# ADVANCES IN CONFIDENCE DISTRIBUTION: INDIVIDUALIZED FUSION LEARNING AND PREDICTIVE DISTRIBUTION FUNCTION

By

JIELI SHEN

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 Regina Y. Liu

and approved by

______________________________

______________________________

______________________________

______________________________

New Brunswick, New Jersey

October, 2017

## ABSTRACT OF THE DISSERTATION

# ADVANCES IN CONFIDENCE DISTRIBUTION: INDIVIDUALIZED FUSION LEARNING AND PREDICTIVE DISTRIBUTION FUNCTION

**By JIELI SHEN**

**Dissertation Director:**

**Minge Xie and Regina Y. Liu**

In this dissertation, we develop new methods for problems for the two fundamental topics of statistical learning - inference and prediction, using the tool of confidence distribution (CD). Specifically, we are interested in i) making efficient and valid statistical inference about an individual subject, by borrowing information from other individual subjects with similar traits, in a heterogeneous database that contains many individual subjects, and ii) effectively and accurately quantifying uncertainties associated with the prediction of future observations from a model estimated based on past observations.

For the first problem, we propose an individualized fusion learning ($i$Fusion) approach, for drawing efficient individualized inference by fusing information from relevant data sources. $i$Fusion is robust for handling heterogeneity arising from diverse sources, and is ideally suited for goal-directed applications such as precision medicine. Specifically, $i$Fusion summarizes individual inferences as CDs, then adaptively forms a clique of individuals that bears relevance to the target individual, and finally combines the CDs from those relevant individuals and draws inference for the target individual based on it. In essence, $i$Fusion "borrows strength" from relevant individuals to improve inference efficiency while retaining inference validity. Computationally, it is parallel in nature and scales up well in comparison with its competitors such as many of the Bayesian methods. Examples in simulations and a real application in financial forecasting are further presented to demonstrate the effectiveness of *i*Fusion.

For the second problem, a general prediction framework is proposed in which prediction is presented in the form of a predictive distribution function. This predictive distribution function is well suited for the notion of confidence subscribed in the frequentist interpretation, and can provide meaningful answers for questions related to prediction. A general approach under this framework is formulated and illustrated by using the concept of CD. This CD-based prediction approach inherits many desirable properties of CD, including its capacity for serving as a common platform for connecting and unifying the existing procedures of predictive inference in Bayesian, fiducial and frequentist paradigms. The theory underlying the CD-based predictive distribution is developed and some related efficiency and optimality issues are addressed. Moreover, a simple yet broadly applicable Monte-Carlo algorithm is proposed for the implementation of the proposed approach. This concrete algorithm together with the proposed definition and associate theoretical development produces a comprehensive statistical inference framework for prediction. Finally, the approach is applied to simulation studies, and a real project on predicting the incoming volume of application submissions to a government agency. The latter shows the applicability of the proposed approach to dependent data settings.

# Acknowledgements

I can never be too grateful to my thesis advisor Professor Minge Xie, a tremendous mentor and trusted friend of mine. The dissertation could not have been accomplished without his support and encouragement. Not only did he open the door to the statistics world for me, but he also taught me how to be a scientific researcher in it. Thank him for continuous and wise advice on my research, career, and life. The moment we spent together exploring this exciting world of statistics will be cherished. It is my honor to work with him. Next, I would like to extend my deepest gratitude to my co-advisor Professor Regina Liu. Her broad knowledge and insight in this area has been a valuable asset to my research as well as career development. Thanks to all of my committee members, including Professor Shou-En Lu and Professor Sijian Wang, for making my defense an enjoyable moment, and for their brilliant comments and suggestions.

I would like take this opportunity to thank the Department of Statistics and Biostatistics of Rutgers University for providing financial support and thank all its faculty and staff for creating such a fantastic research and work environment. Thanks to Professor Rong Chen, Professor Regina Liu, Professor Minge Xie, and Professor Cunhui Zhang for bringing to the collaborative project with Dun & Bradstreet that has partially motivated my research. I also benefit a lot from the insightful suggestions from Professor Zhiqiang Tan. Thanks to our graduate director Professor Harry Crane and Professor Tirthankar Dasgupta for coordinating my defense, and also to Professor John Kolassa, our former graduate director, who is always ready and eager to help year after year during his service. Special thanks to my wonderful friends Linglin He, Xinyu Sun, Liang Wang, and Yilei Zhan, who make me feel so young every day. Also thank my other friends and colleagues at Rutgers including Yi Fan, Long Feng, and Chengrui Li for their company and support. There are a long list of friends; I could not enumerate

them all, but no doubt they are the best ones.

My last thanks go to my parents, my father Huoming Shen, and my mother Yinlian Wang. There aren't enough words or ways to express how grateful I am to you even just for bringing me to the world, not to mention their unconditional and persistent love.

I would like to close my acknowledgements with a quote from a celebrated Chinese elder: "One's fate is dictated by the extent of efforts, and, for that matter, history." This is the best of the time in history, and I will spare no effort in my new journey.

# Dedication

To my parents

&

To a better world.

# Table of Contents

# List of Tables

# List of Figures

# Chapter 1

# Introduction

Inference and prediction are the two primary goals of statistical learning. The research in this dissertation is devoted to addressing two important problems in statistical inference and prediction:

- How to make valid, effective, and efficient statistical inference for each individual subject, by borrowing information from other individual subjects with similar traits, in a heterogeneous database that contains many individual subjects?
- How to effectively and accurately quantify the uncertainties associated with prediction of future observations from a model estimated based on past observations?

The first problem arises in the field of fusion learning, which refers to a collection of methods of synthesizing information from multiple data sources for more powerful finds than those from individual data source alone. Effective fusion learning is of vital importance, especially in light of the automated data acquisition nowadays in many domains. Decision-making processes in many domains such as medicine, life science, and social studies benefit greatly from considering data from different sources. The key challenges in effective fusion learning often stem from massive complex structures and heterogeneity among different data sources. Ignoring such complexity and heterogeneity when making statistical inference may lead to insufficient and even misleading conclusions, especially when the inference goal is a particular individual subject rather than the "average population".

In the first part of this dissertation, we develop a method called individualized fusion learning ($i$Fusion), for drawing efficient individualized inference. $i$Fusion is robust for handling heterogeneity arising from diverse sources, and is ideally suited for goal-directed applications such as precision medicine. Specifically, $i$Fusion first makes inference independently using each individual data source and summarizes them into confidence distributions (CDs); then, it adaptively forms a clique of individuals that bears relevance to the target individual; and finally, it combines the CDs from those relevant individuals. The procedure is simple, yet is effective and efficient. Drawing inference based on $i$Fusion "borrows strength" from other individual data with similar traits, thereby improving inference efficiency while preserving inference validity since it borrows "smartly from only relevant individuals. The approach also enjoys many other nice features such as scalability to big data, adaptability to various configurations of the underlying individual parameter values (essentially without any "parametric" assumption on them), and intrinsic bias-variance tradeoff interpretation, among others, which will be elaborated in the next chapter.

The second problem is in the area of predictive inference. Existing statistical methods for predictive inference in the literature fall into two main categories - Bayesian and frequentist. In the Bayesian category, Bayesian predictive distribution, a distribution function of a future observation integrated over the posterior distribution of the unknown parameter, serves as the main tool. It enjoys the flexibility of distribution functions but does depend on the additional assumptions of priors and usually "does not have clear probability interpretations in finite samples." (Lawless and Fredette, 2005). In the frequentist category, prediction intervals, analogous to that of confidence intervals, are widely used, with a precise and well defined frequentist probabilistic interpretation. But those prediction intervals use only two endpoints of the intervals to describe a future observation, and thus are not as informative or flexible as an entire predictive distribution produced by the Bayesian methods, among others.

In the second part of this dissertation, we extend the concept of prediction intervals to a more general form of predictive distribution functions. It is well suited for the notion of confidence subscribed in the frequentist interpretation, as opposed to the Bayesian predictive distribution. Under this framework, we further propose a general approach to construct predictive distribution functions using CDs. The theory underlying the CD-based predictive distribution is developed and some related efficiency and optimality issues are addressed. Moreover, a simple yet broadly applicable Monte-Carlo algorithm is proposed for implementing the proposed approach. This concrete algorithm together with the proposed definition and associate theoretical development produces a comprehensive statistical inference framework for prediction.

Although the two main parts of this dissertation are self-contained, they are bridged by the common development tool - confidence distribution (CD). A CD can be viewed as a sample-dependent distribution that represents confidence intervals of all levels for a parameter of interest (Cox, 1958; Efron, 1993), and provides "simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)." (Cox, 2013) The two topics covered in this dissertation are yet additional illustrations of how powerful the concept of CD is in many fields of modern statistics.

The rest of this dissertation is organized as follows. Chapter 2 develops the $i$Fusion approach for drawing individualized inference. Chapter 3 presents the CD-based predictive distribution for making predictive inference. Each of the two chapters contains background review and motivation, detailed description of the methodologies along with computing algorithms, and theoretical developments which are further demonstrated by simulation studies under various settings and real-world applications. Chapter 4 concludes the dissertation. Technical proofs are relegated to Appendix.

# Chapter 2

# $i$Fusion: Individualized Fusion Learning

## 2.1 Background and Motivation

Fusion learning refers to synthesizing statistical inferences from multiple data sources to yield more powerful findings than those from individual subjects or sources alone. This is highly sought after, especially in light of the data explosion phenomena in many domains nowadays. The key challenges in effective fusion learning often stem from massive complex structures and heterogeneity among different data sources. Specifically, if the goal is to make inference for a particular individual subject, analysis by simply pooling the data in a database can be insufficient or even misleading, as not all the individuals are relevant. On the other hand, inference based on individual data source alone may be quite inefficient due to loss of information, as some other individuals may be helpful. This chapter presents a new fusion learning approach called individualized fusion learning, abbreviated as $i$Fusion, to effectively merge information from relevant data sources and draw efficient inference for any target individual subject.

This research is initially motivated by a collaborative project with a global consulting firm that provides risk management services for small business worldwide. One of the main objectives of the project is to build forecasting models for each of over 100000 companies using monthly time series data of each company, in conjunction with relevant economic and market indices. To reflect the current status of the companies, only data in past two or three years are used in the analysis. Traditionally, the analysis is carried out by building a model, e.g., an autoregressive model with exogenous variables, for each company using its own data. However, such individual models can be quite unstable due to the small sample size of each company, as reported by their in-house data scientists. With the availability of the large databased of over 100000 companies, there may exist a group of companies that have similar traits to the target company, and the information within the group can be shared (even if only partially) and used to improve the analysis of the target company.

One conventional approach is to assign the companies into subgroups by using an unsupervised clustering method, either directly on the feature space (e.g., company-specific features), or sometimes on the parameter space (after estimating a supervised learning model), and then pool the data in the same subgroup for further analysis. The approach, though leads to increased sample sizes in each subgroup, has some obvious shortcomings. For example, the formation of subgroups can be quite arbitrary as it depends on the number of subgroups specified in the approach, a parameter that is difficult to determine, the type of clustering method used and the metric used for measuring the similarity between companies. Furthermore, all analytical outcome and inference (e.g., estimated parameters, testing) are identical to all individuals in the same subgroup. More importantly, in many situations, there may not be any clear-cut and well divided subgroup structure in the population. In these situations, the conventional subgroup analysis imposes an artificial grouping structure to the population and the analysis often leads to large biases and thus invalid inference in many cases.

Bayesian hierarchical models may also be used to tackle the problem, where each company's model is conditioning on company-specific parameters and the parameters are modeled through a prior distribution. The resulting posterior distributions can then be used to make inference on the individual company-specific parameters. See, e.g., Gelman et al. (2013) and Gustafson et al. (2005), for reviews of Bayesian hierarchical models and their applications. In order to capture the complexity of between-company heterogeneity, a simple prior like Gaussian prior may not be sufficient, but with the help of Monte Carlo Markov Chain (MCMC) techniques, it is possible to consider more complicated models and priors such as finite mixture models. However, the assumption would encounter the same problems of determining the number of clusters (mixtures) and, in the case when the population has no clear-cut and well-divided subgroup structure, artificially imposing a mixture structure to the population. A nonparametric Bayesian approach using, for instance, a Dirichlet process mixture prior in conjugation with generalized linear models (cf. Grün and Leisch 2007 and Hannah et al. 2011) may help mitigate these concerns. But as a Bayesian hierarchical model, it often needs to rely on an iterative MCMC sampling scheme and analyze all companies all together. Considering the large number of companies, this approach can be computationally very intensive and even difficulty to be carried out.

In this chapter, based on recent developments on CD (a brief review of CD is provided in Section 2.2.1), we propose an $i$Fusion approach to draw individualized inference for each individual. $i$Fusion first analyzes the data from each individual company separately and summarizes the inference information into a confidence density function for each company. Then, these individual confidence density functions are fused with respect to a target company, according to a set of target-individual-specific adaptive screening weights (see details in Section 2.2.2), and inference of the target company can be drawn based on it. Like the Bayesian methods, $i$Fusion improves inference efficiency on the target company by "borrowing strength" from other companies/individual data. Unlike the Bayesian methods: i) Inference validity in terms of frequentist properties is guaranteed through the screening weights that ensures only information among relevant individuals is shared. ii) The general framework is suitable and can be adapted for any configurations of the underlying individual parameter values, so it is essentially "nonparametric" and no prior model assumption is needed for the underlying true parameters. iii) $i$Fusion naturally fits into the "divide-and-conquer" scheme and can be scaled up to big data applications like the motivating example with a large number of companies. Specifically, the first step of $i$Fusion separately analyzes individual companies and can be performed without accessing the entire dataset, thus easily allowing distributed implementation and making $i$Fusion a very attractive alternative to the Bayesian methods in big data applications.

The $i$Fusion approach is an adaptive local grouping approach designated for each target individual, and there is an intrinsic bias-variance tradeoff underlying the development. Indeed, as no two individuals (companies) and their performances are exactly the same, introducing other individual data in the analysis will inevitably create bias. But, if the biases are can be ignored or is very small, then the increased size of the data in the $i$Fusion inference can reduce variability (measured by estimation/prediction variance) and thus improve the overall efficiency (smaller mean squared errors) of inference. On the other hand, to ensure the validity of our inference, we certainly need exclude individuals that introduce big biases from our analysis, especially when the reduction in variance by increasing the size of in the $i$Fusion inference cannot overcome those big biases. A very attractive feature of $i$Fusion is that it conducts local searching and pooling, and it does not make any assumption about the underlying individual parameters. To highlight the flexibility of the $i$Fusion over the traditional subgrouping or mixture model approaches and also its scalability to big data application, we include a simulation example (in Section 2.6) of 6000 regression models with the total data size $N = 240000$, in which 6000 pairs of true regression parameters spread evenly on a circle (see the left panel of Figure 2.1). Although the data from most individual regressions should be excluded for the analysis of the data from, say the 1500th regression model (colored in blue in the right panel of Figure 2.1), those data from a few neighboring models (colored in yellow) can help improve the analysis of it. Our theory and simulation support this $i$Fusion practice that drastically improve the efficiency of the individual analysis relying on only the $n_{1500} = 40$ blue data points. Note that there is no clustering or mixture structure in this figure. Conventional methods have difficulty in handling such datasets and providing better results than the individual analysis.

The rest of this chapter is organized as follows. In Section 2.2.1 we provide a brief review of CDs and confidence densities, and how they facilitate fusion learning. In Section 2.2.2 we describe a general $i$Fusion approach, and in Section 2.3 we provide theoretical supports that $i$Fusion can provide asymptotically proper and efficient inference for each individual target. In Section 2.4 we extend $i$Fusion to accommodate more complex and heterogeneous model designs. An efficient and scalable tuning algorithm is described in Section 2.5. To demonstrate the effectiveness of $i$Fusion empirically, we

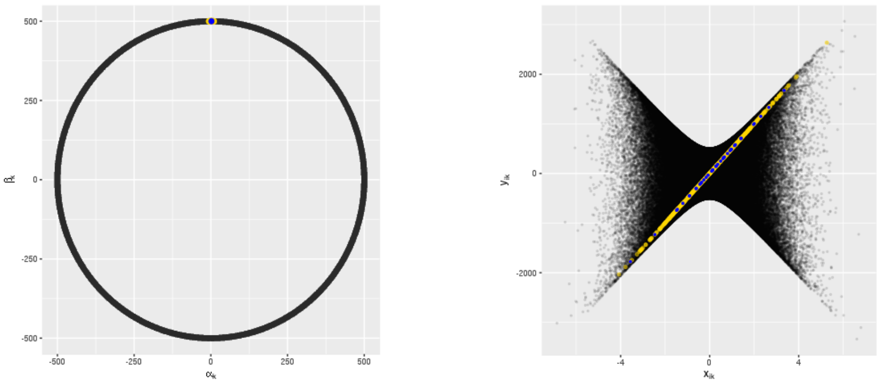

Figure 2.1: Parameter values $(\alpha_k, \beta_k)$ (left) and simulated samples $(x_{ik}, y_{ik})$, $i = 1, \ldots, 50$, (right) for $k = 1, \ldots, 6000$. Here, blue color corresponds to the target individual-1500, and yellow color corresponds to the individuals incorporated in $\mathcal{C}_{1500}$.

present a set of simulation studies in Section 2.6 and a real-world application in Section 2.7. In Section 2.8 we provide further insights and conclude the chapter.

## 2.2 Methodology

### 2.2.1 Review: CD and Fusion Learning

Consider a simple normal example with $x_i \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$, $i = 1, \ldots, n$ for a known $\sigma$, we are interested in making inference of the mean $\theta$. Instead of using a point (sample mean $\bar{x}$) or an interval ($1-\alpha$ level confidence interval $(\bar{x}+\Phi^{-1}(\alpha/2)\sigma/n^{1/2}, \bar{x}+\Phi^{-1}(1-\alpha/2)\sigma/n^{1/2})$), we can use a sample-dependent function $N(\bar{x}, \sigma^2/n)$ to estimate the parameter of interest. Here, $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Such a distribution estimator, referred to as a confidence distribution (CD), can provide meaningful answers to almost all questions related to statistical inference such as point estimation, confidence interval and $p$-values; cf. Xie and Singh (2013), Schweder and Hjort (2016) and references therein. Cox (2013) stated a CD provides "simple and interpretable summaries of what can reasonably be learned from data (and an assumed model)." A CD can be loosely defined as "a sample-dependent distribution that represents confidence intervals of all levels for a parameter of interest (Xie and Singh, 2013). A modern definition of CD is due to Schweder and Hjort (2002) and Singh et al. (2005), and the multivariate version is discussed in Singh et al. (2007) and Schweder and Hjort (2016). If a CD is presented as a density function when appropriate, it is referred to as a confidence density (Efron, 1993; Singh et al., 2007).

The rich information contained in the CD makes it an effective tool for synthesizing information from multiple data sources. Singh et al. (2005) proposed a general framework of combining CDs for a scalar parameter from independent data sources and showed that the combined CD yields valid statistical inference so long as each individual CD is valid, regardless how they are obtained. Xie et al. (2011) showed that the general framework of CD combination can subsume almost all existing meta-analysis approaches as special cases. The nice and general features of CD endue $i$Fusion with great versatility and flexibility, as further described in Section 2.8. Singh et al. (2005) also discussed a framework of combining univariate CDs by multiplying confidence density functions. Liu et al. (2015) extended the framework for fusion learning on multivariate common effects and for heterogeneous study designs, which was also adopted by Tang et al. (2016) and others. A simplified version of their combining formula is

$$h^{(c)}(\boldsymbol{\theta}; \mathcal{S}_1, \ldots, \mathcal{S}_K) = \prod_{k=1}^{K} h_k(\boldsymbol{\theta}; \mathcal{S}_k), \tag{2.1}$$

where $h_k(\boldsymbol{\theta}; \mathcal{S}_k)$ is the confidence density function derived from the $k$th individual subject using data $\mathcal{S}_k$. Liu et al. (2015) showed that the point estimator obtained from $h^{(c)}(\boldsymbol{\theta}; \mathcal{S}_1, \ldots, \mathcal{S}_K)$ enjoys the same estimation efficiency achieved by the maximum likelihood estimator from the analysis of full dataset, but suffices to use individual summary statistics to be implemented.

Most work on combining information in the literature (e.g., Singh et al. (2005) and Liu et al. (2015) and others) are based on the assumption that all the individual parameter values are the same (or at least similar to each other), which can be fairly stringent and hardly hold for many real-world applications. Claggett et al. (2014) relaxed the assumption by allowing unstructured different individual parameter values in a fixed-effects meta-analysis setup, but the development is for the quantiles of the set of individual parameter values and not directly on an individual $\boldsymbol{\theta}_k$.

In next section, with the help of adaptive screening weights, we broaden the framework in Liu et al. (2015) into $i$Fusion for making inference for any target individual parameter. Similar to Claggett et al. (2014), this novel approach is well suited for a very general setting that essentially requires no parametric assumptions on how the individual parameter values are like. Such freedom makes $i$Fusion applicable and useful for tackling many individual-oriented problems like the motivating example in Section 2.1. The framework in Liu et al. (2015), in turn, can be viewed as a special case of $i$Fusion when $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_K$. We close this section by remarking that the original combining formula in Liu et al. (2015) is in fact designed for heterogeneous individual model structures; this motivates our extension of $i$Fusion in Section 2.4.

### 2.2.2 $i$Fusion by Adaptive Combination of CDs

In this section, we formally define the inference problem in math and present our method. Consider a collection of $K$ individual subjects with a dataset $\mathcal{S} = \{\mathcal{S}_1, \ldots, \mathcal{S}_K\}$, where $\mathcal{S}_k$ are samples of size $n_k$ generated independently for the $k$th individual, $k = 1, \ldots, K$. Denote by $n = \sum_{k=1}^{K} n_k$ the sample size of the entire dataset and we assume that $n_k/n \to r_k$ for some constant $r_k \in (0, 1)$ as $n \to \infty$. Suppose that the model for $k$th individual can be characterized by parameter $\boldsymbol{\theta}_k \in \mathbb{R}^{p_k}$, for $k = 1, \ldots, K$. Also, assume that the $K$ individual models have a shared model structure/design (so $p_1 = \cdots = p_K \equiv p$), but the parameter values, $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, can vary across individuals. Any of them may or may not be equal/close to one another, which is completely unknown. Later in Section 2.4, we extend the development to heterogeneous model designs where $p_k$'s can be different from one to the other.

Without loss of generality, individual-1 and thus $\boldsymbol{\theta}_1$ are of our primary interest unless specified otherwise (for convenience, we will use the term individual-$k$, model-$k$, and $\boldsymbol{\theta}_k$ interchangeably), which we refer to as the target individual. Our main research question is: how to make valid and efficient inference about $\boldsymbol{\theta}_1$?

A simple approach is to analyze $\mathcal{S}_1$ directly under the assumed model-1, for which a number of statistical procedures may be used for the task. For the clarity of our presentation and the ease of establishing the asymptotic properties in Section 2.3 and also following Liu et al. (2015), we assume that the individual model can provide us an asymptotic normal confidence distribution, say $N(\hat{\boldsymbol{\theta}}_1, \hat{\Sigma}_1)$, with confidence density

$$h_1(\boldsymbol{\theta}_1; \mathcal{S}_1) = \frac{1}{(2\pi)^{p/2}|\hat{\Sigma}_1|^{1/2}} \exp\big\{ -\frac{1}{2}(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)^t \hat{\Sigma}_1^{-1}(\boldsymbol{\theta}_1 - \hat{\boldsymbol{\theta}}_1)\big\}. \tag{2.2}$$

In the special case of a likelihood inference setup, $\hat{\boldsymbol{\theta}}_1$ is then the maximum likelihood estimate of $\boldsymbol{\theta}_1$, namely, $\hat{\boldsymbol{\theta}}_1 = \arg\max_{\boldsymbol{\theta}} l_1(\boldsymbol{\theta}_1|\mathcal{S}_1)$; and $\hat{\Sigma}_1 = \Sigma_1(\hat{\boldsymbol{\theta}}_1)$ is an estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}_1$ where $\Sigma_1(\boldsymbol{\theta}_1) = [-\partial^2 l_1(\boldsymbol{\theta}_1|\mathcal{S}_1)/\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^t]^{-1}$. We refer to this use of only $\mathcal{S}_1$ for making inference about $\boldsymbol{\theta}_1$ as the individual approach. This individual approach does not utilize any information from other individuals that may be available in the much bigger universe $\mathcal{S}$. This practice is undesirable in some situations such as that described in the motivating example.

We propose an $i$Fusion approach to adaptively fuse information from other individuals to improve the inference efficiency of the individual approach. The first step of $i$Fusion replicates the individual approach for each $k = 1, \ldots, K$, independently, and results in $K$ confidence density functions. Then, it combines these confidence density functions $h_k(\boldsymbol{\theta}_k; \mathcal{S}_k)$, $k = 1, \ldots, K$, according a set of adaptive screening weights,

$$h_1^{(c)}(\boldsymbol{\theta}; \mathcal{S}_1, \ldots, \mathcal{S}_K) = \prod_{k=1}^{K} h_k(\boldsymbol{\theta}; \mathcal{S}_k)^{w_{1k}} \tag{2.3}$$

Here, $h_k(\boldsymbol{\theta}; \mathcal{S}_k)$ is the confidence density function for $\boldsymbol{\theta}_k$ based on $\mathcal{S}_k$, $w_{1k} \in [0, 1]$ is the screening weight for individual-$k$ with respect to individual-1. By introducing the screen weight $w_{1k}$, ideally we would like to include individuals that share the same trait of individual-1 but exclude others that are far different. Further discussions are provided later in this section and also in Section 2.3. From now on we suppress the $\mathcal{S}_k$ in $h_k$ and $\mathcal{S}_1, \ldots, \mathcal{S}_K$ in $h_1^{(c)}$ for notation convenience.

This combined $h_1^{(c)}(\boldsymbol{\theta})$ can be used to derive a combined estimator of $\boldsymbol{\theta}_1$

$$\begin{aligned}\hat{\boldsymbol{\theta}}_1^{(c)} &= \arg\max_{\boldsymbol{\theta}} \log h_1^{(c)}(\boldsymbol{\theta}) \qquad (2.4)\\ &= \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{K} w_{1k} \log h_k(\boldsymbol{\theta})\end{aligned}$$

When the individual confidence density functions take the form of (2.2), simple algebra shows that

$$h_1^{(c)}(\boldsymbol{\theta}) \propto \exp\big\{-\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_1^{(c)})^t(\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_1^{(c)})\big\}, \qquad (2.5)$$

where

$$\hat{\boldsymbol{\theta}}_1^{(c)} = (\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1})^{-1}\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\theta}}_k. \qquad (2.6)$$

The key difference between the combining formulas (2.1) and (2.3) is the use of the screening weights. We visualize this difference in Figure 2.2. When $w_{1k} \equiv 1$ for $\forall k$, (2.3) is the same as (2.1), and $i$Fusion is equivalent to what is shown in the top half. On the other hand with $w_{11} \equiv 1$ and $w_{1k} \equiv 0$, $k = 2, \ldots, K$, $i$Fusion is the same as the individual approach.

