FUSION LEARNING OF DEPENDENT STUDIES BY CONFIDENCE DISTRIBUTION (CD): THEORY AND APPLICATIONS

BY CHENGRUI LI

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

Minge Xie

and approved by

_____

_____

_____

_____

New Brunswick, New Jersey

May, 2017

ABSTRACT OF THE DISSERTATION

Fusion Learning of Dependent Studies by Confidence Distribution (CD): Theory and Applications

by CHENGRUI LI

Dissertation Director: Minge Xie

This dissertation focuses on developing efficient Fusion Learning methodologies for combining information from non-independent sources using *confidence distribution* (CD). The sources hereby are broadly construed as different pieces of information extracted from possibly correlated datasets. This situation typically arises when multiple inferences are performed over different times, locations or experiment settings due to computational and statistical considerations, which encompasses a wide range of scientific and engineering applications (e.g. seismic monitoring and detection, computer experiments). In this dissertation, we develop a general framework to effectively and efficiently combine these correlated information using CD, and furthermore, explore the advantages of this framework in different problems that are of theoretical and practical interests.

The first problem we address is related to analysis of big spatial data. We

propose a sequential split-and-conquer method based on CD to deal with the long-standing issues on computational scalability and predictive uncertainty in Gaussian process (GP) models when the sample size is large. The CD-based combining approach we propose in this work aims to intelligently divide the large-scale problem into smaller sub-problems, and combine the result of each sub-problem strategically without the loss of statistical efficiency under mild conditions. The other problem we visit is the derivation of a general CD-based framework for combining non-independent inference results under a partial linear model scenario, where the inference result is in the form of local estimates from partial linear regression models.

The results discussed in this dissertation further shed light on the potential applications of our CD-based combining framework in the high-level fusion of statistical inference results as a powerful meta-analysis method, or more generally speaking, Fusion Learning.

Acknowledgements

First and foremost, I would like to express my gratitude to my advisor, Professor Minge Xie, who has the attitude to convincingly and continually convey a spirit of adventure in regard to research. I'm inspired by his motivation and enormous dedication to scientific research. I thank him for the guidance, encouragement and help during the past five years. Without his constant support, this dissertation would not have been possible to be finished.

I would like to extend my gratitude to Professor Regina Liu, Professor Ying Hung and Professor Jerry Cheng for their time and efforts to serve on my thesis committee. Especially, I would like to thank Professor Ying Hung, who always patiently and generously offers her help when I have questions. I really appreciate her invaluable direction and brilliant ideas on the projects we worked on together.

My heartfelt thanks also go to the Department of Statistics and Biostatistics at Rutgers University for an excellent research environment, and continuous financial support throughout my graduate study; to our graduate director Professor John Kolassa, who always offers his generous help; to Professor Cun-Hui Zhang for his help and support; to my internship mentors from Bell Labs, Dr. Jin Cao and Dr. Sining Chen, for their guidance on the the application of statistical methods from a practical perspective; to my fellow PhD students, especially Yi Fan, Chuan Liu and Jieli Shen, for their company and morale supports; to Kun Chang, Yang Jiao, Xialu Liu and Yufan Liu for their help on my job search.

Last but not least, I would like to give my deepest gratitude to my family, especially my parents, Junqing Li and Yongzhen Deng. This dissertation stands

as a testament to their unconditional love and encouragement.

Dedication

To my home.

致故乡

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

There have been rapid developments in statistical methodologies on combining information from multiple sources, such as Fusion Learning, Meta-analysis, Divide-and-Conquer algorithm etc, during the past few decades. In particular, the method of confidence distribution (CD) has gained tremendous popularity in fusing inference results from different studies, where the goal is to estimate the common parameters of interest across these studies by combining various CDs constructed individually. However, the majority of prior work heavily relies on the assumption that individual studies are independent, which may not be valid under many circumstances (e.g. spatial-temporal data). More specifically, in this dissertation, we will discuss the following two problems.

- **Problem 1 (Analysis of Big Spatial Data)** The task of analyzing massive spatial data is extremely challenging because of the correlation among the observations. Among the techniques developed for analyzing spatial data, Gaussian process model is one of the most widely used approaches in the literature, see Sacks et al. (1989), Santner et al. (2003) etc. However, two critical issues remain unresolved. One is the computational issue in GP estimation and prediction where intensive manipulations of an $n \times n$ correlation matrix are required and become infeasible for large sample size n. The other is how to improve the naive plug-in predictive distribution which is known to underestimate the uncertainty.

- **Problem 2 (Combining Dependent Studies)** Conventional combining methodologies focus on integrating information without considering the dependency among the individual studies. One specific example is that, when combining log-likelihood functions, usually we take the summation of individual ones. However, this approach will fail if those log-likelihood functions are dependent Fraser and Reid (2015). Another scenario relates to combining the local estimates of parametric components in the partial linear models. The local estimators are derived based on the non-parametric components and local likelihood functions and therefore there will be dependency among those local results. Traditional combining approaches will not be feasible in this case.

To address the two challenging problems above, we adopt the concept of CD to develop innovative approaches to combining information without the assumption of independence. CD is defined as sample-dependent distribution function that can represent the confidence regions of all levels of parameter of interest (Singh et al., 2007). One of the attractive features of CD is that it contains wealth information for making Frequentist inference. Due to this advantage of CD, the CD-based combining approaches are widely employed and adopted in the literature (Claggett et al., 2014; Yang et al., 2014; Liu et al., 2015). In this dissertation, we expand the combining recipe to those dependent studies.

- For Problem 1, we propose a Sequential Split-Conquer-Combine (SSCC) approach to analyze the big spatial data under a GP model setting. The SSCC approach that can tackle the two issues discussed above simultaneously by providing estimators and predictors that maintain the same asymptotic efficiency as the conventional method but reduce the computation dramatically. Moreover, the CD-based predictive distribution contains comprehensive information for statistical inference and provides a

better quantification of predictive uncertainty comparing with the plug-in approach. Simulations are conducted to evaluate the accuracy and computational gains. The proposed framework is demonstrated by a data center example based on tens of thousands of computer experiments generated from a computational fluid dynamic simulator.

- For Problem 2, we propose a general combining approach by using CDs to integrate information from non-independent studies. The idea is inspired by the methodology suggested by Singh et al. (2005) applying to the independent case. We first illustrate the proposed approach on a simple example that combining dependent log-likelihood functions. We also focus on combining the parametric components from partial linear models. The theoretical derivation and numerical validation demonstrate that our general framework can effectively combine non-independent studies and lead to results as if they were derived over the entire raw dataset.

Chapter 2

A Sequential Split-Conquer-Combine Approach for Gaussian Process Modeling in Computer Experiments

2.1 Introduction

Gaussian process (GP) models, also known as kriging, are commonly used in many applications including geostatistics and machine learning. In recent years there has been a growing interest in GP models for the analysis of computer experiments, which is important in science, engineering and medicine (Sacks et al., 1989). Computer experiments refer to the study of real systems using 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. Typically, computer experiments are computationally demanding and their outputs are deterministic in the sense that running the code twice with the same set of input values will produce the same output. Therefore, it is desirable to build an interpolator for the computer experiment outputs and use it as an emulator for the actual computer experiment. In the literature, GP models are extensively used as an interpolator in the analysis of computer experiments. Comparing with conventional applications in geostatistics, computer experiments often involve more variables in their GP models. More discussions of computer experiments can be found in Sacks et al. (1989), Fang et al. (2006), and Santner et al. (2003).

There are two critical issues that have not been solved satisfactorily in GP modeling in the field. The first one is the computational issue. This is because estimation and prediction of GP heavily involve manipulations of the n-by-n correlation matrix, where n is the sample size, that require $O(n^3)$ computations and often result in singularity. This issue has been well recognized in the literature, and the proposed approaches can 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 includes Rue and Tjelmeland (2002), Rue and Held (2005), Cressie and Johannesson (2008), Banerjee et al. (2008), Gramacy and Lee (2008), Wikle (2010), Chang et al. (2014); while approximation approaches includes Nychka (2000), Smola et al. (2001), Nychka et al. (2002), Stein et al. (2004), Snelson and Ghahramani (2005), Furrer et al. (2006),Fuentes (2007), Kaufman et al. (2008), Liang et al. (2013), Gramacy and Apley (2015), and Nychka et al. (2015). Recent studies address the issues by imposing a sparsity constraint on the correlation matrix. Examples including correlation tapering (Kaufman et al., 2008; Stein, 2013) and compact support correlation (Gneiting, 2002; Stein, 2008; Kaufman et al., 2011). However, it has been shown that these methods can display sizable bias in parameter estimation unless the taper/band range is large and, when the taper/band range is large, although the bias is reduced, their computational complexity and burden will be significantly increased (Kaufman et al., 2008; Stein, 2013; Liang et al., 2013). In addition, the connection between the degree of sparsity and computation time is nontrivial.

The second issue is how to accurately quantify the uncertainty in GP modeling. It is well-known that the predictive distributions constructed by substituting the true parameters by estimators, often called plug-in predictive distributions, underestimate the uncertainty (Santner et al., p.98). However, they are still widely used due to the lack of computationally efficient alternatives. Alternative

approaches, for examples, predictive distributions constructed by bootstrap (Sjöstedt-de Luna, 2003; Santner et al., 2003) or Bayesian procedure (Kennedy and O'Hagan, 2001; Schmidt and O'Hagan, 2003) provide better quantification of uncertainty but they require even more intensive computation because more manipulations of the correlation matrix are involved.

Although numerous methods have been proposed to address these issues, to the best of our knowledge, they are developed typically for solving one of the issues leaving the other questionable. So our goal is to introduce an unified framework for GP models that can address both issues simultaneously. This framework is called a *s*equential-*s*plit-*c*onquer-*c*ombine (SSCC) approach, which consists of a sequential split-conquer procedure and an information combining technique using confidence distributions (CDs) and a CD-based predictive distribution (Singh et al., 2005; Yang et al., 2014; Liu et al., 2015). The sequential split-conquer procedure reduces the computational complexity by splitting the data into smaller subsets and allowing estimation to be performed on the subsets individually. Although similar ideas of data splitting are discussed in the literature (Stein, 2013; Chen and Xie, 2014a; Mackey et al., 2015), information from individual subsets are often assumed to be independent which is violated in GP modeling. In contrast, the proposed sequential split-conquer procedure takes into account the data dependency by updating information sequentially from neighborhood subsets so that the estimation efficiency can be enhanced. After splitting the data into subsets, individual information from each subset is combined using a CD technique which provides not only an efficient combined estimate but also a flexible tool for inference. Last, an easy-to-compute GP predictor is introduced and a CD-based predictive distribution is constructed. Apart from the computational reduction, the proposed framework provides combined estimates and predictions that is asymptotically equivalent to the conventional ones under very mild conditions. Furthermore, it provides comprehensive information

for statistical inference and a better quantification of predictive uncertainty comparing with the plug-in approach.

The remainder of this section is organized as follows. In Section 2.2, the standard procedure of estimation and prediction are reviewed for GP models. The unified framework is introduced in Section 2.3. In Section 2.4, simulations are presented to demonstrate the performance of the proposed framework. In Section 2.5, the proposed approach is applied to a data center thermal management study. Summary and concluding remarks are given in Section 2.6.

2.2 Gaussian Process Models

A Gaussian process model can be written as

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

where $y \in \mathbb{R}$ is the output, $\boldsymbol{x} \in \mathbb{R}^p$ is the input, $\mu(\boldsymbol{x})$ is the mean function assumed to be a function of $\boldsymbol{x}$ with unknown parameters $\boldsymbol{\beta} \in \mathbb{R}^p$, say, $\mu(\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{\beta}$. In addition, $Z(\boldsymbol{x})$ is a Gaussian process with mean 0 and $\mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma^2\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$, where $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$ is the correlation function and $\boldsymbol{\theta}$ is a vector of unknown correlation parameters. There are various correlation functions discussed in the literature. Here we focus on a popular choice in computer experiments, a product form of power exponential functions (Sacks et al., 1989; Gramacy and Apley, 2015; Kaufman et al., 2011):

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \prod_{k=1}^{p} R_k(|x_{ik} - x_{jk}|) = \prod_{k=1}^{p} \exp(-\theta_k|x_{ik} - x_{jk}|^\alpha), \tag{2.2}$$

where $0 < \alpha \leq 2$ is a tuning parameter and $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)$ with $\theta_k \geq 0$ for all k. Since the correlation parameters θ_k's are not constrained to be equal, the model can handle different signals in each input dimension which makes (2.2) particularly attractive to the analysis of computer experiments. Note that in (2.2), as long as $|x_{ik} - x_{jk}| \to \infty$ for a single k with $\theta_k > 0$, $\mathrm{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) \to 0$.

Given n realizations $\boldsymbol{y} = (y_1, ..., y_n)^\top$ and the corresponding inputs $X = (\boldsymbol{x}_1^\top, ..., \boldsymbol{x}_n^\top)^\top$, the joint log-likelihood function for (2.1) can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) = -\frac{1}{2\sigma^2}(\boldsymbol{y} - X\boldsymbol{\beta})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\boldsymbol{\beta}) - \frac{1}{2}\log|\Sigma(\boldsymbol{\theta})| - \frac{n}{2}\log(\sigma^2), \quad (2.3)$$

where $\Sigma(\boldsymbol{\theta})$ is the $n \times n$ correlation matrix with the ijth element equals to $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$. The maximum likelihood estimates (MLEs) of $\boldsymbol{\beta}$ and σ can be obtained by

$$\widehat{\boldsymbol{\beta}} = (X^\top \Sigma^{-1}(\boldsymbol{\theta})X)^{-1} X^\top \Sigma^{-1}(\boldsymbol{\theta})\boldsymbol{y}, \quad (2.4)$$

$$\widehat{\sigma}^2 = (\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\widehat{\boldsymbol{\beta}})/n. \quad (2.5)$$

By maximizing the logarithm of the profile likelihood, the MLE of $\boldsymbol{\theta}$ can be obtained by

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \{n\log(\widehat{\sigma}^2) + \log|\Sigma^{-1}(\boldsymbol{\theta})|\}. \quad (2.6)$$

For the estimation of correlation parameters $\boldsymbol{\theta}$, there are some likelihood-based alternatives, including the restricted maximum likelihood (REML) (Irvine et al., 2007) and penalized likelihood approaches (Li and Sudjianto, 2005). In this paper, we focus on the study of MLEs but the results can be further extended to the likelihood-based alternatives.

When the parameters are known, the conditional distribution of y_0 at a new input $\boldsymbol{x}_0$, given the observations $\boldsymbol{y}$, is normal with mean $p_0(\boldsymbol{\beta}, \boldsymbol{\theta})$ and variance $m_0(\boldsymbol{\beta}, \boldsymbol{\theta})$, where

$$p_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{x}_0^\top \boldsymbol{\beta} + \gamma(\boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{\theta})(\boldsymbol{y} - X\boldsymbol{\beta}), \quad (2.7)$$

$$m_0(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sigma^2(1 - \gamma(\boldsymbol{\theta})^\top \Sigma^{-1}(\boldsymbol{\theta})\gamma(\boldsymbol{\theta})), \quad (2.8)$$

and $\gamma(\boldsymbol{\theta})$ is a $n \times 1$ vector with ith element equals to $\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta})$. In practice, when the parameters are unknown, the conventional plug-in approach constructs an predictive distribution by replacing the true parameters by their MLEs. Therefore

the (estimated) plug-in predictive distribution is normally distributed with mean $p_0(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ and variance $m_0(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$.

Calculating MLEs (2.4)-(2.5) and GP predictors (2.7)-(2.8) is computationally intensive because the calculation requires manipulations of a $n \times n$ correlation matrix Σ, such as Σ^{-1} and $|\Sigma|$, which are intractable for moderate sample sizes and infeasible for large sample sizes. On the other hand, ignoring the parameter uncertainty in the construction of plug-in predictive distribution clearly leads to an underestimation of predictive uncertainty.

A commonly used approach to reduce complexity of GP estimation and prediction is to introduce zeros into the correlation matrix and thus computationally efficient sparse matrix techniques (Pissanetzky, 1984; Barry and Pace, 1999) can be used. Methods along this line, such as compactly supported correlation functions and covariance tapering, have received increasing attention in the literature (Gneiting, 2002; Furrer et al., 2006; Kaufman et al., 2008, 2011; Bickel and Levina, 2008; Stein, 2008, 2013; Chu et al., 2011). The compactly supported correlation function introduce zeros into the correlation matrix by assuming

$$R_k(|x_{ik} - x_{jk}|) = 0, \quad \text{if } |x_{ik} - x_{jk}| \geq \tau_k, \tag{2.9}$$

for $\tau_k \geq 0$ and $k = 1, ..., p$. The tuning parameters τ_k is called the range parameters. Another commonly used approach is covariance tapering in which the covariance matrix is multiplied by a tapering function defined by a single range parameter. For these tapering-type of methods, the resulting estimates can display sizable bias when the range parameter is small relative to the true correlation range of the process (Kaufman et al., 2011). Therefore, larger τ_k is preferred for estimation purpose but it leads to a significant increase of computational complexity.

2.3 A Unified Framework by Sequential-Split-Conquer-Combine

2.3.1 Estimation of the mean function coefficient $\boldsymbol{\beta}$

We begin by illustrating the SSCC framework using a simple case where $\boldsymbol{\theta}$ and σ are known. There are two steps: first is to split and sequentially update the data and the second step is information combining using confidence distributions.

Step 1: Sequentially split and conquer

To reduce computation, a key idea of the unified framework is to splitting the data into smaller subsets and allow the estimation to be done individually within each subset. This concept is attractive and is discussed under various settings, including spatial-temporal models (Stein, 2013), matrix factorization in machine learning (Mackey et al., 2015), linear models, and penalized regressions (Chen and Xie, 2014a). However, most of the existing methods cannot take into account the dependency between subsets, which is crucial in the setting of GP models, and thus leads to a significant loss of efficiency in estimation. Therefore, a careful sequential information update is introduced in our procedure to incorporate the dependency between subsets and improve the estimation efficiency.

First, the full data $\boldsymbol{y}$ is split into m disjoint subsets, $(\boldsymbol{y}_1, ..., \boldsymbol{y}_m)$, according to the values of one of the input variables, denoted by the first one X_1 without loss of generality. More specifically, m is defined by $m = \lfloor M_1/\tau \rfloor$, where $M_1 = \max(X_1) - \min(X_1)$ is the range of the first variable and τ is a tuning parameter that is closely connected to the range parameter in tapering. The size of each subset $\boldsymbol{y}_a$ is then denoted by n_a and therefore $\sum_{a=1}^{m} n_a = n$. Theoretically, the procedure and asymptotic results developed in this paper are valid regardless of the choice of the first variable. Some suggestions and discussions regarding how

to choose the first variable in practice and the impacts are given in Section 4.

After splitting, the covariance matrix Σ can be decomposed into blocks indicating the within subset correlations and between subset correlations. A block-wise tapering/thresholding is applied to approximate the correlation matrix. That is, after rearranging the data by the first variable, the correlation matrix Σ is approximated by Σ_t as follows:

$$\Sigma_t = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & O & \cdots & O \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} & \cdots & O \\ & \ddots & \ddots & \ddots & \\ O & O & \cdots & \Sigma_{m(m-1)} & \Sigma_{mm} \end{pmatrix}_{n\times n}, \tag{2.10}$$

where Σ_{ii}, $i = 1, ...m$, captures the correlation within subset $\boldsymbol{y}_i$, Σ_{ij} is a block matrix capturing the correlations between subsets $\boldsymbol{y}_i$ and $\boldsymbol{y}_i$, and O's are matrices with all 0 elements. By replacing Σ by Σ_t in the log-likelihood function (2.3), we have the approximated log-likelihood function denoted by l_t.

Remark 1. *The correlation matrix Σ_t brings in sparsity by introducing zeros to the correlation matrix if the data are neither within the same subset nor in the neighborhood subsets. Such an assumption is relatively mild and desirable comparing with the existing tapering-type of methods because of three reasons. First, the assumption is applied to only one variable while the typical compactly supported correlation and tapering method require the sparsity assumption on all the variables as defined in (2.9). Second, Σ_t maintains the correlation between neighborhood subsets to be estimable while only partial information is estimable in the tapering-type of correlation function. For example, further assuming the off-diagonal matrices Σ_{aa+1} and Σ_{a+1a} to be lower and upper triangles leads to a tapering assumption on the first variable with range parameter τ, i.e., $R_1(|x_{i1} - x_{j1}|) = 0$, if $|x_{i1} - x_{j1}| > \tau$. Another example is the tapering approach discussed by Stein (2013) where none of the correlations between neighborhood subsets are estimable, i.e, the off-diagonal matrices Σ_{aa+1} and Σ_{a+1a} are assumed*

to be zeros. Third, it is computationally affordable for the proposed framework to have a larger τ compared with the tapering-type of methods and thus higher estimation accuracy and efficiency can be achieved. Note that, (2.10) is positive definite with high probability, but it is not guaranteed to be positive definite. In practice, the technique suggested by Cai and Zhou (2012) can be modified and applied here to ensure a positive semi-definite correlation matrix.

Next, we transform $\boldsymbol{y}$ to $\boldsymbol{y}^* = (\boldsymbol{y}_1^*, ..., \boldsymbol{y}_m^*)$ by sequentially updating each subsets as follows so that the correlation between subsets can be incorporated:

$$\boldsymbol{y}_a^* = \boldsymbol{y}_a - L_{a(a-1)}\boldsymbol{y}_{a-1}^*, \tag{2.11}$$

where $L_{(a+1)a} = \Sigma_{t(a+1)a}D_a^{-1}$, $D_a = \Sigma_{aa} - L_{a(a-1)}D_{(a-1)}L_{a(a-1)}^\top$, $L_{(a+1)a}$ and D_a's are solved iteratively by initialing $D_a = \Sigma_{11}$. The update of $\boldsymbol{y}_a^*$ only depends on the $\boldsymbol{y}_a$ and $\boldsymbol{y}_{a-1}^*$, which are all small subsets therefore it is easy to computer. This transformation is in fact a block LDL-decomposition (Fang, 2011) creates independency to the new subsets by carefully removing information sequentially. Based on the following Lemma, the transformed data $\boldsymbol{y}^*$ has an important property that the subsets $\boldsymbol{y}_a^*$ are mutually independent.

Lemma 1. *After transformation, the covariance within each subset $\boldsymbol{y}_a^*$ is D_a and any two subsets are mutually independent. That is, $\boldsymbol{y}^* = (\boldsymbol{y}_1^*, ..., \boldsymbol{y}_m^*)$ has covariance matrix D, where*

$$D = \begin{pmatrix} D_1 & ... & O \\ & \ddots & \\ O & ... & D_m \end{pmatrix}, \; L = \begin{pmatrix} I & & & \\ L_{21} & I & & \\ \vdots & \vdots & \ddots & \\ L_{m1} & L_{m2} & \cdots & I \end{pmatrix},$$

and $LDL^\top = \Sigma_t$.

