Data Fusion for Heterogeneous Data and Its Applications

by

Jingyi Zhang

(Under the Direction of Wenxuan Zhong)

Abstract

With the rapid development of data storage and cloud computing facilities, volume and velocity are no longer the bottlenecks of big data applications. Variety poses more challenges, as the data we obtain may come from extremely heterogeneous sources. Clearly, simple integration of different databases by collating data is not enough. Innovative data fusion approaches open up a wide range of research opportunities in big data research. This thesis will cover data fusion for large-scale data analysis in the following three levels. Feature level fusion through semi-parametric model for heterogeneous data, data level fusion through optimal transport map for medical image data, and decision level fusion through ensemble learning for medical studies.

Index words: Data fusion, dimension reduction, smoothing splines, optimal transport, ensemble learning

Data Fusion for Heterogeneous Data and Its Applications

by

Jingyi Zhang

B.S., Wuhan University, China, 2011
M.S., Wuhan University, China, 2013

A Thesis Submitted to the Graduate Faculty of the
University of Georgia in Partial Fulfillment of the Requirements for the
Degree.

Doctor of Philosophy

Athens, Georgia

2020

Data Fusion for Heterogeneous Data and Its Applications

by

Jingyi Zhang

Major Professor:	Wenxuan Zhong
Committee:	Ping Ma
	Changying Li
	Bing Li

Electronic Version Approved:

Ron Walcott
Interim Dean of the Graduate School
The University of Georgia
May 2020

Table of Contents

List of Figures

List of Tables

Chapter 1

Data Fusion

Big data is also referred to as 3-v data, where the 3-v represents volume, velocity, and variety. With the rapid development of data storage and cloud computing facilities, volume and velocity are no longer the bottlenecks of big data applications. Variety poses more challenges, as the data that we obtain may come from extremely heterogeneous sources. For example, in medical research and environmental sciences, it is a longstanding practice that data are collected locally at individual data centers (Dalgard et al., 2015; Kim et al., 2014; Rahbar et al., 2017; T. Zhang et al., 2016). Clearly, simple integration of different databases by collating data is not enough. Innovative data fusion approaches open up a wide range of research opportunities in big data research.

Example 1.0.1. *Medical research across multiple medical centers.*

In medical research, a crucial task is to examine how some physiological measurements that affect health. The well known physiological measurements are body temperature, blood pressure, blood sugar and blood lipid levels, certain hormone levels, etc. Many medical studies are constrained by the number of human samples that cannot be obtained in a single medical center. Thus, multi-center study is common for large-scale studies Dalgard et al., 2015; Kim et al., 2014; Rahbar et al., 2017. Consider data collected from different departments in hospitals, such as department of cardiology, department of pediatrics and department of gynecology. There are several challenges when dealing with these datasets: (1) it is obviously that the datasets are extremely biased; (2) even when data acquisition is standardized across centers, there are still center specific or method-specific effects on the measurements (H. H. Zhou et al., 2018); and (3) due to concerns such as data privacy, it is impractical to share data, especially when data originate from different medical centers. The first challenge implies that the fitted model will be unreliable if we only use data from one single department. However, the other two challenges prevent us from direct pooling data in a post hoc manner across mul-

tiple centers. It is crucial to effectively integrate information from multi-center data to draw valid conclusions.

Example 1.0.2. *Weather forecast across different cities.*

Weather forecast is based on data collected in weather stations located in different cities. Because of the unique geographical conditions in each city, data from different cities will exhibit heterogeneous patterns. For example, sunlight will have a dramatic impact on inland temperature, but not in coastal cities. Thus we need to apply different prediction models to different cities. On the other hand, because of the spatial correlation, borrowing information from neighboring cities can make the prediction model more accurate.

Example 1.0.3. *Medical image analysis.*

In medical image analysis, researchers usually suggest combining images that are taken under different techniques or from different directions to obtain a more convincing result. One example is the neuroimaging data generated from different magnetic resonance imaging (MRI) techniques, which provide different views of brain function. Another example is the echocardiogram data under different angles, which provide different views of heart structure. In order to understand the story behind the varied images, we need to generate a fused image that can well capture the global structure, and detect abnormalities based on the fused result.

Example 1.0.4. *Disease screening based on various diagnosis.*

We all have such experiences that different clinicians may give different diagnoses using the same examination results. In this situation, instead of simply believing one diagnosis, we are more likely to take the aggregate of multiple conclusions. Moreover, we would probably browse a few websites or go to more clinicians for more diagnoses before we reach a final conclusion.

For those cases shown in the examples, how to fuse information from different datasets or sources are of particular interest. Data fusion can be performed at three different levels: data level, feature level, and decision level. The first two examples are cases on feature level fusion, the third example is a data level fusion, and the last one illustrates a decision level fusion. Compared with the feature level fusion and decision level fusion, data level fusion preserves more original information and is more attractive for information extraction. Alternatively, feature level fusion can discover new patterns and form new insights and consequently is often used for information integration. When systematic decisions need to be made, we need decision fusion methods to focus on the valuable information that can generate optimal decisions.

In this thesis, we introduce a set of statistical tools for data fusion. We first consider feature level fusion through extending multi-index models for

heterogeneous data. Then, we introduce an optimal transport approximation method for data level fusion of medical image data. Finally, we propose an ensemble machine learning method for decision level fusion.

Chapter 2

Feature Space Fusion and Its Application in Heterogeneous Scattered Data

With the rapid development of decentralized computing technology, we are facing data that come from extremely different sources. This type of data is referred to as the scattered data. Scattered data can be very heterogeneous. For example, the patients' age follows completely different distributions in children's hospitals and in senior centers. Clearly, the simple integration of different databases by collating data is not enough. There is an urgent need for data fusion methods to link across different databases.

Multi-index model, which assumes that the response variable depends on q linear combinations of predictors or q hidden features through some unknown link functions η, is intensively studied due to its model interpretability and flexibility. However, how to estimate the feature space and η across different databases is still an open question and is the key that dictates the ultimate performance of data fusion enterprise. In this article, we present a general feature space fusion framework to address the heterogeneity issues for the scattered data. By iteratively estimating and fusing the feature space spanned by the q linear combinations for each source data, we can obtain a fused feature space that includes all regression related information. We show theoretically that the fused feature space is asymptotically consistent under some mild regularity conditions. We also establish the asymptotic convergence rate of the proposed algorithm. As we do not impose any assumption on the link function between the response variable and the predictors for each source data, the decision system obtained

can be considered a model-free decision system. Furthermore, as we allow the predictor distributions and link functions to be variable for different source data, the method can be naturally applied to transfer learning, which can be extremely challenging for regressions that beyond linear or parametric models.

2.1 Introduction

With the fast development of cloud computing and data storage, we quickly step in the big data era. A big challenge that we are currently facing is that data are often collected from extremely different sources. How to extract information from very heterogeneous data sources poses extensive research challenges to the statistician. Clearly, the simple integration of different databases by collating data is not enough. For example, the prognostic effect of dobutamine stress echocardiography differs dramatically among different age groups (Bernheim et al., 2011). Thus, predicting the prognostic effect by collating data from children's hospitals and from general hospitals can be very misleading due to Simpson's paradox, shown in figure 2.1. These challenges become extremely serious in decentralized computing where data have to be computed at individual nodes rather than transmitted to a center due to data privacy, security, ownership, and transmission cost concerns (Fan et al., 2017; H. H. Zhou et al., 2018; H. H. Zhou et al., 2017). How to properly fuse the information of two sources of data open up a wide range of options that may dictate the ultimate performance of big data enterprise.

The multi-index model has been studied for years in statistical literature. It has been intensively studied in econometrics and statistics communities for estimating consumer index and for dimension reduction. Classical theory on the multi-index model requires that both the predictor $\boldsymbol{x}$ and the response y are homogenous. Admittedly, this assumptions do not hold in many instances, especially for modern applications with data coming from multiple centers, such as electronic medical record (EMR) studies in medical researches, price evaluation studies in financial services, real-time monitoring studies in environmental sciences, and etc. (Dalgard et al., 2015; Kim et al., 2014; Rahbar et al., 2017; T. Zhang et al., 2016). Thus, how to obtain the features or indices parameters that can be used for all types of source data is very important.

Recently, with the burgeoning of transfer learning, which assumes different distributions between the training data and testing data, our feature space fusion framework is becoming even more important. For example, if we can obtain a fused feature space that can be used under any geographical locations such as inland locations with less sunshine and coastal locations with more sun-

shine, the weather prediction can be more accurate and simple for local weather stations. Moreover, if we can find common feature space for object images on both sunny days and heavy raining days, transfer learning between sunny days and heavy raining days can be much more simplified, and the image recognition AI tools based on it can be more appropriate for in-field deployment.

To overcome the aforementioned problems, we proposed a regression-based feature space fusion framework, under which we can not only handle the heterogeneity such as the domain shift and center-specific relationship of the different sources of data, but also achieve the same estimation efficiency as we can have for the homogeneous source of data. In particular, we assume that the response of the sth data center y_s depends on the same feature space fused by linear combinations (indices) of domain-specific predictors $\boldsymbol{x}_s$ through domain-specific unknown link functions $\eta_s(\cdot)$. We refer to the proposed model as the multi-center multi-index (MCMI) model. To estimate the MCMI model, we developed a feature space fusion method that integrates the coefficients of the predictors without estimate the unknown source-specific link functions using the gradient methods.

Parameter fusion is pioneered by (Haghighat et al., 2016; Hall & Llinas, 1997; Hall & McMullen, 2004), which proposes to integrate parameters rather than raw observations for information integration. Under the assumption of some pre-specified parametric models, e.g., regression model, or quantitative methods, e.g., gradient descent. Parameter fusion methods are usually conducted through three steps: (1) specify a parametric model or a quantitative method that is suitable for pooled-data; (2) within each data center, calculate the local estimators of the unknown parameters respecting to the pre-specified models or methods; and (3) integrate the local estimators across all local centers to obtain a fused estimator. Theoretically, most of the parameter fusion methods were shown that their estimators are consistent with the pooled-data estimator. As a result, parameter fusion methods have been utilized by many methods as a crucial component. These methods range from distributed gradient descent to divide-and-conquer techniques (Battey et al., 2015; Blanchard & Mücke, 2016; X. Chen & Xie, 2014; Fan et al., 2017; Guo et al., 2017; Lee et al., 2017; Y. Zhang et al., 2015).