One choice of the screening weights is

$$w_{1k} = \mathcal{K}\left(\frac{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2}{b_n}\right)/\mathcal{K}(0), \qquad (2.7)$$

where $\|\cdot\|_2$ stands for the $l^2$ norm, $\mathcal{K}(\cdot)$ is a kernel function, and $b_n$ is a bandwidth parameter that depends on $n$. Some common kernel functions are: i) uniform kernel $\frac{1}{2}\mathbb{1}\{|u| \le 1\}$; ii) Epanechnikov kernel $\frac{3}{4}(1-u^2)\mathbb{1}\{|u| \le 1\}$; iii) quartic kernel $\frac{15}{16}(1-u^2)^2\mathbb{1}\{|u| \le 1\}$; (iv) Gaussian kernel $\frac{1}{\sqrt{2\pi}}e^{-u^2/2}$. Different kernels lead to different finite sample performance. But asymptotically, they are equivalent under suitable regularity conditions on $b_n$, as discussed in Section 2.3.

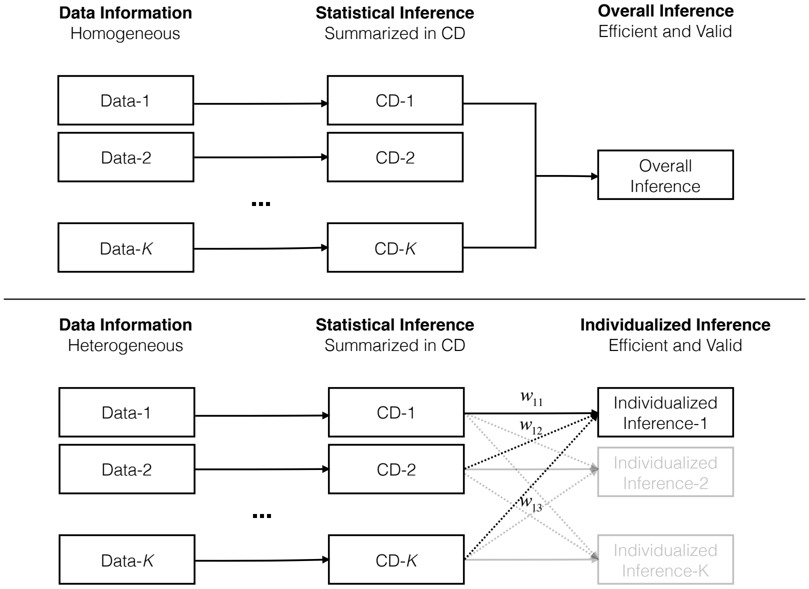

Figure 2.2: Diagrams comparing (top) fusion learning under the assumption that $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_K$, and (bottom) *i*Fusion that produces individualized inference by including screening weights and without assuming that $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_K$. For the bottom half, individual-1 is the inference target.

Provided that the multivariate normal CDs are used, we propose the following formula, slightly modified from (2.7), to further refine its finite sample performance:

$$w_{1k} = \mathcal{K}\left(\frac{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_{(\hat{\Sigma}_1+\hat{\Sigma}_k)^{-1}}}{b_{\bar{n}_{1k}} \cdot (\bar{n}_{1k}p)^{1/2}}\right)/\mathcal{K}(0), \tag{2.8}$$

where $\|\mathbf{x} - \mathbf{y}\|_S = \sqrt{(\mathbf{x}-\mathbf{y})^t S(\mathbf{x}-\mathbf{y})}$ is the Mahalanobis distance with respect to matrix $S$, and $\bar{n}_{1k} = 2n_1 n_k/(n_1 + n_k)$. (2.8) shares the same asymptotic behavior as (2.7), but heuristically has better adaptability to a number of types of variabilities, for example, estimation uncertainty that differs by individual, dimension of $\boldsymbol{\theta}_k$'s as well as scales in different dimensions of $\boldsymbol{\theta}_k$'s.

## 2.3 Theoretical Properties

In this section we establish the theoretical properties of $\hat{\boldsymbol{\theta}}_1^{(c)}$. To facilitate our development, we introduce some concepts and notations. First, define a clique set for individual-1 as

$$\mathcal{C}_1 = \{\boldsymbol{\theta}_k : n^{1/2}\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 = o(1), k = 1, \ldots, K\}. \tag{2.9}$$

The set $\mathcal{C}_1$ always contains $\boldsymbol{\theta}_1$, so $|\mathcal{C}_1| \geq 1$. For any $\boldsymbol{\theta}_k \in \mathcal{C}_1$ and $k \neq 1$, it cannot be distinguished from $\boldsymbol{\theta}_1$ by their $\sqrt{n}$-consistent estimates based on the current sample size. Two extreme cases are: i) $|\mathcal{C}_1| = 1$ indicating that $\boldsymbol{\theta}_1$ is distinguishable from all the other $\boldsymbol{\theta}_k$'s, or ii) $|\mathcal{C}_1| = K$ so all individual parameters cannot be told apart from each other. Between the two extremes is the general situation where $2 \leq |\mathcal{C}_1| \leq K - 1$ ($K \geq 3$), which implies a possible grouping/clustering effect around $\boldsymbol{\theta}_1$. The clique definition follows the "near tie" concept in Xie et al. (2009), Hall and Miller (2010) and Claggett et al. (2014), where the parameters are defined relating to the sample sizes. It can be considered as a "local asymptotic" development (van deer Vaart, 1998) by which "we study the local behavior around a fixed value of the target parameter through a sequence of $\sqrt{n}-$rated parameters" and "measure the performance of an estimator in finer detail and ensure its performance in moderate sample size." (Claggett et al., 2014). Similar asymptotic consideration can also be seen in the high-dimensional regression literatures where it is assumed that the signal level grows at some rate of the sample size, among others. Besides the clique set $\mathcal{C}_1$, we also define a boundary set

$$\mathcal{B}_1 = \{\boldsymbol{\theta}_k : n^{1/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 \to c_k, \text{ for some constant } c_k \in (0, \infty), k = 1, \ldots, K\} \tag{2.10}$$

and the disperse set

$$\mathcal{D}_1 = \{\boldsymbol{\theta}_k : n^{1/2}\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 \to \infty, k = 1, \ldots, K\}. \tag{2.11}$$

Clearly, the parameter set $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\} = \mathcal{C}_1 \cup \mathcal{B}_1 \cup \mathcal{D}_1$ are partitioned into three disjoint sets. An individual $\boldsymbol{\theta}_k$ lies in one and only one of these sets. Define

$$d_1 = \min_k \{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 : \boldsymbol{\theta}_k \in \mathcal{D}_1\} \tag{2.12}$$

as the minimal distance between $\boldsymbol{\theta}_1$ and any parameter inside the disperse set. It immediately follows that $n^{1/2}d_1 \to \infty$.

We start our development by assuming that $\mathcal{B}_1$ is empty, so a $\boldsymbol{\theta}_k$ is either in $\mathcal{C}_1$ or $\mathcal{D}_1$. As such, $d_1$ defined by (2.12) is equivalent to

$$d_1 \equiv \min_k \{\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 : \boldsymbol{\theta}_k \notin \mathcal{C}_1\}. \tag{2.13}$$

We refer to either $\mathcal{B}_1 = \emptyset$ or the equivalence between (2.12) and (2.13) as the separation condition.

The total sample size can be increased when we include additional data from other individuals into our analysis. However, this practice of borrowing information from other individuals will inevitably produce some bias, sometimes ignorable and other times significant. In the terminology of clique, boundary and disperse sets, including individuals in $\mathcal{C}_1$ into analysis introduces little (ignorable) bias for the inference of $\boldsymbol{\theta}_1$, but the bias is not ignorable and even very large by including individuals in $\mathcal{D}_1$. Intuitively, it is attempting to fuse all information from individuals in $\mathcal{C}_1$ to help improve the inference of $\boldsymbol{\theta}_1$. In particular, we define the oracle estimator of $\boldsymbol{\theta}_1$ provided that the membership of $\mathcal{C}_1$ is known in advance

$$\hat{\boldsymbol{\theta}}_1^{(o)} = \arg\max_{\boldsymbol{\theta}} \log h_1^{(o)}(\boldsymbol{\theta}), \tag{2.14}$$

where

$$h_1^{(o)}(\boldsymbol{\theta}) = \prod_{\boldsymbol{\theta}_k \in \mathcal{C}_1} h_k(\boldsymbol{\theta}). \tag{2.15}$$

We also refer to the combination of confidence density functions in (2.15) as the oracle approach. With normal individual confidence densities, it is easy to show that

$$\hat{\boldsymbol{\theta}}_1^{(o)} = (\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1} \sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1} \hat{\boldsymbol{\theta}}_k. \tag{2.16}$$

The following lemma shows that $\hat{\boldsymbol{\theta}}_1^{(o)}$ is consistent, asymptotically normal, and efficient.

**Lemma 2.1.** *Suppose that the membership of $\mathcal{C}_1$ is known. Then, as $n \to \infty$,* *i)* $\hat{\boldsymbol{\theta}}_1^{(o)} \xrightarrow{p} \boldsymbol{\theta}_1$.

*ii)* $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}, \Delta_1^{(o)})$, *where* $\Delta_1^{(o)} = \mathbb{E}[n(\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1}]$.

*iii)* $\hat{\boldsymbol{\theta}}_1^{(o)}$ *is mean squared error (MSE) optimal among all* $\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}$ *given by*

$$\hat{\boldsymbol{\theta}}_1^{\mathcal{F}} = \arg\max_{\boldsymbol{\theta}} \log \prod_{\boldsymbol{\theta}_k \in \mathcal{F}} h_k(\boldsymbol{\theta}_k), \tag{2.17}$$

*where* $\mathcal{F} \subseteq \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$.

The proof of Lemma 2.1 is given in Appendix. Conventional meta-analysis and fusion learning methods often requires that the individuals to be combined share the same parameter values to achieve consistency and asymptotic normality. Part i) and ii) of Lemma 2.1 implies that as long as these individual parameters are close around the target parameter given the sample size, i.e., inside $\mathcal{C}_1$, the above nice properties are preserved. Part iii) further shows that the choice of $\mathcal{F} = \mathcal{C}_1$ leads to the smallest asymptotic MSE, among all the estimators given by the general form of (2.17). Note that the individual estimator $\hat{\boldsymbol{\theta}}_1$ is also a special case of $\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}$ with $\mathcal{F} = \{\boldsymbol{\theta}_1\}$

In reality, the memberships of $\mathcal{C}_1$ is typically unknown. Nevertheless, the oracle approach sets a benchmark for any procedure of making individual inference about $\boldsymbol{\theta}_1$. In fact, we show that the inference about $\boldsymbol{\theta}_1$ led by $i$Fusion is asymptotically the same as the oracle, even without knowing $\mathcal{C}_1$ in advance. Parts i) and ii) of the following theorem provides a set of sufficient conditions (imposed on the screening weights) under which $\hat{\boldsymbol{\theta}}_1^{(c)}$ consistently estimates $\boldsymbol{\theta}_1$ and is also asymptotically normal. Moreover, part ii) shows that $\hat{\boldsymbol{\theta}}_1^{(c)}$ has the same limiting covariance matrix as $\hat{\boldsymbol{\theta}}_1^{(o)}$. Hence, inference using our $i$Fusion approach incurs no loss of efficiency in relative to the oracle approach. The claim applies to MSE as well, as summarized in part iii). A formal proof of the theorem is provided in Appendix.

**Theorem 2.1.** *Suppose that* $w_{1k}$ *satisfies*

$$w_{1k} = \begin{cases} 1 + o_p(n^{-1/2}) & \textit{if } \boldsymbol{\theta}_k \in \mathcal{C}_1; \\ o_p(n^{-1/2}) & \textit{otherwise,} \end{cases} \tag{2.18}$$

*for $k = 1, \ldots, K$. Then, $\hat{\boldsymbol{\theta}}_1^{(c)}$ obtained from* (2.4) *has the following properties: as $n \to \infty$,*

*i)* $\hat{\boldsymbol{\theta}}_1^{(c)} \xrightarrow{p} \boldsymbol{\theta}_1$.

*ii)* $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}, \Delta_1^{(o)})$, *where $\Delta_1^{(o)}$ can be consistently estimated by* $n(\sum_{k=1}^K w_{1k}\hat{\Sigma}_k^{-1})^{-1}(\sum_{k=1}^K w_{1k}^2\hat{\Sigma}_k^{-1})(\sum_{k=1}^K w_{1k}\hat{\Sigma}_k^{-1})^{-1}$.

*iii)* $\hat{\boldsymbol{\theta}}_1^{(c)}$ *has the same MSE as the oracle estimator* $\hat{\boldsymbol{\theta}}_1^{(o)}$.

While Theorem 2.1 outlines a sufficient condition for obtaining a proper and efficient estimator from the $i$Fusion approach, failure to meet this condition risks in invalid inference. The following lemma shows that condition (2.18) is satisfied when formula (2.7) or its modified version (2.8) is used with suitably-chosen bandwidth $b_n$. A proof is provided in Appendix,

**Lemma 2.2.** *$w_{1k}$ given by* (2.7) *or* (2.8) *satisfies* (2.18)*, when any of the following conditions holds:*

*i)* $\mathcal{K}(\cdot)$ *is the uniform kernel, and $b_n$ satisfies that*

$$b_n/d_1 \to 0 \quad \textit{and} \quad n^{1/2}b_n \to \infty. \tag{2.19}$$

*ii)* $\mathcal{K}(\cdot)$ *is the Epanechnikov or the quartic kernel, and $b_n$ satisfies that*

$$b_n/d_1 \to 0 \quad \textit{and} \quad n^{1/4}b_n \to \infty. \tag{2.20}$$

*iii)* $\mathcal{K}(\cdot)$ *is the Gaussian kernel, and $b_n$ satisfies that*

$$(b_n/d_1)^2 \log n \to 0 \quad \textit{and} \quad n^{1/4}b_n \to \infty. \tag{2.21}$$

In the above development, we have assumed that $\mathcal{B}_1 = \emptyset$, under which our $i$Fusion method can yield an estimator that is asymptotically equivalent to the best possible oracle estimator and provide asymptotically the most efficient inference about $\boldsymbol{\theta}_1$. We now turn to the more complicated case that $\mathcal{B}_1 \neq \emptyset$, where the development is not such clean since the parameters in $\mathcal{B}_1$ are not easily separable from those in $\mathcal{C}_1$, making the asymptotic normality infeasible to be used. Note that, for $\boldsymbol{\theta}_k \in \mathcal{B}_1$, bias caused is of the same magnitude of standard deviation. Inclusion of these individuals reduces variance at the price of introducing comparable bias. We have the following theorem to quantify the performance of our *i*Fusion method under this more complicated setting.

**Theorem 2.2.** *Suppose* $w_{1k}$ *satisfies*

$$w_{1k} = \begin{cases} 1 + o_p(n^{-1/2}) & \textit{if } \boldsymbol{\theta}_k \notin \mathcal{D}_1; \\ o_p(n^{-1/2}) & \textit{otherwise.} \end{cases} \tag{2.22}$$

*for* $k = 1, \ldots, K$. *Then,* $\hat{\boldsymbol{\theta}}_1^{(c)}$ *obtained from* (2.4) *has the following properties: as* $n \to \infty$,

*i)* $\hat{\boldsymbol{\theta}}_1^{(c)} \xrightarrow{p} \boldsymbol{\theta}_1$.

*ii)* $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1 - \boldsymbol{B}_1^{(c)}) \xrightarrow{d} N(\mathbf{0}, \bar{\Delta}_1)$, *where* $\boldsymbol{B}_1^{(c)} = (\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1})^{-1} \sum_{\boldsymbol{\theta}_k \in \mathcal{B}_1} \hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)$, *and* $\bar{\Delta}_1 = \mathbb{E}[n(\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1})^{-1}]$.

*iii)* $MSE(\hat{\boldsymbol{\theta}}_1^{(c)}) \leq MSE(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}})$, *if either of the following is met:* $\mathcal{D}^{\mathcal{F}} \neq \emptyset$, *or*

$$\begin{aligned} &\sum_{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2} \in \mathcal{B}_1} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} (\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1})^{-2} \hat{\Sigma}_{k_2}^{-1} (\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + tr\Big\{(\sum_{\boldsymbol{\theta}_k \notin \mathcal{D}_1} \hat{\Sigma}_k^{-1})^{-1}\} \\ \leq &\sum_{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2} \in \mathcal{B}^{\mathcal{F}}} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} (\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-2} \hat{\Sigma}_{k_2}^{-1} (\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + tr\Big\{(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-1}\}. \end{aligned} \tag{2.23}$$

*Here,* $\mathcal{F}$ *is expressed as the union of three disjoint sets* $\mathcal{C}^{\mathcal{F}} \cup \mathcal{B}^{\mathcal{F}} \cup \mathcal{D}^{\mathcal{F}}$ *where* $\mathcal{C}^{\mathcal{F}} \subseteq \mathcal{C}_1$, $\mathcal{B}^{\mathcal{F}} \subseteq \mathcal{B}_1$, *and* $\mathcal{D}^{\mathcal{F}} \subseteq \mathcal{D}_1$.

We remark that part ii) and iii) of Theorem 2.2 are less feasible than their counterparts in Theorem 2.1, because of the unknown true parameter values. In particular, the limiting distribution of $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1 - \boldsymbol{B}_1^{(c)})$ will become nonnormal if $\boldsymbol{B}_1^{(c)}$ is substituted with a $\sqrt{n}$-consistent estimate.

Again, it is important to assure (2.22) before any claim of $\hat{\boldsymbol{\theta}}_1^{(c)}$ in Theorem 2.2 can be asserted. Fortunately, formulas (2.7) and (2.8) are directly applicable to yield qualifying $w_{1k}$'s even $\mathcal{B}_1 \neq \emptyset$ (although the equivalence between (2.13) and (2.12) no longer holds). The result is given by the following lemma, whose proof is essentially the same as Lemma 2.2 and is therefore omitted.

**Lemma 2.3.** *When* $\mathcal{B}_1 \neq \emptyset$*,* $w_{1k}$ *given by* (2.7) *or* (2.8) *satisfies* (2.22)*, if any of conditions* (2.19)*,* (2.21)*, or* (2.20) *holds.*

## 2.4 Extension to Heterogeneous Model Designs

In this section, we show that *i*Fusion can be extended to heterogeneous individual model designs. Simmonds and Higgins (2007) and Liu et al. (2015) provided excellent reviews on model design heterogeneity (although they used the term "parameter heterogeneity" what means varying individual parameter values in our context) encountered in meta-analysis. We use two examples modified from Liu et al. (2015) to demonstrate why the *i*Fusion approach developed in Section 2.2 shall not be directly applied under such heterogeneous model designs. For both examples, we consider $K$ independent linear models implied by clinical trials conducted on $K$ different populations:

$$Y_{ik} = \alpha_k + \beta_k x_{ik} + \gamma_k z_{ik} + \varepsilon_{ik},\ i = 1, \ldots, n_k,\ k = 1, \ldots, K, \tag{2.24}$$

where $Y_{ik}$ is the response for the $i$th observation from the $k$th population, $x_{ik}$ is the treatment status (1/0 for treatment/control), $z_{ik}$ is the drug dosage, and $\varepsilon_{ik} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_k^2)$. Here, $\alpha_k$'s are individual-specific effects, $\beta_k$'s and $\gamma_k$'s measure the sensitivities of response to the treatment and drug dosage, respectively.

**Example 2.1.** The $K$ populations have "distinct gender, race, or disease status."(Liu et al., 2015) Their impact on the response is incorporated in $\alpha_k$'s if they affect the response independently of the treatment and drug dosage. In this case, $\alpha_k$'s are "designed to be heterogeneous." If formula (2.7) or (2.8) is directly applied, no information may be gained because of the discrepancy among $\alpha_k$'s (so the estimated $\hat{\alpha}_k$'s), even if

$(\beta_1, \gamma_1) = \cdots = (\beta_K, \gamma_K)$, thus incurring loss of efficiency.

**Example 2.2.** The drug dosage is not part of the research goal and is hold constant in trial-1, that is, $z_{i1} \equiv z_1$, for $i = 1, \ldots, n_1$. Then, model-1 degenerates to

$$Y_{i1} = (\alpha_1 + \gamma_1 z_1) + \beta_1 x_{i1} + \varepsilon_{i1},\ i = 1, \ldots, n_1.$$

This is known as the "missing covariate designs" pointed out in Simmonds and Higgins (2007), where certain individuals "do not have the design covariate that is of current research interest." Under such situation, neither $\alpha_1$ nor $\gamma_1$ is estimable; formulas (2.7) and (2.8) are not even well-defined.

One of the obstacles to $i$Fusion in the two examples is the unsuitability (Example 2.1) or infeasibility (Example 2.2) of formulating $w_{1k}$ based on full parameter vectors $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_k$. To overcome these problems, we propose a new combining formula and definition of clique set (as well as boundary and disperse sets). Specifically, we partition $\boldsymbol{\theta}_1$ into two disjoint parts, $\boldsymbol{\psi}_1$ and $\boldsymbol{\xi}_1$, where $\boldsymbol{\xi}_1$ corresponds to part of the structure common to all individual models, whereas $\boldsymbol{\psi}_1$ is treated as "nuisance". For instance, in Example 2.1, $\boldsymbol{\xi}_1 = (\beta_1, \gamma_1)^t$, $\psi_1 = \alpha_1$; and in Example 2.2 $\xi_1 = \beta_1$, $\psi_1 = \alpha_1 + \gamma_1 z_1$. Partition $\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_K$ accordingly. We modify the definitions of clique, boundary and disperse sets as:

$$\tilde{\mathcal{C}}_1 = \{\boldsymbol{\xi}_k : n^{1/2} \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_k\|_2 = o(1), k = 1, \ldots, K\}, \tag{2.25}$$

$$\tilde{\mathcal{B}}_1 = \{\boldsymbol{\xi}_k : n^{1/2} \|\boldsymbol{\xi}_k - \boldsymbol{\xi}_1\|_2 \to c_k,\ \text{for some constant } c_k \in (0, \infty), k = 1, \ldots, K\},$$

$$\tilde{\mathcal{D}}_1 = \{\boldsymbol{\xi}_k : n^{1/2} \|\boldsymbol{\xi}_k - \boldsymbol{\xi}_1\|_2 \to \infty, k = 1, \ldots, K\}.$$

Moreover, modify the screening weight (2.7) by substituting $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_k$ with $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_k$, respectively; and (2.8) can be modified with only little extra effort.

To make inference on $\boldsymbol{\theta}_1$, one possible solution is to break the problem according to the partition of $\boldsymbol{\theta}_1$ and then: i) use $h_1(\boldsymbol{\theta})$ to infer $\boldsymbol{\psi}_1$; ii) obtain the confidence density functions for $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K$ by marginalizing the full confidence density functions $h_1(\boldsymbol{\theta}), \ldots, h_K(\boldsymbol{\theta})$, respectively; iii) apply formula (2.3) as before but on the marginalized confidence density functions to infer $\boldsymbol{\xi}_1$. This marginal approach is theoretically justified because the CD combining method is "developed strictly under the frequentist paradigm" and it can "focus directly on the parameter of interest without the additional burden of modeling other parameters" (Xie et al., 2013). In fact, consistency and asymptotic normality can be achieved under this approach, but we show that by considering all parameter components together in one-shot we can achieve more efficient inference.

Assume, without loss of generality, that $w_{1k} \neq 0$, for $k = 1, \ldots, K$ (otherwise if $w_{1k} = 0$, we simply exclude individual-$k$ from the analysis). Also, assume that $\boldsymbol{\psi}_k$ is a scalar so $\boldsymbol{\theta}_k = (\psi_k, \boldsymbol{\xi}_k^t)^t$, for $k = 1, \ldots, K$. Let $\boldsymbol{\eta}_k = (\psi_1, \ldots, \psi_K, \boldsymbol{\xi}_k^t)^t$ and write $\boldsymbol{\eta} = (\psi_1, \ldots, \psi_K, \boldsymbol{\xi}^t)^t$. Denote by $A_k$ be the matrix that maps $\boldsymbol{\eta}$ to $(\psi_k, \boldsymbol{\xi}^t)^t$. The new proposed combining formula

$$h_1^{(c)}(\boldsymbol{\eta}) = \prod_{k=1}^{K} h_k(A_k\boldsymbol{\eta})^{w_{1k}}.$$

Let

$$\hat{\boldsymbol{\theta}}_1^{(c)} = A_1 \hat{\boldsymbol{\eta}}_1^{(c)}. \tag{2.26}$$

where

$$\hat{\boldsymbol{\eta}}_1^{(c)} = \arg\max_{\boldsymbol{\eta}} \log h_1^{(c)}(\boldsymbol{\eta}).$$

We provide theoretical development parallel to that in Section 2.3. Under the assumption that $\tilde{\mathcal{B}}_1 = \emptyset$, we define the oracle estimator of $\boldsymbol{\theta}_1$ given $\tilde{\mathcal{C}}_1$ is known:

$$\hat{\boldsymbol{\theta}}_1^{(o)} = A_1 \hat{\boldsymbol{\eta}}_1^{(o)},$$

where

$$\begin{aligned}\hat{\boldsymbol{\eta}}_1^{(o)} &= \arg\max_{\boldsymbol{\eta}} \log h_1^{(o)}(\boldsymbol{\eta}) \\ &= \arg\max_{\boldsymbol{\eta}} \log \prod_{\boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1} h_k(A_k \boldsymbol{\eta}).\end{aligned}$$

We state, without formal proof, that similar to Lemma 2.1, the oracle estimator $\hat{\boldsymbol{\theta}}_1^{(o)}$ is consistent and asymptotically normally distributed, and has the smallest asymptotical MSE among all estimator of $\boldsymbol{\theta}_1$ given by

$$\hat{\boldsymbol{\theta}}^{\mathcal{F}} = A_1 \hat{\boldsymbol{\eta}}^{\mathcal{F}}, \tag{2.27}$$

where

$$\hat{\boldsymbol{\eta}}^{\mathcal{F}} = \arg\max_{\boldsymbol{\eta}} \prod_{\boldsymbol{\xi}_k \in \mathcal{F}} h_k(A_k \boldsymbol{\eta}),$$

for $\mathcal{F} \subseteq \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K\}$,

The following theorem shows that $\hat{\boldsymbol{\theta}}_1^{(c)}$ consistently estimates $\boldsymbol{\theta}_1$ and is asymptotically normally distributed for suitable choice of $w_{1k}$. Moreover, it has the same limiting covariance matrix as that of $\hat{\boldsymbol{\theta}}_1^{(o)}$ so no loss efficiency is incurred, and the argument applies to MSE as well. The theorem can be viewed the counterpart of Theorem 2.1. Its proof is quite simlar to that of Theorem 2.1 as well with only slight modification for $A_k$ and thereby is omitted.