For individual subsets $\boldsymbol{y}_a^*$, where $a = 1, ..., m$, the log-likelihood function of $\boldsymbol{y}_a^*$ can be written as

$$l_t^{(a)}(\boldsymbol{\beta}) = -\frac{1}{2}\log|D_a| - \frac{1}{2}(C_a\boldsymbol{\beta} - \boldsymbol{y}_a^*)^\top D_a^{-1}(C_a\boldsymbol{\beta} - \boldsymbol{y}_a^*),$$

where $n_a \times p$ matrix $C_a = X_a + \sum_{b=1}^{a-1} B_{ab} X_b$, $n_a \times n_b$ matrix $B_{ab} = \prod_{k=b+1}^{a}(-L_{k(k-1)})$, and $n_a \times p$ matrix X_a is the design matrix corresponding to $\boldsymbol{y}^{(a)}$. By maximizing $l_t^{(a)}(\boldsymbol{\beta})$, we have the MLE of $\boldsymbol{\beta}$ estimated from the ath subset as

$$\widehat{\boldsymbol{\beta}}_a = \arg\max_{\boldsymbol{\beta}} l_t^{(a)}(\boldsymbol{\beta}) = (C_a^\top D_a^{-1} C_a)^{-1} C_a^\top D_a^{-1} \boldsymbol{y}_a^*. \tag{2.12}$$

Under Gaussian process model (2.1) and by a direct calculation, we have $S_a^{-1/2}(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $S_a = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_a) = (C_a^\top D_a^{-1} C_a)^{-1}$. This could be easily verified by Lemma 1.

Step 2: Information combining via confidence distributions

Confidence distribution (CD) refers to any sample-dependent distribution function that can represent confidence intervals/regions of all levels for a parameter of interest (cf., e.g., Xie and Singh (2013)). Conceptually, a CD is not different from a point estimator or a confidence interval, but it uses a sample-dependent distribution function on the parameter space to estimate the parameter of interest. A CD is to "provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)" (Cox, 2013). It can provide meaningful answers for all questions related to statistical inferences and an approach that combines CDs preserves more information than a traditional approach that combines just point estimators (Xie and Singh, 2013; Schweder and Hjort, 2016). Singh et al. (2005) and Xie et al. (2011) described a general framework to combining information based on CDs, which can subsume almost all information combination methods used in the current practice. In this subsection and following Liu et al. (2015); Yang et al. (2014), we provide a combined estimation for the unknown regression parameters based on a set of CDs obtained from the individual subsets.

Specifically, from (2.12), a resulting CD for $\boldsymbol{\beta}$ in the ath subset, expressed in

its density form, is

$$h_n(\boldsymbol{\beta}) \propto \exp\left[-\frac{1}{2\sigma^2}(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta})^T S_a^{-1}(\widehat{\boldsymbol{\beta}}_a - \boldsymbol{\beta})\right]$$

That is, $\mathcal{N}(\widehat{\boldsymbol{\beta}}_a, S_a)$ is a multivariate normal CD for $\boldsymbol{\beta}$; cf., Singh et al. (2007) and Liu et al. (2015) for the formal definition of multivariate normal CD. Then, following Liu et al. (2015) and Section 4 of Singh et al. (2005), a combined point estimator of $\boldsymbol{\beta}$ can be obtained by

$$\widehat{\boldsymbol{\beta}}_c = \arg\max_{\boldsymbol{\beta}} \prod_{a=1}^{m} h_n(\boldsymbol{\beta}). \tag{2.13}$$

By a direct calculation, we have an explicit expression that $\widehat{\boldsymbol{\beta}}_c = (\sum W_a)^{-1}(\sum W_a \widehat{\boldsymbol{\beta}}_a)$, where $W_a = C_a^\top D_a^{-1} C_a$ is the weight matrices. Furthermore, the variance of $\widehat{\boldsymbol{\beta}}_c$ is $S_c = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_c) = (\sum W_a)^{-1}(\sum W_a S_a W_a)(\sum W_a)^{-1} = (\sum W_a)^{-1} = (X^\top \Sigma_t^{-1} X)^{-1}$ and $S_c^{-1/2}(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Again, by the definition of Singh et al. (2007) and Liu et al. (2015), $\mathcal{N}(\widehat{\boldsymbol{\beta}}_c, S_c)$ is a multivariate normal CD for $\boldsymbol{\beta}$. The function $\mathcal{N}(\widehat{\boldsymbol{\beta}}_c, S_c)$ is on the space of $\boldsymbol{\beta}$ and it depends data in all subsets. We call it a *combined CD*. The combined estimator $\widehat{\boldsymbol{\beta}}_c$ and the combined CD $\mathcal{N}(\widehat{\boldsymbol{\beta}}_c, S_c)$ enjoy computational efficiency because they can be calculated based on small subsets and by summations. Following Singh et al. (2005), statistical inference, such as constructing confidence intervals/regions of $\boldsymbol{\beta}$ or calculating p-values, can be easily obtained from the combined CD.

The following theorem states that $\widehat{\boldsymbol{\beta}}_c$ is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}$, the original MLE obtained based on (2.4) without splitting the data. A proof of the theorem can be found in Appendix.

Theorem 1. *Under the regularity conditions in Appendix B and $\tau > O_p(\log n)$, the combined estimator $\widehat{\boldsymbol{\beta}}_c$ is a consistent estimator of $\boldsymbol{\beta}$ and has the following asymptotic distribution:*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_c - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, S),$$

where $S = n\,Cov(\widehat{\boldsymbol{\beta}}_{mle}) = n(X^\top \Sigma^{-1} X)^{-1}$.

2.3.2 Estimation when both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown

We illustrate the unified SSCC framework in a general setting where both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown. In this case, the computation is more demanding compared with the estimation of $\boldsymbol{\beta}$ because MLE can be obtained only by maximizing the likelihood (2.6) without closed form expression and the maximization involves intensive operations of large correlation matrix. Therefore, a computationally efficient estimation procedure is even more critical. We extend the procedure of Section 3.1 to the situation where $\boldsymbol{\theta}$ is also unknown. The idea is to obtain the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ by updating $\boldsymbol{\beta}|\boldsymbol{\theta}$ and $\boldsymbol{\theta}|\boldsymbol{\beta}$ iteratively. Here we describe one of the iteration with details and the completed algorithm is given in the Appendix. This framework can be easily extended to estimate σ. The full expression is rather lengthy but straightforward. So to simplify the notation, we assume σ is known.

Start from an estimation of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^{(t-1)}$, $\boldsymbol{\beta}^{(t)}$ can be estimated by the combined estimator $\widehat{\boldsymbol{\beta}}_c$ given in (2.13) with $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t-1)}$. Given $\boldsymbol{\beta}^{(t)}$, a two-step procedure that is analogue to the one in Section 3.1 is implemented to obtain $\boldsymbol{\theta}^{(t)}$. In Step 1, based on the same splitting $(\boldsymbol{y}_1, ...\boldsymbol{y}_m)$, the sequential updating (2.11) is modified by

$$\boldsymbol{y}_a^*(\boldsymbol{\theta}) = \boldsymbol{y}_a - L_{a(a-1)}(\boldsymbol{\theta})\boldsymbol{y}_{a-1}^*(\boldsymbol{\theta}),$$

where $L_{a(a-1)}(\boldsymbol{\theta}) = \Sigma_{ta(a-1)}(\boldsymbol{\theta})D_{a-1}^{-1}(\boldsymbol{\theta})$ and $D_a(\boldsymbol{\theta}) = \Sigma_{taa}(\boldsymbol{\theta}) - L_{a(a-1)}(\boldsymbol{\theta})D_{a-1}(\boldsymbol{\theta})L_{a(a-1)}^{\top}(\boldsymbol{\theta})$. In step 2, the close form expression of MLE in (2.12) is replaced by maximizing the likelihood

$$l_t^{(a)}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(t)}) = -\frac{1}{2}\log|D_a(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{y}_a^*(\boldsymbol{\theta}) - C_a(\boldsymbol{\theta})\boldsymbol{\beta}^{(t)})^{\top}D_a^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_a^*(\boldsymbol{\theta}) - C_a(\boldsymbol{\theta})\boldsymbol{\beta}^{(t)}), \tag{2.14}$$

where $C_a(\boldsymbol{\theta}) = X_a + \sum_{b=1}^{a-1} B_{ab}(\boldsymbol{\theta})X_b$, $B_{ab}(\boldsymbol{\theta}) = \prod_{k=b+1}^{a}(-L_{k(k-1)}(\boldsymbol{\theta}))$, and X_a is the design matrix for $\boldsymbol{y}_a$. Note that, the calculation of log-likelihood $l_t^{(a)}$ depends only on the current subset $\boldsymbol{y}_a^*$, previous subset $\boldsymbol{y}_{a-1}^*$, and the correlation between these two subsets and thus it is still easy to compute. It is also clear that $l_t =$

$\sum_{a=1}^{m} l_t^{(a)}$.

The estimate of $\boldsymbol{\theta}$ from individual subset $\boldsymbol{y}_a^*$ is denoted by

$$\widehat{\boldsymbol{\theta}}_a^{(t)} = \arg\max_{\boldsymbol{\theta}} l_t^{(a)}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(t)})$$

and the combined estimate for $\boldsymbol{\theta}$ can be calculated by

$$\widehat{\boldsymbol{\theta}}_c^{(t)} = (\sum_{a=1}^{m} S_a^{-1})^{-1}(\sum_{a=1}^{m} S_a^{-1}\widehat{\boldsymbol{\theta}}_a^{(t)}), \tag{2.15}$$

where $S_a = -H_a^{-1}(\widehat{\boldsymbol{\theta}}_a^{(t)})$ and $H_a(\cdot)$ is the ath Hessian matrix derived from $l_t^{(a)}(\boldsymbol{\theta}|\boldsymbol{\beta}^{(t)})$. Therefore, given $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\theta}^{(t)}$ is updated by the combined estimator, i.e., $\boldsymbol{\theta}^{(t)} = \widehat{\boldsymbol{\theta}}_c^{(t)}$. Following Singh et al. (2007) and Liu et al. (2015) and similar to Section 2.3.1, an individual CD from the ath block is $\mathcal{N}(\widehat{\boldsymbol{\theta}}_a^{(t)}, S_a)$ and the combined CD is $\mathcal{N}(\widehat{\boldsymbol{\theta}}_c^{(t)}, S_c^{(t)})$, where $S_c^{(t)} = (\sum_{a=1}^{m} S_a^{-1})^{-1}$. When iteration converges (i.e. $||\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}||$ and $||\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}||$ are both very small), we stop the iteration and denote the combined results $(\boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)})$ as $(\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\theta}}_c)$.

This framework provides significant computational reduction in comparing with the original MLE in (2.4) and (2.6). The combined estimators also maintain desirable asymptotic properties as the original MLE, which is shown in the following theorem.

Theorem 2. *Under the assumptions of Theorem 1 and when $\tau > O_p(n^{1/2})$ as $n \to \infty$, the combined estimator $\widehat{\boldsymbol{\lambda}}_c = (\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\theta}}_c)$ is asymptotically as efficient as MLE $\widehat{\boldsymbol{\lambda}} = (\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\theta}})$ obtained from (2.4) and (2.6).*

Kaufman et al. (2008) points out that estimation based on tapering can be biased. In fact, this can be an issue in most of the tapering-type of methods including the compactly supported correlations and the current method. This is

because, for example when $\boldsymbol{\beta} = 0$, we have

$$\begin{aligned} E\{\frac{\partial\{-l_t(\boldsymbol{\theta})\}}{\partial\boldsymbol{\theta}}\} &= E\{\frac{1}{2}tr(\Sigma_t^{-1}\Sigma_t') + \frac{1}{2}\boldsymbol{y}'\Sigma_t^1\boldsymbol{y}\} \\ &= \frac{1}{2}tr(\Sigma_t^{-1}\Sigma_t') - tr(\Sigma_t^{-1}\Sigma_t'\Sigma_t^{-1}\Sigma) \\ &= \frac{1}{2}tr(\Sigma_t^1(\Sigma - \Sigma_t)) \neq 0, \end{aligned}$$

where $l_t(\cdot)$ denotes the log-likelihood function by compactly supported correlation, $\partial\Sigma_t/\partial\boldsymbol{\theta} = \Sigma_t'$, $\partial\Sigma_t^{-1}/\partial\boldsymbol{\theta} = \Sigma_t^1 = -\Sigma_t^{-1}\Sigma_t'\Sigma_t^{-1}$. This issue can be solved under certain conditions as described in the following corollary.

Corollary 1. *Suppose that $\tau > O_p(\log n)$ as $n \to \infty$. Under the assumptions of Theorem 1, the combined estimator is unbiased.*

2.3.3 Prediction and uncertainty quantification

A CD-based predictive distribution is introduced in this section. It has two advantages. First, it consists a GP predictor that overcomes the computational difficulty in the conventional predictor (2.7) and meanwhile maintains the same asymptotic efficiency. Second, it provides comprehensive information for statistical inference and a better quantification of prediction uncertainty comparing with the plug-in approach.

Based on the sequential split-conquer procedure and the combined estimates obtained from Section 3.2, we approximate the GP predictive mean (2.7) and variance (2.8) by $p_1(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $m_1(\boldsymbol{\beta}, \boldsymbol{\theta})$ as follows:

$$p_1(\boldsymbol{\beta}, \boldsymbol{\theta}) = \boldsymbol{x}_0^\top\boldsymbol{\beta} + \sum_{a=1}^{m}\gamma_a^*(\boldsymbol{\theta})^\top D_a^{-1}(\boldsymbol{\theta})\boldsymbol{y}_a^* + \sum_{a=1}^{m}\gamma_a^*(\boldsymbol{\theta})^\top D_a^{-1}(\boldsymbol{\theta})C_a(\boldsymbol{\theta})\boldsymbol{\beta}, \quad (2.16)$$

$$m_1(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sigma^2(1 - \sum_{a=1}^{m}\gamma_a^*(\boldsymbol{\theta})^\top D_a^{-1}(\boldsymbol{\theta})\gamma_a^*(\boldsymbol{\theta})), \quad (2.17)$$

where $\gamma_a^*(\boldsymbol{\theta}) = \gamma_a(\boldsymbol{\theta}) + L_{a(a-1)}(\boldsymbol{\theta})\gamma_{a-1}^*(\boldsymbol{\theta})$, $\gamma_a(\boldsymbol{\theta})$ is the $n_a \times 1$ vector with ith element equal to $\phi(||\boldsymbol{x}_i - \boldsymbol{x}_0||; \boldsymbol{\theta})$ where $i = \sum_{b=1}^{a-1} n_b + 1, ..., \sum_{b=1}^{a} n_b$. These two

estimates enjoy the computational efficiency because their calculation involves only small correlation matrix with size $n_a \times n_a$, $a = 1, ...m$. Given the computational reduction, the new predictive mean (2.16) and variance (2.17) is shown to be asymptotically equivalent to the conventional ones according to the following theorem.

Theorem 3. *Suppose that $\tau > O_p(n^{1/2})$ as $n \to \infty$. Under the assumptions of Theorem 1, we have*

$$p_1(\boldsymbol{\beta}, \boldsymbol{\theta}) \to p_0(\boldsymbol{\beta}, \boldsymbol{\theta}) and \quad m_1(\boldsymbol{\beta}, \boldsymbol{\theta}) \to m_0(\boldsymbol{\beta}, \boldsymbol{\theta}). \tag{2.18}$$

To provide a better alternative to the plug-in predictive distribution, we construct a CD-based predictive distribution that captures the parameter uncertainty using the combined confidence distributions. The CD-based predictive distribution is first introduced by Shen et al. (2016) under a general setting. Here we extend the idea to GP models and construct a CD-based predictive distribution which is not only more accurate but also easy to compute.

A CD-based predictive distribution is defined by:

$$Q(y_0; \boldsymbol{y}) = \int_{\boldsymbol{\lambda} \in \Theta} G_{\boldsymbol{\lambda}}(y_0) dF_c(\boldsymbol{\lambda}; \boldsymbol{y}), \tag{2.19}$$

where $G_{\boldsymbol{\lambda}}(y_0)$ is the cumulative density function (CDF) of the predictive distribution, i.e., a normal distribution with mean p_1 and variance m_1 given in (2.16) and (2.17). The confidence distribution (CD) of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ is $\mathcal{N}(\widehat{\boldsymbol{\lambda}}_c, S_c^{\lambda})$ where $\boldsymbol{\lambda} = (\boldsymbol{\beta}, \boldsymbol{\theta})$, $\widehat{\boldsymbol{\lambda}}_c = (\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\theta}}_c)$ is the combined estimator of $\boldsymbol{\lambda}$, variance matrix $S_c^{\lambda} = \mathrm{Var}(\widehat{\boldsymbol{\lambda}}_c)$ equals to the corresponding Hessian matrix calculated from log-likelihood function. CD-based predictive distribution is closely related to the Bayesian predictive distribution and bootstrap predictive distribution as discussed by Singh et al. (2005). By direct application of Theorem 4 in Shen et al. (2016), it can be shown that the CD-based predictive distribution outperforms the plug-in approach, measured by the average Kullback-Leibler distance to the true predictive distribution.

To implement the predictive distribution formulated in (2.19), we proposed the following Monte-Carlo algorithm which is simple yet broadly applicable.

Monte-Carlo Algorithm: Obtain T simulated copy of y_0 from $Q(\cdot; \boldsymbol{y})$, denoted by $y_0^{(t)}$ and $t = 1, ...T$, by iteratively perform the following two steps.

1. Simulate a random variable $\boldsymbol{\lambda}^{(t)}|\boldsymbol{y} \sim \mathcal{N}(\widehat{\boldsymbol{\lambda}}_c, S_c^{\lambda})$.

2. Obtain $y_0^{(t)}|\boldsymbol{\lambda}^{(t)} \sim \mathcal{N}(p_1(\boldsymbol{\lambda}^{(t)}), m_1(\boldsymbol{\lambda}^{(t)}))$.

These T copies of y_0 can be used to approximate the predictive distribution in (2.19).

2.4 Simulation

Simulation studies are conducted to examine the performance, including estimation and prediction, of the proposed framework. All simulations are carried out by a machine with a quad-core CPU @ 3.50GHz, 12GB RAM under R 3.3.1 in Windows 10.

To demonstrate the estimation performance, we compare the proposed combined estimator with the regular MLE and the estimator based on the compactly supported correlation (Kaufman et al., 2011), denoted by "Compact". We consider a problem in a four-dimensional input space, $\boldsymbol{x} \in [0, 1]^4$, with three different sample sizes, n = 1000,1500, and 2000, in which $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are unknown. The inputs are generated from a regular grid on $[0, 1]^4$ and the responses are simulated from a Gaussian process model with the mean function coefficient $\boldsymbol{\beta} = (2, 3, 1, 2, 1.5)$ and the correlation function

$$\phi(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\theta}) = \prod_{k=1}^{4} \exp(-\theta_k |x_{ik} - x_{jk}|), \tag{2.20}$$

where $\boldsymbol{\theta} = (15, 1.5, 2, 3)$ and $\sigma^2 = 1$ is assumed to be known. To implement the SSCC framework, we assume $\tau = 0.2$, so the number of blocks is $m = \lfloor M_1/\tau \rfloor =$

$\lfloor 1/0.2 \rfloor = 5$. To emphasize the estimation performance of the parameters, we specify $\alpha = 1$ for the three methods without further tuning. To further increase the flexibility of the GP models, standard tuning methods can be applied. For each sample size, we repeat the simulation 100 times and report the mean, standard deviation, and the computing times, denoted by CT, in Table 2.1. For the cases of $n = 1000, 1500, 2000$, the performance of the three methods are also illustrated by box plots in Figure 2.1, 2.2, 2.3.

	$n = 1000$			$n = 1500$			$n = 2000$		
	MLE	Compact	SSCC	MLE	Compact	SSCC	MLE	Compact	SSCC
$\widehat{\beta}_0$	1.96(0.33)	1.96(0.33)	1.96(0.33)	2.00(0.30)	2.00(0.30)	2.00(0.30)	2.08(0.23)	2.08(0.23)	2.08(0.23)
$\widehat{\beta}_1$	3.03(0.43)	3.02(0.43)	3.02(0.43)	3.01(0.36)	3.01(0.36)	3.01(0.36)	2.90(0.28)	2.90(0.28)	2.90(0.28)
$\widehat{\beta}_2$	1.00(0.23)	1.00(0.23)	1.00(0.23)	0.99(0.21)	0.99(0.21)	0.99(0.21)	1.04(0.20)	1.04(0.20)	1.04(0.20)
$\widehat{\beta}_3$	2.04(0.24)	2.04(0.24)	2.04(0.24)	1.97(0.25)	1.97(0.25)	1.97(0.25)	1.98(0.24)	1.98(0.24)	1.98(0.24)
$\widehat{\beta}_4$	1.53(0.25)	1.53(0.25)	1.53(0.25)	1.52(0.28)	1.52(0.28)	1.52(0.28)	1.49(0.18)	1.49(0.18)	1.49(0.18)
$\widehat{\theta}_1$	14.72(0.42)	14.72(0.42)	14.80(0.43)	14.65(0.39)	14.65(0.39)	14.65(0.49)	14.69(0.45)	14.74(0.45)	14.79(0.46)
$\widehat{\theta}_2$	1.49(0.06)	1.49(0.06)	1.51(0.06)	1.49(0.06)	1.49(0.06)	1.50(0.10)	1.49(0.06)	1.49(0.06)	1.49(0.10)
$\widehat{\theta}_3$	2.00(0.06)	2.00(0.06)	2.01(0.06)	1.98(0.06)	1.98(0.06)	2.00(0.06)	1.99(0.07)	1.99(0.07)	2.00(0.10)
$\widehat{\theta}_4$	3.01(0.06)	3.01(0.06)	2.99(0.06)	3.00(0.06)	3.00(0.06)	3.00(0.08)	3.01(0.06)	3.01(0.06)	3.00(0.07)
CT	32.46(1.29)	30.61(1.34)	4.54(0.11)	99.66(3.83)	95.90(5.44)	10.39(0.53)	227.63(6.96)	222.18(9.33)	20.32(0.95)

Table 2.1: Mean, standard deviation and computing time of estimations by MLE, compact and SSCC methods with simulation studies' sample size $n = 1000, 1500, 2000$.

