} N(0, \boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a}), \quad \forall \boldsymbol{a} \in \mathbb{R}^p$$

After some matrix algebra, we can write,

$$\frac{1}{\sqrt{n}}\boldsymbol{a}^T\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{e} = \frac{1}{\sqrt{n}}\sum_{i=1}^{p} a_i \sum_{\boldsymbol{s}_j} c_i(\boldsymbol{s}_j)e(\boldsymbol{s}_j) = \frac{1}{\sqrt{n}}\sum_{\boldsymbol{s}_j} b(\boldsymbol{s}_j)e(\boldsymbol{s}_j)$$

where $c_i(\boldsymbol{s}_j)$ is the ij^{th} element of the $p \times n$ matrix $\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)$ and $b(\boldsymbol{s}_j) = \sum_{i=1}^{p} a_i c_i(\boldsymbol{s}_j)$. By the spatial CLT in Theorem 3.2.1, we have,

$$\frac{1}{\sqrt{n}}\sum_{\boldsymbol{s}_j} b(\boldsymbol{s}_j)e(\boldsymbol{s}_j) \xrightarrow{D} N\left(0, \lim_{n\to\infty} n^{-1}\sum_{\boldsymbol{s}_i,\boldsymbol{s}_j} \text{Cov}(b(\boldsymbol{s}_i)e(\boldsymbol{s}_i), b(\boldsymbol{s}_j)e(\boldsymbol{s}_j))\right)$$

The covariance term inside the sum above can be calculated as,

$$\sum_{\boldsymbol{s}_i,\boldsymbol{s}_j} \text{Cov}(b(\boldsymbol{s}_i)e(\boldsymbol{s}_i), b(\boldsymbol{s}_j)e(\boldsymbol{s}_j)) = \sum_{k=1}^{p}\sum_{\ell=1}^{p} a_k a_\ell \sum_{\boldsymbol{s}_i,\boldsymbol{s}_j} c_k(\boldsymbol{s}_i)\text{Cov}(e(\boldsymbol{s}_i), e(\boldsymbol{s}_j))c_\ell(\boldsymbol{s}_j)$$

The term $\sum_{\boldsymbol{s}_i,\boldsymbol{s}_j} c_k(\boldsymbol{s}_i)\text{Cov}(e(\boldsymbol{s}_i), e(\boldsymbol{s}_j))c_\ell(\boldsymbol{s}_j)$ can be recognized as the $k\ell^{th}$ element of the $p \times p$ matrix $\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{J}(\boldsymbol{\beta}_0)$. Thus, in the limit as $n \to \infty$, we have shown

that,

$$\frac{1}{\sqrt{n}}\boldsymbol{a}^T\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{e} = \frac{1}{\sqrt{n}}\sum_{\boldsymbol{s}_j} b(\boldsymbol{s}_j)e(\boldsymbol{s}_j) \xrightarrow{D} N(0, \boldsymbol{a}^T\boldsymbol{V}\boldsymbol{a})$$

So by the Cramér Wold device, $\frac{1}{\sqrt{n}}\boldsymbol{J}^T(\boldsymbol{\beta}_0)\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}_0)\boldsymbol{e} \xrightarrow{D} N(0, \boldsymbol{V})$. Finally applying Slutsky's Theorem to the expression given in (3.24),

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_{FGLS} - \boldsymbol{\beta}_0) &= \underbrace{\left(\frac{1}{n}\boldsymbol{J}^T(\hat{\boldsymbol{\beta}}_{FGLS})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{J}(\bar{\boldsymbol{\beta}})\right)^{-1}}_{\xrightarrow{P}\boldsymbol{V}^{-1}} \underbrace{\frac{1}{\sqrt{n}}\boldsymbol{J}^T(\hat{\boldsymbol{\beta}}_{FGLS})\boldsymbol{\Sigma}^{-1}(\hat{\boldsymbol{\theta}})\boldsymbol{e}}_{\xrightarrow{D}N(\boldsymbol{0},\boldsymbol{V})} \\ &\xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{V}^{-1})\end{aligned}$$

Thus, asymptotic normality holds. □

Appendix: Cardinality arguments

The following lemmas are needed in establishing the consistency of the Matheron variogram in Lemma 3.2.2. The arguments here are inspired by Appendix A of Jenish and Prucha (2009).

Lemma 3.5.1. *Any cube with side length* 1 *can contain at most one pair in* $T(\boldsymbol{h})$.

Proof. If there were two such pairs, say $\{\boldsymbol{s}_1, \boldsymbol{s}_2\}$ and $\{\boldsymbol{s}_3, \boldsymbol{s}_4\}$, then the distance between these pairs would be less than 1, which violates Assumption 3.1.1. □

Lemma 3.5.2. *Any ball of radius* M, $B(\boldsymbol{s}, M)$, *can contain at most* $(2M)^d$ *pairs in* $T(\boldsymbol{h})$.

Proof. Such a ball is contained inside a cube of side length $2M$. Each such cube can be partitioned into $(2M)^d$ unit cubes. Then the result follows from Lemma 3.5.1. □

Lemma 3.5.3. *Consider the annulus,* $B(\boldsymbol{s}, M) \setminus B(\boldsymbol{s}, M-1) = \{\boldsymbol{t} \in \mathbb{R}^d : M-1 < d(\boldsymbol{s}, \boldsymbol{t}) \leq M\}$. *This annulus can contain at most* $O(M^{d-1})$ *pairs in* $T(\boldsymbol{h})$.

Proof. From Lemma 3.5.2, and the inequality, $b^d - a^d \leq (b-a)db^{d-1}$ for $a, b \geq 0$, this annulus can contain at most,

$$(2M+2)^d - (2M)^d \leq 2d(2M+2)^{d-1} = 2^d dM^{d-1}(1+M^{-1})^{d-1} \leq CM^{d-1}$$

pairs, where C is a constant independent of $M \geq 1$. □

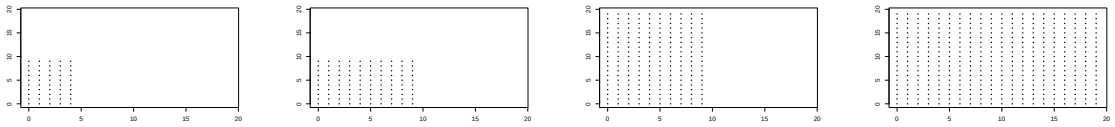
Chapter 4 Numerical study on increasing domain asymptotics

4.1 A comparison of MLE versus least squares variogram estimation

In Section 3.4 of Chapter 3, we gave a brief discussion on maximum likelihood and least squares variogram estimation of the covariance parameters $\boldsymbol{\theta}$. In this section, we perform a simulation study comparing the two methods of estimation. Based on this simulation study, we conjecture that maximum likelihood estimation is in general more efficient than least squares variogram estimation when the errors are Gaussian.

4.1.1 Numerical setup

We take our spatial locations to be a regular rectangular lattice in $\mathbb{R}^2$ with unit spacings. To simulate an increasing domain asymptotics framework, the size of the grid also increases with the number of locations. For our numerical study, we consider sample sizes of $n = 50, 100, 200$ and 400, nested within each other (see Figure 4.1).



(a) 50 locations (b) 100 locations (c) 200 locations (d) 400 locations

Figure 4.1: Increasing domain asymptotics framework

We consider the simple linear spatial regression model,

$$y(\boldsymbol{s}) = \beta_0 + \beta_1 x(\boldsymbol{s}) + e(\boldsymbol{s}) \tag{4.1}$$

where $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ are independent mean zero Gaussian random fields. For the covariate, we consider an exponential covariance $C_x(\boldsymbol{h}) = e^{-\frac{3}{2}||\boldsymbol{h}||}$ and generate one realization $\boldsymbol{x} = (x(\boldsymbol{s}_1), \cdots, x(\boldsymbol{s}_n))^T$ according to this model. For the error, we consider two different covariance functions,

1. Exponential: $C_e(\boldsymbol{h}) = \sigma_e^2 e^{-\theta_e||\boldsymbol{h}||}$

2. Matérn with smoothness parameter $\nu = \frac{3}{2}$: $C_e(\boldsymbol{h}) = \sigma_e^2(1+\sqrt{3}\theta_e||\boldsymbol{h}||)e^{-\sqrt{3}\theta_e||\boldsymbol{h}||}$

We choose as the true parameters $\boldsymbol{\theta}_0 = (\sigma_{e0}^2, \theta_{e0})^T = (\frac{1}{2}, 1)^T$. Finally, for LS variogram estimation, we consider 7 lag vectors $\boldsymbol{h}_1 = (0,1)^T, \boldsymbol{h}_2 = (1,1)^T, \boldsymbol{h}_3 = (1,2)^T, \boldsymbol{h}_4 = (2,2)^T, \boldsymbol{h}_5 = (3,1)^T, \boldsymbol{h}_6 = (2,3)^T$ and $\boldsymbol{h}_7 = (3,3)^T$. Since we are on a regular lattice with spacing one unit apart, we do not require a tolerance region defined in (3.12).

4.1.2 Comparison of asymptotic variances

First, we compare the asymptotic variances as predicted by MLE theory and least squares variogram theory. Recall from Section 3.4 that the asymptotic variance of the MLE is the inverse of the Fisher information as given in (3.18). Recall that the asymptotic variance of the least squares variogram estimator is given in Corollary 3.2.3. We compute these asymptotic covariance matrices at the true $\boldsymbol{\theta}_0$ and present the diagonal elements of each matrix. For the exponential model, the results are given in Table 4.1. For the Matérn model, the results are given in Table 4.2.

	Variance for $\hat{\sigma}_e^2$		Variance for $\hat{\theta}_e$	
n	MLE	Least Squares	MLE	Least Squares
50	0.01717	0.01923	0.14011	0.28313
100	0.00896	0.01098	0.07057	0.14384
200	0.00458	0.00584	0.03546	0.07313
400	0.00234	0.00311	0.01785	0.03692

Table 4.1: Asymptotic variances of $(\hat{\sigma}_e^2, \hat{\theta}_e)^T$ as predicted by MLE and least squares for the exponential variogram model

	Variance for $\hat{\sigma}_e^2$		Variance for $\hat{\theta}_e$	
n	MLE	Least Squares	MLE	Least Squares
50	0.02013	0.02161	0.03787	0.09947
100	0.01054	0.01245	0.01866	0.05107
200	0.00540	0.00664	0.00929	0.02606
400	0.00277	0.00355	0.00463	0.01323

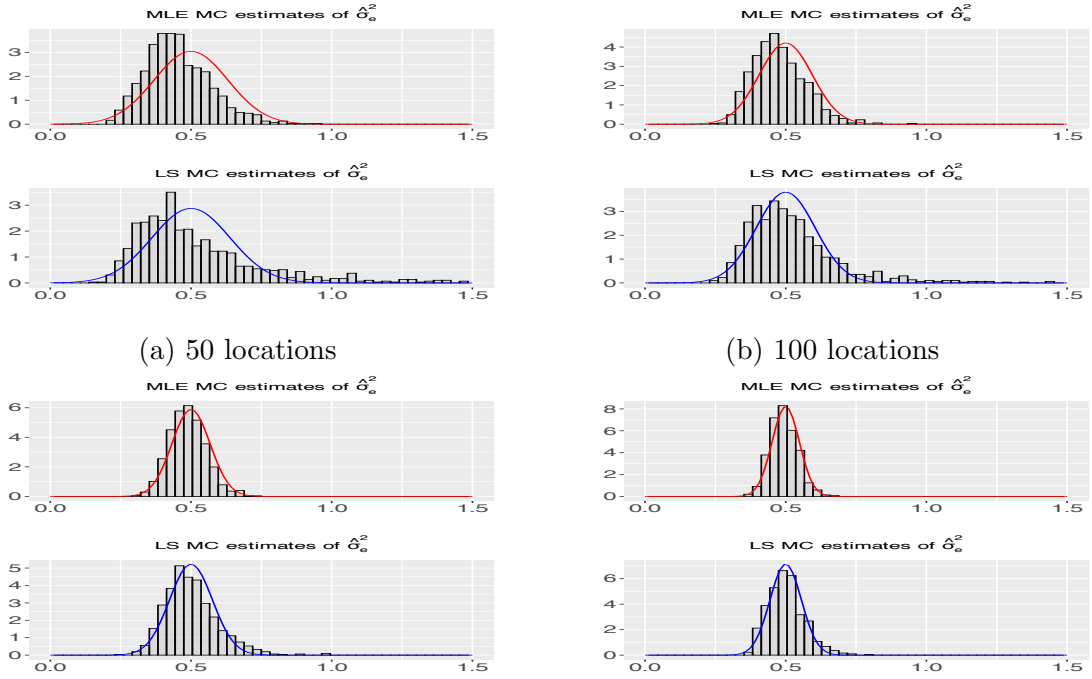
Table 4.2: Asymptotic variances of $(\hat{\sigma}_e^2, \hat{\theta}_e)^T$ as predicted by MLE and least squares for the Matérn ($\nu = \frac{3}{2}$) variogram model

We can see that the MLE asymptotic variances are in general smaller than those of the least squares. In both tables, the relative efficiencies of LS compared to MLE

for $\hat{\sigma}_e^2$ begin roughly at 90% for $n = 50$ but surprisingly drop to approximately 75% when $n = 400$. For $\hat{\theta}_e$, the MLE outperforms LS by a significant amount. The relative efficiencies of LS compared to MLE for $\hat{\theta}_e$ are roughly 50% in Table 4.1 and roughly 35% in Table 4.2.

4.1.3 Comparison of Monte Carlo estimates

Next, we compare Monte Carlo estimates using MLE and least squares. Conditional on $\boldsymbol{x}$, we generate 1000 realizations $\boldsymbol{y}$ according to the simple linear regression model in (4.1), that is, $\boldsymbol{y}|\boldsymbol{x} \sim N(\beta_0\mathbf{1} + \beta_1\boldsymbol{x}, \sigma_e^2\boldsymbol{\Sigma}(\theta_e))$. For our simulations, we arbitrarily choose $(\beta_0, \beta_1)^T$ to be $(6, 3)^T$. Then, we compute MLE and LS estimates for $(\sigma_e^2, \theta_e)^T$ for each realization. We used the nlminb function of the base R package.



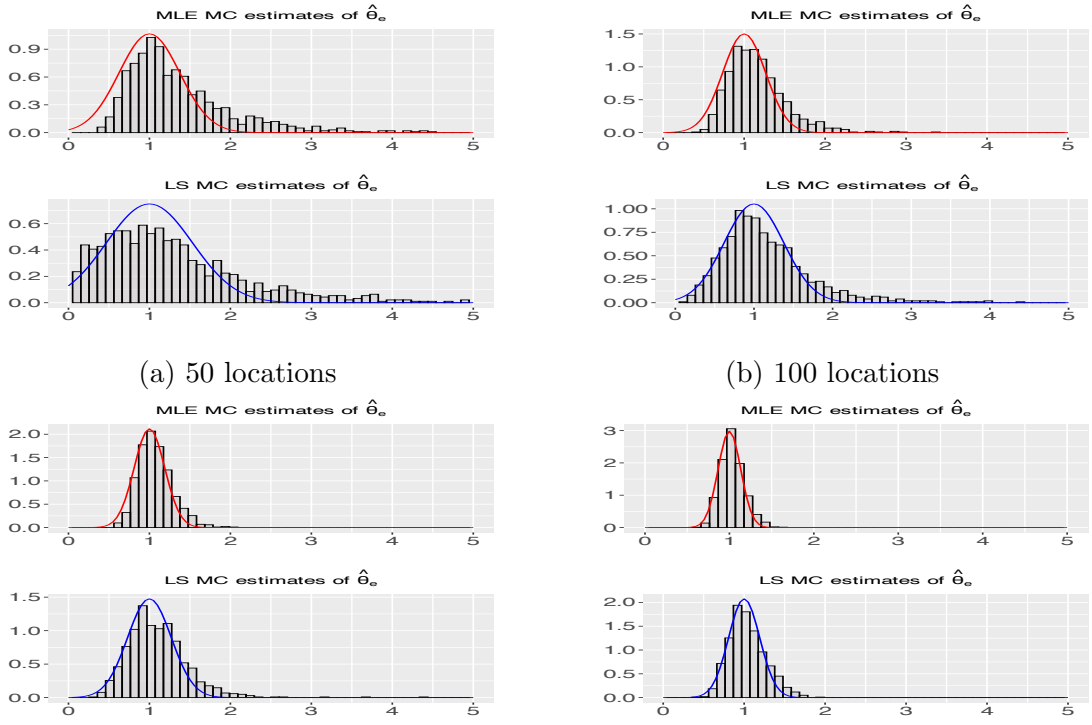
(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

Figure 4.2: MLE and LS estimates for σ_e^2 in the exponential model. In red and blue are the theoretical asymptotic densities predicted by MLE and LS. Histograms are based on 1000 MC simulations of $\boldsymbol{y}|\boldsymbol{x}$ according to (4.1).



(a) 50 locations

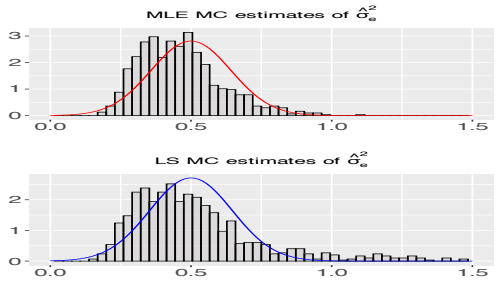
(b) 100 locations

(c) 200 locations

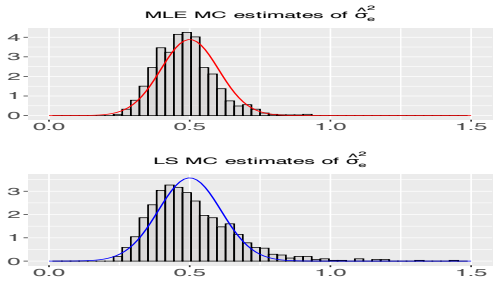
(d) 400 locations

Figure 4.3: MLE and LS histogram estimates for θ_e in the exponential model. In red and blue are the theoretical asymptotic densities predicted by MLE and LS. Histograms are based on 1000 MC simulations of $\boldsymbol{y}|\boldsymbol{x}$ according to (4.1).

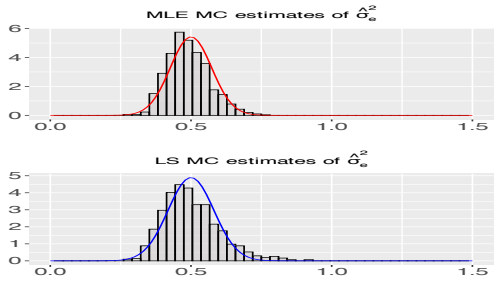
Histograms for the MLE and LS estimates for $(\sigma_e^2, \theta_e)^T$ in the exponential model are given in Figures 4.2 and 4.3 respectively. The histograms for the LS estimates for $(\sigma_e^2, \theta_e)^T$ in general are wider than those of the MLE. Both methods are not well approximated by their theoretical asymptotic densities in small samples. The approximation improves once we reach $n = 200$ observations. We should note that the computation time for the MC estimation of the least squares estimates was orders of magnitude less than MLE estimation, especially as we got to $n = 400$ observations (approximately 5 seconds for LS vs. 25 minutes for MLE on an 8-core laptop computer). This is expected since we are inverting a large matrix in the likelihood function, whereas no such matrix inversion occurs in least squares.



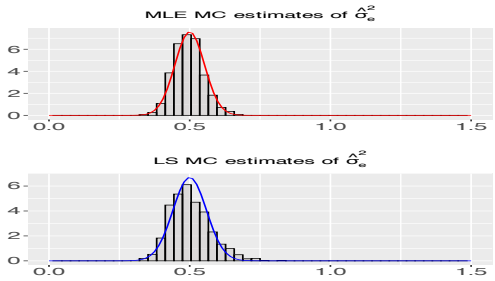
(a) 50 locations



(b) 100 locations

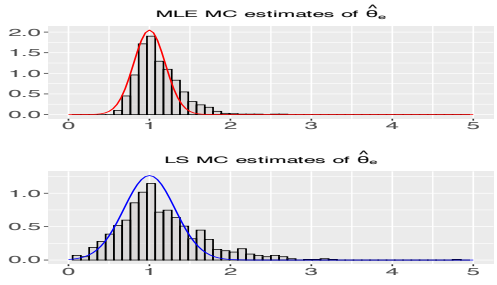


(c) 200 locations

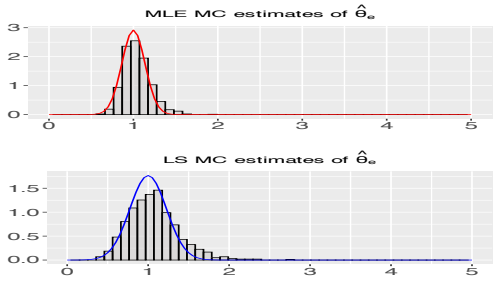


(d) 400 locations

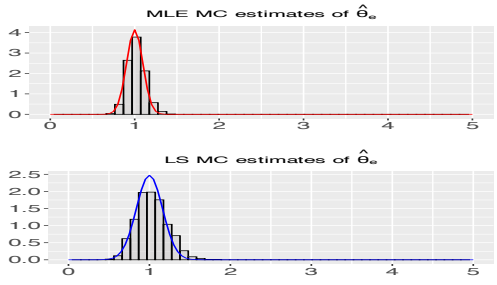
Figure 4.4: MLE and LS histogram estimates for σ_e^2 in the Matérn ($\nu = \frac{3}{2}$) model. In red and blue are the theoretical asymptotic densities predicted by MLE and LS. Histograms are based on 1000 MC simulations of $\boldsymbol{y}|\boldsymbol{x}$ according to (4.1).



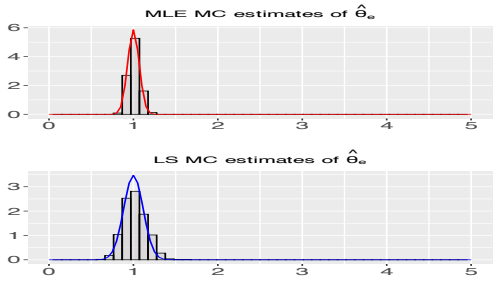
(a) 50 locations



(b) 100 locations



(c) 200 locations



(d) 400 locations

Figure 4.5: MLE and LS histogram estimates for θ_e in the Matérn ($\nu = \frac{3}{2}$) model. In red and blue are the theoretical asymptotic densities predicted by MLE and LS. Histograms are based on 1000 MC simulations of $\boldsymbol{y}|\boldsymbol{x}$ according to (4.1).

Histograms for the MLE and LS estimates for $(\sigma_e^2, \theta_e)^T$ in the Matérn model are given in Figures 4.4 and 4.5 respectively. On top of each subfigure are the MLE estimates and on the bottom are the LS estimates. These estimates show similar behavior as in the exponential model. In general, the MLE histograms are more peaked than the LS ones and once again, the computation time for LS was much less than MLE.

4.2 Real data example: Temperature and pressure in the Pacific Northwest

To demonstrate the estimation procedure in Chapter 3, we explore a weather dataset analyzed by Gneiting et al. (2010).

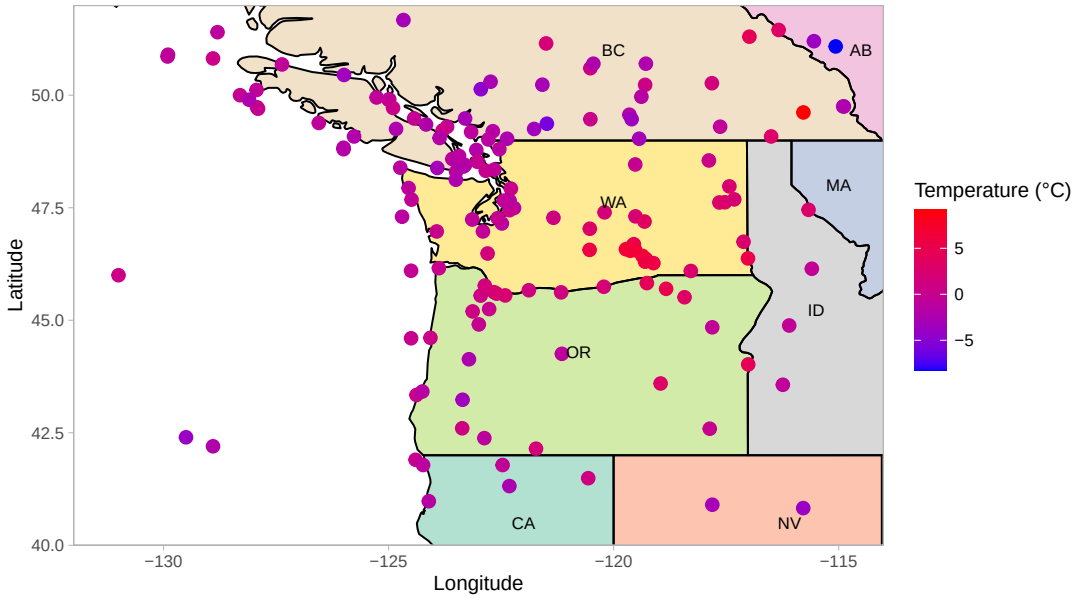


Figure 4.6: Locations of 157 weather stations in the Pacific Northwest region of the United States and Western Canada. Weather stations directly in the Pacific Ocean correspond to buoys and ships.

The dataset contains temperature and pressure forecast errors (forecasts minus observations) from 157 stations in the Northwestern US and Western Canada region (see Figure 4.6). The dataset can be found in RandomFields package in R (Schlather et al. (2022)). To showcase their novel Matérn cross covariance function, Gneiting et al. (2010) modelled the temperature and pressure forecast errors as a bivariate Gaussian random field. We take an another approach and model their relationship through a linear regression model. Let $y(\boldsymbol{s})$ represent the underlying temperature forecast errors (degrees Celsius) and $x_1(\boldsymbol{s})$ represent the underlying pressure forecast errors (kilopascals). From Figure 4.6, we can see that the temperature gradient varies with the location. So we consider as two other covariates $x_2(\boldsymbol{s})$ and $x_3(\boldsymbol{s})$ the east-west spatial coordinate and north-south spatial coordinate of $\boldsymbol{s} \in \mathbb{R}^2$ respectively. Then, consider the model,

$$y(\boldsymbol{s}) = \beta_0 + \beta_1 x_1(\boldsymbol{s}) + \beta_2 x_2(\boldsymbol{s}) + \beta_3 x_3(\boldsymbol{s}) + e(\boldsymbol{s}) \tag{4.2}$$

where $e(\boldsymbol{s})$ is a random field independent of $x(\boldsymbol{s})$. First, we consider the provisional ordinary least squares estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$. In matrix-vector form, the regression model (4.2) is $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Then the OLS estimator is $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = (-0.9936, -7.7696, 0.0482, -0.3169)^T$. Next, after forming the residuals $\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS}$, we perform a variogram analysis to estimate the variance parameters of $e(\boldsymbol{s})$. After doing a map projection of the latitude and longitude coordinates into cartesian coordinates (done with the RFearth2cartesian function in R), we found that the maximum distance between two pairs of points is 16 (in units

of 100 km). For the number of lag vectors K, Crujeiras and van Keilegom (2010) (p. 454) recommend using $K \leq \frac{U}{2}$, where U is the maximum distance between two pairs of points. Thus, in our case, we decided to use 8 lag vectors. The fitted variogram can be found in Figure 4.7.

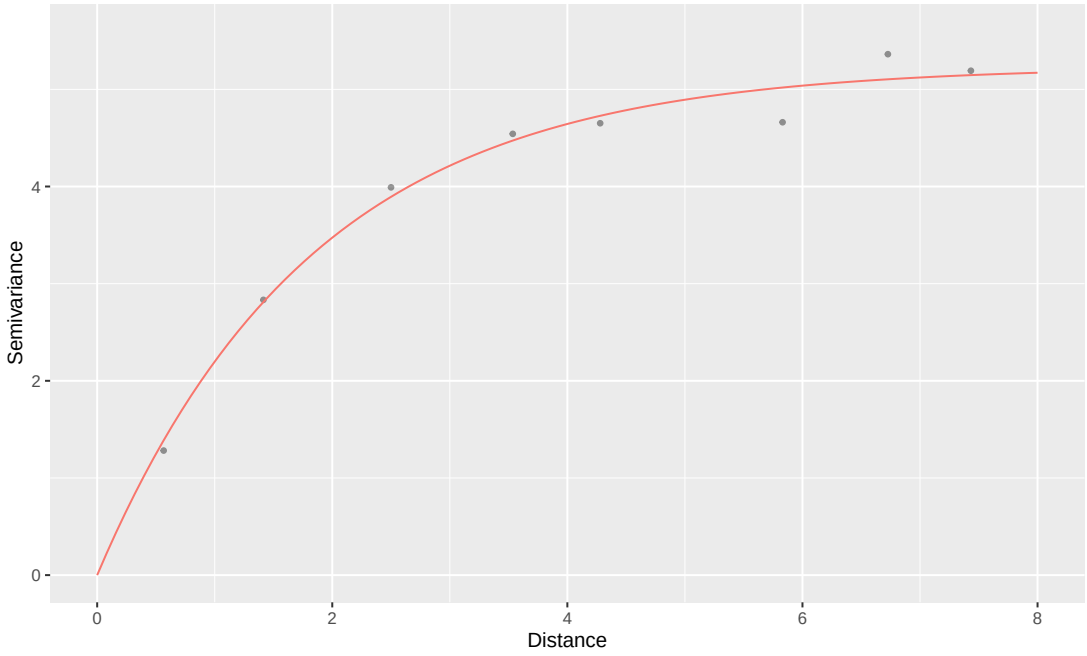


Figure 4.7: Variogram analysis of weather data from Gneiting et al. (2010). Using the OLS residuals from model (4.2), the points represent the empirical variogram values calculated from equation (3.11) at 8 different lag vectors. The red line represents the fitted exponential variogram.

According to Gneiting et al. (2010), temperature forecasts are not smooth, which makes sense geophysically. Using a Matérn covariance, they estimated the smoothness of the temperature random field to be approximately 0.6 which is close to an exponential variogram (smoothness of 0.5). We fitted an exponential variogram to the residuals and found that the model fits reasonably well (Figure 4.7). Thus, we assume that $e(\boldsymbol{s})$ has an exponential variogram structure $\gamma(\boldsymbol{h}; \sigma^2, \theta) = \sigma^2(1 - e^{-\theta||\boldsymbol{h}||})$. The least squares variogram estimates of $\boldsymbol{\theta} = (\sigma^2, \theta)^T$ along with their estimated

standard errors are presented in Table 4.3. The standard errors were computed using Corollary 3.2.3.

	$\hat{\sigma}^2$	$\hat{\theta}$
Estimate (S.E.)	5.2385 (1.9281)	0.5440 (0.3645)

Table 4.3: Least squares variogram estimates along with their estimated standard errors

Finally, we re-estimate $\boldsymbol{\beta}$ using the FGLS procedure. With the estimated $\hat{\boldsymbol{\theta}}$, we compute $\hat{\boldsymbol{\beta}}_{FGLS} = (\boldsymbol{X}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})^{-1}\boldsymbol{y}$, where the i, j^{th} element of $\boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}})$ is given by $\hat{\sigma}^2 e^{-\hat{\theta}||\boldsymbol{s}_i - \boldsymbol{s}_j||}$. The FGLS estimates along with their estimated standard errors are presented in Table 4.4. The standard errors were computed using Proposition 3.3.2.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Estimate (S.E.)	-0.3697 (0.7039)	-3.4384 (0.7640)	-0.0835 (0.1557)	-0.0493 (0.1523)

Table 4.4: FGLS estimates along with their estimated standard errors

For completeness, we now give a comparison with maximum likelihood estimation, assuming Gaussianity like Gneiting et al. (2010). Conditionally given $\boldsymbol{x}$, the distribution of $\boldsymbol{y}|\boldsymbol{x}$ is $N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$. The conditional negative log-likelihood $L(\boldsymbol{\beta}, \boldsymbol{\theta})$ was then minimized in R using the nlminb function. The norm of the gradient was small and the Hessian had positive eigenvalues, indicating that the estimates were local minima. The estimates along with their estimated standard errors are presented in Tables 4.5 and 4.6. The standard errors were computed using the Fisher information (Theorem 3.4.1).

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
Estimate (S.E.)	0.0943 (0.4515)	-4.0186 (0.9912)	0.0139 (0.1225)	-0.0899 (0.1136)

Table 4.5: MLE estimates of $\boldsymbol{\beta}$ along with their estimated standard errors

	$\hat{\sigma}^2$	$\hat{\theta}$
Estimate (S.E.)	5.0345 (0.9218)	1.1158 (0.2533)

Table 4.6: MLE estimates of $(\sigma^2, \theta)^T$ along with their estimated standard errors

Both the FGLS and MLE estimators for $\boldsymbol{\beta}$ suggest that the coefficient for pressure β_1 is the only significant one in the model. In contrast, a summary of the OLS model in R revealed that both the intercept β_0 and β_3, corresponding to the north-south spatial coordinate, were also significant. However, the OLS model assumes independent, identically distributed errors which would be misleading with this spatial data. A comparison of Tables 4.3 and 4.6 shows that MLE estimates of $(\sigma^2, \theta)^T$ have smaller estimated variances, which agrees with the simulation study earlier in this chapter. In fact, it appears that the least squares variogram estimate for σ^2 achieves half the efficiency of the MLE. It is also noteworthy that the least squares estimate for $\hat{\theta}$ is not very well resolved for $n = 157$ locations.

Chapter 5 Confounding in nonlinear spatial regression models

In this chapter, we assume that we have one covariate confounded with the error. Our nonlinear regression model takes the form,

$$y(\boldsymbol{s}) = f(x(\boldsymbol{s}); \boldsymbol{\beta}) + e(\boldsymbol{s}), \quad \boldsymbol{s} \in D \subset \mathbb{R}^d \tag{5.1}$$

where D is a countably infinite lattice in $\mathbb{R}^d$. We assume that $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ are both mean zero stationary Gaussian random fields that are dependent in the sense that their cross covariance function is nonzero. Page et al. (2017) considers a linear regression model in (5.1) where $f(x(\boldsymbol{s}); \boldsymbol{\beta}) = \beta_0 + \beta_1 x(\boldsymbol{s})$. Assuming a specific form of the cross-covariance structure (see Section 5.2.5), they investigate the effect of confounding on generalized least squares estimation of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$. Under the strict assumption that the variance components of $e(\boldsymbol{s})$ are known, they conclude that there can be significant bias in GLS if the confounding is not accounted for. We investigate the confounding in (5.1) under a more general nonlinear setting, and consider several different bivariate cross-covariance structures for $(x(\boldsymbol{s}), e(\boldsymbol{s}))^T$. We generalize the results of Page et al. (2017) a bit further by assuming the variance

components of $e(\boldsymbol{s})$ are unknown.

Recall from Section 2.3 that for a stationary bivariate random field, the cross covariance function is a symmetric positive definite, 2×2 matrix valued function,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} C_{xx}(\boldsymbol{h}) & C_{xe}(\boldsymbol{h}) \\ C_{ex}(\boldsymbol{h}) & C_{ee}(\boldsymbol{h}) \end{pmatrix}$$

For symmetry, it is necessary that $C_{xe}(\boldsymbol{h}) = C_{ex}(-\boldsymbol{h})$ and in general, $C_{xe}(\boldsymbol{h}) \neq C_{ex}(\boldsymbol{h})$, that is, the off-diagonal functions are not equal. Following Paciorek (2010) and Page et al. (2017), we entertain parametric cross-covariance functions of the form,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2 \phi_{xx}(\boldsymbol{h}; \boldsymbol{\theta}_x) & \rho \sigma_x \sigma_e \phi_{xe}(\boldsymbol{h}; \boldsymbol{\theta}_{xe}) \\ \rho \sigma_x \sigma_e \phi_{ex}(\boldsymbol{h}; \boldsymbol{\theta}_{ex}) & \sigma_e^2 \phi_{ee}(\boldsymbol{h}; \boldsymbol{\theta}_e) \end{pmatrix} \tag{5.2}$$

where $\phi_{xx}, \phi_{xe}, \phi_{ee}$ are parametric correlation functions equalling 1 at the origin. The parameters σ_x^2, σ_e^2 and ρ represent the marginal variance of $x(\boldsymbol{s})$, the marginal variance of $e(\boldsymbol{s})$ and the confounding parameter respectively. In a non-confounding context, the parameter ρ is sometimes referred to as a collocated cross-correlation coefficient (Genton and Kleiber (2015)). When $\rho = 0$, this reduces to the case independent covariate and error as discussed in Chapter 3.