Despite the wide applications, the performance of parameter fusion methods highly relies on two critical assumptions, i.e., (1) the predictor variables are i.i.d. across the local centers, and (2) there exists one global model/method that is suitable for the pooled data. It is not surprising that, when these assumptions are violated due to the data heterogeneity, most of these methods may provide misleading results. Consider the Simpson's paradox as an example. In Figure 1,

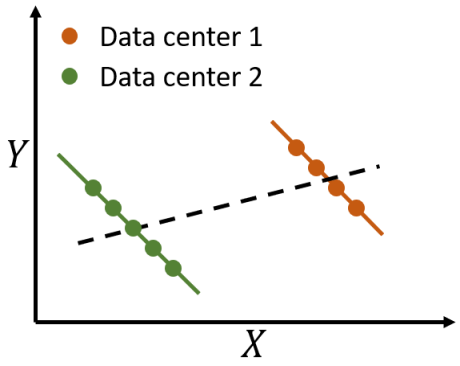

Figure 2.1: Simpson's paradox

let the red dots and green dots represent the data collected in two different local centers, respectively. Suppose a researcher postulates linear regression models in both local centers, and also fits a linear regression model for the pooled-data. One can observe that there is a significant difference between the fitted local models, i.e., marked as red and green lines. One can also observe that the pooled-data regression line (black dashed line) deviates severely from both of the fitted local models. These two observations, as stated in Simpson's paradox, are attributed to the existence of the data heterogeneity in scattered data. As a result, for heterogeneous scattered data, neither fusing the local parameters from the local models nor building a simple global model for pooled-data yield valid results.

The contribution of our space fusion method to the development of multi-source data learning is two-fold. First, it allows the domain shifting of predictor space and also allows center-specific link functions between different centers. This assumption relaxation and the MCMI model that our feature space fusion method relies on includes fully nonparametric models as special cases. Therefore, it can be considered as a complete model-free data fusion procedure applicable for linking any heterogeneous database. Second, as discussed in section 3, we proposed both decentralized algorithms and distributed algorithms for the data fusion procedure to handle big data processing. In general, we believe that our feature space fusion method should become an indispensable member of the repository of transfer learning and scatter data learning and recommend its broad use. As our feature space fusion method can achieve the same estimation efficiency for heterogeneous data as homogenous data, we recommend to use it as a safeguard against possible model-misidentification. In the following, we

outlined the advantages of our feature space fusion (FSF) algorithm over the existing parameter fusion algorithm in more details.

1. FSF considers a general multi-index model, which accommodates both types of data heterogeneity through different predictor distributions and center-specific nonparametric link functions.
2. Rather than fusing the parameters, FSF aims to fuse the feature space. Compared with the fused parameters, the fused feature space takes a weaker form to reflect the "correspondence" among the local models.
3. FSF uses minimum average variance estimation technique to estimate local feature spaces. The idea iteratively searches for the estimates of the feature space and the link function.
4. FSF can be naturally applied on the scenario that the local data centers are connected through a *decentralized* topology. With the fused feature space, FSF can be further applied to *transfer learning*.

Besides the algorithm improvement, we also establish two theoretical results to ensure the consistency and efficiency of the proposed FSF method. We show that in the worst case of data heterogeneity, our method can be as efficient as the "best" local model, i.e., the one with the highest efficiency among all. We also showed that when data is homogeneous, our method can perform as well as using the pooled data.

2.2 Model Setup

Single center multi-index model. Given the observations $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$ for $i = 1, \cdots n$, we first consider the following multi-index model (Elton et al., 1977; B. Li, 2018; K.-C. Li, 2000) for a single data center,

$$y_i = \eta(\boldsymbol{x}_i^T\boldsymbol{\beta}_1, \ldots, \boldsymbol{x}_i^T\boldsymbol{\beta}_q) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where η is an unknown link function , $\boldsymbol{\beta}_j$s are p-dimensional orthogonal regression indices of unit length, q is an integer less than p, and $\{\epsilon_i\}_{i=1}^n$ are stochastic errors with mean zero and variance σ^2. Notice that given $\boldsymbol{x}_i^T\boldsymbol{\beta}_1, \ldots, \boldsymbol{x}_i^T\boldsymbol{\beta}_q$, y_i and $\boldsymbol{x}_i$ are independent. Thus, the subspace spanned by $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ contains all information of $\boldsymbol{x}_i$ that is related to y_i. In the rest of the chapter, we refer the column space spanned by $\mathbf{B}$ as the "feature space" and denote it by $\mathcal{S}(\mathbf{B})$. Model (2.1) is very general, which include single index models (q=1), linear regression models ($\eta(x) = x$) and non-parametric models ($\mathbf{B} = I$) as its special cases.

The multi-index model has been extensively studied in statistics and econometric societies. The main stream research focused on the mean response re-

gression $E(y_i|\boldsymbol{x}_i) = \eta(\boldsymbol{x}_i^T\boldsymbol{\beta}_1, \ldots, \boldsymbol{x}_i^T\boldsymbol{\beta}_q)$. There are primarily two types of methods for identifying the indices, which are the M-estimation methods and the gradient methods. Treating η as an infinite-dimensional nuisance parameter, the M-esimation methods first obtain a nonparametric estimate $\hat{\eta}$ given $\mathbf{B}$, and then minimize a certain target functional of $\mathbf{B}$, $\hat{\eta}$ and the data to generate an estimate of $\mathbf{B}$. Two examples are the semiparametric maximum likelihood estimate and the semiparametric least squares estimate. The properties of these two estimates such as their asymptotic efficiency have been studies in the literature Ichimura and Todd, 2007. In spite of their nice theoretical properties, these estimates are rarely implemented in practice. The main reason for this is that the computation of these estimates require solving hard optimization problems and is easily compromised by the curse of dimensionality.

The second group includes those methods that directly utilize the gradient of $E(y_i|\boldsymbol{x}_i)$, denoted by $m'(\boldsymbol{x}_i)$. Hence, they are referred to as the gradient methods. One basic gradient method is the so called average derivative method studied by Powell et al., 1986. For single index model, the average derivative method is developed based on the observation that $m'(\boldsymbol{x}) = \boldsymbol{\beta}_1\eta'(\boldsymbol{\beta}_1\boldsymbol{x})$. In other words, $\boldsymbol{\beta}_1$ is proportional to $m'(\boldsymbol{x})$ for every $\boldsymbol{x}$. By taking expectation of $m'(\boldsymbol{x})$ and applying integration by parts, we have $\boldsymbol{\beta}_1$ is proportional to $E(m'(\boldsymbol{x})) = -E(yf'(\boldsymbol{x})/f(\boldsymbol{x}))$ where $f(\boldsymbol{x})$ is the marginal density function of $\boldsymbol{x}$. Hence $\boldsymbol{\beta}_1$ can be estimated by the sample average of $yf'(\boldsymbol{x})/f(\boldsymbol{x})$ with f replaced by a nonparametric estimate. The resulted estimate is called the average derivative estimate. To generalize the gradient method to multiple index, Xia et al., 2002 propose to use local linear expansion to approximate $m(\boldsymbol{x})$ and obtain a tentative estimate of $\mathbf{B}$ denoted by $\hat{\mathbf{B}}$. Then, they improve $\hat{\mathbf{B}}$ by minimizing the average variance estimate, i.e. regression residuals $E\{\left[y_i - \eta(\mathbf{B}^T\boldsymbol{x}_i)\right]^2\}$. The updates run iteratively until converges.

More specifically, we use figure 2.2 as a toy example to illustrate how to estimate $\mathbf{B}$ in a single data center. Here, we assume the true function $\eta(\cdot) = (\cdot)^2$ (green surface in panel (a)) and the true $\mathbf{B} = (1, 1)^T$ (red arrow in panel (a)). Here because $\mathbf{B}$ is a vector, we denote it as $\boldsymbol{\beta}$. Then, we generate observations (yellow dots) according to model (2.1). Figure 2.2(b) demonstrates the iterative steps of the estimation algorithm. To be specific, given the estimates $\eta^{[k-1]}$ (gray dashed curve) and $\boldsymbol{\beta}^{[k-1]}$ (red arrow), we first calculate $\boldsymbol{\beta}^{[k]}$ by minimizing the least square functional,

$$\boldsymbol{\beta}^{[k]} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i}^{n} ||y_i - \eta^{[k-1]}(\boldsymbol{x}_i^T\boldsymbol{\beta})||^2,$$

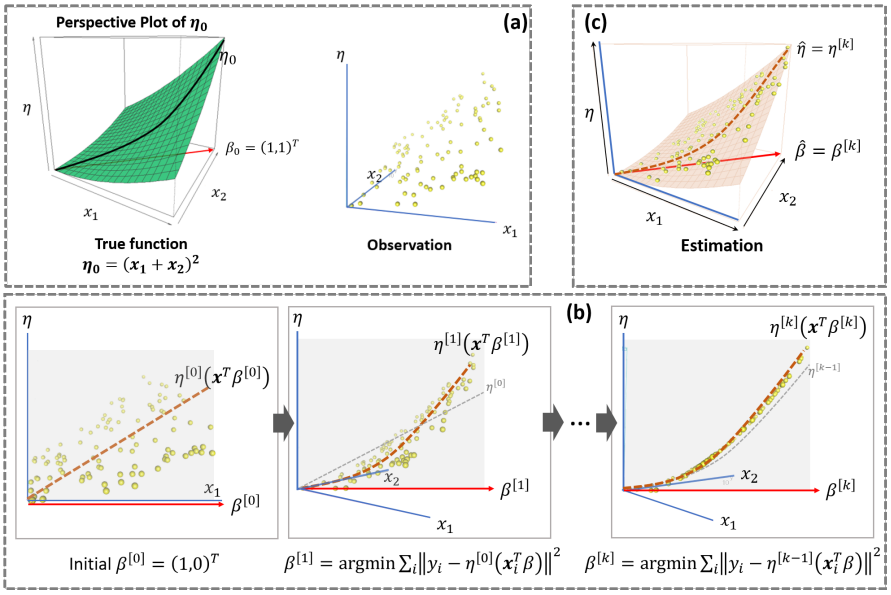

Figure 2.2: Illustration of the estimation algorithm within one single data center.

where $||\cdot||$ denotes the Euclidean norm. Then, we estimate $\eta^{[k]}$ (red dashed curve) using the technique of local linear regression smoother based on the updated $\boldsymbol{x}^T\boldsymbol{\beta}^{[k]}$ (yellow dots). Figure 2.2(c) shows the final output $\widehat{\eta}$ (red dashed curve) and $\widehat{\boldsymbol{\beta}}$ (red arrow), when the algorithm converge. Details of the estimation for model (2.1) are shown in supplementary materials.