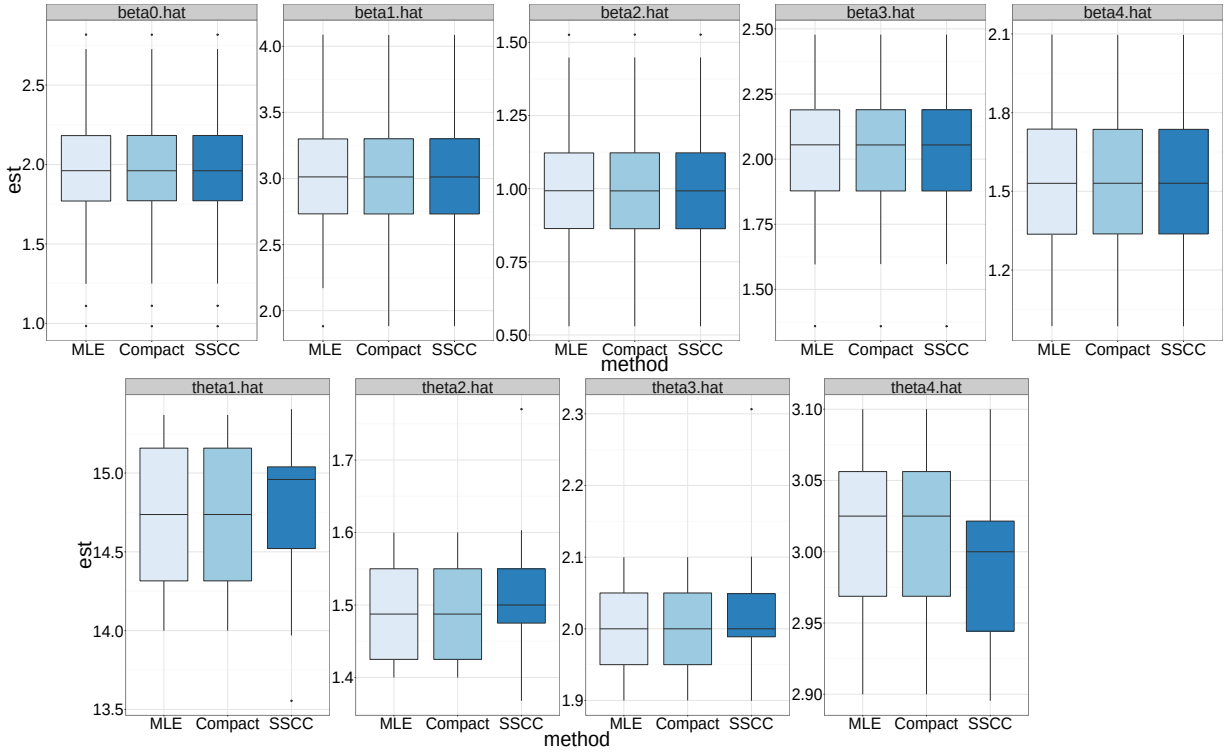

Figure 2.1: Boxplots of estimators by MLE, compact and SSCC methods when sample size $n = 1000$

Based on the results in Table 2.1, the estimation performance of the proposed estimator is compatible with the other two estimators which is consistent with the theoretical results. In terms of the computing time, the proposed method provides a significant reduction comparing with the other two methods, especially when the sample size is large. Specifically, comparing with the original MLE and the compactly supported correlation approach, the computing time is reduced more than 86% by the proposed combined estimator for all the three different sample sizes and this reduction increases with sample sizes.

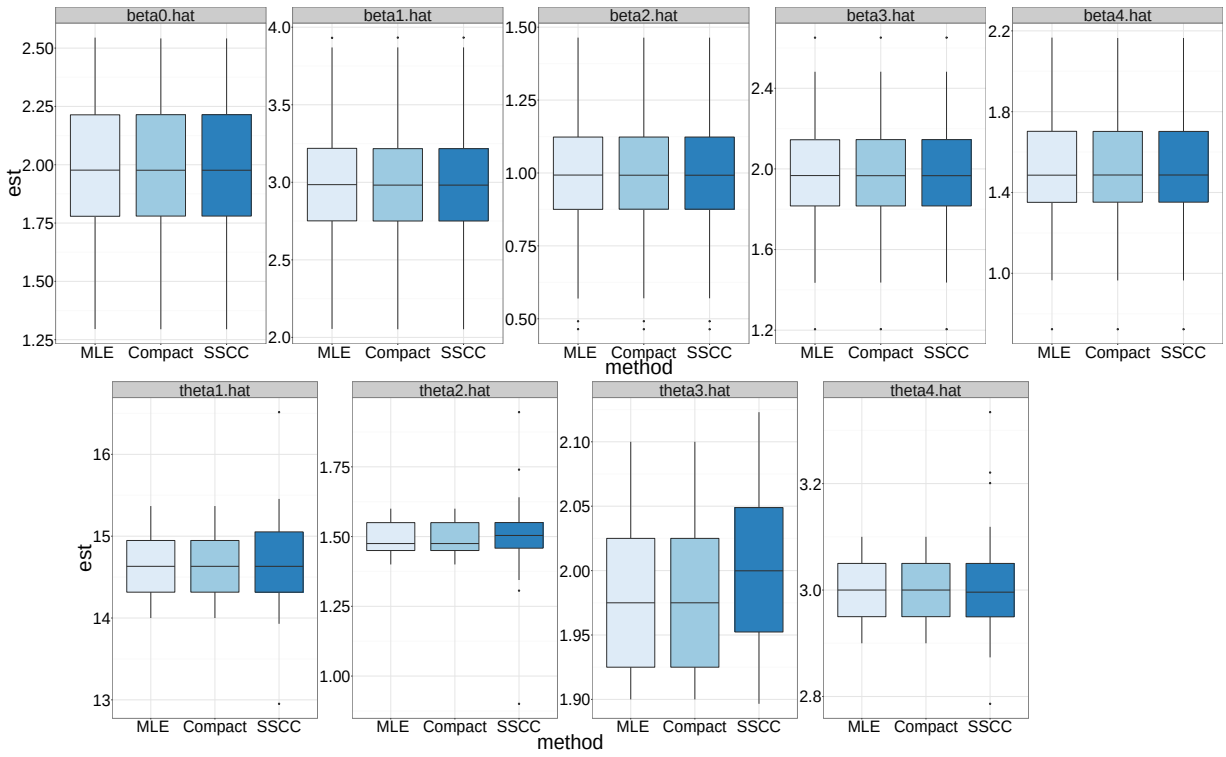

Figure 2.2: Boxplots of estimators by MLE, compact and SSCC methods when sample size $n = 1500$

To illustrate the performance of the proposed predictive distribution, we implement the *Monte-Carlo Algorithm* (Section 3.3) to construct predictive distributions for several untried points following the previous settings with sample size $n = 2000$. We focus on examining the predictive performance by changing the setting of the most important variable, because this is of interest in many applications including the real data analysis in Section 5. Four untried settings are assumed by varying the settings of the most active variables, i.e., changing the setting of the first variable to be 0.2, 0.4, 0.6, and 0.8 respectively. The setting of the other three variables are fixed to be $(0.43, 0.5, 1)$. Based on the estimator in Table 2.1, we have $\widehat{\boldsymbol{\lambda}}_c = (2.08, 2.90, 1.04, 1.98, 1.49, 14.79, 1.49, 2.00, 3.00)$. Here, we construct the CD-based confidence distribution (2.19) according to 1000 copies of y_0 generated by the Monte-Carlo algorithm, denoted by $y_0^{(1)}, \cdots, y_0^{(1000)}$. Figure 2.4 shows the corresponding histograms of y_0 for the four untried settings. The

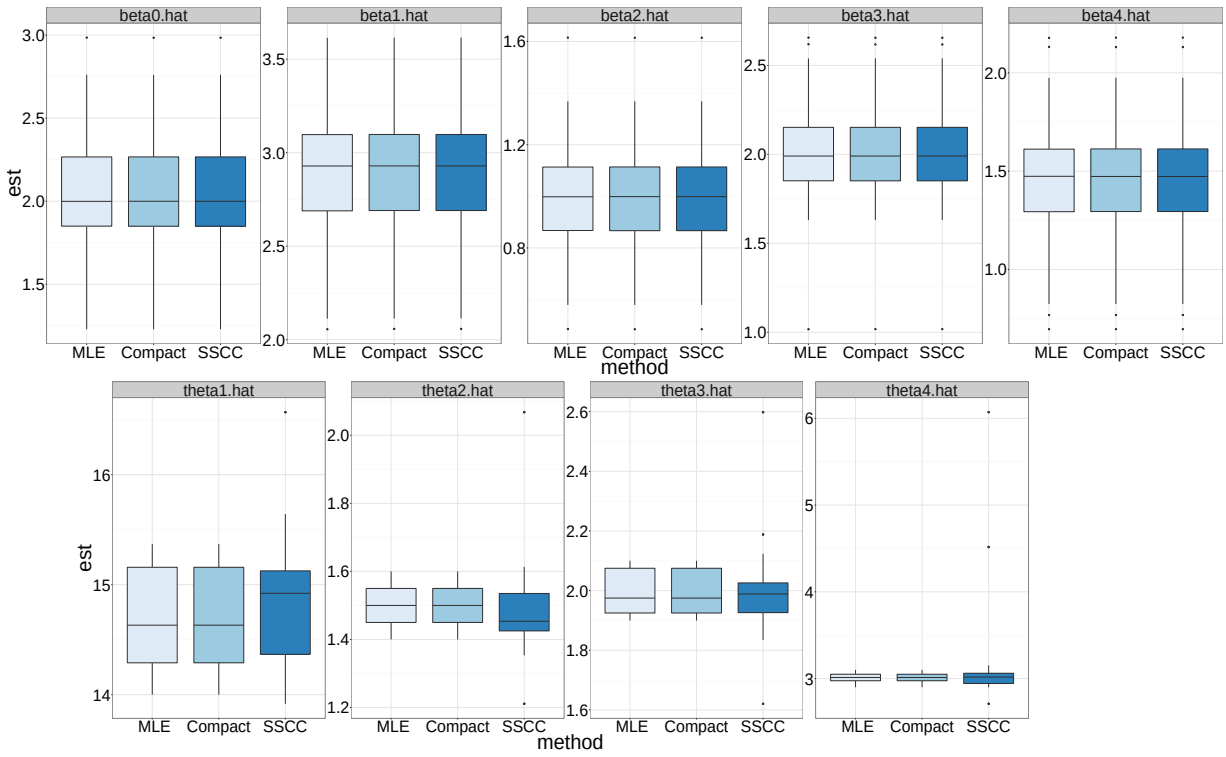

Figure 2.3: Boxplots of estimators by MLE, compact and SSCC methods when sample size $n = 2000$

red dashed lines are the mean function calculated by the true parameters. For example, the first one is calculated by $2 + 3 \times 0.2 + 1 \times 0.43 + 2 \times 0.5 + 1.5 \times 1 = 5.53$. The CD-based predictive distribution not only contain information of the predictive mean but also provides a flexible way to construct predictive intervals with any level of frequentist coverage probability. Furthermore, it provides rich information for statistical inference in practice which is further illustrated by the real example in Section 5.

Note that an efficient setting of the initial mean function coefficients, $\boldsymbol{\theta}^{(0)}$, can significantly reduce the computing time. We suggest to use the coefficients estimated by linear regression models as the initials because they tend to be close to those obtained by GP models. Another important question is how to determine the first variable in practice to implement the splitting procedure. Although any variable can be used theoretically, there are some choices that we found

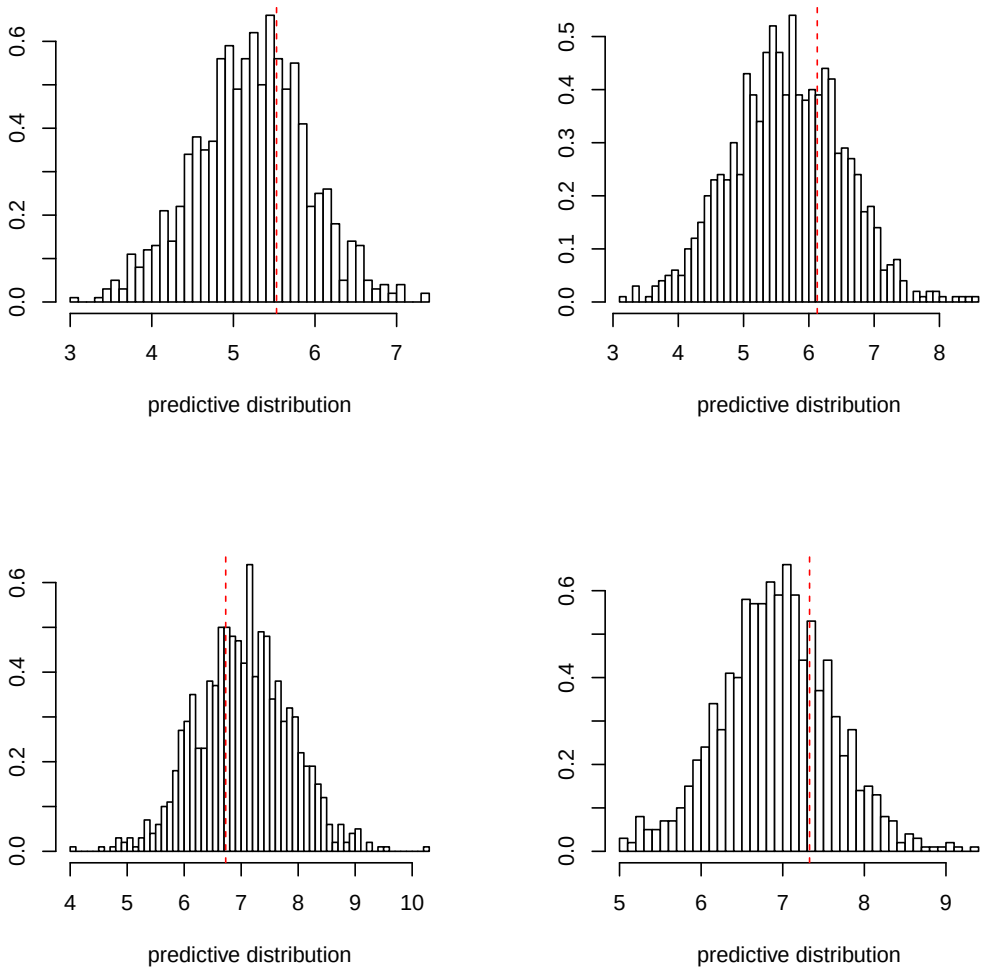

Figure 2.4: Predictive distributions for 4 untried points

empirically more efficient. Ideally, a variable that follows the tapering assumption is a desirable choice. That is, for this particular variable, the correlations between pairs of responses with larger distance are nearly zero, therefore little information is lost in assuming them to be conditionally independent given other variables as described by (2.10). Therefore, a useful way in practice is to empirically examine this assumption by checking the correlation plots for each variable. In our experience, the most active variable often be a reasonable choice, such as the first variable in the simulation study.

2.5 Data Center Thermal Management

A data center is a computing infrastructure facility that houses large amounts of information technology 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. 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, computer experiments 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 Lopez and Hamann (2011). There are 26820 temperature outputs generated from the CFD simulator based on an irregular grid over an 9-dimensional space. The nine variables are listed in Table 2.2. The first six variables control the cooling mechanism, including four computer room air conditioning (CRAC) units with different flow rates $(x_1, ..., x_4)$, the overall room temperature setting (x_5), and the perforated floor tiles with different percentage of open areas (x_6). The last three variables are the spatial location, x-axis, y-axis, and height, in the data center (x_7 to x_9).

Gaussian process model is desirable for the analysis of this problem because it provides a flexible interpolator for the deterministic CFD simulation outputs (Santner et al., 2003). However, it is computationally prohibitive to build a GP model following the standard procedure. Thus, we implemented the proposed

method to this data and for the purpose of comparison, we also performed the original MLE and the compact correlation function in two smaller subsets, $n = 1800$ and $n = 3600$. Estimation results are summarized in Table 2.2, where "-" indicates no result available. For $n = 1800$, we are able to calculate the estimators for the three approaches. The results show that, with a similar estimation performance, the proposed combined estimator reduces the computing time by more than 98% comparing with the other two methods. For $n = 3600$ and the full data, $n = 26860$, the original MLE and the compactly supported correlation approach cannot be carried out due to computational and/or memory limitation.

		$n = 1800$			$n = 3600$			$n = 26820$		
Variable		MLE	Compact	SSCC	MLE	Compact	SSCC	MLE	Compact	SSCC
x_1	$\widehat{\beta}_1$	-8.28(0.10)	-8.29(0.10)	-8.29(0.10)	-	-	-8.05(0.09)	-	-	-7.39(0.08)
	$\widehat{\theta}_1$	0.84(0.01)	0.84(0.01)	0.86(0.01)	-	-	0.85(0.01)	-	-	0.86(0.01)
x_2	$\widehat{\beta}_2$	-9.00(0.10)	-8.99(0.10)	-8.99(0.10)	-	-	-10.14(0.09)	-	-	-9.09(0.08)
	$\widehat{\theta}_2$	0.76(0.01)	0.76(0.01)	0.79(0.01)	-	-	0.77(0.01)	-	-	0.76(0.01)
x_3	$\widehat{\beta}_3$	-6.44(0.10)	-6.45(0.10)	-6.45(0.10)	-	-	-7.08(0.09)	-	-	-6.59(0.09)
	$\widehat{\theta}_3$	1.20(0.01)	1.20(0.01)	1.19(0.01)	-	-	1.14(0.01)	-	-	1.13(0.01)
x_4	$\widehat{\beta}_4$	-5.42(0.11)	-5.41(0.11)	-5.41(0.11)	-	-	-6.52(0.10)	-	-	-5.86(0.10)
	$\widehat{\theta}_4$	1.90(0.01)	1.90(0.01)	1.80(0.01)	-	-	1.70(0.01)	-	-	1.83(0.01)
x_5	$\widehat{\beta}_5$	-0.08(0.13)	-0.07(0.13)	-0.07(0.13)	-	-	-0.68(0.13)	-	-	0.29(0.13)
	$\widehat{\theta}_5$	3.50(0.01)	3.50(0.01)	3.40(0.01)	-	-	3.39(0.01)	-	-	3.50(0.01)
x_6	$\widehat{\beta}_6$	-1.98(0.10)	-1.97(0.10)	-1.97(0.10)	-	-	-2.12(0.10)	-	-	-1.79(0.09)
	$\widehat{\theta}_6$	1.29(0.01)	1.29(0.01)	1.20(0.01)	-	-	1.24(0.01)	-	-	1.28(0.01)
x_7	$\widehat{\beta}_7$	-3.39(0.06)	-3.41(0.06)	-3.41(0.06)	-	-	-4.04(0.04)	-	-	-2.99(0.03)
	$\widehat{\theta}_7$	0.20(0.01)	0.20(0.01)	0.17(0.01)	-	-	0.14(0.01)	-	-	0.15(0.01)
x_8	$\widehat{\beta}_8$	2.80(0.08)	2.80(0.08)	2.80(0.08)	-	-	1.72(0.07)	-	-	0.04(0.06)
	$\widehat{\theta}_8$	0.60(0.01)	0.60(0.01)	0.50(0.01)	-	-	0.62(0.01)	-	-	0.50(0.01)
x_9	$\widehat{\beta}_9$	22.33(0.18)	22.35(0.18)	22.35(0.18)	-	-	24.75(0.18)	-	-	23.90(0.18)
	$\widehat{\theta}_9$	21.90(0.03)	21.90(0.03)	21.45(0.03)	-	-	21.61(0.03)	-	-	21.10(0.03)
CT (seconds)		2768.70	2753.91	55.07	-	-	372.67	-	-	26257.2

Table 2.2: Mean, standard deviation and computing time of estimations by MLE, compact and SSCC methods with subsample size $n = 1800, 3600$ and the entire CFD data

Based on the estimation results in Table 2.2, we can construct the CD-based predictive distribution for some untried settings, which is a crucial step in finding an efficient cooling mechanism in a data center. The prediction performance is

first illustrated by predicting the heat map in the data center by varying the most active variable, height, with the control variables assumed to be: CRAC unit flow rate 6500, unit 2 flow rate 6500, unit 3 flow rate 2750, unit 4 flow rate 2750, room temperature 71° F and tile percentage 75%. Figure 2.5 presents the CD-based predictive heat map at four different heights, i.e., 0, 2.25, 4.25, 6.75. From the heat maps, it is shown that on average, temperature increases with height which agrees with the thermal dynamics in general. Apart from predictive heat map, the CD-based predictive distribution can be used to construct confidence intervals with any level of frequentist coverage probability. It also provides valuable insights of the thermal distribution in the data center. For example, Figure 2.6 shows the predictive distributions for four randomly selected untried settings at location x-axis = 23.5, y-axis = 14.5, with four different heights. At height = 2.25 given other settings, the confidence that the temperature will below 66° F is 99.4%; at height = 6.75, the confidence that the temperatures fall into the interval $(74, 77)$ is 84.2%.

We further compare the CD-based predictive distribution with the plug-in approach in the case when MLE is available, i.e., $n = 1800$. In Figure 2.7, the empirical CD-based predictive density for the first untried setting in Figure 2.5 is given as the black curve and the corresponding plug-in predictive density is given as the red dotted curve. It appears that the plug-in approach slightly underestimate the predictive uncertainty and this underestimation is expected to be larger when the sample size gets smaller. So the empirical result shows that, apart from computational reduction, the CD-based predictive distribution provide a better quantification of predictive uncertainty comparing with the traditional plug-in approach.