When analyzing possible confounding models for $\boldsymbol{C}(\boldsymbol{h})$, we first and foremost must determine if they are valid, that is, if they are non-negative definite. By The-

orem 2.3.1 of Cramér, this requires that the corresponding spectral density matrix,

$$\boldsymbol{f}(\boldsymbol{\omega}) = \begin{pmatrix} \sigma_x^2 f_{xx}(\boldsymbol{\omega};\boldsymbol{\theta}_x) & \rho\sigma_x\sigma_e f_{xe}(\boldsymbol{\omega};\boldsymbol{\theta}_{xe}) \\ \rho\sigma_x\sigma_e f_{ex}(\boldsymbol{\omega};\boldsymbol{\theta}_{ex}) & \sigma_e^2 f_{ee}(\boldsymbol{\omega};\boldsymbol{\theta}_e) \end{pmatrix} \tag{5.3}$$

be non-negative definite, that is, $f_{xx}(\boldsymbol{\omega};\boldsymbol{\theta}_x)f_{ee}(\boldsymbol{\omega};\boldsymbol{\theta}_e) \geq \rho^2 f_{xe}(\boldsymbol{\omega};\boldsymbol{\theta}_{xe})f_{ex}(\boldsymbol{\omega};\boldsymbol{\theta}_{ex})$ for almost all $\boldsymbol{\omega}$. Even for a bivariate model, this criterion may lead to non-trivial restrictions on the parameters, as we see for some of the models discussed below.

5.1 Identifiability in confounding models

Once we have a valid model, we would like to determine if it is feasible for analytical and practical use on real spatial data. We state two identifiability criteria related to estimation of the parameters in the model (5.1).

1. (*Formal identifiability*) Since we are using likelihood estimation, we require that the parameters in model (5.1) be identifiable. When $y(\boldsymbol{s})$ and $x(\boldsymbol{s})$ are observed at locations $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$, the corresponding vector form of (5.1) is,

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) + \boldsymbol{e} \tag{5.4}$$

where,

$$\boldsymbol{y} \quad = \quad (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))^T$$

$$
\begin{aligned}
\boldsymbol{x} &= (x(\boldsymbol{s}_1), ..., x(\boldsymbol{s}_n))^T \\
\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\beta}) &= (f(x(\boldsymbol{s}_1); \boldsymbol{\beta}), ..., f(x(\boldsymbol{s}_n); \boldsymbol{\beta}))^T \\
\boldsymbol{e} &= (e(\boldsymbol{s}_1), ..., e(\boldsymbol{s}_1))^T
\end{aligned}
$$

By assumption, $\boldsymbol{x}$ and $\boldsymbol{e}$ have a joint Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix,

$$
\boldsymbol{\Psi}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_x^2 \boldsymbol{\Phi}_{xx}(\boldsymbol{\theta}_x) & \rho\sigma_x\sigma_e \boldsymbol{\Phi}_{xe}(\boldsymbol{\theta}_{xe}) \\ \rho\sigma_x\sigma_e \boldsymbol{\Phi}_{ex}(\boldsymbol{\theta}_{ex}) & \sigma_e^2 \boldsymbol{\Phi}_{ee}(\boldsymbol{\theta}_e) \end{pmatrix}_{2n \times 2n} \tag{5.5}
$$

where $\{\boldsymbol{\Phi}_{xx}(\boldsymbol{\theta}_x)\}_{i,j} = \phi_{xx}(\boldsymbol{s}_i - \boldsymbol{s}_j; \boldsymbol{\theta}_x)$ and similarly for the other $n \times n$ block matrices. The conditional distribution of $\boldsymbol{y}|\boldsymbol{x}$ in model (5.4) is,

$$
\boldsymbol{y}|\boldsymbol{x} \sim N\left(\boldsymbol{f}(\boldsymbol{x}; \boldsymbol{\beta}) + \boldsymbol{\tau}(\boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\theta})\right) \tag{5.6}
$$

where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^q$ is a vector of covariance parameters containing the confounding parameter ρ. The quantity $\boldsymbol{\tau}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{e}|\boldsymbol{x}]$ is the $n \times 1$ conditional mean vector of $\boldsymbol{e}|\boldsymbol{x}$. By the joint Gaussian assumption, the conditional mean is a linear function of the covariate vector $\boldsymbol{\tau}(\boldsymbol{\theta}) = \boldsymbol{A}(\boldsymbol{\theta})\boldsymbol{x}$, where $\boldsymbol{A}(\boldsymbol{\theta})$ is an $n \times n$ matrix which in some confounding models may simplify to a scalar function of θ multiplied by the identity matrix. The covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta})$ is the $n \times n$ conditional covariance matrix of $\boldsymbol{e}|\boldsymbol{x}$. We define formal identifiability as the likelihood resulting from (5.6) being one-to-one with respect to the unknown parameters $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$, for any set of observed data $\boldsymbol{y}, \boldsymbol{x}$. Gen-

erally, $x(\boldsymbol{s})$ has unknown parameters $(\sigma_x^2, \boldsymbol{\theta}_x^T)^T$ parametrizing its covariance, but since we have observations $\boldsymbol{x} = (x(\boldsymbol{s}_1), \cdots, x(\boldsymbol{s}_n))^T$ available, we assume that they can be consistently estimated. In particular, since $\boldsymbol{x}$ is Gaussian, the likelihood is given by,

$$\boldsymbol{x} \sim N(\mathbf{0}, \sigma_x^2 \boldsymbol{\Sigma}_x(\boldsymbol{\theta}_x)) \tag{5.7}$$

Assumption 5.1.1. *Consistent estimators of the parameters* $(\sigma_x^2, \boldsymbol{\theta}_x^T)^T$ *exist using the* $\boldsymbol{x}$ *data alone, for example, MLE or least squares variogram estimators. Thus, these parameters are identifiable in the sense that the likelihood in (5.7) is one-to-one with respect to* $(\sigma_x^2, \boldsymbol{\theta}_x^T)^T$

2. (*Practical identifiability*) In theory, on an increasing domain asymptotics framework, we expect the MLE estimates of the parameters to converge to a normal distribution with variance given by the inverse Fisher information matrix (Mardia and Marshall (1984)). Through simulations in Chapter 4, we look at the behavior of the Fisher information as the number of observations increase on a regular lattice. We define practical identifiability as good numerical behavior of the model based on the following criteria,

   (1) (*The eigenvalues and condition numbers of the Fisher information matrix*) An ill-conditioned Fisher information and/or Hessian matrix sometimes precludes the use of standard optimization methods like gradient descent and quasi Newton-Raphson.

(2) (*The inverse of the Fisher information*) Standard asymptotic theory suggests the MLE estimates will have an asymptotic covariance matrix equal to the inverse Fisher information. We expect the numbers to be small and to scale down proportionately as the number of observations increase.

(3) (*Empirical distributions of Monte Carlo estimates*) A histogram of MC MLE estimates of each parameter should be well approximated by their theoretical asymptotically normal density. These estimates should also be true minima, that is, the gradients should be small and the Hessian matrix should be positive definite.

The Fisher information is $\boldsymbol{\mathscr{I}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = E_{\boldsymbol{\beta},\boldsymbol{\theta}}[\boldsymbol{H}|\boldsymbol{x}]$, where $\boldsymbol{H}$ is the Hessian matrix of the negative log-likelihood of (5.6). The Fisher information has the general form,

$$\boldsymbol{\mathscr{I}}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{\mathscr{I}}_{\beta\beta} & \boldsymbol{\mathscr{I}}_{\beta\theta} \\ \boldsymbol{\mathscr{I}}_{\beta\theta} & \boldsymbol{\mathscr{I}}_{\theta\theta} \end{bmatrix}_{(p+q)\times(p+q)} \tag{5.8}$$

where the block matrices are given by,

1. $\boldsymbol{\mathscr{I}}_{\beta\beta} = \boldsymbol{J}(\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \boldsymbol{J}(\boldsymbol{\beta})$, where $\boldsymbol{J}(\boldsymbol{\beta})$ is the $n \times p$ Jacobian matrix of $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. In the simple linear regression case, $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\beta}) = \alpha \mathbf{1} + \beta \boldsymbol{x} = \boldsymbol{X}\boldsymbol{\beta}$, this Jacobian simply equals the design matrix $\boldsymbol{X}$.

2. The i^{th} column of $\boldsymbol{\mathscr{I}}_{\beta\theta}$ equals $\boldsymbol{J}(\boldsymbol{\beta})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) \boldsymbol{A}_{\theta_i}(\boldsymbol{\theta}) \boldsymbol{x}$, where $\boldsymbol{A}_{\theta_i}(\boldsymbol{\theta}) = \dfrac{\partial \boldsymbol{A}(\boldsymbol{\theta})}{\partial \theta_i}$ is the element-wise partial derivative of $\boldsymbol{A}(\boldsymbol{\theta})$ with respect to θ_i.

3. The i, j^{th} element of $\boldsymbol{\mathscr{I}}_{\theta\theta}$ equals,

$$\frac{1}{2}\text{tr}(\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}_{\theta_i}(\boldsymbol{\theta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}_{\theta_j}(\boldsymbol{\theta})) + \boldsymbol{x}^T\boldsymbol{A}_{\theta_i}(\boldsymbol{\theta})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})\boldsymbol{A}_{\theta_j}(\boldsymbol{\theta})\boldsymbol{x}$$

where $\boldsymbol{\Sigma}_{\theta_i}(\boldsymbol{\theta})$ is defined similarly to $\boldsymbol{A}_{\theta_i}(\boldsymbol{\theta})$ above.

Due to the presence of $\boldsymbol{\tau}(\boldsymbol{\theta})$ in the mean, the matrix in (5.8) is not block diagonal. Thus, we cannot directly apply Theorem 3.4.1 of Mardia and Marshall (1984) since the conditions they derived were based on the block diagonal structure of the Fisher information (3.18) in the unconfounded model. Sweeting (1980) gives more general conditions on the Fisher information that ensures consistency and asymptotic normality of MLE estimates. These conditions are given in Chapter 7, Theorem 7.4.4, where they are applied in a simpler setting. Since it is difficult to verify the conditions of Mardia and Marshall (1984) and Sweeting (1980) due to the complex structure of the Fisher information in (5.8), we resort to numerical simulations. Nevertheless, based on our simulation study of practical identifiability in Chapter 6, we expect asymptotic normality of MLE estimates to hold with the inverse of (5.8) as the asymptotic covariance matrix.

5.2 A survey of various confounding models

Here we survey various models of $\boldsymbol{C}(\boldsymbol{h})$ by exploring different bivariate cross covariance models from spatial statistics literature. We note that apart from the model by Page et al. (2017), the models below were not intended to describe the

confounding structure between two random fields in a regression model. Rather, they were analyzed for bivariate random fields where both components are observed and data are readily available for both. In our setting, we make the distinction that one of the components of the bivariate random field $(x(\boldsymbol{s}), e(\boldsymbol{s}))^T$ contains the error, which is unobserved.

5.2.1 Separable model

One of the simplest models to consider is the separable model first introduced by Mardia and Goodall (1993), where $\phi_{xx} = \phi_{xe} = \phi_{ex} = \phi_{ee} = \phi$, that is, , there is a shared correlation function ϕ, among all components of $\boldsymbol{C}(\boldsymbol{h})$,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2\phi(\boldsymbol{h};\boldsymbol{\vartheta}) & \rho\sigma_x\sigma_e\phi(\boldsymbol{h};\boldsymbol{\vartheta}) \\ \rho\sigma_x\sigma_e\phi(\boldsymbol{h};\boldsymbol{\vartheta}) & \sigma_e^2\phi(\boldsymbol{h};\boldsymbol{\vartheta}) \end{pmatrix} \tag{5.9}$$

The spectral condition in Theorem (2.3.1) shows that this cross covariance is valid as long as $|\rho| \leq 1$. When taking into account the n spatial locations, the $2n \times 2n$ covariance matrix from (5.5) is,

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_x^2\boldsymbol{\Phi}(\boldsymbol{\vartheta}) & \rho\sigma_x\sigma_e\boldsymbol{\Phi}(\boldsymbol{\vartheta}) \\ \rho\sigma_x\sigma_e\boldsymbol{\Phi}(\boldsymbol{\vartheta}) & \sigma_e^2\boldsymbol{\Phi}(\boldsymbol{\vartheta}) \end{pmatrix} \tag{5.10}$$

Using properties of the multivariate Gaussian distribution discussed in Section 2.1, the conditional mean and variance of $\boldsymbol{e}|\boldsymbol{x}$ are $\boldsymbol{\tau}(\boldsymbol{\theta}) = \dfrac{\rho\sigma_e}{\sigma_x}\boldsymbol{x}$ and $\boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta}) = \sigma_e^2(1 - \rho^2)\boldsymbol{\Phi}(\boldsymbol{\vartheta})$. Paciorek (2010) used this cross covariance when analyzing the effect of

confounding in a linear regression model $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}$. However, in this linear regression model, we can plug in the formulas for $\boldsymbol{\tau}(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta})$ into (5.6) to obtain,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\beta_0\mathbf{1} + \left(\beta_1 + \frac{\rho\sigma_e}{\sigma_x}\right)\boldsymbol{x}, \sigma_e^2(1-\rho^2)\boldsymbol{\Phi}(\boldsymbol{\vartheta})\right)$$

We see that the parameters $(\beta_1, \rho, \sigma_e^2)$ are not identifiable but $\left(\beta_1 + \frac{\rho\sigma_e}{\sigma_x}, (1-\rho^2)\sigma_e^2\right)$ are, akin to the i.i.d. regression setting discussed in Section 1.2. More specifically, it would make no difference to a practitioner if the conditional distribution were,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\beta_0\mathbf{1} + \beta_1^*\boldsymbol{x}, \sigma^2\boldsymbol{\Phi}(\boldsymbol{\vartheta})\right)$$

where $\beta_1^* = \beta_1 + \frac{\rho\sigma_e}{\sigma_x}$ and $\sigma^2 = (1-\rho^2)\sigma_e^2$. In this formulation, there is no confounding present. Thus, a drawback of the separable model is that it cannot be used to model confounding when the trend function is linear. For a more general nonlinear trend, we have,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) + \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{x}, \sigma_e^2(1-\rho^2)\boldsymbol{\Phi}(\boldsymbol{\vartheta})\right) \tag{5.11}$$

Under appropriate linear independence assumptions between $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta})$ and $\boldsymbol{x}$, this identifiability issue is remedied. For example, $\boldsymbol{f}(\boldsymbol{x},\boldsymbol{\beta})$ should not contain a linear term of the form $\gamma\boldsymbol{x}$ for some constant β. Otherwise, the distribution in (5.11) will contain $\left(\gamma + \frac{\rho\sigma_e}{\sigma_x}\right)\boldsymbol{x}$ in the mean, preventing one from identifying γ and $\rho\sigma_e$ separately. More formally, we argue identifiability in the following proposition. We

remark that the parameter $\boldsymbol{\vartheta}$ is identifiable using the $\boldsymbol{x}$ data by Assumption 5.1.1.

Proposition 5.2.1. *Assume that the trend $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\beta})$ in (5.11) does not contain a linear term of the form $\gamma\boldsymbol{x}$ for some constant γ. Moreover, assume that for any $\boldsymbol{x}$, $\boldsymbol{f}$ is one-to-one respect to the regression parameters $\boldsymbol{\beta}$. Then the parameters in (5.11) are identifiable.*

Proof. Let $\boldsymbol{\Omega}_i = (\boldsymbol{\beta}_i^T, \rho_i, \sigma_{ei}^2)^T, i = 1, 2$ be two possible sets of parameters in (5.11). For identifiability, it suffices to consider the conditional moments separately. For the conditional mean, the condition $\mathbb{E}_{\boldsymbol{\Omega}_1}[\boldsymbol{y}|\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{\Omega}_2}[\boldsymbol{y}|\boldsymbol{x}]$ implies that after re-arranging,

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\beta}_1) - \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\beta}_2) = \frac{1}{\sigma_x}(\rho_2\sigma_{e2} - \rho_1\sigma_{e1})\boldsymbol{x}$$

By the assumptions on $\boldsymbol{f}$, this can only hold for all $\boldsymbol{x}$ if $\rho_2\sigma_{e2} = \rho_1\sigma_{e1}$ and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. Finally, setting the conditional variances equal implies that $\sigma_{e1}^2 = \sigma_{e2}^2$ and $\rho_1 = \rho_2$ since we've identified $\rho\sigma_e$. □

5.2.2 Linear model of co-regionalization (LMC)

The linear model of co-regionalization (Banerjee et al. (2014)) assumes that $(x(\boldsymbol{s}), e(\boldsymbol{s}))^T$ has the form,

$$\begin{pmatrix} x(\boldsymbol{s}) \\ e(\boldsymbol{s}) \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} u(\boldsymbol{s}) \\ v(\boldsymbol{s}) \end{pmatrix}$$

where $u(\boldsymbol{s})$ and $v(\boldsymbol{s})$ are independent Gaussian random fields with correlation functions $\phi_1(\boldsymbol{h}; \boldsymbol{\vartheta}_1)$ and $\phi_2(\boldsymbol{h}; \boldsymbol{\vartheta}_2)$ respectively and the coefficient matrix has full rank.

The resulting cross-covariance function is,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} a_{11}^2\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) + a_{12}^2\phi_2(\boldsymbol{h};\boldsymbol{\vartheta}_2) & a_{11}a_{21}\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) + a_{12}a_{22}\phi_2(\boldsymbol{h};\boldsymbol{\vartheta}_2) \\ a_{11}a_{21}\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) + a_{12}a_{22}\phi_2(\boldsymbol{h};\boldsymbol{\vartheta}_2) & a_{21}^2\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) + a_{22}^2\phi_2(\boldsymbol{h};\boldsymbol{\vartheta}_2) \end{pmatrix}$$

We would like to specify the parameters of $x(\boldsymbol{s})$ since we are observing data for this random field. For this purpose, consider the lower triangular case (that is, $a_{12} = 0$) where $a_{11} = \sigma_x$. Letting $a_{21} = \rho\sigma_e$ and $a_{22} = \sqrt{1-\rho^2}\sigma_e$, we obtain the following cross-covariance function,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) & \rho\sigma_x\sigma_e\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) \\ \rho\sigma_x\sigma_e\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) & \rho^2\sigma_e^2\phi_1(\boldsymbol{h};\boldsymbol{\vartheta}_1) + (1-\rho^2)\sigma_e^2\phi_2(\boldsymbol{h};\boldsymbol{\vartheta}_2) \end{pmatrix} \tag{5.12}$$

Once again, Cramér's criterion in Theorem 2.3.1 is satisfied as long as $|\rho| \leq 1$. This model has a very simple interpretation in that $e(\boldsymbol{s})$ is linearly correlated with $x(\boldsymbol{s})$ through the relationship, $e(\boldsymbol{s}) = \frac{\rho\sigma_e}{\sigma_x}x(\boldsymbol{s}) + \sigma_e\sqrt{1-\rho^2}v(\boldsymbol{s})$. The corresponding $2n \times 2n$ covariance matrix from (5.5) is,

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_x^2\boldsymbol{\Phi}_1(\boldsymbol{\vartheta}_1) & \rho\sigma_x\sigma_e\boldsymbol{\Phi}_1(\boldsymbol{\vartheta}_1) \\ \rho\sigma_x\sigma_e\boldsymbol{\Phi}_1(\boldsymbol{\vartheta}_1) & \rho^2\sigma_e^2\boldsymbol{\Phi}_1(\boldsymbol{\vartheta}_1) + (1-\rho^2)\sigma_e^2\boldsymbol{\Phi}_2(\boldsymbol{\vartheta}_2) \end{pmatrix} \tag{5.13}$$

where $\boldsymbol{\Phi}_2(\boldsymbol{\vartheta}_2)$ is correlation matrix of the vector, $\boldsymbol{v} = (v(\boldsymbol{s}_1), ..., v(\boldsymbol{s}_n))^T$. The conditional mean and variance of $\boldsymbol{e}|\boldsymbol{x}$ are $\boldsymbol{\tau}(\boldsymbol{\theta}) = \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{x}$ and $\boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta}) = \sigma_e^2(1 -$

$\rho^2)\boldsymbol{\Phi}_2(\boldsymbol{\vartheta}_2)$ respectively. Thus, the resulting conditional distribution in (5.6) is,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) + \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{x}, \sigma_e^2(1-\rho^2)\boldsymbol{\Phi}_2(\boldsymbol{\vartheta}_2)\right) \tag{5.14}$$

Note the similarity between this model and the separable model in (5.11). As in the Proposition 5.2.1 for the separable model, the parameters can be made identifiable as long as $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta})$ does not contain a linear term in $\boldsymbol{x}$. Thus, for the same reason as the separable model, the LMC cannot be used to study confounding in a linear regression model due to lack of identifiability.

5.2.3 Bivariate Matérn model

If we choose to work with marginal Matérn random fields, then an appropriate cross-covariance model to use would be the bivariate Matérn model (Gneiting et al. (2010)). The cross-covariance function is of the form,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2 M(\boldsymbol{h};\theta_x,\nu_x) & \rho\sigma_x\sigma_e M(\boldsymbol{h};\theta,\nu) \\ \rho\sigma_x\sigma_e M(\boldsymbol{h};\theta,\nu) & \sigma_e^2 M(\boldsymbol{h};\theta_e,\nu_e) \end{pmatrix} \tag{5.15}$$

where $M(\boldsymbol{h};\theta,\nu) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\theta||\boldsymbol{h}||)^\nu K_\nu(\sqrt{2\nu}\theta||\boldsymbol{h}||)$ is the Matérn kernel from Section 2.2.1. This model is flexible in that both marginals and the cross-covariance functions can be separately parametrized. However, this flexibility comes at a cost of parameter restrictions. Using Cramér's theorem, Gneiting et al. (2010) list a set of conditions needed for this model to satisfy the validity criterion. These conditions involve very complicated constraints among the parameters, unless one chooses to

make some extreme assumptions. For example, in the case of $d = 2$, equal scale parameters $\theta_x = \theta_e = \theta$, and $\nu = \frac{1}{2}(\nu_x + \nu_e)$, Gneiting et al. (2010) show that this model is valid if and only if $|\rho| \leq \frac{\sqrt{\nu_x \nu_e}}{\nu}$. Thus, the confounding parameter lies inside an interval smaller than $[-1, 1]$, unless we take $\nu_x = \nu_e$ (but this reduces back to the separable model). The $2n \times 2n$ covariance matrix from (5.5) is,

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_x^2 \boldsymbol{\Phi}_x(\theta_x) & \rho\sigma_x\sigma_e \boldsymbol{\Phi}(\theta) \\ \rho\sigma_x\sigma_e \boldsymbol{\Phi}(\theta) & \sigma_e^2 \boldsymbol{\Phi}_e(\theta_e) \end{pmatrix} \tag{5.16}$$

The conditional mean and variance of $\boldsymbol{x}|\boldsymbol{e}$ is $\boldsymbol{\tau}(\boldsymbol{\theta}) = \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{\Phi}(\theta)\boldsymbol{\Phi}_x^{-1}(\theta_x)\boldsymbol{x}$ and $\boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta}) = \sigma_e^2\boldsymbol{\Phi}(\boldsymbol{\theta}_e) - \boldsymbol{\Phi}(\theta)\boldsymbol{\Phi}_x^{-1}(\theta_x)\boldsymbol{\Phi}(\theta)$. Then, the form of the conditional likelihood is given in (5.6) is given by,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) + \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{\Phi}(\theta)\boldsymbol{\Phi}_x^{-1}(\theta_x)\boldsymbol{x}, \sigma_e^2\boldsymbol{\Phi}(\boldsymbol{\theta}_e) - \rho^2\sigma_e^2\boldsymbol{\Phi}(\theta)\boldsymbol{\Phi}_x^{-1}(\theta_x)\boldsymbol{\Phi}(\theta)\right) \tag{5.17}$$

In either the linear or nonlinear regression models, it is difficult to see if the parameters are formally identifiable based on the above distribution. We believe that the parameters are identifiable based on our numerical study in Chapter 6.

5.2.4 Markov model

Another type of model that we can consider is the so called Markov model (Journel (1999)),

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2 \phi_x(\boldsymbol{h}; \boldsymbol{\theta}_x) & \rho \sigma_x \sigma_e \phi_x(\boldsymbol{h}; \boldsymbol{\theta}_x) \\ \rho \sigma_x \sigma_e \phi_x(\boldsymbol{h}; \boldsymbol{\theta}_x) & \sigma_e^2 \phi_e(\boldsymbol{h}; \boldsymbol{\theta}_e) \end{pmatrix} \tag{5.18}$$

where ϕ_x and ϕ_e are the correlation functions of $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ respectively. If we consider Matérn covariances only, one can see that this is a special case of the bivariate Matérn model where the off-diagonal Matern correlations are equal to the Matern correlation of $x(\boldsymbol{s})$. The term "Markov" model was coined from the following screening hypothesis,

$$\mathbb{E}[e(\boldsymbol{s})|x(\boldsymbol{s}) = x_1, x(\boldsymbol{t}) = x_2] = \mathbb{E}[e(\boldsymbol{s})|x(\boldsymbol{s}) = x_1] \quad \forall \boldsymbol{s}, \boldsymbol{t}$$

that is, the effect of the random field $x(\boldsymbol{s})$ at any location other than $\boldsymbol{s}$ gets screened out. As a consequence of the bivariate Gaussian assumption, this conditional expectation equals $\mathbb{E}[e(\boldsymbol{s})|x(\boldsymbol{s})] = Cx(\boldsymbol{s})$ for some constant C. From the Markov property and Gaussianity, Journel (1999) shows that (5.18) is a consequence of these assumptions with the constant $C = \rho \sigma_x \sigma_e$. By Cramér's theorem, $\boldsymbol{C}(\boldsymbol{h})$ is valid if and only if $f_{ee}(\boldsymbol{\omega}; \boldsymbol{\theta}_e) \geq \rho^2 f_{xx}(\boldsymbol{\omega}; \boldsymbol{\theta}_x)$ for almost all $\boldsymbol{\omega}$, where f_{xx} and f_{ee} are the spectral densities of $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ respectively. Once again, if one decides to use Matérn covariances, Gneiting et al. (2010) give a list of parameter restrictions that can be simplified for this model. The corresponding $2n \times 2n$ covariance matrix from (5.5)

is,,

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_x^2 \boldsymbol{\Phi}_x(\boldsymbol{\theta}_x) & \rho\sigma_x\sigma_e \boldsymbol{\Phi}_x(\boldsymbol{\theta}_x) \\ \rho\sigma_x\sigma_e \boldsymbol{\Phi}_x(\boldsymbol{\theta}_x) & \sigma_e^2 \boldsymbol{\Phi}_e(\boldsymbol{\theta}_e) \end{pmatrix} \tag{5.19}$$

The conditional mean and variance is $\boldsymbol{\tau}(\boldsymbol{\theta}) = \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{x}$ and $\boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta}) = \sigma_e^2 \boldsymbol{\Phi}_e(\theta_e) - \sigma_e^2\rho^2 \boldsymbol{\Phi}_x(\theta_x)$ respectively. Our conditional likelihood in (5.6) then has the form,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) + \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{x}, \sigma_e^2 \boldsymbol{\Phi}_e(\boldsymbol{\theta}_e) - \rho^2\sigma_e^2 \boldsymbol{\Phi}_x(\boldsymbol{\theta}_x)\right) \tag{5.20}$$

Regarding formal identifiability, all parameters are identifiable in this confounding model under linear independence assumptions on $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta})$ and $\boldsymbol{x}$ similar to Assumption 5.2.1 in the separable model and LMC. However, when the trend is linear $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}$, we have partial identifiability of ρ and β_1 if we restrict the sign of ρ.

Proposition 5.2.2. *Suppose the regression model (5.1) is linear, $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}$. Suppose the following assumptions hold.*

(A1) (Linear independence of ϕ_x, ϕ_e) For any $\boldsymbol{\theta}_{e1}, \boldsymbol{\theta}_{e2}$ and fixed $\boldsymbol{h} \in \mathbb{R}^d$, there does not exist $a, b \in \mathbb{R}$ such that $\phi_x(\boldsymbol{h};\boldsymbol{\theta}_x) = a\phi_e(\boldsymbol{h};\boldsymbol{\theta}_{e1}) + b\phi_e(\boldsymbol{h};\boldsymbol{\theta}_{e2})$.

(A2) (Injectivity of ϕ_e) For any fixed $\boldsymbol{h} \in \mathbb{R}^d$, $\phi_e(\boldsymbol{h};\boldsymbol{\theta}_e)$ is one-to-one with respect to $\boldsymbol{\theta}_e$.

(A3) (Sign of ρ) The correlation parameter ρ is known in advance to be either non-positive or to be non-negative.

Then the parameters in (5.20) are identifiable.

As an example of assumptions (A1), (A2), take ϕ_x and ϕ_e both to be exponential, $\phi_x(\boldsymbol{h};\theta_x) = e^{-\theta_x||\boldsymbol{h}||}$, $\phi_e(\boldsymbol{h};\theta_e) = e^{-\theta_e||\boldsymbol{h}||}$. Then if either $\theta_x < \theta_{e1}, \theta_{e2}$ or $\theta_x > \theta_{e1}, \theta_{e2}$, the assumptions hold. Assumption (A3) allows us to identify the parameters ρ and β_1 separately.

Proof. Let $\boldsymbol{\Omega}_i = (\beta_{0i}, \beta_{1i}, \rho_i, \sigma^2_{e_i}, \boldsymbol{\theta}_{e_i}), i = 1, 2$ represent two sets of parameters. To show that the likelihood in (5.20) is one-to-one with respect to $\boldsymbol{\Omega}$, it suffices to consider the conditional moments separately. In particular, $\mathbb{E}_{\boldsymbol{\Omega}_1}[\boldsymbol{y}|\boldsymbol{x}] = \mathbb{E}_{\boldsymbol{\Omega}_2}[\boldsymbol{y}|\boldsymbol{x}]$ implies that,

$$\beta_{01}\mathbf{1} + \left(\beta_{11} + \frac{\rho_1\sigma_{e_1}}{\sigma_x}\right)\boldsymbol{x} = \beta_{02}\mathbf{1} + \left(\beta_{21} + \frac{\rho_2\sigma_{e_2}}{\sigma_x}\right)\boldsymbol{x}$$

Since $\boldsymbol{x}$ and $\mathbf{1}$ are linearly independent (i.e, the design matrix $\boldsymbol{X}$ is full rank), it follows that β_0 and $\beta_1 + \frac{\rho\sigma_e}{\sigma_x}$ are identifiable. Next, $\mathrm{Var}_{\boldsymbol{\Omega}_1}(\boldsymbol{y}|\boldsymbol{x}) = \mathrm{Var}_{\boldsymbol{\Omega}_1}(\boldsymbol{y}|\boldsymbol{x})$ implies that,

$$\sigma^2_{e_1}\boldsymbol{\Phi}_e(\boldsymbol{\theta}_{e_1}) - \rho_1^2\sigma^2_{e_1}\boldsymbol{\Phi}_x(\boldsymbol{\theta}_x) = \sigma^2_{e_2}\boldsymbol{\Phi}_e(\boldsymbol{\theta}_{e_2}) - \rho_2^2\sigma^2_{e_2}\boldsymbol{\Phi}_x(\boldsymbol{\theta}_x)$$

The diagonal elements of these matrices imply that, $\rho_1^2(1-\sigma^2_{e1}) = \rho_2^2(1-\sigma^2_{e2})$ and so $\rho^2(1-\sigma_e^2)$ is identifiable. The off-diagonal terms of these matrices imply that,

$$(\rho_1^2\sigma^2_{e_2} - \rho_2^2\sigma^2_{e_2})\phi_x(\boldsymbol{s}_i - \boldsymbol{s}_j;\boldsymbol{\theta}_x) = \sigma^2_{e_1}\phi_e(\boldsymbol{s}_i - \boldsymbol{s}_j;\boldsymbol{\theta}_{e1}) - \sigma^2_{e_2}\phi_e(\boldsymbol{s}_i - \boldsymbol{s}_j;\boldsymbol{\theta}_{e2})$$

By assumption (A1), this can only occur if $\rho_1^2\sigma_{e_2}^2 = \rho_2^2\sigma_{e_2}^2$ and thus $\rho^2\sigma_e^2$ is identifiable. Together with $\rho^2(1-\sigma_e^2)$, this implies that σ_e^2 and ρ^2 are identifiable. If we restrict the sign of ρ as in assumption (A3), then ρ is identifiable, which in turn, implies that β_1 is identifiable. Finally, if we set the left hand side to 0 above, then $\phi_e(||\boldsymbol{s}_i - \boldsymbol{s}_j||; \boldsymbol{\theta}_{e1}) = \phi_e(||\boldsymbol{s}_i - \boldsymbol{s}_j||; \boldsymbol{\theta}_{e2})$ which implies that $\boldsymbol{\theta}_{e_1} = \boldsymbol{\theta}_{e_2}$ by Assumption (A2). □

We note that because of the restriction of ρ to be non-positive or non-negative, caution should be exercised when applying hypothesis tests for $\rho = 0$ in the linear regression model. This sign restriction cannot in general be relaxed. In the linear regression model, the conditional likelihood in (5.6) has the form,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\beta_0\mathbf{1} + \left(\beta_1 + \frac{\rho\sigma_e}{\sigma_x}\right)\boldsymbol{x}, \sigma_e^2\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e) - \rho^2\sigma_e^2\boldsymbol{\Phi}_x(\boldsymbol{\theta}_x)\right)$$

There are different pairs (β_1, ρ) that lead to the same the likelihood function, for example, (β_1, ρ) and $\left(\beta_1 + 2\frac{\rho\sigma_e}{\sigma_x}, -\rho\right)$.

5.2.5 Page et al model

Page et al. (2017) take a different approach to modeling the cross-covariance. Instead of assuming a cross-covariance function of the underlying bivariate random field $(x(\boldsymbol{s}), e(\boldsymbol{s}))^T$ in (5.2), they decide to look directly at the corresponding $2n \times 2n$ covariance matrix from (5.5). They assume the structure,

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_x^2\boldsymbol{\Phi}_x(\boldsymbol{\theta}_x) & \rho\sigma_x\sigma_e\boldsymbol{L}_x(\boldsymbol{\theta}_x)\boldsymbol{L}_e^T(\boldsymbol{\theta}_e) \\ \rho\sigma_x\sigma_e\boldsymbol{L}_e(\boldsymbol{\theta}_e)\boldsymbol{L}_x^T(\boldsymbol{\theta}_x) & \sigma_e^2\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e) \end{pmatrix} \tag{5.21}$$

where $\boldsymbol{L}, \boldsymbol{L}^T$ are lower and upper triangular matrices which come from the Cholesky decompositions of $\boldsymbol{\Phi}_x(\boldsymbol{\theta}_x) = \boldsymbol{L}_x(\boldsymbol{\theta}_x)\boldsymbol{L}_x^T(\boldsymbol{\theta}_x)$ and $\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e) = \boldsymbol{L}_e(\boldsymbol{\theta}_e)\boldsymbol{L}_e^T(\boldsymbol{\theta}_e)$ respectively. The separable model is a special case of this one when $\boldsymbol{\Phi}_x(\boldsymbol{\theta}_x) = \boldsymbol{\Phi}_e(\boldsymbol{\theta}_e)$, thus it differs by allowing one to model the marginals of $x(\boldsymbol{s})$ and $z(\boldsymbol{s})$ separately. By construction, the off-diagonal cross covariance matrices change depending on the number of spatial locations. As a simple example, suppose we have three locations $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \boldsymbol{s}_3\} = \{(0,0),(1,0),(0,1)\}$ in $\mathbb{R}^2$. Take $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ to have exponential correlation functions $\phi_x(\boldsymbol{s}_i, \boldsymbol{s}_j) = e^{-\frac{1}{2}||\boldsymbol{s}_i - \boldsymbol{s}_j||}$ and $\phi_e(\boldsymbol{s}_i, \boldsymbol{s}_j) = e^{-||\boldsymbol{s}_i - \boldsymbol{s}_j||}$ respectively. Then the correlation matrices are,

$$\boldsymbol{\Phi}_x = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 \end{matrix} \\ \begin{bmatrix} 1.00 & 0.61 & 0.61 \\ 0.61 & 1.00 & 0.49 \\ 0.61 & 0.49 & 1.00 \end{bmatrix} \end{array} \begin{array}{c} \\ \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \\ \boldsymbol{s}_3 \end{array} \qquad \boldsymbol{\Phi}_e = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 \end{matrix} \\ \begin{bmatrix} 1.00 & 0.37 & 0.37 \\ 0.37 & 1.00 & 0.24 \\ 0.37 & 0.24 & 1.00 \end{bmatrix} \end{array} \begin{array}{c} \\ \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \\ \boldsymbol{s}_3 \end{array}$$

After computing the Cholesky decompositions, the upper right cross correlation matrix in (5.21) is,

$$\boldsymbol{L}_x\boldsymbol{L}_e^T = \begin{array}{c} \begin{matrix} \boldsymbol{s}_1 & \boldsymbol{s}_2 & \boldsymbol{s}_3 \end{matrix} \\ \begin{bmatrix} 1.00 & 0.37 & 0.37 \\ 0.61 & 0.96 & 0.32 \\ 0.61 & 0.37 & 0.96 \end{bmatrix} \end{array} \begin{array}{c} \\ \boldsymbol{s}_1 \\ \boldsymbol{s}_2 \\ \boldsymbol{s}_3 \end{array}$$

However, if we do the same calculation using only the locations $\{\boldsymbol{s}_2, \boldsymbol{s}_3\}$, we have,

$$\boldsymbol{L}_x\boldsymbol{L}_e^T = \begin{array}{c} \begin{matrix} \boldsymbol{s}_2 & \boldsymbol{s}_3 \end{matrix} \\ \begin{bmatrix} 1.00 & 0.24 \\ 0.49 & 0.96 \end{bmatrix} \end{array} \begin{array}{l} \boldsymbol{s}_2 \\ \boldsymbol{s}_3 \end{array}$$

We see that the exclusion of $\boldsymbol{s}_1$ changes the dependence structure of $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ at the locations $\{\boldsymbol{s}_2, \boldsymbol{s}_3\}$. This type of triangular array structure that depends on the number of spatial locations is used in spatial autoregressive models (Cressie (1993)). In fact, Page et al. (2017) use an autoregressive model in their simulations when studying confounding, which could have been the motivation behind this cross-covariance. Extra care should be taken with this model as datasets are not easily interpretable as the evaluations at a subset of locations from a well-defined random field.