To estimate model (2.1), consider the local linear smoother of η,

$$\eta(\mathbf{B}^T\boldsymbol{x}_j) = \sum_{i=1}^{n}\{a_j + \boldsymbol{b}_j^T\mathbf{B}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\}w_{ij},$$

where a_j and $\boldsymbol{b}_j$ are the coefficients of Taylor expansion of η at point $\mathbf{B}^T\boldsymbol{x}_j$, and

$$w_{ij} = \frac{K_h\{\mathbf{B}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\}}{\sum_{l=1}^{n} K_h\{\mathbf{B}^T(\boldsymbol{x}_l - \boldsymbol{x}_j)\}}$$

is the kernel weight. Here, $K(\cdot)$ is a kernel function and $K_h = h^q K(\cdot/h)$ with bandwidth h. In practise, we opt to use the Gaussian kernel and choose the bandwidth to be $h = O(n^{-1/(p+4)})$. For simplicity, we denote

$$\zeta(\mathbf{B}, a_j, \boldsymbol{b}_j) := \sum_{j=1}^{n}\sum_{i=1}^{n}\left\{y_i - \left[a_j + \boldsymbol{b}_j^T\mathbf{B}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\right]\right\}^2 w_{ij}.$$

Details of the MAVE algorithm are shown in Algorithm 1.

Algorithm 1 MAVE algorithm

Input: $\{\boldsymbol{x}_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$ as defined in model (1), kernel matrix $K(\cdot)$, bandwidth h.
$k \leftarrow 0$, randomly initialize $\mathbf{B}^{[0]} \in \mathbb{R}^{p\times q}$.
repeat
(i) Given $\mathbf{B}^{[k]}$, for $1 \le i, j \le n$, calculate the kernel weights

$$w_{ij}^{[k+1]} = \frac{(\mathbf{B}^{[k]})^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}{\sum_{i=1}^n K_h(\mathbf{B}^{[k]})^T(\boldsymbol{x}_i - \boldsymbol{x}_j)}.$$

(ii) Estimate $\eta^{[k+1]}$ through kernel method by solving the optimization problem

$$a_j^{[k+1]}, \boldsymbol{b}_j^{[k+1]} = \operatorname{argmin} \zeta(\mathbf{B}^{[k]}, a_j, \boldsymbol{b}_j)$$

through weighted least squares.
(iii) Estimate $\mathbf{B}^{[k+1]}$ through solving the optimization problem

$$\mathbf{B}^{[k+1]} = \underset{\mathbf{B}^T\mathbf{B}=I_q}{\operatorname{argmin}} \zeta(\mathbf{B}, a_j^{[k+1]}, \boldsymbol{b}_j^{[k+1]}).$$

(iv) $k \leftarrow k + 1$.
until converge.

Multiple centers multi-index model. For simplicity, we refer to each individual data center that can store, process, and transport data as a node. Given S nodes, we assume that the observations $\{y_{si}, \boldsymbol{x}_{si}\}_{i=1}^{n_s}$ on node s follows

$$y_{si} = \eta_s(\boldsymbol{x}_{si}^T\boldsymbol{\beta}_1, \ldots, \boldsymbol{x}_{si}^T\boldsymbol{\beta}_q) + \epsilon_{si}, \tag{2.2}$$

where η_s represents the unknown link function for sth node, and $\{\epsilon_{si}\}_{i=1}^{n_s}$ are mean zero variance σ^2 error terms independent of $\boldsymbol{x}_{si}$. It is clear that model (2.2) can well address two different types of nodes' heterogeneity: (a) the predictor variables may have heterogeneous distributions across nodes, (b) each node may have center-specific link functions. The only constraint here is that all nodes share the same feature space, or in another word, the regression indices $\mathbf{B}$ is the same across different data centers. This constraint, we believe is reasonable and necessary in reality. If this constraint is violated, there is no common information shared between nodes and correspondingly, there is no need for data fusion.

2.3 Feature Space Fusion

We propose the Feature Space Fusion (FSF) algorithm to estimate model (2.2). Specifically, FSF aims to fuse the estimated feature space across all local nodes, as illustrated in Figure 2.3. Suppose the sth node is connected to the $(s-1)$th node and the $(s+1)$th node in the data network, so that they can communicate with each other. For the sth node, in the kth iteration, FSF first estimates $\eta_s^{[k]}$ and $\mathbf{B}_s^{[k]}$ as illustrated in figure 2.2. The regression indices $\mathbf{B}_s^{[k]}$ is then transmitted to the connected nodes, illustrated by the red arrows. In the meanwhile, these connected nodes will transmit their local estimates, i.e., $\mathbf{B}_{s-1}^{[k]}$ and $\mathbf{B}_{s+1}^{[k]}$, to the sth node. Then the feature spaces $\mathcal{S}(\mathbf{B}_{s-1}^{[k]})$, $\mathcal{S}(\mathbf{B}_s^{[k]})$, and $\mathcal{S}(\mathbf{B}_{s+1}^{[k]})$ are fused together to obtain the fused feature space $\mathcal{S}(\mathbf{B}_s^{[k+1]})$, then the orthogonal bases of $\mathcal{S}(\mathbf{B}_s^{[k+1]})$ are extracted as the common regression indices $\mathbf{B}_s^{[k+1]}$, illustrated by the green arrows. Here, the matrix $\mathbf{B}_s^{[k+1]}$ should be regarded as a set of basis respecting to the "average" feature space of the feature spaces $\mathcal{S}(\mathbf{B}_{s-1}^{[k]})$, $\mathcal{S}(\mathbf{B}_s^{[k]})$, and $\mathcal{S}(\mathbf{B}_{s+1}^{[k]})$. The fused estimate is then used to calculate the link function $\eta_s^{[k+1]}$, until the algorithm converges.

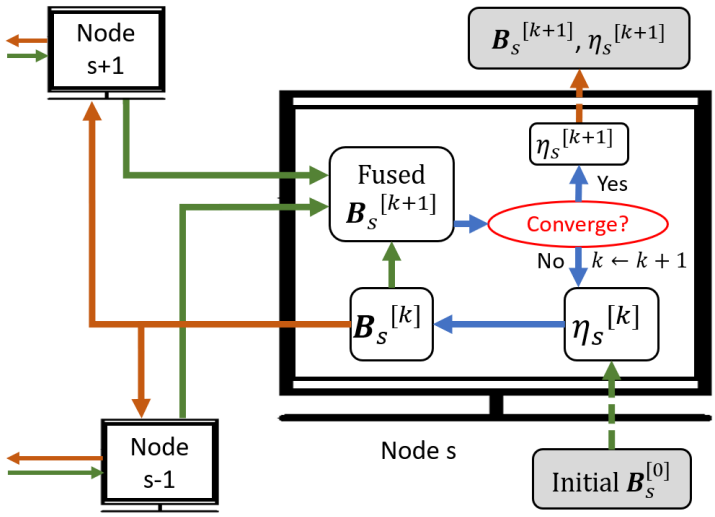

Figure 2.3: Illustration of Algorithm 2.

We now present some essential notations, followed by a detailed algorithm of FSF. Let $\mathbf{J} \in \mathbb{R}^{S\times S}$ be the adjacent matrix of the scattered data structure, i.e. $\mathbf{J}_{ij} = 1$ when there is connection between node i and node j, otherwise, $\mathbf{J}_{ij} = 0$. Let $\mathbf{J}_s = (\mathbf{J}_{s1}, ..., \mathbf{J}_{sS})^T$ be the s-th column of $\mathbf{J}$, and $\{s_1, ..., s_{L_s}\}$ indicate the L_s nodes that are connected to s-th node, i.e $\mathbf{J}_{sk} = 1$, for $k \in \{s, s_1, ..., s_{L_s}\}$, otherwise, $\mathbf{J}_{sk} = 0$. In order to describe the aggregation, we

define a function $\Gamma(\cdot,\cdot)$, such that

$$\Gamma((\mathbf{B}_1,...,\mathbf{B}_S),\mathbf{J}_s)=(\mathbf{B}_s,\mathbf{B}_{s_1},...,\mathbf{B}_{s_{L_s}}),$$

then the fused feature space in sth node can be calculated by

$$\Gamma((\mathbf{B}_1,...,\mathbf{B}_S),\mathbf{J}_s)\mathbf{W}_s,$$

with $\mathbf{W}_s$ as the given node-specified weight matrix. In simple cases, for example, when $q=1$, or when $\mathbf{B}_s$ is sparse enough and the permutation of columns of $\mathbf{B}_s$ is carefully defined, we choose $\mathbf{W}=(\frac{n_s}{\tilde{n}},\frac{n_{s_1}}{\tilde{n}},\ldots,\frac{n_{s_{L_s}}}{\tilde{n}})^T\otimes I_q$, where $\tilde{n}=n_s+\sum_{l=1}^{L_s}n_{s_l}$ is the pooled sample size for node s and its connected nodes. When there is no prior assumption on the structure of $\mathbf{B}_s$, η_s and $\mathbf{B}_s$ are confounding with each other in model (2.2). Consider the following toy example $y=\frac{\boldsymbol{x}^T\boldsymbol{\beta}_1}{\boldsymbol{x}^T\boldsymbol{\beta}_2}+\epsilon$. This model can be written as $y=\eta(\mathbf{B}^T\boldsymbol{x})+\epsilon$, with $\mathbf{B}=(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2)$ and $\eta(z_1,z_2)=z_1/z_2$; however, the model can also be written as $y=\eta^*\big((\mathbf{B}^*)^T\boldsymbol{x}\big)+\epsilon$, with $\mathbf{B}^*=(\boldsymbol{\beta}_2,\boldsymbol{\beta}_1)$ and $\eta^*(z_1,z_2)=z_2/z_1$. In such cases, the weight matrix $\mathbf{W}_s$ will be more complicated. To integrate $\mathbf{B}_{\cdot}^{[k]*}$ from different nodes without mismatching the columns, we choose $\mathbf{W}_s$ to be the eigenvectors corresponding to the first q eigenvalues of the matrix $(\mathbf{B}_s^{[k]*},\mathbf{B}_{s_1}^{[k]*},\ldots,\mathbf{B}_{s_{L_s}}^{[k]*})^T(\mathbf{B}_s^{[k]*},\mathbf{B}_{s_1}^{[k]*},\ldots,\mathbf{B}_{s_{L_s}}^{[k]*})$. The detailed FSF algorithm is provided in Algorithm 2.

Algorithm 2 Feature Space Fusion (FSF) Algorithm

Input: $\{\boldsymbol{x}_{si}\}_{i=1}^{n_s}$, $\{y_{si}\}_{i=1}^{n_s}$ as defined in model (2.2), dimension q, adjacent matrix $\mathbf{J}$, weight matrix $\mathbf{W}_s$.

