RECENT ADVANCES IN COMPUTER EXPERIMENT MODELING

BY YUFAN LIU

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Statistics and Biostatistics

Written under the direction of

Ying Hung

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2014

ABSTRACT OF THE DISSERTATION

Recent Advances in Computer Experiment Modeling

by YUFAN LIU

Dissertation Director: Ying Hung

This dissertation develops methodologies for analysis of computer experiments and its related theories. Computer experiments are becoming increasingly important in science and Gaussian process (GP) models are widely used in the analysis of computer experiments. This dissertation focuses on two settings where massive data are observed on irregular grids or quantiles of correlated data are of interests. In this dissertation, we first develop Latin Hypercube Design-based Block Bootstrap method. Then, we investigate quantiles of computer experiments in which correlated data are observed and propose penalized quantile regression with asymmetric Laplace process.

The computational issue that hinders GP from broader application is recognized, especially for massive data observed on irregular grids. To overcome the computational issue, we introduce an efficient framework based on a novel experimental design based bootstrap method. The main challenge in GP modeling is the estimation of maximum likelihood estimators because it relies heavily on large correlation matrix operations, which are computationally intensive and often intractable for massive data. Using the

idea of design-based data reduction, the proposed framework provides an asymptotically consistent estimation for the parameters in GP with a dramatic reduction in computation. The finite-sample performance is examined through simulation studies. We illustrate the proposed method by a data center example based on tens of thousands of computer experiments generated from a computational fluid dynamics simulator.

GP models and many other existing approaches focus on modeling the conditional mean of the response variable in computer experiments. Little work has been done to study quantile regression model that incorporate data dependence although in practice it is often of substantial interest. In addition, high dimensional data often display heterogeneity and call for models with sparsity in which only a small number of covariates have influence on the conditional distribution of the response. We propose a new modeling framework to model different quantiles in computer experiments and identify important effects for each quantile. The proposed approach utilize asymmetric Laplace process (ALP) instead of Gaussian process modeling. Also, penalized likelihood estimators for ALP are studied. We show that penalized quantile asymmetric Laplace estimator can select true relevant covariates when the number of covariates is large and the number of covariates is able to grow to infinity when the number of observations increase to infinity. Penalized quantile regression with asymmetric Laplace process is demonstrated numerically with simulation and a real data example.

Acknowledgements

First, I would like to express my deepest appreciation to my advisor Professor Ying Hung for her valuable ideas, helpful comments, great support, continuous encouragement and kind patience. I feel very lucky to have worked with her. Without her help and advice, I can not imagine completing this dissertation. I am also grateful to Professor Regina Liu, Professor William Strawderman and Professor Myong K. Jeong for their precious time and efforts to serve on my thesis committee.

My thanks also go to the Department of Statistics and Biostatistics of Rutgers University for providing me support and a great learning and research environment.

Last but definitely not the least, my most heartfelt thanks go to my dearest parents and wife, for their selfless love and support.

Dedication

To my wife, Xueying

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Computer experiments refer to the study of real systems using complex mathematical models. They have been widely used as alternatives to physical experiments, especially for studying complex systems. The reason is, in many situations, a physical experiment is infeasible because it is unethical, impossible, inconvenient or too expensive. A mathematical model of the system can often be developed and input/output pairs can be produced with the help of computers. Computer experiments are widely used in science, engineering and medicine. Typically, computer experiments require a great deal of time and computing to obtain and they are nearly deterministic in the sense that a particular input will produce almost the same output if given to the computer experiment on another occasion. Therefore, it is desirable to build an interpolator for computer experiment outputs and use it as an emulator for the actual computer experiment. More discussions of design and analysis of computer experiments can be found in Fang et al. (2006); Koehler and Owen (1996); Sacks et al. (1989a,b); Santner et al. (2003).

A Gaussian process (GP) model (or called kriging) is a flexible and widely used interpolator in the analysis of computer experiments (Fang et al., 2006; Santner et al., 2003). However, the computational issue that hinders GP from broader applications is well recognized, especially for massive data observed on irregular grids (Gramacy

and Apley, 2013; Gramacy and Lee, 2008; Nychka et al., 2011; Kaufman et al., 2011; Peng and Wu, 2014). This is because modeling and inference of GP heavily involve manipulations of the $n \times n$ correlation matrix that require $O(n^3)$ computations, where n is the sample size. The calculation is computationally intensive and often intractable for massive data, i.e., large n. To overcome the computational difficulties, we propose a Latin Hypercube Design-based (LHD-based) Block Bootstrap method. It is an innovative experimental design-based subsampling plan, which can achieve an accurate and efficient approximation of the maximum likelihood estimators (MLEs) in GP models. The proposed sampling plan provides efficient and flexible data reduction so that the computational complexity is dramatically reduced and the correlation structure among data is still kept. We show that the LHD-based bootstrap estimators are asymptotically consistent to the MLEs using complete data. Details of the proposed approach are provided in Chapter 2.

In Chapter 3, we focus on modeling quantiles of data in computer experiments. The underlying data structure is often of high dimensional and correlated. Despite numerous research on computer experiment modeling, most of the existing approaches focus on modeling the conditional mean of the response variable (Fang et al., 2006; Santner et al., 2003). Little work is done to study quantile regression model that incorporate data dependence although in practice it is often of substantial interest. We propose a new modeling framework to model different quantiles in computer experiments and simultaneously identify important effects for each quantile. To achieve this goal, we extend Gaussian process to asymmetric Laplace process. The choice of asymmetric Laplace process enable us to model the quantiles and incorporate dependence structure. In addition, we regularize the quantile asymmetric Laplace process with a penalty

function, such as L_1 norm penalty (Tibshirani, 1996), the SCAD (Fan and Li, 2001) and the MCP (Zhang, 2010). We show that penalized quantile asymmetric Laplace estimator can select true relevant covariates when the number of covariates is large and able to grow to infinity with the number of observations increasing to infinity.

The rest of this thesis is organized as follows. In Chapter 2, we develop a Latin Hypercube Design-based Block Bootstrap method for Gaussian process model. In Chapter 3, we propose a penalized quantile regression with asymmetric Laplace process.

Chapter 2

Latin Hypercube Design-based Block Bootstrap for Computer Experiment Modeling

2.1 Introduction

In this chapter, we consider the case that massive data are observed on irregular grids in computer experiments. Thus it is impossible or takes too much time to perfrom GP modeling. Several methods are proposed in the literature to address the computational issue in GP modeling; however, to the best of our knowledge, this problem has not been satisfactorily resolved. Apart from computer experiments, this issue has also been recognized in the field of spatial statistics and machine learning. The existing approaches may be characterized broadly as either changing the model to one that is computationally convenient or approximating the likelihood for the original data. Examples of the former include Banerjee et al. (2008), Cressie and Johannesson (2008), Gramacy and Lee (2008),Rue and Held (2005), Rue and Tjelmeland (2002), and Wikle (2010), while approximation approaches includes Fuentes (2007), Furrer et al. (2006), Gramacy and Apley (2013), Kaufman et al. (2008), Nychka (2000), Nychka et al. (1998), Nychka et al. (2002), Smola and Bartlett (2001), Snelson and Ghahramani (2006), and Stein et al. (2004). Despite various methods, most of the existing ones are developed for data sets collected from a regular grid under a low-dimensional geostatistical setting.

These assumptions are often violated in computer experiments because having high-dimensional inputs is common in complex computer experiments and the computational expense often prohibits running computer experiments over a dense grid of input configurations (Fang et al., 2006; Santner et al., 2003; Tang, 1993; Ye, 1998). Recent studies in computer experiments address these issues by imposing a sparsity constraint on the correlation matrix, such as covariance tapering (Kaufman et al., 2008, 2011). However, it has been shown that this method does not work well for purposes of parameter estimation (Liang et al., 2013; Stein, 2013), which is crucial for the construction of GP predictors and statistical inference. In addition, the connection between the degree of covariance matrix sparsity and computation time is nontrivial.

A new framework based on the idea of bootstrap is proposed here to alleviate the computational difficulty of GP modeling, given no loss of estimation consistency asymptotically. Bootstrap is a powerful and increasingly utilize method for statistical inference (DiCiccio and Efron, 1992, 1996; Efron, 1979; Efron and Tibshirani, 1994). Direct extension of existing bootstrap methods to GP models is theoretically attractive but practically inapplicable especially for massive data. This is because conventional bootstrap methods are developed for independent data. Although various block bootstrap methods are proposed for dependent data (Kunsch et al., 1989; Lahiri, 2003; Liu and Singh, 1992; Paparoditis and Politis, 2001), they focus mainly on low-dimensional data such as time series or spatial data and concatenate the sample blocks into a bootstrap sample which has a similar size as the original observations. The application of these methods leads to the same computational difficulties as the standard GP models, including the complexity and singularity in large correlation matrix operations. Because of the high dimensionality of the input space and the massive outputs in computer

experiments, an efficient bootstrap sampling scheme that can tackle the computational issue with GP models is called for.

To address the foregoing issue, a new bootstrap subsampling method called Latin hypercube design-based (LHD-based) block bootstrap is proposed. It is an innovative experimental design-based subsampling plan, which can achieve an accurate and efficient approximation of the maximum likelihood estimators (MLEs) in GP models. The proposed sampling plan provides efficient and flexible data reduction so that the computational complexity is dramatically reduced. Theoretical studies show that the resulting estimators are asymptotically consistent to the MLEs using complete data. Moreover, the proposed approach can be easily parallelized to further speedup the computation. Beyond computer experiments, this framework can be extended to the area of spatial statistics, machine learning, and optimization with massive data.

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the GP modeling framework based on LHD-based block bootstrap. Asymptotic properties are discussed in Section 2.3. In Section 2.4, finite-sample performance of the proposed framework are investigated in a simulation study and, for illustration, the method is applied to a real data set generated from a computational fluid dynamics simulator for a data center thermal management study. Discussion is given in Section 2.5.

2.2 Efficient Gaussian process modeling using bootstrap

2.2.1 Gaussian process models for computer experiments

Consider a computer experiment that has inputs $\boldsymbol{x} \in R^d$ and produces univariate output $y(\boldsymbol{x})$. To analyze the experiments, the output $y(\boldsymbol{x})$ is generally assumed to be

a realization from a stochastic process

$$Y(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}), \tag{2.1}$$

where the mean function is defined as $\mu(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta}$ and $Z(\boldsymbol{x})$ is a weak stationary Gaussian process with mean 0 and covariance function $\sigma^2\psi$. The covariance function is defined by $cov\{Y(\boldsymbol{x}+\boldsymbol{h}), Y(\boldsymbol{x})\} = \sigma^2\psi(\boldsymbol{h};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of correlation parameters and $\psi(\boldsymbol{h};\boldsymbol{\theta})$ is a positive semidefinite function with $\psi(\boldsymbol{0};\boldsymbol{\theta}) = 1$ and $\psi(\boldsymbol{h};\boldsymbol{\theta}) = \psi(-\boldsymbol{h};\boldsymbol{\theta})$. Note that we assume the variables in the mean function are known and such a model is also known as universal kriging. However, the proposed framework is not limited to this assumption. It can be further extended to incorporate various variable selection methods for GP models Chu et al. (2011); Li and Sudjianto (2005).

Suppose n realization are observed and denoted by

$$\begin{aligned}\mathscr{D}_n &= \{\big(\boldsymbol{x}_{t_1}, y(\boldsymbol{x}_{t_1})\big), \ldots, \big(\boldsymbol{x}_{t_n}, y(\boldsymbol{x}_{t_n})\big)\} \\ &= \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}.\end{aligned}$$

Denote $\boldsymbol{y}_n = (y_1, \ldots, y_n)^T$, $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{\phi} = (\theta^T, \boldsymbol{\beta}^T, \sigma^2)^T$ as the vector of all p parameters, and Θ as the parameter space. The likelihood function for (2.1), can be written as

$$f(\boldsymbol{y}_n, \boldsymbol{X}_n; \boldsymbol{\phi}) = \frac{|R_n(\boldsymbol{\theta})|^{-1/2}}{(2\pi\sigma^2)^{n/2}} \exp\{-\frac{1}{2\sigma^2}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})\},$$

where $R_n(\boldsymbol{\theta}) = [\psi(y(\boldsymbol{x}_i), y(\boldsymbol{x}_j); \boldsymbol{\theta}), i, j = 1, \ldots, n]$ is an $n \times n$ correlation matrix. Thus, the log-likehood function, ignoring a constant, is

$$\ell(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) = -\frac{1}{2\sigma^2}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}) - \frac{1}{2}|R_n(\boldsymbol{\theta})| - \frac{n}{2}\log(\sigma^2).$$

Here, the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and σ are unknown. They are estimated using likelihood-based methods such as maximum likelihood or restricted maximum likelihood (REML)

Irvine et al. (2007). In this chapter, we focus on the study of maximum likelihood estimators and the results can be generalized to other approached such as REML, cross validation or Bayesian methods (Kaufman et al., 2011).

For a GP model, the maximum likelihood estimators (MLEs) can be obtained by

$$\hat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n^T R_n^{-1}(\boldsymbol{\theta}) \boldsymbol{X}_n)^{-1} \boldsymbol{X}^T R_n^{-1}(\boldsymbol{\theta}) \boldsymbol{y}_n,$$

$$\hat{\sigma}_n^2 = (\boldsymbol{y}_n - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}_n)^T R_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}_n)/n,$$

and

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}} \{n \log(\hat{\sigma}_n^2) + \log |R_n(\boldsymbol{\theta})|\},$$

where $|R_n(\boldsymbol{\theta})|$ is the determinant of matrix $R_n(\boldsymbol{\theta})$. Based on the estimated MLEs, the predictor for a new point $\boldsymbol{x}_{n+1}$ is given by $y(\boldsymbol{x}_{n+1}) = \boldsymbol{x}_{n+1}^T \hat{\boldsymbol{\beta}}_n + \gamma_n(\hat{\boldsymbol{\theta}}_n)^T R_n^{-1}(\hat{\boldsymbol{\theta}}_n)(\boldsymbol{y}_n - \boldsymbol{X}_n \hat{\boldsymbol{\beta}}_n)$ with $\gamma_n(\hat{\boldsymbol{\theta}}_n)$ being the correlation between the new point and the observations, i.e. $\gamma_n(\hat{\boldsymbol{\theta}}_n) = [\psi(y(\boldsymbol{x}_i), y(\boldsymbol{x}_{n+1}); \hat{\boldsymbol{\theta}}_n), i = 1, \ldots, n]$.

The main challenge in GP modeling is the calculation of MLEs. This is because it relies heavily on the calculation of $R_n^{-1}(\boldsymbol{\theta})$ and $|R_n(\boldsymbol{\theta})|$, which is computationally intensive and often intractable due to numerical issues. It is particularly difficult for massive data (i.e., large n) when they are collected on irregular grids because no Kronecker product techniques can be utilized for computational simplification (Bayarri et al., 2007, ????; Rougier, 2008). Alternatives, such as Bayesian methods suffer from the same difficulty.

2.2.2 Latin hypercube design-based block bootstrap

To overcome the computation issue with GP modeling, especially the estimation of parameters, we propose to estimate MLEs using subsamples collected from LHD-based

block bootstrap. LHD-based block bootstrap is a subsampling plan combining the idea of block bootstrap and Latin hypercube designs (LHDs), a widely used class of space-filling designs in computer experiments (McKay et al., 1979). In general, a m-run LHD in d dimensions can be generated using a random permutation of $\{0, \cdots, m-1\}$ for each dimension. Each permutation leads to a different LHD. LHDs are easy to generate and enjoy a desirable space-filling property called univariate stratification property (Tang, 1993), i.e., the design points are evenly spread across the projection of the experimental region onto any dimension. Because the block bootstrap technique can capture the dependence structure of the underlying response and LHDs are known to be space-filling in high-dimensional space, such a combination provides an efficient subsampling scheme for high-dimensional computer experiment problems.

Assume that we have n observations $\{y_i(\boldsymbol{x}_i)\}$ from a computer model, where $i = 1, ..., n$ and $\boldsymbol{x}_i \in R^d$. Denote the d-dimensional input space by Γ. A LHD-based block bootstrap procedure can be described in the following two steps.

Step 1: *Decompose the d-dimensional experimental region Γ, assumed to be $[0, l]^d$, into disjoint hypercubes. This is achieved by dividing each dimension into m equally spaced intervals so that Γ consists of m^d disjoint hypercubes. Define each hypercube by*

$$\mathcal{B}_n(\boldsymbol{i}) = b(\boldsymbol{i} + \mathcal{U}),$$

where $\boldsymbol{i} = (i_1, ...i_d)$, $i_j \in (0, ..., m-1)$, represents the starting point of each hypercube, $b = l/m$, and $\mathcal{U} = (0, 1]^d$ is the unit hypercube. Let $|\mathcal{B}_n(\boldsymbol{i})|$ be the number of observations in the $\boldsymbol{i}$th hypercube. For simplicity, assume the data points are equally distributed over the blocks, i.e. $|\mathcal{B}_n(\boldsymbol{i})| = n/m^d$.