Since we do not have the form of an underlying cross covariance function $\boldsymbol{C}(\boldsymbol{h})$ or spectral density matrix $\boldsymbol{f}(\boldsymbol{\omega})$, Cramér's theorem cannot be used to determine if model is valid. However, we can still determine if the model is valid by directly verifying if the covariance matrix in (5.21) is positive definite. This will hold if and only if the Schur complement is positive definite (see Lemma 7.4.6 in Chapter 7). Since the Schur complement of the above matrix is $\sigma_e^2(1-\rho^2)\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e)$, the above model is valid iff $|\rho| \leq 1$. The conditional mean and variance of $\boldsymbol{e}|\boldsymbol{x}$ are $\boldsymbol{\tau}(\boldsymbol{\theta}) = \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{L}_e(\boldsymbol{\theta}_e)\boldsymbol{L}(\boldsymbol{\theta}_x)^{-1}\boldsymbol{x}$ and $\boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta}) = \sigma_e^2(1-\rho^2)\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e)$ respectively. Then, the form

of the conditional distribution in (5.6) is,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) + \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{L}_e(\boldsymbol{\theta}_e)\boldsymbol{L}_x(\boldsymbol{\theta}_x)^{-1}\boldsymbol{x}, \sigma_e^2(1-\rho^2)\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e)\right) \tag{5.22}$$

For a general nonlinear trend $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta})$, it is difficult to see if the parameters are formally identifiable. We expect it to be true based on our numerical study of this model in Chapter 6. For a linear regression model $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}$, we can argue identifiability under suitable linear independence assumptions.

Proposition 5.2.3. *Suppose that in our regression model (5.1), the trend function is linear, $\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) = \beta_0\mathbf{1} + \beta_1\boldsymbol{x}$. Assume that for any $\boldsymbol{h} \in \mathbb{R}^d$, the following hold,*

1. *$\phi_e(\boldsymbol{h};\boldsymbol{\theta}_e)$ is injective with respect to $\boldsymbol{\theta}_e$*
2. *$\phi_e(\boldsymbol{h};\boldsymbol{\theta}_e)$ and $\phi_x(\boldsymbol{h};\boldsymbol{\theta}_x)$ are linearly independent functions, that is, not scalar multiples of one another.*

Then the parameters $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ are identifiable.

Proof. First note that the intercept is automatically identifiable as the unconditional mean $\mathbb{E}[\boldsymbol{y}] = \beta_0\mathbf{1}$. For the remaining parameters, let $(\beta_{1i}, \rho_i, \sigma_{ei}^2, \boldsymbol{\theta}_{ei}^T)^T, i = 1, 2$ be two possible sets of parameters. Since the conditional distribution $\boldsymbol{y}|\boldsymbol{x}$ in (5.22) is Gaussian, it suffices to show that the conditional moments are one-to-one with respect to the parameters. Setting the conditional variances equal,

$$\sigma_{e1}^2(1-\rho_1^2)\boldsymbol{\Phi}_e(\boldsymbol{\theta}_{e1}) = \sigma_{e2}^2(1-\rho_2^2)\boldsymbol{\Phi}_e(\boldsymbol{\theta}_{e2})$$

If we compare the diagonal elements of these matrices, the displayed equation implies that $\sigma_{e1}^2(1-\rho_1^2) = \sigma_{e2}^2(1-\rho_2^2)$ and so the parameter $\sigma_e^2(1-\rho^2)$ is identifiable. Then the above implies that $\boldsymbol{\Phi}_e(\boldsymbol{\theta}_{e1}) = \boldsymbol{\Phi}_e(\boldsymbol{\theta}_{e2})$ or equivalently $\phi_e(||\boldsymbol{s}_i - \boldsymbol{s}_j||;\boldsymbol{\theta}_{e1}) = \phi_e(||\boldsymbol{s}_i - \boldsymbol{s}_j||;\boldsymbol{\theta}_{e1})$ for $i,j = 1,...,n$. By assumption 1 above, this implies that $\boldsymbol{\theta}_{e1} = \boldsymbol{\theta}_{e1}$ and thus $\boldsymbol{\theta}_e$ is identified. Next, setting the conditional means equal,

$$\beta_{11}\boldsymbol{x} + \frac{\rho_1\sigma_{e1}}{\sigma_x}\boldsymbol{L}_e(\boldsymbol{\theta}_e)\boldsymbol{L}_x(\boldsymbol{\theta}_x)^{-1}\boldsymbol{x} = \beta_{12}\boldsymbol{x} + \frac{\rho_2\sigma_{e2}}{\sigma_x}\boldsymbol{L}_e(\boldsymbol{\theta}_e)\boldsymbol{L}_x(\boldsymbol{\theta}_x)^{-1}\boldsymbol{x}$$

Setting $\boldsymbol{\gamma} = \frac{1}{\sigma_x}\boldsymbol{L}_x^{-1}(\boldsymbol{\theta}_e)\boldsymbol{x}$, we can rearrange the above expression as,

$$(\beta_{11}-\beta_{12})\sigma_x\boldsymbol{L}_x(\boldsymbol{\theta}_x)\boldsymbol{\gamma} + (\rho_1\sigma_{e1} - \rho_2\sigma_{e2})\boldsymbol{L}_e(\boldsymbol{\theta}_e)\boldsymbol{\gamma} = \boldsymbol{0}$$

By assumption 2 above, since the correlation functions $\phi_e(\boldsymbol{h};\boldsymbol{\theta}_e)$ and $\phi_x(\boldsymbol{h};\boldsymbol{\theta}_x)$ are linearly independent, this implies that the corresponding matrices $\boldsymbol{L}_e(\boldsymbol{\theta}_e)$ and $\boldsymbol{L}_x(\boldsymbol{\theta}_x)$ are linearly independent (as vectors in the space of real $n \times n$ matrices). Thus, $\beta_{11} = \beta_{12}$ and $\rho_1\sigma_{e1} = \rho_2\sigma_{e2}$ and so β_1 and $\rho\sigma_e$ are identifiable. Together with $\sigma_e^2(1-\rho^2)$ being identifiable, this implies that ρ and σ_e^2 are identifiable. □

5.2.6 Asymmetric Markov model

The final model we consider was inspired by the Markov model with a slight modification,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2\phi_x(\boldsymbol{h};\boldsymbol{\theta}_x) & \rho\sigma_x\sigma_e\phi_x(\boldsymbol{h};\boldsymbol{\theta}_x) \\ \rho\sigma_x\sigma_e\phi_e(\boldsymbol{h};\boldsymbol{\theta}_x) & \sigma_e^2\phi_e(\boldsymbol{h};\boldsymbol{\theta}_e) \end{pmatrix} \tag{5.23}$$

Unlike the Markov model in (5.18), this cross covariance function is asymmetric as one of the off-diagonal correlation functions is that of the error. Cramér's Theorem shows that this cross covariance is valid if $|\rho| \leq 1$. Thus, this model does not have parameter restrictions found in the Markov model. The corresponding $2n \times 2n$ covariance matrix from (5.5) is,,

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma_x^2 \boldsymbol{\Phi}_x(\theta_x) & \rho\sigma_x\sigma_e \boldsymbol{\Phi}_x(\theta_x) \\ \rho\sigma_x\sigma_e \boldsymbol{\Phi}_e(\theta_e) & \sigma_e^2 \boldsymbol{\Phi}_e(\theta_e) \end{pmatrix}$$

The conditional mean and variance of $\boldsymbol{e}|\boldsymbol{x}$ are $\boldsymbol{\tau}(\boldsymbol{\theta}) = \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{\Phi}_e(\theta_z)\boldsymbol{\Phi}_x^{-1}(\theta_x)\boldsymbol{x}$ and $\boldsymbol{\Sigma}_{e|x}(\boldsymbol{\theta}) = \sigma_e^2(1-\rho^2)\boldsymbol{\Phi}_e(\theta_e)$ respectively. Then the conditional distribution of $\boldsymbol{y}|\boldsymbol{x}$ is,

$$\boldsymbol{y}|\boldsymbol{x} \sim N\left(\boldsymbol{f}(\boldsymbol{x};\boldsymbol{\beta}) + \frac{\rho\sigma_e}{\sigma_x}\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e)\boldsymbol{\Phi}_x(\boldsymbol{\theta}_x)^{-1}\boldsymbol{x}, \sigma_e^2(1-\rho^2)\boldsymbol{\Phi}_e(\boldsymbol{\theta}_e)\right) \tag{5.24}$$

We note the similarity between this conditional distribution and that of the Page confounding model in (5.22). Due to the similar structure as the Page model, formal identifiability arguments transfer here. Specifically, all parameters are identifiable under in both the linear and nonlinear regression models. In the linear regression model, we refer to Proposition 5.2.3 and its proof for details, replacing the Cholesky decompositions $\boldsymbol{L}_x$ and $\boldsymbol{L}_e$ with the correlation matrices $\boldsymbol{\Phi}_x$ and $\boldsymbol{\Phi}_e$ themselves.

5.2.7 Other confounding models

There are many cross covariance functions in the statistical literature that one may consider as a model of confounding; we gave a description of a select few that

satisfy our identifiability criteria. One other notable model that we omitted is the so-called convolution model (Genton and Kleiber (2015)),

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} (\phi_1 * \phi_1)(\boldsymbol{h}; \boldsymbol{\theta}_x) & (\phi_1 * \phi_2)(\boldsymbol{h}; \boldsymbol{\theta}_x, \boldsymbol{\theta}_e) \\ (\phi_1 * \phi_2)(\boldsymbol{h}; \boldsymbol{\theta}_x, \boldsymbol{\theta}_e) & (\phi_2 * \phi_2)(\boldsymbol{h}; \boldsymbol{\theta}_e) \end{pmatrix}$$

where $(\phi_1 * \phi_2)(\boldsymbol{h}) = \int_{\mathbb{R}^d} \phi_1(\boldsymbol{h} - \boldsymbol{k})\phi_2(\boldsymbol{k})d\boldsymbol{k}$ and ϕ_1, ϕ_2 are correlation functions. The reason for the omission is that in general it is difficult to find closed forms for these convolutions and analyses must typically be done through numerical integrals. A general overview of cross-covariance functions found in spatial statistics literature is given in Genton and Kleiber (2015).

Chapter 6 Numerical study of confounding

6.1 Practical identifiability of confounding models

In Chapter 5, we listed numerical criteria for the practical identifiability of unknown parameters in confounding models. For each of the confounding models listed in Section 5.2, we perform a simulation study based on these criteria to determine if they are suitable for use on real data.

6.1.1 Numerical setup

We use the same spatial locations as in Figure 4.1. In each of the confounding models, we choose exponential covariance functions $C_x(\boldsymbol{h}) = \sigma_x^2 e^{-\theta_x||\boldsymbol{h}||}$ and $C_e(\boldsymbol{h}) = \sigma_e^2 e^{-\theta_e||\boldsymbol{h}||}$ for both $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$. Recall that we treat the covariate parameters $(\sigma_x^2, \theta_x)^T$ as known for our simulations. For our covariate, we arbitrarily choose as the true parameters $(\sigma_x^2, \theta_x)^T = (2, \frac{3}{2})^T$ and generate one realization of $\boldsymbol{x} \sim N(\boldsymbol{0}, \sigma_x^2 \boldsymbol{\Phi}_x(\theta_x))$. For our nonlinear trend, we take $f(x(\boldsymbol{s}); \boldsymbol{\beta}) = \dfrac{\beta_0}{1 + e^{-\beta_1 x(\boldsymbol{s})}}$ with true parameter values of $(\beta_0, \beta_1)^T = (6, 3)^T$. For the remaining covariance parameters and confounding parameter, we arbitrarily take as the true values, $(\rho_0, \sigma_{e0}^2, \theta_{e0})^T = (\frac{1}{4}, 1, \frac{1}{2})^T$. For MC estimation, we generate 1000 realizations of

$\boldsymbol{y}$ conditional on $\boldsymbol{x}$ according to (5.6). For each of these realizations we compute a set of MLE estimates $(\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\theta}}^T)^T$ using the nlminb function in the base R package. This function uses a quasi-Newton method that allows for box-constraints on the parameters. To find evidence that these estimates are true minima, we compute the gradient and Hessian of the negative log-likelihood. We look for small gradient norms and positive eigenvalues of the Hessian. For each set of MC estimates, we plot them as a histogram and overlay the theoretical normal density as predicted by Mardia and Marshall (1984).

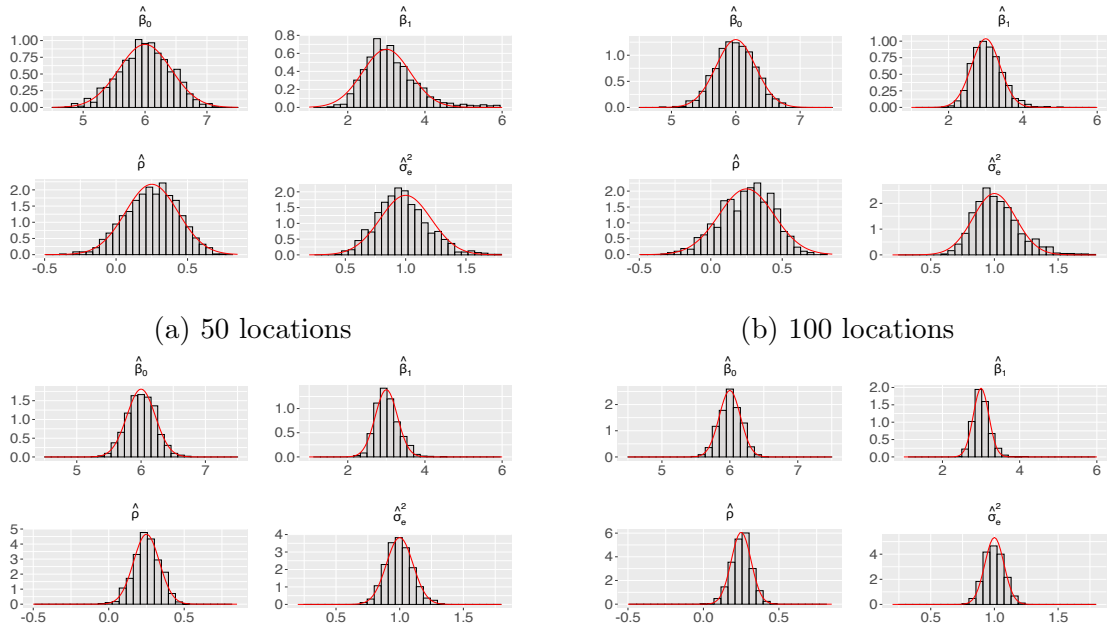
6.1.2 Separable model

Recall that in this model, the covariate and error share the same scale parameter in the exponential correlation function $\phi(\boldsymbol{h}) = e^{-\theta||\boldsymbol{h}||}$. Thus, the scale parameter θ is known here and thus the unknown parameters in this model are $(\beta_0, \beta_1, \rho, \sigma_e^2)^T$. For $n = 50, 100, 200$ and 400 locations, we evaluate the Fisher information matrix for the separable model at these true parameters. The maximum and minimum eigenvalues of the Fisher information and diagonal elements of the inverse Fisher information are given below in Table 6.1. We see that the numbers are generally behaving as expected. By doubling the number of locations, the corresponding elements of the Fisher information and its eigenvalues approximately double as well. The condition numbers remain small as the number of observations increases.

	Eigenvalues		Diagonals of inverse Fisher information			
n	$\lambda_{\max}$	$\lambda_{\min}$	β_0	β_1	ρ	σ_e^2
50	60.78	2.15	0.17289	0.40290	0.04428	0.04759
100	133.80	4.82	0.09069	0.18118	0.01852	0.02277
200	293.24	10.05	0.04656	0.08774	0.00783	0.01098
400	526.48	19.96	0.02506	0.04177	0.00476	0.00573

Table 6.1: Fisher information analysis of the separable model

For MC estimation, we generate 1000 realizations of $\boldsymbol{y}|\boldsymbol{x}$ according to the conditional distribution given in (5.11) and then compute MLE estimates for each. From a separate gradient and Hessian analysis, there is no evidence of large gradients or negative eigenvalues in the Hessians. The histograms of these estimates along with their theoretical densities overlayed are given in Figure 6.1. The MC estimates closely match the theoretical asymptotic normal densities.



(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

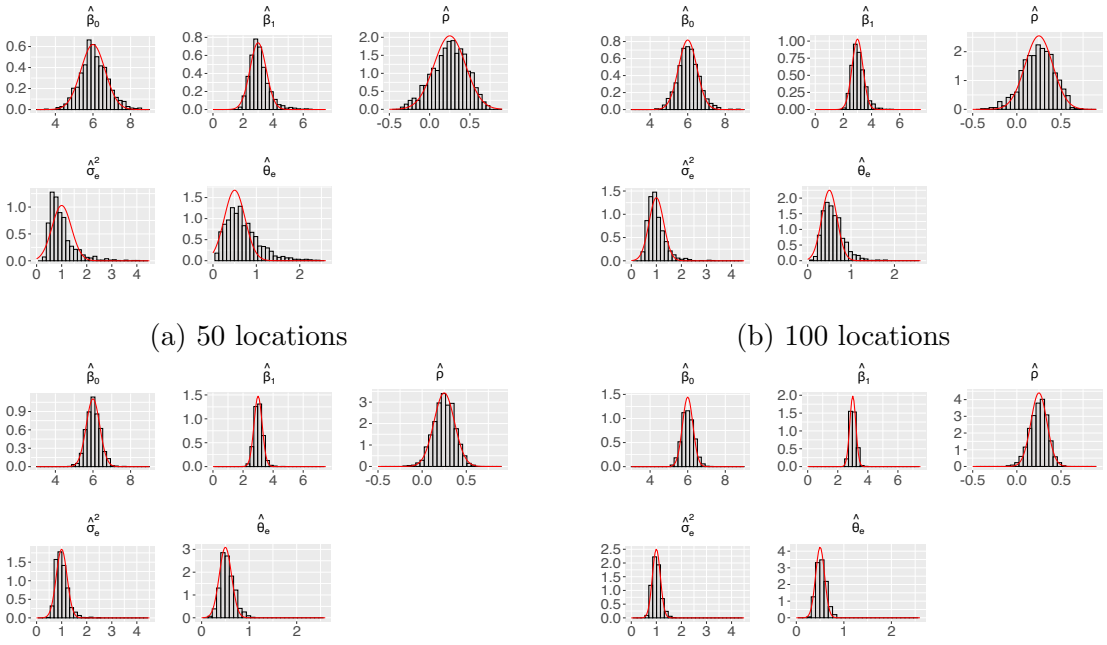
Figure 6.1: Separable model MLE estimates along with the theoretical densities (red). Histograms are based on 1000 simulations of $\boldsymbol{y}|\boldsymbol{x}$ from the model (5.11).

6.1.3 Linear model of coregionalization (LMC)

For the LMC, the marginals of $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ are separately parametrized and do not share a correlation function. The unknown parameters in this model are $(\beta_0, \beta_1, \rho, \sigma_e^2, \theta_e)^T$, with the additional unknown scale parameter θ_e compared to the separable model. The eigenvalues and diagonal elements of the inverse Fisher information matrix are presented in Table 6.2 below.

	Eigenvalues		Diagonals of inverse Fisher information				
n	$\lambda_{\max}$	$\lambda_{\min}$	β_0	β_1	ρ	σ_e^2	θ_e
50	148.62	1.73	0.39418	0.22838	0.04261	0.15117	0.05665
100	243.95	2.55	0.26809	0.13924	0.03291	0.08867	0.03133
200	478.58	5.04	0.13462	0.07225	0.01668	0.04740	0.01660
400	1238.69	10.20	0.06369	0.03864	0.00637	0.02482	0.00884

Table 6.2: Fisher information analysis of the LMC



(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

Figure 6.2: LMC MLE estimates along with the theoretical densities (red). Histograms are based on 1000 simulations of $\boldsymbol{y}|\boldsymbol{x}$ from the model (5.14).

The condition numbers and variances are slightly larger than the separable model, but remain relatively small. Moreover, the numbers show expected scaling behavior with an increasing number of observations. For MC estimation, we generate 1000 realizations of $\boldsymbol{y}|\boldsymbol{x}$ according to the conditional distribution given in (5.14) and then compute MLE estimates for each. Once again, we find no evidence of large gradients or negative eigenvalues of the Hessian matrices. The histogram of these estimates along with their theoretical densities overlayed are given in Figure 6.2. The MC estimates generally are well approximated by their theoretical asymptotic normal densities. This is less evident for $n = 50$, especially for the parameters σ_e^2 and θ_e. However, we get expected asymptotic behavior for larger n.

6.1.4 Bivariate Matérn model

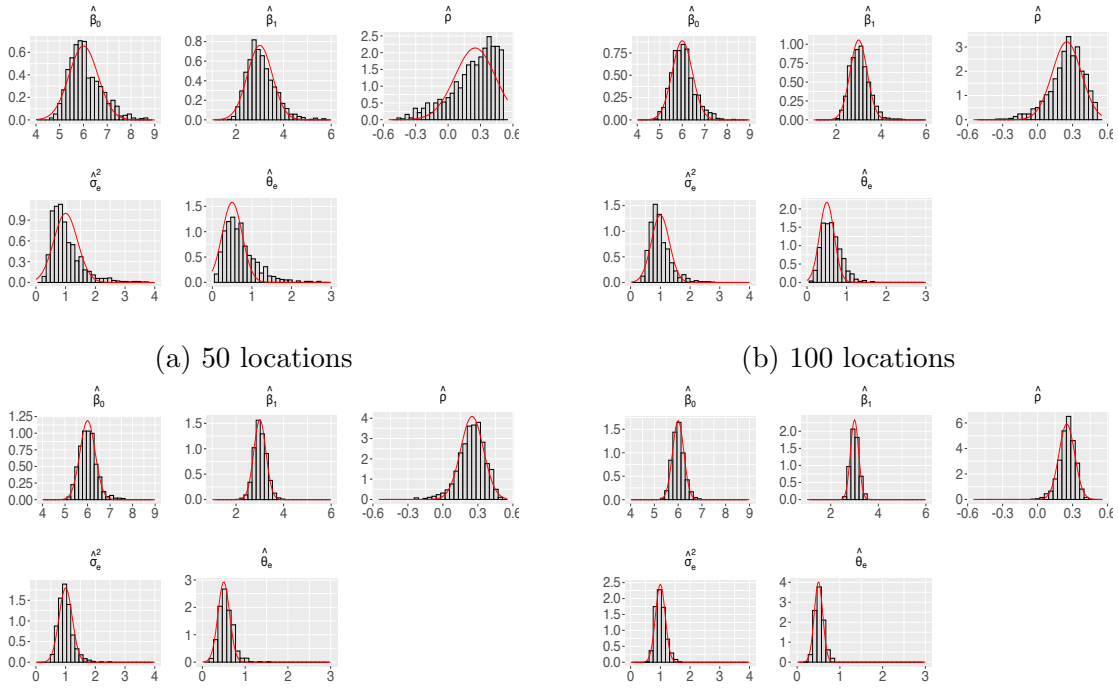
For our simulations, we take the cross correlation function between $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ to also be exponential $\phi(\boldsymbol{h};\theta) = e^{-\theta||\boldsymbol{h}||}$, with $\theta = \theta_x + \theta_e$. From Gneiting et al. (2010), this leads to the nonlinear constraint, $|\rho| \leq \frac{\sqrt{\theta_x\theta_e}}{\theta_x+\theta_e}$. The eigenvalues and diagonal elements of the inverse Fisher information are presented in Table 6.3.

	Eigenvalues		Diagonals of inverse Fisher information				
n	$\lambda_{\max}$	$\lambda_{\min}$	β_0	β_1	ρ	σ_e^2	θ_e
50	260.65	1.86	0.37458	0.18719	0.03491	0.16169	0.06281
100	529.49	3.75	0.18290	0.10431	0.01641	0.09155	0.03412
200	1088.68	6.27	0.10839	0.06162	0.00912	0.04919	0.01847
400	2298.51	13.30	0.04930	0.03063	0.00386	0.02630	0.00949

Table 6.3: Fisher information analysis of the bivariate Matérn model

The condition numbers are larger than the ones in the separable model and LMC,

but still have reasonable magnitude. The variances scale as expected with increasing number of observations. For MC estimation, we generate 1000 observations according to $\boldsymbol{y}|\boldsymbol{x}$ given in (5.17) and compute MLE estimates for each. The nonlinear constraint above is made into a regular box constraint using the reparametrization $\gamma = \dfrac{\rho}{\dfrac{\sqrt{\theta_x\theta_e}}{\theta_x+\theta_e}}$ where $|\gamma| \leq 1$. Finally, we recover the MLE estimate of ρ with the transformation $\hat{\rho} = \hat{\gamma}\dfrac{\sqrt{\theta_x\hat{\theta}_e}}{\theta_x+\hat{\theta}_e}$. The histograms of these estimates along with their theoretical densities are given in Figure 6.3. The estimates for ρ and θ_e are not very well resolved in the $n = 50$ case, but improve for $n = 100, 200, 400$.



(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

Figure 6.3: Bivariate Matérn model MLE estimates along with the theoretical densities (red). Histograms are based on 1000 simulations of $\boldsymbol{y}|\boldsymbol{x}$ from the model (5.17).

This corresponds with our gradient and Hessian analysis, where a small proportion of estimates are found to be on the boundary of the parameter space when $n = 50$.

This proportion becomes smaller as the number of observations double. For sample sizes larger than $n = 100$, the estimates seem to be closely approximated by their theoretical asymptotic densities.

6.1.5 Markov model

When $x(\boldsymbol{s})$ and $z(\boldsymbol{s})$ both have exponential covariances, the constraint for this model becomes either of the following two conditions (Gneiting et al. (2010)),

$$\begin{cases} \rho^2 \leq \dfrac{\theta_e}{\theta_x} & \text{if } \theta_e < \theta_x \\ \rho^2 \leq \left(\dfrac{\theta_x}{\theta_e}\right)^2 & \text{if } \theta_e \geq \theta_x \end{cases}$$

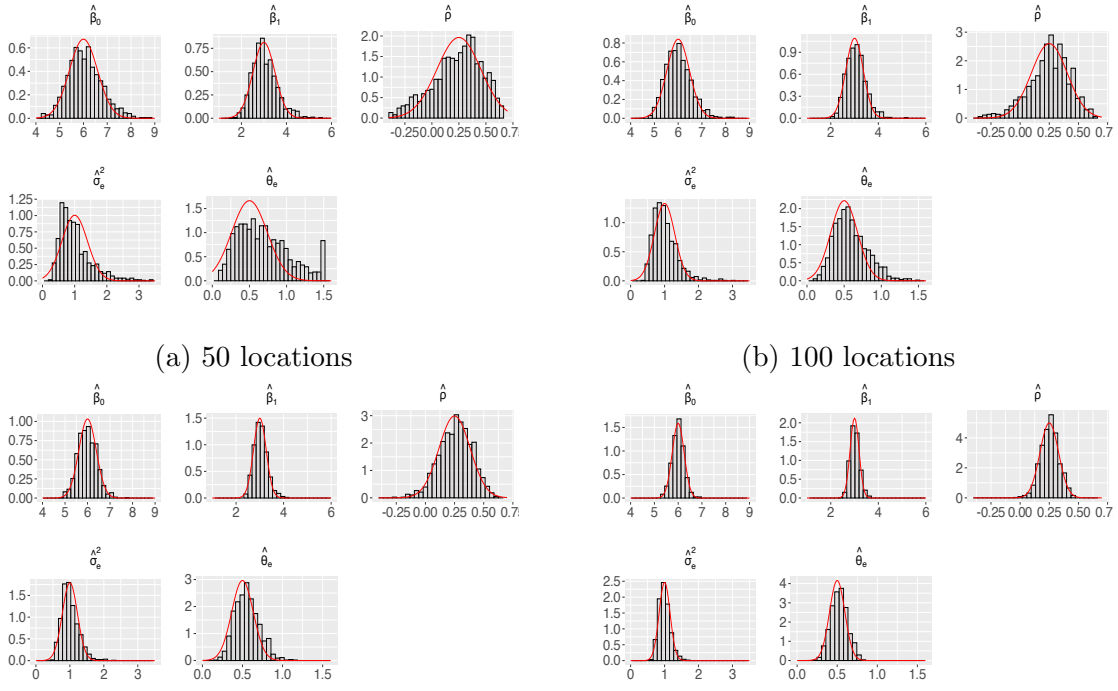
For our simulations, we choose the first condition where $\theta_e < \theta_x$. The other condition yields comparable results and so the analysis is not shown here. We note that our choice of true parameters does satisfy this constraint. The eigenvalues and diagonal elements of the inverse Fisher information are presented in Tables 6.4.

	Eigenvalues		Diagonals of inverse Fisher information				
n	$\lambda_{\max}$	$\lambda_{\min}$	β_0	β_1	ρ	σ_e^2	θ_e
50	235.16	1.58	0.37854	0.32518	0.04138	0.15767	0.05755
100	428.21	2.97	0.22516	0.13303	0.02210	0.08965	0.03205
200	792.10	4.89	0.13842	0.07253	0.01582	0.04904	0.01776
400	1646.67	10.71	0.06246	0.03445	0.00698	0.02613	0.00928

Table 6.4: Fisher information analysis of the Markov model

The condition numbers and variances are comparable to the bivariate Matérn. For MC estimation, we generate 1000 observations according to $\boldsymbol{y}|\boldsymbol{x}$ in (5.20) and compute MLE estimates for each. Similar to the bivariate Matérn model, the nonlinear

constraint above can made into a regular box constraint using the reparametrization $\gamma = \dfrac{\rho}{\sqrt{\dfrac{\theta_e}{\theta_x}}}$. Then we use the nlminb function in R with the box constraints $|\gamma| \leq 1$ and $\theta_e < \theta_x$. Finally, we recover the MLE estimate of ρ with, $\hat{\rho} = \hat{\gamma}\sqrt{\dfrac{\hat{\theta}_e}{\theta_x}}$. The histogram of these estimates along with their theoretical densities overlayed are given in Figure 6.4. There is a noticeable anomaly in the MC estimates for θ_e when $n = 50$. The nlminb function estimated these parameters to be on the boundary $\theta_e = \theta_x$. This behavior disappears once we get to $n = 100$ and larger where the estimates are well approximated by their theoretical asymptotic densities.



(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

Figure 6.4: Markov model MLE estimates along with the theoretical densities (red). Histograms are based on 1000 simulations of $\boldsymbol{y}|\boldsymbol{x}$ from the model (5.20).

6.1.6 Page et al. and asymmetric Markov models

Since the Page et al. and asymmetric Markov models display similar numerics, we present them both here. The eigenvalues and diagonal elements of the inverse Fisher information for both models are presented in Tables 6.5 and 6.6 respectively.

	Eigenvalues		Diagonals of inverse Fisher information				
n	$\lambda_{\max}$	$\lambda_{\min}$	β_0	β_1	ρ	σ_e^2	θ_e
50	97.32	2.18	0.27290	0.23831	0.05944	0.15393	0.05523
100	200.23	3.47	0.18917	0.11988	0.04059	0.08980	0.03011
200	407.29	6.70	0.08929	0.06440	0.01738	0.04703	0.01549
400	820.90	11.57	0.05337	0.03298	0.01242	0.02667	0.00855

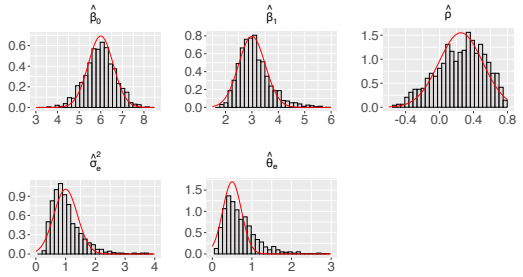
Table 6.5: Fisher information analysis of the Page et al. model

	Eigenvalues		Diagonals of inverse Fisher information				
n	$\lambda_{\max}$	$\lambda_{\min}$	β_0	β_1	ρ	σ_e^2	θ_e
50	103.33	1.78	0.30264	0.27651	0.07343	0.12068	0.04235
100	204.17	4.47	0.12914	0.11387	0.03731	0.07826	0.02723
200	412.95	8.36	0.06756	0.05722	0.01891	0.04294	0.01439
400	834.01	17.92	0.03217	0.02761	0.00902	0.02453	0.00820

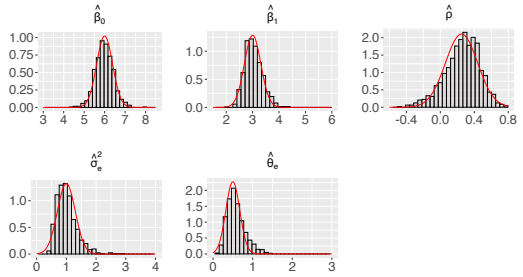
Table 6.6: Fisher information analysis of the asymmetric Markov model

The eigenvalues show no sign of ill-conditioned Fisher information matrices. Moreover, the numbers show expected scaling as the number of observations double.

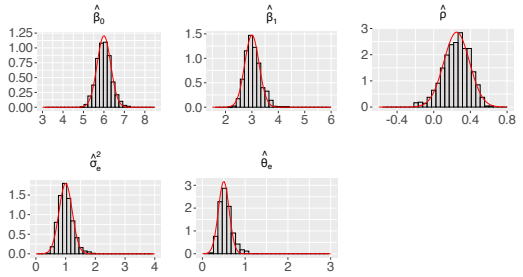
For MC estimation, we generate 1000 observations $\boldsymbol{y}|\boldsymbol{x}$ according to (5.22) for the Page et al. model and (5.24) for the asymmetric Markov model, and compute MLE estimates for each. The histogram of these estimates along with their theoretical densities overlayed are given in Figures 6.5 and 6.6 respectively. A separate gradient and Hessian analysis showed no evidence of non-local minima.



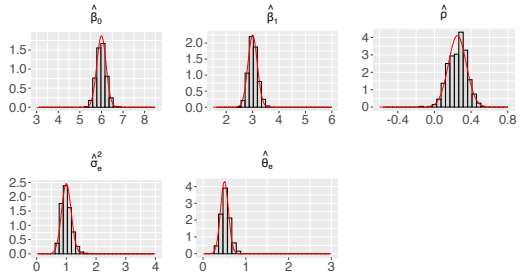
(a) 50 locations



(b) 100 locations

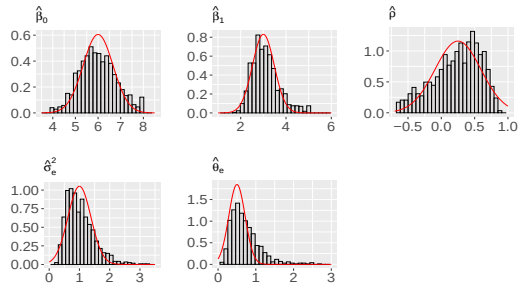


(c) 200 locations

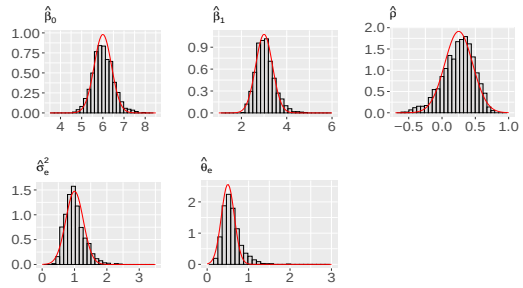


(d) 400 locations

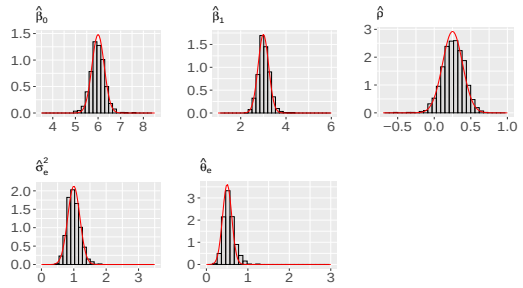
Figure 6.5: Page et al. model MLE estimates along with the theoretical densities (red). Histograms are based on 1000 simulations of $\boldsymbol{y}|\boldsymbol{x}$ from the model (5.22).