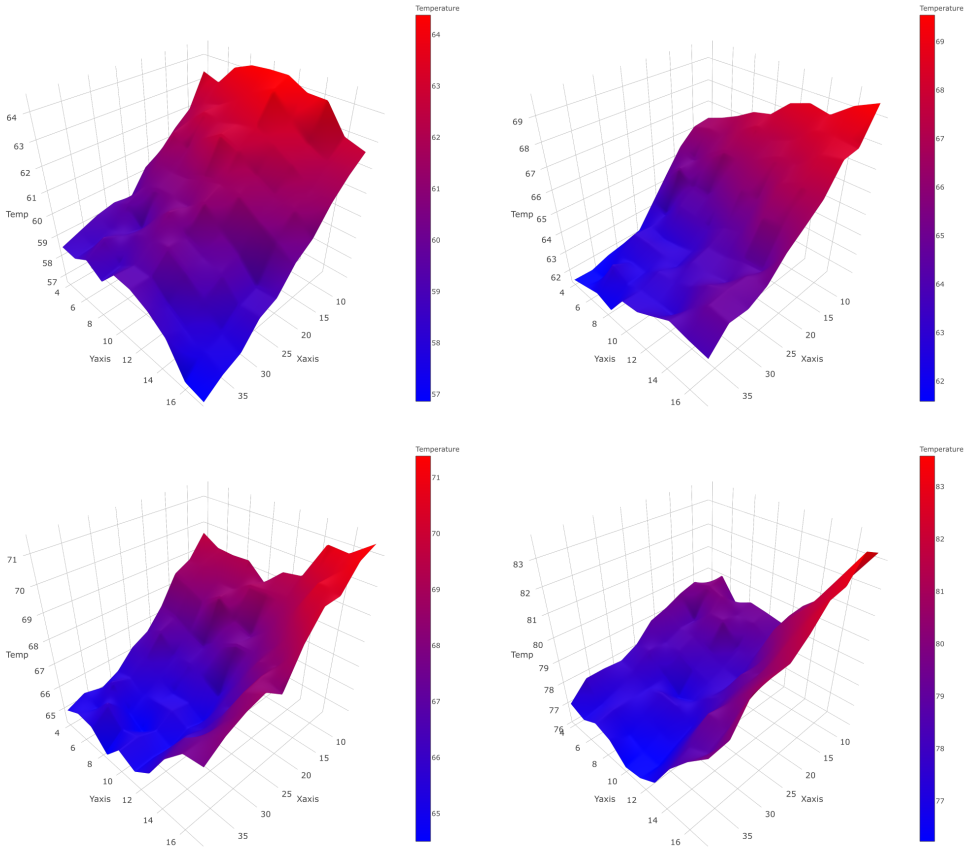

Figure 2.5: Heatmaps: mean of predictive distribution at height 0, 2.25, 4.25 and 6.75 feet

2.6 Summary and Concluding Remarks

In this chapter, we propose a unified sequential split-conquer-combine framework, called SSCC, to tackle two open problems in Gaussian process modeling, the computational difficulty and the underestimation of prediction uncertainty. The proposed method relies on two schemes: tapering (with large tuning parameter τ) and sequential updating. Note that, in most spatial and temporal data, the dependence between data points gets weaker when they are far away. This observation allows us to use the tapering technique with relatively large τ to ignore weak dependence among far away points. However, by tempering alone is not enough since its computation can be still quite complex. We also need to reduce the computing time and account for the strong dependence between

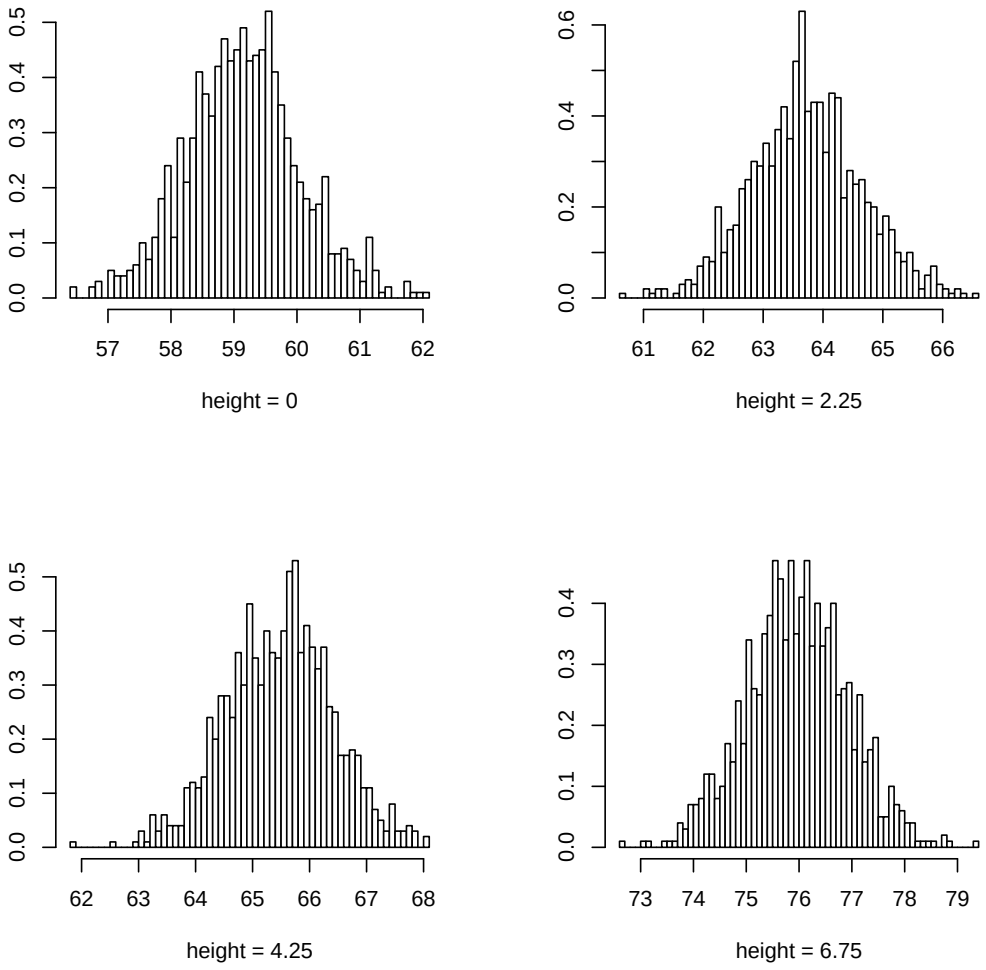

Figure 2.6: Predictive distribution for 4 untried points at different levels of height

neighboring subsets. This is achieved by sequentially updating the subset block using the previous and current blocks and performing analysis on the updated subset one at a time. In a nutshell, we sequentially analyze neighboring blocks and combined all the K estimates obtained to get a combine estimate, say $\widehat{\boldsymbol{\lambda}}_{sscc}$. If we can show $\widehat{\boldsymbol{\lambda}}_{sscc} = \widehat{\boldsymbol{\lambda}}_{taper}$, then, by the fact that $\widehat{\boldsymbol{\lambda}}_{taper} \approx \widehat{\boldsymbol{\lambda}}_{mle}$, we have $\widehat{\boldsymbol{\lambda}}_{sscc} \approx \widehat{\boldsymbol{\lambda}}_{mle}$.

This framework is based on the idea of confidence distribution (CD) and consists of a sequential split-conquer procedure, information combining technique

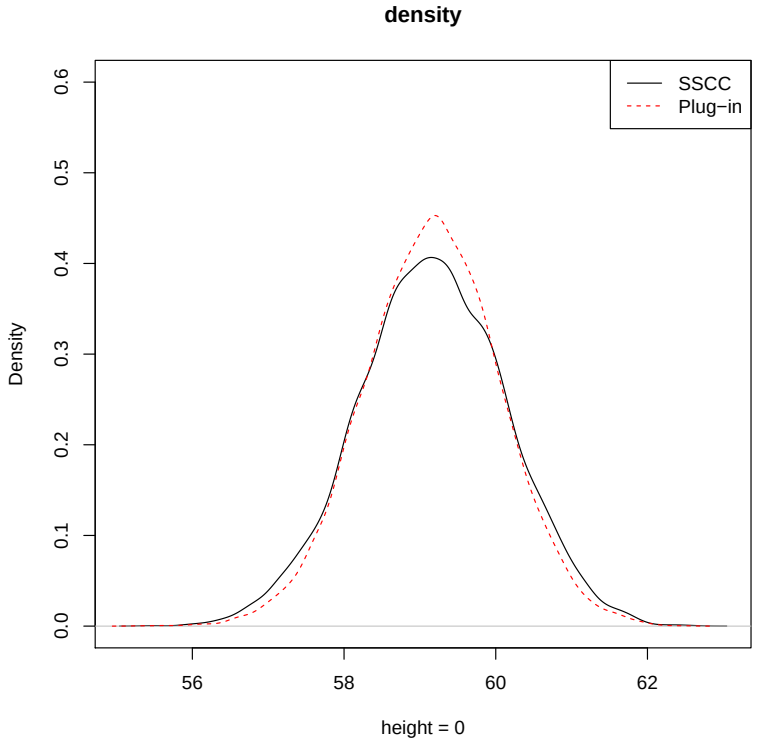

Figure 2.7: Comparison of CD predictive distribution and plug-in predictive distribution when $n = 1800$

using CDs, and a CD-based predictive distribution. A computationally efficient estimation and prediction procedure is introduced. Under mild conditions, the new estimators and predictors are shown to be asymptotically equivalent to the conventional ones using full data, while the computing time is significantly reduced. A Monte-Carlo algorithm is introduced to construct the CD-based predictive distribution which provides rich information for statistical inference and a better quantification of prediction uncertainty comparing with the plug-in approach. The advantages of the proposed method are clearly demonstrated by simulations as well as a data center thermal management problem.

2.7 Appendix A: Proofs

2.7.1 A.1 Proof of Lemma 1

Proof: we first show that $\Sigma_t = LDL^\top$,

$$LDL^\top = \begin{pmatrix} I & & & \\ L_{21} & I & & \\ \vdots & \ddots & \ddots & \\ O & \cdots & L_{m(m-1)} & I \end{pmatrix} \begin{pmatrix} D_1 & \dots & O \\ & \ddots & \\ O & \dots & D_m \end{pmatrix} \begin{pmatrix} I & L_{21}^\top & \cdots & O \\ & I & \ddots & \vdots \\ & & \ddots & L_{m(m-1)}^\top \\ & & & I \end{pmatrix}$$

$$= \begin{pmatrix} D_1 & & & \\ L_{21}D_1 & D_2 & & \\ \vdots & \ddots & \ddots & \\ O & \cdots & L_{m(m-1)}D_{m(m-1)} & D_m \end{pmatrix} \begin{pmatrix} I & L_{21}^\top & \cdots & O \\ & I & \ddots & \vdots \\ & & \ddots & L_{m(m-1)}^\top \\ & & & I \end{pmatrix}$$

$$= \begin{pmatrix} D_1 & D_1L_{21}^\top & \cdots & O \\ L_{21}D_1 & D_2 + L_21D_1L_{21}^\top & \cdots & O \\ \vdots & \ddots & \ddots & \vdots \\ O & \cdots & L_{m(m-1)}D_{m(m-1)} & D_m + L_{m(m-1)}D_{m(m-1)}L_{m(m-1)}^\top \end{pmatrix}.$$

Let $B = L^{-1}$, it is clear that the diagonal block matrices of B are all identity and the lower off diagonal block matrices can be written as

$$B_{ab} = \prod_{k=b+1}^{a} -L_{k(k-1)},$$

where B_{ab} is the abth block matrix. By expanding the updating formula, we have

$$\begin{aligned} \boldsymbol{y}_a^* &= \boldsymbol{y}_a - L_{a(a-1)}\boldsymbol{y}_{a-1}^* = \boldsymbol{y}_a - L_{a(a-1)}(\boldsymbol{y}_{a-1} - L_{(a-1)(a-2)}\boldsymbol{y}_{a-2}^*) \\ &= \cdots = \boldsymbol{y}_a + B_{a(a-1)}\boldsymbol{y}_{a-1} + \cdots + B_{a1}\boldsymbol{y}_1 \end{aligned}$$

Therefore, $B\boldsymbol{y} = \boldsymbol{y}^*$ and $\mathrm{Cov}(\boldsymbol{y}^*) = \mathrm{Cov}(B\boldsymbol{y}) = BL\Sigma_t L^\top B^\top = D$. □

2.7.2 A.2 Proof of Theorem 1

The proof consists of two lemmas that culminate in the final proof.

Lemma (A1)**.** *Under the regularity conditions in Appendix B and $\tau > O(\log n)$, the combined estimator $\widehat{\boldsymbol{\beta}}_c$ is equivalent to $\widehat{\boldsymbol{\beta}}_t = (X^\top \Sigma_t^{-1} X)^{-1} X^\top \Sigma_t^{-1} \boldsymbol{y}$.*

Proof of Lemma (A1): From Lemma 1, we have $\Sigma_t = LDL^\top$. So Σ_t^{-1} can be written as $(L^\top)^{-1} D^{-1} L^{-1} = B^\top D^{-1} B$ and,

$$
\begin{aligned}
&X^\top \Sigma_t^{-1} X = X^\top B^\top D^{-1} B X \\
&= \begin{pmatrix} X_1^\top \\ \vdots \\ X_{m-1}^\top \\ X_m^\top \end{pmatrix}^\top \begin{pmatrix} I & B_{21}^\top & \cdots & B_{m1}^\top \\ & \ddots & \ddots & \vdots \\ & & I & B_{m(m-1)}^\top \\ & & & I \end{pmatrix} \begin{pmatrix} D_1^{-1} & & & \\ & D_2^{-1} & & \\ & & \ddots & \\ & & & D_m^{-1} \end{pmatrix} \\
&\begin{pmatrix} I & & & \\ B_{21} & I & & \\ \vdots & \ddots & \ddots & \\ B_{m1} & \dots & B_{m(m-1)} & I \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_{m-1} \\ X_m \end{pmatrix} \\
&= \begin{pmatrix} X_1^\top & \cdots & X_m^\top + \sum_{j=1}^{m-1} X_j^\top B_{mj}^\top \end{pmatrix} \begin{pmatrix} D_1^{-1} & & \\ & \ddots & \\ & & D_m^{-1} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_m + \sum_{j=1}^{m-1} B_{mj} X_j \end{pmatrix} \\
&= \sum_{i=1}^{m} C_i^\top D_i^{-1} C_i.
\end{aligned}
$$

Similarly, we have $X^\top \Sigma_t^{-1} \boldsymbol{y} = \sum_{i=1}^m C_i^\top D_i^{-1} \boldsymbol{y}_{new}^{(i)}$. Therefore,

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_t &= (X^\top \Sigma_t^{-1} X)^{-1} X^\top \Sigma_t^{-1} \boldsymbol{y} = (\sum_{a=1}^{m} C_a^\top D_a^{-1} C_a)^{-1} \sum_{a=1}^{m} C_a^\top D_a^{-1} \boldsymbol{y}_a^* \\
&= (\sum_{a=1}^{m} C_a^\top D_a^{-1} C_a)^{-1} \sum_{a=1}^{m} (C_a^\top D_a^{-1} C_i)^{-1} \widehat{\boldsymbol{\beta}}_a = \widehat{\boldsymbol{\beta}}_c
\end{aligned}
$$

□

Lemma (A2)**.** *Under the regularity conditions in Appendix B, $\widehat{\boldsymbol{\beta}}_t$ is consistent and asymptotically normal distributed:*

$$
\sqrt{n}(\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}) \xrightarrow{D} \mathcal{N}(0, S),
$$

where $S = nCov(\widehat{\boldsymbol{\beta}}_{mle}) = n(X^\top \Sigma^{-1} X)^{-1}$ *and* $l > O_p(\log n)$.

Proof of Lemma (A2): We first focus on the relation between the true correlation matrix Σ and sparse version Σ_t. According to Golub and Van Loan (1983) and under the regularity conditions in Appendix B, the spectral norm could be bounded by the matrix L_1 norm for any $n \times n$ symmetric matrix A, i.e.

$$\|A\|_2 \leq \|A\|_1 = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}| \tag{2.21}$$

where $\|\cdot\|_2$ denotes the spectral norm of a matrix and $\|\cdot\|_1$ denotes the L_1 norm. we then have

$$\|\Sigma - \Sigma_t\|_2 \leq \|\Sigma - \Sigma_t\|_1 = \max_i \sum_{j:|x_{i1}-x_{j1}|>\tau} |\sigma_{ij}| \leq ne^{-\tau\eta}, \tag{2.22}$$

where $\|\cdot\|_2$ denotes the spectral norm of a matrix and $\|\cdot\|_1$ denotes the L_1 norm. For any two square matrices A and B, we have $\|AB\|_2 \leq \|A\|_2 \|B\|_2$. Thus,

$$\left\|\Sigma^{-1} - \Sigma_t^{-1}\right\|_2 = \left\|\Sigma_t^{-1}(\Sigma - \Sigma_t)\Sigma^{-1}\right\|_2 \leq \left\|\Sigma_t^{-1}\right\|_2 \|\Sigma - \Sigma_t\|_2 \left\|\Sigma^{-1}\right\|_2 \leq ne^{-\tau\eta}/c_1 c_1^*, \tag{2.23}$$

where η, c_1, c_1^* are all positive constants based on the regularity conditions.

Assume that the design matrix $X = O_p(1)$. We have $X^\top X = O_p(n)$ and based on (2.22) and (2.23), it follows

$$\left\|X^\top \Sigma_t^{-1} X - X^\top \Sigma^{-1} X\right\|_2 \leq Mn^2 e^{-\tau\eta},$$
$$\left\|(X^\top \Sigma_t^{-1} X)^{-1} - (X^\top \Sigma^{-1} X)^{-1}\right\|_2 \leq Mn^2 e^{-\tau\eta}/c_1 c_1^*,$$

where M, c_1, c_2^* are all constants. It is also clear that $E(\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta}) = 0$ Next, we show that the covariance matrix of $\sqrt{n}(\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta})$ converges to S, where $S = n\text{Cov}(\widehat{\boldsymbol{\beta}}_{mle})$. First, decompose the covariance matrix by $\text{Cov}(\sqrt{n}(\widehat{\boldsymbol{\beta}}_t - \boldsymbol{\beta})) = n\text{Cov}(\widehat{\boldsymbol{\beta}}_t) = n\text{Cov}(\widehat{\boldsymbol{\beta}}_t) - S + S$ and therefore we only need to show the spectral norm of

$n\text{Cov}(\widehat{\boldsymbol{\beta}}_t) - S$ converges to 0 as the following:

$$\begin{aligned}
\left\|n\text{Cov}(\widehat{\boldsymbol{\beta}}_t) - S\right\|_2 &= n\left\|\text{Cov}(\widehat{\boldsymbol{\beta}}_t) - \text{Cov}(\widehat{\boldsymbol{\beta}}_{mle})\right\|_2 \\
&= n\left\|(X^\top\Sigma_t^{-1}X)^{-1}X^\top\Sigma_t^{-1}\Sigma\Sigma_t^{-1}X(X^\top\Sigma_t^{-1}X)^{-1} - (X^\top\Sigma^{-1}X)^{-1}\right\|_2 \\
&\leq n\left\|(X^\top\Sigma_t^{-1}X)^{-1}X^\top\Sigma_t^{-1}(\Sigma-\Sigma_t)\Sigma_t^{-1}X(X^\top\Sigma_t^{-1}X)^{-1}\right\|_2 \\
&\quad + n\left\|(X^\top\Sigma_t^{-1}X)^{-1} - (X^\top\Sigma^{-1}X)^{-1}\right\|_2 \\
&\leq n\left\|\Sigma-\Sigma_t\right\|_2\left\|\Sigma_t^{-1}\right\|_2\left\|(X^\top\Sigma_t^{-1}X)^{-1}\right\|_2 + Mn^2e^{-\tau\eta}/c_1c_1^*1 \\
&= ne^{-\tau\theta_1}/Mc_1^* + Mn^3e^{-\tau\eta}/c_1c_1^*.
\end{aligned}$$

When $n \to \infty$, in order to guarantee that both terms in the last equation go to 0, τ has to be larger than $3\log(n)/\eta$. Since η is a constant, it is clear that the covariance matrix of $\widehat{\boldsymbol{\beta}}_t$ converges to S when $\tau > O_p(\log(n))$. Thus, $\widehat{\boldsymbol{\beta}}_t$ has the same asymptotic properties as $\widehat{\boldsymbol{\beta}}_{mle}$. □

Proof of Theorem 1: Based on Lemma (A1), we know that the combined estimated $\widehat{\boldsymbol{\beta}}_c$ is equivalent to $\widehat{\boldsymbol{\beta}}_t$. In Lemma (A2), it is shown that $\widehat{\boldsymbol{\beta}}_t$ is asymptotically equivalent to $\widehat{\beta}_{mle}$. Combining the two results, it is clear that $\widehat{\boldsymbol{\beta}}_c$ is asymptotically equivalent to $\widehat{\boldsymbol{\beta}}_{mle}$ under the regularity conditions and when $\tau > O_p(\log(n))$. □

2.7.3 A.3 Proof of Theorem 2

Proof: The proof consists of three parts: i) In each iteration, $\widehat{\boldsymbol{\beta}}_c$ and $\widehat{\boldsymbol{\beta}}_t$ are shown to be asymptotically equivalent given $\boldsymbol{\theta}$, and similarly $\widehat{\boldsymbol{\theta}}_c$ and $\widehat{\boldsymbol{\theta}}_t$ are asymptotically equivalent given $\boldsymbol{\beta}$. ii) Iterative updates of $\boldsymbol{\beta}_t$ and $\boldsymbol{\theta}_t$ converge to which is solved jointly by maximizing $l_t(\boldsymbol{\beta}, \boldsymbol{\theta})$. iii) We show that $(\widehat{\boldsymbol{\beta}}_t, \widehat{\boldsymbol{\theta}}_t)$ has asymptotically the same distribution as the original MLE.

For notational simplicity, we illustrate the proof only for $p = 1$ but it can be easily extended to a general case. Denote $\nabla_\beta l(\beta,\theta) = \partial l(\beta,\theta)/\partial\beta$, $\nabla_\theta l(\beta,\theta) = \partial l(\beta,\theta)/\partial\theta$, $\nabla_{\beta\beta} l(\beta,\theta) = \partial^2 l(\beta,\theta)/\partial\beta^2$, $\nabla_{\theta\theta} l(\beta,\theta) = \partial^2 l(\beta,\theta)/\partial\theta^2$,

and $\nabla_{\beta\theta} l(\beta, \theta) = \partial^2 l(\beta, \theta)/\partial\beta\partial\theta$.

For Part i), Theorem 1 shows that $\hat{\beta}_c$ is equivalent to $\hat{\beta}_t$ given θ. It suffices to show the same result for θ. Assuming that β is fixed, we apply Taylor expansion to $l_t(\theta|\beta)$ at $\theta = \widehat{\theta}_t$,

$$\nabla_\theta l_t(\widehat{\theta}_t|\beta) = \nabla_\theta l_t(\theta_0|\beta) + \nabla_{\theta\theta} l_t(\theta_0|\beta)(\widehat{\theta}_t - \theta_0) + O_p(1).$$

Since $\nabla_\theta l_t(\widehat{\theta}_t|\beta) = 0$, $\widehat{\theta}_t$ can be written as

$$\widehat{\theta}_t = \theta_0 - \nabla_{\theta\theta}^{-1} l_t(\theta_0|\beta)\nabla_\theta l_t(\theta_0|\beta) + O_p(1/n).$$

Similarly, under the regularity conditions in Appendix B, we have:

$$\begin{aligned}
\widehat{\theta}_c &= (\sum_{a=1}^{m} S_a^{-1})^{-1}(\sum_{a=1}^{m} S_a^{-1}\widehat{\theta}_a) \\
&= (\sum_{a=1}^{m} -\nabla_{\theta\theta} l_t^{(a)}(\theta_0|\beta))^{-1}(\sum_{a=1}^{m} -\nabla_{\theta\theta} l_t^{(a)}(\theta_0|\beta)(\theta_0 - \nabla_{\theta\theta}^{-1} l_t^{(a)}(\theta_0|\beta)\nabla_\theta l_t^{(a)}(\theta_0|\beta) + O_p(1/n_a))) \\
&= \theta_0 - \nabla_{\theta\theta}^{-1} l_t(\theta_0|\beta)\nabla_\theta l_t(\theta_0|\beta) + o_p(1/\sqrt{n}) + O_p(1/\tau).
\end{aligned}$$

Because $\tau > O_p(\sqrt{n})$, it holds that $(1/\tau)/(1/\sqrt{n}) \to 0$ when $n \to \infty$. Therefore, given β, $\widehat{\theta}_c$ is asymptotically equivalent to $\widehat{\theta}_t$.