**Theorem 2.3.** *Suppose $w_{1k}$ satisfies*

$$w_{1k} = \begin{cases} 1 + o_p(n^{-1/2}) & \textit{if } \boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1; \\ o_p(n^{-1/2}) & \textit{otherwise,} \end{cases} \tag{2.28}$$

*for $k = 1, \ldots, K$. Then, $\hat{\boldsymbol{\theta}}_1^{(c)}$ obtained from (2.26) has the following properties: as $n \to \infty$,*

*i) $\hat{\boldsymbol{\theta}}_1^{(c)} \xrightarrow{p} \boldsymbol{\theta}_1$.*

*ii) $n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}, \tilde{\Delta}_1)$, where $\tilde{\Delta}_1 = \mathbb{E}[nA_1(\sum_{\boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1} A_1^t]$ and can be* *consistently estimated by* $nA_1(\sum_{k=1}^K w_{1k}A_k^t\hat{\Sigma}_k^{-1}A_k)^{-1}(\sum_{k=1}^K w_{1k}^2A_k^t\hat{\Sigma}_k^{-1}A_k)(\sum_{k=1}^K w_{1k}A_k^t\hat{\Sigma}_k^{-1}A_k)^{-1}A_1^t$.

*iii)* $\hat{\boldsymbol{\theta}}_1^{(c)}$ *has the same MSE as the oracle estimator* $\hat{\boldsymbol{\theta}}_1^{(o)}$*.*

Let $\hat{\psi}_1^{(c)}$ and $\hat{\psi}_1$ be the respecting part of $\hat{\boldsymbol{\theta}}_1^{(c)}$ and $\hat{\boldsymbol{\theta}}_1$ that estimates $\psi_1$. Here, as usual, $\hat{\boldsymbol{\theta}}_1 = \arg\max_{\boldsymbol{\theta}} \log h_1(\boldsymbol{\theta})$. An interesting byproduct of Theorem 2.3 is that $\hat{\psi}_1^{(c)}$ improves upon $\hat{\psi}_1$.

**Corollary 2.1.** *Asymptotically,* $Var(\hat{\psi}_1^{(c)}) \leq Var(\hat{\psi}_1)$ *under the assumptions of Theorem 2.3.*

It shows that there is efficiency gain in the joint approach over the individual approach. It implies that other individuals can contribute to the estimation of $\psi_1$. This may seem counterintuitive at first glance as other individuals do not contain information on $\psi_1$. However, since information for $\psi_1$ and $\boldsymbol{\xi}_1$ is related, improvement in the estimation of $\boldsymbol{\xi}_1$ can in fact improve the estimation of $\psi_1$. As pointed out in Liu et al. (2015), "this phenomenon of borrowing strength is not yet well appreciated in conventional meta-analysis and the individual-specific parameter are generally reported as the final estimators." Despite the more restrictive setting used in Liu et al. (2015) that $\boldsymbol{\xi}_1 = \cdots = \boldsymbol{\xi}_K$ (in our notation), $i$Fusion benefits from the same principle as well even when $\boldsymbol{\xi}_k$'s are not necessarily the same or similar.

## 2.5 Scalable Algorithm for Tuning Screening Weights

The screening weight $w_{1k}$ involves an unknown scaling parameter $b_n$, which can be further decomposed as $b_n = \tau_n b$ for some constant $b$. In practice, we may set $\tau_n$ according to the conditions stated in Lemma 2.2 so that $w_{1k}$ is well-behaved asymptotically. The unknown constant $b$, however, bears a noticeable impact on the performance of $i$Fusion under finite sample size: a very large $b$ results in "over aggressive" inference led by incorrectly including some irrelevant individuals; a very small $b$ leads to the individual approach and no efficiency is gained. In this section, we provide a tuning algorithm based on cross-validation to search for an appropriate $b$:

1. For $k = 1, \ldots, K$, randomly split each $\mathcal{S}_k$ into $V$ equally sized folds $\{\mathcal{S}_k^1, \ldots, \mathcal{S}_k^V\}$. Denote by $\mathcal{S}_k^{-v} = \mathcal{S}_k / \mathcal{S}_k^v$, for $v = 1, \ldots, V$.

2. For a fixed $b$, let $\hat{\boldsymbol{\theta}}_1^{(c)}(b, v)$ be the combined estimator from applying $i$Fusion to $\{\mathcal{S}_1^{-v}, \mathcal{S}_2^{-v}, \ldots, \mathcal{S}_K^{-v}\}$ using $b_n = b\tau_n$ in calculating $w_{1k}$.

3. Compute $\mathbb{L}(b, v)$, the loss of $\hat{\boldsymbol{\theta}}_1^{(c)}(b, v)$ evaluated on $\mathcal{S}_1^v$. The choice of loss function depends on the specific problem. For example, in Simulation I in Section 2.6, the quadratic loss is used:

$$\mathbb{L}(b, v) = \frac{1}{|\mathcal{S}_1^v|} \sum_{Y_{i1} \in \mathcal{S}_1^v} (Y_{i1} - \hat{\boldsymbol{\theta}}_1^{(c)}(b, v))^2.$$

4. Repeat Steps 2 and 3 for $v = 1, \ldots, V$ and average the losses over the $V$ folds:

$$\bar{\mathbb{L}}(b) = \frac{1}{V} \sum_{v=1}^{V} \mathbb{L}(b, v).$$

   Also, compute the standard deviation of $\{\mathbb{L}(b, 1), \ldots, \mathbb{L}(b, V)\}$ denoted by $\mathrm{sd}(\mathbb{L}(b))$.

5. Repeat Steps 2 to 4 along a path of $b \in \mathcal{P}$. Let

$$b^* = \arg\min_{b \in \mathcal{P}} \bar{\mathbb{L}}(b),$$

   and choose $b$ as

$$b^{\mathrm{opt}} = \mathrm{median}\{b : \bar{\mathbb{L}}(b) \leq \bar{\mathbb{L}}(b^*) + \mathrm{sd}(\mathbb{L}(b^*))/\sqrt{V}, b \in \mathcal{P}\}.$$

   Rather than the global minimizer $b^*$, we choose the median of the $b$'s whose corresponding losses are no greater than the minimum loss by one standard error of it. This accommodates for the randomness inherent in $\bar{\mathbb{L}}(b^*)$ and is similar to the one standard error rule of cross validation in Hastie et al. (2010).

The tuning algorithm introduces extra computational burden, but can be accelerated in a number of ways as remarked below.

**Remark 2.1.** In Step 1, when it comes of a large number of individual subjects, a quick pre-screen can be carried out using the ranks of $\{\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_1\|_2\}_{k=1}^K$. Computing $l^2$ norms can be vectorized in many programming languages and so can be done for all $k$ in a shot, in contrast to Mahalanobis distance that requires inverting a matrix and has to be computed one by one. Denote the ranks of $\{\|\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_1\|_2\}_{k=1}^K$ by $\{u_k\}_{k=1}^K$. We set $w_{1k} = 0$ if $u_k > u^*$ for a pre-specified $u^* \in \{1, \dots, K\}$. As such, only a portion of the individuals are carried over to the next steps. The choice of $u^*$ should depends on total number of individual subjects as well as the allocated computing resources.

**Remark 2.2.** In Step 5, it is often not necessary to search the full path $\mathcal{P}$. Typically, loss functions like the quadratic loss roughly exhibit a down-and-up pattern as a function of $b$, due to the intrinsic bias-variance tradeoff. Hence, we may begin with some small $b$, then gradually increase it and stop if there is no further drop of loss. Specifically, let $b^{m*} = \arg\min_{1,\dots,m} \bar{\mathbb{L}}(b_m)$ that corresponds to the running minimum average loss by the $m$th value in $\mathcal{P}$. We stop if $\bar{\mathbb{L}}(b)$ has exceeded $\bar{\mathbb{L}}(b^{m*}) + \mathrm{sd}(\mathbb{L}(b^{m*})/\sqrt{V})$ for consecutively rounds, and then choose

$$b^{\mathrm{opt}} = \mathrm{median}\{b_{m'} : \bar{\mathbb{L}}(b_{m'}) \leq \bar{\mathbb{L}}(b^{m*}) + \mathrm{sd}(\mathbb{L}(b^{m*}))/\sqrt{V}, m' \leq m\}.$$

**Remark 2.3.** The design of *i*Fusion, together with the tuning algorithm, makes it a natural fit for distributed implementation and especially suitable when the individual datasets are stored in different computer clusters. In this case, a central coordinator i) collects the individual confidence density functions that are independently computed using $\mathcal{S}_k^{-v}$ on cluster-$k$, for $k = 1, \dots, K$, ii) computes the combined estimator $\hat{\theta}_1^{(c)}(b, v)$ according to some choice of $b$ and sends back to cluster-1 for evaluation. Such design benefits the most when the target includes every individual and thus in step ii) the coordinator will communicate with every cluster in addition to cluster-1. As such, the algorithm scales up to very large aggregated sample sizes that are too big to be stored/processed in a single computer.

## 2.6 Simulation

In this section, we conduct extensive simulation studies to numerically demonstrate the results established in Sections 2.3 and 2.4. In each study, we compare the performance of *i*Fusion with the individual approach and the oracle approach, as well as some other competitive methods, for example, subgroup analysis (Simulation I) and nonparametric Bayesian method (Simulation II). The approaches are compared thoroughly by the MSEs of point estimators and empirical coverages and width of confidence intervals.

**Simulation I.** We generate random data: $Y_{ik} \overset{\text{i.i.d.}}{\sim} N(\theta_k, 1)$, for $i = 1, \ldots, n_k$, $k = 1, \ldots, 9$, with $\theta_k$ taking values as follows: i) $\theta_k = 0$ for $k = 1, 2, 3$; this forms a clique with equal parameter values. ii) $\theta_k = d + (k-5)/n_k$ for $k = 4, 5, 6$; this also forms a clique according to (2.9) but with varying parameter values. iii) $\theta_k = (k-5)d$ for $k = 7, 8, 9$. Here, $d$ mimics the minimum distance between the parameters that are respectively inside and outside a clique, as defined in (2.13). Furthermore, we allow it to depend on $n_k$ by setting $d = 3n_k^{-1/6}$.

The density function of $N(\hat{\theta}_k, \hat{\sigma}_k^2/n_k)$ is an asymptotic confidence density function for $\theta_k$, where $\hat{\theta}_k = \bar{Y}_{\cdot k} = \sum_{i=1}^{n_k} Y_{ik}/n_k$ and $\hat{\sigma}_k^2 = \sum_{i=1}^{n_k}(Y_{ik} - \bar{Y}_{\cdot k})^2/(n_k - 1)$. To make inference about $\theta_k$, we may directly use this normal confidence density function to draw point estimate (which is simply $\hat{\theta}_k$) and confidence intervals of $\theta_k$; this yields the individual approach. We can also combine the $K$ confidence density functions according to a set of screening weights $w_{1k}$'s, targeting at individual-$k$. Point estimate and confidence intervals can be easily read off as well. This gives the *i*Fusion approach. In its implementation, $w_{1k}$ is calculated using the uniform kernel with $\tau_n = n_k^{-1/3}$ and $b$ tuned via a 5-fold cross-validation under quadratic loss (the same configuration applies to Simulation II and Simulation III).

To measure the long-run performance, we repeat the simulation 500 times for each

$k$ and for $n_k = 40, 400$ that represents small/moderate and large sample sizes respectively. For both the individual approach and $i$Fusion, the following summary statistics are reported: MSE of the point estimate, empirical coverage probability and median width of the confidence interval at the 95% nominal level. Since as a simulation study we have already know the membership of each clique, we can obtain the oracle estimators and confidence intervals by setting the screening weights to matching the memberships of the corresponding clique. For example, when the target is individual-1, $(w_{11}, \ldots, w_{19}) = (1, 1, 1, 0, 0, 0, 0, 0, 0)$, and when the target is individual-8, $(w_{81}, \ldots, w_{89}) = (0, 0, 0, 0, 0, 0, 0, 1, 0)$. As such, we can observe how well our $i$Fusion approach performs in relative to the best possible inferences.

| | | $n_k = 40$ | | | | | | $n_k = 400$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Indiv | *i*Fusion | Oracle | 4-Cls | 5-Cls | 6-Cls | Indiv | *i*Fusion | Oracle | 4-Cls | 5-Cls | 6-Cls |
| MSE | $\theta_1$ | 0.026 | 0.009 | 0.008 | 0.008 | 0.012 | 0.021 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 |
| | $\theta_2$ | 0.023 | 0.010 | 0.008 | 0.008 | 0.013 | 0.018 | 0.003 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 |
| | $\theta_3$ | 0.024 | 0.009 | 0.008 | 0.008 | 0.014 | 0.018 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 | 0.002 |
| | $\theta_4$ | 0.025 | 0.011 | 0.009 | 0.015 | 0.012 | 0.016 | 0.003 | 0.001 | 0.001 | 0.002 | 0.001 | 0.002 |
| | $\theta_5$ | 0.024 | 0.009 | 0.008 | 0.015 | 0.013 | 0.017 | 0.002 | 0.001 | 0.001 | 0.002 | 0.001 | 0.002 |
| | $\theta_6$ | 0.025 | 0.011 | 0.008 | 0.016 | 0.012 | 0.017 | 0.003 | 0.001 | 0.001 | 0.002 | 0.001 | 0.002 |
| | $\theta_7$ | 0.030 | 0.030 | 0.030 | 0.388 | 0.162 | 0.074 | 0.003 | 0.003 | 0.003 | 0.166 | 0.064 | 0.026 |
| | $\theta_8$ | 0.022 | 0.023 | 0.022 | 0.683 | 0.293 | 0.106 | 0.003 | 0.003 | 0.003 | 0.310 | 0.117 | 0.045 |
| | $\theta_9$ | 0.024 | 0.024 | 0.024 | 0.331 | 0.149 | 0.063 | 0.002 | 0.002 | 0.002 | 0.146 | 0.057 | 0.023 |
| Coverage | $\theta_1$ | 0.944 | 0.946 | 0.954 | 0.944 | 0.936 | 0.916 | 0.948 | 0.940 | 0.940 | 0.954 | 0.930 | 0.918 |
| | $\theta_2$ | 0.956 | 0.946 | 0.954 | 0.944 | 0.932 | 0.934 | 0.944 | 0.940 | 0.940 | 0.954 | 0.940 | 0.920 |
| | $\theta_3$ | 0.952 | 0.944 | 0.954 | 0.944 | 0.934 | 0.930 | 0.956 | 0.940 | 0.940 | 0.954 | 0.934 | 0.920 |
| | $\theta_4$ | 0.942 | 0.928 | 0.938 | 0.888 | 0.922 | 0.932 | 0.940 | 0.942 | 0.942 | 0.936 | 0.930 | 0.918 |
| | $\theta_5$ | 0.950 | 0.932 | 0.938 | 0.894 | 0.920 | 0.912 | 0.958 | 0.938 | 0.938 | 0.932 | 0.944 | 0.920 |
| | $\theta_6$ | 0.952 | 0.926 | 0.942 | 0.896 | 0.928 | 0.924 | 0.946 | 0.942 | 0.942 | 0.936 | 0.930 | 0.918 |
| | $\theta_7$ | 0.926 | 0.926 | 0.926 | 0.466 | 0.760 | 0.892 | 0.956 | 0.956 | 0.956 | 0.440 | 0.760 | 0.886 |
| | $\theta_8$ | 0.952 | 0.950 | 0.952 | 0.046 | 0.876 | 0.840 | 0.934 | 0.934 | 0.934 | 0.002 | 0.598 | 0.830 |
| | $\theta_9$ | 0.940 | 0.940 | 0.940 | 0.476 | 0.746 | 0.870 | 0.944 | 0.944 | 0.944 | 0.506 | 0.786 | 0.884 |
| Width | $\theta_1$ | 0.618 | 0.351 | 0.351 | 0.351 | 0.384 | 0.442 | 0.196 | 0.113 | 0.113 | 0.114 | 0.124 | 0.142 |
| | $\theta_2$ | 0.616 | 0.351 | 0.351 | 0.351 | 0.378 | 0.440 | 0.196 | 0.113 | 0.113 | 0.113 | 0.123 | 0.140 |
| | $\theta_3$ | 0.614 | 0.351 | 0.351 | 0.351 | 0.378 | 0.445 | 0.196 | 0.113 | 0.113 | 0.113 | 0.123 | 0.142 |
| | $\theta_4$ | 0.614 | 0.352 | 0.352 | 0.349 | 0.371 | 0.416 | 0.196 | 0.113 | 0.113 | 0.113 | 0.112 | 0.135 |
| | $\theta_5$ | 0.620 | 0.352 | 0.352 | 0.349 | 0.376 | 0.423 | 0.196 | 0.113 | 0.113 | 0.113 | 0.120 | 0.136 |
| | $\theta_6$ | 0.620 | 0.352 | 0.352 | 0.349 | 0.374 | 0.421 | 0.196 | 0.113 | 0.113 | 0.113 | 0.120 | 0.136 |
| | $\theta_7$ | 0.618 | 0.618 | 0.618 | 0.515 | 0.577 | 0.603 | 0.196 | 0.196 | 0.196 | 0.164 | 0.184 | 0.191 |
| | $\theta_8$ | 0.614 | 0.612 | 0.614 | 0.444 | 0.550 | 0.601 | 0.196 | 0.196 | 0.196 | 0.138 | 0.174 | 0.187 |
| | $\theta_9$ | 0.615 | 0.615 | 0.615 | 0.531 | 0.585 | 0.610 | 0.196 | 0.196 | 0.196 | 0.168 | 0.185 | 0.192 |

Table 2.1: Results of Simulation I comparing individual, *i*Fusion, oracle, and subgroup approaches: MSE of point estimates, empirical coverage and median width of 95% confidence intervals. The subgroup approach uses $k$-means clustering to divide the individuals into 4, 5, or 6subgroups and then combines individual confidence densities within each subgroup. The "true" number of subgroups implied by our parameter setting is 5.

Simulation results are shown in Table 3.1. We first note that, given existence of a clique of size greater than one (individuals- 1-6), $i$Fusion always returns point estimates with noticeably less MSE than the individual approach. The confidence intervals from $i$Fusion are also much narrower than those from the individual approach, but preserves approximately the desired nominal coverage probabilities. In fact, $i$Fusion approximate the results of the oracle combination for $n_k = 40$, and are exactly the same as the oracle approach for $n_k = 400$, thus has numerically validated the theoretical results in particular Theorem 2.1 established in Section 2.3. Finally, for an individual as a clique by itself (individuals- 7,8,9), all the three approaches yield quite similar or the same results.

Subgroup analysis is another possible alternative as mentioned in Section 2.1. To be specific, we first use $k$-means clustering on the estimated individual parameters to divide the individuals into a number of subgroups/clusters. Then, within the cluster, the individual confidence density functions are combined to obtain a point estimate and confidence intervals. These results are assumed to be identically applicable to all individuals within that cluster. To reduce the numerical instability of $k$-means clustering algorithm, we randomly set five initial assignments and choose the one that ends up with the smallest sum of squares from points to the assigned cluster centers. Note the number of clusters used in $k$-means clustering need to be determined in advance. In our experiment $4, 5, 6$ cluster setups are tested respectively, where 5-cluster setup aligns with the actual parameter setting.

From in Table 3.1, it is observed that, however, the results on different individuals behave quite differently to the number of clusters. For individuals -1,2,3, the 4-cluster setup works best and approaches the oracle results. For individuals -4,5,6, the 5-cluster setup is in favor of the other two setups, but it is still underperforming the oracle approach with greater MSEs and lower coverage probabilities. For individuals -7,8,9 each of which should ideally constitute a cluster itself, the subgroup approach blows up, with much worse MSEs and coverage probabilities in all cases; it turns out that $k$-means clustering often incorrectly bind, say individual-8, with other individuals, leading to incorrect inference even the cluster number has agreed with the clique configuration.

**Simulation II.** The second simulation study fully expands the illustrative example in Section 2.1, which involves a bivariate setting and a much larger number of individual subjects. Specifically, we simulate 6000 regression datasets according to

$$Y_{ik} = \alpha_k + \beta_k x_{ik} + \varepsilon_{ik}, \quad \varepsilon_{ik} \overset{\text{i.i.d.}}{\sim} N(0,1), \tag{2.29}$$

for $i = 1, \ldots, n_k$ and $k = 1, \ldots, 6000$. The true parameter values of $\{\boldsymbol{\theta}_k = (\alpha_k, \beta_k)^t, k = 1, \ldots, 6000\}$ are set to spread along a circle of radius $R$ by i) generating 1200 points evenly distributed along the circle, ii) replicating each point four times to obtain 6000 points (including the original 1200 points), and iii) adding disturbance to each point. Mathematically, $(\alpha_k, \beta_k) = (R\cos(\lfloor \frac{k-1}{5} \rfloor \frac{2\pi}{1200}) + \frac{U_{k1}}{n}, R\sin(\lfloor \frac{k-1}{5} \rfloor \frac{2\pi}{1200}) + \frac{U_{k2}}{n})$, where $U_{kj} \overset{\text{i.i.d.}}{\sim} U[-1,1]$, for $j = 1, 2$ and $k = 1, \ldots, 6000$. We set $R = 500$ so that the original 1200 points are well separated and $\{\boldsymbol{\theta}_{k-4}, \boldsymbol{\theta}_{k-3}, \ldots, \boldsymbol{\theta}_k\}$ form a clique of size five for $k = 5, 10, \ldots, 6000$. Next, we simulate $x_{ik}$ independently from $N(0, 1.5^2)$ and then generate $Y_{ik}$ according to (2.29).

For the $k$th regression, the probability density function of $N(\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2(X_k^t X_k)^{-1})$ is an asymptotic confidence density function for $\boldsymbol{\theta}_k$. Here, $\hat{\boldsymbol{\theta}}_k$ is the least square estimate of $\boldsymbol{\theta}_k$, $X_k$ is the design matrix of $k$th regression, and $\hat{\sigma}_k^2$ is a consistent estimate of $\sigma_k^2$. We can directly use $N(\hat{\boldsymbol{\theta}}_k, \hat{\sigma}_k^2(X_k^t X_k)^{-1})$ to make inference about $\boldsymbol{\theta}_k$. This gives the individual approach. By applying $i$Fusion targeting at individual-$k$ we expect more efficient combined estimator $\hat{\boldsymbol{\theta}}_k^{(c)}$ and confidence intervals than the individual approach provides. Since $K = 6000$ is very large, we apply the pre-screen introduced in Remark 2.1 to filter out 99.5% individual subjects in implementing $i$Fusion.

Table 2.2 shows the summary statistics of the individual, $i$Fusion, and oracle approaches, based on 500 simulated datasets. The following summary statistics are reported for each approach: MSE of the point estimates, empirical coverage probability and median width of the confidence intervals at the 95% nominal level, of $\alpha_k$ and $\beta_k$, for $n_k = 40$ and 400. We choose result for individuals- 1500, 3000, and 4500 as representatives of the entire 6000 individuals. We note very similar patterns observed in

| | | $n_k = 40$ | | | | $n_k = 400$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Indiv | $i$Fusion | Oracle | NPB | Indiv | $i$Fusion | Oracle | NPB |
| MSE | $\alpha_{1500}$ | 0.026 | 0.007 | 0.005 | 0.005 | 0.002 | 0.0005 | 0.0005 | - |
| | $\beta_{1500}$ | 0.014 | 0.003 | 0.002 | 0.002 | 0.001 | 0.0002 | 0.0002 | - |
| | $\alpha_{3000}$ | 0.026 | 0.007 | 0.005 | 0.007 | 0.003 | 0.0005 | 0.0005 | - |
| | $\beta_{3000}$ | 0.011 | 0.003 | 0.003 | 0.003 | 0.001 | 0.0002 | 0.0002 | - |
| | $\alpha_{4500}$ | 0.028 | 0.007 | 0.006 | 0.006 | 0.002 | 0.0005 | 0.0005 | - |
| | $\beta_{4500}$ | 0.018 | 0.004 | 0.003 | 0.003 | 0.001 | 0.0002 | 0.0002 | - |
| Coverage | $\alpha_{1500}$ | 0.940 | 0.922 | 0.928 | 0.964 | 0.948 | 0.948 | 0.948 | - |
| | $\beta_{1500}$ | 0.944 | 0.928 | 0.930 | 0.972 | 0.950 | 0.966 | 0.966 | - |
| | $\alpha_{3000}$ | 0.932 | 0.922 | 0.930 | 0.920 | 0.944 | 0.952 | 0.952 | - |
| | $\beta_{3000}$ | 0.950 | 0.914 | 0.924 | 0.938 | 0.952 | 0.952 | 0.952 | - |
| | $\alpha_{4500}$ | 0.934 | 0.916 | 0.932 | 0.948 | 0.960 | 0.948 | 0.948 | - |
| | $\beta_{4500}$ | 0.940 | 0.930 | 0.938 | 0.940 | 0.942 | 0.956 | 0.956 | - |
| Width | $\alpha_{1500}$ | 0.624 | 0.278 | 0.271 | 0.277 | 0.196 | 0.088 | 0.088 | - |
| | $\beta_{1500}$ | 0.468 | 0.180 | 0.175 | 0.180 | 0.131 | 0.058 | 0.058 | - |
| | $\alpha_{3000}$ | 0.630 | 0.285 | 0.277 | 0.284 | 0.196 | 0.087 | 0.087 | - |
| | $\beta_{3000}$ | 0.415 | 0.183 | 0.178 | 0.181 | 0.125 | 0.057 | 0.057 | - |
| | $\alpha_{4500}$ | 0.617 | 0.277 | 0.271 | 0.276 | 0.197 | 0.088 | 0.088 | - |
| | $\beta_{4500}$ | 0.531 | 0.217 | 0.212 | 0.216 | 0.127 | 0.059 | 0.059 | - |

Table 2.2: Results of Simulation II comparing individual, $i$Fusion, oracle, and nonparametric Bayes approaches: MSE of point estimates, empirical coverage and median width of 95% confidence intervals. The nonparametric Bayesian (NPB) approach is applied on a subset of individual datasets that have survived the pre-screen procedure of $i$Fusion. Due to the computing cost, the cases when $n_k = 400$ are not run for the NPB approach.

Tables 3.1. In all cases, $i$Fusion returns point estimates with significantly less MSE and narrower confidence intervals than the individual approach. For $n = 40$, $i$Fusion approximates the performance of the oracle approach, and for $n = 400$, $i$Fusion and the oracle combination yield almost the same results. This has again numerically demonstrated the theoretical results in Section 2.3, i.e., Theorem 2.1, under the setting of multivariate parameters and a large number of individual subjects.

In addition, in this study we carry out a nonparametric Bayesian approach to make individualized inference on the target parameters. We use the `DPlmm` function in the `R` package `DPpackage` by (Jara et al., 2011). The function estimates a linear mixed-effects model with a Dirichlet process mixture prior for the distribution of the random effects, and is suitable under our setting where both regression intercept and slope are treated as random effects. In each random simulation, the MCMC samples for the target parameter can be extracted, and posterior mean and credible interval can be computed based these samples. Their frequentist properties can be then examined against the true target parameter values based on the 500 random simulations.

However, this nonparametric Bayesian approach is extremely time-consuming even for a single random simulation, because it simultaneously estimates all the individual parameters rather than a specific target individual parameter. The computation lasts forever in a Late 2013 MacBook Pro with a 2.4 GHz Intel Core i5 processor (10000 MCMC iterations for a single random run). As a compromise, we limit on a subset of data that have survived the pre-screen embeded in $i$Fusion. We emphasize that this practice disobeys the Bayesian principle because all the data and parameters shall be processed as one coherent whole in a Bayesian approach, but by all means it works a practical solution to such big data size.