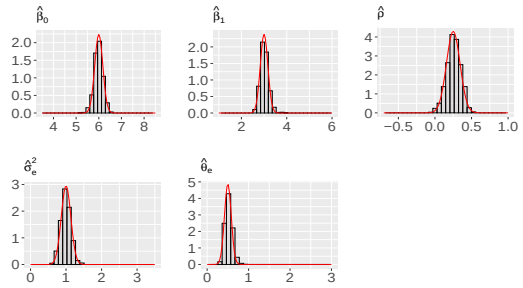
(a) 50 locations



(b) 100 locations



(c) 200 locations



(d) 400 locations

Figure 6.6: Asymmetric Markov model MLE estimates along with the theoretical densities (red). Histograms are based on 1000 simulations of $\boldsymbol{y}|\boldsymbol{x}$ from the model (5.24).

The estimates for ρ are not very well resolved for $n = 50$ in either model, but there is better resolution for larger sample sizes. The MC estimates for both models are approximated well with the normal distribution predicted by Mardia and Marshall (1984).

6.2 Real data example: Housing prices in Boston

We explore the Boston housing dataset first analyzed by Harrison and Rubinfeld (1978). The dataset contains observations related to housing prices in 506 Boston census tracts as determined in the 1970 long-form census (Figure 6.7).

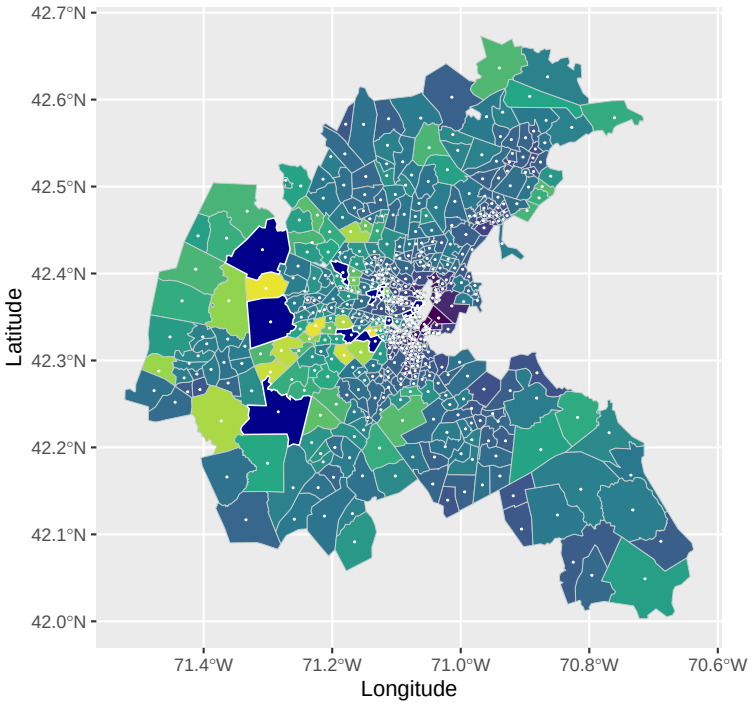


Figure 6.7: Centroids of 506 census tracts in Boston (1970). In dark blue are the locations of the 16 tracts with a censored median house value of $50,000.

The original dataset did not contain any information on spatial coordinates. We use an updated version of the dataset found in the boston dataset of the spData

package in R (Bivand et al. (2022)). This update contains latitude and longitude coordinates along with tract point coordinates projected to UTM zone 19. In our analysis, the median housing value, labelled CMEDV, serves as our response variable $y(\boldsymbol{s})$. We omit the 16 data points median with a censored value of \$50,000 (see Figure 6.7). To highlight the potential for confounding in a nonlinear regression model, we select as our covariate $x(\boldsymbol{s})$, the variable LSTAT. Harrison and Rubinfeld (1978) described this variable as the proportion of population that is perceived to be of lower socioeconomic status, that is, adults without some high school education or classified as laborers.

We first perform an analysis on the covariate LSTAT. After a log-transformation of the LSTAT variable, a histogram of the data was created (Figure 6.8).

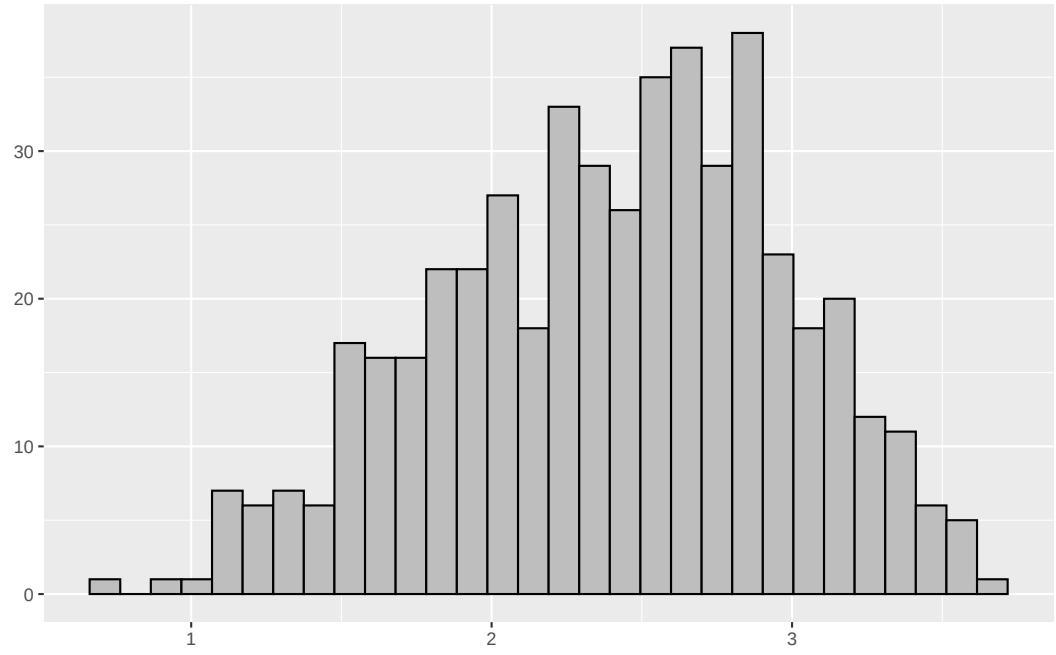


Figure 6.8: Histogram of the log-transformed LSTAT variable from the Boston dataset.

Although a bit skewed, the histogram offers some support to using a Gaussian random field model. Since we are dealing with one realization of a correlated random field, we should note that this exploratory tool is restricted in its use as a diagnostic for normality. However, in the case of a stationary random field (or strictly stationary in the Gaussian case), it may be justified since the distribution at each location is the same.

Next, an empirical variogram was computed using the variogram function from the gstat package (Gräler et al. (2016)) in R (Figure 6.9).

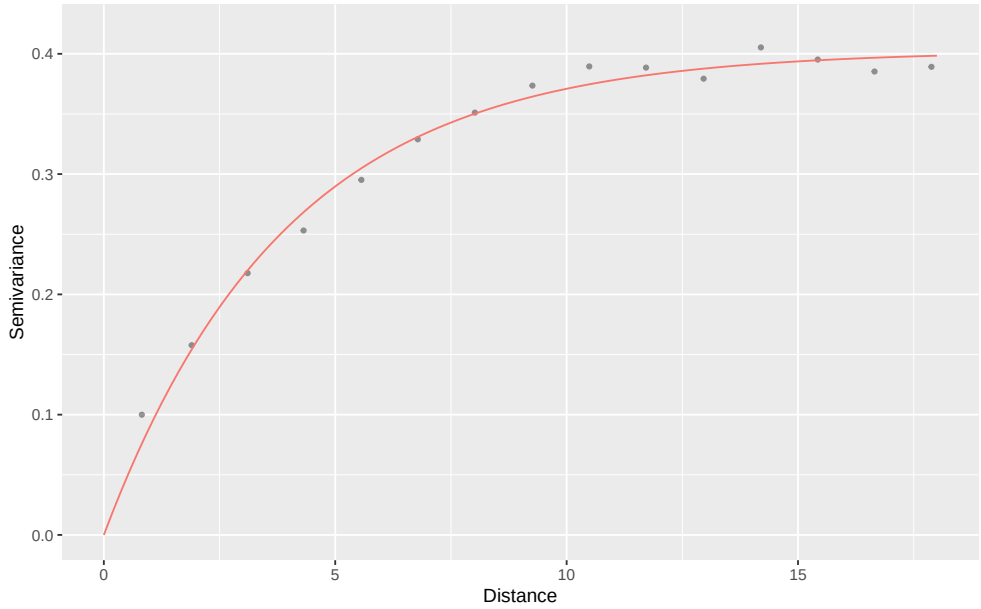


Figure 6.9: Empirical variogram of the log-transformed LSTAT variable from the Boston housing dataset. The empirical variogram was calculated using the Matheron variogram (see equation (3.15)), without any residuals. In red is the fitted exponential variogram curve.

We fitted an exponential variogram curve using the fit.variogram function from the gstat package R. A visual inspection determines that the model fits well and thus, we use the exponential covariance function $C_x(\boldsymbol{h}; \sigma_x^2, \theta_x) = \sigma_x^2 e^{-\theta_x ||\boldsymbol{h}||}$ to model the

dependence structure of $x(\boldsymbol{s})$. Since the covariate is not centered at 0, we also estimate its unknown mean μ_x. The maximum likelihood estimates along with their estimated standard errors (using the Fisher information) is given in Table 6.7.

	$\hat{\sigma}_x^2$	$\hat{\theta}_x$	$\hat{\mu}_x$
Estimate (S.E.)	0.272 (0.028)	0.761 (0.097)	2.027 (0.059)

Table 6.7: Covariate parameter MLE estimates along with their estimated standard errors

Next, we wish to determine the nonlinear relationship between LSTAT and CMEDV. A preliminary scatterplot in Figure 6.10 shows a nonlinear relationship between CMEDV and LSTAT.

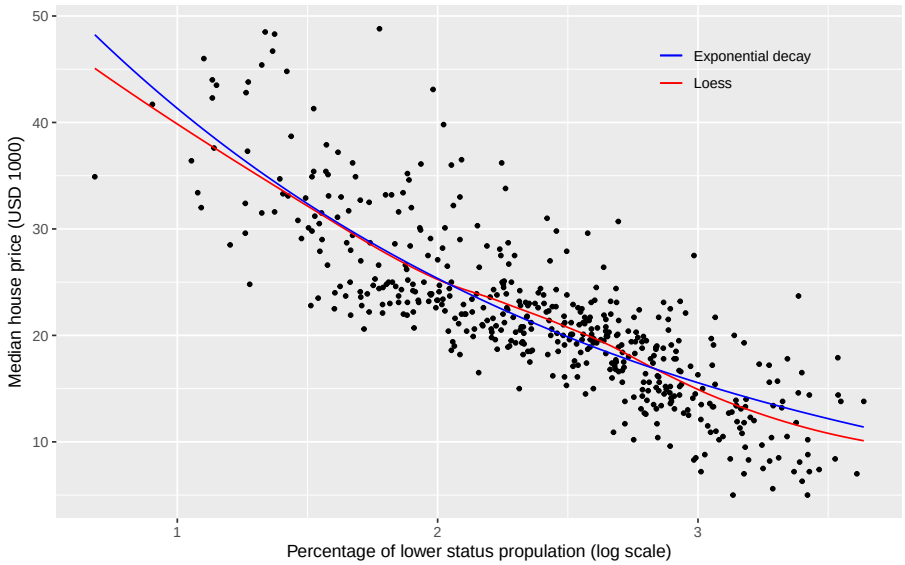


Figure 6.10: Scatterplot of CMEDV against LSTAT from the Boston dataset, with the 16 censored datapoints omitted. In red is a fitted lowess curve and in blue is a fitted exponential decay curve using OLS residuals from model (6.1).

We determined that an exponential decay curve approximated the loess curve well,

and better than a straight line. Thus, the nonlinear model that we consider is,

$$y(\boldsymbol{s}) = \beta_1 e^{-\beta_2 x(\boldsymbol{s})} + e(\boldsymbol{s}) \tag{6.1}$$

Let $\boldsymbol{\theta}$ denote the unknown parameters in the model (6.1) other than $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$. Using the observed data $(\boldsymbol{x}^T, \boldsymbol{y}^T)^T$, we compare estimates of $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ under three different assumptions:

1. Independent, identically distributed errors
2. Spatially correlated errors without confounding
3. Spatially correlated errors with confounding using the confounding models for $(x(\boldsymbol{s}), e(\boldsymbol{s}))^T$ described in Section 5.2.

In the first case, $e(\boldsymbol{s})$ is just Gaussian measurement error, that is, $e(\boldsymbol{s}) \overset{i.i.d}{\sim} N(0, \sigma^2)$ for each $\boldsymbol{s}$. Then, the conditional distribution of $\boldsymbol{y}|\boldsymbol{x}$ is $N\left(\beta_1 e^{-\beta_2 \boldsymbol{x}}, \sigma^2 \boldsymbol{I}\right)$. From this, the maximum likelihood estimates are computed along with their estimated standard errors (using the Fisher information) and given in Table 6.8. The resulting fitted trend is displayed in Figure 6.10.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}^2$
Estimate (S.E.)	67.356 (0.111)	0.489 (0.001)	19.880 (0.898)

Table 6.8: OLS parameter MLE estimates along with their estimated standard errors

For spatially correlated errors without confounding, we must first model the covariance structure of the error random field. Using the fitted OLS estimates $(\hat{\beta}_1, \hat{\beta}_2)^T$, we form the residuals $\hat{\boldsymbol{e}} = \boldsymbol{y} - \hat{\beta}_1 e^{-\hat{\beta}_2 \boldsymbol{x}}$ and perform a variogram analysis.

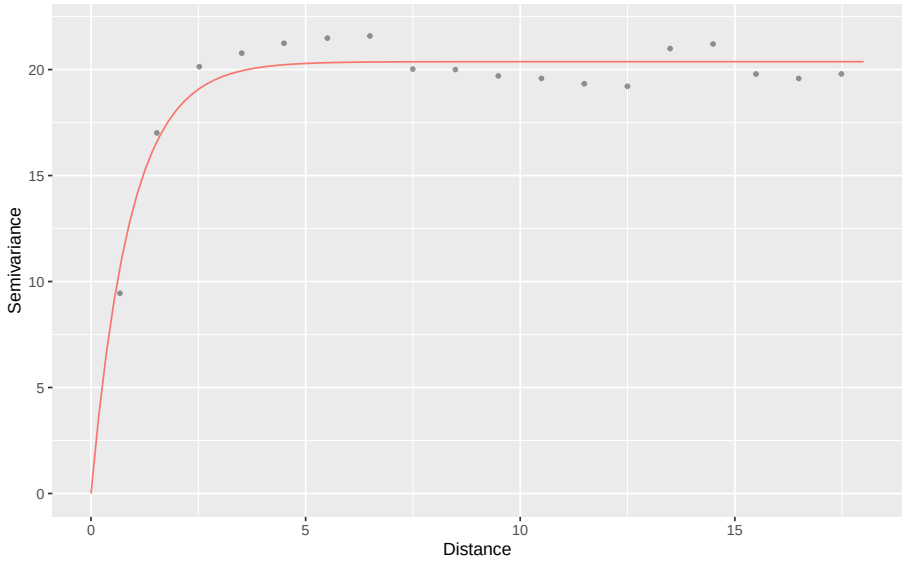


Figure 6.11: Empirical variogram of OLS residuals from the nonlinear regression model (6.1). The fitted exponential variogram is overlayed in red.

An exponential variogram curve was fitted using the fit.variogram function (Figure 6.11). Thus, we take the covariance function of the error random field to be exponential, $C_e(\boldsymbol{h}; \theta_e, \sigma_e^2) = \sigma_e^2 e^{-\theta_e ||\boldsymbol{h}||}$. So the unknown parameters to be estimated without confounding here are $(\beta_1, \beta_2, \sigma_e^2, \theta_e)^T$. The conditional distribution without confounding is $\boldsymbol{y}|\boldsymbol{x} \sim N\left(\beta_1 e^{-\beta_x \boldsymbol{x}}, \sigma_e^2 \boldsymbol{\Sigma}_e(\theta_e)\right)$ where $\{\boldsymbol{\Sigma}_e(\theta_e)\}_{ij} = e^{-\theta_e ||\boldsymbol{s}_i - \boldsymbol{s}_j||}, i, j = 1, ..., 490$. The maximum likelihood estimates along with their estimated standard errors are presented in Table 6.9.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}_e^2$	$\hat{\theta}_e$
Estimate (S.E.)	60.107 (2.207)	0.438 (0.020)	21.171 (1.932)	1.044 (0.128)

Table 6.9: GLS parameter MLE estimates along with their estimated standard errors

Finally, we estimate the parameters $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ taking into account the various confounding models described in Chapter 5. The parameter $\boldsymbol{\theta} = (\sigma_e^2, \theta_e, \rho)^T$ now contains the extra unknown confounding parameter. Recall that the separable model

does not contain an estimate for θ_e since by assumption, the covariate and error share a common correlation function. In general, this assumption is quite restrictive and would not be practical. However, the variogram analyses for LSTAT and the residual $\hat{e}(\boldsymbol{s})$ both show that an exponential correlation for each is a good fit. Moreover, if we compare the MLE estimates $\hat{\theta}_x = 0.761$ from Table 6.7 and $\hat{\theta}_e = 1.044$ from Table 6.9, it appears that the estimated scale parameters are of roughly similar magnitude. Thus, the assumption that they share a common correlation function is at least plausible here and thus, we include the separable model in our analysis.

	Estimates (S.E.)				
Model	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\sigma}_e^2$	$\hat{\theta}_e$	$\hat{\rho}$
1	71.117 (5.789)	0.526 (0.043)	26.931 (2.030)	N/A	0.230 (0.093)
2	70.712 (5.819)	0.522 (0.043)	22.550 (2.273)	1.013 (0.124)	0.232 (0.103)
3	71.273 (5.148)	0.524 (0.038)	21.730 (2.148)	1.076 (0.123)	0.207 (0.076)
4	70.281 (5.813)	0.519 (0.044)	22.647 (2.279)	0.985 (0.115)	0.223 (0.103)
5	72.065 (5.541)	0.530 (0.040)	21.955 (2.195)	1.067 (0.126)	0.242 (0.089)
6	71.275 (5.252)	0.524 (0.038)	21.698 (2.150)	1.060 (0.124)	0.209 (0.080)

Table 6.10: MLE estimates along with their estimated standard errors for the separable model (1), LMC (2), bivariate Matérn model (3), Markov model (4), Page et al. model (5) and asymmetric Markov model (6).

The MLE estimates and estimated standard errors in are given in Table 6.10 for each confounding model. A comparison with Table 6.9 shows that the presence of the confounding parameter significantly changes the MLE estimates for the trend. Page et al. (2017) concluded that in the linear regression model under known covariance parameters, confounding could lead to bias in GLS estimators. This data analysis suggests that something similar could happen with a nonlinear regression model.

For testing $\rho = 0$, we may construct confidence intervals using the above esti-

mates along with their standard errors to a desired confidence level. Alternatively, we may compute Wilks' likelihood ratio statistics along with their p-values. Since the model without confounding is nested within the confounding models, classical theory (van der Vaart (1998), Chapter 16) suggests that the statistic,

$$W = -2 \ln \left(\frac{\sup\limits_{(\boldsymbol{\beta},\boldsymbol{\theta}):\rho=0} L(\boldsymbol{\beta},\boldsymbol{\theta})}{\sup\limits_{(\boldsymbol{\beta},\boldsymbol{\theta})} L(\boldsymbol{\beta},\boldsymbol{\theta})} \right)$$

where $L(\boldsymbol{\beta},\boldsymbol{\theta})$ is the likelihood function of the model, is asymptotically distributed as a chi-squared random variable with one degree of freedom. We present 95% confidence intervals for ρ and Wilks' statistics in Table 6.11. All models appear to give significant evidence of the presence of confounding in this nonlinear regression model.

Model	95% C.I. for ρ	Wilks' statistic	p-value
Separable	(0.048, 0.412)	4.935	0.026
LMC	(0.030, 0.433)	4.054	0.044
Bivariate Matérn	(0.059, 0.356)	6.229	0.013
Markov	(0.020, 0.426)	3.829	0.050
Page et al.	(0.068, 0.417)	6.038	0.014
Asymmetric Markov	(0.053, 0.366)	6.175	0.013

Table 6.11: Confidence intervals for ρ and Wilks' LRT statistics for each confounding model

This was a rudimentary data analysis, but the main purpose was to illustrate how confounding can be potentially be present in spatial regression models, especially if there are several omitted predictors. A more thorough analysis would take into account the effect of other possible covariates, of which there are many in this dataset. In fact, the underlying research question of the original paper by Harri-

son and Rubinfeld (1978) was to determine if pollution affected the housing prices. An analysis involving more covariates would involve modelling multivariate random fields of the form $(x_1(\boldsymbol{s}), ..., x_m(\boldsymbol{s}), e(\boldsymbol{s}))^T$ with a more complicated confounding covariance structure. As we have stated previously, there is literature devoted to the development of valid cross covariance functions for multivariate random fields (Apanasovich et al. (2012)), but none in the context of confounding. This could be possible groundwork for future development in multivariate confounding models.

Chapter 7 Linear regression under infill asymptotics

We motivate the results of this chapter with the following example. Let $x(t)$ be a stationary Gaussian process on $[0,1]$, with unknown mean μ and exponential covariance $C(h) = \sigma^2 e^{-\theta h}$. This is known as the Ornstein-Uhlenbeck process or the continuous-time version of the AR(1) model in time series analysis. For the case $\mu = 0$, Ying (1991) and Abt and Welch (1998) investigated estimation of the covariance parameters (θ, σ^2) when sampling becomes increasingly dense in $[0,1]$. Let $\mathbb{P}_1$ and $\mathbb{P}_2$ denote two Gaussian measures induced by $\{x(t), t \in [0,1]\}$ corresponding to $(\sigma_1^2, \theta_1)^T$ and $(\sigma_2^2, \theta_2)^T$ respectively. It was known to these authors, citing Ibragimov and Rozanov (1978), that $\mathbb{P}_1$ and $\mathbb{P}_2$ are equivalent if and only if $\sigma_1^2\theta_1 = \sigma_2^2\theta_2$. Ying (1991) was then able to prove that $(\sigma^2, \theta)^T$ cannot be consistently estimated individually, but the microergodic parameter $\sigma^2\theta$ can be. The following results on consistency and asymptotic normality of the MLE of $\sigma^2\theta$ can be found in Theorems 1 and 2 of Ying (1991).

Theorem 7.0.1. *Let $\{t_n\}_{n=1}^{\infty}$ be a sequence in $[0,1]$ whose closure equals $[0,1]$. Let $(x(t_1), \cdots, x(t_n))^T$ be a realization of the zero mean OU process based on the first n*

observations. Let $L_n(\sigma^2, \theta)$ *be the likelihood function based on the observations and let* $(\hat{\sigma}_n^2, \hat{\theta}_n)^T = \arg\max_{\sigma^2,\theta} L_n(\sigma^2, \theta)$ *denote the MLE of* $(\sigma^2, \theta)^T$*. Then as* $n \to \infty$*,*

$$\hat{\sigma}_n^2 \hat{\theta}_n \xrightarrow{a.s.} \sigma_0^2 \theta_0$$

$$\sqrt{n}(\hat{\sigma}_n^2 \hat{\theta}_n - \sigma_0^2 \theta_0) \xrightarrow{D} N(0, 2(\sigma_0^2 \theta_0)^2)$$

under the true probability measure, where $(\sigma_0^2, \theta_0)^T$ *are the true parameters.*

Through a Fisher information analysis, Abt and Welch (1998) (section 5, p. 132) conclude a similar result. They first show that the inverse Fisher information for $(\sigma^2, \theta)^T$ does not decay to 0 as the number of observations increase. However, they show that the inverse Fisher information $\mathcal{I}^{-1}_{\sigma^2\theta}$ of the microergodic parameter $\sigma^2\theta$ satisfies $\lim_{n\to\infty} n\mathcal{I}^{-1}_{\sigma^2\theta} = 2(\sigma^2\theta)^2$, matching the asymptotic variance given by Ying.

For estimation of the mean μ, Morris and Ebey (1984) sampled n equally spaced locations $t_i = \frac{i-1}{n-1}, i = 1, ..., n$ in $[0, 1]$. For the corresponding observations $(x(t_1), ..., x(t_n))^T$, they considered as an estimator for μ the sample mean $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x(t_i)$. For any $(\sigma^2, \theta)^T$, they explicitly calculated the limiting variance of the sample mean to be,

$$\lim_{n\to\infty} \text{Var}(\bar{x}_n) = \frac{2\sigma^2}{\theta^2}\left[e^{-\theta} + \theta - 1\right] \tag{7.1}$$

Thus, the asymptotic variance is bounded away from 0. In fact, as we will show in Section 7.3.2, the sample mean converges almost surely to the integral $\int_0^1 x(t)dt$ since the empirical measure of $\{t_1, \cdots, t_n\}$ converges to the Lebesgue measure on

$[0,1]$. This integral is well defined since the sample paths of $x(t)$ are continuous almost surely. Since it is the limit of a Gaussian sequence $\bar{x}_n$, it is also a Gaussian random variable with mean μ and variance $\int_0^1 \int_0^1 \sigma^2 e^{-\theta|s-t|} dsdt$, which has a closed form solution (7.1). These surprising results contrast with their increasing domain counterparts, where for example, consistency and asymptotic normality hold under mild conditions.

In this chapter, we consider the linear regression model,

$$y(\boldsymbol{s}) = \beta_0 + \sum_{k=1}^{p} \beta_k x_k(\boldsymbol{s}) + e(\boldsymbol{s}), \quad \boldsymbol{s} \in D \subset \mathbb{R}^d \tag{7.2}$$

where D is compact. We assume that error $e(\boldsymbol{s})$ is independent of the multivariate random field $(x_1(\boldsymbol{s}), \cdots, x_p(\boldsymbol{s}))^T$ and all fields are mean zero Gaussian. We are interested in the joint estimation of $(\boldsymbol{\beta}^T, \boldsymbol{\theta}^T)^T$ where $\boldsymbol{\theta}$ parametrizes the covariance of $e(\boldsymbol{s})$. For infill sampling, we observe $y(\boldsymbol{s})$ and $\{x_k(\boldsymbol{s})\}_{k=1}^p$ at locations $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$ that become increasingly dense in D. Infill asymptotics literature emphasizes estimation of the covariance parameters and typically assumes that the mean is zero or known. When discussing estimation in an infill setting, Stein (1999), p. 12, states that it is common to assume the mean is highly regular. He further states that smooth covariates have little impact on spatial interpolation, and one does just as well asymptotically by setting the regression coefficients to be 0. In the Matérn case, we clarify these statements with explicit conditions on the smoothness of the covariates relative to the error.

7.1 Equivalence of Gaussian measures with different means

7.1.1 Deterministic mean function

In this section, we give a literature review of some known results on the equivalence and singularity of Gaussian measures with different mean functions. We assume a general form of the regression model,

$$y(\boldsymbol{s}) = m(\boldsymbol{s}) + e(\boldsymbol{s}), \quad \boldsymbol{s} \in D \subset \mathbb{R}^d \tag{7.3}$$

where $e(\boldsymbol{s})$ is a mean zero Gaussian random field with jointly continuous covariance function $C(\boldsymbol{s}, \boldsymbol{t})$ and $m(\boldsymbol{s})$ is a deterministic function representing the mean of $y(\boldsymbol{s})$. In the case of a known covariance function, let $\mathbb{P}_m, \mathbb{P}_0$ be the Gaussian probability measures for $y(\boldsymbol{s})$ corresponding to $m(\boldsymbol{s})$ not identically 0 and $m(\boldsymbol{s}) = 0$ respectively. The solution to the problem of equivalence and mutual singularity of $\mathbb{P}_m$ and $\mathbb{P}_0$ has been known for decades in the probability and time series literature. Cameron and Martin (1944) studied the problem in the special case where $D = [0, 1]$ and $e(\boldsymbol{s})$ is a Brownian motion on D. In Theorem 2 of their paper, they proved that $\mathbb{P}_m \equiv \mathbb{P}_0$ if and only if $m(\boldsymbol{s})$ is absolutely continuous with first derivative being square integrable on $[0, 1]$. As we shall see, it is no coincidence that the smoothness of $m(\boldsymbol{s})$ determines whether or not $\mathbb{P}_m \equiv \mathbb{P}_0$. For broader classes of Gaussian processes, Grenander (1950) derived conditions utilizing the Karhunen-Loève (K-L) representation of $y(\boldsymbol{s})$. The K-L representation of the random field in (7.3) takes

the form of an infinite series (where convergence is defined in L_2),

$$y(\boldsymbol{s}) = m(\boldsymbol{s}) + \sum_{n=1}^{\infty} \xi_n \varphi_n(\boldsymbol{s}) \tag{7.4}$$

where $\{\xi_n\}_{n=1}^{\infty}$ are independent $N(0, \lambda_n)$ and (λ_n, φ_n) satisfy the integral equation,

$$\int_D C(\boldsymbol{s}, \boldsymbol{t}) \varphi_n(\boldsymbol{t}) d\boldsymbol{t} = \lambda_n \varphi_n(\boldsymbol{s}), \quad \varphi_n(\boldsymbol{s}) \in L_2(D) \tag{7.5}$$

The advantage of the representation in (7.4) is that the sigma algebra generated by $\{y(\boldsymbol{s}), \boldsymbol{s} \in D\}$ is equal to the sigma algebra generated by the countable sequence $\{\xi_n\}_{n=1}^{\infty}$ in (7.4). Thus, the measures $\mathbb{P}_m$ and $\mathbb{P}_0$ can be reduced to a countably infinite product of measures induced by $\{\xi_n\}_{n=1}^{\infty}$. A well known theorem of Kakutani (1948) gives necessary and sufficient conditions for two countably infinite product measures to be equivalent or mutually singular. Using the K-L representation of Gaussian random fields and Kakutani's theorem, Grenander proved the following result (see Section 4.4 of Grenander (1950)).

Theorem 7.1.1. *Let (λ_n, φ_n) be the eigenpairs of (7.5) and $m_n = \int_D m(\boldsymbol{s}) \varphi_n(\boldsymbol{s}) d\boldsymbol{s}$. Then the Gaussian measures $\mathbb{P}_0$ and $\mathbb{P}_m$ are either equivalent or mutually singular. They are equivalent if $\sum_{n=1}^{\infty} \frac{m_n^2}{\lambda_n} < \infty$ and mutually singular if $\sum_{n=1}^{\infty} \frac{m_n^2}{\lambda_n} = \infty$.*

Grenander's conditions of Theorem 7.1.1 are in general difficult to verify analytically since the eigenpairs of (7.5) rarely exist in closed form. There are cases such as Brownian motion and the Ornstein-Uhlenbeck process on $\mathbb{R}$, where K-L representations have closed forms, but these are exceptions. Hájek (1958) (and in-

dependently, Feldman (1958)) arrived at this Gaussian dichotomy result in another way, using the entropy distance method described in Section 2.5.2. This method has more statistical relevance because it allows us to formulate equivalence and singularity results in terms of observed data. Let $\{\boldsymbol{s}_1, .., \boldsymbol{s}_n\}$ be a nested sequence of countably dense locations in D and $\boldsymbol{y} = (y(\boldsymbol{s}_1), ..., y(\boldsymbol{s}_n))^T$ be the corresponding observations of $y(\boldsymbol{s})$ at these locations. Then, $\boldsymbol{y}$ is multivariate Gaussian with mean vector $\boldsymbol{m} = (m(\boldsymbol{s}_1), \cdots, m(\boldsymbol{s}_n))^T$ and known covariance matrix $\boldsymbol{\Sigma}$. Denote by p_n, the likelihood ratio of $\mathbb{P}_m$ to $\mathbb{P}_0$ based on the observations $\boldsymbol{y}$. A simple calculation of $\log p_n$ yields,

$$\log p_n = \boldsymbol{m}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y} - \frac{1}{2} \boldsymbol{m}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{m} \tag{7.6}$$

Recall the definition of entropy distance given in (2.10) of Definition 2.5.2. In this case, the entropy distance is $J_n = \mathbb{E}_m[\log p_n] - \mathbb{E}_0[\log p_n] = \boldsymbol{m}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{m}$. By Lemma 2.5.3, Hájek's Gaussian dichotomy result can be stated as follows.

Theorem 7.1.2. *$\mathbb{P}_0$ and $\mathbb{P}_m$ are either equivalent or mutually singular. They are equivalent if $\lim_{n\to\infty} \boldsymbol{m}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{m} < \infty$ and mutually singular if $\lim_{n\to\infty} \boldsymbol{m}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{m} = \infty$.*

There is a relationship between the conditions of Theorems 7.1.1 and 7.1.2. The integral equation in (7.5) of the K-L representation is a continuous analog of the eigenvalue decomposition of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T$. Here $\boldsymbol{P}$ is an orthogonal matrix containing the eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ is a diagonal matrix containing the eigenvalues. Then $\boldsymbol{m}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{m} = \boldsymbol{m}^T \boldsymbol{P}^T \boldsymbol{D}^{-1} \boldsymbol{P} \boldsymbol{m} = \sum_{i=1}^{n} \frac{m_i^2}{\lambda_i}$, where $m_i, i = 1, ..., n$ is the i^{th} entry of $\boldsymbol{P}\boldsymbol{m}$. Letting $n \to \infty$, this has a similar form to

Grenander's condition in Theorem 7.1.1.

In the context of signal detection theory, Parzen (1963) gave an answer to the problem of equivalence and singularity of $\mathbb{P}_0, \mathbb{P}_m$ using the theory of reproducing kernel Hilbert spaces (RKHS). The following definition of RKHS is given in terms of a continuous covariance kernel, but in general applies to any positive definite function. For a treatise on the theory of reproducing kernels, we refer to Aronszajn (1950).

Definition 7.1.3. *Let $C(\boldsymbol{s}, \boldsymbol{t})$ be a jointly continuous covariance function on $D \times D$ where $D \subset \mathbb{R}^d$ is compact. For any fixed $\boldsymbol{t}_0 \in D$, consider the mapping $\boldsymbol{s} \mapsto C(\boldsymbol{s}, \boldsymbol{t}_0)$. Then there exists a unique Hilbert space of functions on D, denoted as $R(C)$, equipped with an inner product $\langle \cdot, \cdot \rangle$ such that,*

1. *$C(\boldsymbol{s}, \boldsymbol{t}_0) \in R(C)$ for any $\boldsymbol{t}_0 \in D$*

2. *$\langle f, C(\cdot, \boldsymbol{t}_0) \rangle = f(\boldsymbol{t}_0)$ for any $f(\boldsymbol{s}) \in R(C)$*

$R(C)$ is called the reproducing kernel Hilbert space (RKHS) with kernel $C(\boldsymbol{s}, \boldsymbol{t})$.

This cursory definition does not give much insight into the structure of the space $R(C)$. For example, it does not give an explicit form of inner product or properties of functions that lie in $R(C)$. The uniqueness property of RKHS implies that as long we can find a Hilbert space with inner product satisfying the properties of Definition 7.1.3, then it will be the RKHS associated with $C(\boldsymbol{s}, \boldsymbol{t})$. It turns out that RKHS are related to the K-L representation (7.4) of a random field. By a theorem of Mercer (1909), the eigenfunctions $\{\varphi_n\}_{n=1}^{\infty}$ in (7.5) form an orthonormal basis of

$L_2(D)$, and the covariance function $C(\boldsymbol{s}, \boldsymbol{t})$ has representation,

$$C(\boldsymbol{s}, \boldsymbol{t}) = \sum_{n=1}^{\infty} \lambda_n \varphi_n(\boldsymbol{s}) \varphi_n(\boldsymbol{t}) \tag{7.7}$$

which converges uniformly on $D \times D$ by continuity. With this representation, we have the following characterization of RKHS (p. 57 of Parzen (1962)), which shows that RKHS are a subspace of $L_2(D)$.

Theorem 7.1.4. *Let $\{\varphi_n\}_{n=1}^{\infty}$ be the orthonormal basis of $L_2(D)$ given in (7.5). The reproducing kernel Hilbert space $R(C)$ is a subspace of $L_2(D)$ consisting of functions with the representation,*

$$f(\boldsymbol{s}) = \sum_{n=1}^{\infty} f_n \varphi_n(\boldsymbol{s}), \quad f_n = \int_D f(\boldsymbol{s}) \varphi_n(\boldsymbol{s}) d\boldsymbol{s} \tag{7.8}$$

such that $||f||^2 := \sum_{n=1}^{\infty} \frac{f_n^2}{\lambda_n} < \infty$. For $f(\boldsymbol{s}), g(\boldsymbol{s}) \in R(C)$, the corresponding inner product is $\langle f, g \rangle := \sum_{n=1}^{\infty} \frac{f_n g_n}{\lambda_n}$.