To prove Part ii), we first perform Taylor expansion at the tth iteration:

$$0 = \nabla_\beta l_t(\beta^{(t)}, \theta^{(t-1)}) = \nabla_\beta l_t(\beta_0, \theta_0) + \nabla_{\beta\theta} l_t(\beta_0, \theta_0)(\theta^{(t-1)} - \theta_0) + \nabla_{\beta\beta} l_t(\beta_0, \theta_0)(\beta^{(t)} - \beta_0) + R_1, \tag{2.24}$$

$$0 = \nabla_\theta l_t(\beta^{(t)}, \theta^{(t)}) = \nabla_\theta l_t(\beta_0, \theta_0) + \nabla_{\theta\beta} l_t(\beta_0, \theta_0)(\beta^{(t)} - \beta_0) + \nabla_{\theta\theta} l_t(\beta_0, \theta_0)(\theta^{(t)} - \theta_0) + R_2, \tag{2.25}$$

where $l_t(\cdot)$ is the approximated log-likelihood, R_1 and R_2 are the remainders of the expansion. Therefore, after solving the last two equations, the tth updates of β and θ can be written as

$$\begin{pmatrix} \beta^{(t)} - \beta_0 \\ \theta^{(t)} - \theta_0 \end{pmatrix} = \boldsymbol{a} + b \begin{pmatrix} \beta^{(t-1)} - \beta_0 \\ \theta^{(t-1)} - \theta_0 \end{pmatrix},$$

where $\boldsymbol{a} = ((R_2 + \nabla_\theta l_t(\beta_0,\theta_0))/(\nabla_{\beta\beta} l_t(\beta_0,\theta_0)\nabla_{\theta\theta} l_t(\beta_0,\theta_0)) - (R_1 + \nabla_\beta l_t(\beta_0,\theta_0))/\nabla_{\beta\beta} l_t(\beta_0,\theta_0)$, $(R_1 + \nabla_\beta l_t(\beta_0,\theta_0))/(\nabla_{\beta\beta} l_t(\beta_0,\theta_0)\nabla_{\theta\theta} l_t(\beta_0,\theta_0)) - (R_2 + \nabla_\theta l_t(\beta_0,\theta_0))/\nabla_{\theta\theta} l_t(\beta_0,\theta_0))^\top$, $b = \nabla^2_{\beta\theta} l_t(\beta_0,\theta_0)/(\nabla_{\beta\beta} l_t(\beta_0,\theta_0)\nabla_{\theta\theta} l_t(\beta_0,\theta_0))$. Therefore we have

$$\begin{pmatrix} \beta^{(t)} - \beta^{(t-1)} \\ \theta^{(t)} - \theta^{(t-1)} \end{pmatrix} = b \begin{pmatrix} \beta^{(t-1)} - \beta^{(t-2)} \\ \theta^{(t-1)} - \beta^{(t-2)} \end{pmatrix} = \cdots = b^{t-1} \begin{pmatrix} \beta^{(1)} - \beta^{(0)} \\ \theta^{(1)} - \theta^{(0)} \end{pmatrix} \tag{2.26}$$

Because Hessian matrix $H(\beta_0,\theta_0)$ should be positive definite i.e. $|H| > 0$, we have $\nabla_{\beta\beta} l_t(\beta_0,\theta_0)\nabla_{\theta\theta} l_t(\beta_0,\theta_0) > \nabla^2_{\beta\theta} l_t(\beta_0,\theta_0)$, indicating that $b < 1$. Thus, when the iteration step t is large, the updates will converge based on (2.26). Assume that after T steps, $(\beta^{(T)},\theta^{(T)})$ converge and from (2.24) and (2.25), we have

$$\begin{pmatrix} \beta^{(T)} \\ \theta^{(T)} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \theta_0 \end{pmatrix} - \begin{pmatrix} \nabla_{\beta\beta} l_t(\beta_0,\theta_0) & \nabla_{\beta\theta} l_t(\beta_0,\theta_0) \\ \nabla_{\theta\beta} l_t(\beta_0,\theta_0) & \nabla_{\theta\theta} l_t(\beta_0,\theta_0) \end{pmatrix}^{-1} \begin{pmatrix} \nabla_\beta l_t(\beta_0,\theta_0) \\ \nabla_\theta l_t(\beta_0,\theta_0) \end{pmatrix} + o_p(1/\sqrt{n}). \tag{2.27}$$

Denote $(\widehat{\beta}_t,\widehat{\theta}_t)$ as the estimator that maximizing $l_t(\beta,\theta)$. Taylor expansion of $l_t(\widehat{\beta}_t,\widehat{\theta}_t)$ leads to the same results for $(\widehat{\beta}_t,\widehat{\theta}_t)$ asymptotically as described in (2.27). Therefore, the result follows.

To prove Part iii), we first show that $l_t(\beta,\theta)$ converges to $l(\beta,\theta)$ as follows:

$$\begin{aligned}
|l(\beta,\theta) - l_t(\beta,\theta)| &= \left| -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{x}\beta)^\top (\Sigma^{-1}(\theta) - \Sigma_t^{-1}(\theta))(\boldsymbol{y} - \boldsymbol{x}\beta) - \frac{1}{2}\log\left|\Sigma(\theta)\Sigma_t^{-1}(\theta)\right| \right| \\
&\le \frac{M}{2\sigma^2} n^2 e^{-\tau\theta_1} + \frac{1}{2}\left|\log|I - (\Sigma_t(\theta) - \Sigma(\theta))\Sigma_t^{-1}(\theta)|\right| \\
&= \frac{M}{2\sigma^2} n^2 e^{-\tau\theta_1} + \frac{1}{2} tr((\Sigma_t(\theta) - \Sigma(\theta))\Sigma_t^{-1}(\theta)) + o(tr((\Sigma_t(\theta) - \Sigma(\theta))\Sigma_t^{-1}(\theta))) \\
&\le \frac{M}{2\sigma^2} n^2 e^{-\tau\theta_1} + \frac{n}{2}\left\|(\Sigma_t(\theta) - \Sigma(\theta))\Sigma_t^{-1}(\theta)\right\|_1 + o(tr((\Sigma_t(\theta) - \Sigma(\theta))\Sigma_t^{-1}(\theta))) \\
&\le \left(\frac{M}{2\sigma^2} + \frac{1}{2}\right) n^2 e^{-\tau\theta_1} + o(n^2 e^{-\tau\theta_1}),
\end{aligned}$$

where $\tau > O_p(\log n)$. Therefore, $|l(\beta,\theta) - l_t(\beta,\theta)| \to 0$ when $n \to \infty$. Since both $l(\beta,\theta)$ and $l_t(\beta,\theta)$ are continuous in β,θ, we have their first and second derivatives converge as well. Therefore, the joint distribution of $(\widehat{\beta}_t,\widehat{\theta}_t)$ is asymptotically equivalent to the the joint distribution of MLE. □

2.7.4 A.4 Proof of Corollary 1

Proof: To show that the combined estimator is unbiased, we first prove that

$$tr(\Sigma_t^1(\Sigma - \Sigma_t)) \to 0 \text{ a.s.},$$

where Σ is the true correlation matrix, $\Sigma_t^1 = -\Sigma_t^{-1}\Sigma_t'\Sigma_t^{-1}$, and $\Sigma_t' = \partial\Sigma_t/\partial\boldsymbol{\theta}$. From Golub and Van Loan (1983), it is known that,

$$tr(A) \leq \sqrt{rank(A)}||A||_F \leq n||A||_2 \leq n||A||_1$$

where A is a full rank positive definite matrix. Therefore,

$$tr(\Sigma_t^1(\Sigma - \Sigma_t)) \leq n||\Sigma_t^1(\Sigma - \Sigma_t)||_1 \leq n(n-l)e^{-\tau\eta}, \tag{2.28}$$

where η is a constant. Because $\tau > O_p(\log n)$, the right-hand side of (2.28) goes to 0. Therefore, we have $E(l_t'(\theta)) = 0$. By Taylor expansion, we have

$$l_t'(\widehat{\theta}_t) = l_t'(\theta) + l_t''(\theta)(\widehat{\theta}_t - \theta) + O_p(1).$$

Since $l_t'(\widehat{\theta}_t) = 0$, we have

$$\widehat{\theta}_t = \theta - l_t'(\theta)/l_t''(\theta) + O_p(1/n). \tag{2.29}$$

By taking expectation on both sides of (2.29), $\widehat{\theta}_t$ is shown to be unbiased, i.e. $E(\widehat{\theta}_t) = \theta$. From the result of Theorem 2, we know that the combined estimator is asymptotically equivalent to $\widehat{\theta}_t$. Therefore, the result for the combined estimator follows. □

2.7.5 A.5 Proof of Theorem 3

Proof: Similarly to Theorem 2, the proof is illustrated by $p = 1$ for notational simplicity. We have

$$\begin{aligned} p_1(\beta,\theta) &= x_0\beta + \sum_{a=1}^{m} \gamma_a^*(\theta)^\top D_a^{-1}(\theta)\boldsymbol{y}_a^* - \sum_{a=1}^{m} \gamma_a^*(\theta)^\top D_a^{-1}(\theta)\boldsymbol{c}_a(\theta)\beta \\ &= x_0\beta + \sum_{a=1}^{m} \gamma_a^*(\theta)^\top D_a^{-1}(\boldsymbol{y}_a^* - \boldsymbol{c}_a(\theta)\beta) \\ &= x_0\beta + \gamma(\theta)^\top B^\top(\theta) D^{-1}(\theta) B(\theta)(\boldsymbol{y} - \boldsymbol{x}\beta) \\ &= x_0\beta + \gamma(\theta)^\top \Sigma_t^{-1}(\theta)(\boldsymbol{y} - \boldsymbol{x}\beta), \end{aligned}$$

where $\boldsymbol{c}_a(\theta)$ is a vector version of $C_a(\theta)$. The difference between $p_1(\beta,\theta)$ and $p_0(\beta,\theta)$ can be written as

$$|p_1(\beta,\theta) - p_0(\beta,\theta)| = |\gamma(\theta)^\top(\Sigma_t^{-1}(\theta) - \Sigma^{-1}(\theta))(\boldsymbol{y} - \boldsymbol{x}\beta)| \leq Mn^2 e^{-\tau\theta}.$$

When $\tau > O_p(\log n)$ and $n \to \infty$, it holds that $p_1(\beta,\theta) \to p_0(\beta,\theta)$. Similarly, we can prove that $m_1(\beta,\theta) \to m_0(\beta,\theta)$ under the assumptions and regularity conditions. □

2.8 Appendix B: Regularity Conditions

B1). The theoretical properties are developed under increasing-domain asymptotics and additionally we assume that

(a) Block size n_a goes to infinity in same rate for $a = 1, ..., m$; Range on each dimension M_k goes to infinity in same rate for $k = 1, ..., p$.

(b) Tuning parameter τ goes to infinity at the same rate as n_a's, i.e.

$$n_a = O_p(\tau) \quad \text{for} \quad a = 1, .., m$$

(c) Range on each dimension M_k goes to infinity as the same rate as sample size n, i.e.

$$M_k = O_p(n) \quad \text{for} \quad k = 1, ..., p$$

B2). Assume that Σ is a symmetric positive definite matrix and it follows the following regularity conditions:

(a) Σ^{-1} is also a symmetric positive definite matrix;

(b) all $\lambda(\Sigma)'s > 0$;

(c) $0 < c_1 \leq \lambda_{min}(\Sigma) \leq \lambda_{max}(\Sigma) \leq c_2$;

(d) $0 < 1/c_2 \leq \lambda_{min}(\Sigma^{-1}) \leq \lambda_{max}(\Sigma^{-1}) \leq 1/c_1$.

where $\lambda(\cdot)$ is the eigenvalue of Σ, $\lambda(\cdot)_{max}, \lambda(\cdot)_{min}$ represent the maximum and minimum eigenvalue respectively, c_1, c_2 are two positive constants.

2.9 Appendix C: Algorithm

Algorithm 1 SSCC estimation of $(\boldsymbol{\beta}, \boldsymbol{\theta})$

SPLIT: $\boldsymbol{y}$ into $(\boldsymbol{y}_1, ..., \boldsymbol{y}_m)$ m blocks based on the choice of τ

Initialization: $(\boldsymbol{\beta}, \boldsymbol{\theta}) = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}), \quad t = 1$

CONQUER $\boldsymbol{\beta}^{(t)}|\boldsymbol{\theta}^{(t-1)}$: $\boldsymbol{y}_1^* = \boldsymbol{y}_1, C_1 = X_1, \Sigma_t = \Sigma_t(\boldsymbol{\theta}^{(t-1)}), D_1 = \Sigma_{t11}, L_{21} = \Sigma_{t21}\Sigma_{t11}^{-1}$

for $a = 2, ..., m$ **do**

  for $b = 1, ..., a-1$ **do**

    $B_{ab} = \prod_{k=b+1}^{a}(-L_{k(k-1)})$

  end for

  $\boldsymbol{y}_a^* = \boldsymbol{y}_a - L_{a(a-1)}\boldsymbol{y}_{a-1}^*$, $C_a = X_a + \sum_{b=1}^{a-1} B_{ab}X_b$ ▷ sequentially update

  $D_a = \Sigma_{taa} - L_{a(a-1)}D_{(a-1)}L_{a(a-1)}^\top$, $L_{(a+1)a} = \Sigma_{t(a+1)a}D_a^{-1}$

  Conquer on the current updated block:

$$\widehat{\boldsymbol{\beta}}_a = (C_a^\top D_a^{-1} C_a)^{-1} C_a^\top D_a^{-1} \boldsymbol{y}_a^*, \quad W_a = C_a^\top D_a^{-1} C_a$$

end for

COMBINE: Update $\boldsymbol{\beta}^{(t)}|\boldsymbol{\theta}^{(t-1)} = (\sum_{a=1}^m W_a)^{-1}(\sum_{a=1}^m W_a\widehat{\boldsymbol{\beta}}_a)$

CONQUER $\boldsymbol{\theta}^{(t)}|\boldsymbol{\beta}^{(t)}$: $\boldsymbol{y}_1^*(\boldsymbol{\theta}) = \boldsymbol{y}_1, C_1(\boldsymbol{\theta}) = X_1, D_1(\boldsymbol{\theta}) = \Sigma_{t11}(\boldsymbol{\theta}), L_{21}(\boldsymbol{\theta}) = \Sigma_{t21}(\boldsymbol{\theta})\Sigma_{t11}^{-1}(\boldsymbol{\theta})$

for $a = 2, ..., m$ **do**

  for $b = 1, ..., a-1$ **do**

    $B_{ab}(\boldsymbol{\theta}) = \prod_{k=b+1}^{a}(-L_{k(k-1)}(\boldsymbol{\theta}))$

  end for

  $\boldsymbol{y}_a^*(\boldsymbol{\theta}) = \boldsymbol{y}_a - L_{a(a-1)}(\boldsymbol{\theta})\boldsymbol{y}_{a-1}^*(\boldsymbol{\theta})$, $C_a(\boldsymbol{\theta}) = X_a + \sum_{b=1}^{a-1} B_{ab}(\boldsymbol{\theta})X_b$

  $D_a(\boldsymbol{\theta}) = \Sigma_{taa}(\boldsymbol{\theta}) - L_{a(a-1)}(\boldsymbol{\theta})D_{(a-1)}(\boldsymbol{\theta})L_{a(a-1)}^\top(\boldsymbol{\theta}), \quad L_{(a+1)a}(\boldsymbol{\theta}) = \Sigma_{t(a+1)a}(\boldsymbol{\theta})D_a^{-1}(\boldsymbol{\theta})$

  Conquer on the current updated block:

$$\widehat{\boldsymbol{\theta}}_a = \arg\max_{\boldsymbol{\theta}}\{-\frac{1}{2}\log|D_a(\boldsymbol{\theta})| - \frac{1}{2}(\boldsymbol{y}_a^*(\boldsymbol{\theta}) - C_a(\boldsymbol{\theta})\boldsymbol{\beta}^{(t)})^\top D_a^{-1}(\boldsymbol{\theta})(\boldsymbol{y}_a^*(\boldsymbol{\theta}) - C_a(\boldsymbol{\theta})\boldsymbol{\beta}^{(t)})\}$$

$$W_a = \widehat{S}_a^{-1} = -H_a(\widehat{\boldsymbol{\theta}}_a)$$

Algorithm 1 SSCC estimation of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ continued

21: **end for**

22: COMBINE: Update $\theta^{(t)}|\beta^{(t)} = (\sum_{a=1}^{m} W_a)^{-1}(\sum_{a=1}^{m} W_a \widehat{\boldsymbol{\theta}}_a)$

23: **if** $|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}| > \varepsilon$ or $|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}| > \varepsilon$ **then**

24: $t = t + 1$

25: Repeat Line 3 to 21

26: **else**

27: $(\widehat{\boldsymbol{\beta}}_c, \widehat{\boldsymbol{\theta}}_c) = (\boldsymbol{\beta}^{(t)}, \boldsymbol{\theta}^{(t)})$

28: **end if**

Chapter 3

Combining Information from Non-independent Studies by Confidence Distribution (CD)

3.1 Introduction

It is important to integrate information efficiently and effectively from multiple sources, especially in the era of data deluge nowadays. During the past decades, there have been rapid developments in statistical methodologies on combining information from multiple sources, such as Meta-analysis (Hedges and Olkin, 1985; Stangl and Berry, 2000; Schulze, 2004), Divide-and-Conquer methodologies from a statistical perspective (Chen and Xie, 2014b; Zhang et al., 2015; Battey et al., 2015) etc. Among most of the previous literatures, the individual studies are assumed to be independent. However, in reality the dependency can not be neglected under some circumstances. In this paper, we propose a combining methodology through a confidence distribution framework without the assumption of independence.

Confidence distribution (CD) refers to any sample-dependent distribution function that can represent confidence intervals/regions of all levels for a parameter of interest (cf., e.g., Xie and Singh (2013)). Conceptually, a CD is not different from a point estimator or a confidence interval, but it uses a sample-dependent distribution function on the parameter space to estimate the parameter of interest. A CD is to "provide simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)" (Cox, 2013). It can provide meaningful answers for all questions related to statistical inferences and

an approach that combines CDs preserves more information than a traditional approach that combines just point estimators (Xie and Singh, 2013; Schweder and Hjort, 2016). Singh et al. (2005) and Xie et al. (2011) described a general framework to combining information based on CDs, which can subsume almost all information combination methods used in the current practice. Recently, this approach is widely employed and adopted in the literatures (Liu et al., 2015; Yang et al., 2014; Claggett et al., 2014). However, these applications highly rely on the assumption that the underlying studies are independent. In this paper, we extend the combining recipe to studies without the assumption of independence.

The remainder of this paper is organized as follows. In Section 3.2, the proposed methodology is introduced along with a copula representation for combining non-independent studies. We also provide the framework through an example on combining dependent likelihood functions. In Section 3.3 we illustrate the framework under the scenario that the individual CD's are constructed based on the local parametric estimators from a partial linear model. Simulation studies and a real data example are included in this section. Summary and concluding remarks are given in Section 3.4.

3.2 Methodology

Suppose that there are K non-independent studies sharing the same underlying parameter of interest θ and its true value is θ_0. We also have the confidence distribution (CD) of θ constructed from each study, $H_i(\theta) = H_i(\theta, \mathbf{X}_i), i = 1, ..K$, where $\mathbf{X}_i$ is the ith sample with size n_i. Please note that $\mathbf{X}_i$'s are not assumed to be mutually independent. By definition, at the true value $\theta = \theta_0$, $H_i(\theta_0)$ is $\mathcal{U}[0, 1]$ distributed. Furthermore, we assume that the joint distribution of $(H_1(\theta_0), ..., H_K(\theta_0))$ is known, say

$$H_1(\theta_0), ..., H_K(\theta_0) \sim F(z_1, ..., z_K) = Pr(H_1(\theta_0) \leq z_1, ..., H_K(\theta_0) \leq z_K). \quad (3.1)$$

Our question is how to combine these CD functions from K non-independent studies. We propose the following combining recipe which is similar to the approach suggested by Singh et al. (2005) assuming independence among studies,

$$H_c^*(\theta) = G_c^*\{g_c^*(H_1(\theta), ..., H_K(\theta))\}, \tag{3.2}$$

where $H_i(\theta)$ is the CD for the ith study. The function $g_c^*(z_1, ..., z_K)$ is any continuous function from the K-dimensional hypercube $[0,1]^K$ to the real line $\mathbb{R} = (-\infty, +\infty)$ which is monotonic on each coordinate. Finally, the function $G_c^*(\cdot)$ is the cumulative distribution function for $g_c^*(U_1, ..., U_K)$ i.e.

$$G_c^*(z) = Pr(g_c^*(U_1, ..., U_K) \leq z) = \int_{g_c^*(z_1,...,z_K)\leq t} dF(z_1, ..., z_K), \tag{3.3}$$

where U_i's are marginally $\mathcal{U}[0,1]$ distributed and jointly follow the distribution whose CDF is $F(\cdot)$. Since $F(\cdot)$ is assumed to be known, $G_c^*(\cdot)$ is well defined. Firstly, we demonstrate that $H_c^*(\theta)$ from (3.2) is a valid CD by the following theorem,

Theorem 4. *Under the settings of* (3.1) *and* (3.2)*, the proposed function* $H_c^*(\theta)$ *is a CD for* θ*.*

Proof of Theorem 1 is provided in Appendix A. From Theorem 1, we then can make valid inferences of θ through the combined CD $H_c^*(\theta)$. Moreover, in Singh et al. (2005), the proposed combined CD $H^{(c)}(\theta)$ from independent studies is demonstrated to be the most efficient in terms of the Bahadur slope by choosing $g_c(z_1, ..., z_K) = \mathrm{DE}^{-1}(z_1) + \cdots + \mathrm{DE}^{-1}(z_K)$, where $\mathrm{DE}(\cdot)$ is the cumulative distribution function of the standard double exponential distribution and K is bounded asymptotically. However, in this paper, we are not focusing on the Bahadur slope because of some technical problems when dealing with nonexact inferences. Instead of talking about the most efficient of the Bahadur slope, in this paper, we view the combined CD $H_c^*(\theta)$ from a more practical perspective, i.e. with certain choices of $g_c^*(\cdot)$, we would be able to make inferences on θ that contain

all the information from K studies without the assumption of independence. Regarding the choice of function $g_c^*(\cdot)$, Xie et al. (2011) studied the following form,

$$g_c^*(z_1, ..., z_K) = w_1 F_0^{-1}(z_1) + \cdots + w_K F_0^{-1}(z_K), \tag{3.4}$$

where $F_0(\cdot)$ is a given cumulative distribution function and $w_i > 0$, with at least one $w_i \neq 0$. The details about choices of weights will be elaborated in the case study section under different settings.