We perform the analysis for the scenarios when $n_k = 40$, based on reduced data that contains only 30 individuals filtered by the pre-screen rule of $i$Fusion. (The analysis for $n_k = 400$ is not run due to the computing limit.) In random simulation, the last 2500 of the total 10000 MCMC samples are used to compute posterior mean and credit intervals. (Despite a much smaller sample size, it still takes around 6 minutes for a single run; for comparison $i$Fusion takes less a second for the same single run.) From Table 2.2, the reported MSEs are very close to or the same as the oracle approach, so it is with the median width of the credible intervals. In terms of coverage probabilities, the other three approaches are even slightly undeforming the nonparametric Bayesian approach (due to the small sample size). The price paid for the nonparametric Bayesian approach to produce comparable outputs to $i$Fusion, however, is far more computing time (in order of magnitude).

**Simulation III.** In the third simulation study, we generate $K = 4$ regression datasets from (2.24) with (similar to the simulation settings in Liu et al. (2015)) to verify Theorem 2.3 and Corollary 2.1 in Section 2.4. For the $k$th regression, $x_{ik}$ is 1 or 0 with equal probability, $z_{ik}$ has three levels of 1, 2, and 5, and each level is assigned with roughly $n_k/3$ observations.

The regression parameters are $\alpha_1 = -1+U_{11}/n_k, \alpha_2 = U_{21}/n_k, \alpha_3 = 1+U_{31}/n_k, \alpha_4 =$

| | | $n_k = 40$ | | | $n_k = 400$ | | |
|---|---|---|---|---|---|---|---|
| | | Indiv | $i$Fusion | Oracle | Indiv | $i$Fusion | Oracle |
| MSE | $\alpha_1$ | 0.097 | 0.059 | 0.056 | 0.011 | 0.005 | 0.005 |
| | $\beta_1$ | 0.136 | 0.046 | 0.037 | 0.011 | 0.004 | 0.003 |
| | $\gamma_1$ | 0.008 | 0.004 | 0.003 | 0.001 | 0.0003 | 0.0003 |
| | $\alpha_4$ | 0.096 | 0.096 | 0.096 | 0.012 | 0.012 | 0.012 |
| | $\beta_4$ | 0.102 | 0.102 | 0.102 | 0.010 | 0.010 | 0.010 |
| | $\gamma_4$ | 0.009 | 0.009 | 0.009 | 0.001 | 0.001 | 0.001 |
| Coverage | $\alpha_1$ | 0.940 | 0.940 | 0.946 | 0.928 | 0.946 | 0.946 |
| | $\beta_1$ | 0.922 | 0.922 | 0.938 | 0.946 | 0.952 | 0.954 |
| | $\gamma_1$ | 0.940 | 0.930 | 0.934 | 0.950 | 0.950 | 0.950 |
| | $\alpha_4$ | 0.960 | 0.960 | 0.960 | 0.944 | 0.944 | 0.944 |
| | $\beta_4$ | 0.944 | 0.944 | 0.944 | 0.954 | 0.954 | 0.954 |
| | $\gamma_4$ | 0.930 | 0.930 | 0.930 | 0.942 | 0.942 | 0.942 |
| Width | $\alpha_1$ | 1.187 | 0.902 | 0.896 | 0.404 | 0.282 | 0.282 |
| | $\beta_1$ | 1.343 | 0.738 | 0.724 | 0.393 | 0.226 | 0.226 |
| | $\gamma_1$ | 0.339 | 0.213 | 0.210 | 0.116 | 0.067 | 0.067 |
| | $\alpha_4$ | 1.313 | 1.313 | 1.313 | 0.413 | 0.413 | 0.413 |
| | $\beta_4$ | 1.256 | 1.256 | 1.256 | 0.392 | 0.392 | 0.392 |
| | $\gamma_4$ | 0.366 | 0.366 | 0.366 | 0.114 | 0.114 | 0.114 |

Table 2.3: Results of Simulation III comparing individual, $i$Fusion, and oracle approaches: MSE of point estimates, empirical coverage and median width of 95% confidence intervals.

$2 + U_{41}/n_k, \beta_1 = 1 + U_{12}/n_k, \beta_1 = 1 + U_{22}/n_k, \beta_3 = 1 + U_{32}/n_k, \beta_4 = -1 + U_{42}/n_k, \gamma_1 = -1 + U_{13}/n_k, \gamma_2 = -1 + U_{23}/n_k, \gamma_3 = -1 + U_{33}/n_k, \gamma_1 = -1 + U_{43}/n_k$, where $U_{kj} \overset{\text{i.i.d.}}{\sim} U[-1, 1]$ for $k = 1, \dots, 4$ and $j = 1, 2, 3$. The configuration follows Example 2.1 in Section 2.4: $(\beta_k, \gamma_k)$ is approximately the same (up to a constant of order $O(1/n_k)$) for $k = 1, 2, 3$; thus individuals- 1,2,3 form a clique according to (2.25) and individual-4 stand alone as a clique, where the cliques are defined based on $(\beta_k, \gamma_k)$ but not $\alpha_k$. We set both individual-1 and individual-4 to be the target Also, let $\sigma_k \equiv 1$, and $n_k \equiv 40$ or 400. For the oracle approach, we set $(w_{11}, w_{12}, w_{13}, w_{14}) = (1, 1, 1, 0)$ and $(w_{41}, w_{42}, w_{43}, w_{44}) = (0, 0, 0, 1)$.

Table 3.2 reports the summary statistics of the individual, $i$Fusion, and oracle approaches, based on 500 random simulations. For individual-1 with $|\mathcal{C}_1| = 3$, $i$Fusion outperforms the individual approach in two regards. First, inference about $\beta_1$ and $\gamma_1$ from $i$Fusion is much more efficient than the individual approach and approximates the oracle approach, with smaller MSEs and width of confidence intervals. These observations agree with Theorem 2.3. A second and more appealing result is on $\alpha_1$, for which $i$Fusion also yields much smaller MSE than the individual approach does, even though $\alpha_1$ is quite different from $\alpha_k$, $k = 2, 3, 4$. This numerically demonstrates the argument in Corollary 2.1, where improvement in estimating $\beta_1$ and $\gamma_1$ is transferred to the estimation of $\alpha_1$. For individual-4 that forms a clique itself, inference from the three approaches are the same. Besides, for all the three approaches and both sample sizes, the empirical coverage probability of the confidence intervals are around the desired nominal level.

## 2.7 Real Data Example

In asset pricing and portfolio management the Fama-French three-factor model is widely used to describe portfolio returns (Fama and French, 1993). It can be expressed in the form of regression:

$$r_{tk} = \alpha_k + b_k r_t^M + s_k \mathrm{SMB}_t + h_k \mathrm{HML}_t + \varepsilon_{tk}, \quad k = 1, \ldots, K. \tag{2.30}$$

Here, $r_{tk}$ is the excessive return on the $k$th portfolio over the risk-free rate at time $t$; $r_t^M$ is the excessive return on the market portfolio; $\mathrm{SMB}_t$ ("small minus big") is the return on a portfolio long small-capitalization stocks and short large-capitalization stocks; $\mathrm{HML}_t$ ("high minus low") is the return on a portfolio long high book-to-price stocks and short low book-to-price stocks (i.e., value stocks versus growth stocks); and $\varepsilon_{tk} \overset{\text{i.i.d.}}{\sim} N(0, \sigma_k^2)$ is the random error. They are calculated with combinations of portfolios composed by ranked stocks and available historical market data. See Grinold and Kahn (1999) for an elegant description on how the factors are constructed.

We download historical daily factor marks from Kenneth French's web page for all the trading days from 2016/01/01 to 2016/12/31. We also collect daily returns of 49 individual portfolios during the same period and from the same source. The portfolios are constructed using stocks listed in NYSE, AMEX, and NASDAQ and their four-digit SIC (Standard Industrial Classification) codes.

The Fama-French model is a powerful explanatory model in explaining the returns by contemporaneous factors and has been empirically validated across different markets;

cf. Fama and French (1993), Fama and French (2012), and Cakici et al. (2013). In this application, we expand it for prediction, that is, at time $t$, we use the factor values to predict the excessive portfolio return at time $t+l$:

$$r_{t+l,k} = \alpha_k + b_k r_t^M + s_k \mathrm{SMB}_t + h_k \mathrm{HML}_t + \varepsilon_{t+l,k}, \quad k = 1, \ldots, K. \tag{2.31}$$

Unlike the simulations in Section 2.6, the true parameter values, if any, are unobservable, hence it is impossible to measure the frequentist properties such as the coverage probability. Instead, we show the effectiveness of $i$Fusion by examining its impact on prediction, in the sense that a more efficiently estimated model typically leads to more accurate prediction. The test framework goes as follows:

1. For the $k$th portfolio, fit model (2.31) using factors and portfolio returns from time $t = 1$ to $m$. Let $\boldsymbol{\theta}_k = (\alpha_k, b_k, s_k, h_k)^t$ and denote by $\hat{\boldsymbol{\theta}}_k$ its least square estimate.

2. Apply $i$Fusion, similar to the procedure of Simulation II, to obtain an "improved" estimate of the regression coefficient $\hat{\boldsymbol{\theta}}_k^{(c)}$.

3. Predict $r_{m+1+l,k}$ given the factors at time $t = m+1$ using either $\hat{\boldsymbol{\theta}}_k$ or $\hat{\boldsymbol{\theta}}_k^{(c)}$, respectively. Calculate the prediction error in terms of difference between the predicted return and the actual return.

4. Repeat Steps 1 to 3 on a rolling basis with a window size $m$, i.e., on $[1, m]$, $[2, m+1]$ and so forth until the last applicable window. Average the prediction errors to obtain the rolling mean squared prediction error (RMSPE),

   $$\mathrm{RMSPE}_k = \frac{1}{S} \sum_{s=1}^{S} (\hat{r}_{m+s+l,k} - r_{m+s+l,k})^2,$$

   for both the individual approach and $i$Fusion. Here, $S$ is the number of available rolling windows.

5. Repeat Steps 1 to 4 for each of the 49 portfolios.

We test for different combinations of rolling window size ($m = 20, 60$) and prediction step ($l = 1, 2, 3$). Note that $m = 20$ is roughly the number of trading days in a month and $m = 60$ in three months. The results are presented in Figure 2.3. Each panel corresponds to a specific combination of $m$ and $l$; and each point is the relative PRMSE, that is, the ratio of RMSPE from $i$Fusion over that from the individual approach, for a portfolio. From the upper-left panel, it is noted that $i$Fusion improves the prediction for almost all the individual portfolios with exception of deterioration for only three of them. For those improved prediction, $i$Fusion reduces RMSPE mostly by more than five percent. This is the similar observation for different prediction steps from the rest two upper panels. We attribute such improvement in prediction accuracy to the improvement in parameter estimation. Notice that there is always an irreducible error associated with a future observation so the reduction of RMSPE is usually not as much as the reduction of MSE (on the parameter estimates) in the simulation examples. Based on the lower panels which use three months' data, reduction in RMSPE is also clear except for only a few portfolios. The improvement on average, however, is not as prominent as in the upper panels. This agrees with our expectation that for fixed number of parameters, the individual approach tends to perform better and differs less from $i$Fusion as the individual sample size increases.

Also under consideration is the Fama-French five-factor model (Fama and French, 2014), where two new factors, investment and profitability, from the dividend discount model, are added. The five-factor model, aimed at capturing the size, value, profitability, and investment patterns in average stock returns, is shown to be superior to the three-factor model in that it lessens the anomaly average returns left unexplained (Fama and French, 2014). The five-factor model for prediction can be expressed as

$$r_{t+l,k} = \alpha_k + b_k r_t^M + s_k \mathrm{SMB}_t + h_k \mathrm{HML}_t + r_k \mathrm{RMW}_t + c_k \mathrm{CMA}_t + \varepsilon_{t+l,k}, \quad k = 1, \ldots, K. \tag{2.32}$$

Here, $\mathrm{RMW}_t$ is the excessive return of the most profitable firms minus the least profitable, and $\mathrm{CMA}_t$ is excessive return spread of firms that invest conservatively minus aggressively.

In analogy to Figure 2.3, Figure 2.4 presents the relative RMSPE for each portfolio

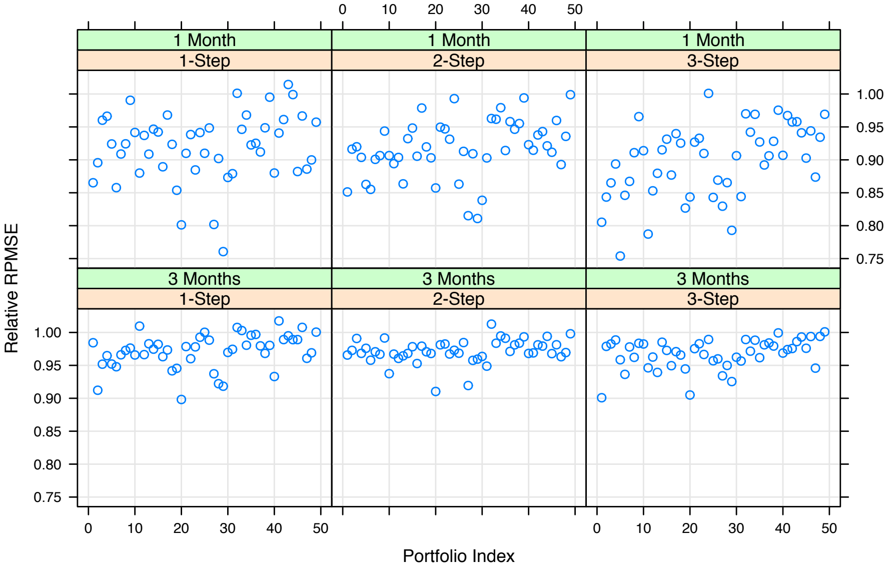

Figure 2.3: $h$-step-ahead PMSPE of $i$Fusion in relative to individual approach, using three-factor model (2.31) with $m$ samples for each portfolio, for $h = 1, 2, 3$, $m = 20, 60$.

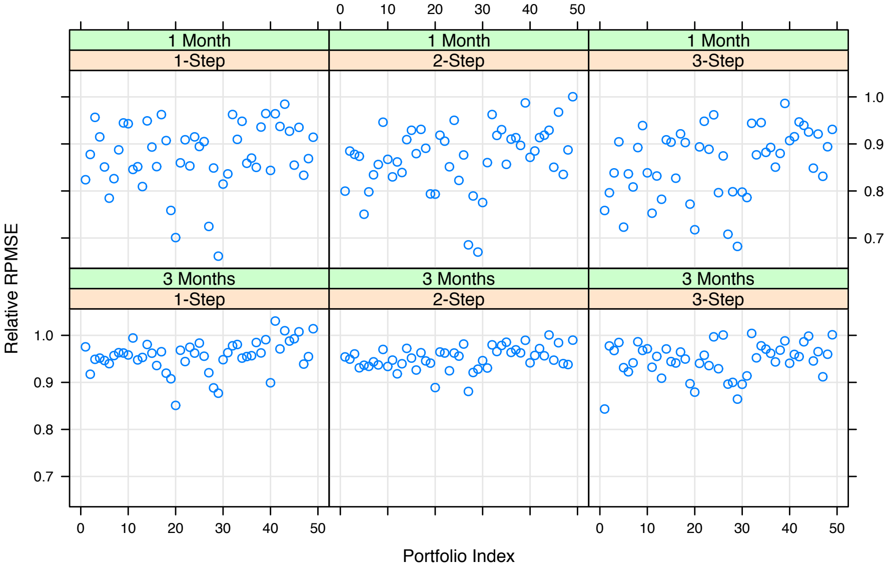

Figure 2.4: $l$-step-ahead RMSPE of $i$Fusion in relative to individual approach, using five-factor model (2.32) with $m$ samples for each portfolio, for $l = 1, 2, 3$, $m = 20, 60$.

for various window sizes and prediction steps. We observe very alike pattern as in Figure 2.3. Once again, the most important take-away is that improvement in prediction accuracy is channeled through the improvement in parameter estimation brought by $i$Fusion.

## 2.8 Further Comments

The key lesson throughout the chapter is the bias-variance tradeoff. Analysis with the fully aggregated data reduces variance but could bias the inference because not all the individuals shares the same/similar underlying truth with respect to the target. On the other hand, inference using only the individual specific data is unbiased (or yields low bias) but could suffer from high variability if its sample size is small. Without additional individual data, the smart way to better infer about the target is to include information from relevant individuals. Inclusion of similar individuals with small biases reduces variance at little cost; inclusion of an individual should be avoided if the bias introduced cannot be overcome by the combined sample size.

$i$Fusion is closely related to meta-analysis and information synthetization using CDs that commonly assumes that $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_K$. As with these methods, it is important to determine in advance if the combination is “suitable and needed”, which is often subject to comprehensive domain knowledge. A variety of quantitative measures have been developed and applied to assess the between-study inconsistency/heterogeneity, such as Cochrans Q (Cochran, 1954), $I^2$ statistic (Higgins et al., 2002, 2003). The “combine or not combine” question is then answered based on subjective judgements and/or the aforementioned quantitative evidence. The adaptive screening weights in $i$Fusion functions similarly as an objective rule on “to combine or not combine”, but are more mathematically rigorous and optimized. It aims at combined inference that is both efficient and valid, the latter of which is largely taken for granted and lacks investigation or validation in practice. Beyond that, $i$Fusion provides additional flexibility under finite sample size in the sense that i) the individuals to be combined do not have to be equally important if a non-uniform kernel function is used, and that ii) the true parameter values within a clique can vary within a specified range as measured in order of magnitude.

The intuition behind $i$Fusion makes it a natural fit for personalized medicine, among many other goal-directed applications. Personalized medicine, sometimes termed precision medicine, is a general terminology that describes a medical procedure/treatment tailor made to the individual patient rather than an "average patient". Conventionally it is accompanied by the subgroup analysis, in which patients are divided into subgroups by one or few baseline characteristics and further analysis is conducted within each subgroup. Essentially it partitions the individuals on its feature space and has natural interpretation, but has no statistical guarantee on the combined inference of model parameters within the subgroup. In comparison, $i$Fusion directly operates on the parameter space and is statistically valid. In practice, with an amount of data, the two procedures can be even jointly used and merits from both sides are preserved; say, partition the individuals into different subgroups according to their features, and then carry out $i$Fusion within the subgroup.

Another attractiveness of $i$Fusion we would like to emphasize is its scalability to big data problems, especially when compared to a Bayesian approach. Developed under the frequentist framework, $i$Fusion allows construction of confidence density functions independently for each individual, without over-worrying about other individuals and any nuisance or less relevant information. This reflects the so-called "division of labor" feature described in Efron (1986) and Wasserman (2007). They also pointed out that in a Bayesian approach, "statistical problems need to be solved as one coherent whole, including assigning priors and conducting analyses with nuisance parameters," and argued that a Bayesian approach is 'not good at division of labor." Compared to the full Bayesian approaches which requires running a large-scale simulation using an MCMC algorithm, such "divide-and-conquer" feature of $i$Fusion makes it scale much better to large applications.

$i$Fusion is a general statistical inference framework that can be applied to a wide range of problems provided the availability of individual (asymptotic) confidence density functions. The numerical examples in Sections 2.6 and 2.7 have demonstrated the effectiveness of $i$Fusion for simple models like linear regression. It is also readily applicable to the inference of more complex models such as time series models, survival models, and high dimensional models. For instance, consider a set of high-dimensional linear regressions corresponding to multiple individual subjects/datasets. Asymptotic confidence densities for the individual regression coefficients can be obtained by the debiased lasso procedure (cf., Javanmard and Montanari 2014, Zhang and Zhang 2014, and van de Geer et al. 2014), and then for a target individual, combined estimate and inference about the target individual regression coefficients can be obtained through $i$Fusion. This extends the divide-and-conquer strategies for inferring high dimensional regression coefficients with multiple datasets (cf. Chen and Xie 2014; Kleiner et al. 2014; Battet et al. To appear; Tang et al. 2016) from an overall to individualized perspective.

We close our discussion by remarking that "the idea of individualized inference is hardly new" (Liu and Meng, 2016) and has been researched under many different names (cf. Cox 1958; Fisher 1959; Berger and Wolpert 1988; Fraser 2004); it is the emergence of big data with the growing capability of data collection, computing power and storage that has renewed vitality of this area and brought many new opportunity as well as challenges. We hope that $i$Fusion will be the right way for both researchers and practitioners who are seeking a statistically efficient, reliable, and computationally scalable inference tool in the era of big (and heterogeneous) data.

# Chapter 3

# Prediction with Confidence

## 3.1 Background and Motivation

Consider the task of predicting the future value of a univariate random variable $Y^*$, based on given samples of size $n$, $\mathbf{Y}_n \equiv \{Y_1, Y_2, \ldots, Y_n\}$. Assume that the vector of the given sample data are from a distribution $G_\theta(\cdot)$ with parameter $\theta$, denoted by $\mathbf{Y}_n \sim G_\theta$, and that the new data point to be predicted is from a distribution $F_\theta(\cdot)$ with the same parameter $\theta$, denoted by $Y^* \sim F_\theta$. Since $G_\theta$ and $F_\theta$ share the same $\theta$, information contained in the observed data $\mathbf{Y}_n$ can be channeled through an estimate of $\theta$ to assist the prediction of $Y^*$. To simplify our presentation, we assume that $Y^*$ and $\mathbf{Y}_n$ are independent, except in Section 3.6 with an example that allows dependence between $Y^*$ and $\mathbf{Y}_n$. Throughout the chapter, the realization of $Y^*$ and $\mathbf{Y}_n$ are denoted by $y^*$ and $\mathbf{y}_n = \{y_1, \ldots, y_n\}$, respectively. Also, when they exist, the corresponding density functions of $F_\theta$ and $G_\theta$ are denoted by $f_\theta$ and $g_\theta$, respectively.

There is a rich literature on predictive inference. Lawless and Fredette (2005) provided an excellent overview on the topic and categorized statistical methods for prediction into two main approaches – frequentist and Bayesian.

I) In frequentist approach, prediction intervals of the specific form $(L_1(\mathbf{Y}_n), L_2(\mathbf{Y}_n))$ are considered, so that the coverage probability

$$\mathrm{CP} \equiv \mathbb{P}_{\mathbb{J}} \left\{ L_1(\mathbf{Y}_n) \leq Y^* \leq L_2(\mathbf{Y}_n) \right\} \tag{3.1}$$

can be specified, exactly or asymptotically. Here, $\mathbb{P}_{\mathbb{J}}$ refers to the joint probability for both random variables $Y^*$ and $\mathbf{Y}_n$. Relevant references include Aitchison and

Dunsmore (1980), Cox (1975), Beran (1990), Barndor-Nielsen and Cox (1996), and Escobar and Meeker (1999), among others.

II) In Bayesian inference, Bayesian predictive distributions of the form

$$Q_B(y^*; \mathbf{y}_n) = \int_{\theta \in \mathbf{\Theta}} F_\theta(y^*) p(\theta | \mathbf{y}_n) d\theta \tag{3.2}$$

are considered, based on data $\mathbf{Y}_n = \mathbf{y}_n$ and a prior distribution for the model parameter $\theta$. Here, $\mathbf{\Theta}$ is the parameter space of $\theta$ and $p(\theta|\mathbf{y}_n)$ is the posterior density of $\theta$ given $\mathbf{Y}_n = \mathbf{y}_n$. Bayesian prediction intervals $(L_1(\mathbf{y}_n), L_2(\mathbf{y}_n))$ can then be obtained from (3.2). Relevant references include Aitchison (1975), Aitchison and Dunsmore (1980), Geisser (1993), Smith (1998) and others.

The classical frequentist approaches in I) have the advantage of having a precise and well defined frequentist probabilistic interpretation, analogous to that of "confidence intervals". But those prediction intervals use only two endpoints of the intervals to describe $Y^*$, and thus are not as informative or flexible as the entire predictive distribution produced by the Bayesian methods in II) (as well as the approach to be proposed in this chapter). This comparative observation is similar to that in comparing inference outcomes from confidence intervals versus confidence distributions (cf. Cox 2013; Xie 2013). Specifically, as stated in Cox (1958, 2013), one often has a sense that "when 95% confidence limits of a normal mean are found then, even if the parameter is outside the calculated range, it will not be too far outside." This sense cannot be captured by the definition of a 95% confidence interval, but can be clearly displayed by a confidence distribution. Similar case can be made for using a full-fledged distribution function to describe the prediction outcome, as to convey fuller the prediction outcome and also sufficiently flexible to admit all forms of prediction outcomes, e.g., point estimates, prediction intervals of all levels, etc.

The Bayesian approach in II) does use a distribution function to describe the prediction of $Y^*$, and enjoys the aforementioned "flexibility". But the Bayesian outcomes depend on the additional assumptions of priors. Lawless and Fredette (2005) pointed out that “objective Bayesian methods do not have clear probability interpretations in finite samples,” and “subjective Bayesian predictions have a clear personal probability interpretation but it is not generally clear how this should be applied to non-personal predictions or decisions.” In addition, many statistical models are developed under non-Bayesian framework and Bayesian predictive distribution methods are not a natural fit for the developments in such practices.

To overcome the above shortcomings of the Bayesian approach Lawless and Fredette (2005) studied frequentist predictive distribution functions in a special setting equipped with pivotal quantities, and referred to this as the pivotal method which actually dates back to Fisher's general approach of fiducial inference (Fisher, 1935). They further proved the superiority of the predictive distributions obtained from the pivotal method, as having a smaller average Kullback-Leibler divergence to the true distribution $f_\theta(y^*)$, over those from the simple plug-in approach by using $f_{\hat{\theta}}(y^*)$ to derive prediction intervals for all $\theta$. Here, $\hat{\theta} \equiv \hat{\theta}(\mathbf{y}_n)$ is the maximum likelihood estimate or any efficient estimate of $\theta$ based on the observed data. A related development is the fiducial predictive distributions studied by Wang et al. (2012), who provided a set of conditions under which the fiducial predictive distributions can be used to construct prediction intervals. The fiducial prediction intervals coincide with the exact pivotal-based intervals when available, and otherwise possess correct frequentist coverage asymptotically.

Following the concept of predictive distribution in Lawless and Fredette (2005), we propose in this chapter a rigorous definition of a predictive distribution function and develop a general approach for constructing a predictive distribution of $Y^*$ using a confidence distribution (CD) of the unknown parameter $\theta$. The resulting predictive distribution can account for both the variability from the future random variable $Y^*$ and that from estimating the unknown parameter $\theta$ using the sample $\mathbf{Y}_n$. It takes the same form as the Bayesian and fiducial predictive distribution functions and thus also enjoys the flexibility of being predictive distribution functions. More importantly, it is anchored on the idea to always provide prediction intervals with clear frequentist probability interpretations. This approach was also considered in Schweder and Hjort (2016) under the name of predictive confidence distribution, which also had a comparison with the Bayesian predictive distribution. In this chapter, we establish theoretical properties for the CD-based predictive distribution, including frequentist coverage probabilities of the prediction intervals and related efficiency and optimality properties. Moreover, we establish the connection of this approach to other existing prediction approaches. In particular, we show that under our formulation the frequentist predictive distribution functions derived from the pivotal method in Lawless and Fredette (2005), the fiducial predictive distributions from Wang et al. (2012), and even the Bayesian predictive distribution all amount to the same equivalent expression. This clearly shows that the CD-based approach can provide a unifying platform linking the existing frequentist, Bayesian and fiducial predictive distribution functions.