For a fixed $\boldsymbol{t}_0 \in D$ and Mercer's representation of $C(\boldsymbol{s}, \boldsymbol{t}_0)$ in (7.7), one may verify that,

1. $||C(\cdot, \boldsymbol{t}_0)||^2 = \sum_{n=1}^{\infty} \frac{(\lambda_n \varphi_n(\boldsymbol{t}_0))^2}{\lambda_n} = C(\boldsymbol{t}_0, \boldsymbol{t}_0) < \infty$
2. $\langle f, C(\cdot, \boldsymbol{t}_0) \rangle = \sum_{n=1}^{\infty} \frac{f_n \lambda_n \varphi_n(\boldsymbol{t}_0)}{\lambda_n} = f(\boldsymbol{t}_0)$ for any $f(\boldsymbol{s}) \in R(C)$.

Thus, $C(\boldsymbol{s}, \boldsymbol{t})$ satisfies the properties of being a reproducing kernel given in Definition 7.1.3. Comparing Theorem 7.1.4 with Grenander's conditions in Theorem 7.1.1, Parzen (1963) formulated his equivalence and singularity result as follows.

Theorem 7.1.5. $\mathbb{P}_m \equiv \mathbb{P}_0$ *if and only if* $m(\boldsymbol{s}) \in R(C)$.

Parzen (1963) states that functions belonging to $R(C)$ must be at least as smooth as $C(\boldsymbol{s}, \boldsymbol{t}_0)$ for any $\boldsymbol{t}_0 \in D$. Thus, in order for the measures $\mathbb{P}_m, \mathbb{P}_0$ to equivalent, the mean function $m(\boldsymbol{s})$ must be at least as smooth as the covariance function $C(\boldsymbol{s}, \boldsymbol{t}_0)$. To see an explicit connection between smoothness of functions and RKHS, we state a few more known results in the spectral domain. When $e(\boldsymbol{s})$ is stationary, the RKHS can be formulated in terms of its spectral density (Wendland (2004), Theorem 10.12).

Theorem 7.1.6. *Let* $C(\boldsymbol{s}, \boldsymbol{t})$ *be a stationary covariance function on* $\mathbb{R}^d$ *with corresponding spectral density* $f(\boldsymbol{\omega})$. *That is,* $C(\boldsymbol{s}, \boldsymbol{t}) = C_0(\boldsymbol{s} - \boldsymbol{t})$ *for some function* C_0. *Then* $R(C)$ *is the subspace of* $L_2(\mathbb{R}^d)$ *consisting of functions* $g(\boldsymbol{s})$ *whose Fourier transform* $\hat{g}(\boldsymbol{\omega}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\boldsymbol{\omega}^T \boldsymbol{s}} g(\boldsymbol{s}) d\boldsymbol{s}$ *satisfies,*

$$||g||^2 := \int_{\mathbb{R}^d} \frac{|\hat{g}(\boldsymbol{\omega})|^2}{f(\boldsymbol{\omega})} d\boldsymbol{\omega} < \infty \tag{7.9}$$

For $g, h \in R(C)$, *the corresponding inner product is* $\langle g, h \rangle := \int_{\mathbb{R}^d} \frac{\hat{g}(\boldsymbol{\omega})\overline{\hat{h}(\boldsymbol{\omega})}}{f(\boldsymbol{\omega})} d\boldsymbol{\omega}$.

Using properties of the Fourier transform, one may verify that for any $\boldsymbol{t}_0 \in D$,

1. $||C(\cdot, \boldsymbol{t}_0)||^2 = \int_{\mathbb{R}^d} f(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty$
2. $\langle g, C(\cdot, \boldsymbol{t}_0) \rangle = \int_{\mathbb{R}^d} \hat{g}(\boldsymbol{\omega}) e^{i\boldsymbol{\omega}^T \boldsymbol{t}_0} d\boldsymbol{s} = g(\boldsymbol{t}_0)$ for any $g(\boldsymbol{s}) \in R(C)$.

Thus, $C(\boldsymbol{s}, \boldsymbol{t})$ satisfies the properties of a reproducing kernel in Definition 7.1.3. This formulation of the RKHS shows that the Fourier transform of any function

$f(\boldsymbol{s}) \in R(C)$ should decay faster than the Fourier transform of $C(\boldsymbol{s}, \boldsymbol{t}_0)$ for any $\boldsymbol{t}_0 \in D$. Since the tail behavior of a Fourier transform of a function $f(\boldsymbol{s})$ determines the smoothness of $f(\boldsymbol{s})$, this confirms that $f(\boldsymbol{s})$ should be smoother than $C(\boldsymbol{s}, \boldsymbol{t}_0)$ for it to belong to $R(C)$. As a corollary to Theorems 7.1.5 and 7.1.6, the following result gives a spectral condition for the equivalence and singularity of $\mathbb{P}_m, \mathbb{P}_0$ and can be found in Theorem 2 (p. 138) of Yadrenko (1983).

Theorem 7.1.7. *Suppose that $e(\boldsymbol{s})$ in (7.3) is stationary with spectral density $f(\boldsymbol{\omega})$. Then $\mathbb{P}_0 \equiv \mathbb{P}_m$ if and only if $m(\boldsymbol{s})$ can be extended to a square integrable function on $\mathbb{R}^d$ and its Fourier transform $\hat{m}(\boldsymbol{\omega})$ satisfies,*

$$\int_{\mathbb{R}^d} \frac{|\hat{m}(\boldsymbol{\omega})|^2}{f(\boldsymbol{\omega})} d\boldsymbol{\omega} < \infty \tag{7.10}$$

Finally, we state another result relating RKHS to Sobolev spaces (see Definitions 2.5.9 and 2.5.10), which makes the role of smoothness become fully apparent. The following result (Wendland (2004), Corollary 10.13) shows that if the spectral density has a precisely polynomial decay, then the RKHS and Sobolev spaces coincide.

Theorem 7.1.8. *Let $\ell > \dfrac{d}{2}$ and suppose the spectral density $f(\boldsymbol{\omega})$ of $e(\boldsymbol{s}), \boldsymbol{s} \in D$ satisfies,*

$$k(1 + ||\boldsymbol{\omega}||^2)^{-\ell} \leq f(\boldsymbol{\omega}) \leq K(1 + ||\boldsymbol{\omega}||^2)^{-\ell}$$

for constants k, K. Then the RKHS $R(C)$ coincides with the Sobolev space $W_2^{\ell}(D)$ and the RKHS norm and Sobolev space norms are equivalent.

This can be seen from (7.9), where the norm of any $g(\boldsymbol{s}) \in R(C)$ is finite iff,

$$||g||^2 := \int_{\mathbb{R}^d} |\hat{g}(\boldsymbol{\omega})|^2 (1 + ||\boldsymbol{\omega}||^2)^{\ell} d\boldsymbol{\omega} < \infty, \quad \ell > \frac{d}{2}$$

and this norm is known to be equivalent to the Sobolev norms given in Definitions 2.5.9 and 2.5.10 (Wendland (2004), p. 133). Once again, we obtain the following equivalence and singularity result as a corollary to Theorems 7.1.5 and 7.1.8. It can be found in Theorems 3 and 4 (p. 140) of Yadrenko (1983), for integer order and fractional order $W_2^{\ell}(D)$ respectively.

Theorem 7.1.9. *Let $\ell > \frac{d}{2}$ and suppose the spectral density $f(\boldsymbol{\omega})$ of $e(\boldsymbol{s})$ satisfies,*

$$k(1 + ||\boldsymbol{\omega}||^2)^{-\ell} \leq f(\boldsymbol{\omega}) \leq K(1 + ||\boldsymbol{\omega}||^2)^{-\ell}$$

for constants k, K. Then $\mathbb{P}_0 \equiv \mathbb{P}_m$ if and only if $m(\boldsymbol{s}) \in W_2^{\ell}(D)$.

The Matérn spectral density in (2.6) is easily seen to satisfy the inequalities in Theorem 7.1.9 with $\ell = \nu + \frac{d}{2}$. Thus, if $e(\boldsymbol{s})$ has Matérn covariance, for any mean function $m(\boldsymbol{s}) \in W_2^{\ell}(D)$, the measures $\mathbb{P}_m$ and $\mathbb{P}_0$ are equivalent.

7.1.2 Stochastic mean function

Now, we give a discussion of the case where $m(\boldsymbol{s})$ is also a random field independent of $e(\boldsymbol{s})$. Assume that $(m(\boldsymbol{s}), y(\boldsymbol{s}))^T$ is a joint random field (not necessarily Gaussian) taking values in a complete separable metric space E (also known as a Polish space). An example of such a space is the set of continuous real valued func-

tions on $D \subset \mathbb{R}^d$ endowed with the supremum norm. Then $(m(\boldsymbol{s}), y(\boldsymbol{s}))^T$ induces a joint probability measure $\mathbb{P}$ on the measurable space $(E, \mathcal{B}(E))$ where $\mathcal{B}(E)$ is the Borel sigma-algebra. The problem is then to determine conditions under which two alternative joint measures $\mathbb{P}_i = \mathbb{P}_i^{M,Y}$ for $i = 1, 2$ are either equivalent or mutually singular. We assume that both measures $\mathbb{P}_i, i = 1, 2$ have the same marginal probability measure $\mathbb{Q} = \mathbb{Q}^M$ for $m(\boldsymbol{s})$. Here, we show that this problem can be reduced to determining whether the conditional probability measures of $y(\boldsymbol{s})$ given $m(\boldsymbol{s})$ are equivalent or mutually singular. By the assumptions on the space E, the regular conditional probabilities $\mathbb{P}_i^{Y|M}(\cdot|m) = \mathbb{P}_i(\cdot|m), i = 1, 2$ exist as measures that are measurable functions of $m(\boldsymbol{s})$ characterized by (Çinlar (2011), Theorem 2.10),

$$\int h(m)\mathbb{P}_i(A|m)d\mathbb{Q}(m) = \int h(m)\mathbb{1}_{y\in A}d\mathbb{P}_i(m, y) \quad A \in \sigma(y(\boldsymbol{s})) \tag{7.11}$$

for all bounded measurable $\sigma(m(\boldsymbol{s}))$ functions h. Then for all $B \in \sigma(m(\boldsymbol{s}), y(\boldsymbol{s}))$, by Fubini and the disintegration theorem (Çinlar (2011), Theorem 2.18), there exists a measurable family of sets $B_m \in \sigma(y(\boldsymbol{s}))$ such that for all bounded measurable $\sigma(m(\boldsymbol{s}))$ functions h,

$$\int h(m)\mathbb{P}_i(B_m|m)d\mathbb{Q}(m) = \int h(m)\mathbb{1}_{(m,y)\in B}d\mathbb{P}_i(m, y) \tag{7.12}$$

Since the measures $\mathbb{P}_i^{Y|M}, i = 1, 2$ are regular conditional probabilities, the Radon-Nikodym derivative,

$$f(m,y) \equiv \frac{d\mathbb{P}_1^{Y|M}}{d(\mathbb{P}_1^{Y|M} + \mathbb{P}_2^{Y|M})}(m,y) \tag{7.13}$$

exists as a jointly measurable function of (m, y) characterized for bounded measurable functions $h(m), g(y)$ by the equation (Çinlar (2011), Theorem 4.44),

$$\begin{aligned} &\int \left[\int f(m,y)g(y)d(\mathbb{P}_1^{Y|M} + \mathbb{P}_2^{Y|M})(y|m) \right] h(m)d\mathbb{Q}(m) \\ &= \int \left[\int g(y)d\mathbb{P}_1^{Y|M}(y|m) \right] h(m)d\mathbb{Q}(m) \end{aligned} \tag{7.14}$$

On the other hand, equation (7.12) implies that the right-hand side of (7.14) is alternately expressed as,

$$\int \left[\int g(y)d\mathbb{P}_1^{Y|M}(y|m) \right] h(m)d\mathbb{Q}(m) = \int g(y)h(m)d\mathbb{P}_1(m,y)$$

while the left-hand side of (7.14) is equal to,

$$\begin{aligned} &\int \left[\int f(m,y)g(y)d(\mathbb{P}_1^{Y|M} + \mathbb{P}_2^{Y|M})(y|m) \right] h(m)d\mathbb{Q}(m) \\ &= \int f(m,y)g(y)h(m)d(\mathbb{P}_1 + \mathbb{P}_2)(m,y) \end{aligned} \tag{7.15}$$

Therefore, equations (7.14) and these last two equations imply that for all bounded measurable g, h,

$$\int f(m,y)g(y)h(m)d\mathbb{P}_1(m,y) = \int f(m,y)g(y)h(m)d(\mathbb{P}_1 + \mathbb{P}_2)(m,y) \tag{7.16}$$

But equation (7.16) uniquely characterizes the Radon-Nikodym derivative of $\mathbb{P}_1$ with respect to $\mathbb{P}_1 + \mathbb{P}_2$,

$$f(m, y) = \frac{d\mathbb{P}_1}{d(\mathbb{P}_1 + \mathbb{P}_2)}(m, y) \tag{7.17}$$

Comparing (7.13) and (7.17) the overall conclusion is that,

$$\frac{d\mathbb{P}_1^{Y|M}}{d(\mathbb{P}_1^{Y|M} + \mathbb{P}_2^{Y|M})}(m, y) \equiv \frac{d\mathbb{P}_1}{d(\mathbb{P}_1 + \mathbb{P}_2)}(m, y) \quad \text{almost surely } \mathbb{P}_1 + \mathbb{P}_2 \tag{7.18}$$

with both sides well defined as jointly measurable $\sigma(m(\boldsymbol{s}), y(\boldsymbol{s}))$ functions. Every assertion above holds without the joint Gaussian assumption. In the Gaussian case, recall that $\mathbb{P}_i, i = 1, 2$ are either equivalent or mutually singular. Then the above discussion leads to the following result.

Proposition 7.1.1. *The joint Gaussian measures $\mathbb{P}_1$ and $\mathbb{P}_2$ are either equivalent or mutually singular.*

(1) $\mathbb{P}_1$ and $\mathbb{P}_2$ are equivalent if and only if the conditional measures $\mathbb{P}_1^{Y|M}$ and $\mathbb{P}_2^{Y|M}$ are equivalent almost surely $\mathbb{Q}$.

(2) $\mathbb{P}_1$ and $\mathbb{P}_2$ are mutually singular if and only if the conditional measures $\mathbb{P}_1^{Y|M}$ and $\mathbb{P}_2^{Y|M}$ are mutually singular almost surely $\mathbb{Q}$.

Proof. For statement (1), $\mathbb{P}_1$ and $\mathbb{P}_2$ are equivalent if and only if the right hand side of (7.18) is strictly between 0 and 1 almost surely with respect to both $\mathbb{P}_1$ and $\mathbb{P}_2$. Define the set $A = \{(m, y) : 0 < f(m, y) < 1\}$. Then by the disintegration

theorem (7.12), $\mathbb{P}_1^{Y|M}(A|m) = \mathbb{P}_2^{Y|M}(A|m) = 1$ almost surely $\mathbb{Q}$. Therefore, the term on the left hand side of (7.18) is strictly between 0 and 1 almost surely $\mathbb{Q}$, which is equivalent to the statement $\mathbb{P}_1^{Y|M} \equiv \mathbb{P}_2^{Y|M}$ almost surely $\mathbb{Q}$.

For statement (2), $\mathbb{P}_1$ and $\mathbb{P}_2$ are mutually singular if and only if the right hand side of (7.18) is equal to 0 almost surely $\mathbb{P}_2$ and 1 almost surely $\mathbb{P}_1$. Define the set $A = \{(m, y) : f(m, y) = 0\}$. Then $\mathbb{P}_1(A) = 0 = \mathbb{P}_2(A^c)$ and by the disintegration theorem (7.12), $\mathbb{P}_1^{Y|M}(A|m) = 0 = \mathbb{P}_2^{Y|M}(A^c|m)$ almost surely $\mathbb{Q}$, which is equivalent to the statement $\mathbb{P}_1^{Y|M} \perp \mathbb{P}_2^{Y|M}$ almost surely $\mathbb{Q}$. □

Proposition 7.1.1 implies that in order to establish equivalence and singularity results for the joint Gaussian probability measures $\mathbb{P}_i, i = 1, 2$, it is equivalent to consider the conditional measures $\mathbb{P}_i^{Y|M}, i = 1, 2$ instead. Then by the results of Parzen (1963), it is a matter of determining if the sample paths of $m(\boldsymbol{s})$ almost surely belong to the RKHS of the covariance kernel of $e(\boldsymbol{s})$. In fact, by a result of Driscoll (1973), this holds with probability 0 or 1, once again highlighting the dichotomy properties of Gaussian measures.

7.2 Microergodicity of the mean in ordinary kriging models

In this section, we consider a Gaussian random field with constant mean parameter, $y(\boldsymbol{s}) = \mu + e(\boldsymbol{s})$, with a known covariance function for $e(\boldsymbol{s})$. This model is also known as the ordinary kriging model in geostatistics (Cressie (1993)). Let $\mathbb{P}_\mu$ and $\mathbb{P}_0$ respectively denote the Gaussian measures parametrized by $\mu \neq 0$ and $\mu = 0$ respectively. The objective of this section is to establish necessary and sufficient

conditions for μ to be microergodic.

7.2.1 Microergodicity and Fisher information

The dichotomy result by Hájek (1958) in Theorem 7.1.2 can be applied directly. In this case, the mean vector is $\boldsymbol{m} = \mu\mathbf{1}$ and the quadratic form becomes $\boldsymbol{m}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{m} = \mu^2\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$. One may notice that the term $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ is also the Fisher information for μ. Using Theorem 7.1.2, we can connect the concepts of microergodicity and Fisher information.

Proposition 7.2.1. *$\mathbb{P}_\mu$ and $\mathbb{P}_0$ are either equivalent or mutually singular. They are mutually singular if and only if $\lim_{n\to\infty} \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1} = \infty$. In other words, μ is microergodic if and only if the Fisher information diverges.*

By Theorem 2.5.5 of Zhang (2004), we know that if μ is not microergodic, it cannot be consistently estimated. Thus, divergence of the Fisher information is a necessary condition for the existence of a consistent estimator of μ. In general, one might expect microergodicity to also be sufficient for a consistent estimator to exist. In particular, if $\{\mathbb{P}_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ is a family of probability measures and $h(\boldsymbol{\theta})$ is microergodic, then it might be plausible that there exists a sequence of estimators $\hat{\boldsymbol{\theta}}$ such that $h(\hat{\boldsymbol{\theta}})$ is consistent for $h(\boldsymbol{\theta})$. However, as Stein (1999) shows (Section 6.2, p. 163), this is generally not the case without further restrictions on the parameter space Θ (for example, if Θ is countable). Fortunately, in the Gaussian setting with known covariance, we can prove that the microergodicity of μ is sufficient for a consistent estimator to exist.

Proposition 7.2.2. *Let $\{y(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a Gaussian random field with mean μ and known covariance function. Then, a consistent estimator of μ exists if and only if μ is microergodic.*

Proof. As already stated in the above paragraph, necessity holds by Theorem 2.5.5. For sufficiency, let $\boldsymbol{y} = (y(\boldsymbol{s}_1), \cdots, y(\boldsymbol{s}_n))^T$ be an observed realization of the Gaussian random field. Then $\boldsymbol{y}$ is multivariate Gaussian with mean $\mu\mathbf{1}$ and covariance matrix $\boldsymbol{\Sigma}$. Consider the maximum likelihood estimator of μ,

$$\hat{\mu} = \frac{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}} \tag{7.19}$$

This estimator is unbiased with variance $\frac{1}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}}$. By Proposition 7.2.1, if μ is microergodic, then the Fisher information $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ diverges. Thus, $\hat{\mu} \xrightarrow{L_2} \mu$ and so $\hat{\mu}$ is consistent. □

Combining Propositions 7.2.1 and 7.2.2 connects the concepts of Fisher information, microergodicity and consistency for estimation of the constant mean of a Gaussian random field.

Corollary 7.2.1. *Let $\{y(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a Gaussian random field with mean μ and known covariance function. The following are equivalent.*

1. *The Fisher information $\lim_{n\to\infty} \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ diverges.*
2. *The mean parameter is microergodic, that is, $\mathbb{P}_\mu \perp \mathbb{P}_0 \iff \mu \neq 0$*
3. *A consistent estimator of μ exists, namely the MLE.*

In general, proving that $\lim_{n\to\infty} \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ diverges or converges is difficult to do analytically. In the Appendix, we show that for the Ornstein-Uhlenbeck process on $[0,1]$, the quantity $\lim_{n\to\infty} \mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ has a closed form and remains finite, confirming the results of Morris and Ebey (1984) on the sample mean. In the next section, we show that for any dimension $d \geq 1$, and for a general class of covariance functions, the quantity $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}$ is necessarily finite, even if we do not have an explicit form of $\boldsymbol{\Sigma}^{-1}$.

7.2.2 Non-microergodicity for a special class of spectral densities

When $e(\boldsymbol{s})$ has Matérn covariance, we stated in a remark after Theorem 7.1.9, that $\mathbb{P}_\mu \equiv \mathbb{P}_0$. Here, we consider a more general class of spectral densities that are bounded below,

$$f(\boldsymbol{\omega}) \geq \frac{A}{(1+||\boldsymbol{\omega}||^2)^\tau} \tag{7.20}$$

for some $A > 0, \tau > \dfrac{d}{2}$. We give an explicit example of an extension of $m(\boldsymbol{s}) = \mu$ satisfying Theorem 7.1.7, thus showing that $\mathbb{P}_\mu \equiv \mathbb{P}_0$ for spectral densities satisfying (7.20). Without loss of generality, we can take the constant μ to be identically 1. If $f(\boldsymbol{\omega})$ satisfies (7.20), then the condition in Theorem 7.1.7 holds if,

$$\int_{\mathbb{R}^d} |\hat{m}(\boldsymbol{\omega})|^2(1+||\boldsymbol{\omega}||^2)^\tau d\boldsymbol{\omega} < \infty$$

We take advantage of the isotropy of the spectral density and find an extension of the mean that is also isotropic. Since D is compact, it is contained within some ball $B(\mathbf{0}, R)$ of radius R. Without loss of generality, assume that $R = 1$. Then, for any $N \in \mathbb{N}$, consider the isotropic extension,

$$m(\boldsymbol{s}) = \mathbb{1}_{||\boldsymbol{s}||\leq 1}(\boldsymbol{s}) + g(||\boldsymbol{s}||)\mathbb{1}_{1<||\boldsymbol{s}||<\infty}(\boldsymbol{s}) \tag{7.21}$$

where $g(r) = e^{-(r-1)^{N+1}}$. This extension is easily seen to be square integrable because of its exponential decay. Before we continue, we state the following lemma which gives a simplified form of the Fourier transform of an isotropic function.

Lemma 7.2.2. *Let $f(\boldsymbol{s})$ be an isotropic function, that is, a function that only depends on $\boldsymbol{s}$ through its Euclidean length $||\boldsymbol{s}||$. Then, the Fourier transform of f is also isotropic and has the form,*

$$\hat{f}(\boldsymbol{\omega}) = \frac{c}{||\boldsymbol{\omega}||^{\frac{d-2}{2}}} \int_0^\infty J_{\frac{d-2}{2}}(r||\boldsymbol{\omega}||) f(r) r^{\frac{d}{2}} dr$$

where c is a constant depending on d and not f or $\boldsymbol{\omega}$. Here, $J_\alpha(z)$ is a Bessel function of the first kind of order α (Watson (1995)).

The above lemma arises from converting the Fourier transform integral into spherical coordinates. An example of such a calculation can be found Stein (1999) (p. 43), so we omit the details here. Next, for our choice of $m(\boldsymbol{s})$ in (7.21), we show that $m(\boldsymbol{s})$ satisfies a certain integral representation.

Lemma 7.2.3. *For any $N \in \mathbb{N}$, consider the extension given in (7.21). Then, for*

$1 \leq k \leq N$, *we have the following integral representation for the Fourier transform* $\hat{m}(\boldsymbol{\omega})$ *of* $m(\boldsymbol{s})$,

$$\hat{m}(\boldsymbol{\omega}) = (-1)^k \frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}+k-1}} \int_1^\infty \frac{\sum_{j=1}^{k} a_j g^{(j)}(r) r^{j-1}}{r^{2k-1}} J_{\frac{d}{2}+k-1}(r||\boldsymbol{\omega}||) r^{\frac{d}{2}+k} dr \tag{7.22}$$

where $a_j \in \mathbb{Z}, a_k = 1$ *and* c *is a constant depending on* d *and not* f *or* $\boldsymbol{\omega}$. *Here,* $g^{(j)}(r)$ *represents the* j^{th} *derivative of* $g(r)$.

The proof, while not complicated, involves repeated integration by parts. We defer the proof to Section 7.5. As a consequence of this integral representation, we have the following decay estimate for the Fourier transform $\hat{m}(\boldsymbol{\omega})$.

Corollary 7.2.4. *For any* $N \in \mathbb{N}$, *there exists a square integrable extension of the function* $\mathbb{1}_D(\boldsymbol{s})$ *to all of* $\mathbb{R}^d$, *whose Fourier transform* $\hat{m}(\boldsymbol{\omega})$ *satisfies,*

$$|\hat{m}(\boldsymbol{\omega})| \leq \frac{C}{||\boldsymbol{\omega}||^{\frac{d-1}{2}+N}} \tag{7.23}$$

where C *is a constant depending on* d, k, N, *but not* f *or* $\boldsymbol{\omega}$.

Proof. Consider the extension given in Lemma 7.2.3 and the integral representation of its Fourier transform for $k = N$. Using the bound $J_\alpha(z) = O(z^{-1/2})$ (Yadrenko (1983), p. 38), we see that,

$$|\hat{m}(\boldsymbol{\omega})| \leq \frac{K}{||\boldsymbol{\omega}||^{\frac{d-1}{2}+N}} \int_1^\infty \frac{\sum_{j=1}^{N} a_j g^{(j)}(r) r^{j-1}}{r^{2N-1}} r^{\frac{d}{2}+N-1/2} dr$$

$$= \frac{K}{||\boldsymbol{\omega}||^{\frac{d-1}{2}+N}} \int_1^\infty \left[\sum_{j=1}^{N} a_j g^{(j)}(r) r^{j-1} \right] r^{\frac{d}{2}-N+1/2} dr$$

Since $g(r) = e^{-(r-1)^{N+1}}$, the integral is seen to be convergent over $(1, \infty)$. □

We can now show that Yadrenko's condition in Theorem 7.1.7 holds.

Proposition 7.2.3. *Let $\{y(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a stationary Gaussian random field with mean parameter μ and spectral density $f(\boldsymbol{\omega})$. Suppose there exist some constants $A > 0, \tau > \frac{d}{2}$ such that $f(\boldsymbol{\omega}) \geq \frac{A}{(1+||\boldsymbol{\omega}||^2)^\tau}$ for all $\boldsymbol{\omega} \in \mathbb{R}^d$. Then μ is not microergodic.*

Proof. Recall the condition (7.10) of Theorem 7.1.7. If $\hat{m}(\boldsymbol{\omega})$ is isotropic, then the condition is equivalent to showing,

$$\int_0^\infty (\hat{m}(r))^2 (1+r^2)^\tau r^{d-1} dr < \infty$$

Now, let $N \in \mathbb{N}$ satisfy $N > \tau + \frac{1}{2}$ and consider the isotropic Fourier transform from Lemma 7.2.3. Splitting the integral from $[0, 1]$ and $[1, \infty)$, we have,

$$\int_0^1 (\hat{m}(r))^2 (1+r^2)^\tau r^{d-1} dr + \int_1^\infty (\hat{m}(r))^2 (1+r^2)^\tau r^{d-1} dr$$

Since Fourier transform of an integrable function is bounded, the first integral above is finite. For the second, we have $|\hat{m}(r)|^2 \leq C^2 r^{-(d-1)-2N}$ by Corollary 7.2.4. Thus, the second integral satisfies,

$$\int_1^\infty (\hat{m}(r))^2 (1+r^2)^\tau r^{d-1} dr \leq C^2 \int_1^\infty \frac{(1+r^2)^\tau}{r^{d-1+2N}} r^{d-1} dr = C^2 \int_1^\infty \frac{(1+r^2)^\tau}{r^{2N}} dr$$

Since the integrand asymptotically behaves like $r^{-2N+2\tau}$, the integral converges by the assumption on N, τ. Thus, by Theorem 7.1.7, the probability measures $\mathbb{P}_\mu, \mathbb{P}_0$ induced by $\mu \neq 0$ and $\mu = 0$ are equivalent. □

We note that since the Matérn spectral density satisfies (7.20), Proposition 7.2.3 gives an alternative proof to the previously known Theorem 7.1.9 in showing that μ is not microergodic. We state this formally as corollary.

Corollary 7.2.5. *Let $\{y(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a constant mean Gaussian random field with mean parameter μ and Matérn covariance. Then μ is not microergodic.*

Thus, Corollaries 7.2.1 and 7.2.5 show that $\lim_{n\to\infty} \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1}$ must remain finite, where $\boldsymbol{\Sigma}$ is the Matérn covariance matrix and so μ cannot be consistently estimated. The implication of this result from a spatial statistics perspective is that one may set the mean to zero in the ordinary kriging model and obtain asymptotically optimal predictions (Stein (1999)).

7.3 Microergodicity and estimation in regression models

In this section, assume now that we have a covariate present in a simple linear regression model,

$$y(\boldsymbol{s}) = \beta_0 + \beta_1 x(\boldsymbol{s}) + e(\boldsymbol{s}) \tag{7.24}$$

where $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ are independent, mean-zero Gaussian random fields with Matérn covariances. We assume in this section that the covariance parameters of $e(\boldsymbol{s})$ are

known and that only $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is unknown.

7.3.1 Microergodicity of the regression parameters

We want to determine conditions under which the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ are microergodic (or non-microergodic). That is, whether two probability measures $\mathbb{P}_{\boldsymbol{\beta}}, \mathbb{P}_{\boldsymbol{\beta}^*}$ parametrized by $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$ are equivalent or singular. We note that the probability measures $\mathbb{P}_{\boldsymbol{\beta}}, \mathbb{P}_{\boldsymbol{\beta}^*}$ are induced by the joint Gaussian random field $(x(\boldsymbol{s}), y(\boldsymbol{s}))^T$. For the intercept, we can use the results from the previous section to conclude that β_0 is not microergodic.

Proposition 7.3.1. *Let $\{y(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be the random field given in (7.24). If $e(\boldsymbol{s})$ has Matérn covariance, then β_0 is not microergodic.*

Proof. Unconditional on the sample paths of $x(\boldsymbol{s})$, $y(\boldsymbol{s})$ is a stationary Gaussian random field with constant mean β_0 and spectral density $f(\boldsymbol{\omega}) = \beta_1^2 f_x(\boldsymbol{\omega}) + f_e(\boldsymbol{\omega})$, where f_x, f_e are the spectral densities of $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ respectively. Then the same extension from Theorem 7.1.7 applies here with $f(\boldsymbol{\omega})$ replacing $f_e(\boldsymbol{\omega})$. □

A consequence of this result is that the intercept β_0 cannot be consistently estimated and thus, we should exclude the intercept from the model. For microergodicity of the slope parameter β_1, by Proposition 7.1.1, it is equivalent to consider the Gaussian probability measures conditional on $x(\boldsymbol{s})$. Conditional on the sample paths of $x(\boldsymbol{s})$, the mean function is $m(\boldsymbol{s}) = \beta_1 x(\boldsymbol{s})$. By Theorem 7.1.9, β_1 is microergodic if and only if the sample paths of $x(\boldsymbol{s})$ do not belong to $W_2^{\ell}(D)$ almost surely. Here $\ell = \nu + \frac{d}{2}$, where ν is the smoothness parameter of the error Matérn covariance. Recalling

Theorems 2.5.11 and 2.5.12 of Scheuerer (2010), we have spectral conditions that determine if the sample paths of $x(\boldsymbol{s})$ belong $W_2^\ell(D)$ almost surely. If $x(\boldsymbol{s})$ also has Matérn covariance, then these spectral conditions lead to a simple inequality involving the Matérn smoothness parameters of $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$.

Proposition 7.3.2. *Let $x(\boldsymbol{s})$ have Matérn covariance with smoothness parameter ν_x and denote $\ell = \nu + \frac{d}{2}$. If ℓ is an integer, then $x(\boldsymbol{s}) \in W_2^\ell(D)$ a.s. if and only if $\nu_x > \ell$. If ℓ is fractional, then the condition $\nu_x > \ell$ is sufficient for $x(\boldsymbol{s}) \in W_2^\ell(D)$ a.s.*

Proof. By Theorems 2.5.11 and 2.5.12, we need to show that,

$$\begin{cases} \int_{\mathbb{R}^d} ||\boldsymbol{\omega}||^{2\ell} f_x(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty & \ell \in \mathbb{N} \\ \int_{\mathbb{R}^d} (\log(1 + ||\boldsymbol{\omega}||))^{1+\alpha} ||\boldsymbol{\omega}||^{2\ell} f_x(\boldsymbol{\omega}) d\boldsymbol{\omega} < \infty & \text{for some } \alpha > 0, \quad \ell \in \mathbb{R}_+ \setminus \mathbb{N} \end{cases}$$

where $f_x(\boldsymbol{\omega}) \asymp (1 + ||\boldsymbol{\omega}||^2)^{-\nu_x - \frac{d}{2}}$. The calculations for either integral lead to the same conclusion that $\nu_x > \ell$. We note that the calculations are very similar to the one given in the proof of Proposition 2.5.1, so we omit the details here. □

The above result shows that if $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ both have Matérn covariances, then the smoothness parameter of $x(\boldsymbol{s})$ cannot be too large compared to that of $e(\boldsymbol{s})$ if we want to include a covariate with consistently estimated coefficient in the regression model (7.24). In particular, Proposition 7.3.2 shows that the quantity $\ell = \nu + \frac{d}{2}$ acts as a critical smoothness for $x(\boldsymbol{s})$. If the smoothness parameter ν_x exceeds this critical smoothness, then the slope parameter is not microergodic. This corroborates

the remarks by Stein (1999) given at the beginning of this chapter. In light of these remarks, it is reasonable to believe that the slope parameter can be consistently estimated as long as $x(\boldsymbol{s}) \notin W_2^{\ell}(D)$ a.s. where $\ell = \nu + \frac{d}{2}$. In the next section, we consider two familiar estimators: the ordinary least squares (OLS) estimator and the maximum likelihood estimator (MLE) of β_1. Somewhat surprisingly, we show that the OLS estimator remains inconsistent under mild conditions regardless of the smoothness of $x(\boldsymbol{s})$.

7.3.2 Inconsistency of the OLS estimator of the slope

First, let us consider the ordinary least squares estimator for β_1. By Proposition 7.3.1, we may set $\beta_0 = 0$ without loss of generality. Then, conditional on $\boldsymbol{x} = (x(\boldsymbol{s}_1), \cdots, x(\boldsymbol{s}_n))^T$, the OLS estimator of β_1 is given by,

$$\tilde{\beta}_{1,n} = \frac{\boldsymbol{x}^T \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{x}} = \frac{\sum_{\boldsymbol{s}_i} x(\boldsymbol{s}_i) y(\boldsymbol{s}_i)}{\sum_{\boldsymbol{s}_i} x^2(\boldsymbol{s}_i)} = \beta_{1,0} + \frac{\sum_{\boldsymbol{s}_i} x(\boldsymbol{s}_i) e(\boldsymbol{s}_i)}{\sum_{\boldsymbol{s}_i} x^2(\boldsymbol{s}_i)} \tag{7.25}$$

where $\beta_{1,0}$ denotes the true β_1 parameter. Note that the asymptotic behavior of the sums in (7.25) depends on the sampling scheme of $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$. In our setting, we assume that the points become increasingly dense in D and fill out the entire region as $n \to \infty$. We formally state the assumption below.

Assumption 7.3.1. *Let $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\} \in D$ represent the sampling locations. For*

any Borel subset $A \subset D$, define the empirical measure of $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$ as,

$$\mu_n(A) = \frac{1}{n} \sum_{\boldsymbol{s}_i} \delta_{\boldsymbol{s}_i}(A) \tag{7.26}$$

where $\delta_{\boldsymbol{s}}(A)$ is the Dirac delta measure of the set A. Then μ_n converges weakly (van der Vaart (1998)) to a measure μ with a strictly positive and bounded density. That is,

$$\lim_{n\to\infty} \int_D f(\boldsymbol{s}) d\mu_n(\boldsymbol{s}) = \int_D f(\boldsymbol{s}) d\mu(\boldsymbol{s})$$

for any continuous function f on D. If the empirical measure is random, then we assume that conditional on $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$, weak convergence holds almost surely.