$k\leftarrow 0$, randomly initialize $\mathbf{B}_s^{[0]}\in\mathbb{R}^{p\times q}$.

repeat

(i) Given $\mathbf{B}_s^{[k]}$, calculate $\eta_s^{[k]}$ through kernel methods with details relegated to algorithm 1.

(ii) Calculate $\mathbf{B}_s^{[k]*}$ based on $\eta_s^{[k]}$.*

(iii) Let $\mathbf{J}_s$ be the s-th column of $\mathbf{J}$. Calculate

$$\mathbf{B}_s^{[k+1]}=\Gamma((\mathbf{B}_1^{[k]*},...,\mathbf{B}_S^{[k]*}),\mathbf{J}_s)\mathbf{W}_s.$$

(iv) $k\leftarrow k+1$.

until converge on all nodes.

Implementation details. Since the dimension q is unknown, we estimate it through minimizing

$$CV(q) = \frac{1}{n}\sum_{j=1}^{n}\left(y_j - \frac{\sum_{i=1,i\neq j}^{n} K_{h_q}^{(i,j)} y_i}{\sum_{i=1,i\neq j}^{n} K_{h_q}^{(i,j)}}\right)^2,$$

where $K_{h_q}^{(i,j)} = K_{h_q}\{\hat{\boldsymbol{\beta}}_1^T(\boldsymbol{x}_i - \boldsymbol{x}_j), \ldots, \hat{\boldsymbol{\beta}}_d^T(\boldsymbol{x}_i - \boldsymbol{x}_j)\}$ with $K_{h_q} = h^q K(\cdot/h)$ being a kernel function with bandwidth h, for $q = 1, \ldots, p$. In practise, we opt to use the Gaussian kernel and choose the bandwidth to be $h = O(n^{-1/(p+4)})$.

Under the scattered data scenario, for each $q \in \{1, 2, \ldots, p\}$, we first calculate $CV_s(q)$, in sth node. Then we exchange $CV_s(q)$ with its L_s connected nodes to obtain the average,

$$\overline{CV}_s(q) = \frac{1}{L_s + 1}\left(CV_s(q) + \sum_{l=1}^{L_s} CV_{s_l}(q)\right).$$

When Algorithm 2 converges, all nodes will be synchronized with the same $\overline{CV}(q)$. As a result, we choose q by

$$\hat{q} = \underset{q\in\{1,\ldots,p\}}{\operatorname{argmin}}\, \overline{CV}(q).$$

2.4 Theoretical Results

We use the following metric, originally used in Xia et al., 2002, to measure the distance between the estimated feature space $\mathcal{S}(\widehat{\mathbf{B}}) \in \mathbb{R}^{\hat{q}\times p}$ and the true feature space $\mathcal{S}(\mathbf{B}_0) \in \mathbb{R}^{q\times p}$,

$$m(\mathcal{S}(\widehat{\mathbf{B}}), \mathcal{S}(\mathbf{B}_0)) = \begin{cases} ||(\mathbf{I}_q - \mathbf{B}_0\mathbf{B}_0^T)\widehat{\mathbf{B}}|| & \text{if } \hat{q} < q \\ ||(\mathbf{I}_{\hat{q}} - \widehat{\mathbf{B}}\widehat{\mathbf{B}}^T)\mathbf{B}_0|| & \text{if } \hat{q} \geq q \end{cases}$$

Let $\widehat{\mathbf{B}}_D$ be the estimator of $\mathbf{B}_0$ through Algorithm 2 and $\widehat{\mathbf{B}}_s$, $s = 1, \ldots, S$, be the estimator of $\mathbf{B}_0$ in the sth node using its local data. We further denote $m_D = m(\mathcal{S}(\widehat{\mathbf{B}}_D), \mathcal{S}(\mathbf{B}_0)))$ and $m_s = m(\mathcal{S}(\widehat{\mathbf{B}}_s), \mathcal{S}(\mathbf{B}_0)))$.

To achieve the main theoretical results in this chapter, we need the following regularity conditions. These conditions are widely used in sufficient dimension reduction literature, and we refer to B. Li, 2018 for details.

Condition 2.4.1. *(i)* $E(\|\boldsymbol{x}_s\|^k < \infty)$ *for all* $k > 0, s = 1, \ldots, S$*;*

(ii) The third derivatives of $E(\boldsymbol{x}_s|y_s)$ *and* $E(\boldsymbol{x}_s\boldsymbol{x}_s^T|y_s)$ *are bounded and continuous for* $s = 1, \ldots, S$*.*

Condition 2.4.2. *Denote $f_{\boldsymbol{x}_s}$ and f_{y_s} as the density function of $\boldsymbol{x}_s$ and y_s for $s = 1, \ldots, S$.*

(i) $f_{\boldsymbol{x}_s}$ has bounded fourth derivative;

(ii) $f_{\boldsymbol{x}_s}$ is bounded away from 0 in a neighborhood $\mathcal{D}$ around 0;

(iii) f_{y_s} has bounded derivative;

(iv) f_{y_s} is bounded away from 0 on a compact support.

Condition 2.4.3. *(i) The conditional density $f_{\boldsymbol{x}_s|y_s}$ is bounded for all $s = 1, \ldots, S$;*

(ii) The conditional density $f_{(\boldsymbol{x}_{s0}, \boldsymbol{x}_{sk})|(y_{s0}, y_{sk})}$ is bounded for all $k \geq 1$ and $s = 1, \ldots, S$.

Condition 2.4.4. *The third derivatives of η_s is bounded and continuous for $s = 1, \ldots, S$.*

Condition 2.4.5. *The kernel function $K(\cdot)$ is a spherical symmetric density function with a bounded derivative, and all the moments of $K(\cdot)$ exist.*

Condition 2.4.2 is needed for the uniform rate of consistency of the kernel smoothing methods. Condition 2.4.3 is needed for kernel estimation of dependent data. Condition 2.4.4 is imposed to meet the continuous requirement for kernel smoothing. Condition 2.4.5 is satisfied by most of the commonly used kernel functions.

Convergence of the FSF algorithm. Suppose model (2.2) holds and there are $n_i, i = 1, \ldots, S$ observations in the sth node, respectively. The following theorem gives the convergence results of the FSF algorithm.

Theorem 2.4.1. *Under conditions 2.4.1 - 2.4.5, Algorithm 2 converges. Furthermore, one has*

$$\lim_{\min\{n_1, \ldots, n_S\} \to \infty} P(m_D - m_s > 0) = 0, s = 1, \ldots, S.$$

Theorem 2.4.1 shows the proposed algorithm converges and is consistently superior to all the local estimations.

Theorem 2.4.2. *Suppose in each node s, $\boldsymbol{x}_s$ in model (2.2) has a density with compact support, $s = 1, \ldots, S$. Assume conditions 2.4.1 - 2.4.5 hold, when $n_s \to \infty$, for $s = 1, ..., S$, one has*

$$\hat{q} \xrightarrow{p} q,$$

where "$\xrightarrow{p}$" means converging in probability.

Theorem 2.4.2 shows the estimated dimension for the proposed algorithm converges to the true dimension.

Convergence rate of the feature space estimator. Let O_p be the order in probability, which is similar to O but for random variables. Let $\tilde{n} = n_s + \sum_{l=1}^{L_s} n_{s_l}$ be the pooled sample size for node s and its connected nodes. We denote $n = \sum_{s=1}^{S} n_s$ as the pooled sample size for all nodes. When $q = 1$, or when the permutation of columns of $\mathbf{B}_s$ are carefully defined, we mentioned that the node-specified weight matrix $\mathbf{W}_s$ can be defined as $\mathbf{W}_s = (n_s/\tilde{n}, n_{s_1}/\tilde{n}, \ldots, n_{s_{L_s}}/\tilde{n})^T \otimes I_d$. The following theorem gives the convergence rate of the proposed feature space estimator.

Theorem 2.4.3. *Suppose* $h_s = O(n_s^{-1/(p+4)})$ *for* $s = 1, \ldots, S$, *and* $\mathbf{W}_s = (n_s/\tilde{n}, n_{s_1}/\tilde{n}, \ldots, n_{s_{L_s}}/\tilde{n})^T \otimes \mathbf{I}_d$. *Under conditions 2.4.1 - 2.4.5, when* $p \geq 2$, *one has*

$$m(\mathcal{S}(\widehat{\mathbf{B}}_D), \mathcal{S}(\mathbf{B}_0)) = O_p\left(n^{-\frac{3}{p+4}} \log n\right).$$

Theorem 2.4.3 states that under the regularity conditions, the estimation of the proposed method can achieve efficiency at the same rate as that of the pooled sample estimator.

2.5 Experimental Studies

Simulations without data heterogeneity. We first consider three cases under the scenario that the data heterogeneity does not exist. Under this scenario, the estimation using pooled-data is expected to be the best. We set the pooled-data estimation as the baseline to evaluate our method.

Case 1. We consider the following model with $q = 2$ and the sample size equals 500 in each node

$$\begin{aligned} &y_{si} = \boldsymbol{\beta}_1^T \boldsymbol{x}_{si}(\boldsymbol{\beta}_1^T \boldsymbol{x}_{si} + \boldsymbol{\beta}_2^T \boldsymbol{x}_{si} + 1) + 0.5\epsilon_{si}, \\ &i = 1, \ldots, 500, \quad s = 1, 2. \end{aligned}$$

Let $p = 10$, $\boldsymbol{\beta}_1 = (0, 1, 0, \ldots, 0)^T$, $\boldsymbol{\beta}_2 = (1, 0, \ldots, 0)^T$ and $\mathbf{B}_0 = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$. The predictor variables $\boldsymbol{x}_{si}$ are generated from a multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_p)$. For each s, $\{\epsilon_{si}\}_{i=1}^n$ are i.i.d. random errors with the mean equals zero and the variance equals one.

Case 2. We consider the following model

$$\begin{aligned} &y_{si} = \frac{\boldsymbol{\beta}_1^T \boldsymbol{x}_{si}}{0.5 + (\boldsymbol{\beta}_2^T \boldsymbol{x}_{si} + 1.5)^2} + 0.5\epsilon_{si}, \\ &i = 1, \ldots, 500, \quad s = 1, 2, \end{aligned}$$

with other settings analogous to the ones in Case 1.