The rest of this chapter is organized as follows. Section 3.2 defines predictive distribution functions and formulates a CD-based predictive approach. Section 3.3 examines the theoretical properties of the CD-based predictive distribution function and shows its connections to the Bayesian and fiducial predictive functions, and the frequentist predictive distribution function studied in Lawless and Fredette (2005). This section also presents several properties concerning the efficiency and optimality. Section 3.4 contains a simple yet broadly applicable Monte-Carlo algorithm for carrying out the CD-based approach. Section 3.5 demonstrates the effectiveness of the proposed CD-based approach using a simulation study under the linear and nonlinear regression models. Section 3.6 presents a real project on predicting the future volume of application submissions to a government agency, showing that the proposed approach applies even to settings with dependent observations. Section 3.7 provides further comments and discussions.

## 3.2 Predictive Distribution Function and Its General Formulation Based on CD

Let $\mathcal{Y}^*$ be the sample space of $Y^*$ and $\mathcal{Y}^n$ the sample space of $\mathbf{Y}_n$. Recall that $\mathbf{Y}_n \equiv \{Y_1, Y_2, \ldots, Y_n\} \sim G_\theta$ ; $Y^* \sim F_\theta$, and $\theta \in \mathbb{R}^p$ is the unknown parameter with parameter space $\mathbf{\Theta}$. Denote by $\theta_0$ the true parameter value of $\theta$. We define a predictive distribution function for $Y^*$ based on the sample data $\mathbf{Y}_n$ as follows.

**Definition 3.1.** *A function $Q(\cdot;\cdot)$ on $\mathcal{Y}^* \times \mathcal{Y}^n \longrightarrow (0,1)$ is called* A PREDICTIVE DISTRIBUTION FUNCTION FOR A NEW OBSERVATION $Y^*$ *if it satisfies the two requirements below:*

*R1) For each given $\mathbf{Y}_n = \mathbf{y}_n \in \mathcal{Y}^n$, $Q_{\mathbf{y}_n}(\cdot) = Q(\cdot;\mathbf{y}_n)$ is a cumulative distribution function on $\mathcal{Y}^*$;*

*R2) $Q(Y^*;\mathbf{Y}_n)$, as a function of both random sample $Y^*$ and $\mathbf{Y}_n$, satisfies the following equation:*

$$\mathbb{P}_{\mathbb{J}}(Q(Y^*;\mathbf{Y}_n) \leqslant \alpha) = \alpha, \quad \textit{for any } 0 < \alpha < 1, \tag{3.3}$$

*where $\mathbb{P}_{\mathbb{J}}(\cdot)$ is the joint probability measure w.r.t. $Y^*$ and $\mathbf{Y}_n$. Also, the function $Q(\cdot;\cdot)$ is called an asymptotic predictive distribution if the statement in* (3.3) *holds asymptotically.*

Requirement R1) in Definition 3.1 implies that, in principle, any sample-dependent distribution function on the space of the future random variable $Y^*$ can be used to predict $Y^*$ (i.e., to describe the performance of $Y^*$). To draw meaningful prediction inference, the additional requirement R2) is imposed to ensure that the statements of our prediction have the desired frequentist interpretations. In particular, requirement R2) ensures that the coverage probability (CP for short) defined in (3.1) equals $\alpha$, $0 < \alpha < 1$, for $L_1(\mathbf{Y}_n) = Q^{-1}_{\mathbf{Y}_n}(\alpha/2)$ and $L_2(\mathbf{Y}_n) = Q^{-1}_{\mathbf{Y}_n}(1-\alpha/2)$.

Note that Definition 3.1 of prediction functions bears striking resemblance to the definition of confidence distributions (CDs), except that the parameter $\theta$ and the corresponding parameter space $\mathbf{\Theta}$ in CDs are now replaced, respectively, by the "future observation" $Y^*$ and its sample space $\mathcal{Y}^*$. More precisely, a sample-dependent function defined on the parameter space $\mathbf{\Theta}$ is called a CD for $\theta$ if it satisfies the following two requirements: R1$^c$) For each given sample, it is a distribution function on the parameter space $\mathbf{\Theta}$; R2$^c$) It can provide confidence intervals (regions) of all levels for $\theta$; cf. Xie and Singh (2013), Schweder and Hjort (2016) and references therein. See also Schweder and Hjort (2002) and Singh et al. (2005) for a formal definition of CD. In general, a CD is a distribution estimate, instead of the usual point or interval estimate, of the parameter of interest.

The statement of Definition 3.1 is an abstract definition without concrete procedures for constructing predictive distribution functions. We exploit the similarities between the concepts of CDs and predictive distributions, in terms of their capability of summarizing information and quantifying uncertainty, to devise a precise formulation based on CD for constructing predictive distribution functions.

As stated in Cox (2013), a CD provides "a simple and interpretable summary of what can reasonably be learned from data (and an assumed model)." It quantifies both the information and uncertainty about the parameter $\theta$ from the observed data, and thus should naturally be first and key ingredient for constructing a predictive distribution function for $Y^*$. This link of CDs to the construction of predictive distributions will later be seen as desirable in many practices. More specifically, for a given CD for $\theta$ derived from the data $\mathbf{y}_n$, denoted by $H_n(\cdot) = H(\cdot; \mathbf{y}_n)$, we can apply the formula below to obtain a predictive distribution function:

$$Q(y^*; \mathbf{y}_n) = \int_{\theta \in \boldsymbol{\Theta}} F_\theta(y^*) dH(\theta; \mathbf{y}_n). \tag{3.4}$$

In Schweder and Hjort (2016) the same formula was also suggested along with some examples. Strictly speaking, $Q(y^*; \mathbf{y}_n)$ obtained by using (3.4) may not always satisfy requirement R2), but our theoretical results in Section 3.3 show that R2) holds exactly under some additional conditions on $F_\theta(y^*)$ and $H(\theta; \mathbf{y}_n)$ and asymptotically under mild conditions.

We now use two simple examples to illustrate the construction formula (3.4). The first example assumes i.i.d. sequence from the same distribution, but the second allows $Y^*$ and $Y_i$'s to have different distribution to show the flexibility and generality of the proposed approach. These two examples will serve as working examples for illustrating key steps in our development throughout the chapter.

**Example 3.1.** (*Normal distribution with known variance*) Let $Y_1, \ldots, Y_n$ and $Y^*$ be independent copies from $N(\theta, \sigma^2)$ with a known $\sigma^2$. A CD for $\theta$ based on the sample $\mathbf{y}_n$ is $N(\bar{y}, \sigma^2/n)$, where $\bar{y} = \sum_{i=1}^n y_i/n$ is the sample mean. This yields $F_\theta(y^*) = \Phi((y^* - \theta)/\sigma)$ and $H(\theta; \mathbf{y}_n) = \Phi((\theta - \bar{y})/(\sigma/\sqrt{n}))$. Thus, by (3.4), it follows immediately

$$Q(y^*; \mathbf{y}_n) = \int_{-\infty}^{\infty} \Phi\left(\frac{y^* - \theta}{\sigma}\right) d\Phi\left(\frac{\theta - \bar{y}}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{y^* - \bar{y}}{\sigma\sqrt{1 + 1/n}}\right). \tag{3.5}$$

Since $Q(Y^*; \mathbf{Y}_n) = \Phi((Y^* - \bar{Y})/(\sigma/\sqrt{1 + 1/n})) \sim \text{Uniform}(0, 1)$, the requirements in Definition 3.1 are satisfied. Note that this $Q(y^*; \mathbf{y}_n)$ is exactly the well-known predictive distribution $N(\bar{y}, \sigma^2(1 + 1/n))$ as well as the Bayesian predictive distribution with a flat prior for $\theta$.

**Example 3.2.** (*Exponential distribution*) Let $Y_1, \ldots, Y_n$ be independent copies from an exponential distribution with scale $\alpha\theta$ where $\alpha > 0$ is a known acceleration parameter (as in an accelerated life testing). Then, the joint density function of $\mathbf{Y}_n \equiv \{Y_1, Y_2, \ldots, Y_n\}$ is $g_\theta(\mathbf{y}_n) = (\alpha\theta)^{-n} e^{-n\bar{y}/(\alpha\theta)}$, where $\bar{y} = \sum_{i=1}^n y_i/n$ is the sample mean. Let $Y^*$ follow an exponential distribution with scale $\theta$, i.e., with the density function $f_\theta(y^*) = \theta^{-1} e^{-y^*/\theta}$. A CD for $\theta$ based on the sample $\mathbf{y}_n$ is $H_n(\theta) = H(\theta; \mathbf{y}_n) = 1 - \Gamma_{n,1}(n\bar{y}/(\alpha\theta))$, where $\Gamma_{n,1}(\cdot)$ is the cumulative distribution function of Gamma$(n, 1)$ distribution. With $F_\theta(y^*) = 1 - e^{-y^*/\theta}$, for $y^* > 0$, it follows from (3.4) with straightforward calculation that

$$Q(y^*; \mathbf{y}_n) = \int_0^{\infty} F_\theta(y^*) dH(\theta; \mathbf{y}_n) = 1 - \left\{1 + \frac{\alpha y^*}{n\bar{y}}\right\}^{-n}. \tag{3.6}$$

Clearly, the two requirements in Definition 3.1 hold, since $\alpha Y^*/\bar{Y}$ follows an $F$-distribution and $Q(Y^*; \mathbf{Y}_n) = \mathcal{F}_{2,2n}(Y^*/\bar{Y}) \sim \text{Uniform}(0, 1)$. Here, $\mathcal{F}_{2,2n}(t) = 1 - (1 + t/n)^{-n}$ is the cumulative distribution function of the $F$-distribution with degrees of freedom $(2, 2n)$. Note that this same predictive distribution can also be obtained using the Bayesian approach with the Jeffreys' prior $\pi(\theta) \propto 1/\theta$.

Note that there are many ways to derive a CD, say from, for instance, normalized likelihood, fiducial distribution, Bayesian posterior distribution, bootstrap distribution,

$p$-value function, among others; cf. Xie and Singh (2013) and references therein. The same paper also stated, “Any approach, regardless of being frequentist, fiducial or Bayesian, can potentially be unified under the concept of confidence distributions, as long as it can be used to build confidence intervals of all levels, exactly or asymptotically”. This useful property that CD can provide a unified framework to encompass inference procedures from different paradigms is readily inherited by the framework of predictive distribution functions. This makes formula (3.4) broadly applicable in many general settings, which will be further elaborated in the next sections.

## 3.3 Theoretical Properties

In this section, we investigate theoretical properties of the predictive distribution $Q(y^*; \mathbf{y}_n)$ constructed using formula (3.4). For ease of presentation, we focus on the case of scalar $\theta$ with $p = 1$ in this section. We will provide comments on extensions to the case of a multivariate $\theta$ with $p > 1$ at the end of the section.

The mean, the median and the mode of a CD $H_n(\cdot) = H(\cdot; \mathbf{y}_n)$ have been shown in Singh et al. (2007) to be consistent estimators of the unknown parameter $\theta$ under Condition (A) below:

(A) For any $\delta$, $0 < \delta < 1/2$, $L_n(\delta) = H_n^{-1}(1-\delta) - H_n^{-1}(\delta) \to 0$, in probability, as the sample size $n \to \infty$.

Later, Xie et al. (2011) proved that this is equivalent to Condition (A’) below:

(A’) For any fixed $\epsilon > 0$, $H_n(\theta_0 - \epsilon) \to 0$ and $H_n(\theta_0 + \epsilon) \to 1$, in probability, as $n \to \infty$,

where $\theta_0$ is the true value of $\theta$. These two conditions can be interpreted as: as the sample size $n$ increases, the probability mass of the CD $H_n(\theta)$ becomes more concentrated around $\theta_0$.

We establish the following theorem to show that, if $H_n(\theta)$ satisfies Condition (A) or (A’), then $Q(y^*; \mathbf{y}_n)$ in (3.4) is an asymptotic predictive distribution function for $Y^*$. Thus, $Q(y^*; \mathbf{y}_n)$ based on $H_n(\theta)$ has valid frequentist interpretations asymptotically. A proof of the theorem is given in Appendix.

**Theorem 3.1.** *Assume that the CD $H_n(\cdot)$ used for constructing the predictive function in (3.4) satisfies Condition (A), and also that $F_\theta(\cdot)$ is continuous in $\theta$ in a neighborhood of $\theta_0$:*

$$\sup_t |F_\theta(t) - F_{\theta_0}(t)| \leq C\,|\theta - \theta_0|\,, \tag{3.7}$$

*for some constant $C > 0$. Then,*

$$Q(Y^*; \mathbf{Y}_n) = U + o_p(1), \tag{3.8}$$

*where $U \sim$ Uniform$(0, 1)$.*

Theorem 3.1 ensures an asymptotic coverage in (3.3) for a broad range of settings, though in some cases such as in Examples 3.1 and 3.2, $Q(Y^*; \mathbf{Y}_n)$ follows exactly Uniform$(0, 1)$ independent of the sample size. Next, we provide a set of sufficient conditions, under which the predictive distribution $Q(Y^*; \mathbf{Y}_n)$ always has exact coverage probability. Specifically, consider a condition on the distribution function $F_{\theta_0}(y^*)$:

(I) Suppose that there exists a monotonic mapping $s_1 : \mathcal{Y}^* \times \boldsymbol{\Theta} \to \mathcal{Y}^*$ and a monotonic mapping $s_2 : \boldsymbol{\Theta} \times \boldsymbol{\Theta} \to \boldsymbol{\Theta}$ such that $F_{\theta_0}(y^*)$ is invariant to the transformations $s_1$ and $s_2$ in the sense that, for any $\theta \in \boldsymbol{\Theta}$,

$$F_{\theta_0}(y^*) = F_{s_2(\theta_0,\theta)}(s_1(y^*, \theta)). \tag{3.9}$$

Condition (I) is satisfied in both Examples 3.1 and 3.2. For instance, in Example 3.1, with $s_1(y^*, \theta) = y^* - \theta$, $s_2(\theta_0, \theta) = \theta_0 - \theta$ and $\mathcal{Y}^* \equiv \boldsymbol{\Theta} \equiv (-\infty, \infty)$, we can verify (3.9), since $F_{\theta_0}(y^*) = \Phi((y^* - \theta_0)/\sigma) = \Phi(\{(y^* - \theta) - (\theta_0 - \theta)\}/\sigma) = F_{\theta_0 - \theta}(y^* - \theta)$ for any $\theta \in (-\infty, \infty)$. Similarly, in Example 3.2, with $s_1(y^*, \theta_0) = y^*/\theta_0$, $s_2(\theta_0, \theta) = \theta_0/\theta$ and $\mathcal{Y}^* \equiv \boldsymbol{\Theta} \equiv (0, \infty)$, we immediately have (3.9), since $F_{\theta_0}(y^*) = 1 - e^{-y^*/\theta_0} = 1 - e^{-(y^*/\theta)/(\theta_0/\theta)} = F_{\theta_0/\theta}(y^*/\theta)$ for any $\theta \in (0, \infty)$.

Without loss of generality and to simplify our presentation, we assume from now on that $s_2(\theta_0, \theta)$ is increasing in $\theta_0$ and decreasing in $\theta$. Denote by $S_{\theta_0}(t) = \mathbb{P}_{\theta_0}(s_1(Y^*, \theta_0) \leq$

$t$) and $R_{\theta_0}(t) = \mathbb{P}_{\theta_0}(s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0) \leq t)$, where $\hat{\theta}(\mathbf{Y}_n)$ is the maximum likelihood estimate or some other efficient estimate of $\theta_0$ derived from the observed data. It follows immediately that $R_{\theta_0}(s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0)) \sim$ Uniform$(0, 1)$. If $s_1(Y^*, \theta_0)$ and $s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0)$, are pivotal quantities, then $S_{\theta_0}(t)$ and $R_{\theta_0}(t)$ are independent of $\theta_0$, and thus can be written as $S(t)$ and $R(t)$. In this case, a CD for $\theta_0$ can be obtained by

$$H_R(\theta; \hat{\theta}(\mathbf{y}_n)) = 1 - R(s_2(\hat{\theta}(\mathbf{y}_n), \theta)).$$

Following (3.4), a corresponding predictive distribution is

$$Q_R(y^*; \mathbf{y}_n) = \int_{\theta \in \boldsymbol{\Theta}} F_\theta(y^*) dH_R(\theta; \hat{\theta}(\mathbf{y}_n)). \tag{3.10}$$

The following theorem states that the function $Q_R(\cdot; \cdot)$ expressed in (3.10) is an exact predictive distribution function. This theorem covers a class of cases including Examples 3.1 and 3.2. The proof of the theorem is also given in Appendix.

**Theorem 3.2.** *Assume that condition (I) holds, and that* $s_1(Y^*, \theta_0)$ *and* $s_2(\hat{\theta}(\mathbf{Y}_n), \theta_0)$ *are pivotal quantities. Then,* $Q_R(Y^*; \mathbf{Y}_n) \sim$ *Uniform*$(0, 1)$.

The proposed CD-based prediction framework has broad implications. In particular, we present two corollaries which indicate that the CD-based prediction framework can be applied broadly to encompass several existing Bayesian, fiducial and frequentist prediction procedures.

First, note that the fiducial and Bayesian posterior distributions are sample-dependent distribution functions on the parameter space. If their corresponding fiducial or credible intervals have valid frequentist probability coverages (which is a goal in many developments on the topics of fiducial and (objective) Bayes), they satisfy the definition of CDs; cf. Xie and Singh (2013). In this context, Bayesian predictive distributions defined in (3.2) and the fiducial predictive distributions defined in Wang et al. (2012) are in fact the same as (or treated as special cases of) the general formulation (3.4). Thus, an immediate result from Theorems 3.1 and 3.2 is that the predictive intervals obtained from these fiducial and Bayesian predictive distributions have valid frequentist coverage. This observation is summarized as a corollary below.

**Corollary 3.1.** *If a Bayesian posterior or a fiducial distribution of $\theta$ can be justified as a CD, then its corresponding predictive distribution also has the valid frequentist probability coverage as defined in Definition 3.1.*

Note that the predictive distribution by the pivotal method of Lawless and Fredette (2005) can also be linked to the general formulation (3.4), even though it is quite different in appearance. The pivotal method relies on the random variable $W = F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*)$, which is required to be a pivotal quantity so that its cumulative distribution function $K(t) = \mathbb{P}_{\mathbb{J}}(W \leq t)$ is parameter-free. By defining our predictive distribution function as

$$Q_{\text{piv}}(y^*; \mathbf{y}_n) \equiv K(F_{\hat{\theta}(\mathbf{y}_n)}(y^*)), \tag{3.11}$$

we obtain the predictive distribution function proposed in Lawless and Fredette (2005). Clearly, $Q_{\text{piv}}(y^*; \mathbf{y}_n)$ satisfies the requirements in Definition 3.1. The next corollary states that $Q_{\text{piv}}(y^*; \mathbf{y}_n)$ can actually be expressed in the general formula (3.4). A proof of Corollary 3.2 can be found in Appendix.

**Corollary 3.2.** *Under the condition of Theorem 3.2, $Q_{piv}(y^*; \mathbf{y}_n)$ defined in (3.11) can be expressed as*

$$Q_{piv}(y^*; \mathbf{y}_n) = \int_{\theta \in \mathbf{\Theta}} F_\theta(y^*) dH_R(\theta; \hat{\theta}(\mathbf{y}_n)),$$

*where $H_R(\theta; \hat{\theta}(\mathbf{y}_n))$ is a CD obtained based on $\hat{\theta}(\mathbf{y}_n)$.*

Altogether, Corollaries 3.1 and 3.2 suggest that the general formulation of predictive distributions in (3.4) through CDs provides a common link or a unifying platform for most, if not all, existing frequentist, fiducial and Bayesian predictive distributions.

We next discuss two optimality results regarding choices of different predictive distribution functions. Like in the CD development where multiple CDs exist for the same estimation problem, there may also exist different predictive distribution functions for the same prediction problem. We start with reviewing some definitions and properties on the relative efficiency of CD and then make extensions to predictive distributions.

Following Singh et al. (2001), Schweder and Hjort (2002), Singh et al. (2007) and Xie and Singh (2013), a CD $H_1(\cdot) \equiv H_1(\cdot; \mathbf{Y}_n)$ is considered more efficient than another CD $H_2(\cdot) \equiv H_2(\cdot; \mathbf{Y}_n)$ at $\theta = \theta_0$, if for all $\varepsilon > 0$,

$$H_1(\theta_0 - \varepsilon) \overset{\text{sto}}{\leq} H_2(\theta_0 - \varepsilon) \quad \text{and} \quad 1 - H_1(\theta_0 + \varepsilon) \overset{\text{sto}}{\leq} 1 - H_2(\theta_0 + \varepsilon). \tag{3.12}$$

Here, the symbol $\overset{\text{sto}}{\leq}$ stands for stochastically less than or equal to. This definition has an equivalent form: $H_1(\cdot)$ is more efficient than $H_2(\cdot)$ at $\theta = \theta_0$, if for all $u \in (0, 1)$,

$$(H_1^{-1}(u) - \theta_0)^+ \overset{\text{sto}}{\leq} (H_2^{-1}(u) - \theta_0)^+ \quad \text{and} \quad (H_1^{-1}(u) - \theta_0)^- \overset{\text{sto}}{\leq} (H_2^{-1}(u) - \theta_0)^-. \tag{3.13}$$

The inequalities (3.12)) and (3.13) are interpreted in Singh et al. (2007) as that $H_1(\cdot)$ is more "concentrated" around the true parameter $\theta_0$ than $H_2(\cdot)$. A natural notion of MSE for a CD $H(\cdot)$ is defined in Xie and Singh (2013):

$$\text{MSE}(H) = \mathbb{E}_{\mathbf{Y}_n} \int_{\theta \in \mathbf{\Theta}} (\theta - \theta_0)^2 dH(\theta) = \mathbb{E}_{\mathbf{Y}_n, U}(H^{-1}(U) - \theta_0)^2, \tag{3.14}$$

where $U \sim \text{Uniform}(0, 1)$. It follows by elementary exercise that

$$\text{MSE}(H_1) \leq \text{MSE}(H_2), \tag{3.15}$$

if $H_1(\cdot)$ is more efficient than $H_2(\cdot)$ at $\theta = \theta_0$. In addition to the MSE of CD, the discussions on the optimality issues surrounding CDs in Xie and Singh (2013) and Schweder and Hjort (2016) indicate that a better CD typically leads to a better point estimator and hypothesis test, and vice versa.

The relative efficiency of predictive distributions can be defined by extending (3.13). Specifically, we say one predictive distribution $Q_1(\cdot; \mathbf{Y}_n)$ is more efficient than another predictive distribution $Q_2(\cdot; \mathbf{Y}_n)$, if for all $u \in (0, 1)$,

$$(Q_1^{-1}(u) - F_{\theta_0}^{-1}(u))^+ \overset{\text{sto}}{\leq} (Q_2^{-1}(u) - F_{\theta_0}^{-1}(u))^+, \tag{3.16}$$

and

$$(Q_1^{-1}(u) - F_{\theta_0}^{-1}(u))^- \overset{\text{sto}}{\leq} (Q_2^{-1}(u) - F_{\theta_0}^{-1}(u))^-. \tag{3.17}$$

Now, suppose that $Q_1(\cdot; \mathbf{Y}_n)$ and $Q_2(\cdot; \mathbf{Y}_n)$ are two predictive distribution functions induced through the general form (3.4) using two different CDs, $H_1(\cdot)$ and $H_2(\cdot)$. A natural question is whether a better CD leads to a better predictive distribution function. The following theorem provides an affirmative answer under a set of suitable conditions.

**Theorem 3.3.** *Suppose that $Q_i^{-1}(u) = \int_{\theta\in\boldsymbol{\Theta}} F_\theta^{-1}(u) dH_i(\theta)$ for $i = 1, 2$, and that $F_\theta^{-1}(u)$ is nondecreasing in $\theta$, for any given $u \in (0, 1)$. If $H_1(\cdot)$ is more efficient than $H_2(\cdot)$ at $\theta = \theta_0$, then $Q_1(\cdot; \mathbf{Y}_n)$ is more efficient than $Q_2(\cdot; \mathbf{Y}_n)$.*

We can also define MSE of a predictive distribution $Q(\cdot; \mathbf{Y}_n)$ analogously:

$$\text{MSE}(Q) = \mathbb{E}_{\mathbf{Y}_n, U}(Q^{-1}(U) - F_{\theta_0}^{-1}(U))^2, \tag{3.18}$$

where $U \sim \text{Uniform}(0, 1)$. In essence, MSE$(Q)$ quantifies the expected squared deviation between the quantiles of $Q(\cdot; \mathbf{Y}_n)$ and $F_{\theta_0}(\cdot)$. The counterpart of (3.15) can be immediately established under the same setting of Theorem 3.3.

**Corollary 3.3.** *Under the setting of Theorem 3.3,*

$$MSE(Q_1) \leq MSE(Q_2). \tag{3.19}$$

Consider the setting of Example 3.1. Singh et al. (2007) showed that $H_1(\theta) = \Phi((\theta - \bar{Y})/(\sigma/\sqrt{n}))$ is the most efficient CD for $\theta_0$. We also consider a CD derived from the sample median $M$. Since $\sqrt{n}(M - \theta_0) \to N(0, \pi\sigma^2/2)$ in distribution, as $n \to \infty$, $H_2(\theta) = \Phi((\theta - M)/(\sigma/\sqrt{2n/\pi}))$ is an asymptotic CD for $\theta_0$. Although $H_2$ may be more robust, it is known to be less efficient than $H_1$. Applying (3.4), the predictive distribution functions based on $H_1$ and $H_2$ can be obtained. They are $Q_1(Y^*; \mathbf{Y}_n) = \Phi((Y^* - \bar{Y})/(\sigma/\sqrt{1 + 1/n}))$ and $Q_2(Y^*; \mathbf{Y}_n) = \Phi((Y^* - M)/(\sigma/\sqrt{1 + \pi/2n}))$, respectively. It is easy to verify that the requirement in Theorem 3.3 is satisfied. Thus, Corollary 3.3 implies that the MSE($Q_1$) is smaller than MSE($Q_2$). Indeed, simple algebra gives MSE$(Q_1) = \frac{2}{n}\sigma^2$ and MSE$(Q_2) \approx \frac{\pi}{n}\sigma^2$ for some large $n$.

If there is a family of uniformly most powerful unbiased (UMPU) tests for testing $K_0 : \theta \leq c$ versus $K_1 : \theta > c$, for every $c \in \mathbf{\Theta}$, Theorem 2.2 of Singh et al. (2007) states that the CD corresponding to the $p$-value function of the UMPU tests is the most efficient. Combining this observation with Theorem 3.3, we immediately have:

**Corollary 3.4.** *Under the setting of Theorem 3.3 and assume that a CD is derived from a p-value function of a UMPU test, then the corresponding predictive distribution function obtained by using (3.4) has the smallest MSE.*

Finally, we discuss the plug-in predictive distribution $F_{\hat{\theta}}(y^*)$ which has often been used as an approximation to the true distribution $F_{\theta_0}(y^*)$. Although the plug-in predictive distribution has valid asymptotic coverage probability similar to that of (3.8), it fails to account for the uncertainty in the estimation of $\theta$ and typically cannot achieve exact coverage probability in comparison with the result of Theorem 3.2. In fact, Lawless and Fredette (2005) showed that when the pivot method applies, the predictive distribution $Q_{\text{piv}}(y^*; \mathbf{y}_n)$ in (3.11) is always better than the plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, as measured by the average Kullback-Leibler divergence to the true distribution $F_{\theta_0}(y^*)$; cf. Theorem 1 of Lawless and Fredette (2005).