The empirical measure defined in (7.26) can be deterministic or random depending on whether $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$ are deterministic or random. For example, if the locations are sampled through an inhomogeneous Poisson point process with intensity measure μ (not necessarily Lebesgue), then the convergence defined in Assumption 7.3.1 will hold almost surely with limiting measure μ. As a second example, Lahiri (1996) assumes the following deterministic sampling scheme. Since the region $D \subset \mathbb{R}^d$ is compact, it is contained within a hyperrectangle $R = [a_1, b_1] \times ... \times [a_d, b_d]$ with smallest possible volume. For every n, we may subdivide each $[a_j, b_j]$, $1 \leq j \leq d$ into N_{jn} equally spaced intervals with length $\frac{b_j - a_j}{N_{jn}}$. This will partition R into smaller hypercubes Δ_n, each with equal volume $|\Delta_n| = \prod_{j=1}^{n} \left(\frac{b_j - a_j}{N_{jn}} \right)$ converging uniformly to 0. Within each cube we pick a point that falls inside in D. This

also satisfies the convergence assumption since this is how the Lebesgue measure is typically constructed.

Under the second sampling scheme described above, Lahiri (1996) shows with deterministic covariates that the OLS estimator is inconsistent by showing that the sums in (7.25) converge to integrals of random fields with respect to the Lebesgue measure. We take a similar approach using the sampling scheme described in Assumption 7.3.1 and random covariates. Before we state the inconsistency result, we first give a definition of the integral of a random field.

Definition 7.3.2. *Let $\{z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ be a random field on D. Assume that the sampling locations $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$ satisfy Assumption 7.3.1. If $z(\boldsymbol{s})$ has continuous sample paths a.s. on D, then the integral of $z(\boldsymbol{s})$ is defined as the a.s. limit,*

$$\int_D z(\boldsymbol{s}) d\mu(\boldsymbol{s}) := \lim_{n\to\infty} \frac{1}{n} \sum_{\boldsymbol{s}_i} z(\boldsymbol{s}_i)$$

We can apply this to the OLS estimator in (7.25). If we normalize the sums in the numerator and denominator by n, we obtain,

$$\tilde{\beta}_{1,n} - \beta_{1,0} = \frac{\frac{1}{n} \sum_{\boldsymbol{s}_i} x(\boldsymbol{s}_i) e(\boldsymbol{s}_i)}{\frac{1}{n} \sum_{\boldsymbol{s}_i} x^2(\boldsymbol{s}_i)} \tag{7.27}$$

By Definition 7.3.2, the numerator and denominator on the right hand side of (7.27) each converge to integrals almost surely. The following proposition is a variant Theorem 1 of Lahiri (1996), who considered the specific sampling scheme described

above with the limiting measure to be the Lebesgue measure.

Proposition 7.3.3. *Assume that the sampling locations* $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$ *satisfy Assumption 7.3.1. Suppose that* $x(\boldsymbol{s})$ *and* $e(\boldsymbol{s})$ *have Matérn covariances. Then,*

$$\tilde{\beta}_{1,n} - \beta_{1,0} \xrightarrow{a.s.} \frac{\int_D x(\boldsymbol{s})e(\boldsymbol{s})d\mu(\boldsymbol{s})}{\int_D x^2(\boldsymbol{s})d\mu(\boldsymbol{s})} \tag{7.28}$$

which is a non-degenerate random variable.

Proof. If $x(\boldsymbol{s}), e(\boldsymbol{s})$ have Matérn covariances, then the sample paths of $x(\boldsymbol{s}), e(\boldsymbol{s})$ are bounded almost surely by Proposition 2.5.1. Thus, the integrals on the right-hand side of (7.28) are separately and jointly non-degenerate random variables. □

Thus, the OLS estimator $\tilde{\beta}_{1,n}$ cannot be consistent, regardless of whether or not $x(\boldsymbol{s}) \in W_2^\ell(D)$ a.s.

7.3.3 Consistency of MLE of the slope

Now we investigate the maximum likelihood estimator $\hat{\beta}_{1,n}$. Once again, we set the intercept β_0 to zero without loss of generality. With a known error covariance matrix $\boldsymbol{\Sigma}$, the negative log-likelihood of $\boldsymbol{y}|\boldsymbol{x}$ is, up to a constant,

$$L(\beta) = \frac{1}{2}(\boldsymbol{y} - \beta_1\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \beta_1\boldsymbol{x}) \tag{7.29}$$

From this expression, the MLE (also the GLS estimator) of β_1 is calculated as $\hat{\beta}_{1,n} = \dfrac{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y}}{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}$. Conditional on $\boldsymbol{x}$, the MLE $\hat{\beta}_{1,n}$ has distribution,

$$\hat{\beta}_{1,n} \sim N\left(\beta_{1,0}, \frac{1}{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}}\right)$$

Note that the quantity $\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$ is the conditional Fisher information for β_1. If this quantity remains finite as $n \to \infty$, then the asymptotic distribution of $\hat{\beta}_{1,n}$ is a non-degenerate Gaussian random variable and so $\hat{\beta}_{1,n}$ cannot be consistent. On the other hand, if this quantity diverges, the variance of $\hat{\beta}_{1,n}$ goes to 0 and thus it is consistent. In general, since $\boldsymbol{\Sigma}^{-1}$ does not have an analytical form (except in cases like the OU process on the real line), it is difficult to determine if this quantity diverges analytically. However, we can use similar arguments as in the constant mean case, in particular with the entropy distance in Lemma 2.5.3. Let $\mathbb{P}_{\beta_1}$ and $\mathbb{P}_{\beta_1^*}$ be the conditional probability measures induced by β_1 and β_1^* respectively. Then from the form of the likelihood in (7.29), we can calculate the log-likelihood ratio of $\mathbb{P}_{\beta_1}$ to $\mathbb{P}_{\beta_1^*}$ to be,

$$\log p_n = (\beta_1 - \beta_1^*)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{y} + \frac{1}{2}((\beta_1^*)^2 - \beta_1^2)\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

Then, it is straightforward to calculate the conditional entropy distance as,

$$J_n = \mathbb{E}_{\beta_1}[\log p_n|\boldsymbol{x}] - E_{\beta_1^*}[\log p_n|\boldsymbol{x}] = (\beta_1 - \beta_1^*)^2\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

which is proportional to the conditional Fisher information. By Lemma 2.5.3, $\mathbb{P}_{\beta_1} \perp \mathbb{P}_{\beta_1^*}$ if and only if the conditional entropy distance diverges. The above remarks, combined with Theorem 7.1.9, give necessary and sufficient conditions for the consistency of $\hat{\beta}_{1,n}$.

Proposition 7.3.4. *Let $y(\boldsymbol{s}) = \beta_1 x(\boldsymbol{s}) + e(\boldsymbol{s})$, where $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ are independent Gaussian random fields and β_1 unknown. Suppose that $e(\boldsymbol{s})$ has a known Matérn covariance structure with smoothness ν. Finally, suppose that $x(\boldsymbol{s}), y(\boldsymbol{s})$ are observed at an increasingly dense, nested set of locations $\{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$ with corresponding data vectors $\boldsymbol{x}, \boldsymbol{y}$. Then, the following statements are equivalent:*

1. *The conditional Fisher information $\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}$ for β_1 diverges almost surely.*
2. *$x(\boldsymbol{s}) \notin W_2^{\ell}(D)$ a.s. where $\ell = \nu + \dfrac{d}{2}$.*
3. *The slope parameter is microergodic, that is, $\mathbb{P}_{\beta_1} \perp \mathbb{P}_{\beta_1^*} \iff \beta_1 \neq \beta_1^*$.*

Under any of these conditions, the MLE $\hat{\beta}_{1,n} = \dfrac{\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x}}$ is consistent for β_1.

7.4 Joint estimation of the slope and covariance parameters

The previous sections assumed that the Matérn covariance parameters of the error were known. In this section, we assume only that the Matérn smoothness parameter is known. The case of unknown ν is not addressed in this thesis. We should note however, that it has been recently proved that ν can be consistently estimated under infill asymptotics (Loh et al. (2021)). If joint estimation of β_1 and the Matérn covariance parameters $(\sigma^2, \theta)^T$ is of interest, then one would have to establish the

equivalence or singularity of the conditional measures $\mathbb{P}_1, \mathbb{P}_2$ induced by $(\beta_1, \sigma_1^2, \theta_1)^T$ and $(\beta_1^*, \sigma_2^2, \theta_2)^T$ respectively. Fortunately, as the following result from Stein (1999) states, we can determine the equivalence or singularity of $\mathbb{P}_1$ and $\mathbb{P}_2$ by separately considering β_1 and $(\sigma^2, \theta)^T$. First, let $\mathbb{P}_{\beta_1}, \mathbb{P}_{\beta_1^*}$ be the conditional Gaussian measures induced by β_1 and β_1^* respectively under known $(\sigma^2, \theta)^T$. Next, let $\mathbb{Q}_{\sigma_1^2,\theta_1}, \mathbb{Q}_{\sigma_2^2,\theta_2}$ by the zero-mean Gaussian measures induced by $(\sigma_1^2, \theta_1)^T$ and $(\sigma_2^2, \theta_2)^T$ respectively. The following result is a version of Corollary 5 (p. 117) in Stein (1999).

Lemma 7.4.1. $\mathbb{P}_1 \equiv \mathbb{P}_2$ *if and only if* $\mathbb{P}_{\beta_1} \equiv \mathbb{P}_{\beta_1^*}$ *and* $\mathbb{Q}_{\sigma_1^2,\theta_1} \equiv \mathbb{Q}_{\sigma_2^2,\theta_2}$.

Our version is slightly different than that Corollary 5 of Stein (1999), who considers deterministic mean functions. But since we are conditioning on the sample paths of $x(\boldsymbol{s})$, we can use Stein's result here due to Proposition 7.1.1.

For a zero mean Gaussian random field with Matérn covariance, there are known results concerning estimation of the covariance parameters under infill asymptotics. Generalizing the result from Ying (1991) to higher spatial dimensions and known smoothness ν, Zhang (2004) proved that the Matérn covariance parameters $(\sigma^2, \theta)^T$ are not individually microergodic on a bounded domain $D \subset \mathbb{R}^d, d = 1, 2, 3$, and thus cannot be consistently estimated. However, it is shown by Zhang that under known smoothness ν, the quantity $\sigma^2\theta^{2\nu}$ is microergodic using results on the equivalence of Gaussian measures from Yadrenko (1983) and Stein (1999). The following result is a combination of Theorem 3 (consistency) of Zhang (2004) and Theorem 3 (asymptotic normality) of Wang and Loh (2011).

Theorem 7.4.2. *Let* $d \leq 3$ *and* $(\sigma^2, \theta)^T$ *be the Matérn covariance parameters of*

a zero mean Gaussian random field on $D \subset \mathbb{R}^d$. *Let* $L_n(\sigma^2, \theta)$ *be the likelihood function based on a nested sequence of observations* D_n *in* D. *For a fixed* θ, *let* $\hat{\sigma}_n^2 = \arg\max_{\sigma^2} L_n(\sigma^2, \theta)$ *be the maximum likelihood estimator of* σ^2. *If* $(\sigma_0^2, \theta_0)^T$ *are the true parameters, then under the true probability measure* $\mathbb{P}_{\sigma_0^2, \theta_0}$,

$$\begin{aligned} \hat{\sigma}_n^2 \theta^{2\nu} &\xrightarrow{a.s.} \sigma_0^2 \theta_0^{2\nu} \\ \sqrt{n}(\hat{\sigma}_n^2 \theta^{2\nu} - \sigma_0^2 \theta_0^{2\nu}) &\xrightarrow{D} N(0, 2(\sigma_0^2 \theta_0^{2\nu})^2) \end{aligned}$$

Du et al. (2009) also proved the above asymptotic normality result but in the case $d = 1$ on a bounded interval. The above results prove that the scale parameter can be fixed to any value θ and we would still get consistency and asymptotic normality. Kaufman and Shaby (2013) extended these results by establishing consistency and asymptotic normality of $\hat{\sigma}_n^2 \hat{\theta}_n^{2\nu}$, where $\hat{\theta}_n$ is the maximum likelihood estimator of θ. The following result can be found in Theorem 2 of their paper.

Theorem 7.4.3. *Let* $d \leq 3$ *and* $(\sigma^2, \theta)^T$ *be the Matérn covariance parameters of a zero mean Gaussian random field on* $D \subset \mathbb{R}^d$. *Let* $L_n(\sigma^2, \theta)$ *be the likelihood function based on a nested sequence of observations* D_n *in* D. *Let* $(\hat{\sigma}_n^2, \hat{\theta}_n)^T = \arg\max_{\sigma^2, \theta} L_n(\sigma^2, \theta)$ *be the MLE of* $(\sigma^2, \theta)^T$. *If* $(\sigma_0^2, \theta_0)^T$ *are the true parameters, then under the true probability measure* $\mathbb{P}_{\sigma_0^2, \theta_0}$,

$$\begin{aligned} \hat{\sigma}_n^2 \hat{\theta}_n^{2\nu} &\xrightarrow{a.s.} \sigma_0^2 \theta_0^{2\nu} \\ \sqrt{n}(\hat{\sigma}_n^2 \hat{\theta}_n^{2\nu} - \sigma_0^2 \theta_0^{2\nu}) &\xrightarrow{D} N(0, 2(\sigma_0^2 \theta_0^{2\nu})^2) \end{aligned}$$

We should note that both theorems above hold when $d \leq 3$. It has been shown that when $d \geq 5$, the parameters θ and σ^2 are individually microergodic and can be consistently estimated (Anderes (2009)). The case $d = 4$ remains an open problem.

If we combine the previous results on the regression slope and the results of Zhang (2004) on the Matérn covariance parameters, Lemma 7.4.1 suggests that if $x(\boldsymbol{s}) \notin W_2^{\ell}(D)$, then the microergodic parameters are $(\beta_1, \sigma^2\theta^{2\nu})^T$. The next step is to determine whether or not the joint MLE of $(\beta_1, \sigma^2\theta^{2\nu})^T$ is consistent and asymptotically normal. The conditional negative log-likelihood of $\boldsymbol{y}|\boldsymbol{x}$ is up to a constant,

$$L_n(\beta_1, \sigma^2, \theta) = \frac{1}{2}\log\big(\det\big(\sigma^2\boldsymbol{\Sigma}(\theta)\big)\big) + \frac{1}{2\sigma^2}(\boldsymbol{y} - \beta_1\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y} - \beta_1\boldsymbol{x})$$

Consider a re-parametrization of the likelihood in terms of the microergodic parameter $\phi = \sigma^2\theta^{2\nu}$,

$$L_n(\beta_1, \phi, \theta) = \frac{1}{2}\log\left(\det\left(\frac{\phi}{\theta^{2\nu}}\boldsymbol{\Sigma}(\theta)\right)\right) + \frac{\theta^{2\nu}}{2\phi}(\boldsymbol{y} - \beta_1\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y} - \beta_1\boldsymbol{x}) \quad (7.30)$$

7.4.1 Fixed scale parameter θ

We first consider the situation where the non-microergodic scale parameter θ is fixed, similar to Theorem 7.4.2. Stein (1999), p. 175, gives a discussion on estimation in the presence of non-microergodic parameters, citing results by Crowder (1976). Let $\boldsymbol{\tau} = (\boldsymbol{\tau}_1^T, \boldsymbol{\tau}_2^T)^T$ denote the parameters of interest, with $\boldsymbol{\tau}_1$ being microergodic, and $\boldsymbol{\tau}_1$ being non-microergodic. Stein (1999) conjectures that if $\boldsymbol{\tau}_2$ is set to some

fixed value and not estimated, the asymptotic behavior of the MLE of $\boldsymbol{\tau}_1$ will be the same as if $\boldsymbol{\tau}_2$ were known. We verify this statement in our setting where $\boldsymbol{\tau} = (\beta_1, \sigma^2, \theta)^T$ and θ is the non-microergodic parameter. Let $L_\theta(\beta_1, \phi) = L_n(\beta_1, \phi, \theta)$ denote the likelihood in (7.30) for any $\theta > 0$. Then the 2×2 Hessian matrix (or observed Fisher information) of $L_\theta(\beta_1, \phi)$ is readily calculated as,

$$\boldsymbol{H}_\theta(\beta_1, \phi) = \begin{pmatrix} \frac{\theta^{2\nu}}{\phi}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x} & \frac{\theta^{2\nu}}{\phi^2}(\boldsymbol{y} - \beta_1\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x} \\ \frac{\theta^{2\nu}}{\phi^2}(\boldsymbol{y} - \beta_1\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x} & -\frac{n}{2\phi^2} + \frac{\theta^{2\nu}}{\phi^3}Q \end{pmatrix} \tag{7.31}$$

where $Q = (\boldsymbol{y} - \beta_1\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y} - \beta_1\boldsymbol{x})$. After taking the expectation conditional on $\boldsymbol{x}$, the conditional Fisher information is,

$$\boldsymbol{\mathscr{I}}_\theta(\beta_1, \phi) = \mathbb{E}\left[\boldsymbol{H}_\theta(\beta_1, \phi)\right] = \begin{pmatrix} \frac{\theta^{2\nu}}{\phi}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x} & 0 \\ 0 & \frac{n}{2\phi^2} \end{pmatrix} \tag{7.32}$$

For the MLE $(\hat{\beta}_{1,n}(\theta), \hat{\phi}_n(\theta))^T = \underset{\beta_1,\phi}{\arg\min}\, L_\theta(\beta_1, \phi)$, there are known sufficient regularity conditions that ensure consistency and asymptotic normality (Mardia and Marshall (1984), Sweeting (1980)). In particular, by quoting Mardia and Marshall (1984), p. 138, we require continuity, growth and convergence of the Hessian matrix. The following general theorem from Sweeting (1980) gives conditions that ensure consistency and asymptotic normality of MLE estimators.

Theorem 7.4.4. *Let $L(\boldsymbol{\tau})$ be the negative log-likelihood function depending on a parameter $\boldsymbol{\tau} \in \mathbb{R}^p$. Assume that the true parameter $\boldsymbol{\tau}_0$ is known to lie in the interior of a compact subset of $\mathbb{R}^p$. Let $\mathbb{P}_0$ denote true probability measure parametrized by*

$\boldsymbol{\tau}_0$. *Suppose the following conditions hold.*

1. *(Continuity) The Hessian* $\boldsymbol{H}(\boldsymbol{\tau})$ *matrix of second order derivatives is continuous in a neighbourhood around* $\boldsymbol{\tau}_0$.

2. *(Growth) Under the true probability measure* $\mathbb{P}_0$*, the smallest eigenvalue of the Fisher information* $\boldsymbol{\mathcal{I}}(\boldsymbol{\tau}) = \mathbb{E}[H(\boldsymbol{\tau})]$ *diverges in probability to* ∞.

3. *(Convergence) Under the true probability measure* $\mathbb{P}_0$*,*

$$\boldsymbol{\mathcal{I}}^{-1/2}\boldsymbol{H}\boldsymbol{\mathcal{I}}^{-1/2} \xrightarrow{P} \boldsymbol{I}_p$$

where $\boldsymbol{I}_p$ *is the* $p \times p$ *identity matrix.*

Then the MLE $\hat{\boldsymbol{\tau}} = \arg\min_{\boldsymbol{\tau}} L(\boldsymbol{\tau})$ *is consistent under the true probability measure* $\mathbb{P}_0$ *and,*

$$\boldsymbol{\mathcal{I}}_0^{1/2}(\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}_0) \xrightarrow{D} N(\boldsymbol{0}, \boldsymbol{I}_p)$$

where $\boldsymbol{\mathcal{I}}_0 = \boldsymbol{\mathcal{I}}(\boldsymbol{\tau}_0)$ *is the Fisher information evaluated at true parameters.*

For our problem, the first condition is verifiable from (7.31), recalling that the non-microergodic parameter θ is being treated as a constant. The second condition is also verifiable from (7.32) since by Proposition 7.3.4, the first diagonal element of (7.32) diverges almost surely under the true probability measure and it is clear that the second diagonal diverges as well. For the final condition, we can easily calculate the matrix product $\boldsymbol{\mathcal{I}}_\theta^{-1/2}\boldsymbol{H}_\theta\boldsymbol{\mathcal{I}}_\theta^{-1/2}$ since these are 2×2 matrices and $\boldsymbol{\mathcal{I}}_\theta$ is diagonal.

After some minor simplifications, the matrix product equals,

$$\begin{pmatrix} 1 & \frac{\sqrt{2}\theta^{\nu}}{\phi\sqrt{n}} \frac{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y}-\beta_1\boldsymbol{x})}{\sqrt{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x}}} \\ \frac{\sqrt{2}\theta^{\nu}}{\phi\sqrt{n}} \frac{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y}-\beta_1\boldsymbol{x})}{\sqrt{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x}}} & \frac{2\theta^{2\nu}}{\phi n}Q - 1 \end{pmatrix} \tag{7.33}$$

We require that the bottom right diagonal $\frac{2\theta^{2\nu}}{n\phi}Q - 1$ converge in probability to 1 and the off-diagonal $\frac{\sqrt{2}\theta^{\nu}}{\phi\sqrt{n}} \frac{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y}-\beta_1\boldsymbol{x})}{\sqrt{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x}}}$ converge in probability to 0 under the true measure $\mathbb{P}_0$. In the proof of the following result, we show these conditions hold.

Proposition 7.4.1. *Suppose that the sample paths of $x(\boldsymbol{s})$ do not belong to $W_2^{\ell}(D)$ almost surely, where $\ell = \nu + \frac{d}{2}, d \leq 3$ and ν is the Matérn smoothness of $e(\boldsymbol{s})$. Then the conditions of Theorem 7.4.4 are satisfied. Thus, for any $\theta > 0$, the MLE of $(\beta_1(\theta), \phi(\theta))^T$ is consistent under the true probability measure $\mathbb{P}_0$ and,*

$$\begin{pmatrix} v^{1/2}(\hat{\beta}_{1,n}(\theta) - \beta_{1,0}) \\ \sqrt{n}(\hat{\phi}_n(\theta) - \phi_0) \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 2\phi_0^2 \end{pmatrix} \right) \tag{7.34}$$

where $v = \frac{\theta_0^{2\nu}}{\phi_0}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta_0)\boldsymbol{x}$, $\hat{\phi}_n(\theta) = \hat{\sigma}^2\theta^{2\nu}$ and $\phi_0 = \sigma_0^2\theta_0^{2\nu}$.

Proof. From the preceding discussion, we need to show that (7.33) converges to the 2×2 identity matrix under the true measure $\mathbb{P}_0$. Under $\mathbb{P}_0$, note that,

$$Q = (\boldsymbol{y} - \beta_{1,0}\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\theta_0)(\boldsymbol{y} - \beta_{1,0}\boldsymbol{x}) = \boldsymbol{e}^T\boldsymbol{\Sigma}^{-1}(\theta_0)\boldsymbol{e}$$

After re-parametrizing back in terms of σ^2, we have $\frac{2\theta_0^{2\nu}}{n\phi_0}Q = \frac{2}{n\sigma_0^2}\boldsymbol{e}^T\boldsymbol{\Sigma}^{-1}(\theta_0)\boldsymbol{e}$. From properties of the multivariate normal distribution, we know that this quadratic form equals $\frac{2}{n}\sum_{i=1}^{n}\epsilon_i^2$ where $\epsilon_i, i = 1, \cdots, n$ are i.i.d $N(0,1)$. By the LLN, this converges in probability to 2 and so the diagonal term $\frac{2\theta^{2\nu}}{n\phi}Q - 1$ converges in probability to 1 as desired. Next, under $\mathbb{P}_0$ the off-diagonal term of (7.33) is proportional to,

$$\frac{1}{\sqrt{n}}\frac{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta_0)(\boldsymbol{y} - \beta_{1,0}\boldsymbol{x})}{\sqrt{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta_0)\boldsymbol{x}}} = \frac{1}{\sqrt{n}}\frac{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta_0)\boldsymbol{e}}{\sqrt{\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta_0)\boldsymbol{x}}} \sim N\left(0, \frac{1}{n}\right)$$

conditionally given $\boldsymbol{x}$. Thus the off-diagonal term converges to 0 in probability and $\boldsymbol{\mathscr{I}}_\theta^{-1/2}\boldsymbol{H}_\theta\boldsymbol{\mathscr{I}}_\theta^{-1/2} \xrightarrow{P} \boldsymbol{I}$ verifying the convergence conditions of Theorem 7.4.4. □

Note that as a simple corollary, we get automatically get consistency and asymptotic normality of $\hat{\sigma}_n^2 = \frac{\hat{\phi}_n}{\theta^{2\nu}}$ by fixing θ, even though this parameter is non-microergodic otherwise. In particular, under $\mathbb{P}_0$,

$$\begin{aligned} \hat{\sigma}_n^2 &\xrightarrow{P} \frac{\sigma_0^2\theta_0^{2\nu}}{\theta^{2\nu}} \\ \sqrt{n}\left(\hat{\sigma}_n^2 - \frac{\sigma_0^2\theta_0^{2\nu}}{\theta^{2\nu}}\right) &\xrightarrow{D} N\left(0, 2\left(\frac{\sigma_0^2\theta_0^{2\nu}}{\theta^{2\nu}}\right)^2\right) \end{aligned}$$

We perform a simulation study in Chapter 8 to illustrate these results.

7.4.2 Estimated scale parameter θ

Now we consider the effect of estimating θ. A Fisher information analysis as in Theorem 7.4.4 for the full parameter set $(\beta_1, \phi, \theta)^T$ would be difficult in this situation for a few reasons. First, we would need to compute second order derivatives of the

likelihood with respect to θ. However, derivatives of Matérn covariance functions with respect to θ are generally (unless the smoothness is a half integer) intractable. Moreover, even if closed form derivatives exist, we would need to ensure that the extra information added in estimating θ does not disrupt the regularity of the Fisher information for $(\beta_1, \phi)^T$ alone. Fortunately, we can still prove without using Fisher information that consistency and asymptotic normality of $\sigma^2\theta^{2\nu}$ holds in the case of estimated θ. This method is inspired by Kaufman and Shaby (2013) (see Theorem 7.4.3). First, note that for any $\theta > 0$, the MLE of $(\beta_1, \phi)^T$ has a closed form by setting derivatives of $L_\theta(\beta_1, \phi)$ in (7.30) to zero,

$$\hat{\beta}_{1,n} = (\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{y} \tag{7.35}$$

$$\hat{\phi}_n = \frac{\theta^{2\nu}}{n}(\boldsymbol{y} - \hat{\beta}_{1,n}\boldsymbol{x})^T\boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y} - \hat{\beta}_{1,n}\boldsymbol{x}) \tag{7.36}$$

Inserting the equation (7.35) into (7.36) we can show that the MLE of ϕ has an expression in terms of θ alone,

$$\hat{\phi}_n(\theta) = \frac{\theta^{2\nu}}{n}\boldsymbol{y}^T\left[\boldsymbol{\Sigma}^{-1}(\theta) - \boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x}(\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\theta)\right]\boldsymbol{y} \tag{7.37}$$

In the zero mean case (corresponding to $\boldsymbol{x} = \boldsymbol{0}$), Kaufman and Shaby (2013) showed that the function $\hat{\phi}_n(\theta)$ is monotonic with respect to θ. With some extra work, we show that the quadratic form in $\boldsymbol{y}$ given in (7.37) is also monotonic in θ. First we state some preliminary lemmas.

Lemma 7.4.5. *Let $\boldsymbol{A}, \boldsymbol{B}$ be real symmetric positive definite $n \times n$ matrices. If*

$\boldsymbol{A} - \boldsymbol{B}$ *is positive semidefinite, then* $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$ *is positive semidefinite.*

Lemma 7.4.6. *Consider the real symmetric* $(m+n) \times (m+n)$ *block matrix,*

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} \\ \boldsymbol{M}_{12}^T & \boldsymbol{M}_{22} \end{pmatrix}$$

where $\boldsymbol{M}_{22}$ *a positive definite* $n \times n$ *matrix. Then* $\boldsymbol{M}$ *is positive definite if and only if* $\boldsymbol{M}_{11}$ *and the Schur complement* $\boldsymbol{M}_{11} - \boldsymbol{M}_{12}\boldsymbol{M}_{22}^{-1}\boldsymbol{M}_{12}^T$ *are positive definite.*

Lemmas 7.4.5 and 7.4.6 can be found in Corollary 7.7.4 and Theorem 7.7.7 of Horn and Johnson (2013) respectively. These results together imply the following.

Proposition 7.4.2. *Let* $\boldsymbol{A}, \boldsymbol{B}$ *be real symmetric positive definite* $n \times n$ *matrices and suppose that* $\boldsymbol{A} - \boldsymbol{B}$ *is positive semidefinite. Then for any real* $n \times m$ *matrix* $\boldsymbol{C}$ *of full rank, the difference*

$$\left[\boldsymbol{A} - \boldsymbol{A}\boldsymbol{C}(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{A}\right] - \left[\boldsymbol{B} - \boldsymbol{B}\boldsymbol{C}(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{B}\boldsymbol{C})^{-1}\boldsymbol{C}^T\boldsymbol{B}\right]$$

is also positive semidefinite.

The proof can be found in Section 7.5. Finally, we state the main result of Kaufman and Shaby (2013). It can be found in Lemma 1 of their paper.

Lemma 7.4.7. *Let* $0 < \theta_1 < \theta_2$ *and suppose* $\boldsymbol{\Sigma}(\theta_2), \boldsymbol{\Sigma}(\theta_1)$ *are the corresponding symmetric positive definite Matérn* $n \times n$ *correlation matrices. Then the difference* $\theta_1^{-2\nu}\boldsymbol{\Sigma}(\theta_1) - \theta_2^{-2\nu}\boldsymbol{\Sigma}(\theta_2)$ *is positive semidefinite and by Lemma 7.4.5, the difference* $\theta_2^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_2) - \theta_1^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_1)$ *is also positive semidefinite.*

With these technical lemmas on hand, we are now ready to prove that $\hat{\phi}_n(\theta)$ defined in (7.37) is monotone in θ.

Proposition 7.4.3. *Let $\hat{\phi}_n(\theta)$ be defined as in (7.37). Then $\hat{\phi}_n(\theta)$ is a nondecreasing function of $\theta > 0$.*

Proof. Let $0 < \theta_1 < \theta_2$ and consider the difference $\hat{\phi}_n(\theta_2) - \hat{\phi}_n(\theta_1)$. This difference is nonnegative if and only if the matrix,

$$\begin{aligned} & \left[\theta_2^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_2) - \theta_2^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_2)\boldsymbol{x}(\boldsymbol{x}^T(\theta_2^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_2))\boldsymbol{x})^{-1}\boldsymbol{x}^T\theta_2^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_2)\right] \\ - & \left[\theta_1^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_1) - \theta_1^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_1)\boldsymbol{x}(\boldsymbol{x}^T(\theta_1^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_1))\boldsymbol{x})^{-1}\boldsymbol{x}^T\theta_1^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_1)\right] \end{aligned}$$

is positive semidefinite. By Lemma 7.4.7, the difference $\theta_2^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_2) - \theta_1^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_1)$ is positive semidefinite. Then, Proposition 7.4.2 applied to the above quantity with $\boldsymbol{A} = \theta_2^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_2), \boldsymbol{B} = \theta_1^{2\nu}\boldsymbol{\Sigma}^{-1}(\theta_1)$ and $\boldsymbol{C} = \boldsymbol{x}$ yields the result. □

Since $\hat{\phi}_n(\theta)$ is a monotone function of θ, we can now prove the consistency and asymptotic normality of $\hat{\phi}_n(\hat{\theta}_n) = \hat{\sigma}^2\hat{\theta}_n^{2\nu}$ where $\hat{\theta}_n$ is any sequence of bounded estimators of θ.

Proposition 7.4.4. *Let $[\theta_1, \theta_2] \subset (0, \infty)$ be some compact interval containing the true parameter θ_0. Suppose that $\hat{\theta}_n$ is any sequence of estimators that lie in $[\theta_1, \theta_2]$ for all n. Then, under $\mathbb{P}_0$,*

$$\begin{aligned} \hat{\phi}_n(\hat{\theta}_n) &\xrightarrow{P} \phi_0 \\ \sqrt{n}\left(\hat{\phi}_n(\hat{\theta}_n) - \phi_0\right) &\xrightarrow{D} N\left(0, 2\phi_0^2\right) \end{aligned}$$

Proof. For consistency, by Proposition 7.4.1, we have pointwise convergence $\hat{\phi}_n(\theta) \xrightarrow{P} \phi_0$ for any $\theta \in [\theta_1, \theta_2]$. But then by Proposition 7.4.3, since $\hat{\phi}_n(\theta)$ is monotone in θ, this implies that the convergence is uniform on $[\theta_1, \theta_2]$. Thus, for any sequence of estimators $\hat{\theta}_n \in [\theta_1, \theta_2]$,

$$|\hat{\phi}_n(\hat{\theta}_n) - \phi_0| \leq \sup_{\theta \in [\theta_1, \theta_2]} |\hat{\phi}_n(\theta) - \phi_0| \xrightarrow{P} 0$$

For asymptotic normality, first denote $\Phi(t)$ as the CDF of a standard Gaussian random variable. Then by Proposition 7.4.1, we have pointwise convergence,

$$\mathbb{P}(\sqrt{n}(\hat{\phi}_n(\theta) - \phi_0) \leq t) \xrightarrow{P} \Phi\left(\frac{t}{\sqrt{2\phi_0^2}}\right) \tag{7.38}$$

for any $\theta \in [\theta_1, \theta_2]$ and $t \in \mathbb{R}$. By Proposition 7.4.3, since $\hat{\phi}_n(\theta)$ is monotone in θ, the sequence of distribution functions $\mathbb{P}(\sqrt{n}(\hat{\phi}_n(\theta) - \phi_0) \leq t)$ is also monotone in θ and so the convergence in (7.38) is uniform on $[\theta_1, \theta_2]$. So for any sequence $\hat{\theta}_n \in [\theta_1, \theta_2]$ and $t \in \mathbb{R}$,

$$\begin{aligned} &\left| \mathbb{P}(\sqrt{n}(\hat{\phi}_n(\hat{\theta}_n) - \phi_0) \leq t) - \Phi\left(\frac{t}{\sqrt{2\phi_0^2}}\right) \right| \\ \leq &\sup_{\theta \in [\theta_1, \theta_2]} \left| \mathbb{P}(\sqrt{n}(\hat{\phi}_n(\theta) - \phi_0) \leq t) - \Phi\left(\frac{t}{\sqrt{2\phi_0^2}}\right) \right| \xrightarrow{P} 0 \end{aligned}$$

□

Our simulation studies in Chapter 8 show that estimating θ results in less bias in the MLE of $\phi = \sigma^2\theta^{2\nu}$ as opposed to fixing θ to some arbitrary value. This is

similar to what Kaufman and Shaby (2013) concluded in the zero mean case. Thus, it might be preferable to use an estimator $\hat{\theta}_n$ in small samples even if it cannot be a consistent estimator. For the MLE of the slope parameter β_1, we do not have a similar proof in the case of estimated θ. For any $\theta > 0$, the MLE of β_1 has the form,

$$\hat{\beta}_{1,n}(\theta) = \arg\min_{\beta_1} (\boldsymbol{y} - \beta_1 \boldsymbol{x})^T \boldsymbol{\Sigma}^{-1}(\theta)(\boldsymbol{y} - \beta_1 \boldsymbol{x}) = \frac{\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}(\theta) \boldsymbol{y}}{\boldsymbol{x}^T \boldsymbol{\Sigma}^{-1}(\theta) \boldsymbol{x}}$$

Unlike the parameter $\hat{\phi}_n(\theta)$, this is not a monotone function in θ and thus a uniform convergence result like above cannot be readily established. However, based on our simulation study in Chapter 8, we conjecture that consistency and asymptotic normality still hold with any sequence $\hat{\theta}_n$ serving as a plug-in estimator.

In our simulations, we also consider the case of multiple covariates in the regression model (7.2). In light of the previous discussion involving one covariate, we expect that similar results as Propositions 7.4.1 and 7.4.4 to hold when there are other covariates in the model. Without loss of generality, we can partition the covariates in terms of their smoothness. Specifically, let $x_k(\boldsymbol{s}), k = 1, ..., m$ be covariates with rough sample paths not in $W_2^\ell(D)$ and let $x_k(\boldsymbol{s}), k = m+1, .., p$ be the covariates with smooth sample paths in $W_2^\ell(D)$. Then, we expect the regression coefficients $(\beta_1, ..., \beta_m)^T$ to be microergodic. Denoting $\hat{\boldsymbol{\beta}}_n$ as the MLE of $(\beta_1, ..., \beta_m)^T$ and $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_m \end{bmatrix}_{n \times m}$ as the design matrix, we expect $(\hat{\boldsymbol{\beta}}^T, \hat{\sigma}^2 \hat{\theta}_n^{2\nu})^T$ to be consistent and,

$$\begin{pmatrix} \boldsymbol{V}^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ \sqrt{n}(\hat{\sigma}_n^2\hat{\theta}_n^{2\nu} - \sigma_0^2\theta_0^{2\nu}) \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{I}_m & 0 \\ 0 & 2(\sigma_0^2\theta_0^{2\nu})^2 \end{pmatrix} \right) \tag{7.39}$$

where $\boldsymbol{V} = \frac{1}{\sigma_0^2}\boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}(\theta_0)\boldsymbol{X}$. The above proofs can be adapted by replacing the vector $\boldsymbol{x}$ with the design matrix $\boldsymbol{X}$.