3.2.1 Copula Representation

To derive the form of combined CD $H_c^*(\theta)$, we need to know the function $G_c^*(\cdot)$ which is determined by $g_c^*(\cdot)$. Because often times the joint distributions of multiple random variables do not have explicit forms and the marginal distribution of each component in $g_c^*(\cdot)$ is $\mathcal{U}[0,1]$ distributed, we then can employ copula to represent the joint distribution $F(\cdot)$. The joint distribution function $F(\cdot)$ can be represented by copula according to Sklar's theory,

Lemma 2. *(Sklar, 1973) Suppose that $F(\cdot)$ is a distribution function on $\mathbb{R}^K$ with marginal distribution functions $F_1(\cdot), ..., F_K(\cdot)$ on each direction. Then there is a copula $C(\cdot)$ such that:*

$$C(F_1(x_1), ..., F_K(x_K)) = F(x_1, ..., x_K). \tag{3.5}$$

If $F(\cdot)$ is continuous, then the copula $C(\cdot)$ satisfies:

$$C(z_1, ..., z_K) = F(F_1^{-1}(z_1), ..., F_K^{-1}(z_K)), \tag{3.6}$$

where $0 < z_i < 1$, for $i = 1, ..., K$.

Since we are focusing on the joint distribution of CD functions which follow marginal $\mathcal{U}[0,1]$ distributions, the following corollary is useful.

Corollary 2. *If a joint distribution $F(z_1, \ldots, z_K)$ is marginally $\mathcal{U}[0,1]$ distributed on each direction, the following holds:*

$$C(z_1, ..., z_K) = F(z_1, \ldots, z_K), \tag{3.7}$$

where $C(\cdot)$ is a copula.

Based on Corollary 2, the problem of joint distribution function of CD functions can be equivalently substituted by a valid copula. Given the copula $C(z_1, ..., z_K)$, the function $G_c^*(\cdot)$ can be represented as follows,

$$G_c^*(z) = Pr(g_c^*(U_1, ..., U_K) \leq z) = \int_{g_c^*(z_1,...,z_K)\leq z} dC(z_1, ..., z_K). \tag{3.8}$$

For different choices of $g_c^*(\cdot)$'s, there may not be explicit formula for the corresponding $G_c^*(\cdot)$. Assuming that we know how to generate random samples from copula $C(z_1, ..., z_K)$, we may use *Monte Carlo* integration method to get $G_c^*(\cdot)$. We propose the following algorithm to approximate $G_c^*(\cdot)$ in practice:

1. For $j = 1, .., N$, simulate a random sample $(z_1^j, ..., z_K^j)$ from copula $C(z_1, ..., z_K)$;

2. Then $G_c^*(z)$ can be approximated by the empirical CDF as following:

$$\widetilde{G}_c^*(z) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{1}_{\{g_c^*(z_1^j,...,z_K^j)\leq z\}} \tag{3.9}$$

   where $\mathbf{1}_{\{\cdot\}}$ is indicator function.

Obviously, it is not desirable to resort to numerical approximations of $G_c^*(\cdot)$ due to considerations on statistical and computational efficiency. Fortunately, there exist some convenient and valid choices of copulas that can bypass the above numerical procedures to obtain a closed-form expression of $G_c^*(\cdot)$. In this paper, we adopt the Gaussian copula which is one of the most commonly used copulas in practice. For the Gaussian copula, the individual functions $F_i(\cdot)$'s are CDFs of standard normal variables while $F(\cdot)$ is therefore a multivariate normal

CDF. According to Lemma 4 and Corollary 2, this is equivalent to using a copula of the following form,

$$C(z_1, ..., z_K) = \mathbf{\Phi}_R(\Phi^{-1}(z_1), ..., \Phi^{-1}(z_K)), \tag{3.10}$$

where $\mathbf{\Phi}_R(\cdot)$ is CDF of a K-dimensional multivariate normal distribution with mean 0 and covariance matrix R which is assumed known in this paper.

3.2.2 Example: Combining Dependent Likelihood by CD-based Approach

In this sub-section, we are going to illustrate the proposed CD based combining approach through the likelihood context. Fraser and Reid (2015) proposed a method of combining likelihood functions through score functions in composite likelihood context. Inspired by the importance of score function in likelihood inferences, our combining framework also uses score functions to construct individual CDs. In the following, we present how our combining recipe subsumes the approach proposed by Fraser and Reid (2015) (See Appendix B).

Suppose that $\ell_1(\theta), ..., \ell_K(\theta)$ are K dependent log-likelihood functions and the score function is $s_i(\theta) = \ell_i'(\theta)$ for the ith study. Based on the Taylor expansion of the score function around the true value $\theta = \theta_0$, we have

$$s(\theta) \approx s(\theta_0) + s'(\theta_0)(\theta - \theta_0). \tag{3.11}$$

The expectation of the score function can be easily derived as

$$E(s(\theta)) = E(s'(\theta_0))\theta - E(s'(\theta_0))\theta_0. \tag{3.12}$$

Let $\boldsymbol{s}^0 = \boldsymbol{s}(0)$, where $\boldsymbol{s}(\theta) = (s_1(\theta), ..., s_K(\theta))^\top$ and $E(s_i^0) = -E(s'(\theta_0))\theta_0$. The score function vector $\boldsymbol{s}^0$ then follows an K-dimensional asymptotic multivariate normal distribution,

$$\boldsymbol{s}^0 \sim \mathcal{N}_K(V\theta, W), \tag{3.13}$$

where W's ith element of $v_i = -E(s_i'(\theta_0)) = V_{ii}$, the covariance matrix V's ijth element is assumed to be known as $V_{ij} = \text{Cov}(s_i(\theta_0), s_j(\theta_0))$. By assumption, each element of $\boldsymbol{s}^0$ follows an asymptotic normal distribution,

$$s_i^0 \sim \mathcal{N}(V_{ii}\theta, V_{ii}), \tag{3.14}$$

where s_i^0 is the ith element of $\boldsymbol{s}^0$, V_{ii} is the Fisher's information for ith study. Then for $i = 1, \ldots, K$, we construct a CD function for parameter θ as follows,

$$H_i(\theta) = \Phi\left(\frac{V_{ii}\theta - s_i^0}{\sqrt{V_{ii}}}\right) = \Phi\left(\sqrt{V_{ii}}(\theta - s_i^0/V_{ii})\right), \tag{3.15}$$

where $\Phi(\cdot)$ is standard normal CDF. Given the choice of $g_c^*(\cdot)$ in (3.4) from Xie et al. (2011) and $F_0(\cdot) = \Phi(\cdot)$, $g_c^*(\cdot)$ is written as:

$$g_c^*(z_1, ..., z_K) = w_1\Phi^{-1}(z_1) + \cdots + w_K\Phi^{-1}(z_K), \tag{3.16}$$

where w_i's are the weights. To derive the cumulative distribution function $G_c^*(\cdot)$ for $g_c^*(\cdot)$, we represent the joint distribution of $(\Phi^{-1}(U_1), ..., \Phi^{-1}(U_K))$ by the Gaussian copula (3.10) introduced previously, where U_i's are $\mathcal{U}[0,1]$ distributed random variables. The distribution of $g_c^*(\cdot)$ can be easily derived as,

$$g_c^*(U_1, ..., U_K) \sim N(0, \boldsymbol{w}^\top R\boldsymbol{w}), \tag{3.17}$$

where $R = S^{-1}WS^{-1}$ and $S = \text{diag}(\sqrt{V_{11}}, \ldots, \sqrt{V_{KK}})$. Therefore, the CDF $G_c^*(z) = \Phi(z/\sqrt{\boldsymbol{w}^\top R\boldsymbol{w}})$. After plugging in the individual CD functions $H_i(\theta)$'s, $g_c^*(\cdot)$ and $G_c^*(\cdot)$ into the proposed combining recipe (3.2), the combined CD is given by,

$$H_c^*(\theta) = \Phi\left(\frac{1}{\sqrt{\boldsymbol{w}^\top R\boldsymbol{w}}}\sum_{i=1}^{K} w_i\sqrt{V_{ii}}(\theta - s_i^0/V_{ii})\right) = \Phi\left(\frac{\boldsymbol{w}^\top S^{-1}V}{\sqrt{\boldsymbol{w}^\top R\boldsymbol{w}}}(\theta - \widehat{\theta}_c)\right), \tag{3.18}$$

where $\widehat{\theta}_c = (\boldsymbol{w}^\top S^{-1}V)^{-1}\boldsymbol{w}^\top S^{-1}\boldsymbol{s}^0$, $\text{Var}(\widehat{\theta}_c) = (\boldsymbol{w}^\top S^{-1}V)^{-1}\boldsymbol{w}^\top S^{-1}WS^{-1}\boldsymbol{w}(\boldsymbol{w}^\top S^{-1}V)^{-1}$. If we choose $\boldsymbol{w} = SW^{-1}V$, then

the combined estimator becomes $\widehat{\theta}_c = (V^\top W^{-1} V)^{-1} V^\top W^{-1} \boldsymbol{s}^0$, which is identical to the result from Fraser and Reid (2015). The new combined likelihood function $\ell_{\text{new}}(\theta)$ in (3.63) could be obtained accordingly.

Remark 2. *In this section, we only provide one way to construct individual CD function $H_i(\theta)$ by employing the corresponding score function. For likelihood function, Singh et al. (2005) proved that normalized likelihood function is a valid CD function as well. Therefore, the combined CD function $H_c^*(\theta)$ can also be obtained through this definition. The details will not be discussed in this paper but the results are identical which could be easily proved.*

3.3 Combining Parametric Components from Partial Linear Models

In this section, we illustrate the proposed methodology through a partial linear model framework along with numerical studies. A partial linear model or semi-parametric model is described by two sets of parameters, where one is parametric (finite-dimensional) and the other is non-parametric (infinite-dimensional). Partial linear models have various applications in many fields such as economics, medicine, biology, etc. Examples of former studies include Green and Yandell (1985), Dinse and Lagakos (1983), Green and Silverman (1994), Schmalensee and Stoker (1999), Speckman (1988), Engle et al. (1986) etc. Before we proceed with the specifics of our approach, we first provide some preliminaries of partial linear models and related work.

3.3.1 Partial Linear Models and Estimation

Assume that we have the following partial linear model,

$$y_i = x_i\beta + \boldsymbol{z}_i^\top \boldsymbol{\gamma} + \eta(t_i) + \varepsilon_i, i = 1, ..., n, \tag{3.19}$$

where y is the dependent variable, $(x, \boldsymbol{z}, t)$ are explanatory variables and ε_i's are i.i.d random variables following $\mathcal{N}(0, \sigma^2)$. In model (3.19), we assume that β is a scalar and $\boldsymbol{\gamma}$ is a $p \times 1$ vector. The parametric component of this model therefore is $(\beta, \boldsymbol{\gamma})$ and the non-parametric component is the unknown function $\eta(\cdot)$. In this article, we only focus making inference on β and treat $\boldsymbol{z}$ as 'nuisance' variable.

Model (3.19) is a simple version of the generalized partially linear single-index model (Carroll et al., 1997) that can be solved by the quasi-likelihood method (Severini and Staniswalis, 1994). A standard approach to estimate $(\beta, \boldsymbol{\gamma})$ under partial linear models consists of two steps. First, we estimate the non-parametric part as a function of $(\beta, \boldsymbol{\gamma})$ by maximizing the likelihood function weighted by a kernel function. Then, we plug in the non-parametric estimator in terms of $(\beta, \boldsymbol{\gamma})$ into the full likelihood function. Mathematically, this two-step procedure can be described as follows.

Step 1 *Denote the weighted log-likelihood function by*

$$\ell_{\mathrm{w}}(a, t) = -\frac{1}{2n\sigma^2} \sum_{i=1}^{n} K_h(t_i - t)(y_i - x_i\beta - \boldsymbol{z}_i^\top \boldsymbol{\gamma} - a)^2, \tag{3.20}$$

where $a = \eta(t)$ and $K_h(\cdot)$ is a kernel function with bandwidth h. Therefore, the non-parametric part $\eta(t, \beta)$ is estimated by solving the equation $\partial \ell_{\mathrm{w}}(a, t)/\partial a = 0$, i.e.

$$\widehat{\eta}(t, \beta, \boldsymbol{\gamma}) = \arg\max_a \ell_{\mathrm{w}}(a, t) = \boldsymbol{v}(t)^\top (\boldsymbol{y} - \boldsymbol{x}\beta - Z\boldsymbol{\gamma}), \tag{3.21}$$

where $\boldsymbol{v}(t) = (v_1(t), \ldots, v_n(t))^\top$, $v_i(t) = K_h(t_i - t)/\sum_{i=1}^n K_h(t_i - t)$, $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$, $Z = (\boldsymbol{z}_1^\top, \ldots, \boldsymbol{z}_n^\top)^\top$.

Step 2 *Denote the global likelihood function for $\beta, \boldsymbol{\gamma}$ by*

$$\ell_{\mathrm{global}}(\beta, \boldsymbol{\gamma}) = -\frac{1}{2n\sigma^2} \sum_{i=1}^{n} (y_i - x_i\beta - \boldsymbol{z}_i^\top \boldsymbol{\gamma} - \widehat{\eta}(t_i, \beta, \boldsymbol{\gamma}))^2, \tag{3.22}$$

where $\widehat{\eta}(t_i, \beta, \boldsymbol{\gamma})$, the non-parametric estimate in (3.21), takes the place of $\eta(t_i)$ in the full log-likelihood function. Then $(\beta, \boldsymbol{\gamma})$ can be estimated globally

by solving the equation $\partial \ell_{\text{global}}(\beta, \boldsymbol{\gamma})/\partial(\beta, \boldsymbol{\gamma}) = 0$*, which leads to*

$$\begin{pmatrix} \widehat{\beta}_{\text{global}} \\ \widehat{\boldsymbol{\gamma}}_{\text{global}} \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{x}}^\top \widetilde{\boldsymbol{x}} & \widetilde{\boldsymbol{x}}^\top \widetilde{Z} \\ \widetilde{\boldsymbol{x}}^\top \widetilde{Z} & \widetilde{Z}^\top \widetilde{Z} \end{pmatrix}^{-1} \widetilde{X}^\top \widetilde{\boldsymbol{y}}, \tag{3.23}$$

where $\widetilde{\boldsymbol{x}} = (I - U)\boldsymbol{x}$*,* $\widetilde{Z} = (I - U)Z$*,* $\widetilde{X} = (\widetilde{\boldsymbol{x}}, \widetilde{Z})$*,* $\widetilde{\boldsymbol{y}} = (I - U)\boldsymbol{y}$*,* $U = (\boldsymbol{v}(t_1)^\top, \ldots, \boldsymbol{v}(t_n)^\top)^\top$*.*

This global estimator $(\widehat{\beta}_{\text{global}}, \widehat{\boldsymbol{\gamma}}_{\text{global}})$ is shown to be consistent and asymptotically normal (Carroll et al., 1997), which is also equivalent to the least squares estimate in Härdle et al. (2012). Finally, with the estimate of the paramatric part, the nonparametric part can be estimated by

$$\widehat{\eta}(t) = \frac{\sum_{i=1}^n K_h(t_i - t)(y_i - x_i \widehat{\beta}_{\text{global}} - \boldsymbol{z}_i^\top \widehat{\boldsymbol{\gamma}}_{\text{global}})}{\sum_{i=1}^n K_h(t_i - t)}. \tag{3.24}$$

A similar estimation procedure can be found in Boente et al. (2006) by changing their loss functions into the above likelihood functions (i.e., $\ell_{\text{w}}(a, \beta, t)$ in (3.20) and $\ell_{\text{global}}(\beta, \boldsymbol{\gamma})$ in (3.22)).

The global estimation procedure for the parametric component can be regarded as estimating the linear regression model as follows,

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{x}}\beta + \widetilde{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{3.25}$$

In this study, since only the scalar β is parameter of interest, we employ the canonical form (Rao, 2009) of model (3.25) by transforming the column spaces of $\widetilde{\boldsymbol{x}}$ an $\widetilde{Z}$ into two orthogonal subspaces as follows,

$$\widetilde{\boldsymbol{y}} = \widetilde{\boldsymbol{x}}^*\beta + \widetilde{Z}\boldsymbol{\gamma}^* + \boldsymbol{\varepsilon}, \tag{3.26}$$

where $\widetilde{\boldsymbol{x}}^* = \widetilde{\boldsymbol{x}} - \widetilde{Z}(\widetilde{Z}^\top \widetilde{Z})^{-1}\widetilde{Z}^\top \widetilde{\boldsymbol{x}} = (I - P_{\widetilde{Z}})\widetilde{\boldsymbol{x}}$, $P_{\widetilde{Z}} = \widetilde{Z}(\widetilde{Z}^\top \widetilde{Z})^{-1}\widetilde{Z}^\top$ and $\boldsymbol{\gamma}^*$ is the corresponding parameter of $\widetilde{Z}$, and therefore, $\widetilde{Z}^\top \widetilde{\boldsymbol{x}}^* = (0, \ldots, 0)^\top$. Following the two-step conventional estimation procedure illustrated above, we can obtain the global estimator of $(\beta, \boldsymbol{\gamma}^*)$,

$$\begin{pmatrix} \widehat{\beta}_{\text{global}} \\ \widehat{\boldsymbol{\gamma}}^*_{\text{global}} \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{x}}^{*\top} \widetilde{\boldsymbol{x}}^* & \widetilde{\boldsymbol{x}}^{*\top} \widetilde{Z} \\ \widetilde{\boldsymbol{x}}^{*\top} \widetilde{Z} & \widetilde{Z}^\top \widetilde{Z} \end{pmatrix}^{-1} (\widetilde{\boldsymbol{x}}^{*\top}, \widetilde{Z}^\top)\widetilde{\boldsymbol{y}} = \begin{pmatrix} (\widetilde{\boldsymbol{x}}^{*\top} \widetilde{\boldsymbol{x}}^*)^{-1} \widetilde{\boldsymbol{x}}^{*\top} \widetilde{\boldsymbol{y}} \\ (\widetilde{Z}^\top \widetilde{Z})^{-1} \widetilde{Z}^\top \widetilde{\boldsymbol{y}} \end{pmatrix}. \tag{3.27}$$

Hence, the global estimator of β is $\widehat{\beta}_{\text{global}} = (\widetilde{\boldsymbol{x}}^{*\top}\widetilde{\boldsymbol{x}}^{*})^{-1}\widetilde{\boldsymbol{x}}^{*\top}\widetilde{\boldsymbol{y}}$. It is easily to prove that $\widehat{\beta}_{\text{global}}$ is equivalent to the one derived from (3.23).

3.3.2 The CD-based Combining Approach

In our framework, the goal is to obtain a combined estimator which is equivalent to $\widehat{\beta}_{\text{global}}$ from local estimates. Briefly speaking, the difference between global and local estimation is to employ different likelihood functions which are built on either the entire or partial dataset. More specifically, we first obtain local estimates derived in individual studies, and then combine the local results using our framework. Similar to the structure we laid out earlier for the global estimator $\widehat{\beta}_{\text{global}}$, we first derive the local maximum likelihood estimator for β based on *local likelihood function* through the two-step procedure. Next, we construct individual CDs based on local estimators. Finally, we illustrate how to achieve the global estimator by combining individual CDs using our recipe. The numerical results based on real datasets also corroborate that the combined result achieves the same performance vis-a-vis global estimator.

Now we elaborate on the two-step procedure for estimating $\widehat{\beta}$ locally in each study. Our approach is motivated by the technique proposed by Tibshirani and Hastie (1987), called *local likelihood estimation*. This technique was initially applied to generalized linear models and the *proportional hazards model*. A similar method that hinges on local likelihood estimation can also be found in Fan et al. (1997), which formulates the local log-likelihood (with kernel smoothing) as

$$\ell(\beta, \theta) = n^{-1} \sum_{i=1}^{n} l_i(\beta, \theta) K_h(X_i - x), \tag{3.28}$$

where $l_i(\beta, \theta)$ in the summand is the ith log-likelihood function, β and θ are the parameters of proportional hazards model and $K_h(t) = h^{-1}K(t/h)$ is some chosen kernel function with bandwidth h. For other applications of the local likelihood estimation approach, see Staniswalis (1989), Hjort and Glad (1995), Hjort and

Jones (1996), Loader et al. (1996) and Fan et al. (1998).