The next theorem reports a slightly more general result. Let $H_R(\theta; \hat{\theta}(\mathbf{y}_n))$ be a CD for $\theta$ obtained based on $\hat{\theta} = \hat{\theta}(\mathbf{y}_n)$. To simplify the notations, we let $Q_{\hat{\theta}}(t) = Q(t; \hat{\theta})$, $q_{\hat{\theta}}(t) = \frac{d}{dt}Q_{\hat{\theta}}(t)$ and $f_{\hat{\theta}}(t) = \frac{d}{dt}F_{\hat{\theta}}(t)$. The theorem below shows that the predictive distribution function $Q_{\hat{\theta}}(y^*) = Q(y^*; \hat{\theta}(\mathbf{y}_n))$ obtained using $H_R(\theta; \hat{\theta}(\mathbf{y}_n))$ is better than the naive plug-in predictive distribution function $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, as measured by the average Kullback-Leibler divergence to the true distribution $F_{\theta_0}(y^*)$. The proof is provided in Appendix.

**Theorem 3.4.** *Assume that*

$$\mathbb{E}_{\mathbb{J}} \left\{ \frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)} \right\} \leq 1. \tag{3.20}$$

*Then,*

$$\bar{D}_{KL}(f_{\theta_0}|q_{\hat{\theta}}) \leq \bar{D}_{KL}(f_{\theta_0}|f_{\hat{\theta}}),$$

*where* $\bar{D}_{KL}(f_{\theta_0}|g_{\hat{\theta}}) = \mathbb{E}_{\mathbb{J}}\left\{\log \frac{f_{\theta_0}(Y^*)}{g_{\hat{\theta}}(Y^*)}\right\}$ *is the average Kullback-Leibler divergence between* $f_{\theta_0}$ *and any density function of the form* $g_{\hat{\theta}}$*.*

In the pivot example in Lawless and Fredette (2005), $Q_{\rm piv}(y^*; \mathbf{y}_n) = K(F_{\hat{\theta}(\mathbf{y}_n)}(y^*))$. So, $q_{\hat{\theta}}(t) = \frac{\partial}{\partial t}Q_{\rm piv}(t; \mathbf{y}_n) = k(F_{\hat{\theta}(\mathbf{y}_n)}(t))f_{\hat{\theta}(\mathbf{y}_n)}(t)$, where $k(s) = \frac{\partial}{\partial s}K(s)$ is the density function corresponding to the cumulative distribution function $K(\cdot)$. It follows from direct calculation that $\mathbb{E}_{\mathbb{J}}\left\{\frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)}\right\} = \mathbb{E}_{\mathbb{J}}\{1/k(F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*))\} = \mathbb{E}_U\{1/k(U)\} = 1$, where the last two equations are obtained by variable transformation and the observation that $U = F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*) \sim k(\cdot)$. Thus, (3.20) holds and Theorem 3.4 covers the result of Lawless and Fredette (2005) as a special case.

We close this section by addressing the potential extensions of the above theoretic developments to a multivariate $\boldsymbol{\theta} \in \mathbb{R}^p$ setting with $p > 1$. (From now on we use the bold $\boldsymbol{\theta}$ whenever $p > 1$.) Although, on the outset, we note that Definition 3.1 of the predictive distribution, the general formulation (3.4) in Section 3.2 and even the algorithm to be proposed in Section 3.4 can be applied directly in the multivariate setting, there remains a technical difficulty in defining a general multivariate CD and thus a rigorous presentation of all theoretical results in Section 3.3 for the general multivariate $\boldsymbol{\theta}$ setting is still being sought. In principle, the concept of a multivariate CD is straightforward (i.e., a sample-dependent distribution function on the multivariate parameter space that can produce confidence regions of all levels), however a precise definition with explicit mathematical formulation to cover general cases thus far remains elusive. But partial progress can still be made, since under asymptotic settings or wherever the usual likelihood inference or bootstrap theory applies, multivariate CDs can be applied with ease. For instance, under the general setup of likelihood inference, the multivariate normal distribution $N(\hat{\boldsymbol{\theta}}, \hat{\Sigma})$ serves as a first-order asymptotic CD for $\boldsymbol{\theta}$ where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ and $\hat{\Sigma}$ is the inverse of the observed Fisher's information using the entire $n$ observations; cf. Yang et al. (2014) and Liu et al. (2015). In addition, if we limit ourselves to center-outwards confidence regions (instead of all

Borel sets) in the parameter space, concepts such as the $c$-CDs considered in Singh et al. (2007) and the confidence curve considered in Schweder (2007) and Schweder and Hjort (2016) offer coherent notions of multivariate CDs in the exact sense. In these cases, we still can generalize most of the theoretical developments to the multivariate setting. This fact has been used in some of our examples, e.g., in Section 3.6. See also Schweder and Hjort (2016) for related discussions. Last but not least, extension can be made to establish predictive distributions for a multivariate observation, that is, multivariate predictive distribution (MPD). MPD utilizes the concept of data depth (cf. Liu et al. 1999) to characterize its frequentist properties and is currently under development by our research group.

## 3.4 Algorithm for Simulating from Predictive Distributions

To implement the approach formulated in (3.4), we propose a Monte-Carlo algorithm for computing predictive distributions and prediction intervals. This algorithm is simple yet applicable to a wide range of problems. Specifically, given $\mathbf{Y}_n = \mathbf{y}_n$, a CD $H_n(\cdot) = H(\cdot; \mathbf{y}_n)$ is a distribution function on the parameter space $\boldsymbol{\Theta}$. Conditional on $\mathbf{Y}_n = \mathbf{y}_n$, we can simulate a CD-random variable $\theta_{\text{CD}}$ by $\theta_{\text{CD}} \big| \mathbf{y}_n \sim H_n(\cdot)$. The concrete algorithm is as follows.

[**Monte-Carlo Algorithm**] Obtain a simulated copy of $y_S^*$ from $Q(\cdot; \mathbf{y}_n)$ by: first simulate a CD-random variable $\theta_{\text{CD}} \big| \mathbf{y}_n \sim H_n(\cdot)$, and then simulate a $y_S^*$ from $y_S^* \big| \theta_{\text{CD}} \sim f_{\theta_{\text{CD}}}(\cdot)$. Repeat this procedure a large number of times, say $N$ times, to obtain $N$ copies of simulated $y_S^*$. The histogram of these $N$ copies of $y_S^*$ are then used to approximate a predictive distribution of $Y^*$ and hence its prediction intervals of all levels.

This algorithm applies to any CD $H_n(\cdot) = H(\cdot, \mathbf{y}_n)$ for $\boldsymbol{\theta} \in \mathbb{R}^p$. Note that, any approach, regardless being frequentist, fiducial or Bayesian, can be used to construct CDs, as long as the produced CDs can be used to build confidence intervals of all levels, exact or asymptotically; cf. Xie and Singh (2013) and references therein. Hence this algorithm is quite general and can be applied broadly.

As a special case, this algorithm can be carried out using a bootstrap method, noting that a bootstrap distribution is known to be also a CD (see e.g., Efron 1998; Xie and Singh 2013). In particular, we can simply simulate a future observation $y^*$ by $y^*|\theta_{\text{boot}} \sim f_{\theta_{\text{boot}}}(\cdot)$, where $\theta_{\text{boot}}$ is the bootstrap estimate of the parameter $\theta$. Obviously, this simulation method makes the proposed prediction approach very useful in practice, as it is simple and general.

Clearly, the prediction intervals and predictive distributions obtained by using the proposed algorithm above have valid frequentist interpretations, following Theorems 3.1 and 3.2 (and their extensions to multivariate $\theta$ as discussed at the end of Section 3.3).

## 3.5 Simulation

In this section, we use two simulation examples to demonstrate the proposed approach and computing algorithm for constructing predictive distribution functions, and then examine their frequentist properties. The first example involves a simple linear regression model with zero-intercept, for which exact predictive distribution function is well-known and can be obtained explicitly. We report and compare the numerical results from this explicit predictive distribution function and those from our computing algorithm. The second example relates to a nonlinear regression, for which an exact CD for the underlying parameter does not exist, neither does an exact predictive distribution function. Nonetheless, we can apply our computing algorithm with several different asymptotic CD functions to perform predictions and study their numerical performance.

**Simulation I.** Consider a simple linear regression model with zero-intercept:

$$y_i = \theta x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. Let $\hat{\theta} = \sum_{i=1}^n y_i x_i / \sum_{i=1}^n x_i^2$ be the ordinary least squares estimate of $\theta$ and $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{\theta} x_i)^2$. For a new independent observation $Y^*$

associated with covariate $x^*$, there exists the well-known predictive distribution

$$Q_t(y^*; \mathbf{y}_n) = T_{n-1}\left(\frac{y^* - \hat{\theta}x^*}{\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}}\right), \tag{3.21}$$

where $T_{n-1}(\cdot)$ is the cumulative distribution function of $t$-distribution with degrees of freedom $n-1$. It is easy to verify that $Q_t(y^*; \mathbf{y}_n)$ satisfies Definition 3.1. This is the same as the predictive distribution considered in Schweder and Hjort (2016).

Alternatively, straightforward calculation yields

$$H(\theta) = \Phi\left(\frac{\theta - \hat{\theta}}{\hat{\sigma}/\sqrt{\sum_{i=1}^n x_i^2}}\right)$$

as an asymptotic CD for $\theta$. The corresponding predictive distribution for $Y^*$ using formula (3.4) is then

$$Q(y^*; \mathbf{y}_n) = \int_{-\infty}^{\infty} \Phi\left(\frac{y^* - \xi x^*}{\sigma}\right) d\Phi\left(\frac{\xi - \hat{\theta}}{\sigma/\sqrt{\sum_{i=1}^n x_i^2}}\right) = \Phi\left(\frac{y^* - \hat{\theta}x^*}{\sigma\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}}\right),$$

and hence

$$Q_a(y^*; y_n) = \Phi\left(\frac{y^* - \hat{\theta}x^*}{\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}}\right) \tag{3.22}$$

is an asymptotic predictive distribution.

We can also construct the predictive distribution using bootstrap distribution of $\theta$, since bootstrap distribution is an asymptotic CD (as demonstrated earlier) with which the bootstrap estimator is the corresponding CD-random variable. Specially, we first bootstrap the residuals $e_i = y_i - \hat{\theta}x_i$, denoted by $e_{i,\text{boot}}$, and then compute the bootstrap least squares estimate of $\hat{\theta}_{\text{boot}}$ using the new samples $\{(y_{i,\text{boot}}, x_i)\}_{i=1}^n$, where $y_{i,\text{boot}} \equiv \hat{\theta}x_i + e_{i,\text{boot}}$. Finally, a sample from the predictive distribution of $Y^*_{\text{boot}}$, say $Q_{\text{boot}}(\cdot; \mathbf{y}_n)$, can be obtained empirically by first generating $\varepsilon^* \sim N(0, \hat{\sigma}^2)$ and then computing $y^*_{\text{boot}} = \hat{\theta}_{\text{boot}}x^* + \varepsilon^*$. Repeat these four steps for a large number of times to get sufficient many copies of $y^*_{\text{boot}}$. These copies of $y^*_{\text{boot}}$ are then used to construct a predictive distribution function as well as prediction intervals.

We compare the empirical coverage probabilities of the prediction intervals from the four different predictive distributions: i) the naive plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*) = \Phi((y^* - \hat{\theta}x^*)/\hat{\sigma})$, ii) the exact predictive distribution $Q_t(y^*; \mathbf{y}_n)$ in (3.21), iii) the asymptotic predictive distributions $Q_a(y^*; \mathbf{y}_n)$ in (3.22) and iv) $Q_{\text{boot}}(y^*; \mathbf{y}_n)$ described above. The prediction intervals are obtained by taking the upper and lower $\alpha/2$ quantiles of the corresponding predictive distributions. Comparisons are made with different choices of $\alpha$ and sample sizes in order to provide a general picture of their performance. The numerical settings are as follows: $\theta = 1$, $\sigma = 1$, $x_i \sim U[-2, 2]$ are fixed once they have been generated, and $x^* = 2$. For $Q_{\text{boot}}(y^*; \mathbf{y}_n)$, $1,000$ bootstrap samples are utilized. Three sample sizes are considered: $n = 10, 100, 1,000$. For each sample size, the analysis is repeated 5,000 times with $y_1, \ldots, y_n, y^*$ being simulated anew accordingly.

Table 3.1 shows the empirical coverage probabilities and median widths of the prediction intervals. Note that the widths of the prediction intervals from $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, $Q_t(y^*; \mathbf{y}_n)$ and $Q_a(y^*; \mathbf{y}_n)$ can be assessed without simulation. They are, respectively, $2z_{\alpha/2}\hat{\sigma}$, $2t_{n-1,\alpha/2}\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}$ and $2z_{\alpha/2}\hat{\sigma}\sqrt{1 + (x^*)^2/\sum_{i=1}^n x_i^2}$. Here, $t_{n-1,\alpha/2}$ and $z_{\alpha/2}$ are the $(1 - \alpha/2)$th percentiles of $t$-distribution with degrees of freedom $n - 1$ and the standard normal distribution, respectively. From Table 3.1, at all nominal levels, the prediction intervals from $Q_t(y^*; \mathbf{y}_n)$ has the correct frequentist coverage probability as expected. For small sample size (such as $n = 10$), the empirical coverage of the prediction intervals from $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$ is far below its nominal level. This is because those prediction intervals do not take into account the uncertainty stemming from the estimation of the unknown parameter and such uncertainty is large relative (or at least comparable) to the noise level ($\sigma^2$) for small $n$. The empirical coverages of the prediction intervals from $Q_a(y^*; \mathbf{y}_n)$ and $Q_{\text{boot}}(y^*; \mathbf{y}_n)$ improve significantly upon those from $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$, though are still below the nominal level for small $n$. This is due to the facts that in $Q_a(y^*; \mathbf{y}_n)$ the estimated $\hat{\sigma}$ is used to approximate the actual $\sigma$ and that for $Q_{\text{boot}}(y^*; \mathbf{y}_n)$ the bootstrap distribution works well for at least moderate sample size $n$. For moderate or large sample size, such as $n = 100$ or $n = 1,000$, the coverage probabilities of all the four types of prediction intervals approximate well their nominal

| $n$ | $1-\alpha$ | $F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*)$ | | $Q_t(Y^*;\mathbf{Y}_n)$ | | $Q_a(Y^*;\mathbf{Y}_n)$ | | $Q_{\text{boot}}(Y^*;\mathbf{Y}_n)$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | Coverage | Width | Coverage | Width | Coverage | Width | Coverage | Width |
| 10 | 0.80 | 0.695 | 2.482 | 0.793 | 3.096 | 0.762 | 2.869 | 0.746 | 2.786 |
| 10 | 0.90 | 0.803 | 3.185 | 0.895 | 4.103 | 0.860 | 3.682 | 0.845 | 3.579 |
| 10 | 0.95 | 0.871 | 3.796 | 0.950 | 5.064 | 0.916 | 4.387 | 0.906 | 4.258 |
| 100 | 0.80 | 0.795 | 2.555 | 0.805 | 2.619 | 0.803 | 2.601 | 0.798 | 2.595 |
| 100 | 0.90 | 0.893 | 3.279 | 0.903 | 3.370 | 0.899 | 3.339 | 0.897 | 3.326 |
| 100 | 0.95 | 0.941 | 3.907 | 0.947 | 4.028 | 0.944 | 3.978 | 0.943 | 3.958 |
| 1000 | 0.80 | 0.801 | 2.562 | 0.802 | 2.568 | 0.801 | 2.566 | 0.799 | 2.562 |
| 1000 | 0.90 | 0.901 | 3.288 | 0.902 | 3.296 | 0.901 | 3.293 | 0.900 | 3.284 |
| 1000 | 0.95 | 0.948 | 3.918 | 0.948 | 3.929 | 0.948 | 3.924 | 0.942 | 3.903 |

Table 3.1: Comparison of different predictive distributions in Simulation I: 80%, 90% and 95% prediction intervals.

levels.

**Simulation II.** Consider a nonlinear regression in the form of

$$y_i = h(x_i, \boldsymbol{\theta}) + \varepsilon_i, \quad i = 1, \dots, n,$$

where

$$h(x_i, \boldsymbol{\theta}) = \frac{\theta_1 x_i}{\theta_2 + x_i},$$

and $\varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$. The parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)^t$ can be estimated by nonlinear least squares (NLS) and solved iteratively using Gauss-Newton algorithm. Denote by $\hat{\boldsymbol{\theta}}$ the NLS estimate of $\theta$ and let $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^n (y_i - h(x_i, \hat{\boldsymbol{\theta}}))^2$. Although no explicit expression of the exact sampling distribution of $\hat{\boldsymbol{\theta}}$ exists, it can be approximated by $N\big(\boldsymbol{\theta}, \sigma^2(A(\mathbf{x}, \boldsymbol{\theta})^t A(\mathbf{x}, \boldsymbol{\theta}))^{-1}\big)$ where $A(\mathbf{x}, \boldsymbol{\theta})$ is the $n \times 2$ matrix with its $i$th row being $\Big(\frac{\partial}{\partial \theta_1} h(x_i, \boldsymbol{\theta}), \frac{\partial}{\partial \theta_2} h(x_i, \boldsymbol{\theta})\Big) = \Big(\frac{x_i}{\theta_2 + x_i}, -\frac{\theta_1 x_i}{(\theta_2 + x_i)^2}\Big)$. Therefore, the cumulative distribution function of $N\big(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2(A(\mathbf{x}, \hat{\boldsymbol{\theta}})^t A(\mathbf{x}, \hat{\boldsymbol{\theta}}))^{-1}\big)$ can be used as an asymptotic CD function for $\boldsymbol{\theta}$. In this formula, the unknown values are replaced by their corresponding estimates.

For a new independent observation $Y^*$ associated with covariate $x^*$, we can construct asymptotic predictive distribution by using the above asymptotic CD and taking advantage of the approximation

$$h(x^*, \boldsymbol{\theta}) \approx h(x^*, \hat{\boldsymbol{\theta}}) + a(x^*, \hat{\boldsymbol{\theta}})^t (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}),$$

| $n$ | $1-\alpha$ | $F_{\hat{\theta}(\mathbf{Y}_n)}(Y^*)$ | | $Q_a(Y^*;\mathbf{Y}_n)$ | | $Q_{\mathrm{boot}}(Y^*;\mathbf{Y}_n)$ | |
|---|---|---|---|---|---|---|---|
| | | Coverage | Width | Coverage | Width | Coverage | Width |
| 10 | 0.80 | 0.698 | 2.438 | 0.764 | 2.817 | 0.751 | 2.742 |
| 10 | 0.90 | 0.809 | 3.129 | 0.862 | 3.615 | 0.854 | 3.512 |
| 10 | 0.95 | 0.872 | 3.728 | 0.913 | 4.308 | 0.903 | 4.177 |
| 100 | 0.80 | 0.782 | 2.552 | 0.791 | 2.610 | 0.791 | 2.603 |
| 100 | 0.90 | 0.888 | 3.275 | 0.896 | 3.350 | 0.894 | 3.335 |
| 100 | 0.95 | 0.941 | 3.902 | 0.946 | 3.992 | 0.943 | 3.973 |
| 1000 | 0.80 | 0.794 | 2.564 | 0.795 | 2.569 | 0.793 | 2.564 |
| 1000 | 0.90 | 0.895 | 3.291 | 0.896 | 3.298 | 0.896 | 3.285 |
| 1000 | 0.95 | 0.949 | 3.922 | 0.949 | 3.929 | 0.947 | 3.907 |

Table 3.2: Comparison of different predictive distributions in Simulation II: 80%, 90% and 95% prediction intervals.

where $a(x^*,\boldsymbol{\theta}) = \frac{\partial h(x^*,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$. Applying formula (3.4) with some simple algebra, one can obtain the asymptotic predictive distribution

$$Q_a(y^*; y_n) = \Phi\left(\frac{y^* - h(x^*,\hat{\boldsymbol{\theta}})}{\hat{\sigma}\sqrt{1 + a(x^*,\hat{\boldsymbol{\theta}})^t (A(\mathbf{x},\hat{\boldsymbol{\theta}})^t A(\mathbf{x},\hat{\boldsymbol{\theta}}))^{-1} a(x^*,\hat{\boldsymbol{\theta}})}}\right).$$

Alternatively, we can construct the bootstrap-based predictive distribution, denoted by $Q_{\mathrm{boot}}(y^*;\mathbf{y}_n)$, following almost the same procedure as in Simulation I.

We proceed to compare the empirical coverage probabilities of the prediction intervals from the three different predictive distributions: i) the naive plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*) = \Phi((y^* - h(x^*,\hat{\boldsymbol{\theta}}))/\hat{\sigma})$, the asymptotic predictive distributions ii) $Q_a(y^*;\mathbf{y}_n)$, and iii) $Q_{\mathrm{boot}}(y^*;\mathbf{y}_n)$. Comparisons are made at $\alpha = 0.8, 0.9, 0.95$ and $n = 10, 100, 1,000$ with 5,000 repetitions for each sample size. The numerical settings are: $\theta_1 = 15$, $\theta_2 = 5$, $\sigma = 1$, $x_i \overset{\text{i.i.d.}}{\sim} U[0,30]$ are fixed once generated, $x^* = 40$. For the bootstrap-based approach, $1,000$ bootstrap samples are generated. Similar to Table 3.1, Table 3.2 lists the empirical coverage probabilities and median widths of the prediction intervals. In the case of small sample size ($n = 10$), the empirical coverage of the prediction intervals from all the three approaches are below the nominal level since they are all approximate methods. However, both the CD-based predictive distributions, either $Q_a(y^*;\mathbf{y}_n)$ derived from the multivariate normal CD or $Q_{\mathrm{boot}}(y^*;\mathbf{y}_n)$ derived from the bootstrap CD, have outperformed the plug-in predictive distribution $F_{\hat{\theta}(\mathbf{y}_n)}(y^*)$ in terms of empirical coverage. This is because the CD-based methods have incorporated the uncertainty in the parameter estimation. Again, for moderate or large sample size, such as $n = 100$ or $n = 1,000$, the coverage probabilities of all the three prediction intervals are close to the corresponding nominal levels.

## 3.6 Real Data Example

In this section, we provide a real data example, in which the predictive inference is applied to data from a complex time series. We can envision that the development of predictive distributions be applied and generalized to other complex situations such as survival analysis, multiple regressions and any other fields and applications that involve forecasting and prediction.

Before we start our real data example, we extend the general formula (3.4) discussed in Section 3.2 to cover the case that $Y^*$ and $\mathbf{Y}_n$ are dependent; for instance, a time series data in which $\mathbf{Y}_n$ are sample observations up to data and $Y^*$ is a future response at the time series. Specifically, we propose to consider the conditional distribution of $Y^*$ given $\mathbf{Y}_n$ and modify the general formula (3.4) to be

$$Q_c(y^*; \mathbf{y}_n) = \int_{\theta \in \boldsymbol{\Theta}} F_\theta(y^* | \mathbf{y}_n) dH(\theta; \mathbf{y}_n). \tag{3.23}$$

In fact, formula (3.4) can now be viewed as a special case of (3.23) when $F_\theta(y^*|\mathbf{y}_n) \equiv F_\theta(y^*)$. Many of the theoretical results developed in Section 3.3 can be extended straightforwardly. For example, if we modify (3.7) to be $\sup_t |F_\theta(t|\mathbf{y}_n) - F_{\theta_0}(t|\mathbf{y}_n)| \leq C\,|\theta - \theta_0|$ for some positive constant $C$, then the result of Theorem 3.1 applies to $Q_c(y^*; \mathbf{y}_n)$ for the dependent case. This means that the predictive distribution function $Q_c(y^*; \mathbf{y}_n)$ for the dependent case also has valid frequentist interpretations, under a set of very mild conditions.

The real data example is from a research project partially sponsored by the US Department Homeland Security (DHS) through its academic research center DHS University Center of Excellence for Command, Control, and Interoperability (CCICADA) based at Rutgers University. This data example specifically focuses on the analysis of the monthly volume of applications for a certain type of government benefit (the name

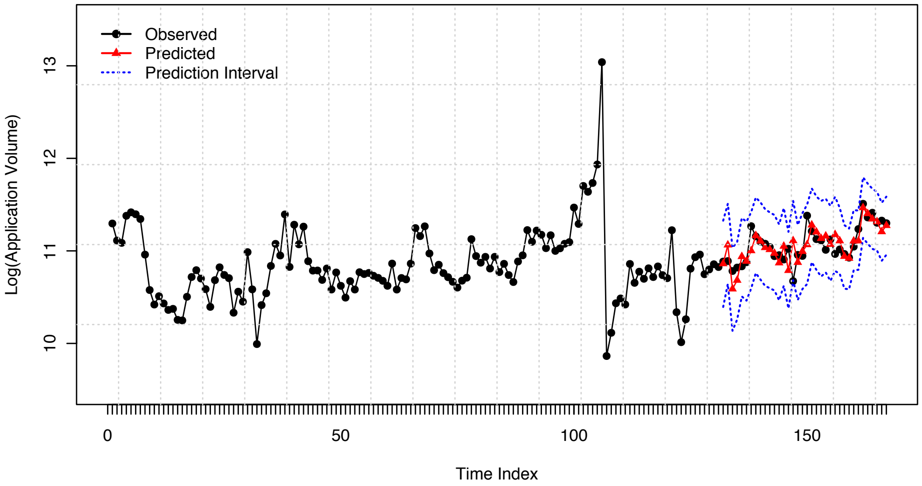

Figure 3.1: Time series plot of monthly application volumes for a government benefit and 95% one-step ahead prediction intervals starting from $t = 132$ to 167 on a rolling basis with window width $d = 120$. The red points show the predicted values and the blue dotted lines are the upper and lower limits of the corresponding 95% prediction intervals.

of the governmental program is masked per a confidentiality agreement).

The main objective of the project is to seek more effective statistical methods that can substantially improve upon the current benchmark model used by the agency in gaining accuracy of forecast. This gain can allow the agency to optimize the human resource allocation and minimize the cost of management.

The data set contains 167 months of application volume. The logarithm transformation of the 167 observed volumes are shown in Figure 3.1. We denote the transformed series by $\{y_t\}_{t=1}^{167}$.