For regression models, results in spatial statistics literature concerning estimation of the regression coefficients are scarce. There is such a result in the previously mentioned paper by Ying (1991). In addition to the analysis of the zero mean Ornstein-Uhlenbeck process on $[0, 1]$, Ying considers the spatial regression model

$$y(t) = \boldsymbol{x}(t)^T\boldsymbol{\beta} + e(t), \quad t \in [0, 1]$$

where $\boldsymbol{x}(t) = (x_1(t), ..., x_p(t))$ is a deterministic multivariate function and $e(t)$ is a OU process on $[0, 1]$. Ying showed that the MLE $\hat{\sigma}^2\hat{\theta}_n$ is consistent and asymptotically normal just as in Theorem 7.0.1 without a regression term. Under certain smoothness conditions on $\boldsymbol{x}(t)$, akin to each component $x_i(t)$ being in the Sobolev space $W_2^1([0, 1])$, Ying showed that the MLE $\hat{\boldsymbol{\beta}}_n$ is asymptotically normal. However, a consistency result is not given since the smoothness conditions on $\boldsymbol{x}(t)$ prevent $\boldsymbol{\beta}$ from being microergodic. The proof technique by Ying exploited the Markov property of the OU process on $[0, 1]$, giving a more useable form of the likelihood function for $(\boldsymbol{\beta}^T, \sigma^2\theta)^T$. Evidently, this technique does not work for higher spatial dimensions or covariance functions other than the exponential. Results on regression

models in these settings are not available to our knowledge.

7.5 Proofs of results

Proof of Lemma 7.2.3

Proof. First consider the case $k = 1$. By Lemma 7.2.2, the first term of $m(\boldsymbol{\omega})$ in (7.21) has its Fourier transform calculated as,

$$\int_{\mathbb{R}^d} e^{i\boldsymbol{\omega}^T \boldsymbol{s}} \mathbb{1}_{||\boldsymbol{s}||\leq 1}(\boldsymbol{s}) d\boldsymbol{s} = \frac{c}{||\boldsymbol{\omega}||^{\frac{d-2}{2}}} \int_0^1 J_{\frac{d-2}{2}}(r||\boldsymbol{\omega}||) r^{\frac{d}{2}} dr = \frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}}} J_{\frac{d}{2}}(||\boldsymbol{\omega}||) \quad (7.40)$$

The second equality comes from the relation $\int_0^1 z^{\alpha+1} J_\alpha(\tau z) dz = \frac{1}{\tau} J_{\alpha+1}(x)$ (see Yadrenko (1983), p. 13). Now, we continue the calculation for second term in (7.21) using Lemma 7.2.2,

$$\int_{1<||\boldsymbol{s}||<\infty} g(||\boldsymbol{s}||) e^{i\boldsymbol{\omega}^T \boldsymbol{s}} d\boldsymbol{s} = \frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}-1}} \int_1^\infty g(r) J_{\frac{d-2}{2}}(r||\boldsymbol{\omega}||) r^{\frac{d}{2}} dr$$

Using the identity $\frac{d}{dz}(z^\alpha J_\alpha(z)) = z^\alpha J_{\alpha-1}(z)$ (Watson (1995), p. 45), we integrate by parts,

$$\int_{1<||\boldsymbol{s}||<\infty} g(||\boldsymbol{s}||) e^{i\boldsymbol{\omega}^T \boldsymbol{s}} d\boldsymbol{s} = \frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}}} \left[g(r) J_{\frac{d}{2}}(r||\boldsymbol{\omega}||) r^{\frac{d}{2}} \Big|_1^\infty - \int_1^\infty g'(r) J_{\frac{d}{2}}(r||\boldsymbol{\omega}||) r^{\frac{d}{2}} dr \right]$$

Since $g(1) = 1$ and $\lim_{r\to\infty} g(r) r^{\frac{d}{2}} = 0$, the first term above equals $-\frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}}} J_{\frac{d}{2}}(||\boldsymbol{\omega}||)$, thereby cancelling with the first term of the Fourier transform calculation in (7.40).

Thus, we are left with the integral,

$$\hat{m}(\boldsymbol{\omega}) = -\frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}}}\int_1^\infty g'(r)J_{\frac{d}{2}}(r||\boldsymbol{\omega}||)r^{\frac{d}{2}}dr = -\frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}}}\int_1^\infty \frac{g'(r)}{r}J_{\frac{d}{2}}(r||\boldsymbol{\omega}||)r^{\frac{d}{2}+1}dr$$

proving that (7.22) holds for $k = 1$. Now suppose it holds for some $k \in \{1, .., N-1\}$,

$$\hat{m}(\boldsymbol{\omega}) = (-1)^k\frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}+k-1}}\int_1^\infty \frac{\sum_{j=1}^{k} a_j g^{(j)}(r)r^{j-1}}{r^{2k-1}}J_{\frac{d}{2}+k-1}(r||\boldsymbol{\omega}||)r^{\frac{d}{2}+k}dr$$

Performing an integration by parts,

$$\hat{m}(\boldsymbol{\omega}) = (-1)^k\frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}+k}}\left[\frac{\sum_{j=1}^{k} a_j g^{(j)}(r)r^{j-1}}{r^{2k-1}}J_{\frac{d}{2}+k}(r||\boldsymbol{\omega}||)r^{\frac{d}{2}+k}\right|_1^\infty$$

$$-\int_1^\infty \frac{\sum_{j=1}^{k+1} b_j g^{(j)}(r)r^{j-1}}{r^{2k+1}}J_{\frac{d}{2}+k}(r||\boldsymbol{\omega}||)r^{\frac{d}{2}+k+1}dr\Bigg]$$

where $b_j \in \mathbb{Z}$. Specifically,

$$\begin{aligned} b_1 &= (1-2k)a_1 \\ b_j &= (j-2k)a_j + a_{j-1},\ j = 2, \cdots, k \\ b_{k+1} &= a_k = 1 \end{aligned}$$

Recalling the chosen form of $g(r) = e^{-(r-1)^{N+1}}$, we can see that the first set of terms vanishes since $\lim_{r\to 1} g^{(j)}(r) = 0$ and $\lim_{r\to\infty} g^{(j)}(r)r^m = 0$ for any power $m > 0$ and all

$j = 1, ..., N-1$. Thus, we are left with the integral,

$$\hat{m}(\boldsymbol{\omega}) = (-1)^{k+1} \frac{c}{||\boldsymbol{\omega}||^{\frac{d}{2}+k}} \left[\int_1^\infty \frac{\sum_{j=1}^{k+1} b_j g^{(j)}(r) r^{j-1}}{r^{2k+1}} J_{\frac{d}{2}+k}(r||\boldsymbol{\omega}||) r^{\frac{d}{2}+k+1} dr \right]$$

Thus, the result holds for $k+1$ and so by induction on the set $\{1, .., N\}$, the result follows. □

Proof of Proposition 7.4.2

Proof. Consider the $2n \times 2n$ symmetric block matrix,

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A} - \boldsymbol{B} & (\boldsymbol{A} - \boldsymbol{B})\boldsymbol{C} \\ \boldsymbol{C}^T(\boldsymbol{A} - \boldsymbol{B}) & \boldsymbol{C}^T(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{C} \end{pmatrix}$$

Since $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite, there exists a matrix $\boldsymbol{R}$ such that $\boldsymbol{A} - \boldsymbol{B} = \boldsymbol{R}^T\boldsymbol{R}$. Then $\boldsymbol{M}$ must be positive semidefinite because it has a decomposition,

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{R}^T \\ \boldsymbol{C}^T\boldsymbol{R}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{R} & \boldsymbol{R}\boldsymbol{C} \end{pmatrix}$$

But since $\boldsymbol{M}$ can also be written as a difference of two real symmetric matrices,

$$\boldsymbol{M} = \underbrace{\begin{pmatrix} \boldsymbol{A} & \boldsymbol{A}\boldsymbol{C} \\ \boldsymbol{C}^T\boldsymbol{A} & \boldsymbol{C}^T\boldsymbol{A}\boldsymbol{C} \end{pmatrix}}_{\boldsymbol{M}_A} - \underbrace{\begin{pmatrix} \boldsymbol{B} & \boldsymbol{B}\boldsymbol{C} \\ \boldsymbol{C}^T\boldsymbol{B} & \boldsymbol{C}^T\boldsymbol{B}\boldsymbol{C} \end{pmatrix}}_{\boldsymbol{M}_B}$$

the difference $\boldsymbol{M}_A - \boldsymbol{M}_B$ must also be positive semidefinite. Note that we cannot use Lemma 7.4.5 yet since $\boldsymbol{M}_A$ and $\boldsymbol{M}_B$ are only positive semidefinite and their inverses do not exist. To circumvent this issue, for any $\epsilon > 0$, consider a perturbation $\boldsymbol{M}_A(\epsilon) = \boldsymbol{M}_A + \epsilon \boldsymbol{I}_{2n}$ where $\boldsymbol{I}_{2n}$ is the $2n \times 2n$ identity matrix (and similarly define $\boldsymbol{M}_B(\epsilon)$). Now $\boldsymbol{M}_A(\epsilon)$ and $\boldsymbol{M}_B(\epsilon)$ are both positive definite and their difference $\boldsymbol{M}_A(\epsilon) - \boldsymbol{M}_B(\epsilon)$ is positive semidefinite. By Lemma 7.4.5, this implies that $\boldsymbol{M}_B^{-1}(\epsilon) - \boldsymbol{M}_A^{-1}(\epsilon)$ is positive semidefinite. Since positive semidefiniteness is preserved for all principal submatrices of $\boldsymbol{M}_B^{-1}(\epsilon) - \boldsymbol{M}_A^{-1}(\epsilon)$, it follows after using block matrix inversion formulas (Horn and Johnson (2013), p. 18) that the principal submatrix,

$$\begin{aligned} & \left[\boldsymbol{B} + \epsilon \boldsymbol{I}_n - \boldsymbol{B}\boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{B}\boldsymbol{C} + \epsilon \boldsymbol{I}_n)^{-1}\boldsymbol{C}^T\boldsymbol{B}\right]^{-1} \\ - & \left[\boldsymbol{A} + \epsilon \boldsymbol{I}_n - \boldsymbol{A}\boldsymbol{C}(\boldsymbol{C}^T\boldsymbol{A}\boldsymbol{C} + \epsilon \boldsymbol{I}_n)^{-1}\boldsymbol{C}^T\boldsymbol{A}\right]^{-1} \end{aligned}$$

is also positive semidefinite. By Lemma 7.4.6, both inverse matrices are well defined since the matrices being inverted are positive definite. Finally applying Lemma 7.4.5 again, we get that,

$$\left[\boldsymbol{A} - \boldsymbol{A}\boldsymbol{C}(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{A}\boldsymbol{C} + \epsilon \boldsymbol{I}_n)^{-1}\boldsymbol{C}^T\boldsymbol{A}\right] - \left[\boldsymbol{B} - \boldsymbol{B}\boldsymbol{C}(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{B}\boldsymbol{C} + \epsilon \boldsymbol{I}_n)^{-1}\boldsymbol{C}^T\boldsymbol{B}\right]$$

is positive semidefinite. Since this is true for any $\epsilon > 0$, the result holds. □

Appendix: Calculations for the Ornstein-Uhlenbeck process

Suppose $y(t)$ is a stationary Gaussian process on $[0,1]$ with exponential covariance function $C(h) = \sigma^2 e^{-\theta h}$. Let $t_1 < .... < t_n$ be arbitrary sampling locations in $[0,1]$ such that the empirical measure of $\{t_1, \cdots, t_n\}$ converges to the Lebesgue measure on $[0,1]$ as $n \to \infty$. Then, the random vector $\boldsymbol{y} = (y(t_1), ..., y(t_n))^T$ is multivariate normal with covariance matrix, $\{\Sigma\}_{i,j} = \sigma^2 e^{-\theta(t_j - t_i)}, 1 \leq i \leq j \leq n$. By exploiting the Markov property of this process, $z_1 = y_1$ and $z_k = y_k - e^{-\theta(t_k - t_{k-1})} y_{k-1}, k = 2, ..., n$ are independent normal random variables. In matrix notation, this transformation can be written as,

$$\underbrace{\begin{pmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_n \end{pmatrix}}_{\boldsymbol{z}} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -e^{-\theta(t_2-t_1)} & 1 & 0 & 0 & \cdots & 0 \\ 0 & -e^{-\theta(t_3-t_2)} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & -e^{-\theta(t_n-t_{n-1})} & 1 \end{pmatrix}}_{\boldsymbol{L}} \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}}_{\boldsymbol{y}}$$

where $\boldsymbol{L}$ is a lower bi-diagonal matrix with 1's on the diagonal and $e^{-t_k - t_{k-1}}, k = 2, ..., n$ on the off-diagonal. Since $\boldsymbol{Ly}$ is an affine transformation of a Gaussian vector, the matrix $\boldsymbol{L}$ defined above satisfies the property, $\boldsymbol{L\Sigma L}^T = \sigma^2 \boldsymbol{D}$, where $\boldsymbol{D} = \text{diag}(1, 1 - e^{-\theta(t_2 - t_1)}, ..., 1 - e^{-\theta(t_n - t_{n-1})})$ is the covariance matrix of $\boldsymbol{z}$. Then, the formula for the inverse of $\boldsymbol{\Sigma}$ can be found by $\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2} \boldsymbol{L}^T \boldsymbol{D}^{-1} \boldsymbol{L}$. The matrix product $\boldsymbol{L}^T \boldsymbol{D}^{-1} \boldsymbol{L}$ ends up being tri-diagonal. Letting $\Delta_{k,k-1} = t_k - t_{k-1}$, the

diagonal elements of this tri-diagonal matrix are,

$$(\boldsymbol{L}^T\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}_{k,k} = \begin{cases} \frac{1}{1-e^{-2\theta\Delta_{2,1}}} & k=1 \\ \frac{1}{1-e^{-2\theta\Delta_{n,n-1}}} & k=n \\ \frac{1}{1-e^{-2\theta\Delta_{k-1,k-2}}} + \frac{e^{-2\theta\Delta_{k,k-1}}}{1-e^{-2\theta\Delta_{k,k-1}}} & k=3,...,n \end{cases}$$

The off-diagonal entries are $(\boldsymbol{L}^T\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}_{k,k-1} = (\boldsymbol{L}^T\boldsymbol{D}^{-1}\boldsymbol{L})^{-1}_{k-1,k} = -\frac{e^{-\theta\Delta_{k,k-1}}}{1-e^{-2\theta\Delta_{k,k-1}}}, k = 2,...,n$. Summing the entries of $\boldsymbol{\Sigma}^{-1}$ together and simplifying yields,

$$\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1} = \frac{1}{\sigma^2}\left[1 + \sum_{k=2}^{n}\frac{1-e^{-\theta(t_k-t_{k-1})}}{1+e^{-\theta(t_k-t_{k-1})}}\right]$$

By a Taylor expansion argument, we have the bounds $\frac{x}{2} - \frac{x^3}{24} \leq \frac{1-e^{-x}}{1+e^{-x}} \leq \frac{x}{2}$ for all $x \geq 0$. Then, an upper bound for the partial sum above is,

$$\sum_{k=2}^{n}\frac{1-e^{-\theta(t_k-t_{k-1})}}{1+e^{-\theta(t_k-t_{k-1})}} \leq \frac{\theta}{2}\sum_{k=2}^{n}(t_k - t_{k-1}) = \frac{\theta}{2}(t_n - t_1)$$

Similarly, a lower bound for the above partial sum is,

$$\sum_{k=2}^{n}\frac{1-e^{-\theta(t_k-t_{k-1})}}{1+e^{-\theta(t_k-t_{k-1})}} \geq \frac{\theta}{2}(t_n - t_1) - \frac{\theta^3}{24}\sum_{k=2}^{n}(t_k - t_{k-1})^3$$

By the assumptions on $\{t_1,..,t_n\}$, there exists an M_n such that $t_k - t_{k-1} \leq M_n$ uniformly in k and $M_n \to 0$. Applying this to the lower bound, we obtain,

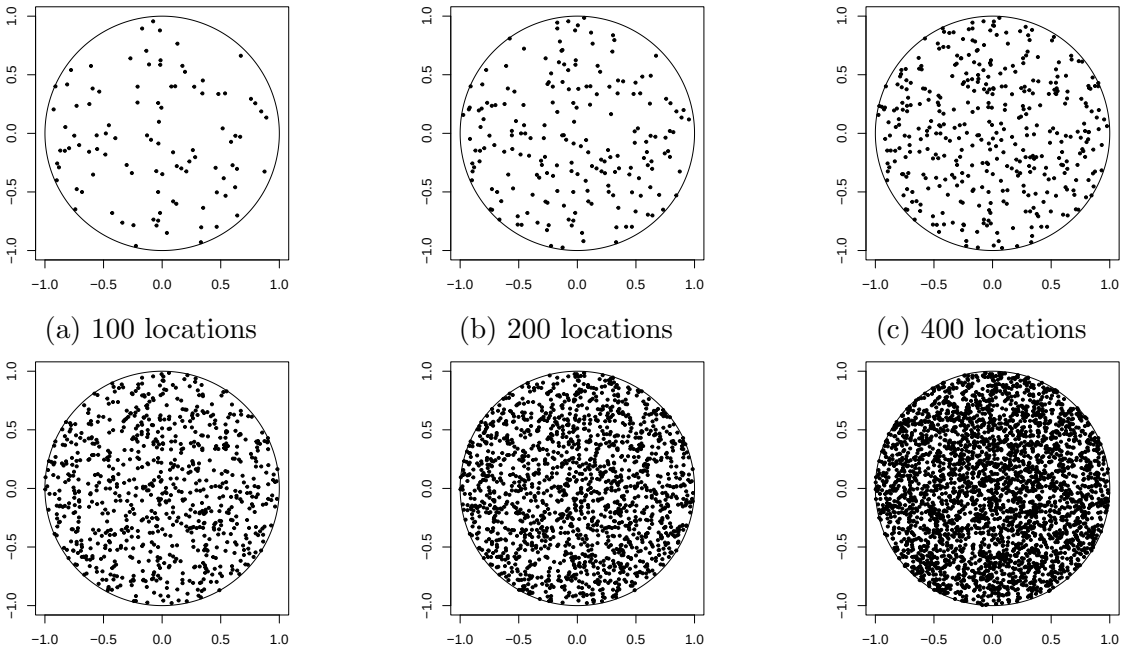
$$\frac{\theta}{2}(t_n - t_1) - \frac{\theta^3 M_n^2}{24}(t_n - t_1) \leq \sum_{k=2}^{n} \frac{1 - e^{-\theta(t_k - t_{k-1})}}{1 + e^{-\theta(t_k - t_{k-1})}} \leq \frac{\theta}{2}(t_n - t_1)$$

Since $t_n - t_1 \to 1$, the limit equals $\lim_{n\to\infty} \sum_{k=2}^{n} \frac{1 - e^{-\theta(t_k - t_{k-1})}}{1 + e^{-\theta(t_k - t_{k-1})}} = \frac{\theta}{2}$. Thus, we see that the sum of the inverse matrix elements equals,

$$\lim_{n\to\infty} \mathbf{1}^T \boldsymbol{\Sigma}^{-1} \mathbf{1} = \frac{1}{\sigma^2} \left[1 + \lim_{n\to\infty} \sum_{k=2}^{n} \frac{1 - e^{-\theta(t_k - t_{k-1})}}{1 + e^{-\theta(t_k - t_{k-1})}} \right] = \frac{1}{\sigma^2} \left[1 + \frac{\theta}{2} \right] = \frac{2 + \theta}{2\sigma^2} < \infty$$

Chapter 8 Numerical study on infill asymptotics

In Chapter 7, we gave results and conjectures pertaining to the microergodicity and estimation of regression coefficients and variance parameters under infill asymptotics. In this chapter, we perform numerical simulations to illustrate the assertions and support the conjectures. We take as our sampling domain the unit disk in $\mathbb{R}^2$ (see Figure 8.1). To simulate an infill asymptotics framework, we generate a



(a) 100 locations (b) 200 locations (c) 400 locations

(d) 800 locations (e) 1600 locations (f) 3200 locations

Figure 8.1: Infill asymptotics asymptotics framework

nested sequence of locations uniform in the unit disk and consider sample sizes of

$n = 100, 200, 400, 800, 1600, 3200$.

8.1 The behavior of OLS estimates

In this section, we investigate the behavior of OLS estimates under infill asymptotics. Consider the simple linear regression model,

$$y(\boldsymbol{s}) = \beta_0 + \beta_1 x(\boldsymbol{s}) + e(\boldsymbol{s}), \quad \boldsymbol{s} \in D \tag{8.1}$$

where $D \subset \mathbb{R}^d$ is compact. We assume that $x(\boldsymbol{s})$ and $e(\boldsymbol{s})$ are independent Gaussian random fields, each with their own Matérn covariances parametrized by $(\sigma_x^2, \theta_x, \nu_x)^T$ and $(\sigma^2, \theta, \nu)^T$ respectively. In Chapter 7, we showed that the OLS estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ remains inconsistent regardless of the smoothness of the covariate. Here we give a simulation study to demonstrate this. For $e(\boldsymbol{s})$, we arbitrarily choose Matérn covariance parameters $(\sigma^2, \theta, \nu)^T = (1, 1, 1)^T$. For the $x(\boldsymbol{s})$ Matérn covariance parameters, we choose $(\sigma_x^2, \theta_x, \nu_x)^T = (1, 1, \frac{1}{2})^T$ and generate one zero mean Gaussian random vector $\boldsymbol{x}$ according to these values. We note that $\nu_x = \frac{1}{2}$ was chosen to be less than the "critical" smoothness $\ell = 2$. To make the observations nested, we generate $\boldsymbol{x}$ for the largest set of locations ($n = 3200$) and consider subsets $(x(\boldsymbol{s}_1), ..., x(\boldsymbol{s}_n))^T$ for $n = 100, 200, 400, 800, 1600, 3200$. Consider the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$, where $\boldsymbol{X} = \begin{bmatrix}\mathbf{1} & \boldsymbol{x}\end{bmatrix}_{n\times 2}$. Conditional on $\boldsymbol{x}$, the variance is,

$$\text{Var}(\hat{\boldsymbol{\beta}}_{OLS}|\boldsymbol{x}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Sigma}(\theta)\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \tag{8.2}$$

where $\boldsymbol{\Sigma}(\theta)$ is the covariance matrix of $e(\boldsymbol{s})$. For each sample size, we calculate the variances in (8.2). The results are given in Table 8.1.

n	β_0	β_1
100	0.546877	0.110918
200	0.553042	0.098621
400	0.553196	0.104220
800	0.551528	0.103709
1600	0.556859	0.107505
3200	0.559162	0.106006

Table 8.1: Variances for the OLS estimator of $\boldsymbol{\beta}$. For each n, values were computed using (8.2) and taking nested subsets of values from the single simulated Gaussian vector $\boldsymbol{x} = (x(\boldsymbol{s}_1), \cdots, x(\boldsymbol{s}_{3200}))^T$.

As expected, the variances of the OLS estimate for β_0 show no decay because we proved that this parameter is non-microergodic (Proposition 7.3.1). The variances of the OLS estimate for β_1 also show little to no decay despite choosing $x(\boldsymbol{s})$ to have rough sample paths compared to $e(\boldsymbol{s})$. This is consistent with Proposition 7.3.3.

8.2 The effect of smoothness on Fisher information

In this section, we investigate the effect of the smoothness of the covariate on the behavior of the Fisher information for the slope β_1. We consider the same linear regression model in (8.1). For $e(\boldsymbol{s})$, we arbitrarily choose Matérn covariance parameters $(\sigma^2, \theta, \nu)^T = (1, 1, 1)^T$. For the $x(\boldsymbol{s})$ Matérn covariance parameters, we arbitrarily choose $(\sigma_x^2, \theta_x)^T = (1, 1)^T$. For the smoothness parameter, we consider different values of $\nu_x \in \left\{1, \frac{3}{2}, 2, \frac{5}{2}, 3\right\}$. We generate five Gaussian random vectors $\boldsymbol{x}_i, i = 1, 2, 3, 4, 5$ according to these smoothness parameters and calculate the in-

verse Fisher information,

$$\frac{1}{\boldsymbol{x}_i^T \boldsymbol{\Sigma}(\theta)^{-1} \boldsymbol{x}_i}, \quad i = 1, 2, 3, 4, 5 \tag{8.3}$$

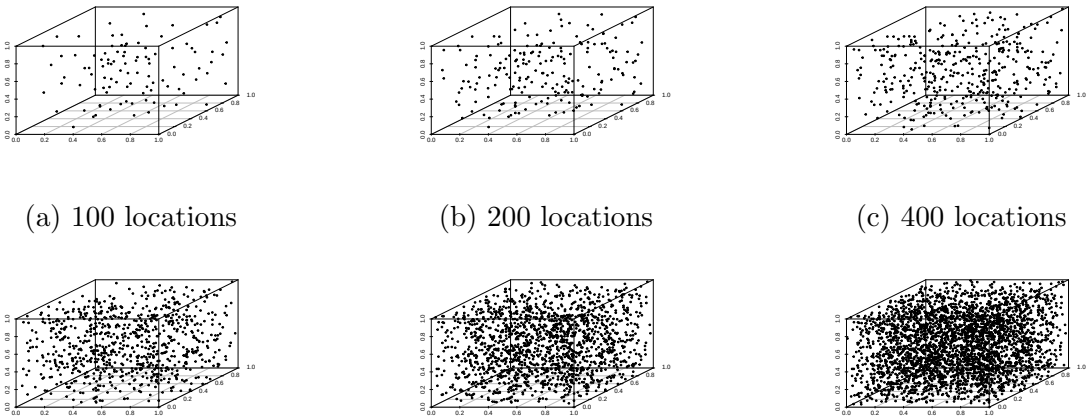
in each case.

	Smoothness ν_x				
n	1	$\frac{3}{2}$	2	$\frac{5}{2}$	3
100	0.008846	0.034859	0.040558	0.286650	0.179624
200	0.004848	0.026896	0.038488	0.253762	0.171720
400	0.002657	0.020257	0.036712	0.243199	0.167509
800	0.001324	0.015646	0.035217	0.236432	0.165679
1600	0.000666	0.011749	0.034057	0.232442	0.164278
3200	0.000318	0.008797	0.032975	0.229535	0.163516

Table 8.2: Inverse Fisher information for β_1 in the case $d = 2$ for different ν_x values. For each n, values were computed using (8.3) and taking nested subsets of values from the single simulated Gaussian vector $\boldsymbol{x}_i = (x_i(\boldsymbol{s}_1), \cdots, x_i(\boldsymbol{s}_{3200}))^T$. Each $\boldsymbol{x}_i, i = 1, 2, 3, 4, 5$ corresponds to a different ν_x.

In Proposition 7.3.2, we showed that the term $\ell = 2$ acts as a "critical" smoothness parameter. Any value ν_x greater than this quantity implies that the slope β_1 is non-microergodic. As shown in Table 8.2, when $\nu_x = 1$, we get expected decay of the inverse Fisher information as the number of observations doubles. In the case $\nu_x = \frac{3}{2}$, there is also decay albeit at a slower rate. Once we reach the critical smoothness value of $\nu_x = 2$ and above, the decay seems to slow down. This is especially true in the case $\nu_x = 3$, as there does not seem to be any noticeable decay at all, corroborating statements made in Chapter 7.

We repeat this simulation in the case $d = 3$ by generating points uniformly in the unit cube $[0, 1]^3 \subset \mathbb{R}^3$ (Figure 8.2).



(a) 100 locations (b) 200 locations (c) 400 locations

(d) 800 locations (e) 1600 locations (f) 3200 locations

Figure 8.2: Infill asymptotics asymptotics in the unit cube in $\mathbb{R}^3$

We use the same Matérn covariance parameters for $e(\boldsymbol{s})$, specifically $(\sigma^2, \theta, \nu)^T = (1, 1, 1)^T$. We consider covariate smoothness parameters $\nu_x \in \left\{\frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \frac{7}{2}, \frac{9}{2}\right\}$. We generate five Gaussian random vectors $\boldsymbol{x}_i, i = 1, 2, 3, 4, 5$ according to these smoothness parameters and calculate the inverse Fisher information in (8.3) in each case.

	Smoothness ν_x				
n	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$	$\frac{9}{2}$
100	0.002068	0.026120	0.058248	0.074221	0.052362
200	0.000830	0.017631	0.049612	0.067673	0.048607
400	0.000331	0.011576	0.042790	0.064927	0.047015
800	0.000121	0.007412	0.038407	0.061629	0.043801
1600	0.000049	0.004737	0.034617	0.059126	0.042189
3200	0.000019	0.003112	0.032106	0.057852	0.041526

Table 8.3: Inverse Fisher information for β_1 in the case $d = 3$ for different ν_x values. For each n, values were computed using (8.3) and taking nested subsets of values from the single simulated Gaussian vector $\boldsymbol{x}_i = (x_i(\boldsymbol{s}_1), \cdots, x_i(\boldsymbol{s}_{3200}))^T$. Each $\boldsymbol{x}_i, i = 1, 2, 3, 4, 5$ corresponds to a different ν_x.

In this case, the critical smoothness is $\ell = \frac{5}{2}$. We can see from Table 8.3 that the inverse Fisher information values follow the same pattern as in the previous case

where $d = 2$. When ν_x is far less than ℓ, there is variance decay. When ν_x is close to or greater than ℓ, there is no noticeable decay. Note that the critical smoothness here $\ell = \frac{5}{2}$ is not an integer. Recall from Chapter 7, Proposition 7.3.2 that Scheuerer (2010) only gave a sufficient condition in the fractional ℓ case for $x(\boldsymbol{s}) \in W_2^{\ell}(D)$ a.s., namely $\nu_x > \ell$. This simulation might suggest it is also a necessary condition.

8.3 Behavior of maximum likelihood estimates

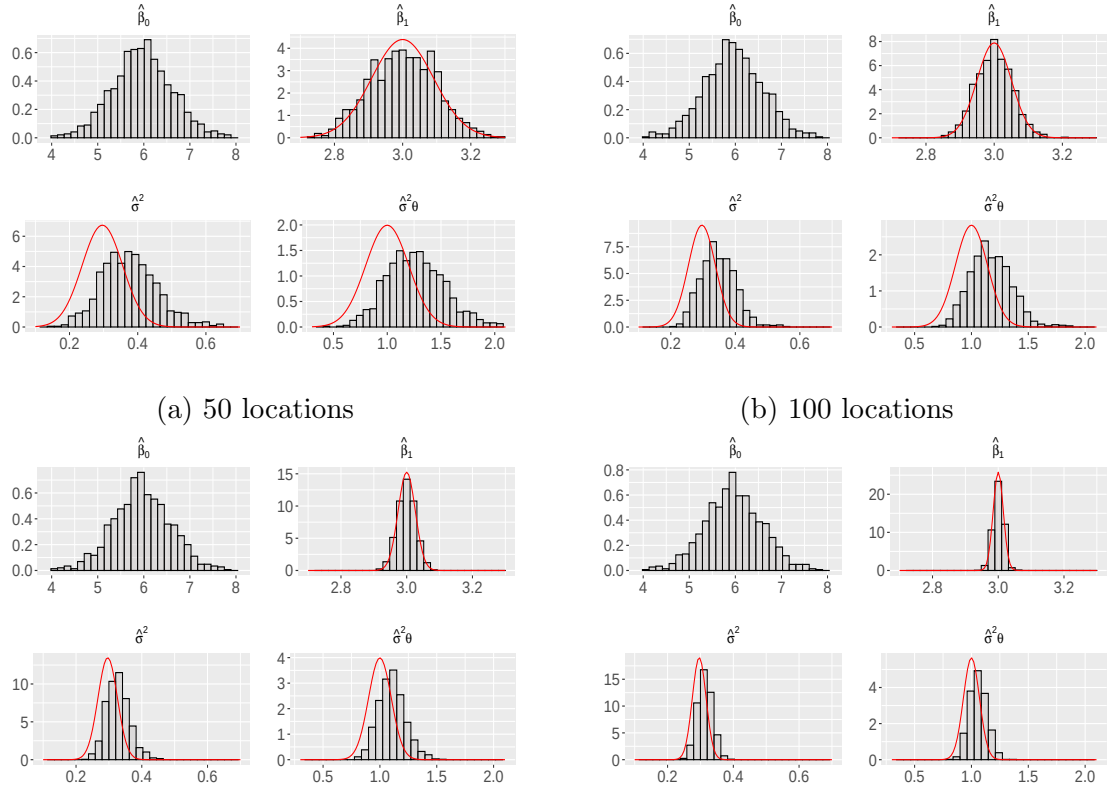
In this section, we perform a Monte Carlo simulation study of maximum likelihood estimation. In $\mathbb{R}^2$, we consider locations in the unit circle with sample sizes of $n = 50, 100, 200$ and 400 as in Figure 8.1. We arbitrarily choose as our true Matérn $e(\boldsymbol{s})$ covariance parameters $(\sigma^2, \theta, \nu)^T = (1, 1, \frac{3}{2})^T$. Thus, the critical smoothness in this case is $\ell = \frac{5}{2}$.

8.3.1 One covariate

We consider the same linear regression model in (8.1). For the regression parameters in (8.1), we arbitrarily choose $(\beta_0, \beta_1)^T = (6, 3)^T$. For the covariate, we choose as the true parameters $(\sigma_x^2, \theta_x)^T = (1, 1)^T$. We compare MLE estimation when the Matérn scale parameter θ is both fixed and estimated. We choose a Matérn smoothness parameter of $\nu_x = 1$ for the covariate and generate one Gaussian vector $\boldsymbol{x}$. Then, we generate 1000 realizations of $\boldsymbol{y}|\boldsymbol{x}$ according to the regression model (8.1).

8.3.1.1 Fixed θ

First, we fix $\theta = \frac{3}{2}$ and compute MLE estimates of $(\beta_0, \beta_1, \sigma^2, \sigma^2\theta^{2\nu})^T$. The empirical distributions of the MLE estimates are given in Figure 8.3.



(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

Figure 8.3: Histograms of MLE estimates of $(\beta_0, \beta_1, \sigma^2, \sigma^2\theta^{2\nu})^T$ when θ is fixed to an incorrect value $\theta = \frac{3}{2}$ (true value is $\theta_0 = 1$). In red are the theoretical asymptotic densities predicted by Proposition 7.4.1. Estimates were computed from 1000 MC simulations of $\boldsymbol{y}|\boldsymbol{x}$ based on the regression model (8.1).

Since we chose $\nu_x < \ell$, the parameter β_1 is microergodic. This is reflected in the histograms of $\hat{\beta}_{1,n}$, where empirical distributions become more peaked as the number of observations doubles. The empirical distributions of MLE estimates for $(\sigma^2, \sigma^2\theta^{2\nu})^T$ also display this behavior. The empirical variances of the MLE estimates for all parameters are given in Table 8.4. We note that variances of the

estimates for β_0 show no decay, consistent with the theory that this parameter is non-microergodic.

n	β_0	β_1	σ^2	$\sigma^2\theta^{2\nu}$
50	0.423230	0.008741	0.006673	0.076014
100	0.399379	0.002474	0.002814	0.032054
200	0.392289	0.000674	0.001135	0.012924
400	0.386582	0.000246	0.000499	0.005682

Table 8.4: Empirical variances of MC estimates from Figure 8.3.

When comparing the histograms with their theoretical densities, it appears that the estimates for $(\sigma^2, \sigma^2\theta^{2\nu})^T$ are heavily biased in smaller sample sizes unlike the estimates for β_1. This bias becomes less pronounced when $n = 200$. Table 8.5 displays the empirical absolute biases of these estimates.

n	σ^2	$\sigma^2\theta^{2\nu}$
50	0.078094	0.263567
100	0.050493	0.170413
200	0.029595	0.099883
400	0.016403	0.055361

Table 8.5: Empirical absolute biases of MC estimates for $(\sigma^2, \sigma^2\theta^{2\nu})^T$ from Figure 8.3. These values were computed by calculating the mean absolute difference between the MC estimates and the true parameter values of $\sigma_0^2 = 1$ and $\sigma_0^2\theta_0^{2\nu} = 1$

We attempted this simulation again with different fixed values of θ. Figure 8.4 shows boxplots of the empirical distributions of the MLE of $\sigma^2\theta^{2\nu}$ for fixed values of $\theta \in \left\{\frac{1}{2}, \frac{3}{4}, 1, \frac{5}{4}, \frac{3}{2}\right\}$. The results indicate that for fixed values of θ further away from the true parameter $\theta = 1$, larger sample sizes are required to estimate $\sigma^2\theta^{2\nu}$ with less bias. This is consistent with the simulations performed by Kaufman and Shaby (2013), who showed that in the zero mean model, small sample bias occurs in the MLE of $\sigma^2\theta^{2\nu}$ by fixing θ.