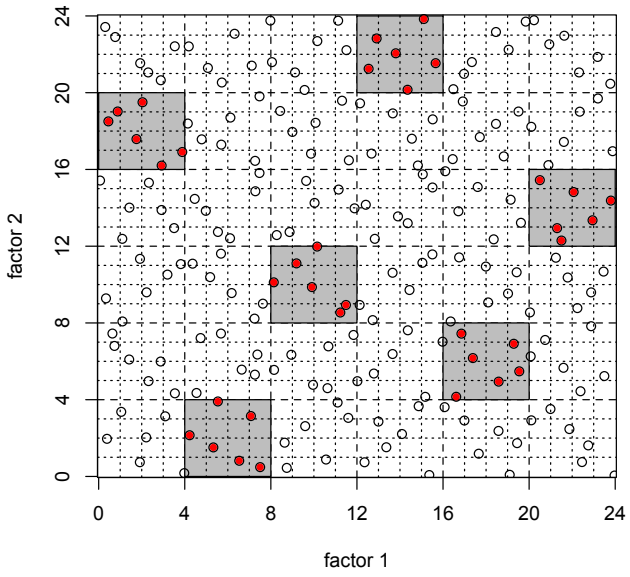

Figure 2.1: An example of LHD-based block bootstrap

Step 2: *A LHD-based block bootstrap sample is defined by selecting indices* $\boldsymbol{i}$ *according to a randomly generated* m*-run LHD in a* d*-dimensional space, denoted by* $\boldsymbol{i}_1^*,\ldots,\boldsymbol{i}_m^*$*. More specifically, let* $\boldsymbol{\pi}_i = (\pi_i(1),\ldots,\pi_i(m))$*,* $1 \le i \le d$*, be independent random permutations of* $\{0,\ldots,m-1\}$*, which is uniformly distributed over* $m!$ *possible permutations. Thus,* $\boldsymbol{i}_j^* = (\boldsymbol{\pi}_1(j),\ldots,\boldsymbol{\pi}_d(j))$*,* $j = 1,\ldots,m$*. The* m *selected hypercubes are denoted by* $\mathcal{B}_n(\boldsymbol{i}_1^*),\ldots,\mathcal{B}_n(\boldsymbol{i}_m^*)$*. The bootstrap samples with size* $N = \sum_{i=1}^m |\mathcal{B}_n(\boldsymbol{i}_i^*)|$*, denoted by* $y_1^*(\boldsymbol{x}_1^*),\ldots,y_N^*(\boldsymbol{x}_N^*)$*, are the observations in the selected cubes. Based on the* N *subsamples,* $\hat{\boldsymbol{\phi}}_N^*$ *is obtained by maximizing the likelihood.*

Figure 2.1 illustrates an example of LHD-based block bootstrap samples with $d = 2$, $l = 24$, $b = 4$, $m = 6$, $|\mathcal{B}_n(\boldsymbol{i})| = 6$, $N = 36$, and $n = 216$. The 6-run 2-dimensional LHD is denoted by $\boldsymbol{i}_1^* = (0,4)$, $\boldsymbol{i}_2^* = (1,0)$, $\boldsymbol{i}_3^* = (2,2)$, $\boldsymbol{i}_4^* = (3,5)$, $\boldsymbol{i}_5^* = (4,1)$, and

$\boldsymbol{i}_6^* = (5, 3)$. The observations, denoted by circles, are located irregularly on the grid. The gray areas are the LHD-based blocks. It is clear that the univariate stratification property of LHDs is preserved; therefore, if we project the selected blocks onto any dimension, they are evenly spread out. The red dots are the resulting subsample from this LHD-based block bootstrap.

Different from the existing block bootstrap approaches (Kunsch et al., 1989; Lahiri, 1995, 1999, 2003; Liu and Singh, 1992; Politis and Romano, 1994), the subsamples obtained by LHD-based block bootstrap contain an attractive space-filling structure with a significantly smaller sample size to efficiently achieve data reduction and thus reduce computation.

2.3 Consistency of the LHD-based block Bootstrap Estimators

2.3.1 Notations

Recall that $\boldsymbol{\phi}$ is a vector representing all the parameters in a GP model. It is estimated by maximizing the likelihood function, which is the most computationally intensive step. Based on n outputs, $y_1(\boldsymbol{x}_1), ..., y_n(\boldsymbol{x}_n)$, from a computer experiment, the MLE of $\boldsymbol{\phi}$ is denoted by $\hat{\boldsymbol{\phi}}_n$. A bootstrap version of $\hat{\boldsymbol{\phi}}_n$, denoted by $\hat{\boldsymbol{\phi}}_N^*$, is obtained by calculating the MLE for each LHD-based block bootstrap subsample collected with $\mathbf{X}_N^* = (\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_N^*)^T$ and $\boldsymbol{y}_N^* = (y_1^*, \ldots, y_N^*)^T$. Our primary interest is to investigate the asymptotic properties of the estimated parameters under increasing domain asymptotic (Chu et al., 2011; Cressie and Cassie, 1993; Mardia and Marshall, 1984), which is suitable for computer experiments. In contrast, properties under fixed domain asymptotic (Stein, 1999; Ying, 1993; Zhang, 2004) deserve further studies and they are left for future investigation.

We first introduce a mathematical formalization of the LHD-based block bootstrap procedure. Given the underlying probability space $(\Omega, \mathcal{F}, P)$ of a Gaussian process, a sample of size n with settings $\boldsymbol{x}_1(\omega), ..., \boldsymbol{x}_n(\omega)$ and corresponding $y(\boldsymbol{x})$'s are observed from a given realization $\omega \in \Omega$. Let $(\Lambda, \mathcal{G})$ be a measurable space on the realization. For each $\omega \in \Omega$, denote $P^*_{N,\omega}$ as the probability measure induced by the m-run LHD bootstrap on $(\Lambda, \mathcal{G})$. The proposed bootstrap is a method to generate new dataset on $(\Lambda, \mathcal{G}, P^*_{N,\omega})$ conditional on the n original observations. Let $\tau_t : \Lambda \to \{1, ..., n\}$ denote a random index generated by the LHD-based block bootstrap. So, τ_t is the tth index in the intersect index of observations and $\{\mathcal{B}_n(\boldsymbol{i}^*_1), ..., \mathcal{B}_n(\boldsymbol{i}^*_m)\}$, where $(\boldsymbol{i}^*_1, ..., \boldsymbol{i}^*_m)$ is a randomly generated m-run LHD. Therefore, for $(\lambda, \omega) \in \Lambda \times \Omega$, we have the tth bootstrap sample: $\boldsymbol{x}^*_t(\lambda, \omega) \equiv \boldsymbol{x}_{\tau_t(\lambda)}(\omega)$.

2.3.2 Consistency of the LHD-based block bootstrap mean

Before studying the asymptotic performance of MLEs, this section focuses on understanding properties of the LHD-based block bootstrap mean, which is an important foundation to the theoretical development for $\hat{\boldsymbol{\phi}}^*_N$ later. Suppose $\{Y(\boldsymbol{x}_t), t \in R\}$ follows a Gaussian process with mean μ. Given n observations, the sample estimation of mean μ is

$$\bar{y}_n = \frac{1}{n} \sum_{s=1}^{n} y_s,$$

and the LHD-based block bootstrap mean with N samples is given by

$$\bar{y}^*_N = \frac{1}{N} \sum_{s=1}^{N} y^*_s.$$

With a slight abuse of notation, we replace the notation of random variable Y by its realization y unless otherwise specified. In addition, $\boldsymbol{E}(\cdot)$ and $\boldsymbol{Cov}(\cdot,\cdot)$ denote the expectation and variance under P while $\boldsymbol{E}^*_{N,\omega}(\cdot)$ and $\boldsymbol{Cov}^*_{N,\omega}(\cdot,\cdot)$ denote the expectation and variance under $P^*_{N,\omega}$.

The properties of LHD-based block bootstrap mean are investigated in the following lemmas which lead to a proof of the distribution consistency of $\bar{y}^*_N$.

Lemma 2.1 *LHD-based block bootstrap mean is unbiased, i.e.,*

$$\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = \bar{y}_n.$$

Proof: Since the data points are equally distributed over all the blocks, we have

$$\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = \sum_{i_1,\ldots,i_d} \frac{1}{m^d} \bar{y}_{i_1,\ldots,i_d} = \bar{y}_n. \square$$

The next two lemmas show the consistency of LHD-based block bootstrap variance. Denote the population variance of Gaussian process by $\tau_n^2 = \frac{1}{n}\sum_{s,t=1}^{n} \boldsymbol{Cov}(Y_s(\boldsymbol{x}_s), Y_t(\boldsymbol{x}_t))$. Given the following regularity conditions, Lemma 2.2 provides an infeasible consistent estimator of τ_n^2.

(A.1) $\frac{n}{m^d}\boldsymbol{Cov}\{(\bar{y}_{\boldsymbol{i}} - \mu)^2, (\bar{y}_{\boldsymbol{j}} - \mu)^2\} = O(1)$.

(A.2) $|\tau_n^2| = O(1)$.

Lemma 2.2 *Let* $\bar{y}_{\boldsymbol{i}} = \frac{1}{\mathcal{B}_n(\boldsymbol{i})}\sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i})} y_s$, $\forall \boldsymbol{i} = (i_1, \ldots, i_d)$. *Under (A.1) and (A.2), we have*

$$\frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2 - \tau_n^2 \xrightarrow{\mathrm{P}} 0.$$

Proof: Denote by $A_n = \frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2$. We will show that $\boldsymbol{Cov}(A_n, A_n) = 0$ and $\boldsymbol{E}(A_n) = \tau_n^2$. Variance of A_n is calculated as the following:

$$
\begin{aligned}
& \boldsymbol{Cov}(A_n, A_n) \\
= \; & \boldsymbol{Cov}(\frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2, \frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} (\bar{y}_{i_1,\ldots,i_d} - \mu)^2) \\
= \; & \frac{n^2}{m^{4d}} \sum_{i_1,\ldots,i_d} \sum_{j_1,\ldots,j_d} \boldsymbol{Cov}((\bar{y}_{i_1,\ldots,i_d} - \mu)^2, (\bar{y}_{j_1,\ldots,j_d} - \mu)^2) \\
= \; & \frac{n^2}{m^{4d}} \frac{m^{4d}}{n^4} \sum_{\boldsymbol{i}} \sum_{\boldsymbol{j}} \sum_{\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2} \in \mathcal{B}_n(\boldsymbol{i})} \sum_{\boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{j})} \boldsymbol{Cov}\{(y_{s_1} - \mu)(y_{s_2} - \mu), (y_{t_1} - \mu)(y_{t_2} - \mu)\} \\
= \; & \frac{1}{n^2} \sum_{\boldsymbol{i}} \sum_{\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2}, \boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{Cov}\{(y_{s_1} - \mu)(y_{s_2} - \mu), (y_{t_1} - \mu)(y_{t_2} - \mu)\} \\
& + \frac{1}{n^2} \sum_{\boldsymbol{i} \neq \boldsymbol{j}} \sum_{\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2} \in \mathcal{B}_n(\boldsymbol{i})} \sum_{\boldsymbol{x}_{t_1}, \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{j})} \boldsymbol{Cov}\{(y_{s_1} - \mu)(y_{s_2} - \mu), (y_{t_1} - \mu)(y_{t_2} - \mu)\}
\end{aligned}
$$

By separately expanding the two terms above, we can rewrite $\boldsymbol{Cov}(A_n, A_n)$ as

$$
\begin{aligned}
& \frac{1}{n^2} \sum_{\boldsymbol{i}} \sigma^4 \{2|\mathcal{B}_n(\boldsymbol{i})| + \sum_{\boldsymbol{x}_{s_1} \neq \boldsymbol{x}_{s_2} \in \mathcal{B}_n(\boldsymbol{i})} 2\psi^2(y(\boldsymbol{x}_{t_1}), y(\boldsymbol{x}_{t_2})) \\
& + \sum_{\boldsymbol{x}_{s_1} \neq \boldsymbol{x}_{s_2} \neq \boldsymbol{x}_{t_1} \in \mathcal{B}_n(\boldsymbol{i})} 2\psi(y(\boldsymbol{x}_{t_1}), y(\boldsymbol{x}_{s_1}))\psi(y(\boldsymbol{x}_{t_1}), y(\boldsymbol{x}_{s_2})) \\
& + \sum_{\boldsymbol{x}_{s_1} \neq \boldsymbol{x}_{s_2} \neq \boldsymbol{x}_{t_1} \neq \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{i})} \psi(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_1}))\psi(y(\boldsymbol{x}_{s_2}), y(\boldsymbol{x}_{t_2})) \\
& + \psi(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_2}))\psi(y(\boldsymbol{x}_{s_2}), y(\boldsymbol{x}_{t_1}))\} \\
& + \frac{1}{n^2} \sum_{\boldsymbol{i} \neq \boldsymbol{j}} \sigma^4 \{ \sum_{\boldsymbol{x}_{s_1} \in \mathcal{B}_n(\boldsymbol{i}) \boldsymbol{x}_{t_1} \in \mathcal{B}_n(\boldsymbol{j})} 2\psi^2(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_1})) \\
& + \sum_{\boldsymbol{x}_{s_1} \neq \boldsymbol{x}_{s_2} \in \mathcal{B}_n(\boldsymbol{i}) \boldsymbol{x}_{t_1} \in \mathcal{B}_n(\boldsymbol{j})} 2\psi(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_1}))\psi(y(\boldsymbol{x}_{s_2}), y(\boldsymbol{x}_{t_2})) \\
& + \sum_{\boldsymbol{x}_{s_1} \in \mathcal{B}_n(\boldsymbol{i}), \boldsymbol{x}_{t_1} \neq \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{j})} 2\psi(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_1}))\psi(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_2})) \\
& + \sum_{\boldsymbol{x}_{s_1} \neq \boldsymbol{x}_{s_2} \in \mathcal{B}_n(\boldsymbol{i}), \boldsymbol{x}_{t_1} \neq \boldsymbol{x}_{t_2} \in \mathcal{B}_n(\boldsymbol{j})} \psi(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_1}))\psi(y(\boldsymbol{x}_{s_2}), y(\boldsymbol{x}_{t_2})) \\
& + \psi(y(\boldsymbol{x}_{s_1}), y(\boldsymbol{x}_{t_2}))\psi(y(\boldsymbol{x}_{s_2}), y(\boldsymbol{x}_{t_1}))\} \\
= \; & O(\frac{1}{n} + \frac{m^d}{n}) \to 0,
\end{aligned}
$$

where $m = o(n^{1/d})$. In addition, the expectation of A_n is

$$\begin{aligned}
& \boldsymbol{E}(A_n) - \tau_n^2 = \frac{n}{m^{2d}} \sum_{i_1,\ldots,i_d} \boldsymbol{E}(\bar{y}_{i_1,\ldots,i_d} - \mu)^2 - \frac{1}{n}\sum_{s,t=1}^{n} \boldsymbol{Cov}(y_s, y_t) \\
= \; & \frac{1}{n}\sum_{\boldsymbol{i}} \{ \sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{E}(y_s - \mu)^2 + \sum_{\boldsymbol{x}_s \neq \boldsymbol{x}_t \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{E}(y_s - \mu)(y_t - \mu) \} \\
& -\sigma^2 - \frac{1}{n}\sum_{s \neq t} \boldsymbol{E}(y_s - \mu)(y_t - \mu) \\
= \; & \frac{1}{n}\sum_{\boldsymbol{i}} \sum_{\boldsymbol{x}_s \neq \boldsymbol{x}_t \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{E}(y_s - \mu)(y_t - \mu)\} - \frac{1}{n}\sum_{s \neq t} \boldsymbol{E}(y_s - \mu)(y_t - \mu) \\
= \; & \frac{1}{n}\sum_{\boldsymbol{i} \neq \boldsymbol{j}} \sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}), \boldsymbol{x}_t \in \mathcal{B}_n(\boldsymbol{j})} \sigma^2 \psi(y(\boldsymbol{x}_s), y(\boldsymbol{x}_t)) = o(1).
\end{aligned}$$

Thus, we have $A_n - \tau_n^2 \xrightarrow{\mathrm{P}} 0$. □

Denote ${\tau_N^*}^2 = \boldsymbol{Cov}^*_{N,\omega}(\bar{y}^*_N, \bar{y}^*_N)$ as the bootstrap variance under $P^*_{N,\omega}$. The following lemma reflects the convergence rate difference between $\bar{y}_n$ and $\bar{y}^*_N$ with different sample sizes.

Lemma 2.3 *Assume (A.1)- (A.2), then*

$$n{\tau_N^*}^2/m^{d-1} - \tau_n^2 \xrightarrow{\mathrm{P}} 0.$$

Proof: Based on the definition of $n{\tau_N^*}^2/m^{d-1}$, we have

$$n{\tau_N^*}^2/m^{d-1} = \; \tfrac{n}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_1^*}) + 2\tfrac{n(m-1)}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_2^*}).$$

we compute the two terms on the right hand side separately as follows.