Case 3. We consider the case when $p = 10$ and $q = 4$,

$$y_{si} = \boldsymbol{\beta}_1^T \boldsymbol{x}_{si} (\boldsymbol{\beta}_2^T \boldsymbol{x}_{si})^2 + (\boldsymbol{\beta}_3^T \boldsymbol{x}_{si})(\boldsymbol{\beta}_4^T \boldsymbol{x}_{si}) + 0.5\epsilon_{si},$$
$$i = 1, \ldots, 500, \quad s = 1, 2,$$

where

- $\boldsymbol{\beta}_1 = (1, 2, 3, 4, 0, 0, 0, 0, 0, 0)^T / \sqrt{30}$,
- $\boldsymbol{\beta}_2 = (-2, 1, -4, 3, 1, 2, 0, 0, 0, 0)^T / \sqrt{35}$,
- $\boldsymbol{\beta}_3 = (0, 0, 0, 0, 2, -1, 2, 1, 2, 1)^T / \sqrt{15}$,
- $\boldsymbol{\beta}_4 = (0, 0, 0, 0, 0, 0, -1, -1, 1, 1)^T / 2$,

and $\mathbf{B}_0 = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4)$. The setting for $\boldsymbol{x}_{si}$ and ϵ_{si} are analogous to the ones in Case 1.

For each case, we first generated 1000 replicated samples, then randomly divided the samples into two nodes in each replication. We took 100 replications and plotted the boxplots for the space distance between the estimated feature space $\mathcal{S}(\widehat{\mathbf{B}})$ and $\mathcal{S}(\mathbf{B}_0)$ of each method. The results are shown in Figure 2.4, where the solid red line represents the result for pooled-data. We observe that, when the data heterogeneity does not exist, the performance of the proposed FSF method is similar to the performance of the pooled-data estimator and is significantly superior to all the local estimators.

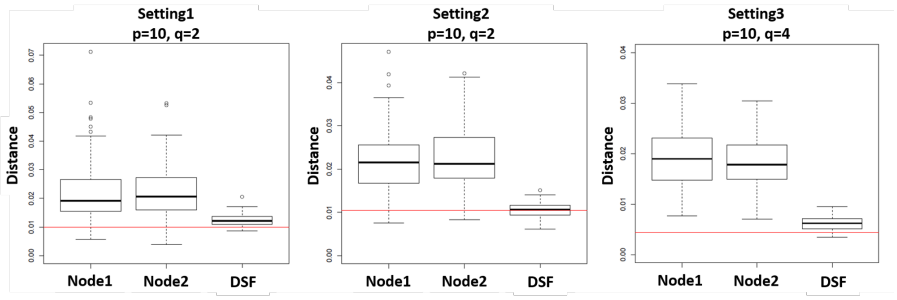

Figure 2.4: Boxplot of the distance between $\mathcal{S}(\widehat{\mathbf{B}})$ and $\mathcal{S}(\mathbf{B}_0)$ for each method under two settings respectively. The results obtained by MAVE applied to full data is set as a baseline (the red line).

Simulations with the first type of data heterogeneity. We then consider three cases where the first type of data heterogeneity exists, i.e., the predictor variables have different probability distribution across the nodes. For each case, two model settings are considered, which are analogous to the ones in Case 2 and Case 3. We set the sample size to be 300 in each node. More details are relegated to the supplementary material.

Case 4. For half of the nodes, the predictor variables are i.i.d. generated from the multivariate uniform distribution $[0, 1]^d$, and the other half are i.i.d, generated from the multivariate normal distribution $N(\mathbf{0}, \mathbf{I}_p)$.
Case 5. In this case, the independent predictor variables are normally distributed in each node with the same covariance matrix $\mathbf{I}_p$ and different means $\mu \cdot \mathbf{1}_p$.
Case 6. In this case, the independent predictor variables are normally distributed in each node with the same mean $\mathbf{0}$ and different covariance matrices $\sigma^2 \cdot \mathbf{I}_p$.

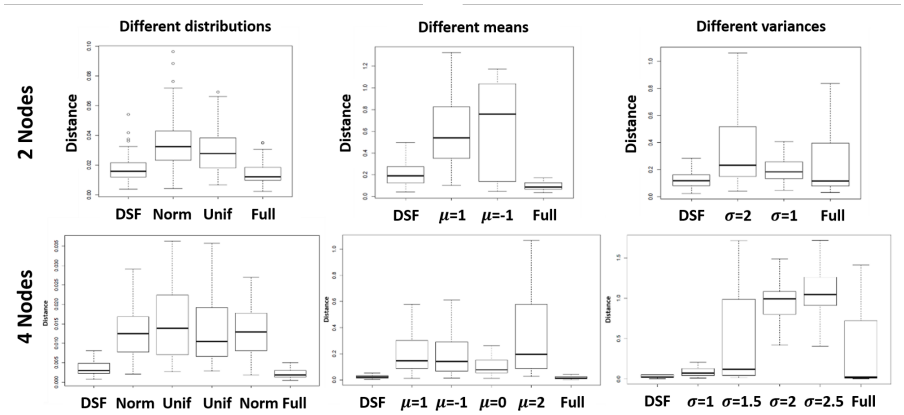

Figure 2.5: Boxplot of the distance between $\mathcal{S}(\widehat{\mathbf{B}})$ and $\mathcal{S}(\mathbf{B}_0)$ for each method. The results for Case 4, Case 5, and Case 6 are shown in the left column, middle column, and the right column, respectively. The above row and the lower row show the results for two-node cases and four-node cases, respectively.

For each case, we fitted the model with the proposed method and compared the results to MAVE within each individual node and MAVE applied to the pooled-data. The simulated results for 100 replicates are shown in Figure 2.5, where the upper row represents the two-node cases, and the lower row represents the four-node cases. We first observe that the estimation using pooled-data, at times, performs worse than local estimations. Such an observation can be attributed to the existence of the data heterogeneity, as stated in the Simpson's paradox. We then observe that the proposed FSF method performs significantly better than all the local estimators. We attribute such a success to the fact that the proposed method effectively integrates the information from local nodes, resulting in more accurate estimations.

Simulations with the second type of data heterogeneity. Finally, we consider a more complicated case that the link function $\eta_s(\cdot)$'s take different forms in different nodes. The data in node one and node two are generated with the settings that are analogous to the settings, as in Case 1 and Case 2, respectively. The results for 100 replications are showed in Figure 2.6. We observe

that the pooled-data estimation has the worst performance, due to the fact that the link functions are node-specific. We also observe that the proposed method outperforms the local estimations and the pooled-data estimation in both bias and variance.

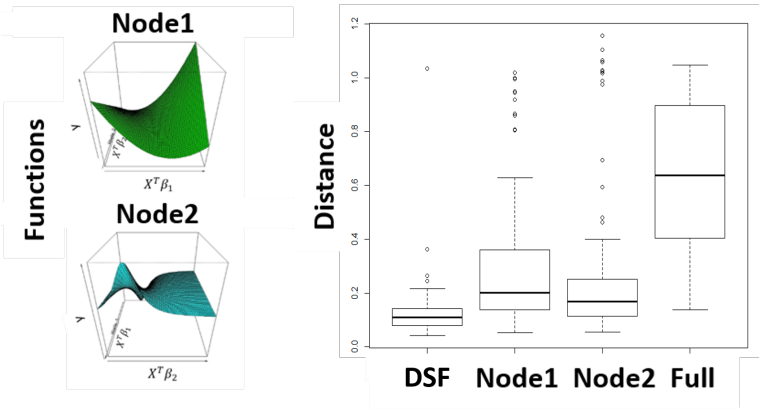

Figure 2.6: Left pannel: visualizations of different link functions. Right panel: the boxplot of the distance between $\mathcal{S}(\widehat{\mathbf{B}})$ and $\mathcal{S}(\mathbf{B}_0)$ for each method.

2.6 Real Data Analysis

Multi-department medical data. Due to some practical concerns, it is sensible to transfer patients' medical records between different departments. Most times, the records can only be shared under the agreement of patients. For such medical researches, data cannot be directly gathered or pooled. The clinical response of patient i in department s, i.e., y_{si}, can be considered as multivariate functions or mappings of q biomarkers that quantified by the projected data $\mathbf{B}^T\boldsymbol{x}_{si}$, where $\mathbf{B} \in \mathbb{R}^{q\times p}$ does not depends on either s or i. As the medical protocols are often different across different departments, such as the department of cardiology, the department of pediatrics, and the department of gynecology. To describe the data heterogeneity, we allow the unknown function $\eta_s(\cdot)$ to be different in each department with data $\boldsymbol{x}_{si}$ not identically distributed for each $s = 1, \ldots, S$. The data record 96 factors of medical examination information and the coagulation factor for 1,172 patients from seven different departments in hospitals in Shanghai. Because there are fewer than 30 records collected in the third, fifth, and seventh department, we did not take those part of data. By treating each department as one node, we applied the proposed FSF method to fit the model. Under each replication, we randomly chose a sample of size 30 from node 1 (the first department) as our testing set. The rest data were used for training.

Weather report data. Weather report data for 49 cities in Australia from Dec 12, 2008, to Jun 24, 2017, have been recorded. There are 142,000 records with 24 variables. Among the variables, there are categorical variables such as wind direction, level of cloudiness, whether rains today (y/n), etc. The quantitative variables are temperature, wind speed, humidity, pressure, etc. We randomly chose 14 cities to train the model by treating each city as a node in the scattered data system, i.e., $S = 14$. The region-specific geographical conditions result in the data heterogeneity. We chose six quantitative variables (wind speed, humidity, and pressure, each recorded daily at 9 AM and 3 PM) as the predictor $\boldsymbol{x}_{si}$, to predict if it will rain in the next day. The response y_{si} is the risk of raining the next day (quantitative). We randomly chose 100 observations from the 12th node as our testing set and used the rest data of the 14 nodes for training.

House price data. The data records the house prices in Melbourne from the year 2016 to 2017, including house type, distance from the central business district (CBD), land size, building size, longitude, latitude, etc.. We focus on one certain house type (marked as "h" in the dataset), and choose the distance from CBD, land size and building size as the predictors. Using the longitude and latitude information, we create a new predictor indicates the distance from the bay. The data is partitioned to 13 areas, as shown in Figure 2.7 by K-means clustering. Each area is treated as a node. We randomly picked 100 data points from the 10th node as our testing data and used the rest data for training.