In this work, we make a first attempt in incorporating this line of approaches in our CD-based framework by extending its application in partial linear models. Note that due to the kernel smoothing imposed in the local likelihood function (3.28), there are dependencies among those local estimators. Denote the local estimator at t for $(\beta, \boldsymbol{\gamma})$ as $(\widehat{\beta}(t), \widehat{\boldsymbol{\gamma}}(t))$, the local estimation procedure is described as follows,

Step 1 *For each $(\beta, \boldsymbol{\gamma}, t)$, the nonparametric part can be estimated in the same way by solving $\partial \ell_{\text{profile}}(a,t)/\partial a = 0$,*

$$\widehat{\eta}(t, \beta, \boldsymbol{\gamma}) = \arg\max_a \ell_{\text{profile}}(a,t) = \boldsymbol{v}(t)^\top (\boldsymbol{y} - \boldsymbol{x}\beta - Z\boldsymbol{\gamma}). \tag{3.29}$$

Step 2 *Then the local log-likelihood function for β is constructed as*

$$\ell_{\text{local}}(\beta, \boldsymbol{\gamma}, t) = -\frac{1}{2n\sigma^2} \sum_{i=1}^{n} (y_i - x_i\beta - \boldsymbol{z}_i^\top \boldsymbol{\gamma} - \widehat{\eta}(t_i, \beta))^2 K_b(t_i - t) \tag{3.30}$$

$$= -\frac{1}{2n\sigma^2} (\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{x}}\beta - \widetilde{Z}\boldsymbol{\gamma})^\top K(t) (\widetilde{\boldsymbol{y}} - \widetilde{\boldsymbol{x}}\beta - \widetilde{Z}\boldsymbol{\gamma}), \tag{3.31}$$

where $K(t) = \text{diag}(K_b(t_1 - t), \ldots, K_b(t_n - t))$, $K_b(\cdot)$ is kernel function with bandwidth equals to b. Please note that the kernel function in ℓ_{local} doesn't have to be the same as the one for estimating the non-parametric part. Then local estimators of $(\beta, \boldsymbol{\gamma})$ can be solved simultaneously by maximizing (3.31) *which leads the results,*

$$\begin{pmatrix} \widehat{\beta}(t) \\ \widehat{\boldsymbol{\gamma}}(t) \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{x}}^\top K(t) \widetilde{\boldsymbol{x}} & \widetilde{\boldsymbol{x}}^\top K(t) \widetilde{Z} \\ \widetilde{Z}^\top K(t) \widetilde{\boldsymbol{x}} & \widetilde{Z}^\top K(t) \widetilde{Z} \end{pmatrix}^{-1} \widetilde{X}^\top K(t) \widetilde{\boldsymbol{y}}. \tag{3.32}$$

Similar to the global estimation procedure, the local procedure can also be seen as estimating the linear regression model in (3.25). However, the error term $\boldsymbol{\varepsilon}$ is assumed to follow a multivariate normal distribution with covariance matrix $K(t)^{-1}\sigma^2$. We then consider the linear regression model as follows,

$$\widetilde{\boldsymbol{y}}_t = \widetilde{\boldsymbol{x}}_t \beta + \widetilde{Z}_t \boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{3.33}$$

where $\widetilde{\boldsymbol{y}}_t = K(t)^{1/2}\widetilde{\boldsymbol{y}}$, $\widetilde{\boldsymbol{x}}_t = K(t)^{1/2}\widetilde{\boldsymbol{x}}$, $\widetilde{Z}_t = K(t)^{1/2}\widetilde{Z}$, $K(t)^{1/2}K(t)^{1/2} = K(t)$, $\boldsymbol{\varepsilon}$ is then can be assumed to be i.i.d $\mathcal{N}(0, \sigma^2)$. We then assume the canonical form for the local model as follows,

$$\widetilde{\boldsymbol{y}}_t = \widetilde{\boldsymbol{x}}_t^*\beta + \widetilde{Z}_t\boldsymbol{\gamma}^{**} + \boldsymbol{\varepsilon} \tag{3.34}$$

where $\widetilde{\boldsymbol{x}}_t^* = \widetilde{\boldsymbol{x}}_t - \widetilde{Z}_t(\widetilde{Z}_t^\top\widetilde{Z}_t)^{-1}\widetilde{Z}_t^\top\widetilde{\boldsymbol{x}}_t = (I - P_{\widetilde{Z}_t})\widetilde{\boldsymbol{x}}_t$, $P_{\widetilde{Z}_t} = \widetilde{Z}_t(\widetilde{Z}_t^\top\widetilde{Z}_t)^{-1}\widetilde{Z}_t^\top$ and $\boldsymbol{\gamma}^{**}$ is the corresponding parameter of $\widetilde{Z}_t$, and therefore, $\widetilde{Z}_t^\top\widetilde{\boldsymbol{x}}_t^* = (0, \ldots, 0)^\top$. Also, following the local estimation procedure , we can obtain the local estimator of $(\beta, \boldsymbol{\gamma}^{**})$,

$$\begin{pmatrix} \widehat{\beta}(t) \\ \widehat{\boldsymbol{\gamma}}^{**}(t) \end{pmatrix} = \begin{pmatrix} \widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{\boldsymbol{x}}_t^* & \widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{Z}_t \\ \widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{Z}_t & \widetilde{Z}_t^\top\widetilde{Z}_t \end{pmatrix}^{-1} (\widetilde{\boldsymbol{x}}_t^{*\top}, \widetilde{Z}_t^\top)\widetilde{\boldsymbol{y}}_t = \begin{pmatrix} (\widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{\boldsymbol{x}}_t^*)^{-1}\widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{\boldsymbol{y}}_t \\ (\widetilde{Z}_t^\top\widetilde{Z}_t)^{-1}\widetilde{Z}_t^\top\widetilde{\boldsymbol{y}}_t \end{pmatrix}. \tag{3.35}$$

Therefore, the local estimator of β is $\widehat{\beta}(t) = (\widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{\boldsymbol{x}}_t^*)^{-1}\widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{\boldsymbol{y}}_t = (\widetilde{\boldsymbol{x}}_t^{*\top}\widetilde{\boldsymbol{x}}_t^*)^{-1}\widetilde{\boldsymbol{x}}_t^{*\top}K(t)^{1/2}(I - U)\boldsymbol{y}$. From Fan et al. (1997), it is proved that the estimators of β and θ derived from (3.28) achieve a convergence rate of $(nh)^{-1/2}$, and that $(nh)^{-1/2}(\widehat{\beta}(t) - \beta_0)$ follows an asymptotic normal distribution with mean 0 and a constant variance, where β_0 is the true value of β. Therefore, for $i = 1, ..., n$, the individual CDs could be constructed as following,

$$H_i(\beta) = \Phi\left(\frac{\beta - \widehat{\beta}_i}{\widehat{\sigma}_i}\right), \tag{3.36}$$

where $\widehat{\beta}_i = \widehat{\beta}(t_i)$ and $\widehat{\sigma}_i$ is the standard deviation of $\widehat{\beta}_i$. Before proceeding to the combining part, we study the variance and dependency among $\widehat{\beta}_i$'s. Define $\widehat{\boldsymbol{\beta}}_{\text{local}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_n)^\top$, and $\widehat{\boldsymbol{\beta}}_{\text{local}}$ can be represented in terms of $\boldsymbol{x}, \boldsymbol{y}, K(\cdot)$ and U as follows,

$$\widehat{\boldsymbol{\beta}}_{\text{local}} = \begin{pmatrix} (\widetilde{\boldsymbol{x}}_{t_1}^{*\top}\widetilde{\boldsymbol{x}}_{t_1}^*)^{-1}\widetilde{\boldsymbol{x}}_{t_1}^{*\top}K(t_1)^{1/2} \\ \cdots \\ (\widetilde{\boldsymbol{x}}_{t_n}^{*\top}\widetilde{\boldsymbol{x}}_{t_n}^*)^{-1}\widetilde{\boldsymbol{x}}_{t_n}^{*\top}K(t_n)^{1/2} \end{pmatrix} (I - U)\boldsymbol{y}. \tag{3.37}$$

Denote V as the covariance matrix for $\widehat{\boldsymbol{\beta}}_{\text{local}}$, therefore, the ijth element of V is

$$V_{ij} = \text{Cov}(\widehat{\beta}_i, \widehat{\beta}_j) = (\widetilde{\boldsymbol{x}}_{t_i}^{*\top}\widetilde{\boldsymbol{x}}_{t_i}^*)^{-1}\widetilde{\boldsymbol{x}}_{t_i}^{*\top}K(t_i)^{1/2}(I - U)(I - U)^\top K(t_j)^{1/2}\widetilde{\boldsymbol{x}}_{t_j}^*(\widetilde{\boldsymbol{x}}_{t_j}^{*\top}\widetilde{\boldsymbol{x}}_{t_j}^*)^{-1}\sigma^2, \tag{3.38}$$

then $\widehat{\sigma}_i = \sqrt{V_{ii}}$, where σ is assumed known in this context but can be estimated through the mean squared error of the estimator. Let $\boldsymbol{w}$ be the weight vector for $g_c^*(\cdot)$, $F_0(\cdot) = \Phi(\cdot)$, $g_c^*(z_1, ..., z_n) = \sum_{i=1}^n w_i \Phi^{-1}(z_i)$, and similar to the example we illustrated previously, we apply Gaussian copula. Hence the CDF of $g_c^*(U_1, ..., U_n)$ can be easily derived as,

$$G_c^*(z) = \Phi(z/\sqrt{\boldsymbol{w}^\top R \boldsymbol{w}}), \tag{3.39}$$

where $R = S^{-1}VS^{-1}$ and $S = \text{diag}\,(\widehat{\sigma}_1, ..., \widehat{\sigma}_n)$. Then the combined CD can be obtained as

$$H_c^*(\beta) = \Phi\left(\frac{1}{\sqrt{\boldsymbol{w}^\top R \boldsymbol{w}}} \sum_{i=1}^n w_i \frac{\beta - \widehat{\beta}_i}{\widehat{\sigma}_i}\right) = \Phi\left(\frac{\boldsymbol{w}^\top S^{-1}\mathbf{1}}{\sqrt{\boldsymbol{w}^\top R \boldsymbol{w}}}(\beta - \widehat{\beta}_c)\right), \tag{3.40}$$

where $\mathbf{1} = (1, ..., 1)^\top$, the combined estimator derived from the combined CD

$$\widehat{\beta}_c = (\boldsymbol{w}^\top S^{-1}\mathbf{1})^{-1} \boldsymbol{w}^\top S^{-1} \widehat{\boldsymbol{\beta}}_{\text{local}}. \tag{3.41}$$

Among different choices of $\boldsymbol{w}$, we then have the following theorem,

Theorem 5. *If the weight vector $\boldsymbol{w}^\top = \boldsymbol{c}^\top B^{-1} S$, the combined estimator $\widehat{\beta}_c$ is equivalent to $\widehat{\beta}_{\text{global}}$, where $S = diag\,(\widehat{\sigma}_1, ..., \widehat{\sigma}_n)$, and*

$$B = \begin{pmatrix} (\widetilde{\boldsymbol{x}}_{t_1}^{*\top} \widetilde{\boldsymbol{x}}_{t_1}^*)^{-1} \widetilde{\boldsymbol{x}}_{t_1}^{*\top} K(t_1)^{1/2} \\ \cdots \\ (\widetilde{\boldsymbol{x}}_{t_n}^{*\top} \widetilde{\boldsymbol{x}}_{t_n}^*)^{-1} \widetilde{\boldsymbol{x}}_{t_n}^{*\top} K(t_n)^{1/2} \end{pmatrix}, \quad \boldsymbol{c} = \widetilde{\boldsymbol{x}}^* (\widetilde{\boldsymbol{x}}^{*\top} \widetilde{\boldsymbol{x}}^*)^{-1} \tag{3.42}$$

Detailed proof is provided in the appendix. In Theorem 5, though the calculation of the optimal weight vector can be computationally inefficient because of the inversion of the $n \times n$ matrix B, we can apply the kernel with bounded domain on $K_b(\cdot)$, e.g. the triangular kernel and then B becomes a banded matrix. The computing complexity of inversion of a banded matrix is $O(ln^2)$, where l is the width of band (Kavcic and Moura, 2000). Theorem 5 demonstrates that through the local estimators and our CD-based combining recipe, we can easily obtain the exact same result as the global estimator. We describe $\widehat{\beta}_c$ as follows:

$$\widehat{\beta}_c = \boldsymbol{c}^\top B^{-1} \widehat{\boldsymbol{\beta}}_{\text{local}} = \boldsymbol{w}^{*\top} \widehat{\boldsymbol{\beta}}_{\text{local}}, \tag{3.43}$$

where $w_i^* = \sum_{j=1}^{n} c_j b^{ji}$ with c_j being the jth element of $\boldsymbol{c}$ and b^{ji} being the jith element of B^{-1}. The combined estimator is a linear combination of local estimators and $\text{Var}(\widehat{\beta}_c) = \boldsymbol{c}^\top B^{-1} V (B^{-1})^\top \boldsymbol{c}$.

3.3.3 Numerical Studies

In this section, we illustrate the performance of our framework under partial linear regression models using a simulation study and a real data example. Both examples indicate that the proposed methodology can achieve the same performance as that of the global estimator.

Simulation Study

Firstly, we assume that the data is generated from the following partial linear regression model

$$y_i = 3x_i + z_{1i} - 2z_{2i} + \sin(t_i) + \varepsilon_i, i = 1, ..., n, \tag{3.44}$$

where $x_i = 0$ if i is odd, and $x_i = 1$ if i is even, $t_i = 5(i-1)/99$, z_{1i} and z_{2i}'s are independently generated from $\mathcal{N}(0,1)$ distribution, the standard deviation $\sigma = 0.1$, therefore ε_i's are i.i.d. random variables following $\mathcal{N}(0, 0.01)$. We define the kernel functions have the following forms, the Gaussian kernel $K_h(t) = h^{-1}\phi(t/h)$ where $h = 0.5$ and the rectangular kernel $K_b(t) = (2b)^{-1}\mathbb{I}_{t\in[-b,b]}$ where $b = 0.5$ and $\mathbb{I}$ is an indicator function.

In this simulation study, we perform 100 replications and estimate both the global the local estimators for different sample sizes, $n = 100, 200, 500$.. In Table 3.1, we compare global and combined estimators by taking mean and standard deviation of the absolute difference. The results indicate that, those two estimators are equivalent given the minimal difference. To be more concise, Figure 3.1 shows the one specific simulation's local estimators $\widehat{\beta}(t)$'s along with the global estimator marked by the red dotted line and true value as dashed line,

where $\widehat{\beta}_{\text{global}} = 3.003$ and $\widehat{\beta}_c = 3.003$. Figure 3.2 gives the boxplots for both global and combined estimators of 100 replications under different sample size settings.

	n=100	n=200	n=500
$\lvert\widehat{\beta}_{\text{global}} - \widehat{\beta}_c\rvert$	3.81e-14(1.03e-15)	7.86e-15(1.26e-15)	1.86e-14(2.08e-15)
$\lvert\beta_0 - \widehat{\beta}_c\rvert$	3.61e-3(1.84e-3)	1.33e-3(8.85e-4)	9.47e-4(7.62e-4)

Table 3.1: Mean and standard deviations of the mean absolute difference of global and combined estimators of β with different sample size, true value and combined estimators $n = 100, 200, 500$

.

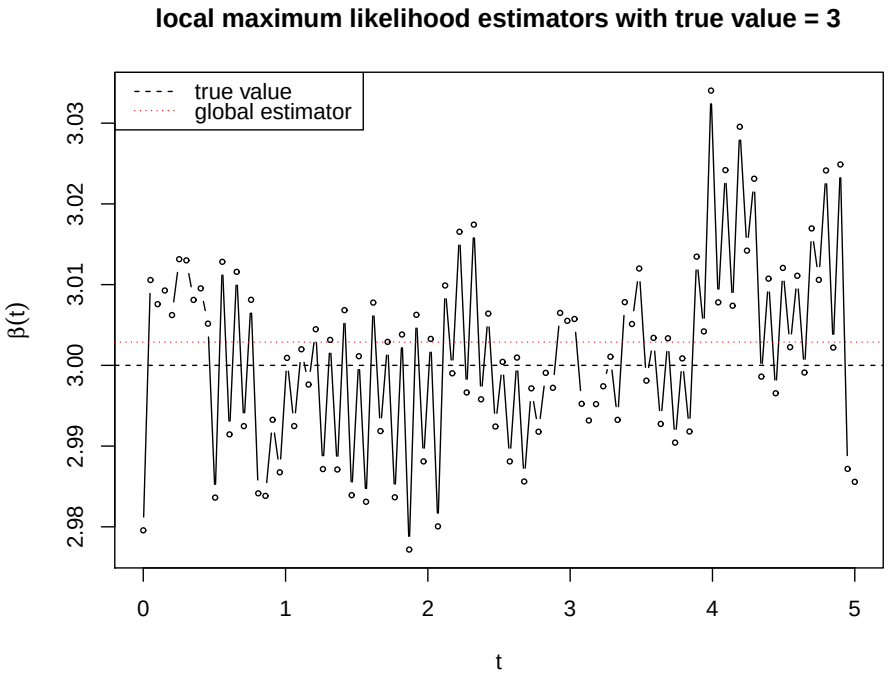

Figure 3.1: One single run of simulation's local estimators versus t_i with $n = 100$.

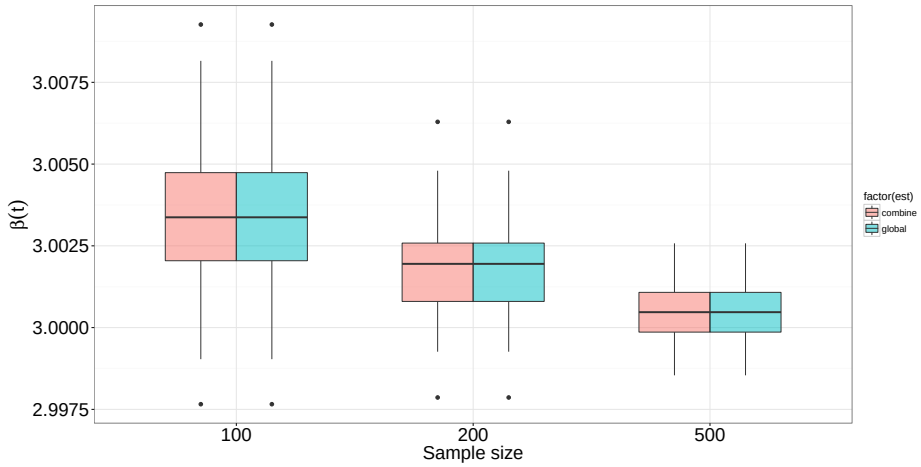

Figure 3.2: Boxplots for global and combined estimators of β based on 100 runs under the sample size settings $n = 100, 200, 500$.

Real Data Example

We now apply the proposed methodology to a real data example. The dataset is called onions data originally taken from Ratkowsky (1983). This dataset has been used in a multitude of previous studies, see Young and Bowman (1995), Bowman and Azzalini (1997) and Ruppert et al. (2003) etc. The onions dataset contains 84 observations from an experiment involving the production of white Spanish onions in two South Australian locations: Purnong Landing and Virginia, South Australia. The objective of this experiment is to establish the relationship between the yield of onions (g/plant) and the areal density (plants/m^2) and location. The partial linear model is built as follows,

$$y_i = \beta x_i + \eta(t_i) + \varepsilon_i, \tag{3.45}$$

where y_i is equal to $\log(\text{yield}_i)$; $x_i = 1$, if the ith observation is from Virginia, $x_i = 0$, if the ith observation is from Purnong Landing; t_i is the ith observation's areal density. Similar to the simulation study, we apply the Gaussian kernel for both non-parametric part and local estimates such that, $K_h(t) = h^{-1}\phi(t/h)$ with bandwidth $h = 4.5$ and $K_b(t) = b^{-1}\phi(t/b)$ with bandwidth $b = 4.5$. Therefore,

the global estimator of $\widehat{\beta}_{\text{global}} = -0.3286346$ and $\text{Var}(\widehat{\beta}_{\text{global}}) = 0.011$. Figure 3.3 presents the observations and fitted line based on the model setting. Figure 3.4 gives the local estimators and the corresponding 95% confidence interval. The combined estimator can be easily derived from the combined CD such that $\widehat{\beta}_c = -0.3286332$.

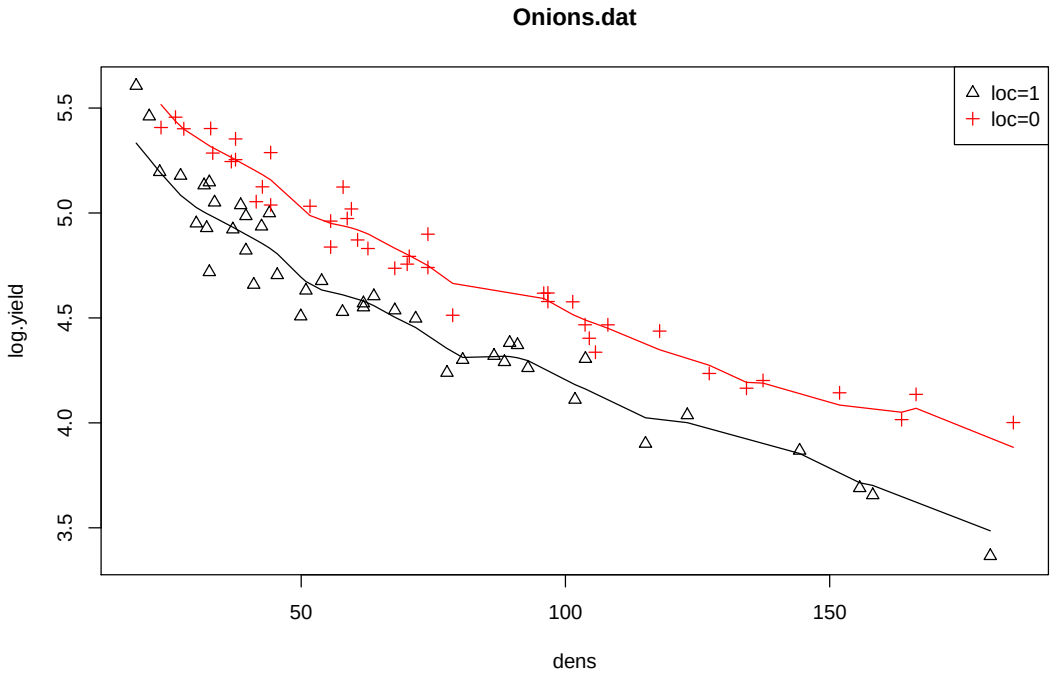

Figure 3.3: Onion dataset with fitted lines generated by global estimator

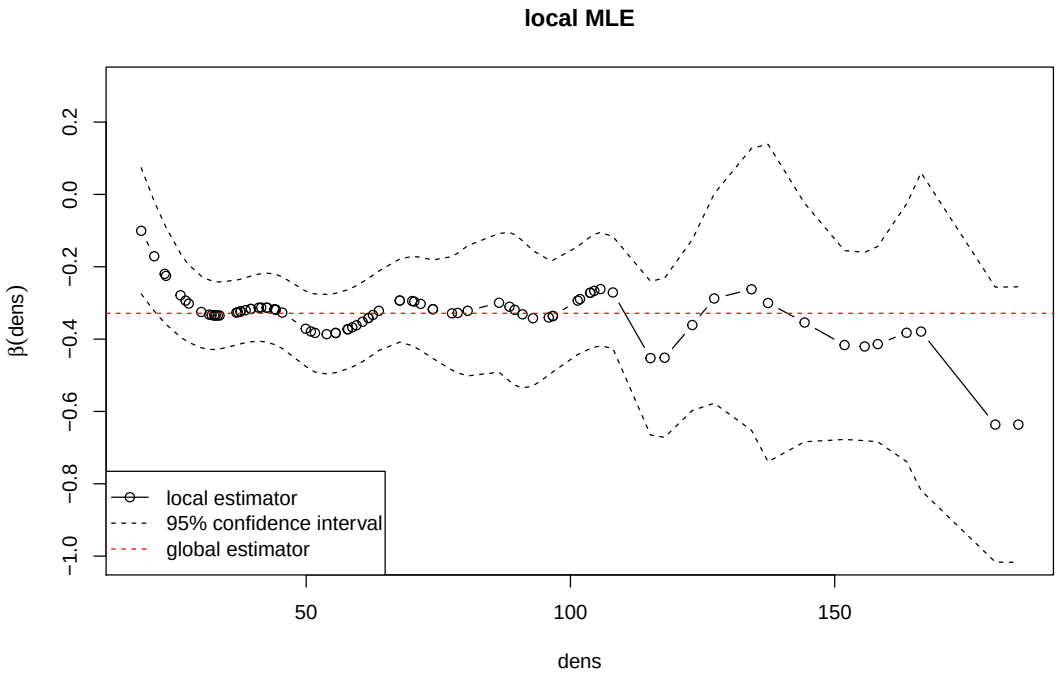

Figure 3.4: Local maximum likelihood estimators versus dens$_i$'s

3.4 Concluding Remarks

In this paper, we proposed a CD-based combining approach to effectively integrate information from multiple studies without assuming their independence. The underpinning to our general framework is the construction of the combined CD function from individual CDs that fully exploit the merits of copulas based on the Sklar's theorem. The copula method is widely used to model dependencies when the marginal distributions of a joint distribution are known. Although only Gaussian copula is considered in this work, there are many other copulas that can be applied in our general framework, for instance, the t-copula (Embrechts et al., 2001; Fang et al., 2002; Demarta and McNeil, 2005) and empirical copulas (Bouyé et al., 2000). As shown previously, with different choices of $g_c^*(\cdot)$ functions, there always exists a combined CD to fuse local estimates. Therefore, if the local/individual studies and their dependency information is given, the combined results can be obtained through the combining recipe. This enables the

flexibility of fusing local inference results by abstracting their dependencies onto the covariate structure imposed in various copulas, and hence circumvents the necessity of dealing with dependencies during local estimation. Most importantly, it is shown that the combined result from our framework is also globally optimal under various scenarios.