First, calculate $\frac{n}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_1^*})$:

$$\begin{aligned}
& \frac{n}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_1^*}) = \frac{n}{m^d}\boldsymbol{E}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*} - \bar{y}_n)^2 \\
= & \frac{n}{m^d}\sum_{i_1,\ldots,i_d}\frac{1}{m^d}(\bar{y}_{i_1,\ldots,i_d} - \bar{y}_n)^2 = \frac{n}{m^{2d}}\sum_{i_1,\ldots,i_d}(\bar{y}_{i_1,\ldots,i_d} - \mu + \mu - \bar{y}_n)^2 \\
= & \frac{n}{m^{2d}}\sum_{i_1,\ldots,i_d}\{(\bar{y}_{i_1,\ldots,i_d} - \mu)^2 - 2(\bar{y}_n - \mu)(\bar{y}_{i_1,\ldots,i_d} - \mu) + (\bar{y}_n - \mu)^2\} \\
= & \frac{n}{m^{2d}}\sum_{i_1,\ldots,i_d}(\bar{y}_{i_1,\ldots,i_d} - \mu)^2 - \frac{n}{m^d}(\bar{y}_n - \mu)^2 \\
= & A_n - B_n.
\end{aligned}$$

By Lemma 2.2, we have $A_n - \tau_n^2 \xrightarrow{\mathrm{P}} 0$. For $B_n = \frac{n}{m^d}(\bar{y}_n - \mu)^2$, by the central limit theorem for $\bar{y}_n$, we have $B_n \xrightarrow{\mathrm{P}} 0$.

Finally, it suffices to show that $\frac{n(m-1)}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_2^*})$ converges to 0 in probability under P. The following double summation $\sum_{i_1,\ldots,j_d,j_1,\ldots,j_d}$ are taken over $\boldsymbol{i} = (i_1, \ldots, i_d)$ and $\boldsymbol{j} = (j_1, \ldots, j_d)$ such that $\mathcal{B}_n(\boldsymbol{i})$ and $\mathcal{B}_n(\boldsymbol{j})$ are not equal and are selected together.

$$\begin{aligned}
& \frac{n(m-1)}{m^d}\boldsymbol{Cov}^*_{N,\omega}(\bar{y}_{\boldsymbol{i}_1^*}, \bar{y}_{\boldsymbol{i}_2^*}) \\
= & \frac{n(m-1)}{2m^d}\frac{1}{\binom{m^d}{2} - m^{d-1}d\binom{m}{2}}\sum_{\boldsymbol{i}\neq\boldsymbol{j}}(\bar{y}_{\boldsymbol{i}} - \bar{y}_n)(\bar{y}_{\boldsymbol{j}} - \bar{y}_n) \\
= & \frac{n(m-1)}{m^{2d}}\frac{1}{m^d - 1 - d(m-1)}\sum_{\boldsymbol{i}\neq\boldsymbol{j}}(\bar{y}_{\boldsymbol{i}} - \mu)(\bar{y}_{\boldsymbol{j}} - \mu) \\
& -2\frac{n(m-1)}{m^{2d}}\frac{m^{d-1}}{m^d - 1 - d(m-1)}\sum_{\boldsymbol{i}}(\bar{y}_{\boldsymbol{i}} - \mu)(\bar{y}_n - \mu) + \frac{n(m-1)}{m^d}(\bar{y}_n - \mu)^2 \\
= & \frac{n(m-1)}{m^{2d}}\frac{1}{m^d - 1 - d(m-1)}\sum_{\boldsymbol{i}\neq\boldsymbol{j}}(\bar{y}_{\boldsymbol{i}} - \mu)(\bar{y}_{\boldsymbol{j}} - \mu) \\
& +\frac{n(m-1)}{m^d}[1 - \frac{2m^d}{m\{m^d - 1 - d(m-1)\}}](\bar{y}_n - \mu)^2 \\
= & C_n + D_n.
\end{aligned}$$

Similar to A_n and B_n, we can show that $C_n \xrightarrow{\mathrm{P}} 0$ and $D_n \xrightarrow{\mathrm{P}} 0$. The result follows immediately. □

Based on the results from Lemmas 2.1-2.3, we develop the asymptotic consistency

of the LHD-based block bootstrap mean in the following theorem.

Theorem 2.1 *Under (A.1)-(A.2), if* $m \to \infty$ *and* $m = o(n^{1/d})$*, then*

$$\sup_x |P^*_{N,\omega}(\sqrt{n/m^{d-1}}(\bar{y}^*_N - \bar{y}_n)/\tau_n \le x) - P(\sqrt{n}(\bar{y}_n - \mu)/\tau_n \le x)| \xrightarrow{\mathrm{P}} 0,$$

when $n \longrightarrow \infty$.

Proof: It suffices to show that (1) $\boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N) = \bar{y}_n$; (2) $n{\tau^*_N}^2/m^{d-1} - \tau_n^2 \xrightarrow{\mathrm{P}} 0$; and (3) $\sup_x |P^*_{N,\omega}((\bar{y}^*_N - \boldsymbol{E}^*_{N,\omega}(\bar{y}^*_N))/\tau^*_N \le x) - \Phi(x)| \xrightarrow{\mathrm{P}} 0$, where $\Phi(\cdot)$ denotes standard normal distribution function and ${\tau^*_N}^2 = \boldsymbol{Cov}^*_{N,\omega}(\bar{y}^*_N, \bar{y}^*_N)$.

Lemma 2.1 and Lemma 2.3 proved before imply the results in (1) and (2). Note that $\bar{y}^*_N = \frac{1}{m}\sum_{j=1}^m \bar{y}_{\boldsymbol{i}^*_j}$ and $(\bar{y}_{\boldsymbol{i}^*_1}, \ldots, \bar{y}_{\boldsymbol{i}^*_m})$ follows Latin Hypercube sampling distribution. According to Loh (1996), we have the Berry-Essen type of bound for Latin Hypercube sampling

$$\sup_x |P^*_{N,\omega}((\bar{y}^*_N - \bar{y}_n)/\tau^*_N \le x) - \Phi(x)| \le c^* m^{-1/2},$$

where c^* is a constant that depends only on d, given $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}^*_1}\|^3 < \infty$. So we only need to show that $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}^*_1}\|^3$ is bounded uniformly in probability under P. Since $\boldsymbol{E}^*_{N,\omega}\|\bar{y}_{\boldsymbol{i}_1}\|^3 = \frac{1}{m^d}\sum_{\boldsymbol{i}} \bar{y}^3_{\boldsymbol{i}}$ and according to Minkowski's inequality, it follows that

$$\frac{1}{m^d}\sum_{\boldsymbol{i}} \boldsymbol{E}\{\bar{y}^3_{\boldsymbol{i}}\} \le \frac{1}{m^d}\sum_{\boldsymbol{i}} \frac{1}{|\mathcal{B}_n(\boldsymbol{i})|^3}\{\sum_{\boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i})} \boldsymbol{E}(y_s)\}^3 < \infty.$$

□

2.3.3 Consistency of MLEs using LHD-based block bootstrap

To investigate the asymptotic properties of the estimators from LHD-based block bootstrap, we first decompose the likelihood function by blocks. For each block, denote

$\boldsymbol{y_i} = (y_s(\boldsymbol{x}_s), \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}))$, $\boldsymbol{X_i} = (\boldsymbol{x}_s, \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}))^T$, $R_{\boldsymbol{i},\boldsymbol{j}}(\boldsymbol{\theta}) = \big[\psi(y(\boldsymbol{x}_s), y(\boldsymbol{x}_t); \boldsymbol{\theta}), \boldsymbol{x}_s \in \mathcal{B}_n(\boldsymbol{i}), \boldsymbol{x}_t \in \mathcal{B}_n(\boldsymbol{j})\big]$ and $\boldsymbol{z_i} = R_{\boldsymbol{i},\boldsymbol{i}}^{-1/2}(\boldsymbol{\theta})(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})$. Then, we can rewrite the normalized log-likelihood function as

$$
\begin{aligned}
& Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) = \frac{1}{n}\ell(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) \\
= \; & -(2n\sigma^2)^{-1} \sum_{\boldsymbol{i}=(i_1,\ldots,i_d)} (\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta})^T R_{\boldsymbol{i}}^{-1}(\boldsymbol{\theta})(\boldsymbol{y_i} - \boldsymbol{X_i}\boldsymbol{\beta}) \\
& +(2n)^{-1} \sum_{\boldsymbol{i}=(i_1,\ldots,i_d)} \log(|R_{\boldsymbol{i},\boldsymbol{i}}(\boldsymbol{\theta})|) - 2^{-1}\log(\sigma^2) + n^{-1} r_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) \\
= \; & -(2n\sigma^2)^{-1} \sum_{s=1}^{n} z_s^2 - (2n)^{-1} \sum_{s=1}^{n} \log(\lambda_s) \\
& -(2n)^{-1} \sum_{s=1}^{n} \log(\sigma^2) + n^{-1} r_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) \\
= \; & n^{-1} \sum_{s=1}^{n} q_s(z_s, \boldsymbol{\phi}) + n^{-1} r_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}) \\
= \; & n^{-1} \sum_{s=1}^{n} q_s(\omega, \boldsymbol{\phi}) + n^{-1} r_n(\omega, \boldsymbol{\phi}),
\end{aligned}
$$

where $\{\lambda_s, s = 1, \ldots, n\} = \{\text{eigenvalues of } |R_{\boldsymbol{i},\boldsymbol{i}}(\boldsymbol{\theta})|, \boldsymbol{i} = (i_1, \ldots, i_d)\}$ with $(i_1, \ldots, i_d)$ in lexicographical order and eigenvalues from the largest to the smallest. Note that $r_n(\omega, \boldsymbol{\phi})$ contains all terms involving the off block-diagonal terms. Define $D_n(\boldsymbol{\theta}) = \text{diag}(R_{\boldsymbol{i},\boldsymbol{i}}(\boldsymbol{\theta}))$ and $E_n(\boldsymbol{\theta}) = R_n(\boldsymbol{\theta}) - D_n(\boldsymbol{\theta})$. Assuming that $E_n(\boldsymbol{\theta}) = U_n(\boldsymbol{\theta})U_n^T(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
r_n(\omega, \boldsymbol{\phi}) \; = \; & \frac{1}{2\sigma^2(1+g)} (\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T D_n^{-1}(\boldsymbol{\theta}) E_n(\boldsymbol{\theta}) D_n^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}) \\
& +\frac{1}{2}\log|I_n + U_n^T(\boldsymbol{\theta}) D_n^{-1}(\boldsymbol{\theta}) U_n(\boldsymbol{\theta})|,
\end{aligned}
$$

where $g = \text{trace}(E_n(\boldsymbol{\theta})D_n^{-1}(\boldsymbol{\theta}))$.

Then maximum likelihood estimator is given by

$$
\hat{\boldsymbol{\phi}}_n = \arg\max_{\boldsymbol{\phi}} Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi}).
$$

Analogue to the decomposition for $Q_n(\boldsymbol{X}_n, \boldsymbol{y}_n, \boldsymbol{\phi})$, the log-likelihood function for LHD-based block bootstrap samples can be written as

$$Q_N^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi}) \;=\; N^{-1}\sum_{s=1}^{N} q_s^*(\cdot, \omega, \boldsymbol{\phi}) + N^{-1} r_N^*(\cdot, \omega, \boldsymbol{\phi}), \tag{2.2}$$

where $r_N^*(\cdot, \omega, \boldsymbol{\phi})$ contains all terms involving the off block-diagonal terms with bootstrapped samples. Specifically,

$$\begin{aligned} r_N^*(\cdot, \omega, \boldsymbol{\phi}) \;=\; & \frac{1}{2\sigma^2(1+g^*)}(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta})^T {D_N^*}^{-1}(\boldsymbol{\theta}) E_N^*(\boldsymbol{\theta}) {D_N^*}^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_N^* - \boldsymbol{X}_N^*\boldsymbol{\beta}) \\ & + \frac{1}{2}\log|I_N + {U_N^*}^T(\boldsymbol{\theta}) {D_N^*}^{-1}(\boldsymbol{\theta}) U_N^*(\boldsymbol{\theta})|, \end{aligned}$$

where $D_N^*(\boldsymbol{\theta}) = \mathrm{diag}(R_{\boldsymbol{i}_j^*, \boldsymbol{i}_j^*}(\boldsymbol{\theta}), j = 1, \ldots, m)$ and $E_N^*(\boldsymbol{\theta}) = R_N^*(\boldsymbol{\theta}) - D_N^*(\boldsymbol{\theta})$ with $E_N^*(\boldsymbol{\theta}) = U_N^*(\boldsymbol{\theta}) {U_N^*}^T(\boldsymbol{\theta})$; $g^* = \mathrm{trace}(E_N^*(\boldsymbol{\theta}) {D_N^*}^{-1}(\boldsymbol{\theta}))$. The bootstrapped version of $\hat{\boldsymbol{\phi}}_n$ is

$$\hat{\boldsymbol{\phi}}_N^* = \arg\max_{\boldsymbol{\phi}} Q_N^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi}).$$

The following assumptions are required for studying the convergence of the bootstrap estimator $\hat{\boldsymbol{\phi}}_N^*$.

(A.3) $\lim_{n\to\infty} \sup_{\boldsymbol{\theta}} \lambda_{\max}(E_n(\boldsymbol{\theta})) = 0$, when the block space $b = l/m \to \infty$.

(A.4) $\forall\ \boldsymbol{\phi}_1,\ \boldsymbol{\phi}_2 \in \Theta,\ |q_s(\cdot, \boldsymbol{\phi}_1) - q_s(\cdot, \boldsymbol{\phi}_2)| \le L_s|\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2| a.s.P$, where L_s is Lipschitz constant and $\sup_n\{n^{-1}\sum_{s=1}^n \boldsymbol{E}L_s\} = O(1)$.

(A.5) Θ is compact.

(A.6) The functions $q_s(\omega, \boldsymbol{\phi})$ and $r_n(\omega, \boldsymbol{\phi})$ are such that $q_s(\cdot, \boldsymbol{\phi})$ and $r_n(\cdot, \boldsymbol{\phi})$ are measurable for all $\boldsymbol{\phi} \in \Theta$, a compact subset of R^p. In addition, $q_s(\omega, \cdot) : \Theta \longrightarrow R$ and $r_n(\omega, \cdot) : \Theta \longrightarrow R$ are continuous on Θ a.s.-P, $s = 1, \cdots, n$, $n = 1, 2, \cdots$.

(A.7) $Q_n(\omega, \cdot) : \Theta \to R$ is continuously differentiable of order 2 on Θ a.s. P.

(A.8) There exists a sequence $H_n(\boldsymbol{\phi}) : \Theta \to R^{p\times p}$ such that $\nabla^2 Q_n(\cdot, \boldsymbol{\phi}) - H_n(\boldsymbol{\phi}) \xrightarrow{\text{P}} 0$ as $n \to \infty$ uniformly on Θ.

(A.9) $H_n(\boldsymbol{\phi}^0)$ is $O(1)$ and uniformly nonsingular.

(A.10) $Q_N^*(\lambda, \omega, \cdot) : \Theta \to R$ are continuously differentiable of order 2 on Θ a.s. P. Also, function $\nabla^2 Q_n(\omega, \boldsymbol{\phi})$ is such that $\nabla^2 Q_n(\cdot, \boldsymbol{\phi})$ is measurable for all $\boldsymbol{\phi} \in \Theta$ and $\nabla^2 Q_n(\omega, \cdot) : \Theta \to R$ is continuous on Θ a.s.-P.

(A.11) $\forall\ \boldsymbol{\phi}_1,\ \boldsymbol{\phi}_2 \in \Theta, |\nabla^2 Q_n(\cdot, \boldsymbol{\phi}_1) - \nabla^2 Q_n(\cdot, \boldsymbol{\phi}_2)| \leq M_s |\boldsymbol{\phi}_1 - \boldsymbol{\phi}_2| a.s.P$, where M_s is Lipschitz constant and $\sup_n\{n^{-1}\sum_{s=1}^n \boldsymbol{E}M_s\} = O(1)$.

Assumption (A.3) controls the correlation between bootstrapped blocks. (A.4) and (A.5) are required in order to achieve uniform convergency of the bootstrapped likelihood function. (A.6) ensures the existence of the estimators. (A.7)-(A.9) are regularity conditions for standard MLE consistency in GP models, which is analogue to the conditions in Mardia and Marshall (1984). (A.10) ensures the existence of covariance matrix. (A.11) is the Global Lipschitz condition for $\nabla^2 Q_n(\omega, \cdot)$ which guarantees the convergence of LHD-based block bootstrap covariance matrix.

Theoretical properties of the LHD-based block bootstrap likelihood function (3.3) are established in the following two lemmas, which leads to a proof of convergence properties of the bootstrap estimator $\hat{\boldsymbol{\phi}}_N^*$. Lemma 2.4 below first established the pointwise weak law of large numbers for the LHD-based block bootstrap likelihood functions.