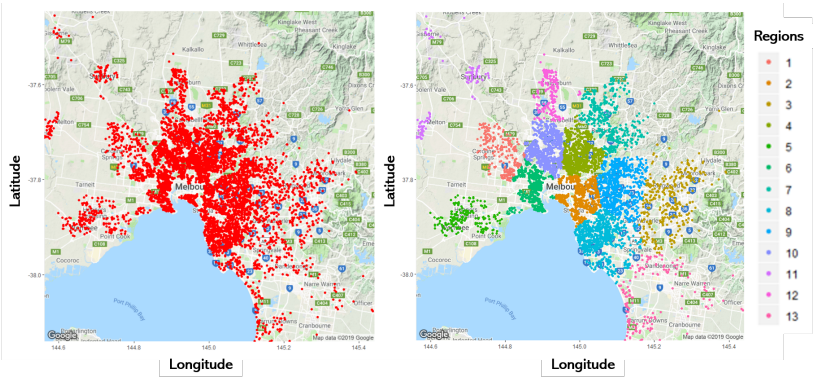

Figure 2.7: Visualization of the house price data. The right panel show the partition of the dataset using K-means clustering.

For all three real data analyses, we compared the testing MSE of our method with those of support vector regression (SVR), simple neural network (SNN), and minimum average variance estimation (MAVE). All three methods were applied to the pooled-data. We also list the testing MSEs for each individual node based on its own data. To evaluate our method, we reported the relative im-

provements of the decentralized method over other methods, which is defined as

$$IMP = (MSE_{others} - MSE_{FSF})/MSE_{others}.$$

An IMP greater than 0 implies that the proposed FSF method performs better. The results listed in Table 2.1 show the mean improvements after 20 replicates. We also listed the first decile (10-quantile) of the results. For the weather report data, it took too much time to run SVR since the pooled-data volume is too big. From the result, we see that most time, the decentralized method can significantly outperform other methods. For the house price data, the decentralized method performs similarly as the 10th and 11th nodes (marked as red) and significantly better than others. Those "non-significance" in 10th and 11th nodes may be caused by two reasons, 1) the testing data came from the 10th node, and 2) the 11th node is far away from others.

Table 2.1: The mean improvement (Mean IMP) of the proposed FSF method over other methods, with the first decile presented as well.

Medical Research							
	pooled-data			Single node			
Method	SVR	SNN	MAVE	#1	#2	#3	#4
Mean IMP	.57	.68	.42	.34	.41	.51	.39
1^{st} Decile	.45	.52	.09	.06	.20	.39	.18

Weather Report																
	pooled-data			Single node												
Method	SVR	SNN	MAVE	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14
Mean IMP	–	.51	.33	.30	.28	.24	.34	.30	.27	.29	.31	.32	.37	.28	.28	.30	.32
1^{st} Decile	–	.24	.05	.07	.01	.02	.08	.06	.03	.02	.02	.03	.08	.04	.00	.04	.06

Housing Price																
	pooled-data			Single node												
Method	SVR	SNN	MAVE	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13
Mean IMP	.49	.89	.36	.38	.41	.44	.60	.42	.44	.44	.45	.59	.39	.27	.38	.45
1^{st} Decile	.30	.83	.17	.08	.17	.17	.43	.21	.18	.12	.16	.40	-.11	-.07	.01	.20

2.7 Concluding Remarks

We address the data heterogeneity in data fusion and propose a feature space fusion method for scattered datasets. We confirm the asymptotic convergence of the algorithm, the efficiency of the estimation, and the consistency of the estimated dimension theoretically. We show that our method works well on heterogeneous scattered data. We would like to extend the algorithm to a more complicated network and consider the asynchronous algorithms to further reduce the computational cost in the near future.

Chapter 3

Image Fusion through Optimal Transport Map and Its Application in Echocardiogram

There is a long and rich history of Optimal Transport Map (OTM): initiated by Monge in the 18th century and reinvigorated by modern machine learning applications like generative networks and transfer learning. Though the mathematical properties of OT have been extensively studied, the Wasserstein distance induced by an empirical OTM suffers from a slow convergence rate when the dimensionality is large. In high dimensional regime, the empirical distribution summarized from a random sample with a fixed sample size is usually atypical of the population due to the "curse of dimensionality". Without any smoothness constraint, the empirical OTM that pursues a one-to-one map to minimize the Wasserstein distance between two high-dimensional random samples will inevitably lead to severe over-fitting. To address this issue, we propose a novel estimator of OTM named Smoothed Monge Map (SMM). SMM tackles the "curse of dimensionality" problem by applying coordinate-wise smoothing spline to the empirical OTM. By imposing smoothness penalties, SMM balances the in-sample goodness of fit and the roughness of the transport map. When the dimensionality $d > 4$, under mild conditions, SMM can effectively alleviate the over-fitting issue and improve the convergence rate from $O(n^{-1/d})$ to $O(n^{-1/2})$. Numerical studies on synthetic datasets justified the superior performance of SMM in comparison with mainstream competitors. Further, we apply SMM to a challenging echocardiogram analysis testing problem.

3.1 Introduction

These years, the rapid development of computer vision and machine learning techniques has triggered a medical technology revolution. The deep learning algorithms have been developed for medical image processing, especially for echocardiograms S. Chen et al., 2019; Ghorbani et al., 2020; Madani et al., 2018; Ouyang et al., 2019; J. Zhang et al., 2018. In echocardiogram analysis, one major task is to tracing the myocardial movement of the endocardium of the left ventricle (LV-Endo). Deep learning methods, such as CNN, have been applied in recent years. However, such methods require large quantity of labeled image. With huge image diversity, labeling LV-Endo could be super labor-intensive. To address the image diversity problem, we propose a reference-based tracing method. Instead of labeling on individual echocardiogram, we label on a reference image which can well capture the myocardial structure. The reference image is generated through image fusion. One possible choice is the "Wasserstein barycenter" (Cuturi & Doucet, 2014). To obtain the Wasserstein barycenter as the reference image of echocardiogram, the key is to well approximate the Optimal Transport Map (OTM).

Nowadays, as a powerful tool to quantify the minimum "distance" between two metric spaces, OTM has been reinvigorated in a remarkable proliferation of modern data science applications, including machine learning (Alvarez-Melis et al., 2018; Arjovsky et al., 2017; Canas & Rosasco, 2012; Courty et al., 2016; Flamary et al., 2018; Meng et al., 2019; Peyré, Cuturi, et al., 2019), statistics (Cazelles et al., 2018; Del Barrio et al., 2019; Panaretos & Zemel, 2019), computer vision (Ferradans et al., 2014; Peyré, Cuturi, et al., 2019; Rabin et al., 2014; Su et al., 2015), and so on.

Despite its popularity, the Wasserstein distance induced by an empirical OTM suffers from a slow convergence rate when the dimensionality is large. A stylized feature of high-dimensional data is that the empirical distribution summarized from a random sample with a fixed sample size is usually atypical of the population due to the "curse of dimensionality". Suppose that we observe two i.i.d. samples $\{\boldsymbol{a}_i\}_{i=1}^n \in \mathbb{R}^d$ and $\{\boldsymbol{b}_i\}_{i=1}^n \in \mathbb{R}^d$ from two continuous probability distributions α and β, respectively. When $d > 4$, the 2-Wasserstein distance induced by the empirical OTM between $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_i\}_{i=1}^n$ converges to its population counterpart at a slow rate of order $O(n^{-1/d})$ (Dudley, 1969; Fournier & Guillin, 2015). Unfortunately, this rate is tight in the sense that the upper and lower bounds meet except for a positive constant (Weed & Bach, 2019). Like many well-studied nonparametric methods, the empirical OTM that minimizes the in-sample goodness of fit without any smoothness

constraints will inevitably run into severe over-fitting issues. The slow convergence issue has already hindered the broad applications of OTM with high-dimensional datasets.

A good amount of existing literature has been devoted to addressing the "curse of dimensionality" issue of empirical OTM. The convergence rate can be refined when the measures are supported on low-dimensional sub-domains: Weed and Bach, 2019 proposed to find an implicit dimension d_0 of the data and proved that the convergence rate could be refined when $d_0 \ll d$. However, the calculation of d_0 is not straight forward and d_0 may not necessarily be much smaller than d. Genevay et al., 2019 studied the sinkhorn divergence, a regularized variant of the Wasserstein distance. The authors showed that the converges rate of sinkhorn divergences is getting closer to the order $O(n^{-1/2})$, as the regularization parameter diverges. Nevertheless, the sinkhorn divergence pays the price on creating a bias term that can dominate the variance part in some real applications (Genevay et al., 2019). Forrow et al., 2019 proposed an estimator for Wasserstein distances using factored couplings as regularization. Though the proposed estimator converges to its expectation at the rate of $O(n^{-1/2})$ for fixed d, there is no theoretical guarantee that this estimator can overcome the "curse of dimensionality" issue in general high-dimensional setups.

To address the aforementioned problems, we propose a novel estimator of OTM named Smoothed Monge Map (SMM). SMM tackles the "curse of dimensionality" issue by applying coordinate-wise smoothing spline to the empirical OTM. By imposing smoothness penalties, SMM balances the in-sample goodness of fit and the roughness of the transport map. When the dimensionality $d > 4$, under mild conditions, SMM can effectively alleviate the over-fitting issue and improve the convergence rate of the p-Wasserstein distance from $O(n^{-1/d})$ to $O(n^{-1/2})$. Thus, SMM can induce a consistent empirical Wasserstein distance in the high-dimensional regime. Numerically, we show that SMM outperforms several state-of-the-art OTM estimators through extensive synthetic experiments. With such empirical Wasserstein distance , we can approximate the Wasserstein barycenter more precisely. A more precise Wasserstein barycenter of the echocardiogram provides a better reference image which can represent the "standard heart". We also prove the asymptotic normality of the empirical Wasserstein distance which paves the way for applying SMM to test the distributional equivalence of two samples. In echocardiogram analysis, researchers study echocardiogram from different directions and fuse them together to recover the 3-D myocardial movement features of the heart. The current echocardiogram classification heavily relies on clinicians' experiences

and domain-specific knowledge, and thus are highly subjective (Michel et al., 2017). The testing enables us a quantitative-based classification method.

3.2 Problem setup

3.2.1 Monge map and Wasserstein distance

The optimal transportation theory has been widely studied in mathematics, probability, and economics, see (Ferradans et al., 2014; Reich, 2013; Su et al., 2015) and references therein. Let $\alpha \in \mathbb{R}^d$ and $\beta \in \mathbb{R}^d$ be two continuous probability measures defined on a proper probability space. To focus on the contexts that are most relevant to machine learning applications, we assume that at least one of the two measures α and β has a continuous density with respect to the Lebesgue measure. The Monge problem with an L_2 transport cost aims to find the an optimal transport map (OTM)[1] $\phi^* : \mathbb{R}^d \to \mathbb{R}^d$ that solves the following minimization problem

[1] Also called the Monge map.

$$\phi^* = \inf_{\phi \in \Phi} \int_{\mathbb{R}^d} \|a - \phi(a)\|^2 \mathrm{d}\alpha(a), \tag{3.1}$$

where $\|\cdot\|$ is the vector norm, and the set Φ includes all push forward maps[2] $\phi_{\#}(\cdot)$ that satisfy $\phi_{\#}(\alpha) = \beta$ and $\phi_{\#}^{-1}(\beta) = \alpha$. Throughout this chapter, we work on the moderate or high dimensional case such that $d > 4$. Under the above setups, the Brenier's theorem (Brenier, 1991) guarantees the existence of the Monge map and proves that the Monge problem is equivalent to the Kantorovich formulation of the optimal transport problem (Kantorovich, 2006; Kantorovitch, 1958).