It was noted in Chang (2015) the known outliers at $t = 105$, 106, 107 due to policy changes in the application process. Thus, we filter out these outliers with three indicator variables $\mathbb{1}_t^{(105)}$, $\mathbb{1}_t^{(106)}$, and $\mathbb{1}_t^{(107)}$, where $\mathbb{1}_t^{(k)} = \mathbb{1}\{t = k\}$. Also, the series in seasonal nature exhibits a cyclical pattern with periodicity of 12 that is modeled with seasonal terms. In addition, there is a strong linear relationship between $y_t$ and another type of benefit application $x_t$. Taking all this information into account and also building upon the work by Chang (2015), we propose the following seasonal ARMA model with

| | $\phi_1$ | $\Phi_1$ | $\Theta_1$ | $\beta_1$ | $\beta_2$ | $\beta_3$ | $\beta_4$ |
|---|---|---|---|---|---|---|---|
| Estimate | 0.784 | 0.998 | -0.966 | 0.975 | 0.633 | 2.160 | -0.696 |
| Std. Error | 0.048 | 0.011 | 0.111 | 0.014 | 0.175 | 0.200 | 0.175 |

Table 3.3: Coefficient estimates and their standard errors of model (3.24).

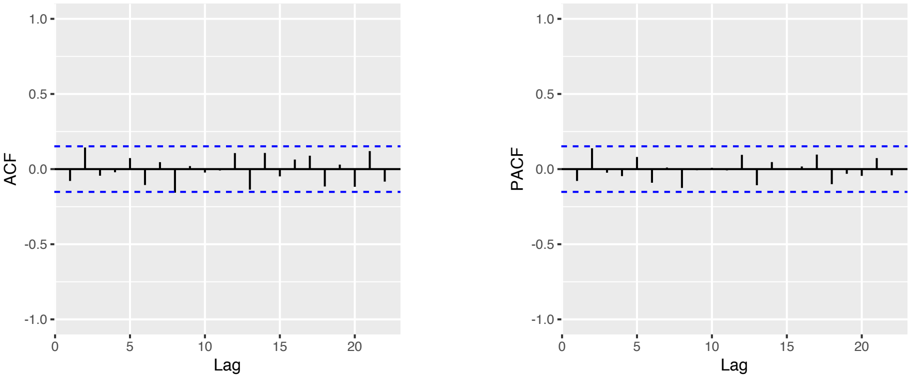

Figure 3.2: Sample ACF and PACF plots of the residuals from model (3.24).

exogenous variables,

$$(1-\phi_1 B)(1-\Phi_1 B^{12})(y_t - \beta_1 x_t - \beta_2 \mathbb{1}_t^{(105)} - \beta_3 \mathbb{1}_t^{(106)} - \beta_4 \mathbb{1}_t^{(107)}) = (1+\Theta_1 B^{12})\varepsilon_t. \quad (3.24)$$

Here, $\{\varepsilon_t\}$ is a white noise series with variance $\sigma_\varepsilon^2$, $B$ is the backshift operator such that $B^s y_t = y_{t-s}$ for an integer $s > 0$. Also, denote by $\boldsymbol{\theta} = (\phi_1, \Phi_1, \Theta_1, \beta_1, \beta_2, \beta_3, \beta_4)$ the associated coefficients.

Table 3.3 summarizes the coefficient estimates and their standard errors from model (3.24). It is easy to see that all the coefficients are significant at the 95% significance level. Figure 3.2 shows the sample autocorrelation function (ACF) and partial autocorrelation function (PACF) plots of the residuals from model (3.24). With no significant autocorrelation and partial autocorrelation, we conclude that model (3.24) is adequate in capturing the patterns of $\{y_t\}_{t=1}^{167}$.

Our ultimate goal is to make prediction on future application volumes given the past observations and to construct the corresponding prediction intervals and predictive distributions. More specifically, we need predict a sequence of $y_{167+h}$ for $h = 1, 2, \ldots,$

based on past observations up to time $t = 167$. On the other hand, since we do not know the values of the future observations after $t = 167$, we cannot really tell how good these predictions are. To this end, we demonstrate the effectiveness of our proposed method by formulating our predictions as of length $h > 0$ steps away, on a rolling basis with a rolling window size $d$ (e.g., $d = 120$ corresponding to the data of the past ten years). That is, at time $t$, we predict $y_{t+l}$ based on the most recent $d$ observations, compare the prediction with the actual value, and then increase $t$ by one and repeat the procedure until $t = 167 - h$. It is well-known that the coverage of the prediction intervals by the so-called plug-in method (described in Section 3.3) is typically below the nominal level because they fail to consider the uncertainty in parameter estimation, among others. Using our approach, however, it is possible to capture this type of uncertainty, and thus show substantial improvement.

The process to derive simulated predictive distribution of $y_{t+h}$, given $\{y_{t-d+1}, \ldots, y_t\}$ for any prediction length $h$, is outlined in four steps as follows.

1. Estimate model (3.24) using, e.g., maximum likelihood method, and $\{y_{t-d+1}, \ldots, y_t\}$. Denote by $\hat{\boldsymbol{\theta}} = (\hat{\phi}_1, \hat{\Phi}_1, \hat{\Theta}_1, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ the estimated coefficients, $\hat{\Sigma}$ the covariance matrix of $\hat{\boldsymbol{\theta}}$, and $\hat{\sigma}_\varepsilon^2$ the estimated variance of the error term. Let $\hat{y}_s$ be the fitted values of $y_s$ and $e_s = y_s - \hat{y}_s$, for $s \leq t$.

2. As demonstrated in Section 3.2, the multivariate normal distribution $N(\hat{\boldsymbol{\theta}}, \hat{\Sigma})$ serves as a first-order asymptotic CD for $\boldsymbol{\theta} = (\phi_1, \Phi_1, \Theta_1, \beta_1, \beta_2, \beta_3, \beta_4)$ for a reasonable $d$. Thus, we can simulate

$$\hat{\boldsymbol{\theta}}_{\text{CD}} = (\phi_{1,\text{CD}}, \Phi_{1,\text{CD}}, \Theta_{1,\text{CD}}, \beta_{1,\text{CD}}, \beta_{2,\text{CD}}, \beta_{3,\text{CD}}, \beta_{4,\text{CD}}) \sim N(\hat{\boldsymbol{\theta}}, \hat{\Sigma}).$$

   We also draw $\varepsilon_{t+1}^*, \ldots, \varepsilon_{t+h}^* \overset{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2)$, with the unknown $\sigma_\varepsilon^2$ replaced by $\hat{\sigma}_\varepsilon^2$ under a reasonable $d$.

3. Recursively solve for $y_{t+h}^*$ through $(1-\phi_{1,\text{CD}}B)(1-\Phi_{1,\text{CD}}B^{12})(y_{t+h}^* - \beta_{1,\text{CD}}x_{t+h} - \beta_{2,\text{CD}}\mathbb{1}_{t+h}^{(105)} - \beta_{3,\text{CD}}\mathbb{1}_{t+h}^{(106)} - \beta_{4,\text{CD}}\mathbb{1}_{t+h}^{(107)}) = (1+\Theta_{1,\text{CD}}B^{12})\varepsilon_{t+h}^*$, where $y_s^* = y_s$ and $\varepsilon_s^* = e_s$ for $s \leq t$.

4. Repeat Steps 1 to 3 for, say, $N = 5,000$ times and get $N$ copies of prediction value of $y^*_{t+h}$. These copies of $y^*_{t+h}$ can be used to form a predictive distribution and prediction intervals for $y_{t+h}$.

Following the algorithm above, we can now make one-step ahead prediction, i.e., $h = 1$, for our dataset, rolling from $t = 131$ to 166 (representing three years) with rolling window size $d = 120$. The blue dotted lines in Figure 3.1 show the upper and lower limits of the 95% prediction intervals.

We also plot in Figure 3.3 the predictive predictions at, for example, $t = 141, 142, 143$ and 144, respectively, with the black lines indicating the actual values of $y_t$. The predictive distributions provided in our prediction contain a wealth of information and can facilitate the quantification of uncertainty in prediction. Take $t = 141$ for example, we are able to gain insight into issues such as: i) What is the prediction interval at 90% confidence level? (The 90% prediction interval is [10.7,11.4].) ii) What confident levels are associated with the statements that the untransformed application volume will be greater than 40,000, 50,000 or 60,000? (The confidence is 98.3%, 84.4% and 54.0% respectively.) iii) What is the lowest predicted application volume of original scale at 90% confidence level? (It is with 90% confidence level that the application volume will exceed 47,332.) These are all important questions concerning government officials in their planning of allocating manpower for handling applications.

## 3.7 Further Comments

This framework developed in this chapter is very general and the proposed CD-based formulation is broadly applicable, as the CD concept covers a broad range of examples, including: fiducial distribution, bootstrap distributions, likelihood functions (after normalization), $p$-value functions, and Bayesian posterior distributions. Regardless of different statistical paradigms, these examples can all be used as CDs as long as they provide valid frequentist probability coverage. This entails that the proposed predictive distribution has the desirable property to be flexible and all encompassing. Case in point is that the Bayesian posterior distribution is often a CD, either asymptotically

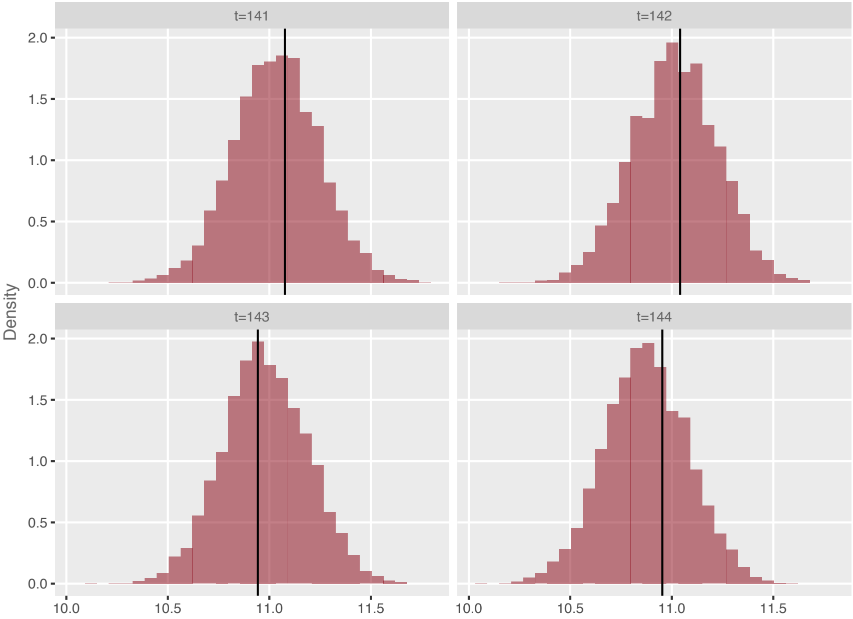

Figure 3.3: One-step ahead predictive distribution of $Y_t$ for $t = 141, 142, 143$ and 144.

under the Bernstein-von Mises type theorems or exact using probability matching priors. Noting that $Q(y^*; \mathbf{y}_n)$ has the same form of the Bayesian predictive distribution in (3.2), the Bayesian predictive distribution can be simply viewed as a special case of our CD-based predictive distribution. Similar arguments apply to the fiducial predictive distributions defined in Wang et al. (2012). All these observations show that the general formulation of $Q(y^*; \mathbf{y}_n)$ through CDs provides an ideal platform to unify most of, if not all, the existing frequentist, fiducial and Bayesian predictive distributions.

There are ample discussions in literature on the great generality and utility of CD as an inference tool. Given that CD has succeeded in providing solutions to problems surrounding difficult complex settings such as making inference from combining heterogeneous studies (e.g., Liu et al. 2015; Claggett et al. 2014; Yang et al. 2014) or studies that fail to produce well-defined point or interval estimates (e.g., Liu et al. 2014), it would seem natural to expect that our proposed CD-based approach can be applied to make inference in predictions for such complex problem settings as well. This should be worth studying further.

Finally, there are also some literatures that treat "predictive distributions" as estimators of $F_\theta(y^*|\mathbf{y}_n)$, the distribution function of $Y^*$ given $\mathbf{Y}_n = \mathbf{y}_n$, see, e.g., Aitchison (1975), Murray (1977), Ng (1980), Lejeune and Faulkenberry (1982), Harris (1989), and Vidoni (1998). But, as pointed out by Lawless and Fredette (2005), although an estimator of $F_\theta(y^*|\mathbf{y}_n)$, say $\tilde{F}(y^*|\mathbf{y}_n)$, provides probability statements about the future random variable $Y^*$, given $\mathbf{Y}_n = \mathbf{y}_n$, the probability statements for $Y^*$ do not have a frequentist interpretation in terms of repeated sampling. For example, even if $a^* = L(\mathbf{y}_n)$ is chosen so that $\tilde{F}(a^*|\mathbf{y}_n) = 0.95$, it is not true in general that $\mathbb{P}_{\mathbb{J}}(Y^* < L(\mathbf{Y}_n)) = 0.95$; see Lawless and Fredette (2005) for further elaborations. Furthermore, there are developments of "predictive likelihood function" (see, e.g., Bjornstad 1990 and references therein), which rely on a so-called likelihood principle for prediction (Berger and Wolpert, 1988). The general idea here is to eliminate the "nuisance" parameter $\theta$ in the joint likelihood function $L(\theta|y^*, \mathbf{y}_n)$ by using different techniques to obtain a new "likelihood" $L(y^*|\mathbf{y}_n)$ which is free of $\theta$, and then use it to make predictive inference. Depending on the techniques use, different versions of predictive likelihood functions can be obtained, and their performance naturally varies. Some may meet the frequentist probability coverage criterion discussed in this chapter, but many may not (cf. Bjornstad 1990). Finally, even though in some special cases the method of the predictive likelihood function coincides with the predictive distribution function developed in this chapter, this method does not stress the need of providing a predictive distribution function that has suitable frequentist probabilistic interpretations.

# Chapter 4
# Concluding Remarks

In this dissertation, we use the concept of CD to develop two novel and general frameworks/approaches for statistical fusion learning and predictive inference, that is, *i*Fusion, and CD-based predictive distribution function.

*i*Fusion is a statistical fusion learning framework for drawing efficient individualized inference, through adaptive combination of confidence density functions from individual subjects. Such effective *i*Fusion "borrows strength" from other individuals, while preserves inference validity by "smartly" borrowing only from individuals that bears relevance to the target individual and filtering out unrelated ones. Under suitable definition of cliques and the separation condition, statistical inference derived from the combined confidence density function is shown to achieve the best allowable asymptotic efficiency. Extensions are further made to accommodate a wide range of model design heterogeneity.

The CD-based predictive distribution framework is used for predictive inference by: i) providing a formal definition of predictive distribution functions, ii) presenting a general approach based on CDs for constructing such predictive distribution functions, and finally, iii) proposing a Monte-Carlo algorithm to implement the CD-based approach. We also establish the supporting theories for the proposed approach, and discuss the optimality issues and the connections to other existing prediction approaches, including Bayesian, fiducial and the frequentist pivotal-based predictive distribution proposed in Lawless and Fredette (2005) as well as the CD-based method by Schweder and Hjort (2016). The proposed approach is shown to have several desirable features. Particularly notable is its ability to afford a valid frequentist interpretation and yield prediction intervals of all levels with desired frequentist coverage probability.

# Appendix A

# Proofs

## Proof of Lemma 2.1

*Proof.*

i) We prove a more concrete result:

$$\mathbb{P}\Big(n^{\alpha}\|\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1\|_2 \geq \varepsilon\Big) \to 0$$

for any $\alpha \in (0, 1/2)$ and $\varepsilon > 0$. Define

$$\boldsymbol{\theta}_1^{(o)} = (\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1} \sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1} \boldsymbol{\theta}_k. \tag{A.1}$$

Then

$$\boldsymbol{\theta}_1^{(o)} - \boldsymbol{\theta}_1 = (\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1} \sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1) = \boldsymbol{o}_p(n^{-1/2}). \tag{A.2}$$

Since each $\hat{\boldsymbol{\theta}}_k$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_k$, it follows that

$$\begin{aligned} \mathbb{P}\Big(n^{\alpha}\|\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1\|_2 \geq \varepsilon\Big) &\leq \mathbb{P}\Big(n^{\alpha}\|\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1^{(o)}\|_2 \geq \varepsilon/2\Big) + \mathbb{P}\Big(n^{1/2}\|\boldsymbol{\theta}_1^{(o)} - \boldsymbol{\theta}_1\|_2 \geq \varepsilon/2\Big) \\ &\leq \mathbb{P}\Big(O_p(n^{\alpha-1/2}) \geq \varepsilon/2\Big) + \mathbb{P}\Big(o_p(1) \geq \varepsilon/2\Big) \to 0. \end{aligned}$$

ii) Write

$$n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1) = n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1^{(o)}) + n^{1/2}(\boldsymbol{\theta}_1^{(o)} - \boldsymbol{\theta}_1).$$

For the first term,

$$n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1^{(o)}) \xrightarrow{d} N(\mathbf{0}, n(\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1}),$$

where

$$\lim_{n\to\infty} n(\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1} = \Delta_1^{(o)}.$$

For the second term,

$$n^{1/2}(\boldsymbol{\theta}_1^{(o)} - \boldsymbol{\theta}_1) = \boldsymbol{o}_p(1)$$

based on (A.2). Together these lead to

$$n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(o)} - \boldsymbol{\theta}_1) \xrightarrow{d} N(\mathbf{0}, \Delta_1^{(o)}).$$

iii) By some simple calculation, the MSE of $\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}$ can be decomposed as the sum of its squared bias and trace of its covariance:

$$\text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) = \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2} \in \mathcal{F}} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} (\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-2} \hat{\Sigma}_{k_2}^{-1} (\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \text{tr}\Big\{(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-1}\Big\}, \tag{A.3}$$

asymptotically (we use the term "asymptotically" since $\hat{\Sigma}_k$'s are the sample covariance matrix rather than its population version). Also, asymptotically,

$$\text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) = \text{tr}\Big\{(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-1}\Big\} = O(n^{-1})$$

if $\mathcal{F} \subseteq \mathcal{C}_1$, because the squared bias in (A.3) is of order $o(n^{-1})$ and is dominated by the trace when $\boldsymbol{\theta}_k \in \mathcal{C}_1$. In contrast, if any $\boldsymbol{\theta}_k \notin \mathcal{C}_1$, i.e., $\boldsymbol{\theta}_k \in \mathcal{D}_1$ (since $\mathcal{B}_1 = \emptyset$), is included in $\mathcal{F}$, then asymptotically

$$\text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) = \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2} \in \mathcal{F}} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} (\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-2} \hat{\Sigma}_{k_2}^{-1} (\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1)$$

and $n\text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) \to \infty$. Thus, the MSE-optimal $\mathcal{F}$ should be a subset of $\mathcal{C}_1$. On the other hand, because $\text{tr}\Big\{(A+B)^{-1}\Big\} \leq \text{tr}\Big\{A^{-1}\Big\}$ for any two positive definite matrix $A$ and $B$, we have

$$\text{tr}\Big\{(\sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} \hat{\Sigma}_k^{-1})^{-1}\Big\} \leq \text{tr}\Big\{(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-1}\Big\},$$

for $\forall \mathcal{F} \subseteq \mathcal{C}_1$. So the choice of $\mathcal{F} = \mathcal{C}_1$, in other words, $\hat{\boldsymbol{\theta}}_1^{(o)}$ has the smallest asymptotical MSE among all estimators in the form of $\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}$.

□

### Proof of Theorem 2.1

*Proof.*

i) Define

$$\boldsymbol{\theta}_1^{(c)} = (\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1})^{-1}\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1}\boldsymbol{\theta}_k. \tag{A.4}$$

It follows that

$$\begin{aligned}
\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1 &= (\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1})^{-1}\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1) \\
&= (\sum_{k=1}^{K} w_{1k}\hat{\Sigma}_k^{-1})^{-1}\Big(\sum_{\boldsymbol{\theta}_k \notin \mathcal{C}_1} w_{1k}\hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1) + \sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1} w_{1k}\hat{\Sigma}_k^{-1}(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\Big) \\
&= \Big(\sum_{k=1}^{K} w_{1k}(n\hat{\Sigma}_k^{-1})\Big)^{-1}\Big(\sum_{\boldsymbol{\theta}_k \notin \mathcal{C}_1} o_p(n^{-1/2})(n\hat{\Sigma}_k^{-1})(\boldsymbol{\theta}_k - \boldsymbol{\theta}_1) \\
&\quad + \sum_{\boldsymbol{\theta}_k \in \mathcal{C}_1}(1 + o_p(n^{-1/2}))(n\hat{\Sigma}_k^{-1})\boldsymbol{o}_p(n^{-1/2})\Big) \\
&= \boldsymbol{o}_p(n^{-1/2}).
\end{aligned} \tag{A.5}$$

Since $\hat{\boldsymbol{\theta}}_1^{(c)}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_1^{(c)}$, it follows that, for any $\alpha \in (0, 1/2)$ and $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}\Big(n^{\alpha}\|\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1\|_2 \geq \varepsilon\Big) &\leq \mathbb{P}\Big(n^{\alpha}\|\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1^{(c)}\|_2 \geq \varepsilon/2\Big) + \mathbb{P}\Big(n^{1/2}\|\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1\|_2 \geq \varepsilon/2\Big) \\
&\leq \mathbb{P}\Big(O_p(n^{\alpha-1/2}) \geq \varepsilon/2\Big) + \mathbb{P}\Big(o_p(1) \geq \varepsilon/2\Big) \to 0.
\end{aligned}$$

ii) Similar to the proof for part ii) of Lemma 2.1, this is a direct result of

$$n^{1/2}(\hat{\boldsymbol{\theta}}_1^{(c)} - \boldsymbol{\theta}_1^{(c)}) \xrightarrow{d} N(\mathbf{0}, n(\sum_{k=1}^{K} w_{1k}\Lambda_k)^{-1}(\sum_{k=1}^{K} w_{1k}^2\Lambda_k)(\sum_{k=1}^{K} w_{1k}\Lambda_k)^{-1})$$

where the covariance matrix converges to $\Delta_1^{(o)}$ in probability, and $n^{1/2}(\boldsymbol{\theta}_1^{(c)} - \boldsymbol{\theta}_1) = \boldsymbol{o}_p(1)$.

iii) Asymptotically, $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) = \mathrm{tr}\Big\{\mathrm{Var}(\hat{\boldsymbol{\theta}}_1^{(c)})\Big\}$ and $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(o)}) = \mathrm{tr}\Big\{\mathrm{Var}(\hat{\boldsymbol{\theta}}_1^{(o)})\Big\}$. But from part ii) of Theorem 2.1, we have shown that $\hat{\boldsymbol{\theta}}_1^{(c)}$ and $\hat{\boldsymbol{\theta}}_1^{(o)}$ have the same limiting covariance matrix. Therefore asymptotically we have $\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) = \mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(o)})$.

□

## Proof of Lemma 2.2

*Proof.*

i) On one hand, when $\boldsymbol{\theta}_k \notin \mathcal{C}_1$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 \geq d_1$. For any $\varepsilon > 0$ and $b_n$ satisfying (2.19),

$$\begin{aligned}
& \mathbb{P}\Big(n^{1/2}\mathbb{1}\{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n \leq 1\} \leq \varepsilon\Big) \\
= \;& \mathbb{P}\Big(\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n > 1\Big) \\
= \;& \mathbb{P}\Big(\big(\|(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) - (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\|\big)/b_n \geq 1\Big) \\
\geq \;& \mathbb{P}\Big(\big(\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 - \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 - \|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_h\|_2\big)/b_n \geq 1\Big) \\
\geq \;& \mathbb{P}\Big(\big(1 - \frac{O(n^{-1/2})}{d_1}\big)\frac{d_1}{b_n} \geq 1\Big) \to 1.
\end{aligned}$$

On the other hand, when $\boldsymbol{\theta}_k \in \mathcal{C}_1$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 = o(n^{-1/2})$. For any $\forall\varepsilon > 0$ and $b_n$ satisfying (2.19),

$$\begin{aligned}
& \mathbb{P}\Big(n^{1/2}|\mathbb{1}\{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n \leq 1\} - 1| \leq \varepsilon\Big) \\
= \;& \mathbb{P}\Big(\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2/b_n \leq 1\Big) \\
= \;& \mathbb{P}\Big(\|(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) - (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\|_2/b_n \leq 1\Big) \\
\geq \;& \mathbb{P}\Big((\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 + \|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k\|_2 + \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2)/b_n \leq 1\Big) \\
= \;& \mathbb{P}\Big(\frac{O(n^{-1/2})}{b_n} \leq 1\Big) \to 1,
\end{aligned}$$

as $n \to \infty$. This has completed the proof that $w_{1k} = 1 + o_p(n^{-1/2})$ when $\boldsymbol{\theta}_k \in \mathcal{C}_1$ and that $w_{1k} = o_p(n^{-1/2})$ when $\boldsymbol{\theta}_k \notin \mathcal{C}_1$ under the uniform kernel.

ii) We show for the Epanechnikov kernel only. In fact, the part when $\boldsymbol{\theta}_k \notin \mathcal{C}_1$ is the same as that under the uniform kernel and is omitted.

When $\boldsymbol{\theta}_k \in \mathcal{C}_1$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 = o(n^{-1/2})$. For any $\forall \varepsilon > 0$ and $b_n$ satisfying (2.19),

$$\begin{aligned}
& \mathbb{P}\Big(n^{1/2}|(1 - \frac{|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2^2}{b_n^2})\mathbb{1}\{\frac{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2}{b_n} \leq 1\} - 1| \leq \varepsilon\Big) \\
= \; & \mathbb{P}\Big(n^{1/2}\frac{\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2^2}{b_n^2} \leq \varepsilon\Big) \\
= \; & \mathbb{P}\Big(n^{1/2}\|(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) - (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\|_2^2/b_n^2 \leq \varepsilon\Big) \\
\geq \; & \mathbb{P}\Big(n^{1/2}(\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 + \|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k\|_2 + \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2)^2/b_n^2 \leq \varepsilon\Big) \\
= \; & \mathbb{P}\Big(\frac{O(n^{-1/2})}{b_n^2} \leq \varepsilon\Big) \to 1,
\end{aligned}$$

as $n \to \infty$. This has completed the proof that $w_{1k} = 1 + o_p(n^{-1/2})$ when $\boldsymbol{\theta}_k \in \mathcal{C}_1$ and that $w_{1k} = o_p(n^{-1/2})$ when $\boldsymbol{\theta}_k \notin \mathcal{C}_1$ under the Epanechnikov kernel.

iii) On one hand, when $\boldsymbol{\theta}_k \notin \mathcal{C}_1$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\| \geq d_1$. For any $\forall \varepsilon > 0$ and $b_n$ satisfying (2.21),

$$\begin{aligned}
& \mathbb{P}\Big(n^{1/2}\exp\{-\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2^2/(2b_n^2)\} \leq \varepsilon\Big) \\
= \; & \mathbb{P}\Big(\|\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_k\|_2^2/(2b_n^2) \geq \log(\frac{n}{\varepsilon^2})\Big) \\
= \; & \mathbb{P}\Big(\|(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1) + (\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_k) - (\boldsymbol{\theta}_k - \boldsymbol{\theta}_1)\|_2^2/(2b_n^2) \geq \log(\frac{n}{\varepsilon^2})\Big) \\
\geq \; & \mathbb{P}\Big((\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_1\|_2 - \|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1\|_2 - \|\boldsymbol{\theta}_k - \hat{\boldsymbol{\theta}}_h\|_2)^2/(2b_n^2) \geq \log(\frac{n}{\varepsilon^2})\Big) \\
\geq \; & \mathbb{P}\Big((1 - \frac{O(n^{-1/2})}{d_1})^2(\frac{d_1}{b_n})^2 \geq 2\log(\frac{n}{\varepsilon^2})\Big) \to 1.
\end{aligned}$$

The last equation is true because $\left(\frac{d_1}{b_n}\right)^2/\log n \to \infty$ by (2.21) .