Lemma 2.4 *(Pointwise Weak Law of Large Numbers) Under (A.1)-(A.3), for each*

$\boldsymbol{\phi} \in \Theta$,

$$\lim_{n\to\infty} P\Big[P^*_{N,\omega}\big(|N^{-1}\sum_{s=1}^{N} q^*_s(\cdot,\omega,\boldsymbol{\phi}) + N^{-1}r^*_N(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi}) - n^{-1}r_n(\omega,\boldsymbol{\phi})| > \delta\big) > \xi\Big] = 0.$$

Proof: Rewrite the bootstrapped likelihood function as

$$\begin{aligned} & N^{-1}\sum_{s=1}^{N} q^*_s(\cdot,\omega,\boldsymbol{\phi}) + N^{-1}r^*_N(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi}) - n^{-1}r_n(\omega,\boldsymbol{\phi}) \\ = \;& N^{-1}\sum_{s=1}^{N}\{q^*_s(\cdot,\omega,\boldsymbol{\phi}) - \boldsymbol{E}^* q^*_s(\cdot,\omega,\boldsymbol{\phi})\} + \{N^{-1}\sum_{s=1}^{N}\boldsymbol{E}^* q^*_s(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi})\} \\ & + N^{-1}r^*_N(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}r_n(\omega,\boldsymbol{\phi}) \\ = \;& I_1 + I_2 + I_3. \end{aligned}$$

By Lemma 2.3, $I_2 = 0$. With respect to I_3, we will show that $n^{-1}r_n(\omega,\boldsymbol{\phi}) \to 0$ in P and $N^{-1}r^*_N(\cdot,\omega,\boldsymbol{\phi}) \to 0$, prob-$P^*_{N,\omega}$ prob-P. For notation simplicity, we miss $\boldsymbol{\theta}$ in the following computation. The expectation and variance of $n^{-1}r_n(\omega,\boldsymbol{\phi})$ are:

$$\begin{aligned} & |\boldsymbol{E}\{n^{-1}r_n(\omega,\boldsymbol{\phi})\}| \\ \le \;& |\boldsymbol{E}\{\frac{1}{2n\sigma^2(1+g)}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})^T D_n^{-1}E_n D_n^{-1}(\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta})\}| + \big|\log|I_n + U_n^T D_n^{-1}U_n|\big| \\ \le \;& \frac{1}{2n\sigma^2(1+g)}\lambda_{\max}(E_n)\lambda_{\max}(D_n^{-1})\boldsymbol{E}\{\|\boldsymbol{y}_n - \boldsymbol{X}_n\boldsymbol{\beta}\|_2^2\}|\log(|I_n| + |U_n^T D_n^{-1}U_n|)| \\ \le \;& \frac{1}{2n\sigma^2(1+g)}\lambda_{\max}(E_n)\lambda_{\max}(D_n^{-1}) + |\log\{1 + \lambda^n_{\max}(E_n)|D_n^{-1}\}| \\ = \;& o(1) \end{aligned}$$

and

$$\begin{aligned}
& \boldsymbol{Var}(n^{-1}r_n(\omega,\boldsymbol{\phi})) = \boldsymbol{Var}\{\frac{1}{2n\sigma^2(1+g)}(\boldsymbol{y}_n-\boldsymbol{X}_n\boldsymbol{\beta})^T D_n^{-1}E_nD_n^{-1}(\boldsymbol{y}_n-\boldsymbol{X}_n\boldsymbol{\beta})\} \\
\leq\ & \frac{1}{4(1+g)^2\sigma^4n^2}\boldsymbol{Var}\{\sum_{i=1}^n(\sum_{j=1}^n u_{ij}\varepsilon_j)^2\} \\
\leq\ & \frac{1}{4(1+g)^2\sigma^4n^2}\sum_{i=1}^n\boldsymbol{Var}\{(\sum_{j=1}^n u_{ij}^2)(\sum_{j=1}^n\varepsilon_j^2)\} \\
\leq\ & \frac{c_n}{4(1+g)^2\sigma^4n^2}\sum_{i=1}^n\sum_{j=1}^n\boldsymbol{Var}(\varepsilon_j^2) \\
=\ & o(1),
\end{aligned}$$

where ε_j is the i^{th} entry of $D_n^{-1}(\boldsymbol{y}_n-\boldsymbol{X}_n\boldsymbol{\beta})$ and $\boldsymbol{u}_i=(u_{ij})$ is the i^{th} row of U_n; $c_n=\max_i\{\sum_{j=1}^n u_{ij}^2\}$.

In addition, as $\lambda_{\max}(E_N^*)\leq\lambda_{\max}(E_n)$ and $\lambda_{\max}({D_N^*}^{-1})\leq\lambda_{\max}(D_n^{-1})$, we have

$$\begin{aligned}
& \frac{1}{2\sigma^2(1+g^*)}(\boldsymbol{y}_N^*-\boldsymbol{X}_N^*\boldsymbol{\beta})^T{D_N^*}^{-1}E_N^*{D_N^*}^{-1}(\boldsymbol{y}_N^*-\boldsymbol{X}_N^*\boldsymbol{\beta}) \\
\leq\ & \frac{1}{2\sigma^2}\lambda_{\max}(E_n)\lambda_{\max}(D_n^{-1})\|\boldsymbol{y}_N^*-\boldsymbol{X}_N^*\boldsymbol{\beta}\|_2^2.
\end{aligned}$$

According to Theorem 2.1, we have $N^{-1}\|\boldsymbol{y}_N^*-\boldsymbol{X}_N^*\boldsymbol{\beta}\|_2^2-n^{-1}\|\boldsymbol{y}_n-\boldsymbol{X}_n\boldsymbol{\beta}\|_2^2\to 0$ prob-$P_{N,\omega}^*$ prob-P. Similarly, we can bound $\log|I_N+{U_N^*}^T{D_N^*}^{-1}U_N^*|$. As $\lambda_{\max}(E_n)\to 0$, we have $\frac{1}{N}r_N^*(\cdot,\omega,\boldsymbol{\phi})\to 0$, prob-$P_{N,\omega}^*$ prob-P.

So when n is sufficiently large, we only need to show that $\lim_{n\to\infty}P\big[P_{N,\omega}^*(|I_1|>\delta)>\xi\big]=0$. By Chebyshev's inequality,

$$P_{N,\omega}^*(|I_1|>\delta)\leq\frac{1}{\delta^2}\boldsymbol{Var}_{N,\omega}^*(\bar{q}_N^*(\cdot,\omega,\boldsymbol{\phi})).$$

By Lemma 2.1, $r^{-1}\boldsymbol{Var}_{N,\omega}^*(\bar{q}_N^*(\cdot,\omega,\boldsymbol{\phi}))=O_p(1)$, together with the fact that $N=$

n/m^{d-1}

$$\begin{aligned} & P\big[P^*_{N,\omega}(|I_1| > \delta) > \xi\big] \\ \leq\ & P\big[\frac{n}{m^{d-1}}\frac{1}{\delta^2}\boldsymbol{Var}^*_{N,\omega}(\bar{q}^*_N(\cdot,\omega,\boldsymbol{\phi})) > \xi\frac{n}{m^{d-1}}\big] \\ =\ & O(m^{2d-2}/n^2) \to 0. \end{aligned}$$

□

The next lemma further extends Lemma 2.4 to the uniform weak law of large numbers for the LHD-based block bootstrap likelihood functions.

Lemma 2.5 *(Uniform Weak Law of Large Numbers) Under (A.1)-(A.5),* $\forall\ \delta, \xi > 0$,

$$\begin{aligned} \lim_{n\to\infty} P\Big[P^*_{N,\omega}(\sup_{\boldsymbol{\phi}\in\Theta} |N^{-1}\sum_{s=1}^{N} q^*_s(\cdot,\omega,\boldsymbol{\phi}) + N^{-1}r^*_N(\cdot,\omega,\boldsymbol{\phi}) \\ -n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi}) - n^{-1}r_n(\omega,\boldsymbol{\phi})| > \delta) > \xi\Big] = 0. \end{aligned}$$

Proof: By Lemma 2.4, $|n^{-1}r_n(\omega,\boldsymbol{\phi}) - N^{-1}r^*_N(\cdot,\omega,\boldsymbol{\phi})|$ can be arbitrarily small as n is large enough uniformly over Θ. We only need to show that

$$\lim_{n\to\infty} P\big[P^*_{N,\omega}(\sup_{\boldsymbol{\phi}\in\Theta} |N^{-1}\sum_{s=1}^{N} q^*_s(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi})| > \delta) > \xi\big] = 0.$$

Given $\epsilon > 0$ that will be selected later, let $\{\eta(\boldsymbol{\phi}_j,\epsilon), j = 1,\ldots,K\}$ be a finite cover of Θ, where $\eta(\boldsymbol{\phi}_i,\epsilon) = \{\boldsymbol{\phi}\in\Theta : |\boldsymbol{\phi}-\boldsymbol{\phi}_j| < \epsilon\}$. Then

$$\begin{aligned} & \sup_{\boldsymbol{\phi}} |N^{-1}\sum_{s=1}^{N} q^*_s(\cdot,\omega,\boldsymbol{\phi}) - n^{-1}\sum_{s=1}^{n} q_s(\omega,\boldsymbol{\phi})| \\ =\ & \max_{j=1}^{K} \sup_{\boldsymbol{\phi}\in\eta(\boldsymbol{\phi}_j,\epsilon)} |\bar{q}^*_N(\cdot,\omega,\boldsymbol{\phi}) - \bar{q}_n(\omega,\boldsymbol{\phi})|. \end{aligned}$$

It follows that $\forall\ \delta > 0$ with fixed ω,

$$\begin{aligned} & P_{N,\omega}\big(\sup_{\boldsymbol{\phi}\in\Theta} |\bar{q}^*_N(\cdot,\omega,\boldsymbol{\phi}) - \bar{q}_n(\omega,\boldsymbol{\phi})| > \delta\big) \\ \leq\ & \sum_{j=1}^{K} P_{N,\omega}\big(\sup_{\boldsymbol{\phi}\in\eta(\boldsymbol{\phi}_j,\epsilon)} |\bar{q}^*_N(\cdot,\omega,\boldsymbol{\phi}) - \bar{q}_n(\omega,\boldsymbol{\phi})| > \delta\big). \end{aligned}$$

For $\forall\ \boldsymbol{\phi} \in \eta(\boldsymbol{\phi}_j, \epsilon)$, by Global Lipschitz condition,

$$\begin{aligned}
& |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi})| \\
\leq\ & |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}_j) - \bar{q}_n(\omega, \boldsymbol{\phi}_j)| + |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}_j) - \bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi})| \\
& + |\bar{q}_n(\omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi}_j)| \\
\leq\ & |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}_j) - \bar{q}_n(\omega, \boldsymbol{\phi}_i)| + N^{-1} \sum_{s=1}^{N} L_s^* \epsilon + n^{-1} \sum_{s=1}^{n} L_s \epsilon,
\end{aligned}$$

where L_s^* is the bootstrapped Lispchitz coefficient.

By Markov inequality and $\sup_n \{n^{-1} \sum_{s=1}^{n} \boldsymbol{E} L_s\} = O(1)$,

$$P(n^{-1} \sum_{s=1}^{n} L_s > \delta/3) \leq 3\epsilon\Delta/\delta \leq \xi/3,$$

where Δ is a large constant. If we choose $\epsilon < \xi\delta/(9\Delta)$, we have

$$\begin{aligned}
& P\big[P_{N,\omega}^*(\sup_{\boldsymbol{\phi} \in \eta(\boldsymbol{\phi}_j, \epsilon)} |\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) - \bar{q}_n(\omega, \boldsymbol{\phi})| > \delta) > \xi\big] \\
\leq\ & P\big[P_{N,\omega}^*(|\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}_j) - \bar{q}_n(\omega, \boldsymbol{\phi}_j)| > \delta) > \xi/3\big] \\
& + P\big[P_{N,\omega}^*(N^{-1} \sum_{s=1}^{N} L_s^* \epsilon > \delta/3) > \xi/3\big] + P[n^{-1} \sum_{s=1}^{n} L_s \epsilon > \delta/3] \\
=\ & K_1 + K_2 + K_3.
\end{aligned}$$

According to Lemma 2.4, $K_1 \leq \xi/3$ when n is large enough. By Markov's inequality, we have

$$P_{N,\omega}^*(N^{-1} \sum_{s=1}^{N} L_s^* \epsilon > \delta/3) \leq N^{-1} \sum_{s=1}^{N} \boldsymbol{E}^* L_s^*/(\delta/3\epsilon) = n^{-1} \sum_{s=1}^{n} L_s/(\delta/3\epsilon).$$

The last equality holds because of Lemma 2.1. Thus, $K_2 < \xi/3$ as well as K_3. □

In the following theorem, we show that $\hat{\boldsymbol{\phi}}_N^*$ converges in probability to $\hat{\boldsymbol{\phi}}_n$, conditional on all samples with probability tending to one. For any LHD-based block bootstrapped statistic $\hat{T}_N^*$, we write $\hat{T}_N^* \to 0\ \ prob - P_{N,\omega}^*, prob - P$ if for any $\epsilon > 0$ and any $\delta > 0$, $\lim_{n\to\infty} P\{P_{N,\omega}^*(|\hat{T}_N^*| > \epsilon| > \delta)\} = 0$.

Theorem 2.2 *Under (A.1)- (A.6), if* $m = o(n^{1/d})$ *and* $m \to \infty$, *then*

$$\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n \to 0 \quad prob - P_{N,\omega}^*, prob - P.$$

Proof: With the full preparation of the likelihood convergence developed in Lemmas 2.4 and 2.5, the convergence of bootstrap parameter estimation follows immediately given the existence of $\hat{\boldsymbol{\phi}}_n$ and $\hat{\boldsymbol{\phi}}_N^*$.

Denote $\bar{q}_N^*(\cdot, \omega, \boldsymbol{\phi}) = N^{-1} \sum_{i=1}^N q_i^*(\cdot, \omega, \boldsymbol{\phi})$ and $\bar{q}_n(\omega, \boldsymbol{\phi}) = n^{-1} \sum_{i=1}^n q_i(\omega, \boldsymbol{\phi})$. By (A.6), $q_s^*(\cdot, \omega, \cdot) : \Lambda \times \Theta \to R$ and $r_N^*(\cdot, \omega, \cdot) : \Lambda \times \Theta \to R$ are measurable-$\mathcal{G}$ for each $\boldsymbol{\phi} \in \Theta$. In addition, $q_s^*(\lambda, \omega, \cdot)$ and $r_N^*(\lambda, \omega, \cdot)$ are continuous on Θ for all λ. Thus, we have $\hat{\boldsymbol{\phi}}_N^*(\cdot, \omega)$ exists as a measurable-$\mathcal{G}$ function by Jennrich (1969).

Following the procedure in Goncalves and White Gonçalves and White (2004), for any subsequence $\{n'\}$, given that $\hat{\boldsymbol{\phi}}_{n'}$ is identifiable and unique, there exists a further subsequence $\{n''\}$ such that $\hat{\boldsymbol{\phi}}_{n''}$ is identifiably unique with respect to $\{Q_{n''}\}$ for all $\omega \in F$ in some $F \in \mathcal{F}$ with $P(F) = 1$. By condition (A.6), there exists $G \in \mathcal{F}$ with $P(G) = 1$ such that for all $\omega \in G$, $\{Q_{N''}^*(\cdot, \omega, \boldsymbol{\phi})\}$ (N'' is corresponding bootstrapped sample size of n'') is a sequence of random function on $(\Lambda, \mathcal{G}, P_{N,\omega}^*)$ continuous on Θ for all $\lambda \in \Lambda$. Hence, by White White (1996), for fixed $\omega \in G$, there exists $\hat{\boldsymbol{\phi}}_{N''}^*(\cdot, \omega) : \Lambda \to \Theta$ measurable-$\mathcal{G}$ and $\hat{\boldsymbol{\phi}}_{N''}^*(\cdot, \omega) = \arg\max_{\boldsymbol{\phi}} Q_{N''}^*(\cdot, \omega, \boldsymbol{\phi})$. By the uniform weak law of large numbers for $Q_N^*(\boldsymbol{X}_N^*, \boldsymbol{y}_N^*, \boldsymbol{\phi})$ obtained from Lemma 2.5, we have $Q_{N''}^*(\cdot, \omega, \boldsymbol{\phi}) - Q_{n''}(\omega, \boldsymbol{\phi}) \to 0$ as $n'' \to \infty$ $prob - P_{N,\omega}^*$ $prob - P$ uniformly on Θ. Hence, there exists a further subsequence $\{n'''\}$ such that $Q_{N'''}^*(\cdot, \omega, \boldsymbol{\phi}) - Q_{n'''}(\omega, \boldsymbol{\phi}) \to 0$ as $n'' \to \infty$ $prob - P_{N,\omega}^*$ $prob - P$ for all ω in some $H \in \mathcal{F}$ with $P(H) = 1$. Choose $\omega \in F \cap G \cap H$, by White (1996), we have $\hat{\boldsymbol{\phi}}_{N'''}^* - \hat{\boldsymbol{\phi}}_{n'''} \to 0$ as $n''' \to \infty$ $prob - P_{N,\omega}^*$ $prob - P$. Since this is true for any subsequence $\{n'\}$, we have $P(F \cap G \cap H) = 1$. □

Under some regularity conditions given by Mardia and Marshall Mardia and Marshall (1984), it can be shown that the original estimator $\hat{\boldsymbol{\phi}}_n$ based on full data is consistent and it converges to $\boldsymbol{\phi}^0$, where $\boldsymbol{\phi}^0$ is the unique maximizer of $\bar{Q}_n(\boldsymbol{\phi})$ which satisfies $Q_n(\omega, \boldsymbol{\phi}) - \bar{Q}_n(\boldsymbol{\phi}) \longrightarrow 0$ a.s. P. We first establish a generalization of Theorem 1 in Mardia and Marshall (1984) under assumptions (A.7)-(A.9).