[2] For all $\Omega \subset \mathbb{R}^d$, $\phi_{\#}(\alpha)(\Omega) = \alpha(\phi^{-1}(\Omega))$

The Monge map naturally induces a meaningful distance, named Wasserstein distance (Peyré, Cuturi, et al., 2019; Villani, 2008), between the two measures. With the Monge map in (3.1), the 2-Wasserstein distance between α and β is defined as

$$W_2(\alpha, \beta) = \left(\int_{\mathbb{R}^d} \|a - \phi^*(a)\|^2 \mathrm{d}\alpha(a) \right)^{1/2}.$$

In practice, the underlying measures α and β are usually not observable. Instead, solving the optimal transport problem between two empirical measures has been considered as the essential component in many machine learning applications, such as computer vision and domain adaptation (Bhushan Damodaran et al., 2018; Courty et al., 2017; Courty et al., 2014; Courty et al., 2016; Perrot

et al., 2016). Suppose that we observe two i.i.d. samples $\{\boldsymbol{a}_i\}_{i=1}^{n_a} \in \mathbb{R}^d$ and $\{\boldsymbol{b}_i\}_{i=1}^{n_b} \in \mathbb{R}^d$ from the measures α and β, respectively. For ease of presentation, we assume $n_a = n_b = n$ and all the observations are equally weighted. The methodology and theory developed in this chapter can be easily extended to the case when $n_a \neq n_b$.

The empirical Monge map $\widehat{\phi} : \mathbb{R}^d \to \mathbb{R}^d$ with an L_2 transport cost is defined as

$$\widehat{\phi} = \underset{\phi \in \Phi_n}{\operatorname{argmin}} \sum_{i=1}^{n} \|\boldsymbol{a}_i - \phi(\boldsymbol{a}_i)\|^2, \tag{3.2}$$

where Φ_n is the set that contains all one-to-one maps from $\{\boldsymbol{a}_i\}_{i=1}^n$ to $\{\boldsymbol{b}_i\}_{i=1}^n$. The combinatorial optimization of the matching problem in (3.2) has been long studied. Some successful algorithms include the Hungarian algorithm (Bertsimas & Tsitsiklis, 1997; Kuhn, 1955) and auction algorithm (Bertsekas, 1981, 1992), among others. Then, the empirical 2-Wasserstein distance can be calculated as

$$W_2(\boldsymbol{a}_n, \boldsymbol{b}_n) = \left(\frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{a}_i - \widehat{\phi}(\boldsymbol{a}_i)\|^2 \right)^{1/2}. \tag{3.3}$$

3.2.2 Curse of dimensionality

Like many machine learning, statistics, and optimization problems that involve probability measures in $\mathbb{R}^d$, the convergence of empirical 2-Wasserstein distance suffers from the so-called "curse of dimensionality" issue Bellman, 2015. When the dimension d increases, the empirical measures summarized from the samples $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_i\}_{i=1}^n$ become decreasingly representative to their population counterparts. Hence, the convergence of $W_2(\boldsymbol{a}_n, \boldsymbol{b}_n)$ to $W_2(\alpha, \beta)$ is slow. It has been shown that the convergence rate of the empirical 1-Wasserstein distance is lower bounded by the order $O(n^{-1/d})$, and this order is asymptotically tight Dudley, 1969. Some recent studies (Fournier & Guillin, 2015; Weed & Bach, 2019) have justified that such a lower bound is inevitable in general $d > 4$ settings. In this subsection, we elaborate on this "curse of dimensionality" phenomenon by unveiling the connection between the empirical Monge map and the nearest neighbor distance.

Let $\boldsymbol{a}_0$ be a fixed point in $[0, 1]^d$ and $\{\boldsymbol{b}_i\}_{i=1}^n$ be a sample drawn from the uniform density with domain on $[0, 1]^d$. The nearest neighbor of $\boldsymbol{a}_0$ in sample $\{\boldsymbol{b}_i\}_{i=1}^n$ is the point that minimizes a pre-specified distance metric. Suppose the distance is measured by the Euclidean norm, the nearest neighbor $\boldsymbol{b}^*$ of $\boldsymbol{a}_0$

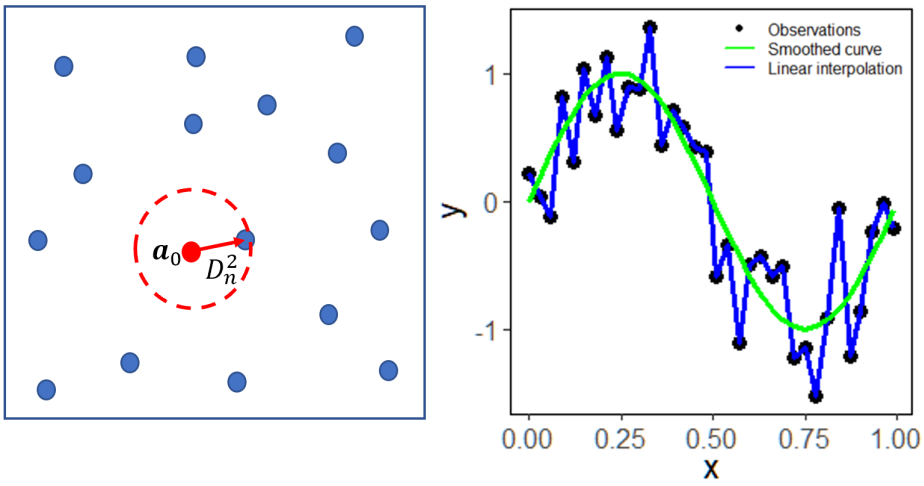

Figure 3.1: Left panel: the square of the nearest neighbor distance D_n^2 converges to 0 at the rate of $O(n^{-2/d})$. Right panel: rather than linear interpolation, non-parametric regression method provides a smoothed curve that avoids over-fitting.

is defined as

$$\boldsymbol{b}^* = \underset{\boldsymbol{b}\in\{\boldsymbol{b}_i\}_{i=1}^n}{\operatorname{argmin}} \|\boldsymbol{a}_0 - \boldsymbol{b}\|^2.$$

Moreover, the random variable $D_n^2 := ||\boldsymbol{a}_0 - \boldsymbol{b}^*||^2$ is called the nearest distance of $\boldsymbol{a}_0$ to the sample $\{\boldsymbol{b}_i\}_{i=1}^n$. The definition of the nearest neighbor and the nearest neighbor distance in a $d = 2$ case is illustrated in the left panel of Figure 3.1. The red point is $\boldsymbol{a}_0$, the blue points are $\{\boldsymbol{b}_i\}_{i=1}^n$, the red solid line connects $\boldsymbol{a}_0$ to its nearest neighbor $\boldsymbol{b}^*$, and D_n^2 is indicated as the length of the red solid line.

Intuitively, one would expect the nearest neighbor distance D_n^2 converges to zero as $n \to \infty$(Evans et al., 2002; Percus & Martin, 1998). However, the convergence rate will suffer from the increase of dimension d dramatically due to the natural of space filling. It is shown in Evans et al., 2002 that D_n^2 uniformly converges to 0 at the rate of $O(n^{-2/d})$, as $n \to \infty$. Indeed, the "curse of dimensionality" is a long-standing issue in high-dimensional nearest neighbor algorithms that cause various difficulties in data mining, machine learning, and database technologies (Andoni & Indyk, 2006; Hinneburg et al., 2000; Muja & Lowe, 2014; Salton, 1989).

Similarly, the empirical Monge map suffers from the "curse of dimensionality" as it requires a one-to-one map between two observed samples. Suppose that

we draw two i.i.d. samples $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_i\}_{i=1}^n$ from the same uniform density function μ with domain on $[0,1]^d$. For each point in $\{\boldsymbol{a}_i\}_{i=1}^n$, we can find its Monge mapped point $\widehat{\phi}(\boldsymbol{a}_i)$ and its nearest neighbor $\boldsymbol{b}^*(\boldsymbol{a}_i)$ both in $\{\boldsymbol{b}_i\}_{i=1}^n$, respectively. Notice that, $\widehat{\phi}(\boldsymbol{a}_i)$ and $\boldsymbol{b}^*(\boldsymbol{a}_i)$ are not necessarily coincide as the nearest map is not a one-to-one map. Besides, one can build the folloiwg connection between the empirical 2-Wasserstein distance and the nearest neighbor distance as follows

$$\begin{aligned}\text{var}\left(W_2(\boldsymbol{a}_n,\boldsymbol{b}_n)\right) &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{a}_i - \widehat{\phi}(\boldsymbol{a}_i)\|^2\right] \\ &\geq \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{a}_i - \boldsymbol{b}^*(\boldsymbol{a}_i)\|^2\right] \\ &= \mathbb{E}\left[D_n^2(\boldsymbol{a}_i)\right] = O(n^{-2/d}).\end{aligned}$$

The above results show that the variance of the empirical 2-Wasserstein distance is lower bounded by the convergence rate of the nearest neighbor distance, which is of order $O(n^{-2/d})$. The cause of the slow convergent of variance is the nature of the one-to-one map, which perfectly minimizes the in-sample distance but fails to accurately infer the underlying distance between two probability measures when the samples are less representative in high-dimensional space. An analogy is to consider fitting a non-parametric regression without any smoothness constraints. We will end up with an over-fitting model that minimizes the in-sample fitting error with a zigzag curve that produces a huge variance.

To overcome the aforementioned "curse of dimensionality" issue, we propose to add smoothness constraints to the Monge problem. The new method, named Smoothed Monge Map (SMM), provides a non-parametric smoothed variant of the empirical Monge map. SMM is less prone to over-fitting when d is large. We introduce the details, implementations, and advantages of SMM in the next section.