In this work, we first illustrated an example on combining dependent likelihood functions. There were more than one way to combine dependent log-likelihood functions by specifying $g_c^*(\cdot)$ with different weights. Second, we applied the proposed framework on combining the parametric components of partial linear models. Instead of performing inferences over the entire dataset, the global estimator could be also obtained by local estimators based on local pieces of information. Last but not least, this modular breakdown of inferences and fusion under our framework also potentially provides an elegant solution to decompose large-scale problems into small ones that can be later combined without loss of optimality.

3.5 Appendix A: Proofs

3.5.1 A.1 Proof of Theorem 4

Proof: For each set of fixed samples across K studies, since $g_c^*(\cdot)$ is monotonic (say increasing) on each coordinate, $H_c^*(\cdot) = G_c^*(\cdot)$ is always a cumulative distribution function. At the true value $\theta = \theta_0$, $H_c^*(\theta_0) = G_c^*\{g_c^*(H_1(\theta_0), ..., H_K(\theta_0))\}$. By the definition of $G_c^*(t)$, we have

$$\begin{aligned} Pr(H_c^*(\theta_0) \leq t) &= Pr(G_c^*\{g_c^*(H_1(\theta_0), ..., H_K(\theta_0))\} \leq t) \\ &= Pr(g_c^*(H_1(\theta_0), ..., H_K(\theta_0)) \leq (G_c^*)^{-1}(z)) \\ &= G_c^*((G_c^*)^{-1}(z)) = z \end{aligned}$$

Therefore, $H_c^*(\theta_0)$ is also U[0,1] distributed and $H_c^*(\theta)$ is a CD for θ. □

3.5.2 A.2 Proof of Corollary 2

Proof: In our case, we have the marginal U[0,1] distribution functions $F_i(z) = \mathbf{1}_{\{z\in[0,1]\}}t + \mathbf{1}_{\{z\in(1,\infty)\}}$. Then we can get the following result from Lemma 1:

$$C(z_1,...,z_K) = F(F_1^{-1}(z_1),...,F_K^{-1}(z_K)) = F(z_1,...,z_K), \text{ for } 0 < z_i < 1, i = 1,...,K \tag{3.46}$$

□

3.5.3 A.3 Proof of Theorem 5

Proof: Firstly, we write the global estimator $\widehat{\beta}_{\text{globl}}$ in terms of $\boldsymbol{x}, \boldsymbol{y}, K(\cdot)$ and U as follows,

$$\widehat{\beta}_{\text{global}} = (\widetilde{\boldsymbol{x}}^{*\top}\widetilde{\boldsymbol{x}}^{*})^{-1}\widetilde{\boldsymbol{x}}^{*\top}\widetilde{\boldsymbol{y}} \tag{3.47}$$

$$= (\boldsymbol{x}(I-U)^{\top}(I-P_Z)^{\top}(I-P_Z)(I-U)\boldsymbol{x})^{-1}\boldsymbol{x}(I-U)^{\top}(I-P_Z)^{\top}(I-U)\boldsymbol{y} \tag{3.48}$$

$$= (\boldsymbol{x}(I-U)^{\top}(I-P_Z)(I-U)\boldsymbol{x})^{-1}\boldsymbol{x}(I-U)^{\top}(I-P_Z)(I-U)\boldsymbol{y} \tag{3.49}$$

$$= \boldsymbol{c}^{\top}(I-U)\boldsymbol{y} \tag{3.50}$$

When $\boldsymbol{w} = \boldsymbol{c}^{\top}B^{-1}S$, the combined estimator equals to

$$\widehat{\beta}_c = (\boldsymbol{w}^{\top}S^{-1}\mathbf{1})^{-1}\boldsymbol{w}^{\top}S^{-1}\widehat{\boldsymbol{\beta}}_{\text{local}} \tag{3.51}$$

$$= (\boldsymbol{c}^{\top}B^{-1}\mathbf{1})^{-1}\boldsymbol{c}^{\top}B^{-1}\widehat{\boldsymbol{\beta}}_{\text{local}} \tag{3.52}$$

$$= (\boldsymbol{c}^{\top}B^{-1}\mathbf{1})^{-1}\boldsymbol{c}^{\top}B^{-1}B(I-U)\boldsymbol{y} \tag{3.53}$$

$$= (\boldsymbol{c}^{\top}B^{-1}\mathbf{1})^{-1}\boldsymbol{c}^{\top}(I-U)\boldsymbol{y} \tag{3.54}$$

Since we have

$$B\widetilde{\boldsymbol{x}} = \begin{pmatrix} (\widetilde{\boldsymbol{x}}_{t_1}^{*\top}\widetilde{\boldsymbol{x}}_{t_1}^{*})^{-1}\widetilde{\boldsymbol{x}}_{t_1}^{*\top}K(t_1)^{1/2} \\ \vdots \\ (\widetilde{\boldsymbol{x}}_{t_n}^{*\top}\widetilde{\boldsymbol{x}}_{t_n}^{*})^{-1}\widetilde{\boldsymbol{x}}_{t_n}^{*\top}K(t_n)^{1/2} \end{pmatrix}\widetilde{\boldsymbol{x}} \tag{3.55}$$

$$= \begin{pmatrix} (\widetilde{\boldsymbol{x}}K(t_1)^{1/2}(I-P_{Z_{t_1}})^\top(I-P_{Z_{t_1}})K(t_1)^{1/2}\widetilde{\boldsymbol{x}})^{-1}\widetilde{\boldsymbol{x}}K(t_1)^{1/2}(I-P_{Z_{t_1}})^\top K(t_1)^{1/2} \\ \vdots \\ (\widetilde{\boldsymbol{x}}K(t_n)^{1/2}(I-P_{Z_{t_n}})^\top(I-P_{Z_{t_n}})K(t_n)^{1/2}\widetilde{\boldsymbol{x}})^{-1}\widetilde{\boldsymbol{x}}K(t_n)^{1/2}(I-P_{Z_{t_n}})^\top K(t_n)^{1/2} \end{pmatrix}\widetilde{\boldsymbol{x}} \tag{3.56}$$

$$= \begin{pmatrix} (\widetilde{\boldsymbol{x}}K(t_1)^{1/2}(I-P_{Z_{t_1}})K(t_1)^{1/2}\widetilde{\boldsymbol{x}})^{-1}\widetilde{\boldsymbol{x}}K(t_1)^{1/2}(I-P_{Z_{t_1}})K(t_1)^{1/2} \\ \vdots \\ (\widetilde{\boldsymbol{x}}K(t_n)^{1/2}(I-P_{Z_{t_n}})K(t_n)^{1/2}\widetilde{\boldsymbol{x}})^{-1}\widetilde{\boldsymbol{x}}K(t_n)^{1/2}(I-P_{Z_{t_n}})K(t_n)^{1/2} \end{pmatrix}\widetilde{\boldsymbol{x}} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \tag{3.57}$$

$$\boldsymbol{c}^\top\widetilde{\boldsymbol{x}} = (\boldsymbol{x}(I-U)^\top(I-P_Z)(I-U)\boldsymbol{x})^{-1}\boldsymbol{x}(I-U)^\top(I-P_Z)\widetilde{\boldsymbol{x}} \tag{3.58}$$

$$= (\boldsymbol{x}(I-U)^\top(I-P_Z)(I-U)\boldsymbol{x})^{-1}\boldsymbol{x}(I-U)^\top(I-P_Z)(I-U)\boldsymbol{x} = 1 \tag{3.59}$$

Therefore, $\boldsymbol{c}^\top B^{-1}\mathbf{1} = \boldsymbol{c}^\top B^{-1}B\widetilde{\boldsymbol{x}} = \boldsymbol{c}^\top\widetilde{\boldsymbol{x}} = 1$, which indicates that $\widehat{\beta}_c = (\boldsymbol{c}^\top B^{-1}\mathbf{1})^{-1}\boldsymbol{c}^\top(I-U)\boldsymbol{y} = \boldsymbol{c}^\top(I-U)\boldsymbol{y} = \widehat{\beta}_{\text{global}}$. □

3.6 Appendix B: Combining Dependent Likelihood in Composite Likelihood Context

Suppose that $\ell_1(\boldsymbol{\theta}), ..., \ell_K(\boldsymbol{\theta})$ are K dependent log-likelihood functions, where the parameters of interests $\{\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^p\}$. The score function for the ith log-likelihood function is given as $s_i(\boldsymbol{\theta}) = \partial\ell_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$. Based on the Taylor expansion of the score function around the true value $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, we have

$$s(\boldsymbol{\theta}) \approx s(\boldsymbol{\theta}_0) + s'(\boldsymbol{\theta}_0)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) = s(\boldsymbol{\theta}_0) + \left.\frac{\partial^2\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}(\boldsymbol{\theta}-\boldsymbol{\theta}_0). \tag{3.60}$$

The expectation of the score function can be easily derived as

$$E(s(\boldsymbol{\theta})) = E(s'(\boldsymbol{\theta}_0))\boldsymbol{\theta} - E(s'(\boldsymbol{\theta}_0))\boldsymbol{\theta}_0. \tag{3.61}$$

Let $\boldsymbol{s}^0 = \boldsymbol{s}(0)$, where $\boldsymbol{s}(\boldsymbol{\theta}) = (s_1^\top(\boldsymbol{\theta}), \ldots, s_K^\top(\boldsymbol{\theta}))^\top$ and $E(s_i(0)) = -E(s'(\boldsymbol{\theta}_0))\boldsymbol{\theta}_0$. The score function vector $\boldsymbol{s}^0$ then follows an pK-dimensional asymptotic multivariate normal distribution,

$$\boldsymbol{s}^0 \sim \mathcal{N}_{pK}(V\boldsymbol{\theta}, W), \tag{3.62}$$

where $V = (V_{11}^\top, \ldots, V_{KK}^\top)^\top$ is a $pK \times p$ dimensional matrix stacked by K matrices of dimension $p \times p$, such that $V_{ii} = -E(s_i'(\boldsymbol{\theta}_0))$, the covariance matrix W is a $pK \times pK$ matrix with the ith block diagonal matrix equals to V_{ii} and the ijth block matrix is assumed to be known as $V_{ij} = \mathrm{Cov}(s_i(\boldsymbol{\theta}_0), s_j(\boldsymbol{\theta}_0))$. Give the joint distribution (3.62), the new MLE of θ is derived by $\widehat{\boldsymbol{\theta}}^0 = (V^\top W^{-1} V)^{-1} V^\top W^{-1} \boldsymbol{s}^0$ and $Var(\widehat{\boldsymbol{\theta}}^0) = (V^\top W^{-1} V)^{-1}$. Finally, the new log-likelihood function constructed by Fraser and Reid (2015) is written as in the equivalent form,

$$\ell_{\text{new}}(\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^0)^\top (V^\top W^{-1} V)(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}^0). \tag{3.63}$$

More specifically, when $p = 1$, i.e. the parameter of interest is a scalar, say θ, the new log-likelihood function can be rewritten as a linear combination of the K individual ones,

$$\ell_{\text{new}}(\theta) = V^\top W^{-1} \boldsymbol{s}^0 \theta = V^\top W^{-1} (\ell_1(\theta), \ldots, \ell_K(\theta))^\top. \tag{3.64}$$

Chapter 4
Concluding Remarks

In this dissertation, we focus on developing methodologies for combining information from dependent studies. The developments are both based on combination of CDs, which gains popularity in fusing inferences form different studies. We have shown that the proposed framework has desirable properties in both making statistical inferences and solving computing issues.

More concretely, in Chapter 2, we propose a sequential split-conquer-combine (SSCC) approach to analyze big spatial data under a Gaussian process model setting. The unified framework consists of a sequential split-conquer procedure, information combining technique using CDs, and a CD-based predictive distribution. Under mild assumptions, the combined estimators and predictors are shown to be asymptotically equivalent to the ones derived by using the entire dataset, while the computing time is significantly reduced. As a byproduct, we also introduced a Monte-Carlo algorithm to construct the CD-based predictive distribution which provides rich information for statistical inference and a better quantification of prediction uncertainty comparing with the plug-in approach. In Chapter 3, we propose a general combining approach based on individual univariate CDs to integrate information from multiple studies without assuming their independence. The combined results can be obtained through the combining recipe, if the individual studies and their dependency information is known. This enables the flexibility of fusing local inference results by abstracting their dependencies onto the covariate structure imposed in various copulas, and hence circumvents the necessity of dealing with dependencies during

local estimation. Most importantly, it is shown that the combined result from our framework is also globally optimal under various scenarios. We illustrated an example on combining the parametric components of partial linear models. Instead of performing inferences over the entire dataset, the combined estimator, which is shown to be equivalent to the global one, could be obtained by local estimators based on local pieces of information. Last but not least, this modular breakdown of inferences and fusion under our framework also potentially provides an elegant solution to decompose large-scale problems into small ones that can be later combined without loss of optimality.

Bibliography

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.

Barry, R. P. and Pace, R. K. (1999), "Monte Carlo estimates of the log determinant of large sparse matrices," *Linear Algebra and its applications*, 289, 41–54.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2015), "Distributed estimation and inference with statistical guarantees," *arXiv preprint arXiv:1509.05457*.

Bickel, P. and Levina, E. (2008), "Regularized estimation of large covariance matrices," *Annals of Statistics*, 36, 199–227.

Boente, G., He, X., and Zhou, J. (2006), "Robust estimates in generalized partially linear models," *The Annals of Statistics*, 2856–2878.

Bouyé, E., Durrleman, V., Nikeghbali, A., Riboulet, G., and Roncalli, T. (2000), "Copulas for finance-a reading guide and some applications," .

Bowman, A. W. and Azzalini, A. (1997), *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, vol. 18, OUP Oxford.

Cai, T. T. and Zhou, H. H. (2012), "Optimal rates of convergence for sparse covariance matrix estimation," *Annals of Statistics*, 40, 2389–2420.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized partially

linear single-index models," *Journal of the American Statistical Association*, 92, 477–489.

Chang, W., Haran, M., Olson, R., Keller, K., et al. (2014), "Fast dimension-reduced climate model calibration and the effect of data aggregation," *The Annals of Applied Statistics*, 8, 649–673.

Chen, X. and Xie, M. (2014a), "A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data," *Statistica Sinica*, 24, 1655–1684.

Chen, X. and Xie, M.-g. (2014b), "A split-and-conquer approach for analysis of extraordinarily large data," *Statistica Sinica*, 1655–1684.

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

Claggett, B., Xie, M., and Tian, L. (2014), "Meta-analysis with fixed, unknown, study-specific parameters," *Journal of the American Statistical Association*, 109, 1660–1671.

Cox, D. R. (2013), "Discussion of "Confidence distribution, the frequentist distribution estimator of a parameter"," *International Statistical Review*, 81, 40–41.

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.

Demarta, S. and McNeil, A. J. (2005), "The t copula and related copulas," *International Statistical Review/Revue Internationale de Statistique*, 111–129.

Dinse, G. E. and Lagakos, S. (1983), "Regression analysis of tumour prevalence data," *Applied statistics*, 236–248.

Embrechts, P., Lindskog, F., and McNeil, A. (2001), "Modelling dependence with copulas," *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

Engle, R. F., Granger, C. W., Rice, J., and Weiss, A. (1986), "Semiparametric estimates of the relation between weather and electricity sales," *Journal of the American statistical Association*, 81, 310–320.

Fan, J., Farmen, M., and Gijbels, I. (1998), "Local maximum likelihood estimation and inference," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 591–608.

Fan, J., Gijbels, I., King, M., et al. (1997), "Local likelihood and local partial likelihood in hazard regression," *The Annals of Statistics*, 25, 1661–1690.

Fang, H. (2011), "Stability Analysis of Block LDL^T Factorizations for Symmetric Indefinite Matrices," *IMA J Numer Anal*, 528–555.

Fang, H.-B., Fang, K.-T., and Kotz, S. (2002), "The meta-elliptical distributions with given marginals," *Journal of Multivariate Analysis*, 82, 1–16.

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

Fraser, D. A. S. and Reid, N. (2015), "Combining log-likelihood or significance functions: Higher accuracy for composite likelihood," .

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.

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

Golub, G. H. and Van Loan, C. F. (1983), *Matrix Computations*, Johns Hopkins University Press.

Gramacy, R. B. and Apley, D. W. (2015), "Local Gaussian process approximation for large computer experiments," *Journal of Computational and Graphical Statistics*, 24, 561–578.

Gramacy, R. B. and Lee, H. K. (2008), "Gaussian processes and limiting linear models," *Computational Statistics & Data Analysis*, 53, 123–136.

Green, P. and Silverman, B. (1994), "Nonparametric regression and generalized linear models. Number 58 in Monographs on Statistics and Applied Probability," .

Green, P. J. and Yandell, B. S. (1985), "Semi-parametric generalized linear models," in *Generalized linear models*, Springer, pp. 44–55.

Härdle, W., Liang, H., and Gao, J. (2012), *Partially linear models*, Springer Science & Business Media.

Hedges, L. V. and Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, New York: Academic Press.

Hjort, N. L. and Glad, I. K. (1995), "Nonparametric density estimation with a parametric start," *The Annals of Statistics*, 882–904.

Hjort, N. L. and Jones, M. (1996), "Locally parametric nonparametric density estimation," *The Annals of Statistics*, 1619–1647.

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.

Kaufman, C., Bingham, D., Habib, S., Heitmann, K., and Frieman, J. A. (2011), "Efficient Emulators of Computer Experiments using Compactly Supported Correlation Functions, with an Application to Cosmology," *Annals of Applied Statistics*, 5, 2470–2492.

Kaufman, C., Schervish, M., and Nychka, D. (2008), "Covariance Tapering for Likelihood-based Estimation in Large Spatial Datasets." *J. Amer. Statist. Assoc.*, 103, 1545–1555.

Kavcic, A. and Moura, J. M. (2000), "Matrices with banded inverses: Inversion algorithms and factorization of Gauss-Markov processes," *IEEE transactions on Information Theory*, 46, 1495–1509.

Kennedy, M. C. and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models," *Journal of the Royal Statistical Society*, 63, 425–464.

Li, R. and Sudjianto, A. (2005), "Analysis of Computer Experiments Using Penalized Likelihood," *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, D., Liu, R., and Xie, M. (2015), "Multivariate Meta-Analysis of Heterogeneous Studies Using Only Summary Statistics: Efficiency and Robustness." *J. Amer. Statist. Assoc.*, 110, 326–340.

Loader, C. R. et al. (1996), "Local likelihood density estimation," *The Annals of Statistics*, 24, 1602–1618.

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

Mackey, L., Talwalkar, A., and Jordan, M. I. (2015), "Distributed matrix completion and robust factorization," *Journal of Machine Learning Research*, 16, 913–960.

Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015), "A multiresolution Gaussian process model for the analysis of large spatial datasets," *Journal of Computational and Graphical Statistics*, 24, 579–599.

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.

Pissanetzky, S. (1984), *Sparse Matrix Technology-electronic edition*, Academic Press.

Rao, C. R. (2009), *Linear statistical inference and its applications*, vol. 22, John Wiley & Sons.

Ratkowsky, D. A. (1983), *Nonlinear regression modelling: A unified practical approach*, Dekker.

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

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

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric regression*, no. 12, Cambridge university press.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn., H. P. (1989), "Design and Analysis of Computer Experiments," *Statistical Scicence*, 4, 409–423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, Springer.

Schmalensee, R. and Stoker, T. M. (1999), "Household gasoline demand in the United States," *Econometrica*, 67, 645–662.

Schmidt, A. M. and O'Hagan, A. (2003), "Bayesian inference for non-stationary spatial covariance structure via spatial deformations," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65, 743–758.

Schulze, R. (2004), *Meta-analysis-A comparison of approaches*, Hogrefe Publishing.

Schweder, T. and Hjort, N. (2016), *Confidence, Likelihood and Probability*, Cambridge, U.K.: Cambridge University Press.

Severini, T. A. and Staniswalis, J. G. (1994), "Quasi-likelihood estimation in semiparametric models," *Journal of the American statistical Association*, 89, 501–511.

Shen, J., Liu, R., and Xie, M. (2016), "Prediction with confidence - a unifying framework for prediction," *Journal of Statistical Planing and Inference (under revision)*.

Singh, K., Xie, M., and Strawderman, W. E. (2007), "Confidence distribution (CD): distribution estimator of a parameter," *Complex Datasets and Inverse Problems: Tomography, Networks and Beyond. IMS Lecture Notes-Monograph Series*, 45, 132–150.

Singh, K., Xie, M., Strawderman, W. E., et al. (2005), "Combining information from independent sources through confidence distributions," *The Annals of Statistics*, 33, 159–183.

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

Sklar, A. (1973), "Random variables, joint distribution functions, and copulas," *Kybernetika*, 9, 449–460.

Smola, A. J., Bartlett, P., et al. (2001), "Sparse greedy Gaussian process regression," *Advances in neural information processing systems*, 13, 619–625.

Snelson, E. and Ghahramani, Z. (2005), "Sparse Gaussian processes using pseudo-inputs," in *Advances in neural information processing systems*, pp. 1257–1264.

Speckman, P. (1988), "Kernel smoothing in partial linear models," *Journal of the Royal Statistical Society. Series B (Methodological)*, 413–436.

Stangl, D. and Berry, D. A. (2000), *Meta-analysis in medicine and health policy*, CRC Press.

Staniswalis, J. G. (1989), "The kernel estimate of a regression function in likelihood-based models," *Journal of the American Statistical Association*, 84, 276–283.

Stein, M. L. (2008), "A modeling approach for large spatial datasets," *J. Korean Statist. Soc.*, 37, 3–10.

— (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.

Tibshirani, R. and Hastie, T. (1987), "Local likelihood estimation," *Journal of the American Statistical Association*, 82, 559–567.

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

Xie, M. and Singh, K. (2013), "Confidence distribution, the frequentist distribution estimator of a parameter (with discussions)," *International Statistical Review*, 81, 3–39.

Xie, M., Singh, K., and Strawderman, W. E. (2011), "Confidence Distributions and a Unifying Framework for Meta-analysis." *J. Amer. Statist. Assoc.*, 106(493), 320–333.

Yang, G., Liu, D., Liu, R. Y., Xie, M., and Hoaglin, D. (2014), "A confidence distribution approach for an efficient network meta-analysis," *Statistical Methodology*, 20, 105–125.

Young, S. G. and Bowman, A. W. (1995), "Non-parametric analysis of covariance," *Biometrics*, 920–931.

Zhang, Y., Duchi, J., and Wainwright, M. (2015), "Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates," *J. Mach. Learn. Res*, 16, 3299–3340.