On the other hand, when $\boldsymbol{\theta}_k \in \mathcal{C}_1$, $\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_k\|_2 = o(n^{-1/2})$. For any $\varepsilon > 0$ and $b_n$

satisfying (2.21),

$$\begin{aligned}
& \mathbb{P}\Big(n^{1/2}|\exp\{-\|\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_k\|_2^2/(2b_n^2)\}-1|\le\varepsilon\Big)\\
\ge\ & \mathbb{P}\Big(n^{1/2}\|\hat{\boldsymbol{\theta}}_1-\hat{\boldsymbol{\theta}}_k\|_2^2/(2b_n^2)\le\varepsilon\Big)\\
=\ & \mathbb{P}\Big(n^{1/2}\|(\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1)+(\boldsymbol{\theta}_k-\hat{\boldsymbol{\theta}}_k)-(\boldsymbol{\theta}_k-\boldsymbol{\theta}_1)\|_2^2/(2b_n^2)\le\varepsilon\Big)\\
\ge\ & \mathbb{P}\Big(n^{1/2}(\|\hat{\boldsymbol{\theta}}_1-\boldsymbol{\theta}_1\|_2+\|\boldsymbol{\theta}_k-\hat{\boldsymbol{\theta}}_k\|_2+\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_1\|_2)^2/(2b_n^2)\le\varepsilon\Big)\\
=\ & \mathbb{P}\Big(\frac{O(n^{-1/2})}{b_n^2}\le 2\varepsilon\Big)\to 1
\end{aligned}$$

as $n\to\infty$. The last equation is true because $n^{1/2}b_n^2\to\infty$ by (2.21). This has completed the proof that $w_{1k}=1+o_p(n^{-1/2})$ when $\boldsymbol{\theta}_k\in\mathcal{C}_1$ and that $w_{1k}=o_p(n^{-1/2})$ when $\boldsymbol{\theta}_k\notin\mathcal{C}_1$ under the Gaussian kernel.

□

## Proof of Theorem 2.2

*Proof.* We only prove part iii); part i) and ii) can be proved by the same argument in the proof of Theorem 2.1. Define $\boldsymbol{\theta}_1^{(c)}$ as (A.4). For large $n$,

$$\hat{\boldsymbol{\theta}}_1^{(c)}=(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-1}\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\hat{\boldsymbol{\theta}}_k$$

and

$$\boldsymbol{\theta}_1^{(c)}=(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-1}\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1}\boldsymbol{\theta}_k.$$

Thus, asymptotically,

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) &= (\boldsymbol{\theta}_1^{(c)}-\boldsymbol{\theta}_1)^t(\boldsymbol{\theta}_1^{(c)}-\boldsymbol{\theta}_1)+\mathrm{tr}\Big\{\mathrm{Var}(\hat{\boldsymbol{\theta}}_1^{(c)})\Big\}\\
&= \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2}\notin\mathcal{D}_1}(\boldsymbol{\theta}_{k_1}-\boldsymbol{\theta}_1)^t\hat{\Sigma}_{k_1}^{-1}(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-2}\hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2}-\boldsymbol{\theta}_1)+\mathrm{tr}\Big\{(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-1}\Big\}\\
&= \sum_{\boldsymbol{\theta}_{k_1},\boldsymbol{\theta}_{k_2}\in\mathcal{B}_1}(\boldsymbol{\theta}_{k_1}-\boldsymbol{\theta}_1)^t\hat{\Sigma}_{k_1}^{-1}(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-2}\hat{\Sigma}_{k_2}^{-1}(\boldsymbol{\theta}_{k_2}-\boldsymbol{\theta}_1)+\mathrm{tr}\Big\{(\sum_{\boldsymbol{\theta}_k\notin\mathcal{D}_1}\hat{\Sigma}_k^{-1})^{-1}\Big\}.
\end{aligned}$$

The last equation holds because if either $\boldsymbol{\theta}_{k_1}$ or $\boldsymbol{\theta}_{k_2} \in \mathcal{C}_1$, then the squared bias vanishes in relative to the trace as $n \to \infty$.

Now, for any $\mathcal{F} \subseteq \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, if $\mathcal{D}^{\mathcal{F}} \neq \emptyset$, then from (A.3), $n\text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) \to \infty$, but $\text{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) = O(n^{-1})$. So asymptotically $\text{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) < \text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}})$. On the other hand, if $\mathcal{D}^{\mathcal{F}} = \emptyset$, then from (A.3),

$$\begin{aligned}
\text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}}) &= \sum_{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2} \in \mathcal{C}^{\mathcal{F}} \cup \mathcal{B}^{\mathcal{F}}} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} (\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-2} \hat{\Sigma}_{k_2}^{-1} (\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \text{tr}\Big\{(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-1}\Big\} \\
&= \sum_{\boldsymbol{\theta}_{k_1}, \boldsymbol{\theta}_{k_2} \in \mathcal{B}^{\mathcal{F}}} (\boldsymbol{\theta}_{k_1} - \boldsymbol{\theta}_1)^t \hat{\Sigma}_{k_1}^{-1} (\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-2} \hat{\Sigma}_{k_2}^{-1} (\boldsymbol{\theta}_{k_2} - \boldsymbol{\theta}_1) + \text{tr}\Big\{(\sum_{\boldsymbol{\theta}_k \in \mathcal{F}} \hat{\Sigma}_k^{-1})^{-1}\Big\}.
\end{aligned}$$

Thus we have established the asymptotic equivalence between $\text{MSE}(\hat{\boldsymbol{\theta}}_1^{(c)}) \leq \text{MSE}(\hat{\boldsymbol{\theta}}_1^{\mathcal{F}})$ and (2.23) when $\mathcal{D}^{\mathcal{F}} = \emptyset$. □

## Proof of Corollary 2.1

*Proof.* Asymptotically,

$$\text{Var}(\hat{\boldsymbol{\eta}}_1^{(c)}) = (\sum_{\boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1}.$$

It sufficies to show that

$$\{(\sum_{\boldsymbol{\xi}_k \in \tilde{\mathcal{C}}_1} A_k^t \hat{\Sigma}_k^{-1} A_k)^{-1}\}_{(1,1)} \leq \{\hat{\Sigma}_1\}_{(1,1)}.$$

Without loss of generality we assume that $K = 2$. If $\boldsymbol{\xi}_2 \notin \tilde{\mathcal{C}}_1$ then the equality holds. If $\boldsymbol{\xi}_2 \in \tilde{\mathcal{C}}_1$, we need show

$$\{(A_1^t \hat{\Sigma}_1^{-1} A_1 + A_2^t \hat{\Sigma}_2^{-1} A_2)^{-1}\}_{(1,1)} \leq \{\hat{\Sigma}_1\}_{(1,1)}.$$

Partition $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ as

$$\hat{\Sigma}_1 = \begin{pmatrix} a_1 & \boldsymbol{b}_1^t \\ \boldsymbol{b}_1 & C_1 \end{pmatrix}, \quad \hat{\Sigma}_2 = \begin{pmatrix} a_2 & \boldsymbol{b}_2^t \\ \boldsymbol{b}_2 & C_2 \end{pmatrix},$$

where $C_1$ and $C_2$ are $(p-1)\times(p-1)$ matrices. By definition,

$$A_1 = \begin{pmatrix} 1 & 0 & \mathbf{0}^t_{(p-1)\times 1} \\ \mathbf{0}^t_{(p-1)\times 1} & \mathbf{0}^t_{(p-1)\times 1} & I_{p-1} \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 1 & \mathbf{0}^t_{q\times 1} \\ \mathbf{0}^t_{(p-1)\times 1} & \mathbf{0}^t_{(p-1)\times 1} & I_{p-1} \end{pmatrix},$$

where $I_{p-1}$ is an identity matrix of size $p-1$. Some linear algebra with blockwise matrix inversion formula gives

$$\{(A_1^t\hat{\Sigma}_1^{-1}A_1 + A_2^t\hat{\Sigma}_2^{-1}A_2)^{-1}\}_{(1,1)} = a_1 - \boldsymbol{b}_1^t C_1^{-1}\boldsymbol{b}_1 + \boldsymbol{b}_1^t C_1^{-1}(C_1^{-1}+C_2^{-1})^{-1}C_1^{-1}\boldsymbol{b}_1.$$

By Lemma A.3 in Liu et al. (2015): for two $q\times q$ positive definite matrices $W_1$ and $W_2$ and $\boldsymbol{v}\in\mathbb{R}^q$,

$$\boldsymbol{v}^t(W_1+W_2)^{-1}\boldsymbol{v} \le \boldsymbol{v}W_1^{-1}\boldsymbol{v}.$$

Therefore

$$\{(A_1^t\hat{\Sigma}_1^{-1}A_1 + A_2^t\Lambda_2\hat{\Sigma}_2^{-1})^{-1}\}_{(1,1)} \le a_1 = \{\hat{\Sigma}_1\}_{(1,1)}.$$

□

## Proof of Theorem 3.1

*Proof.* By condition (A) and (3.7), we have, for any $\epsilon > 0$,

$$\begin{aligned} & \left| \int_{\theta\in\boldsymbol{\Theta}} \{F_\theta(Y^*) - F_{\theta_0}(Y^*)\}\, dH(\theta;\mathbf{Y}_n) \right| \\ \le \quad & \int_{\theta\in\boldsymbol{\Theta}} |F_\theta(Y^*) - F_{\theta_0}(Y^*)|\, dH(\theta;\mathbf{Y}_n) \\ = \quad & \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} |F_\theta(Y^*) - F_{\theta_0}(Y^*)|\, dH(\theta;\mathbf{Y}_n) + 2H(\theta_0-\epsilon) + 2(1-H(\theta_0+\epsilon)) \\ \le \quad & C\epsilon\int_{\theta_0-\epsilon}^{\theta_0+\epsilon} dH(\theta;\mathbf{Y}_n) + o_p(1) \le C\epsilon + o_p(1). \end{aligned}$$

It follows that

$$\int_{\theta\in\boldsymbol{\Theta}} \{F_\theta(Y^*) - F_{\theta_0}(Y^*)\}\, dH(\theta;\mathbf{Y}_n) = o_p(1).$$

Thus, we have

$$\begin{aligned} Q(Y^*;\mathbf{Y}_n) &= \int_{\theta\in\mathbf{\Theta}} F_\theta(Y^*)dH(\theta;\mathbf{Y}_n) = F_{\theta_0}(Y^*) + \int_{\theta\in\mathbf{\Theta}} \left\{F_\theta(Y^*) - F_{\theta_0}(Y^*)\right\} dH(\theta;\mathbf{Y}_n) \\ &= U + o_p(1). \end{aligned}$$

□

## Proof of Theorem 3.2

*Proof.* First, we note that

$$F_\theta(y^*) = \mathbb{P}_\theta(Y^* \leq y^*) = \mathbb{P}_\theta(s_1(Y^*,\theta) \leq s_1(y^*,\theta)) = S(s_1(y^*,\theta)).$$

Therefore,

$$\int_{\theta\in\mathbf{\Theta}} F_\theta(y^*)dH_R(\theta;\hat{\theta}(\mathbf{y}_n)) = \int_{\theta\in\mathbf{\Theta}} S(s_1(y^*,\theta))dH_R(\theta;\hat{\theta}(\mathbf{y}_n)). \tag{A.6}$$

Let $(W,V)$ be a transformation from $(Y^*,\hat{\theta}(\mathbf{Y}_n))$ such that

$$\begin{cases} W = F_{s_2(\hat{\theta}(\mathbf{Y}_n),\theta_0)}(s_1(Y^*,\theta_0)) \\ V = s_2(\hat{\theta}(\mathbf{Y}_n),\theta_0), \end{cases}$$

and let $w = F_{s_2(\hat{\theta}(\mathbf{y}_n),\theta_0)}(s_1(y^*,\theta_0))$ be a realization of $W$. By the invariance condition

$$w = F_{s_2(\hat{\theta}(\mathbf{y}_n),\theta)}(s_1(y^*,\theta)). \tag{A.7}$$

Plugging (A.7) into the right hand side of (A.6) yields

$$\int_{\theta\in\mathbf{\Theta}} F_\theta(y^*)dH_R(\theta;\hat{\theta}(\mathbf{y}_n)) = \int_{\theta\in\mathbf{\Theta}} S(F^{-1}_{s_2(\hat{\theta}(\mathbf{Y}_n),\theta)}(w))dH_R(\theta;\hat{\theta}(\mathbf{y}_n)). \tag{A.8}$$

On the other hand, the cumulative distribution function of $W$ is

$$\begin{aligned}
\mathbb{P}(W \leq w) &= \int_v \mathbb{P}(W \leq w | V = v) dR(v) \\
&= \int_v \mathbb{P}(F_v(s_1(Y^*, \theta_0)) \leq w) dR(v) \qquad \text{(A.9)} \\
&= \int_v \mathbb{P}(s_1(Y^*, \theta_0) \leq F_v^{-1}(w)) dR(v) \\
&= \int_v S(F_v^{-1}(w)) dR(v) \\
&= \int_{\theta \in \mathbf{\Theta}} S(F^{-1}_{s_2(\hat{\theta}(\mathbf{Y}_n), \theta)}(w)) dH_R(\theta; \hat{\theta}(\mathbf{y}_n)). \qquad \text{(A.10)}
\end{aligned}$$

Here, (A.9) is true because given $V = v$, $F_V(s_1(Y^*, \theta_0))$ is independent of $V$; and (A.10) is true following the transfer of randomness from $v$ to $\theta$ through $v = s_2(\hat{\theta}(\mathbf{y}_n), \theta)$.

By (A.8) and (A.10), we have that

$$Q_R(y^*; \mathbf{y}_n) = \int_{\theta \in \mathbf{\Theta}} F_\theta(y^*) dH_R(\theta; \hat{\theta}(\mathbf{y}_n)),$$

and it is equivalent to $F_W(w) = \mathbb{P}(W \leq w)$, where $F_W(\cdot)$ is the cumulative distribution function of $W$. Therefore, $Q_R(Y^*; \mathbf{Y}_n) = F_W(W)$ and it is uniformly distributed on $(0, 1)$. □

## Proof of Corollary 3.2

*Proof.* By the invariance condition

$$F_{\hat{\theta}}(y^*) = F_{s_2(\hat{\theta}(\mathbf{y}_n), \theta_0)}(s_1(y^*, \theta_0)).$$

It immediately follows from the proof of Theorem 3.2 that $K(F_{\hat{\theta}(\mathbf{y}_n)}(y^*))$, or equivalently, $K(F_{s_2(\hat{\theta}(\mathbf{y}_n), \theta_0)}(s_1(y^*, \theta_0)))$, can be expressed as $\int_{\theta \in \mathbf{\Theta}} F_\theta(y^*) dH_R(\theta; \hat{\theta}(\mathbf{y}_n))$. □

## Proof of Theorem 3.3

*Proof.* Since $F_\theta^{-1}(u)$ is nondecreasing in $\theta$ for any given $u \in (0, 1)$, $\{F_\theta^{-1}(u) - F_{\theta_0}^{-1}(u) > \varepsilon\}$ has non-zero probability only if $\theta > \theta_0$, for any $\varepsilon \geq 0$. Therefore, from (3.13) we have

$$(F_{\theta_{\mathrm{CD},1}}^{-1}(u) - F_{\theta_0}^{-1}(u))^{+} \overset{\mathrm{sto}}{\leq} (F_{\theta_{\mathrm{CD},2}}^{-1}(u) - F_{\theta_0}^{-1}(u))^{+}. \tag{A.11}$$

Since $Q_i^{-1}(u) = \int_{\theta \in \boldsymbol{\Theta}} F_{\theta}^{-1}(u) dH_i(\theta)$, for $i = 1, 2$, taking expectation on the two sides of (A.11) with respect to $\theta_{\mathrm{CD},1} \sim H_1(\cdot)$ and $\theta_{\mathrm{CD},2} \sim H_2(\cdot)$ respectively leads to (3.16). (3.17) can be derived in the same way. □

## Proof of Corollary 3.3

*Proof.* (3.16) and (3.17) jointly imply

$$|Q_1^{-1}(u) - F_{\theta_0}^{-1}(u)| \overset{\mathrm{sto}}{\leq} |Q_2^{-1}(u) - F_{\theta_0}^{-1}(u)|.$$

or equivalently,

$$(Q_1^{-1}(u) - Q_{\theta_0}^{-1}(u))^2 \overset{\mathrm{sto}}{\leq} (Q_2^{-1}(u) - Q_{\theta_0}^{-1}(u))^2.$$

This further implies

$$\mathbb{E}_{\mathbf{Y}_n}(Q_1^{-1}(u) - Q_{\theta_0}^{-1}(u))^2 \overset{\mathrm{sto}}{\leq} \mathbb{E}_{\mathbf{Y}_n}(Q_2^{-1}(u) - Q_{\theta_0}^{-1}(u))^2.$$

Since this holds for any $u \in (0, 1)$, it immediately follows the result of (3.19) by substituting $u$ with $U \sim$ Uniform$(0, 1)$ and taking expectation on both sides. □

## Proof of Theorem 3.4

*Proof.* The average Kullback-Leibler distance of any density function in the form of $g_{\hat{\theta}}(\cdot)$ to $f_{\theta_0}(\cdot)$ can be expressed as $\bar{D}_{\mathrm{KL}}(f_{\theta_0}|g_{\hat{\theta}}) = \mathbb{E}_{\mathbb{J}}\left\{\log \frac{f_{\theta_0}(Y^*)}{g_{\hat{\theta}}(Y^*)}\right\}$, where the expectation is taken jointly over $\mathcal{Y}^* \times \mathcal{Y}^n$ at the true parameter value $\theta_0$, and thus (3.20) implies

$$\bar{D}_{\mathrm{KL}}(f_{\theta_0}|q_{\hat{\theta}}) - \bar{D}_{\mathrm{KL}}(f_{\theta_0}|f_{\hat{\theta}}) = \mathbb{E}_{\mathbb{J}}\left\{\log \frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)}\right\} \leq \log \mathbb{E}_{\mathbb{J}}\left\{\frac{f_{\hat{\theta}}(Y^*)}{q_{\hat{\theta}}(Y^*)}\right\} \leq 0.$$

□

# Bibliography

Aitchison, J., 1975. Goodness of prediction fit. Biometrika 62, 547–554.

Aitchison, J., Dunsmore, I. R., 1980. Statistical Prediction Analysis. Cambridge University Press, Cambridge, U.K.

Barndor-Nielsen, O. E., Cox, D. R., 1996. Prediction and asymptotics. Bernoulli 2, 319–340.

Battet, H., Fan, J., Liu, H., Lu, J., Zhu, Z., To appear. Distributed estimation and inference with statistical guarantees. The Annals of Statistics.

Beran, R., 1990. Calibrating prediction regions. Journal of the American Statistical Association 85, 715–723.

Berger, J. O., Wolpert, R. L., 1988. The Likelihood Principle, 2nd Edition. Institute of Mathematical Statistics, Haywood, CA.

Bjornstad, J. F., 1990. Predictive likelihood: a review (with discussion). Statistical Science 5, 242–265.

Cakici, N., Fabozzi, F. J., Tan, S., 2013. Size, value, and momentum in emerging stock returns. Emerging Markets Review 16, 46–65.

Chang, K., 2015. Topics in compositional, seasonal and spatial-temporal time series. Ph.D. thesis, Rutgers University.

Chen, X., Xie, M., 2014. A split-and-conquer appproach for analysis of extraordinarily large data. Statistica Sinica 24, 1655–1684.

Claggett, B., Xie, M., Tian, L., 2014. Meta analysis with fixed, unknown, study-specific parameters. Journal of the American Statistical Association 109, 1667–1671.

Cochran, W. G., 1954. The combination of estimates from different experiments. Biometrics 19, 101–129.

Cox, D. R., 1958. Some problems connected with statistical inference. The Annals of Mathematical Statistics 29, 357–372.

Cox, D. R., 1975. Prediction intervals and empirical bayes condence intervals. In: Gani, J. (Ed.), Perspectives Probability and Statistics. Academic Press, London.

Cox, D. R., 2013. Discussion of "confidence distribution, the frequentist distribution estimator of a parameter: a review". International Statistical Review 81, 40–41.

Efron, B., 1986. Why isn't everyone a bayesian. The American Statistician 40, 262–266.

Efron, B., 1993. Bayes and likelihood calculations from confidence intervals. Biometrika 80, 3–26.

Efron, B., 1998. R. a. fisher in the 21st century. Statistical Science 13, 95–122.

Escobar, L. A., Meeker, W. Q., 1999. Statistical prediction based on censored life data. Technometrics 41, 113–124.

Fama, E. F., French, K. R., 1993. Common risk factors in the returns on stocks and bonds. Journal of Financial Economics 33, 3–56.

Fama, E. F., French, K. R., 2012. Size, value, and momentum in international stock returns. Journal of Financial Economics 105, 457–472.

Fama, E. F., French, K. R., 2014. A five-factor asset pricing model. Journal of Financial Economics 116, 1–22.

Fisher, R. A., 1935. The fiducial argument in statistical inference. The Annals of Eugenics 6, 91–98.

Fisher, R. A., 1959. Mathematical probability in the natural sciences. Technometrics 1, 21–29.

Fraser, D. A. S., 2004. Ancillaries and conditional inference. Statistical Science 19, 333–369.

Geisser, S., 1993. Predictive Inference: An Introduction. Chapman & Hall, New York.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., Rubin, D. B., 2013. Bayesian Data Analysis, 3rd Edition. Chapman & Hall/CRC.

Grinold, R. C., Kahn, R. N., 1999. Active Portfolio Management: A Quantitative Approach for Producing Superior Returns and Controlling Risk, 2nd Edition. McGraw-Hill.

Grün, B., Leisch, F., 2007. Finite mixtures of generalized linear regression models. Tech. rep., Department of Statistics, University of Munich.

Gustafson, P., Hossain, S., McCandless, L., 2005. Innovative bayesian methods for biostatistics and epidemiology. Handbook of Statistics, Volume 25 : Bayesian Thinking, Modeling and Computation.

Hall, P., Miller, H., 2010. Bootstrap confidence intervals and hypothesis tests for extrema of parameters. Biometrika 97, 881–892.

Hannah, L. A., Blei, D. M., Powell, W. B., 2011. Dirichlet process mixtures of generalized linear models. Journal of Machine Learning Research 1, 1–33.

Harris, I. R., 1989. Predictive fit for natural exponential families. Biometrika 76, 675–684.

Hastie, T., Tibshirani, R., Friedman, J., 2010. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition. Springer, New York.

Higgins, J., Thompson, S., Deeks, J., Altman, D., 2002. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. Journal of Health Services Research & Policy 7, 51–61.

Higgins, J., Thompson, S. G., Deeks, J. J., Altman, D. G., 2003. Measuring inconsistency in meta-analyses. British Medical Journal 327, 557–560.

Jara, A., Hanson, T. E., Quintana, F. A., 2011. Dppackage: Bayesian semi- and nonparametric modeling in r. Journal of Statistical Software 40.

Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research 15, 2869–2909.

Kleiner, A., Talwalkar, A., Sarkar, P., Jordan, M. I., 2014. A scalable bootstrap for massive data. Journal of the Royal Statistical Society: Series B 76, 795–816.

Lawless, F., Fredette, M., 2005. Frequentist prediction intervals and predictive distributions. Biometrika 92, 529–542.

Lejeune, M., Faulkenberry, G. D., 1982. A simple predictive density function. Journal of the American Statistical Association 77, 654–657.

Liu, D., Liu, R., Xie, M., 2014. Exact meta-analysis approach for discrete data and its application to 2 × 2 tables with rare events. Journal of the American Statistical Association 109, 1450–1465.

Liu, D., Liu, R. Y., Xie, M., 2015. Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. Journal of the American Statistical Association 110, 326–340.

Liu, K., Meng, X. L., 2016. There is individualized treatment. why not individualized inference? Annals of Review of Statistics and its Applications 3, 79–111.

Liu, R. Y., Parelicus, J., Singh, K., 1999. Multivariate analysis by data depth: descriptive statistics, graphics and inference. The Annals of Statistics 27, 783–858.

Murray, G. D., 1977. A note on the estimation of probability density functions. Biometrika 64, 150–152.

Ng, V. M., 1980. On the estimation of parametric density functions. Biometrika 67, 505–506.

Schweder, T., 2007. Confidence nets for curves. Advances in Statistical Modeling and Inference. Essays in Honor of Kjell A. Doksum, 593–609.

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

Schweder, T., Hjort, N. L., 2002. Confidence and likelihood. Scandinavian Journal of Statistics 29, 309–332.

Simmonds, M. C., Higgins, J. P. T., 2007. Covariate heterogeneity in meta-analysis: criteria for deciding between meta-regression and individual patient data. Statistics in Medicine 26, 2982–2999.

Singh, K., Xie, M., Strawderman, W. E., 2001. Confidence distributions - concept, theory and applications. Tech. rep., Deptartment of Statistics and Biostatistics, Rutgers University.

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

Singh, K., Xie, M., Strawderman, W. E., 2007. Confidence distribution (cd) - distribution estimator of a parameter. IMS Lecture Notes-Monograph Series 54, 132–150.

Smith, R. L., 1998. Bayesian and frequentist approaches to parametric predictive inference. In: Bayesian Statistics 6. pp. 589–612.

Tang, L., Zhou, L., Song, P. X.-K., 2016. Method of divide-and-combine in regularised generalised linear models for big data.

van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics 42, 1166–1202.

van deer Vaart, A. W., 1998. Asymptotic Statistics. Vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, U.K.

Vidoni, P., 1998. A note on modified estimative prediction limits and distributions. Biometrika 85, 949–953.

Wang, J. C. M., Hannig, J., Iyer, H. K., 2012. Fiducial prediction intervals. Journal of Statistical Planning and Inference 142, 1980–1990.

Wasserman, L., 2007. Why isn't everyone a bayesian? In: Morris, C. N., Tibshirani, R. J. (Eds.), The Science of Bradley Efron. Springer, New York, pp. 260–261.

Xie, M., 2013. Rejoinder of "confidence distribution, the frequentist distribution estimator of a parameter: a review". International Statistical Review 81, 68–77.

Xie, M., Liu, R. Y., Damaraju, C. V., Olson, W. H., 2013. Incorporating external information in analyses of clinical trials with binary outcomes. The Annals of Applied Statistics 7, 342–368.

Xie, M., Singh, K., 2013. Confidence distribution, the frequentist distribution estimator of a parameter: a review. International Statistical Review 81, 3–39.

Xie, M., Singh, K., Strawderman, W. E., 2011. Confidence distributions and a unifying framework for meta-analysis. Journal of the American Statistical Association 106, 320–333.

Xie, M., Singh, K., Zhang, C., 2009. Confidence intervals for population ranks in the presence of ties and near ties. Journal of the American Statistical Association 104, 775–788.

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

Zhang, C., Zhang, S. S., 2014. Confidence intervals for low-dimensional parameters in high-dimensional linear models. Journal of the Royal Statistical Society: Series B 76, 217–242.