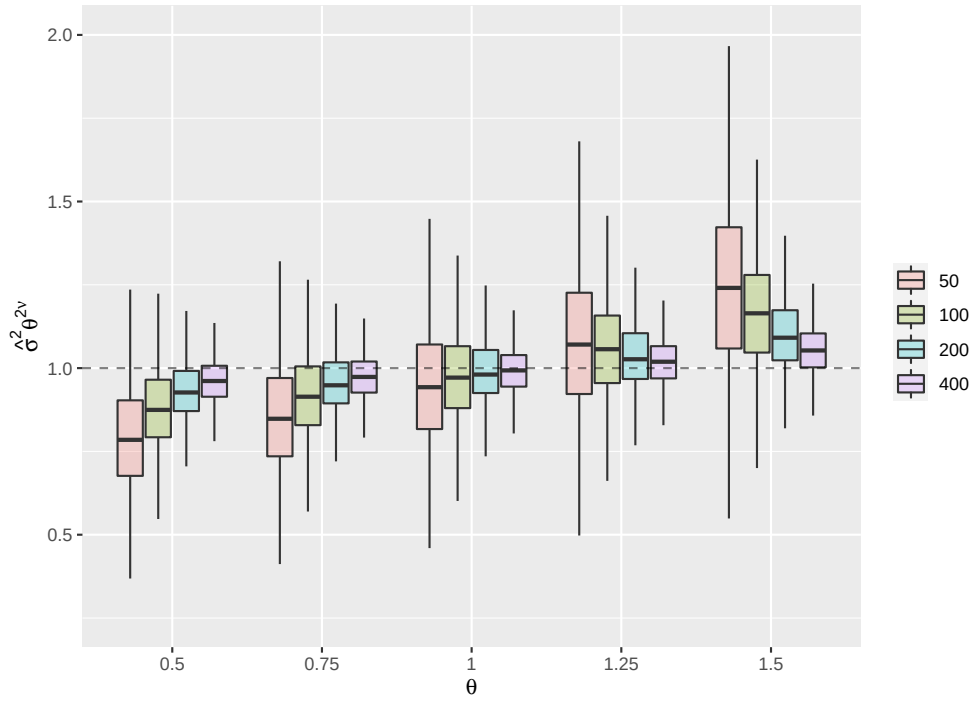
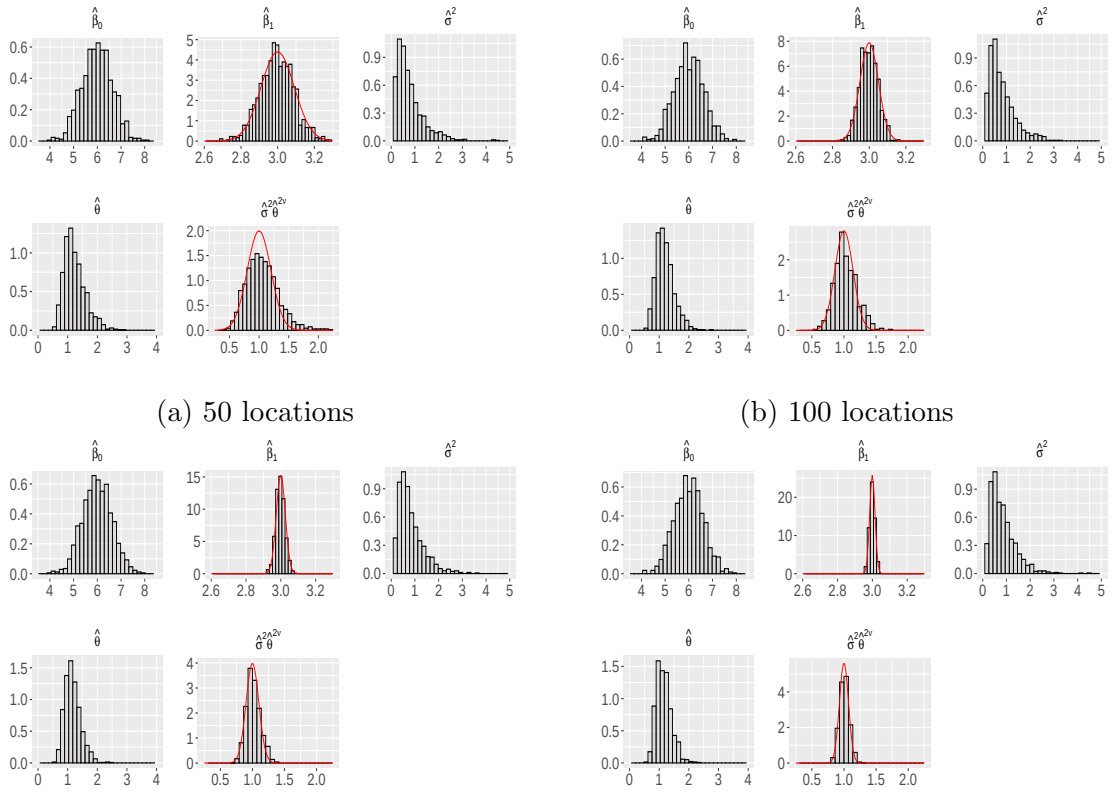


Figure 8.4: Boxplots of empirical distributions for $\hat{\sigma}^2\theta^{2\nu}$ when θ is fixed at different values. The four boxplots in each group correspond to sample sizes of $n = 50, 100, 200, 400$ going from left to right. The dashed line indicates the true value of $\sigma^2\theta^{2\nu} = 1$.

8.3.1.2 Estimated θ

We now consider the effect of estimating θ on the MLE of $\sigma^2\theta^{2\nu}$. Using the same generated data $\boldsymbol{x}$ and $\boldsymbol{y}|\boldsymbol{x}$ as in the fixed θ case, we compute MLE estimates of all parameters $(\beta_0, \beta_1, \sigma^2, \theta, \sigma^2\theta^{2\nu})^T$. Note that we use the term MLE very loosely here for the microergodic parameters $(\sigma^2, \theta)^T$. We are taking them to be computed minimizers, not necessarily unique, of the likelihood function over a bounded interval. Histograms of the MC estimates are given in Figure 8.5. In Figure 8.5, we can see that the empirical distributions for the MLEs of $(\beta_1, \sigma^2\theta^{2\nu})^T$ become more peaked with increasing number of observations.



(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

Figure 8.5: Histograms of MLE estimates of $(\beta_0, \beta_1, \sigma^2, \sigma^2\theta^{2\nu})^T$ when θ is estimated by its (pseudo) MLE. In red are the theoretical asymptotic densities predicted by Proposition 7.4.1. Estimates were computed from 1000 MC simulations of $\boldsymbol{y}|\boldsymbol{x}$ based on the regression model (8.1).

The same cannot be said for the MLEs of the non-microergodic parameters $(\beta_0, \sigma^2, \theta)^T$. This is consistent with the theory that $(\beta_1, \sigma^2\theta^{2\nu})^T$ are the only two microergodic parameters when $\nu_x < \ell$. Unlike in the fixed θ case, the empirical distributions of $\sigma^2\theta^{2\nu}$ do not show bias in comparison with their theoretical asymptotic densities, even in the $n = 50$ case. Once again, this is consistent with the simulations shown in Kaufman and Shaby (2013), who considered the problem with no regression parameters. Table 8.6 presents the variances of the above MLE empirical distributions for each sample size.

n	β_0	β_1	σ^2	θ	$\sigma^2\theta^{2\nu}$
50	0.404041	0.008461	0.369989	0.132107	0.076724
100	0.381767	0.002398	0.389967	0.092195	0.030815
200	0.377246	0.000667	0.343704	0.071875	0.012140
400	0.371659	0.000245	0.302005	0.064599	0.005585

Table 8.6: Empirical variances of MC estimates from Figure 8.5.

The variances of $(\hat{\beta}_{1,n}, \hat{\sigma}_n^2\hat{\theta}_n^{2\nu})^T$ both decay rapidly when the number of observations are doubled, whereas the remaining variances do not show such rapid decay. Comparing Tables 8.4 and 8.6, the empirical variances of the MLEs of $(\beta_1, \sigma^2\theta^{2\nu})^T$ are remarkably similar whether or not θ is fixed.

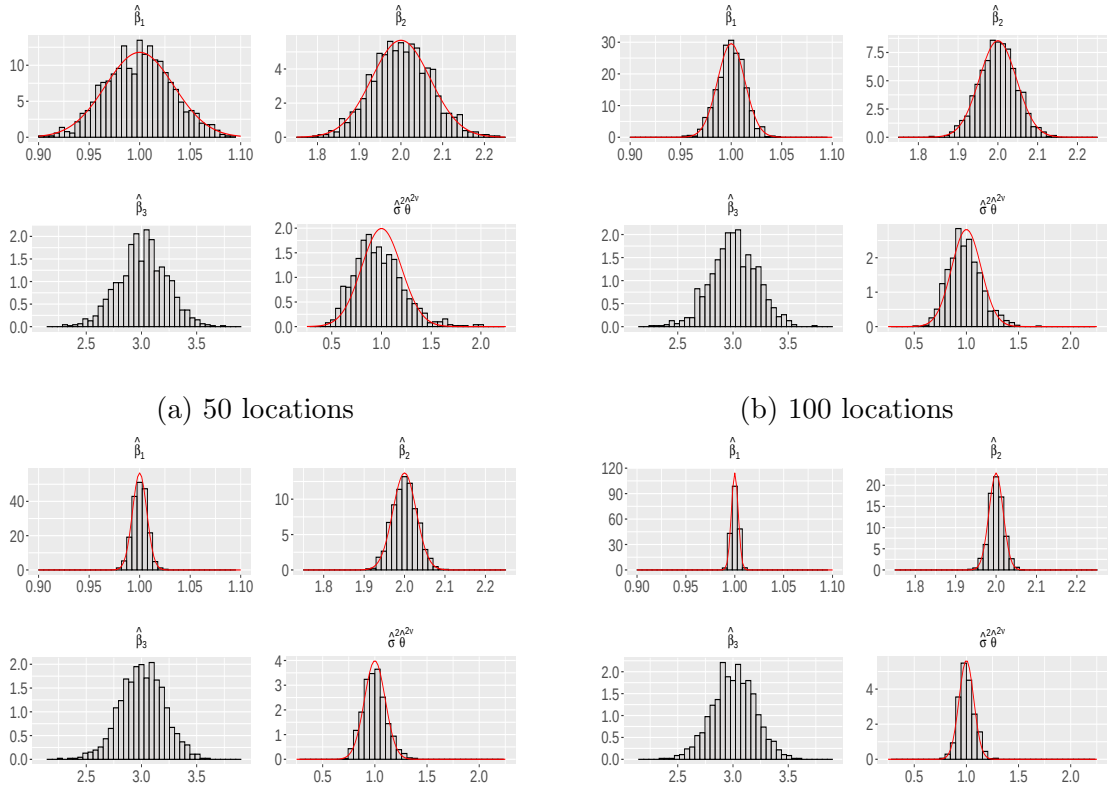
8.3.2 Multiple covariates

Finally, we consider adding a couple covariates to the model,

$$y(\boldsymbol{s}) = \beta_1 x_1(\boldsymbol{s}) + \beta_2 x_2(\boldsymbol{s}) + \beta_3 x(\boldsymbol{s}) + e(\boldsymbol{s}), \quad \boldsymbol{s} \in D \tag{8.4}$$

where $e(\boldsymbol{s})$ is independent of $x_k(\boldsymbol{s}), k = 1, 2, 3$ and D is the same unit circle in $\mathbb{R}^2$. We exclude the intercept since we know it cannot be consistently estimated. We take $e(\boldsymbol{s})$ to have the same Matérn covariance parameters $(\sigma^2, \theta, \nu)^T = (1, 1, \frac{3}{2})^T$ as before. Thus, the critical smoothness parameter remains at $\ell = \frac{5}{2}$. We give the covariates Matérn covariances with the same sill and scale parameters $(\sigma_x^2, \theta_x)^T = (1, 1)^T$. To illustrate the effect of smoothness on estimation, we choose smoothness parameters $\nu_{x1} = \frac{1}{2}, \nu_{x2} = 1$ and $\nu_{x3} = 3$ for $x_1(\boldsymbol{s}), x_2(\boldsymbol{s})$ and $x_3(\boldsymbol{s})$ respectively. For the true regression parameters, we arbitrarily choose $(\beta_1, \beta_2, \beta_3)^T = (1, 2, 3)^T$. Then, we

generate 1000 realizations of $\boldsymbol{y}|\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ based on (8.4) and compute MLE estimates for each realization. The histograms of these estimates are displayed in Figure 8.6.



(a) 50 locations

(b) 100 locations

(c) 200 locations

(d) 400 locations

Figure 8.6: Histograms of MLE estimates for $(\beta_1, \beta_2, \beta_3, \sigma^2\theta^{2\nu})^T$. In red, are the theoretical asymptotic densities of the microergodic parameters as predicted by equation (7.39). Estimates were computed from 1000 MC simulations of $\boldsymbol{y}|\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ based on the model (8.4).

For illustrative purposes, we omit the histograms for the non-microergodic parameters $(\sigma^2, \theta)^T$ and only focus on $(\beta_1, \beta_2, \beta_3, \sigma^2\theta^{2\nu})^T$. As expected, the empirical distributions of the MLEs of the microergodic parameters $(\beta_1, \beta_2, \sigma^2\theta^{2\nu})^T$ become more peaked with increasing number of observations. The empirical distribution of the MLE for β_3 does not show this behavior. The variances of the empirical distributions are given in Table 8.7.

n	β_1	β_2	β_3	$\sigma^2\theta^{2\nu}$
50	0.001085	0.004994	0.047734	0.059185
100	0.000183	0.002146	0.043962	0.024694
200	0.000051	0.000891	0.039362	0.012086
400	0.000012	0.000332	0.037022	0.005820

Table 8.7: Empirical variances of MC estimates from Figure 8.6

Note that the rate of decay in the variance of the MLE of β_2 is slower than that of β_1, even though both parameters are microergodic. This is likely because the Matérn smoothness for $x_2(\boldsymbol{s})$ is $\nu_{x2} = 1$, which is closer to the critical smoothness $\ell = \frac{5}{2}$ than that of $x_1(\boldsymbol{s})$, which was chosen to be $\nu_x = \frac{1}{2}$. This suggests that the rate of convergence may depend on the smoothness parameter and is not in general $\sqrt{n}$. Apart from this, the table generally agrees with the theoretical expectation that the variances of $(\beta_1, \beta_2, \sigma^2\theta^{2\nu})^T$ decay whereas the variance of β_3 does not. All histograms seem to be approximated well by their theoretical asymptotic densities.

Chapter 9 Conclusions and perspectives

In this dissertation, we addressed parametric estimation in linear and nonlinear spatial regression models under different asymptotic frameworks. Under increasing domain asymptotics, we considered regression models with and without confounding. In the case where there is no confounding, we brought attention to and bridged a gap in the spatial statistics and econometrics literature. Using asymptotic theory for spatial random fields developed by econometricians, we were able to adequately prove the consistency and asymptotic normality of estimators in commonly used in spatial statistics. In the presence of confounding, we expanded on existing literature by considering models with a nonlinear trend and unknown covariance parameters. We showed that it is possible to jointly estimate the unknown parameters in Gaussian spatial regression models under different confounding models. These parameters were shown to be well resolved under maximum likelihood estimation, even for moderately sized samples. Under infill asymptotics, we looked at estimation in linear regression models. Existing literature generally does not address estimation of the mean under infill asymptotics, but we showed that it is possible to consistently estimate regression parameters if the covariates have rougher sample paths compared to the error. We conclude with a summary of possible directions for future research.

Multiple covariates under confounding

In our investigation of confounding in spatial regression models, we only considered one covariate. A natural extension of this would be to include more covariates in the model, say for example,

$$y(\boldsymbol{s}) = f(x_1(\boldsymbol{s}), \cdots, x_p(\boldsymbol{s}); \boldsymbol{\beta}) + e(\boldsymbol{s})$$

where one or more of the covariates are now correlated with $e(\boldsymbol{s})$. This requires a valid multivariate cross covariance function $(x_1(\boldsymbol{s}), \cdots, x_p(\boldsymbol{s}), e(\boldsymbol{s}))^T$ with more than two components. It is not hard to generate valid covariances at specified locations, but finding reasonably well-motivated multivariate models including confounding without drastically restricting the parameters does present a challenge. Moreover it may not be simple or even possible to find closed forms of the resulting likelihood function of $\boldsymbol{y}|\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p$ like we did in the one covariate case $\boldsymbol{y}|\boldsymbol{x}$.

Confounding under infill asymptotics

We did not consider the effect of confounding in linear regression models under infill asymptotics. This would require some result on the equivalence and mutual singularity of probability measures induced by multivariate Gaussian random fields. An an example, suppose $(x(\boldsymbol{s}), e(\boldsymbol{s}))^T$ is a bivariate Gaussian random field with cross

covariance function,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2\phi(\boldsymbol{\theta}_x) & \rho\sigma_x\sigma_e\phi(\boldsymbol{\theta}_e) \\ \rho\sigma_x\sigma_e\phi(\boldsymbol{\theta}_e) & \sigma_e^2\phi(\boldsymbol{\theta}_e) \end{pmatrix}$$

To our knowledge, results on the equivalence and singularity of measures in the spirit of Zhang (2004) are scarce. In another co-authored paper by Zhang (Zhang and Cai (2015)), sufficient conditions for the equivalence of measures are given in a special case of the bivariate Matérn model,

$$\boldsymbol{C}(\boldsymbol{h}) = \begin{pmatrix} \sigma_x^2 M(\boldsymbol{h};\theta,\nu) & \rho\sigma_x\sigma_e M(\boldsymbol{h};\theta,\nu) \\ \rho\sigma_x\sigma_e M(\boldsymbol{h};\theta,\nu) & \sigma_e^2 M(\boldsymbol{h};\theta,\nu) \end{pmatrix}$$

where $M(\boldsymbol{h};\theta,\nu)$ is a Matérn kernel. Note that this is also a separable model. It is shown that for a fixed smoothness and spatial dimension $d \leq 3$, two measures $\mathbb{P}_1, \mathbb{P}_2$ induced by $(\sigma_{xi}^2, \sigma_{ei}^2, \theta_i, \rho_i)^T, i = 1, 2$ are equivalent if $\rho_1 = \rho_2, \sigma_{x1}^2\theta^{2\nu} = \sigma_{x2}^2\theta^{2\nu}$ and $\sigma_{e1}^2\theta^{2\nu} = \sigma_{e2}^2\theta^{2\nu}$. A similar result for one spatial dimension $d = 1$ is established in Velandia et al. (2017). These authors consider a separable model where each component is an exponential covariance function,

$$\boldsymbol{C}(t) = \begin{pmatrix} \sigma_x^2 e^{-\theta t} & \rho\sigma_x\sigma_e e^{-\theta t} \\ \rho\sigma_x\sigma_e e^{-\theta t} & \sigma_e^2 e^{-\theta t} \end{pmatrix}$$

Here, it is shown that two measures $\mathbb{P}_1, \mathbb{P}_2$ induced by $(\sigma_{xi}^2, \sigma_{ei}^2, \theta_i, \rho_i)^T, i = 1, 2$ are equivalent if and only if $\rho_1 = \rho_2, \sigma_{x1}^2\theta = \sigma_{x2}^2\theta$ and $\sigma_{e1}^2\theta = \sigma_{e2}^2\theta$. The results in

these specific models suggest it plausible that the confounding parameter ρ can be consistently estimated under infill asymptotics.

A hybrid asymptotic framework

We also considered a hybrid asymptotic framework that has features of both increasing domain and infill asymptotics. Suppose that we have the regression model,

$$y(\boldsymbol{s}) = \sum_{k=1}^{p} \beta_k x_k(\boldsymbol{s}) + e(\boldsymbol{s}), \quad \boldsymbol{s} \in D$$

where the error $e(\boldsymbol{s})$ is independent of the covariates $x_k(\boldsymbol{s}), k = 1, ..., p$ and D is a compact subset of $\mathbb{R}^d$. As in the infill asymptotics framework, we observe $y(\boldsymbol{s})$ and $x_k(\boldsymbol{s}), k = 1, \cdots, p$, at an increasing sequence of dense subsets $D_n = \{\boldsymbol{s}_1, \cdots, \boldsymbol{s}_n\}$ in D. To mimic aspects of the increasing domain asymptotics framework, we assume that the covariance functions of the covariates rapidly decay as the number of observations grows. To achieve this, take $\delta_n \nearrow \infty$ to be a sequence of increasing positive real numbers and assume that for any $\boldsymbol{s}_i, \boldsymbol{s}_j \in D_n$,

$$\mathbb{E}[x_k(\boldsymbol{s}_i, \boldsymbol{s}_j)] = C_k(\delta_n \boldsymbol{s}_i, \delta_n \boldsymbol{s}_j) \to 0, \quad k = 1, \cdots, p$$

as $n \to \infty$. Heuristically, the correlation decay allows us to assume mixing conditions on the covariates as in Chapter 2. Since we are no longer assuming a lattice where the locations are at a fixed distance apart, the arguments from Jenish and Prucha (2009) need to be modified slightly. We impose the following condition, which precludes the scaled sampling locations from becoming too dense.

Assumption 9.0.1. *For any n, define the set, $A_n = \{\{\boldsymbol{s}_i, \boldsymbol{s}_j\} : \delta_n ||\boldsymbol{s}_i - \boldsymbol{s}_j|| \leq 1\}$. Then, $\lim_{n\to\infty} \frac{1}{n^2}|A_n| = 0$ where $|\cdot|$ represents the cardinality of A_n.*

As an example of such a sampling scheme, suppose that for each $n \geq 1$, the locations $\{s_1, .., s_n\}$ fall inside the unit interval at equal spacings, $s_i = \frac{i}{n}, i = 1, ..., n$. Then, choosing $\delta_n = n$, we see that $A_n = \{\{i, j\} \in \{1, .., n\} : |i - j| \leq 1\}$. The number of pairs within a distance of 1 of each other is $O(n)$ and so $\frac{1}{n^2}|A_n| = 0$.

We now give a brief explanation as to why these assumptions restore consistency of the OLS estimates of the regression coefficients not allowed under infill asymptotics. In vector notation, the regression model can be written as,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \qquad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 & \cdots \boldsymbol{x}_p \end{bmatrix}_{n\times p}, \quad \boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T \tag{9.1}$$

where $\boldsymbol{y}$ is the vector of observations for the response at $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$ and similarly for $\boldsymbol{x}_j, j = 1, ..., p$ and $\boldsymbol{e}$. Letting $\boldsymbol{\beta}_0$ denote the true regression parameter vector, we have,

$$\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 = \left(\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i e_i \tag{9.2}$$

where $\boldsymbol{X}_i = (x_1(\boldsymbol{s}_i), \cdots, x_p(\boldsymbol{s}_i))^T$ is the i^{th} row of $\boldsymbol{X}$. Consider the vector,

$$\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i e_i = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} x_1(\boldsymbol{s}_i)e(\boldsymbol{s}_i) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_p(\boldsymbol{s}_i)e(\boldsymbol{s}_i) \end{pmatrix}$$

As in the proof of Proposition 3.1.1, we examine the asymptotic behavior of both the matrix $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T$ and vector $\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i e_i$. First, notice that the matrix $\frac{1}{n}\sum_{\boldsymbol{s}_i} \boldsymbol{X}_i \boldsymbol{X}_i^T$ contains terms of the form,

$$\frac{1}{n}\sum_{\boldsymbol{s}_i} x_j(\boldsymbol{s}_i), \qquad \frac{1}{n}\sum_{\boldsymbol{s}_i} x_j^2(\boldsymbol{s}_i), \qquad \frac{1}{n}\sum_{\boldsymbol{s}_i} x_j(\boldsymbol{s}_i)x_k(\boldsymbol{s}_i), \quad 1 \leq j,k \leq p$$

Using similar arguments and assumptions from Proposition 3.1.1 and Assumption 9.0.1, it can be shown that $\lim_{n\to\infty} \frac{1}{n}\sum_{\boldsymbol{s}_i} \boldsymbol{X}_i \boldsymbol{X}_i^T = \lim_{n\to\infty} \frac{1}{n}\sum_{\boldsymbol{s}_i} \mathbb{E}[\boldsymbol{X}_i \boldsymbol{X}_i^T]$, which we assume to exist and is non-singular. Now, consider the vector $\frac{1}{n}\sum_{\boldsymbol{s}_i} \boldsymbol{X}_i e_i$. These terms are of the form,

$$\frac{1}{n}\sum_{\boldsymbol{s}_i} e(\boldsymbol{s}_i)x_k(\boldsymbol{s}_i), \quad k = 1, .., p$$

It can be shown by mimicking the proof of Proposition 3.1.1, together with assumption 9.0.1, that each of these terms goes to 0 in probability. Thus, by (9.2), $\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0 \xrightarrow{P} \mathbf{0}$.

The next natural direction to consider for this framework would be maximum likelihood estimation. In Chapter 7, we determined that certain roughness conditions were needed for the consistency and asymptotic normality of the MLE of $\boldsymbol{\beta}$ to hold. It is reasonable to believe that in this asymptotic framework, these roughness conditions are not needed since by allowing correlation decay, we are essentially imitating mixing conditions. One may also consider how estimators of the variance parameters of $e(\boldsymbol{s})$ behave in the Matérn case. Since the error $e(\boldsymbol{s})$ does not experience correlation decay, we expect the same parameter $\sigma^2\theta^{2\nu}$ to be microergodic.

Bibliography

[1] Abrahamsen, Petter. *A Review of Gaussian Random Fields and Correlation Functions*. Norwegian Computing Center, Oslo, 1997.

[2] Abt, Markus and Welch, William J. Fisher information and maximum-likelihood estimation of covariance parameters in gaussian stochastic processes. *The Canadian Journal of Statistics*, 26(1):127–137, 1998.

[3] Adler, Robert J. *The Geometry of Random Fields*. John Wiley & Sons, New York, 1981.

[4] Anderes, Ethan B. On the consistent separation of scale and variance for gaussian random fields. *Annals of Statistics*, 38:870–893, 2009.

[5] Apanasovich, Tatiyana V., Genton, Marc G., and Sun, Ying. A valid matérn class of cross-covariance functions for multivariate random fields with any number of components. *Journal of the American Statistical Association*, 107(497): 180–193, 2012.

[6] Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[7] Banerjee, Sudipto, Carlin, Bradley P., and Gelfand, Alan E. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, New York, 2014.

[8] Bhat, B. R. On the method of maximum-likelihood for dependent observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):48–53, 1974.

[9] Billingsley, Patrick. *Convergence of Probability Measures*. Wiley, New York, 2 edition, 1968.

[10] Bivand, Roger, Nowosad, Jakub, and Lovelace, Robin. *spData: Datasets for Spatial Analysis*, 2022. URL `https://jakubnowosad.com/spData/`. R package version 2.2.0.

[11] Bolthausen, Erwin. On the Central Limit Theorem for Stationary Mixing Random Fields. *The Annals of Probability*, 10(4):1047 – 1050, 1982.

[12] Bradley, Richard C. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

[13] Cameron, R. H. and Martin, W. T. Transformations of weiner integrals under translations. *Annals of Mathematics*, 45(2):386–396, 1944.

[14] Çinlar, Erhan. *Probability and Stochastics*. Graduate Texts in Mathematics. Springer, New York, NY, 1st edition, 2011.

[15] Chen, Huann-Sheng, Simpson, Douglas G., and Ying, Zhiliang. Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica*, 10(1):141–156, 2000.

[16] Clayton, David G., Bernardinelli, Luisa, and Montomoli, Cristina. Spatial correlation in ecological analysis. *International Journal of Epidemiology*, 22(6): 1193–1202, 1993.

[17] Cressie, Noel and Lahiri, Soumendra Nath. The asymptotic distribution of REML estimators. *Journal of Multivariate Analysis*, 45(2):217–233, 1993.

[18] Cressie, Noel and Lahiri, Soumendra Nath. Asymptotics for REML estimation of spatial covariance parameters. *Journal of Statistical Planning and Inference*, 50(3):327–341, 1996.

[19] Cressie, Noel A.C. *Statistics for Spatial Data*. John Wiley & Sons, New Jersey, 1993.

[20] Crowder, Martin J. Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(1):45–53, 1976.

[21] Crujeiras, Rosa M. and van Keilegom, Ingrid. Least squares estimation of nonlinear spatial trends. *Computational Statistics & Data Analysis*, 54(2):452–465, 2010.

[22] Davis, Bruce M. and Borgman, Leon E. A note on the asymptotic distribution of the sample variogram. *Journal of the International Association for Mathematical Geology*, 14(2):189–193, 1982.

[23] Doukhan, Paul. *Mixing. Properties and Examples*. Springer-Verlag, New York, 1994.

[24] Driscoll, Michael F. The reproducing kernel hilbert space structure of the sample paths of a gaussian process. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 26:309–316, 1973.

[25] Du, Juan, Zhang, Hao, and Mandrekar, V. S. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *The Annals of Statistics*, 37 (6A):3330 – 3361, 2009.

[26] Feldman, Jacob. Equivalence and perpendicularity of Gaussian processes. *Pacific Journal of Mathematics*, 8(4):699 – 708, 1958.

[27] Gallant, A. Ronald and Goebel, J. Jeffery. Nonlinear regression with autocorrelated errors. *Journal of the American Statistical Association*, 71(356):961–967, 1976.

[28] Gelfand, Alan E., Diggle, Peter, Guttorp, Peter, and Fuentes, Montserrat. *Handbook of Spatial Statistics*. Taylor & Francis, Florida, 2010.

[29] Genton, Marc G. and Kleiber, William. Cross-Covariance Functions for Multivariate Geostatistics. *Statistical Science*, 30(2):147 – 163, 2015.

[30] Gneiting, Tilmann, Kleiber, William, and Schlather, Martin. Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177, 2010.

[31] Gräler, Benedikt, Pebesma, Edzer, and Heuvelink, Gerard. Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218, 2016. URL `https://journal.r-project.org/archive/2016/RJ-2016-014/index.html`.

[32] Grenander, Ulf. Stochastic processes and statistical inference. *Arkiv f́or Matematik*, 1(3):195 – 277, 1950.

[33] Guyon, Xavier. *Random Fields on a Network*. Springer, New York, 1995.

[34] Hájek, Jaroslav. On a property of normal distributions of an arbitrary stochastic process. *Czechoslovak Math. J.*, 8:610–618, 1958.

[35] Hanks, Ephraim M. Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics*, 26(4):243–254, 2015.

[36] Harrison, David and Rubinfeld, Daniel L. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978.

[37] Hayashi, Fumio. *Econometrics*. Princeton Univ. Press, Princeton, NJ, 2000.

[38] Heyde, Chris and Hall, Peter Gavin. *Martingale Limit Theory and Its Application*. Academic Press, New York, 1980.

[39] Hodges, James S. and Reich, Brian J. Adding spatially correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334, 2010.

[40] Hodges, James S., Reich, Brian J., and Zadnik, Vesna. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.

[41] Horn, Roger A. and Johnson, Charles R. *Matrix Analysis*. Cambridge University Press, Cambridge; New York, 2nd edition, 2013.

[42] Hughes, John and Haran, Murali. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B*, 75(1):139–159, 2013.

[43] Ibragimov, Ildar A. and Rozanov, Yuri A. *Gaussian Random Processes*. Springer, New York, 1978.

[44] Jenish, Nazgul and Prucha, Ingmar. Central limit theorems and uniform laws of large numbers for arrays of random fields. *Journal of Econometrics*, 150(1): 86–98, 2009.

[45] Journel, A.G. Markov models for cross-covariances. *Mathematical Geology*, 31 (8):955–964, 1999.

[46] Kakutani, Shizuo. On equivalence of infinite product measures. *Annals of Mathematics*, 49(1):214–224, 1948.

[47] Kaufman, C. G. and Shaby, B. A. The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484, 2013.

[48] Lahiri, Soumendra Nath. On inconsistency of estimators based on spatial data under infill asymptotics. *Sankhya: The Indian Journal of Statistics, Series A*, 58(3):403–417, 1996.

[49] Lahiri, Soumendra Nath, Lee, Yoondong, and Cressie, Noel. On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference*, 103(1): 65–85, 2002.

[50] Loh, Wei-Liem, Sun, Saifei, and Wen, Jun. On fixed-domain asymptotics, parameter estimation and isotropic Gaussian random fields with Matérn covariance functions. *The Annals of Statistics*, 49(6):3127 – 3152, 2021.

[51] Mardia, Kanti V. and Goodall, Colin R. *Spatial-temporal analysis of multivariate environmental monitoring data.*, chapter 16, pages 347 – 386. Multivariate Environmental Statistics. Elsevier, 1993.

[52] Mardia, Kanti V. and Marshall, Roger J. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146, 1984.

[53] Mercer, James. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441-458):415–446, 1909.

[54] Morris, Max D. and Ebey, Sherwood F. An interesting property of the sample mean under a first-order autoregressive model. *The American Statistician*, 38 (2):127–129, 1984.

[55] Newey, Whitney K. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59(4):1161–1167, 1991.

[56] Paciorek, Christopher J. The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25(1):107–125, 2010.

[57] Page, Garritt L., Liu, Yajun, He, Zhuoqiong, and Sun, Donchu. Estimation and prediction in the presence of spatial confounding for spatial linear models. *Scandinavian Journal of Statistics*, 44(3):780–797, 2017.

[58] Parzen, Emanuel. Extraction and detection problems and reproducing kernel hilbert spaces. *Journal of the Society for Industrial and Applied Mathematics Series A Control*, 1(1):35–62, 1962.

[59] Parzen, Emanuel. Probability density functionals and reproducing kernel hilbert spaces. In *Proceedings of the Symposium on Time Series Analysis*, volume 196, pages 155–169. Wiley, New York, 1963.

[60] Pötscher, Benedikt M. and Prucha, Ingmar R. Generic uniform convergence and equicontinuity concepts for random functions: An exploration of the basic structure. *Journal of Econometrics*, 60(1):23–63, 1994.

[61] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

[62] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2019. URL `http://www.rstudio.com/`.

[63] Scheuerer, Michael. Regularity of the sample paths of a general second order random field. *Stochastic Processes and their Applications*, 120(10):1879–1897, 2010.

[64] Schlather, Martin, Freudenberg, Alexander, Furrer, Reinhard, Kroll, Martin, Ripley, Brian D, and Ratcliff, John W. *RandomFieldsUtils: Utilites for the Simulation and Analysis of Random Fields*, 2022. URL `https://cran.r-project.org/package=RandomFieldsUtils`. R package version 1.2.5.

[65] Skorokhod, Anatoliy V. and Yadrenko, Myhailo I. On absolute continuity of measures corresponding to homogeneous gaussian fields. *Theory of Probability & Its Applications*, 18(1):27–40, 1973.

[66] Stein, Michael L. Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics*, 16(1):55 – 63, 1988.

[67] Stein, Michael L. *Interpolation of Spatial Data*. Springer-Verlag, New York, 1999.

[68] Sweeting, T. J. Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics*, 8(6):1375 – 1381, 1980.

[69] Tang, Wenpin, Zhang, Lu, and Banerjee, Sudipto. On identifiability and consistency of the nugget in Gaussian spatial process models. *Journal of the Royal Statistical Society Series B*, 83(5):1044–1070, 2021.

[70] van der Vaart, Aad. *Asymptotic Statistics*. Cambridge University Press, Cambridge, 1998.

[71] Velandia, Daira, Bachoc, François, Bevilacqua, Moreno, Gendre, Xavier, and Loubes, Jean-Michel. Maximum likelihood estimation for a bivariate Gaussian process under fixed domain asymptotics. *Electronic Journal of Statistics*, 11 (2):2978 – 3007, 2017.

[72] Wang, Daqing and Loh, Wei-Liem. On fixed-domain asymptotics and covariance tapering in Gaussian random field models. *Electronic Journal of Statistics*, 5:238 – 269, 2011.

[73] Watson, G.N. *A Treatise on the Theory of Bessel Functions*. Cambridge Mathematical Library. Cambridge University Press, 1995.

[74] Wendland, Holger. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2004.

[75] Yadrenko, Myhailo I. *Spectral Theory of Random Fields*. Optimization Software, New York, 1983.

[76] Ying, Zhiliang. Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36(2):280–296, 1991.

[77] Zhang, Hao. Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.

[78] Zhang, Hao and Cai, Wenxiang. When Doesn't Cokriging Outperform Kriging? *Statistical Science*, 30(2):176 – 180, 2015.

[79] Zimmerman, Dale L. and Zimmerman, M. Bridget. A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, 33(1):77–91, 1991.