Proposition 2.1 *Under (A.7)-(A.9), the asymptotic normality and weak consistency of $\hat{\boldsymbol{\phi}}_n$ hold, i.e.,*

$$\sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}_0) \to N(0, H^{-1}(\boldsymbol{\phi}_0)).$$

Next we study the distribution consistency of $\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n$ in Theorem 2.3. This result provides a strong theoretical support to the proposed LHD-based block bootstrap framework. It shows that this framework guarantees the asymptotic consistency of the bootstrapped MLEs to the MLEs using full data, given the advantages that this procedure reduces computation and avoids numerical issues in calculating MLEs.

Theorem 2.3 *Under (A.1)-(A.11), if $m = o(n^{1/d})$ and $m \to \infty$, then*

$$P\big[\omega : \sup_x |P^*_{N,\omega}(\sqrt{n/m^{d-1}}(\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n) \le x) - P(\sqrt{n}(\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}^0) \le x)| > \epsilon\big] \to 0,$$

where "$\le$" applies to each component of the vectors.

Proof: Define $B_n^0 = \boldsymbol{Var}\{n^{-1}\sum_{s=1}^n \nabla q_s(z_s, \boldsymbol{\phi}^0)\}$. We first show that

$$\sqrt{n/m^{d-1}}\, B_n^{0\,-1/2} \nabla Q_N^*(\cdot, \omega, \hat{\boldsymbol{\phi}}_N^*) \to N(0, I_p)$$

under $P^*_{N,\omega}$.

To show this, denote $\bar{h}_N^*(\boldsymbol{\phi}) = N^{-1}\sum_{s=1}^{N}\nabla q_s^*(z_s^*,\boldsymbol{\phi})$ and $\bar{h}_n(\boldsymbol{\phi}) = n^{-1}\sum_{s=1}^{n}\nabla q_s(z_s,\boldsymbol{\phi})$. Then

$$\begin{aligned}
& \sqrt{n/m^{d-1}}[\bar{h}_N^*(\hat{\boldsymbol{\phi}}_N^*) - \bar{h}_n(\hat{\boldsymbol{\phi}}_n)] \\
= \; & \sqrt{n/m^{d-1}}[\bar{h}_N^*(\hat{\boldsymbol{\phi}}_N^*) - \bar{h}_N^*(\hat{\boldsymbol{\phi}}_n)] + \sqrt{n/m^{d-1}}[\bar{h}_N^*(\hat{\boldsymbol{\phi}}_n) - \bar{h}_N^*(\boldsymbol{\phi}^0)] \\
& + \sqrt{n/m^{d-1}}[\bar{h}_N^*(\boldsymbol{\phi}^0) - \bar{h}_n(\boldsymbol{\phi}^0)] + \sqrt{n/m^{d-1}}[\bar{h}_n(\boldsymbol{\phi}^0) - \bar{h}_n(\hat{\boldsymbol{\phi}}_n) \\
= \; & J_1 + J_2 + J_3 + J_4.
\end{aligned}$$

Since $\bar{h}_n$ and $\bar{h}_N^*$ are functions whose secondary derivative are continuous, $J_1 \to 0$ $prob-P_{N,\omega}^*, prob-P$ as $\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n \to 0$ $prob-P_{N,\omega}^*, prob-P$ in Theorem 2.2. $J_2 + J_4 \xrightarrow{\text{P}} 0$ as $\hat{\boldsymbol{\phi}}_n - \boldsymbol{\phi}^0 \to 0$ by Proposition 1. In addition, the two terms in J_3 are both evaluated at $\boldsymbol{\phi}^0$ which is a fixed value, by Theorem 2.1, we have $B_n^{0\,1/2} J_3 \to N(0, I_p)$ in $P_{N,\omega}^*$.

By condition (A.10) and follow a similar proof as Lemma 2.5, we have

$$\nabla^2 Q_N^*(\cdot,\omega,\boldsymbol{\phi}) - \nabla^2 Q_n(\omega,\boldsymbol{\phi}) \to 0 \;\; prob-P_{N,\omega}^*, prob-P.$$

Let $\hat{H}_n(\omega) = \nabla^2 Q_n(\omega, \hat{\boldsymbol{\phi}}_n)$. According to White White (1996), given the result $\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n \to 0$ $prob-P_{N,\omega}^*, prob-P$, and assumption (A.8), we have

$$\begin{aligned}
\sqrt{N}(\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n) &= -\hat{H}_n^{-1}(\omega)\sqrt{n}\nabla Q_N^*(\cdot,\omega,\hat{\boldsymbol{\phi}}_n) + o_{P_{N,\omega}^*}(1) \\
&= -H_n(\boldsymbol{\phi}^0)^{-1}(\omega)\sqrt{n}\nabla Q_N^*(\cdot,\omega,\hat{\boldsymbol{\phi}}_n) + o_{P_{N,\omega}^*}(1).
\end{aligned}$$

Given the fact that under $P_{N,\omega}^*$

$$\sqrt{n/m^{d-1}}\, B_n^{0\,-1/2}\nabla Q_N^*(\cdot,\omega,\hat{\boldsymbol{\phi}}_N^*) \to N(0, I_p),$$

we have

$$B_n^{0\,-1/2} H_n(\boldsymbol{\phi}^0)\sqrt{N}(\hat{\boldsymbol{\phi}}_N^* - \hat{\boldsymbol{\phi}}_n) \to N(0, I_p)$$

under $P_{N,\omega}^*$. By the method of subsequence as in Theorem 2.2, together with the consistency results of $\hat{\boldsymbol{\phi}}_n$ in Proposition 1, the result follows immediately. □

2.4 Examples

2.4.1 Simulation study

The finite-sample performance of the proposed framework is demonstrated by numerical examples in this section. The objective is to compare the MLEs obtained from LHD-based block bootstrap with those obtained using full data. All simulations are performed using a W35653 20GHz, 2G RAM workstation under R 2.15.2 in Windows 7.

Simulations are conducted for three sample sizes, $n = 400, 900, 2500$ and data are generated from a regular grid in $[0, 1]^2$. Note that the proposed method is particularly useful for data collected from irregular grids. The reason to generate the simulations from a regular grid is because under this setting, the MLE calculation using full data can be further simplified by Kronecker product techniques and some matrix singularity can be avoided (Bayarri et al., 2007, ????; Rougier, 2008). But these techniques are only applicable to data sets collected from a regular grid. Therefore, a favorable comparison of the proposed method would make an even stronger case for the design-based subsampling procedure.

Data are generated from Gaussian process with the mean function coefficients set to be $\boldsymbol{\beta} = (1, 1)$. Choose the correlation function to be

$$\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-\sum_{i=1}^{2} |x_{1i} - x_{2i}|/\theta_i),$$

where $\theta_1 = \theta_2 = 1$ and $\sigma^2 = 1$. For each choice of sample size, a total of 100 data sets are simulated. For each simulated data set, 20 LHD-based block bootstrap samples are collected. Three different settings, $m = 1, 4$, and 6, of LHD-based block bootstrap are performed. When $m = 1$, the results refer to the conventional GP modeling using

Table 2.1: LHD-based Bootstrap parameter estimation

	m	Parameter estimation					Time
		θ_1	θ_2	β_1	β_2	σ^2	
n=400	1	0.94 (0.24)	0.97 (0.24)	0.89 (0.78)	1.04 (0.76)	0.92 (0.32)	58.42 (11.73)
	4	0.94 (0.29)	0.93 (0.27)	0.87 (0.92)	1.01 (0.87)	1.15 (0.48)	56.88 (7.96)
	6	0.89 (0.30)	0.93 (0.29)	0.87 (0.79)	1.01 (0.82)	0.80 (0.31)	32.18 (4.02)
n=900	1	0.96 (0.18)	0.95 (0.16)	0.86 (0.80)	1.04 (0.78)	0.92 (0.22)	306.40 (62.50)
	4	0.95 (0.20)	0.94 (0.18)	0.81 (0.92)	1.05 (0.86)	0.90 (0.25)	280.31 (33.44)
	6	0.94 (0.21)	0.92 (0.20)	0.82 (0.86)	1.05 (0.82)	0.86 (0.24)	145.35 (17.11)
n=2500	1	0.98 (0.13)	0.98 (0.14)	1.01 (0.80)	1.02 (0.79)	0.96 (0.20)	2258.60 (688.52)
	4	0.98 (0.16)	0.96 (0.17)	1.01 (0.84)	0.97 (0.83)	0.95 (0.23)	1871.21 (214.36)
	6	0.98 (0.16)	0.97 (0.17)	1.03 (0.86)	0.98 (0.81)	0.95 (0.24)	792.50 (82.59)

full data. To show the empirical performance of the LHD-based block bootstrap approach, we report the computing time, mean and standard deviation of the parameter estimation. The results are summarized in Table 2.1. In addition, the empirical density function for each parameter is shown in Figures 2.2 and 2.3 for sample size $n = 900$. Similar plots are obtained for the other two sample sizes, therefore they are omitted.

The results in Table 2.1 demonstrate that the estimated parameters based on LHD-based block bootstrap are consistent to the one obtained from full data, i.e., $m = 1$ cases. In general, the standard deviations increase slightly with the number of blocks m in estimating the correlation parameters, while decrease in estimating the rest of the parameters. In terms of computing time, LHD-based block bootstrap is much faster than the conventional GP modeling and such advantage is particularly significant when the total sample size n is large. For example, the percentage of computational time saved using the proposed method is increased from 45% to 65% when sample size increases. Figures 2.2 and 2.3 show that the two empirical distributions generated by the LHD-based block bootstrap are approximately normal and perform similarly to the

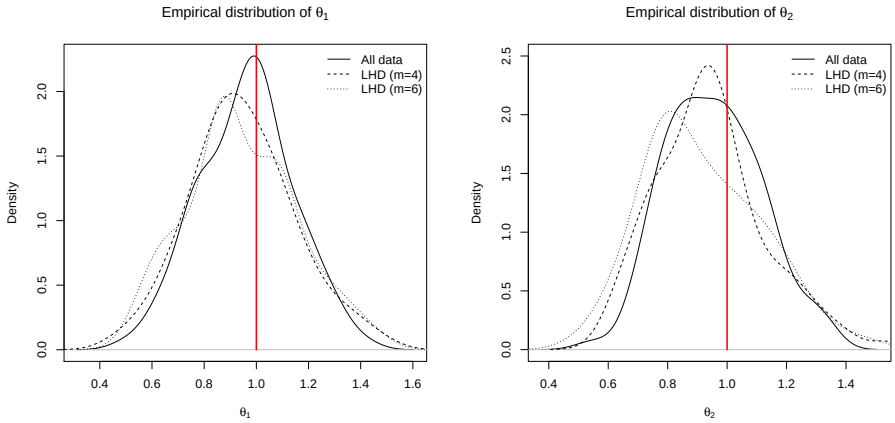

Figure 2.2: Comparison of the empirical distributions of $\hat{\theta}_n$ and $\hat{\theta}_N^*$ (n=900)

one from full data.

2.4.2 Application to data center thermal management study

A data center is a computing infrastructure facilities that house large amounts of information technology (IT) equipment used to process, store, and transmit digital information. Data center facilities constantly generate large amounts of heat to the room, which must be maintained at an acceptable temperature for reliable operation of the equipment. More discussions of data center can be found in Hung et al. (2012). A significant fraction of the total power consumption in a data center is for heat removal; therefore, determining the most efficient cooling mechanism has become a major challenge. The objective of a thermal management study is to model the thermal distribution in a data center and the final goal is to design a data center with an efficient heat-removal mechanism.

For a data center thermal study, physical experiments are not always feasible because some settings are highly dangerous and expensive to perform. Therefore, simulations

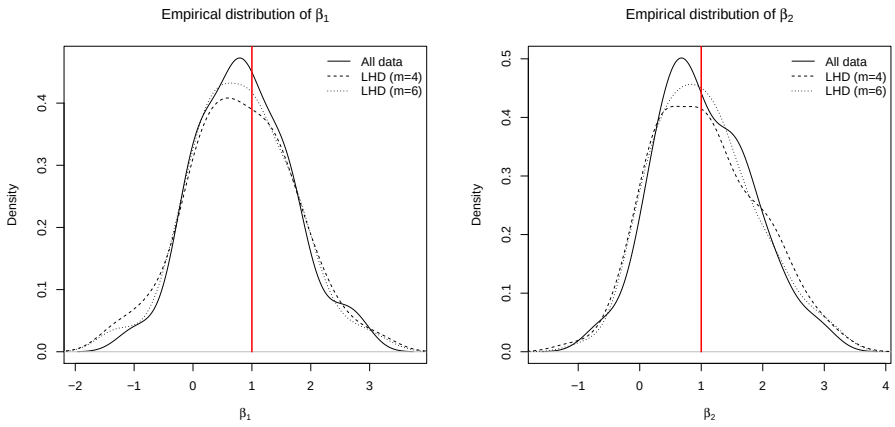

Figure 2.3: Comparison of the empirical distributions of $\hat{\beta}_n$ and $\hat{\beta}_N^*$ (n=900)

based on computational fluid dynamics (CFD) are widely used. In this example, CFD simulations are conducted at IBM T. J. Watson Research Center based on a real data center layout. Detailed discussions about the CFD simulations can be found in López and Hamann (2011). The first three columns in Table 2.2 list nine factors and their levels in the CFD simulations, including four computer room air conditioning (CRAC) units with different flow rates $(x_1, ..., x_4)$, the overall room temperature setting (x_5), the perforated floor tiles with different percentage of open areas (x_6), and spatial location in the data center (x_7 to x_9). There are 27,000 temperatures simulated from the CFD simulator and these temperature outputs are obtained from an irregular grid over the 9-dimensional experimental space.

It is computationally infeasible to fit a GP model using all the CFD simulation outputs. Therefore, we reduce the computation by the proposed LHD-based block bootstrap approach with $m = 3$ for variables x_6, x_7 and x_9, which are the top three factors with highest levels. The fitted GP model is summarized by the last two columns

Table 2.2: LHD Bootstrap analysis of thermal management data

	Variable	Levels	$\hat{\beta}$	$\hat{\theta}$
x_1	CRAC unit 1 flow rate (cfm)	(0,7000,8500,10000 11500,13000)	-8.58(0.96)	0.85(0.17)
x_2	CRAC unit 2 flow rate (cfm)	(0,7000,8500,10000 11500,13000)	-11.12(1.26)	0.77(0.23)
x_3	CRAC unit 3 flow rate (cfm)	(0,2500,4000,5500)	-6.83(0.80)	1.14(0.27)
x_4	CRAC unit 4 flow rate (cfm)	(0,2500,4000,5500)	-6.26(0.98)	1.70(0.71)
x_5	Room temperature setting (F)	(65,67,69,71,73, 75)	-0.82(0.66)	3.39(0.94)
x_6	Tile open area percentage (%)	$(15, 25, 35, 45$ $(55, 65, 75)$	0.15(3.63)	1.24(0.91)
x_7	Location in x-axis	8 unequally spaced	-5.09(2.72)	0.14(0.11)
x_8	Location in y-axis	4 unequally spaced	3.70(2.18)	0.62(0.25)
x_9	Height	18 equally spaced	33.43(3.90)	21.61(0.22)

of Table 2.2, where $\hat{\beta}$ represents the estimated mean function coefficients and $\hat{\theta}$ represents the correlation parameters estimated based on exponential covariance function. From the fitted model, it appears that the height (x_9) in a data center has relatively larger effect in the mean function and the overall room temperature setting (x_5) has a larger estimate in its correlation parameter. These results agree with the general understanding of thermal dynamics in a data center that temperature increases with height and the overall room temperature setting has significant impact on controlling the temperature. It is worth noting that the mean function in the fitted GP model contains the linear effects of the nine factors, which is predetermined without any variable selection mechanism. If the objective is to further specify important factors, penalized likelihood approaches developed in Chu et al. (2011); Li and Sudjianto (2005) can be incorporated to the proposed framework for variable selection.

2.5 Discussion

We present the practically useful LHD-based block bootstrap procedure to tackle computational difficulties in GP modeling. It borrows the strength of space-filling designs to provide an efficient subsampling plan and reduce computational complexity. We prove a very general result to support the asymptotic consistency of the estimators obtained from the proposed procedure. Finite-sample performance is examined through simulation studies. The proposed procedure is applied to a data center thermal management study and an efficient GP model is obtained based on 27,000 computer outputs generated from CFD simulator.

Future work on the LHD-based block bootstrap procedure will be explored in the following directions. First, extensions of the proposed procedure to optimal designs with better space-filling properties is intuitively appealing. For example, it is known that randomly generated LHDs can contain some structure. To further enhance desirable space-filling properties, various modifications are proposed (Fang et al., 2006; Owen, 1994; Qian et al., 2009; Qian and Wu, 2009; Tang, 1993, 1994; Ye, 1998). Numerical comparisons and theoretical developments of the generalization to different types of optimal space-filling designs will be carefully studied. Second, an interesting and important issue of the LHD-based block bootstrap is to determine the optimal block size. This topic has been discussed for conventional block bootstrap methods (Hall et al., 1995; Lahiri, 1999; Nordman et al., 2007), however the solutions therein are not directly applicable to GP models. We plan to study the optimal block size for the propose procedure based on a new criterion defined for GP. Third, we plan to extend the LHD-based block bootstrap idea to construct bootstrap predictive distributions. This is a promising direction because it not only overcomes the drawback of GP plug-in

predictors (Santner et al. (2003), p.98) but also addresses the computational issue with the conventional bootstrap predictive distribution (Sjöstedt-de Luna, 2003).

Chapter 3

Penalized Quantile Regression with Asymmetric Laplace Process Model for Computer Experiments

3.1 Introduction

Despite numerous research on computer experiment modeling, most of the existing approaches focus on modeling the conditional mean of the response variable (Fang et al., 2006; Santner et al., 2003). Little work is done to study quantile regression model that incorporate data dependence although in practice it is often of substantial interest. Classic quantile regression model usually consider the case that observations are all independent. For example, Koenker and Bassett Jr (1978), Gutenbrunner and Jurecková (1992), Chaudhuri et al. (1997) and Chernozhukov (2005) investigate quantile regression models with fixed number of covariates. Recently, Reich et al. (2011), Lum et al. (2012) and Boukouvalas et al. (2012) extend quantile regression models to incorporate spatial dependence in Bayesian framework. However, none of the aforementioned work investigate high dimensional data which is commonly seen in computer experiments with increasing availability of large size data and computing power. High dimensional data often display heterogeneity (Wang et al., 2012) and calls for models with sparsity in which only a small number of covariates have influence on the conditional distribution of response. Penalization has emerged as a successful technique for model selection and it is been extended to classic quantile regression as well. Belloni et al. (2011) investigate

L_1-penalized quantile regression and Wang et al. (2012) consider nonconvex penalties. Zou and Yuan (2008) propose penalized composite quantile regression and show that it enjoys oracle model selection property. It is worth noting that Belloni et al. (2011) and Wang et al. (2012) consider possibly infinite collection quantile regression models with different quantiles and assume the set of covariates that can impact the conditional distribution of response may differ when different quantiles or segments of conditional distribution are studied. On the other hand, Zou and Yuan (2008) study finite number of quantile regression models and they all have the same set of relevant covariates. In this chapter, we adopt the former concept, that is, relevant covariates may be different for various quantiles of response's conditional distribution.

This research is motivated by a data center thermal study. In this study, large amounts of heat are constantly generated to the room, which must be maintained at an acceptable temperature for reliable operation of the equipment. A significant fraction of the total power consumption is for heat removal. Therefore, we are interested in understanding the effect of different types of cooling approaches and the goal is to find the most efficient and reliable heat removal mechanism. To achieve this goal, we need to identify active factors and their effects on extreme temperature quantiles. Because better control on higher temperature quantile can prevent significant damage to servers and bringing up lower quantile temperature properly can dramatically reduce unnecessary power consumption.

In this research, a new modeling framework is proposed to model different quantiles in computer experiments and simultaneously identify important effects for each quantile. To achieve this goal, we extend Gaussian process to asymmetric Laplace process. The choice of asymmetric Laplace process enable us to model the quantiles

and incorporate dependence structure. In addition, we regularize the quantile asymmetric Laplace process with a penalty function, such as L_1 norm penalty (Tibshirani, 1996), the SCAD (Fan and Li, 2001) and the MCP (Zhang, 2010). We show that penalized quantile asymmetric Laplace estimator can select true relevant covariates when the number of covariates is large and able to grow to infinity when the number of observations increase to infinity.

The reminder of this chapter is organized as follows. In Section 3.2, we introduce quantile regression with asymmetric Laplace process and its penalized estimator. In Section 3.3, we investigate asymptotic properties of the proposed penalized estimator. In Section 3.4, we propose an algorithm for penalized quantile regression with asymmetric Laplace process. Section 3.5 presents simulation studies and a real data example. Final discussions are given in Section 3.6 and proofs are in Section 3.7.

3.2 Penalized quantile regression with asymmetric Laplace process

Suppose that we have a set of data $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ with input $\boldsymbol{x}_i \in \mathbb{R}^p$ and output $y_i \in \mathbb{R}$. The $100\tau\%$ quantile of $y|\boldsymbol{x}$, $\tau \in (0,1)$ can be recovered (Koenker and Bassett Jr, 1978) by minimizing a check function as

$$\min \sum_{i=1}^{n} \rho_\tau(y_i - \boldsymbol{x}_i\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ and $\rho_\tau(\cdot)$ is the "check function" defined by $\rho_\tau(t) = t[\tau - I(t < 0)], \tau \in (0,1)$. As minimizing L_2 loss is associated with normal errors in linear models, minimizing check function corresponds to assuming asymmetric Laplace errors. That is, assume for $i = 1, \ldots, n$,

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i,$$

where ϵ_i are independent and follow Asymmetric Laplace Process (ALP) with $100\tau\%$ quantile being 0. The probability density function of ϵ_i is

$$f_\tau(\epsilon_i \mid \sigma) = \frac{\tau(1-\tau)}{\sigma} \exp\left[-\frac{\epsilon_i}{\sigma}(\tau - I(\epsilon_i \leq 0))\right] = \{\tau(1-\tau)/\sigma\} \exp(-\rho_\tau(\epsilon_i)\ \sigma). \quad (3.1)$$

As a result, the maximum likelihood estimation (MLE) is equivalent to estimators minimizing the check function.

To construct a nonlinear parametric model for quantile analysis in computer experiments, we incorporate a representation of the ALP provided by Kozubowski and Podgorski (2000). A random variable ϵ_τ following ALP with density (3.1) can be represented by (Kozubowski and Podgorski, 2000)

$$\epsilon_\tau = \sigma\sqrt{\frac{2W}{\tau(1-\tau)}}Z + \sigma\frac{1-2\tau}{\tau(1-\tau)}W,$$

where Z and W are independent, Z follows standard normal distribution and W follows standard exponential distribution.

Using ALP, a new parametric model incorporating data dependence can be constructed by (Lum et al., 2012)

$$Y(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}_\tau(\boldsymbol{x}), \quad (3.2)$$

where $\boldsymbol{x}\boldsymbol{\beta}$ is the mean function, $\boldsymbol{\epsilon}_\tau(\boldsymbol{x}) = \sigma\sqrt{\frac{2W}{\tau(1-\tau)}}Z(\boldsymbol{x}) + \sigma\frac{1-2\tau}{\tau(1-\tau)}W$, $Z(\boldsymbol{x})$ and W are independent, W follows standard exponential distribution, and $Z(\boldsymbol{x})$ is a stationary Gaussian process with mean 0 and correlation function ψ, i.e. $corr\{Z(\boldsymbol{x}+\boldsymbol{h}), Z(\boldsymbol{x})\} = \psi(\boldsymbol{h};\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of correlation parameters and $\psi(\boldsymbol{h};\boldsymbol{\theta})$ is positive semidefinite function with $\psi(\mathbf{0};\boldsymbol{\theta}) = 1$ and $\psi(\boldsymbol{h};\boldsymbol{\theta}) = \psi(-\boldsymbol{h};\boldsymbol{\theta})$. Note that $Z(\boldsymbol{x})$ is used to incorporate correlations among observations. In this chapter, we consider a common exponential distribution W for n observations. In fact, quantile ALP model can be

extended to incorporate independent standard exponential distribution or spatial exponential distribution (Lum et al., 2012). For simplicity, we use common W in this chapter.

With n observations collected from computer experiments, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, denote $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{X} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$. Let $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$, the likelihood function of (3.2) can be written as

$$L_\tau(\boldsymbol{\epsilon}) = \frac{2\exp(\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau)}{(2\pi)^{n/2}|\Sigma_\tau|^{1/2}}\Big(\frac{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}}{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\Big)^{v/2}K_v\Big(\sqrt{(2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau)(\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})}\Big), \quad (3.3)$$

where $v = (2-n)/2$, $\boldsymbol{\mu}_\tau = \sigma\frac{1-2\tau}{\tau(1-\tau)}\mathbf{1}_n$, $\Sigma_\tau = \frac{\sigma^2}{\tau(1-\tau)}\Psi$ with $\mathbf{1}_n$ being a vector of length n and all entries equaling to 1 and $\Psi = (\psi(\boldsymbol{x}_i, \boldsymbol{x}_j), i, j = 1, \ldots, n)$ and $K_v(\cdot)$ is the modified Bessel function

$$K_v(u) = \frac{1}{2}\Big(\frac{u}{2}\Big)^v\int_0^\infty z^{-v-1}\exp\Big(-z-\frac{u^2}{4z}\Big)dz,$$

valid for complex u with the non-negative real part of u^2.

For $i = 1, \ldots, n$, since all $\epsilon_i = y_i - \boldsymbol{x}_i\boldsymbol{\beta}$ have common W, error terms are always correlated even Φ is identity matrix. Explicitly, covariance between ϵ_i and ϵ_j is

$$cov(\epsilon_i, \epsilon_j) = \sigma^2\frac{2\phi(\boldsymbol{x}_i - \boldsymbol{x}_j; \boldsymbol{\theta})}{\tau(1-\tau)} + \sigma^2\frac{(1-2\tau)^2}{\tau^2(1-\tau)^2}.$$

To estimate parameters and identify important variables simultaneously, we implement a penalized likelihood approach to the proposed model. Define penalized log-likelihood as

$$Q_\tau(\boldsymbol{\beta}) = \ell_\tau(\boldsymbol{\beta}) - \sum_{j=1}^p \rho_{\lambda_\tau}(|\beta_j|),$$

where $\ell_\tau(\boldsymbol{\beta}) = n^{-1}\log L_\tau(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ and $\rho_{\lambda_\tau}(\cdot)$ is penalty function with tuning parameter λ_τ and penalized likelihood estimator is defined as

$$\hat{\boldsymbol{\beta}}_\tau = \arg\max_{\boldsymbol{\beta}\in\Omega} Q_\tau(\boldsymbol{\beta}),$$

where $\Omega = \{\boldsymbol{\beta} : \ell_\tau(\boldsymbol{\beta}) < \infty\}$.

3.3 Asymptotic properties

In this section, we investigate the asymptotic property of $\hat{\boldsymbol{\beta}}_\tau$. Suppose the true coefficients for each τ is $\boldsymbol{\beta}_\tau^0$. Denote $A_\tau = \{j : \beta_{\tau,j} \neq 0\}$ is the index set of non-zero coefficients and A_τ^c is its complement. It's worth noting that for $\boldsymbol{\beta} \in \Omega$, $\ell_\tau(\cdot)$ is differentiable everywhere. Following Fan and Peng (2004), we have the following theorems.

Assume the following regularity conditions for $\boldsymbol{\beta} \in \Omega$.

(A1) There exists constants $C_1, C_2 > 0$, such that $I_\tau(\boldsymbol{\beta}) = nE_{\boldsymbol{\beta}}\{(\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})(\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})^T\}$ satisfies

$$0 < C_1 < \lambda_{\min}\{I_\tau(\boldsymbol{\beta})\} \leq \lambda_{\max}\{I_\tau(\boldsymbol{\beta})\} < C_2 < \infty,$$

and $J_\tau(\boldsymbol{\beta}) = -E_{\boldsymbol{\beta}}\{\nabla^2 \ell_\tau(\boldsymbol{\beta})\}$ satisfies

$$0 < C_1 < \lambda_{\min}\{J_\tau(\boldsymbol{\beta})\} \leq \lambda_{\max}\{J_\tau(\boldsymbol{\beta})\} < C_2 < \infty.$$

(A2) There exists constants $C_3, C_4 > 0$ such that for $j, k = 1, \ldots, p$,

$$E_{\boldsymbol{\beta}}\{\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \beta_j}\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \beta_k}\}^2 < C_3/n^2 < \infty$$

and

$$E_{\boldsymbol{\beta}}\{\frac{\partial^2 \ell_\tau(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\}^2 < C_4/n < \infty.$$

(A3) For $i, j, k = 1, \ldots, p$, assume there are random variables $M_{i,j,k}$ such that

$$|\frac{\partial^3 \ell_\tau(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j \partial \beta_k}| < M_{i,j,k}$$

with probability 1 and there exists a constant $C_5 > 0$ such that

$$E_\beta M_{i,j,k}^2 < C_5 < \infty.$$

Condition A1 and A2 ensure that the second moment of the likelihood function and Fisher's information matrix is positive definite. Due to the complexity of the likelihood function, we do not have $I_\tau(\boldsymbol{\beta}) = J_\tau(\boldsymbol{\beta})$. So eigenvalues of $I_\tau(\boldsymbol{\beta})$ and $J_\tau(\boldsymbol{\beta})$ are need to be controlled. Condition A3 ensures that higher orders of the Taylor expansion for the likelihood function are small enough.

Furthermore, we assume the following conditions on the penalty function.

(A4) $\liminf_{n\to\infty} \liminf_{x\to 0+} \rho'_{\lambda_\tau}(x)/\lambda_\tau > 0$

(A5) There are constant D_1 and D_2 such that $\forall\ x_1, x_2 > D_1\lambda_\tau$, $|\rho''_{\lambda_\tau}(x_1) - \rho''_{\lambda_\tau}(x_2)| < D_2|x_1 - x_2|$.

Condition A4 and A5 are regular conditions on the penalty function. It includes a large family of penalty functions, including L_1 norm penalty, the SCAD and the MCP.

Conditions on the tuning parameter and true coefficients $\boldsymbol{\beta}^0_\tau$ are given below.

(A6) $\max_{j\in A_\tau} |\rho'_{\lambda_\tau}(|\beta^0_{\tau,j}|)| = o(n^{-1/2})$.

(A7) $\max_{j\in A_\tau} |\rho''_{\lambda_\tau}(|\beta^0_{\tau,j}|)| = o(1)$.

Condition A6 and A7 guarantee proper choice of the tuning parameter with respect to the strength of true non-zero coefficients.

We establish the existence of the penalized likelihood estimator.

Theorem 3.1 *Suppose that conditions A1 to A7 are satisfied. If* $p = o(n^{1/4})$, *then there is a local maximizer* $\hat{\boldsymbol{\beta}}_\tau$ *of* $Q_\tau(\cdot)$ *such that* $\|\hat{\boldsymbol{\beta}}_\tau - \boldsymbol{\beta}^0_\tau\|_2 = O_p(\sqrt{p/n})$.

The sparsity property of proposed penalized estimator is given below.

Theorem 3.2 *Suppose conditions A1 to A7 are satisfied. If* $\sqrt{p/n}/\lambda_\tau = o(1)$ *and* $p = o(n^{1/4})$*, then with probability approaching 1, the root-n/p consistent estimator* $\hat{\boldsymbol{\beta}}_\tau$ *satisfies* $\hat{\boldsymbol{\beta}}_{\tau, A_\tau^c} = 0$.

Together with theorem 3.1 and theorem 3.2, we show that penalized quantile asymmetric Laplace estimator can select true relevant covariates when the number of covariates is large and able to grow to infinity when the number of observations increase to infinity.

3.4 Algorithm

In this section, we propose an algorithm to compute penalized ALP estimators. Let $\beta_{\tau,j} = \beta_{\tau,j}^+ - \beta_{\tau,j}^-$, where $\beta_{\tau,j}^+ \geq 0$ and $\beta_{\tau,j}^- \geq 0$, $j = 1, \ldots, p$. Then the optimization problem is equivalent to

$$(\boldsymbol{\beta}^+, \boldsymbol{\beta}^-) = \arg\max_{\boldsymbol{\beta}^+, \boldsymbol{\beta}^-} Q_\tau(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^+ - \boldsymbol{X}\boldsymbol{\beta}^-),$$

subject to $\beta_{\tau,j}^+, \beta_{\tau,j}^- \geq 0$, where $\boldsymbol{\beta}^+ = (\beta_{\tau,j}^+, j = 1, \ldots, p)$ and $\boldsymbol{\beta}^- = (\beta_{\tau,j}^-, j = 1, \ldots, p)$.

The aforementioned algorithm is computationally intensive because of the complicated likelihood function. Following Zou and Li (2008), given an initial value $\boldsymbol{\beta}^{int}$, we approximate the log-likelihood function by

$$\ell_\tau(\boldsymbol{\beta}) \approx \ell_\tau(\boldsymbol{\beta}^{int}) + \nabla^T \ell_\tau(\boldsymbol{\beta}^{int})(\boldsymbol{\beta} - \boldsymbol{\beta}^{int}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{int})^T \nabla^2 \ell_\tau(\boldsymbol{\beta}^{int})(\boldsymbol{\beta} - \boldsymbol{\beta}^{int}).$$

Take $\boldsymbol{\beta}^{int}$ as the maximum likelihood estimate $\tilde{\boldsymbol{\beta}}$. Since $\nabla \ell_\tau(\tilde{\boldsymbol{\beta}}) = 0$, one-step sparse estimator is given by

$$\hat{\boldsymbol{\beta}} = \arg\min \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{int})^T \{-\nabla^2 \ell_\tau(\boldsymbol{\beta}^{int})\}(\boldsymbol{\beta} - \boldsymbol{\beta}^{int}) + \sum_{j=1}^{p} \rho_{\lambda_\tau}(|\beta_j|),$$

where

$$\begin{aligned}\nabla \ell_\tau(\boldsymbol{\beta}) &= -\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau - 2v\frac{\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}}{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}} \\ &\quad -\frac{\tilde{K}_v'(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}{\tilde{K}_v(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}\frac{\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})}{\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}}}\end{aligned}$$

and

$$\begin{aligned}\nabla^2\ell_\tau(\boldsymbol{\beta}) &= -2v\frac{\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{X}}{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}} - 2v\frac{(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})^T}{(\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})^2} \\ &\quad -\frac{2\tilde{K}_v''(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}{\tilde{K}_v(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}\frac{(2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau)(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})^T}{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}} \\ &\quad -\frac{2\{\tilde{K}_v'(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})\}^2}{\tilde{K}_v^2(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}\frac{(2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau)(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})^T}{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}} \\ &\quad -\frac{\tilde{K}_v'(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}{\tilde{K}_v(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}\frac{\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{X})}{\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}}} \\ &\quad -\frac{2\tilde{K}_v'(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}{\tilde{K}_v(\sqrt{2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau}\sqrt{\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon}})}\frac{(2+\boldsymbol{\mu}_\tau^T\Sigma_\tau^{-1}\boldsymbol{\mu}_\tau)^{1/2}(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})(\boldsymbol{X}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})^T}{(\boldsymbol{\epsilon}^T\Sigma_\tau^{-1}\boldsymbol{\epsilon})^{3/2}}\end{aligned}$$

with

$$\tilde{K}_v(u) = u^{-v}K_v(u) = \frac{1}{2^{v+1}}\int_0^\infty z^{-v-1}\exp(-z-\frac{u^2}{4z})dz,$$

and $\tilde{K}_v'(u) = -u\tilde{K}_{v-1}(u)/2$, $\tilde{K}_v''(u) = -\tilde{K}_{v-1}(u)/2 + u^2\tilde{K}_{v-2}(u)/4$.

Numerous literatures discuss computing the integral in $L_\tau(\cdot)$, see e.g. Amos (1974) and Jin and Zhang (1996). When n is odd, say $n = 2r+3$, we can further reduce the computational burden by writing the integral in closed form

$$K_v\big(u\big) = \sqrt{\frac{\pi}{2u}}e^{-u}\sum_{k=0}^{r}\frac{(r+k)!}{(r-k)!k!}(2u)^{-k}.$$

3.5 Numerical studies

3.5.1 Simulation studies

In this section, we present several examples that demonstrate the model selection property and quantile estimation of the proposed model. We simulate n observations $(\boldsymbol{x}_i, y_i), i = 1, \ldots, n$ from the following model

$$y(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}_\tau(\boldsymbol{x}),$$

where $\boldsymbol{\epsilon}_\tau(\boldsymbol{x}) = \sigma\sqrt{\frac{2W}{\tau(1-\tau)}}Z(\boldsymbol{x}) + \sigma\frac{1-2\tau}{\tau(1-\tau)}W$, $Z(\boldsymbol{x})$ and W are independent, $Z(\boldsymbol{x})$ is a stationary Gaussian process and W follows standard exponential distribution. Assume there are p explanatory variables and the stationary Gaussian process $Z(\boldsymbol{x})$ has mean 0 and covariance structure $cov(Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2)) = \exp(-\sum_{i=1}^{p}|x_{1i} - x_{2i}|/\theta_i)$, where θ_i, $i = 1, \ldots, p$ are parameters.

Similar to Wang et al. (2012), we use quantiles $\tau = 0.5, 0.3$ and 0.7 to demonstrate our approach. Common penalty functions such as L_1 norm penalty, the SCAD penalty and the MCP are applied. In example 1, we take $n = 200$, $p = [n^{1/4}] = 4$ and $s = [\sqrt{p}] = 2$. In example 2, we consider larger number of variables and take $p = 50$ and $s = 5$ to further explore the performance of proposed method. Non-zero coefficients $\boldsymbol{\beta}^0_{A_\tau} \approx \sqrt{2p/n}$. Other model parameters are chosen as $\theta_i = 1, i = 1, \ldots, p$ and $\sigma = 1$. BIC is used for tuning parameter selection. Based on 100 replications for a given model, the performance of proposed approach is evaluated by the following criterion: model size which is the number of variables with non-zero estimation, variable selection sensitivity which is defined as the proportion of variables in A_τ that are selected, variable selection specificity which is the proportion of variables in A^c_τ are excluded and mean square error (MSE) of $\boldsymbol{\beta}$, defined as $\sum_{i=1}^{p}(\hat{\beta}_i - \beta^0_i)^2/p$. The results are shown in Table 3.1.

Table 3.1: Model selection for quantile Gaussian process

		model size	variable selection sensitivity	variable selection specificity	MSE ($\boldsymbol{\beta}$)
Example 1: $n = 200$ $p = 4$ $s = 2$					
$\tau = 0.5$	LASSO	2.11 (0.37)	99.00% (7.04%)	93.50% (16.90%)	0.10 (0.24)
	SCAD	2.02 (0.32)	98.00% (9.85%)	97.00% (11.93%)	0.09 (0.27)
	MCP	1.97 (0.22)	98.00% (9.85%)	99.50% (5.00%)	0.09 (0.26)
$\tau = 0.3$	LASSO	1.99 (0.39)	95.50% (16.04%)	96.00% (13.63%)	0.20 (0.45)
	SCAD	1.86 (0.38)	92.00% (19.75%)	99.00% (7.04%)	0.26 (0.54)
	MCP	1.90 (0.36)	94.00% (16.33%)	99.00% (7.04%)	0.23 (0.54)
$\tau = 0.7$	LASSO	2.02 (0.47)	95.00% (15.08%)	94.00% (16.33%)	0.20 (0.42)
	SCAD	1.84 (0.39)	91.50% (18.88%)	99.50% (5.00%)	0.26 (0.53)
	MCP	1.82 (0.39)	91.00% (19.31%)	100% (0%)	0.29 (0.55)
Example 2: $n = 200$ $p = 50$ $s = 5$					
$\tau = 0.5$	LASSO	6.28 (3.06)	98.20% (12.74%)	96.96% (6.59%)	0.20 (0.93)
	SCAD	5.31 (1.84)	98.60% (10.73%)	99.16% (3.93%)	0.17 (0.88)
	MCP	5.28 (1.78)	98.60% (10.73%)	99.20% (3.80%)	0.17 (0.86)
$\tau = 0.3$	LASSO	18.40 (10.05)	100% (0%)	70.25% (22.26%)	0.09 (0.11)
	SCAD	15.65 (10.85)	97.71% (7.03%)	77.73% (21.62%)	1.26 (2.33)
	MCP	15.57 (10.78)	97.71% (7.03%)	77.96% (21.48%)	1.26 (2.34)
$\tau = 0.7$	LASSO	19.37 (10.40)	98.37% (11.37%)	67.41% (22.68%)	0.13 (0.51)
	SCAD	11.38 (9.70)	91.22% (16.51%)	85.53% (19.44%)	1.76 (3.13)
	MCP	11.25 (9.68)	91.43% (15.99%)	85.85% (19.39%)	1.73 (3.05)

According to Table 3.1, the proposed approach with different penalty function can select the true variables with high probability and small estimation bias. The performance is the best when $\tau = 0.5$, as shown by higher variable selection sensitivity and specificity as well as smaller MSE. When the number of variables p is larger, the performance of median estimation is still comparable to the case when p is smaller. However, when $\tau = 0.3$ or $\tau = 0.7$, it is harder to distinguish the noise variables which are in A_τ^c. Variable selection specificity drops from more than 90% to about 70% for all penalty functions. MSE is larger than the scenario with $p = 4$. In addition, for all examples, model with SCAD penalty performs similar to model with MCP penalty. Model with LASSO penalty usually intends to select more variables comparing with model with SCAD or MCP penalty. All the observations are consistent with the existing theories.

3.5.2 Data center example

A data center is a computing infrastructure facilities that house large amounts of information technology (IT) equipment used to process, store, and transmit digital information. Data center facilities constantly generate large amounts of heat to the room, which must be maintained at an acceptable temperature for reliable operation of the equipment. More discussions of data center can be found in Hung et al. (2012). A significant fraction of the total power consumption in a data center is for heat removal; therefore, determining the most efficient cooling mechanism has become a major challenge. The objective of a thermal management study is to model the thermal distribution in a data center and the final goal is to design a data center with an efficient heat-removal mechanism.

For a data center thermal study, physical experiments are not always feasible because

some settings are highly dangerous and expensive to perform. Therefore, simulations based on computational fluid dynamics (CFD) are widely used. In this example, CFD simulations are conducted at IBM T. J. Watson Research Center based on a real data center layout. Detailed discussions about the CFD simulations can be found in López and Hamann (2011). The first three columns in Table 3.2 list nine factors and their levels in the CFD simulations, including four computer room air conditioning (CRAC) units with different flow rates $(x_1, ..., x_4)$, the overall room temperature setting (x_5), the perforated floor tiles with different percentage of open areas (x_6), and spatial location in the data center (x_7 to x_9). There are 27,000 temperatures simulated from the CFD simulator and these temperature outputs are obtained from an irregular grid over the 9-dimensional experimental space. It is of interest to know which CRAC would impact the room temperatures, especially the place with extreme temperatures. Therefore, we conduct quantile regression with ALP analysis with $\tau = 0.1$, $\tau = 0.5$ and $\tau = 0.9$. Since the dataset is huge and it is infeasible to apply quantile regression with ALP to the entire dataset. We randomly select 500 samples for illustration. L_1-norm penalty is applied. The results are shown in Table 3.2.

According to Table 3.2, tile open area percentage, location in y-axis and height have impact on 10% temperature, median temperature and 90% temperature. However, unit 2 can influence the areas with low temperature and median temperature but not areas with high temperature. On the other hand, unit 1, unit 3 can influence the areas with high temperature and median temperature but not areas with low temperature. In addition, unit 4 and location in x-axis can only reduce the temperature in areas with high temperature.

Table 3.2: Quantile analysis of thermal management data

	Variable	Levels	$\tau = 0.1$	$\tau = 0.5$	$\tau = 0.9$
x_0	Intercept	-	0.99	39.60	56.20
x_1	CRAC unit 1 flow rate (cfm)	(0,7000,8500,10000 11500,13000)	-	-5.39	-7.53
x_2	CRAC unit 2 flow rate (cfm)	(0,7000,8500,10000 11500,13000)	-5.04	-8.14	-
x_3	CRAC unit 3 flow rate (cfm)	(0,2500,4000,5500)	-	-2.77	-5.89
x_4	CRAC unit 4 flow rate (cfm)	(0,2500,4000,5500)	-	-	-4.90
x_5	Room temperature setting (F)	(65,67,69,71,73, 75)	-	-	-
x_6	Tile open area percentage (%)	$(15, 25, 35, 45$ $(55, 65, 75)$	0.05	0.26	0.05
x_7	Location in x-axis	8 unequally spaced	-	-	-3.38
x_8	Location in y-axis	4 unequally spaced	4.91	4.68	4.35
x_9	Height	18 equally spaced	33.18	34.46	32.71

3.6 Discussion

In this chapter, we present a penalized quantile regression with asymmetric Laplace process that not only incorporate correlations among observations but also select relevant covariates simultaneously. The proposed model is built under the assumption that for different quantiles, the set of covariates that can impact conditional distribution of response can be different. We also establish asymptotic model selection consistency and provide an algorithm for the proposed penalized estimators.

Future work can be explored in the following directions. First, considering infinite collection of quantiles with different relevant covariates, no oracle property has been established for penalized quantile regression with or without data dependence. Asymptotic distributions of penalized estimators are not yet known. Second, due to complicated likelihood function and penalty functions, it is computationally intensive

to obtain penalized estimators for quantile regression with asymmetric Laplace distribution. An efficient algorithm is desired for practical use on large datasets.

3.7 Appendix

3.7.1 Lemma

Lemma 3.1 *Under condition A2 and assume $p = o(n^{1/4})$, we have*

$$\|\nabla^2 \ell_\tau(\boldsymbol{\beta}) - J_\tau(\boldsymbol{\beta})\|_2 = o_p(1/p)$$

and

$$\|n(\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})(\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})^T - I_\tau(\boldsymbol{\beta})\|_2 = o_p(1/p).$$

Proof: By Chebyshev's inequality, for any $\epsilon > 0$,

$$\begin{aligned} & P(\|\nabla^2 \ell_\tau(\boldsymbol{\beta}) - J_\tau(\boldsymbol{\beta})\|_2 \geq \epsilon/p) \\ \leq \; & \frac{p^2}{\epsilon^2} \sum_{i,j=1}^{p} E[\frac{\partial^2 \ell_\tau(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} - E\{\frac{\partial^2 \ell_\tau(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}\}]^2 = O(p^4/n). \end{aligned}$$

Similarly, we can prove $\|n(\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})(\frac{\partial \ell_\tau(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}})^T - I_\tau(\boldsymbol{\beta})\|_2 = o_p(1/p)$.

3.7.2 Proof of Theorem 3.1

Let $\alpha_n = \sqrt{p/n}$ and set $\|\boldsymbol{w}\|_2 = C$ where C is a large enough constant. To prove theorem 1, we only need to show that for any given ϵ, there exists a C such that

$$P(\sup_{\|\boldsymbol{w}\|_2 = C} Q_\tau(\beta_\tau^0 + \alpha_n \boldsymbol{w}) < Q_\tau(\boldsymbol{\beta}_\tau^0)) \geq 1 - \epsilon.$$

Since $\rho_{\lambda_\tau}(0) = 0$, we have

$$\begin{aligned} & Q_\tau(\beta_\tau^0 + \alpha_n \boldsymbol{w}) - Q_\tau(\boldsymbol{\beta}_\tau^0)) \\ \leq \; & \{\ell_\tau(\beta_\tau^0 + \alpha_n \boldsymbol{w}) - \ell_\tau(\beta_\tau^0)\} - \sum_{j \in A_\tau} \{\rho_{\lambda_\tau}(|\beta_{\tau,j}^0 + \alpha_n w_j|) - \rho_{\lambda_\tau}(|\beta_{\tau,j}^0|)\} \\ = \; & I_1 + I_2. \end{aligned}$$

Using Taylor expansion, we have

$$\begin{aligned} I_1 &= \alpha_n \nabla^T \ell_\tau(\boldsymbol{\beta}_\tau^0)\boldsymbol{w} + \frac{\alpha_n^2}{2}\boldsymbol{w}^T\nabla^2\ell_\tau(\boldsymbol{\beta}_\tau^0)\boldsymbol{w} + \frac{\alpha_n^3}{6}\nabla\{\boldsymbol{w}^T\nabla^2\ell_\tau(\boldsymbol{\beta}_\tau^*)\boldsymbol{w}\}\boldsymbol{w} \\ &= I_{11} + I_{12} + I_{13}. \end{aligned}$$

By condition A1, $|I_{11}| \le \alpha_n\|\nabla^T\ell_\tau(\boldsymbol{\beta}_\tau^0)\|_2\|\boldsymbol{w}\|_2 = \alpha_n O_p(\sqrt{p/n})\|\boldsymbol{w}\|_2$. With respect to I_{12}, by Lemma 3.1, we have

$$\begin{aligned} I_{12} &= \frac{\alpha_n^2}{2}\boldsymbol{w}^T\{\nabla^2\ell_\tau(\boldsymbol{\beta}_\tau^0) - E\nabla^2\ell_\tau(\boldsymbol{\beta}_\tau^0)\}\boldsymbol{w} - \frac{\alpha_n^2}{2}\boldsymbol{w}^T J_\tau(\boldsymbol{\beta}_\tau^0)\boldsymbol{w} \\ &= -\frac{\alpha_n^2}{2}\boldsymbol{w}^T J_\tau(\boldsymbol{\beta}_\tau^0)\boldsymbol{w} + O_p(1/p)\alpha_n^2\|\boldsymbol{w}\|_2^2. \end{aligned}$$

By condition A3, we have

$$\begin{aligned} |I_{13}| &= |\frac{\alpha_n^3}{6}\nabla\{\boldsymbol{w}^T\nabla^2\ell_\tau(\boldsymbol{\beta}_\tau^*)\boldsymbol{w}\}\boldsymbol{w}| \\ &\le \frac{\alpha_n^3}{6}(\sum_{i,j,k} M_{i,j,k}^2)^{1/2}\|\boldsymbol{w}\|_2^3 = O_p(p^{3/2}\alpha_n^3)\|\boldsymbol{w}\|_2^3. \end{aligned}$$

In addition,

$$\begin{aligned} I_2 &= -\sum_{j\in A_\tau}[\rho'_{\lambda_\tau}(|\beta_{\tau,j}^0|)\mathrm{sgn}(\beta_{\tau,\mathrm{j}}^0)\alpha_\mathrm{n}\mathrm{w_j} + \rho''_{\lambda_\tau}(|\beta_{\tau,\mathrm{j}}|)\mathrm{w_j^2}\alpha_\mathrm{n}^2\{1+\mathrm{o}(1)\}] \\ &= I_{21} + I_{22}. \end{aligned}$$

By condition A6 and A7, we have

$$|I_{21}| \le \sum_{j\in A_\tau}|\rho'_{\lambda_\tau}(|\beta_{\tau,j}^0|)\alpha_n w_j| \le \sqrt{s_\tau}\alpha_n\|\boldsymbol{w}\|_2\max_{j\in A_\tau}|\rho'_{\lambda_\tau}(|\beta_{\tau,j}^0|)| = o(\alpha_n\sqrt{p/n}),$$

and $|I_{22}| \le \alpha_n^2\|\boldsymbol{w}\|_2^2\max_{j\in A_\tau}|\rho''_{\lambda_\tau}(|\beta_{\tau,j}^0|)| = o(\alpha_n^2)$.

In total, all the terms are dominated by I_{12} which is negative.

3.7.3 Proof of Theorem 3.2

We first show the sparsity, that is $\hat{\boldsymbol{\beta}}_{A_\tau^c} = 0$. Let $\epsilon = C\sqrt{p/n}$. It is sufficient to show that with probability approaching 1 as $n \to 1$, for any $\boldsymbol{\beta}_{A_\tau}$ such that $\|\boldsymbol{\beta}_{A_\tau} - \boldsymbol{\beta}_{\tau,A_\tau}^0\|_2 =$

$O_p(\sqrt{p/n})$, we have for $i \in A_\tau^c$,

$$\frac{\partial Q_\tau(\boldsymbol{\beta})}{\partial \beta_i} < 0, \quad \text{for} \quad 0 < \beta_i < \epsilon$$
$$\frac{\partial Q_\tau(\boldsymbol{\beta})}{\partial \beta_i} > 0, \quad \text{for} \quad -\epsilon < \beta_i < 0.$$

By Taylor expansion, we have

$$\begin{aligned}
& \frac{\partial Q_\tau(\boldsymbol{\beta})}{\partial \beta_i} = \frac{\ell_\tau(\boldsymbol{\beta})}{\partial \beta_i} - \rho'_{\lambda_\tau}(|\beta_i|)\mathrm{sgn}(\beta_\mathrm{i}) \\
= & \frac{\ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i} + \sum_{j=1}^{p} \frac{\partial^2 \ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i \partial \beta_j}(\beta_j - \beta_{\tau,j}^0) \\
& + \sum_{j,k=1}^{p} \frac{\partial^3 \ell_\tau(\boldsymbol{\beta}_\tau^*)}{\partial \beta_i \partial \beta_j \partial \beta_k}(\beta_j - \beta_{\tau,j}^0)(\beta_k - \beta_{\tau,k}^0) - \rho'_{\lambda_\tau}(|\beta_i|)\mathrm{sgn}(\beta_\mathrm{i}) \\
= & I_1 + I_2 + I_3 + I_4,
\end{aligned}$$

where $\boldsymbol{\beta}_\tau^*$ lies between $\boldsymbol{\beta}_\tau^0$ and $\boldsymbol{\beta}$.

Using Chebyshev's inequality, we can show that $|I_1| = O_p(1/\sqrt{n})$. Rewrite I_2, we have

$$\begin{aligned}
|I_2| &= \sum_{j=1}^{p} \{\frac{\partial^2 \ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i \partial \beta_j} - E\frac{\partial^2 \ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i \partial \beta_j}\}(\beta_j - \beta_{\tau,j}^0) + \sum_{j=1}^{p} E\frac{\partial^2 \ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i \partial \beta_j}(\beta_j - \beta_{\tau,j}^0) \\
&= I_{21} + I_{22}.
\end{aligned}$$

By Cauchy-Schwarz inequality and Lemma 3.1, we have

$$|I_{21}| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}_\tau^0\|_2 [\sum_{j=1}^{p} \{\frac{\partial^2 \ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i \partial \beta_j} - E\frac{\partial^2 \ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i \partial \beta_j}\}^2]^{1/2} = O_p(n^{-1/2}p^{-1/2}),$$

and since the eigenvalues of $J_\tau(\boldsymbol{\beta})$ are bounded

$$|I_{22}| \le \|\nabla \frac{\partial \ell_\tau(\boldsymbol{\beta}_\tau^0)}{\partial \beta_i}\|_2 \|\boldsymbol{\beta} - \boldsymbol{\beta}_\tau^0\|_2 = O_p(\sqrt{p/n}).$$

Similarly, rewrite I_3 as

$$\begin{aligned} I_3 &= \sum_{j,k=1}^{p} \{\frac{\partial^3 \ell_\tau(\boldsymbol{\beta}_\tau^*)}{\partial\beta_i\partial\beta_j\partial\beta_k} - E\frac{\partial^3 \ell_\tau(\boldsymbol{\beta}_\tau^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}\}(\beta_j - \beta_{\tau,j}^0)(\beta_k - \beta_{\tau,k}^0) \\ &\quad + \sum_{j,k=1}^{p} E\frac{\partial^3 \ell_\tau(\boldsymbol{\beta}_\tau^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}(\beta_j - \beta_{\tau,j}^0)(\beta_k - \beta_{\tau,k}^0) \\ &= I_{31} + I_{32}. \end{aligned}$$

By condition A3 and Cauchy-Schwarz inequality, we have

$$|I_{31}| \leq [\sum_{j,k=1}^{p} \{\frac{\partial^3 \ell_\tau(\boldsymbol{\beta}_\tau^*)}{\partial\beta_i\partial\beta_j\partial\beta_k} - E\frac{\partial^3 \ell_\tau(\boldsymbol{\beta}_\tau^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}\}^2]^{1/2}\|\boldsymbol{\beta} - \boldsymbol{\beta}_\tau^0\|_2^2 = O_p(p^2/n) = o_p(\sqrt{p/n}),$$

and

$$|I_{32}|^2 \leq [\sum_{j,k=1}^{p} E\{\frac{\partial^3 \ell_\tau(\boldsymbol{\beta}_\tau^*)}{\partial\beta_i\partial\beta_j\partial\beta_k}\}^2]\|\boldsymbol{\beta} - \boldsymbol{\beta}_\tau^0\|_2^4 = O_p(p^4/n^2).$$

Therefore, we have $I_1 + I_2 + I_3 = O_p(\sqrt{p/n})$. By condition A4, from

$$\begin{aligned} \frac{\partial Q_\tau(\boldsymbol{\beta})}{\partial\beta_i} &= -\rho'_{\lambda_\tau}(|\beta_i|)\mathrm{sgn}(\beta_\mathrm{i}) + \mathrm{O_p}(\sqrt{\mathrm{p/n}}) \\ &= -\lambda_\tau\{\rho'_{\lambda_\tau}(|\beta_i|)\mathrm{sgn}(\beta_\mathrm{i})/\lambda_\tau + \mathrm{O_p}(\sqrt{\mathrm{p/n}}/\lambda_\tau)\}, \end{aligned}$$

we can see that the sign of β_i determines the sign of the derivative completely.

Bibliography

Amos, D. (1974), "Computation of modified Bessel functions and their ratios," *Mathematics of Computation*, 28, 239–251.

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

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J., and Walsh, D. (2007), "Computer model validation with functional output," *The Annals of Statistics*, 35, 1874–1906.

Bayarri, M., Berger, J. O., Kennedy, M. C., Kottas, A., Paulo, R., Sacks, J., Cafeo, J. A., Lin, C.-H., and Tu, J. (????), "Predicting vehicle crashworthiness: Validation of computer models for functional and hierarchical data," *Journal of the American Statistical Association*, 104, 929–943.

Belloni, A., Chernozhukov, V., et al. (2011), "1-penalized quantile regression in high-dimensional sparse models," *The Annals of Statistics*, 39, 82–130.

Boukouvalas, A., Barillec, R., and Cornford, D. (2012), "Gaussian Process Quantile Regression using Expectation Propagation," *arXiv preprint arXiv:1206.6391*.

Chaudhuri, P., Doksum, K., Samarov, A., et al. (1997), "On average derivative quantile regression," *The Annals of Statistics*, 25, 715–744.

Chernozhukov, V. (2005), "Extremal quantile regression," *The Annals of Statistics*, 23, 806–839.

Chu, T., Zhu, J., Wang, H., et al. (2011), "Penalized maximum likelihood estimation and variable selection in geostatistics," *The Annals of Statistics*, 39, 2607–2625.

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

Cressie, N. A. and Cassie, N. A. (1993), *Statistics for spatial data*, vol. 900, Wiley New York.

DiCiccio, T. and Efron, B. (1992), "More accurate confidence intervals in exponential families," *Biometrika*, 79, 231–245.

DiCiccio, T. J. and Efron, B. (1996), "Bootstrap confidence intervals," *Statistical Science*, 11, 189–212.

Efron, B. (1979), "Bootstrap methods: another look at the jackknife," *The Annals of Statistics*, 7, 1–26.

Efron, B. and Tibshirani, R. J. (1994), *An introduction to the bootstrap*, vol. 57, Chapman and Hall/CRC press, New York.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Peng, H. (2004), "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, 32, 928–961.

Fang, K.-T., Li, R., and Sudjianto, A. (2006), *Design and modeling for computer experiments*, Chapman and Hall/CRC press, New York.

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

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

Gonçalves, S. and White, H. (2004), "Maximum likelihood and the bootstrap for nonlinear dynamic models," *Journal of Econometrics*, 119, 199–219.

Gramacy, R. B. and Apley, D. W. (2013), "Local Gaussian process approximation for large computer experiments," *arXiv preprint arXiv:1303.0383*.

Gramacy, R. B. and Lee, H. K. (2008), "Bayesian treed Gaussian process models with an application to computer modeling," *Journal of the American Statistical Association*, 103, 1119–1130.

Gutenbrunner, C. and Jurecková, J. (1992), "Regression rank scores and regression quantiles," *The Annals of Statistics*, 20, 305–330.

Hall, P., Horowitz, J. L., and Jing, B.-Y. (1995), "On blocking rules for the bootstrap with dependent data," *Biometrika*, 82, 561–574.

Hung, Y., Qian, P. Z. G., and Wu, C. F. J. (2012), "Statistical design and analysis methods for data center thermal management," *Energy efficient thermal management of data centers*, (J. Yogendra and K. Pramod eds.).

Irvine, K. M., Gitelman, A. I., and Hoeting, J. A. (2007), "Spatial designs and properties of spatial correlation: effects on covariance estimation," *Journal of agricultural, biological, and environmental statistics*, 12, 450–469.

Jennrich, R. I. (1969), "Asymptotic properties of non-linear least squares estimators," *The Annals of Mathematical Statistics*, 40, 633–643.

Jin, J. and Zhang, S. (1996), *Computation of special functions*, Wiley-Interscience, New York.

Kaufman, C. G., Bingham, D., Habib, S., Heitmann, K., Frieman, J. A., et al. (2011), "Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology," *The Annals of Applied Statistics*, 5, 2470–2492.

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

Koehler, J. and Owen, A. (1996), "Computer experiments," *Handbook of statistics*, 13, 261–308.

Koenker, R. and Bassett Jr, G. (1978), "Regression quantiles," *Econometrica: journal of the Econometric Society*, 46, 33–50.

Kozubowski, T. J. and Podgorski, K. (2000), "A multivariate and asymmetric generalization of Laplace distribution," *Computational Statistics*, 15, 531–540.

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

Lahiri, S. (1995), "On the asymptotic behaviour of the moving block bootstrap for normalized sums of heavy-tail random variables," *The Annals of Statistics*, 23, 1331–1349.

Lahiri, S. N. (1999), "Theoretical comparisons of block bootstrap methods," *The Annals of Statistics*, 27, 386–404.

— (2003), *Resampling methods for dependent data*, Springer.

Li, R. and Sudjianto, A. (2005), "Analysis of computer experiments using penalized likelihood in Gaussian Kriging models," *Technometrics*, 47, 111–120.

Liang, F., Cheng, Y., Song, Q., Park, J., and Yang, P. (2013), "A Resampling-Based Stochastic Approximation Method for Analysis of Large Geostatistical Data," *Journal of the American Statistical Association*, 108, 325–339.

Liu, R. Y. and Singh, K. (1992), "Moving blocks jackknife and bootstrap capture weak dependence," *Exploring the limits of bootstrap*, 225, 226–249.

Loh, W.-L. (1996), "On Latin hypercube sampling," *The Annals of Statistics*, 24, 2058–2080.

López, V. and Hamann, H. F. (2011), "Heat transfer modeling in data centers," *International Journal of Heat and Mass Transfer*, 54, 5306–5318.

Lum, K., Gelfand, A. E., et al. (2012), "Spatial quantile multiple regression using the asymmetric Laplace process," *Bayesian Analysis*, 7, 235–258.

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

McKay, M. D., Beckman, R. J., and Conover, W. J. (1979), "Comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, 21, 239–245.

Nordman, D. J., Lahiri, S. N., and Fridley, B. L. (2007), "Optimal block size for variance estimation by a spatial block bootstrap method," *Sankhyā: The Indian Journal of Statistics*, 69, 468–493.

Nychka, D., Wikle, C., and Royle, J. A. (2002), "Multiresolution models for nonstationary spatial covariance functions," *Statistical Modelling*, 2, 315–331.

Nychka, D. W. (2000), "Spatial-process estimates as smoothers," *Smoothing and regression: approaches, computation, and application*, 393–424.

Nychka, D. W., Haaland, P., O'Connell, M., and Ellner, S. (1998), "FUNFITS, data analysis and statistical tools for estimating functions," *Lecture notes in statistics*, 159–180.

— (2011), "Accurate emulators for large-scale computer experiments," *The Annals of Statistics*, 39, 2974–3002.

Owen, A. B. (1994), "Controlling correlations in Latin hypercube samples," *Journal of the American Statistical Association*, 89, 1517–1522.

Paparoditis, E. and Politis, D. N. (2001), "Tapered block bootstrap," *Biometrika*, 88, 1105–1119.

Peng, C.-Y. and Wu, C. J. (2014), "On the choice of nugget in kriging modeling for deterministic computer experiments," *Journal of Computational and Graphical Statistics*, 23, 151–168.

Politis, D. N. and Romano, J. P. (1994), "The stationary bootstrap," *Journal of the American Statistical Association*, 89, 1303–1313.

Qian, P. Z., Ai, M., and Wu, C. J. (2009), "Construction of nested space-filling designs," *The Annals of Statistics*, 37, 3616–3643.

Qian, P. Z. and Wu, C. J. (2009), "Sliced space-filling designs," *Biometrika*, 96, 945–956.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011), "Bayesian spatial quantile regression," *Journal of the American Statistical Association*, 106, 6–20.

Rougier, J. (2008), "Efficient emulators for multivariate deterministic functions," *Journal of Computational and Graphical Statistics*, 17, 827–843.

Rue, H. and Held, L. (2005), *Gaussian Markov random fields: theory and applications*, Chapman and Hall/CRC Press, Boca Raton.

Rue, H. and Tjelmeland, H. (2002), "Fitting Gaussian Markov random fields to Gaussian fields," *Scandinavian Journal of Statistics*, 29, 31–49.

Sacks, J., Schiller, S. B., and Welch, W. J. (1989a), "Designs for computer experiments," *Technometrics*, 31, 41–47.

Sacks, J., Welch, W. J., Mitchell, T. J., Wynn, H. P., et al. (1989b), "Design and analysis of computer experiments," *Statistical science*, 4, 409–423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The design and analysis of computer experiments*, Springer.

Sjöstedt-de Luna, S. (2003), "The bootstrap and kriging prediction intervals," *Scandinavian journal of statistics*, 30, 175–192.

Smola, A. J. and Bartlett, P. L. (2001), "Sparse greedy Gaussian process regression," *Advances in Neural Information Processing Systems*, 13, 619–625.

Snelson, E. and Ghahramani, Z. (2006), "Sparse Gaussian processes using pseudo-inputs," *Advances in neural information processing systems*, 18, 1257–1264.

Stein, M. L. (1999), *Interpolation of spatial data: some theory for kriging*, Springer.

— (2013), "Statistical properties of covariance tapers," *Journal of Computational and Graphical Statistics*, 22, 866–885.

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

Tang, B. (1993), "Orthogonal array-based Latin hypercubes," *Journal of the American Statistical Association*, 88, 1392–1397.

— (1994), "A theorem for selecting OA-based Latin hypercubes using a distance criterion," *Communications in Statistics-Theory and Methods*, 23, 2047–2058.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 267–288.

Wang, L., Wu, Y., and Li, R. (2012), "Quantile regression for analyzing heterogeneity in ultra-high dimension," *Journal of the American Statistical Association*, 107, 214–222.

White, H. (1996), *Estimation, inference and specification analysis*, no. 22, Cambridge university press.

Wikle, C. K. (2010), "Low-rank representations for spatial processes," *Handbook of Spatial Statistics*, 107–118.

Ye, K. Q. (1998), "Orthogonal column Latin hypercubes and their application in computer experiments," *Journal of the American Statistical Association*, 93, 1430–1439.

Ying, Z. (1993), "Maximum likelihood estimation of parameters under a spatial sampling scheme," *The Annals of Statistics*, 21, 1567–1590.

Zhang, C. (2010), "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38, 894–942.

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

Zou, H. and Li, R. (2008), "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of statistics*, 36, 1509–1533.

Zou, H. and Yuan, M. (2008), "Composite quantile regression and the oracle model selection theory," *The Annals of Statistics*, 36, 1108–1126.