3.3 Methodology

3.3.1 Nonparametric regression and smoothing splines

Suppose that we fit a random sample $\{y_i, \boldsymbol{x}_i\}_{i=1}^n \in \mathbb{R}^{1+d}$ with a nonparametirc regression model

$$y_i = \eta(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1,\ldots,n,$$

where η is an unknown function to be estimated and $\{\epsilon_i\}_{i=1}^n$ are independent and zero mean random errors. A large family of nonparametric regression methods (e.g. Fan, 2018; Györfi et al., 2006; Wasserman, 2013) estimate the function η by minimizing a penalized least squares loss function

$$\frac{1}{n}\sum_{i=1}^{n}\{y_i - \eta(\boldsymbol{x}_i)\}^2 + \lambda J(\eta), \tag{3.4}$$

where $J(\eta)$ is a roughness penalty (Gu, 2013; Wahba, 1990; X. Wang et al., 2011). The smoothing parameter λ, which can be selected based on the generalized cross-validation criterion (Wahba & Craven, 1978), controls the trade-off between the goodness-of-fit and the "roughness" of η. The minimization of (3.4) is over the functional forms of η in a reproducing kernel Hilbert space $\mathcal{H}$, which can yield a smoothing spline estimate for η[3].

[3] Similar estimates can be obtained by other nonparametric regression methods, like kernel smoothing and wavelet methods.

The right panel of Figure 3.1 gives a toy example of the smoothing spline method. Compared with the linear interpolation (blue line), the smoothing spline (green curve) can "smooth-out" the variance contributed by the errors $\{\epsilon_i\}_{i=1}^n$ and is less prone to over-fit the responses (black dots)

3.3.2 Estimation of smoothing splines

The standard formulation of smoothing splines minimizes (3.4) in a reproducing kernel Hilbert space $\mathcal{H} = \{\eta : J(\eta) < \infty\}$, where $J(\cdot)$ is a squared semi-norm. Let $\mathcal{N}_J = \{\eta : J(\eta) = 0\}$ be the null space of $J(\eta)$. Further, we assume that $\mathcal{N}_J$ is a finite-dimensional linear subspace of $\mathcal{H}$ with basis $\{\xi_i\}_{i=1}^m$, where m is the dimension of $\mathcal{N}_J$. Let $\mathcal{H}_J$ denote the orthogonal complement of $\mathcal{N}_J$ in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{N}_J \oplus \mathcal{H}_J$. Then, $\mathcal{H}_J$ is also a reproducing kernel Hilbert space with $J(\cdot)$ being the squared norm. The reproducing kernel of $\mathcal{H}_J$ is denoted by $R_J(\cdot,\cdot)$.

The well-known representer theorem (Wahba, 1990) states: although the original penalized least squares problem for smoothing splines is formulated in the infinite-dimensional space $\mathcal{H}$, the solution of it lies in a finite-dimensional space. Specifically, there exist $\boldsymbol{d} = (d_1,\ldots,d_m)^T$ and $\boldsymbol{c} = (c_1,\ldots,c_n)^T$, such that the minimizer of (3.4) in $\mathcal{H}$ is given by

$$\eta(x) = \sum_{k=1}^{m} d_k\xi_k(x) + \sum_{i=1}^{n} c_i R_J(x_i, x).$$

Let $\mathbf{Y} = (y_1,\ldots,y_n)^T$ be the response vector, $\mathbf{S}$ be an $n \times m$ matrix whose (i,j)th entry is denoted as $\xi_j(x_i)$, and $\mathbf{R}$ be an $n \times n$ matrix whose (i,j)th

entry is denoted as $R_J(x_i, x_j)$. According to the representer theorem, solving (3.4) is equivalent to minimize

$$\underset{\boldsymbol{d}\in\mathbb{R}^m, \boldsymbol{c}\in\mathbb{R}^n}{\operatorname{argmin}} \frac{1}{n}(\mathbf{Y} - \mathbf{S}\boldsymbol{d} - \mathbf{R}\boldsymbol{c})^T(\mathbf{Y} - \mathbf{S}\boldsymbol{d} - \mathbf{R}\boldsymbol{c}) + \lambda \boldsymbol{c}^T\mathbf{R}\boldsymbol{c}.$$

The solution of the above minimization problem admits a closed form as the minimizer $(\hat{\boldsymbol{d}}, \hat{\boldsymbol{c}})$ is the solution of the following linear system of equations

$$\begin{pmatrix} \mathbf{S}^T\mathbf{S} & \mathbf{S}^T\mathbf{R} \\ \mathbf{R}^T\mathbf{S} & \mathbf{R}^T\mathbf{R} + n\lambda\mathbf{R} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{d}} \\ \hat{\boldsymbol{c}} \end{pmatrix} = \begin{pmatrix} \mathbf{S}^T\mathbf{Y} \\ \mathbf{R}^T\mathbf{Y} \end{pmatrix}.$$

3.3.3 Smoothed Monge map

The success of the smoothing spline inspires us to improve the empirical Monge map $\widehat{\phi}$ by considering a "smoothed" variant $\widetilde{\phi}$. The proposed method, named Smoothed Monge Map (SMM), aims to fit univariate smoothing splines for each dimension of the empirical Monge map.

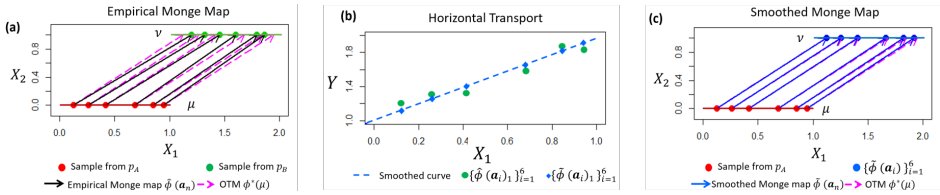

Figure 3.2: Illustration of the smoothed Monge map.

We illustrate the idea of Smoothed Monge Map through a toy example illustrated in Figure 3.2. Let μ and ν be two uniform densities on $[(0,0)^T, (1,0)^T]$ and $[(1,1)^T, (2,1)^T]$, respectively. Note that the supports of μ and ν are essentially one-dimensional and the Monge map from μ to ν, i.e. $\phi^*(\mu)$, can be explicitly written as $\phi^*((x_1, x_2)^T) = (x_1 + 1, x_2 + 1)^T$. We generate $\{\boldsymbol{a}_i\}_{i=1}^6$ and $\{\boldsymbol{b}_i\}_{i=1}^6$ as two random samples from μ and ν, and calculate their empirical Monge map $\widehat{\phi}(\boldsymbol{a}_n)$. In Figure 3.2(a), $\{\boldsymbol{a}_i\}_{i=1}^6$ and $\{\boldsymbol{b}_i\}_{i=1}^6$ are plotted as the red dots and green dots. Besides, $\phi^*(\mu)$ and $\widehat{\phi}(\boldsymbol{a}_n)$ are plotted as the pink dashed arrows and solid black arrows, respectively. The deviations between the pink and black arrows indicate that the empirical Monge map over-fits the random sample. Then, we marginally fit nonparametric regressions between $\{\widehat{\phi}(\boldsymbol{a}_i)\}_{i=1}^6$ and $\{\boldsymbol{a}_i\}_{i=1}^6$. The fitted curve (actually a straight line in this case) of the first dimension is plotted in Figure 3.2(b). The collection of the marginal

Algorithm 3 Smoothed Monge Map

Input: data points $\{\boldsymbol{a}_i\}_{i=1}^n \in \mathbb{R}^d$, $\{\boldsymbol{b}_i\}_{i=1}^n \in \mathbb{R}^d$
Step 1: Calculate the empirical Monge map between $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_i\}_{i=1}^n$, denote the map as $\widehat{\phi}$;
Step 2:
for j in $1:d$ **do**

Calculate the smoothing spline estimator $\widetilde{\eta}_{(j)} : \mathbb{R}^d \to \mathbb{R}$ of the regression problem

$$\widehat{\phi}(\boldsymbol{a}_i)_j = \eta_{(j)}(\boldsymbol{a}_i) + \epsilon_{ij}, i = 1, \ldots, n.$$

end for
Step 3: For any $\boldsymbol{a}_0 \in \mathbb{R}^d$ on the support of μ, the density of $\{\boldsymbol{a}_i\}_{i=1}^n$; Define the smoothed Monge map $\widetilde{\phi}$ as

$$\widetilde{\phi}(\boldsymbol{a}_0) = (\widetilde{\eta}_{(1)}(\boldsymbol{a}_0), \ldots, \widetilde{\eta}_{(j)}(\boldsymbol{a}_0))^T.$$

Step 4: The smoothed 2-Wasserstein distance is calculated as

$$\widetilde{W}_2 = \left(\frac{1}{n} \sum_{i=1}^n \|\boldsymbol{a}_i - \widetilde{\phi}(\boldsymbol{a}_i)\|^2 \right)^{1/2}.$$

estimators forms a smoothed Monge map $\widetilde{\phi}(\boldsymbol{a}_n)$. In Figure 3.2(c), we replace the empirical Monge map $\widehat{\phi}(\boldsymbol{a}_n)$ by the Smoothed Monge Map $\widetilde{\phi}(\boldsymbol{a}_n)$ (blue solid arrows). As we can see, the Smoothed Monge Map avoids the over-fitting issue and well match the population ones. The details of the Smoothed Monge map is summarized in Algorithm 3 below.

3.3.4 Implementation details and computational cost

In Algorithm 3, the computational cost mainly resides in Step 1 and Step 2. In Step 1, the computational cost for calculating the empirical Monge map with the auction algorithm is of order $O(n^2 \log(n))$ (Schwartz, 1994). In each iteration of Step 2, the computational cost of solving the smoothing spline estimation is of order $O(n^3)$ when $d \geq 4$.

The computation of Algorithm 3 can be accelerated by a basis selection method. The basis selection method uses $q < n$ sampled basis functions instead of n basis functions to approximate the full-sample estimator, resulting in a

computational cost of order $O(nq^2)$. Then, the overall computational cost of Algorithm 1 is reduced to the order $O(n^2 \log(n) + dnq^2)$. As suggested in (Ma et al., 2015), we can choose q to be $Cn^{2/9}$ for some positive constant C. When the dimensionality d does not diverge too fast with the sample size n (i.e. $d = O(n^{1/2})$), the computational cost of Algorithm 1 is of order $O(n^2 \log(n))$.

3.4 Theoretical Results

In this section, we present the major theoretical results of this chapter. Due to the space limitation, more technical lemmas and detailed proofs are deferred to a supplemental file. To begin with, we list the technical assumptions required for the delivery of the theoretical results.

Assumption 3.4.1. *(a) α and β are Borel probabilities on $\mathbb{R}^d$ with positive densities in the interior of their convex support.*

(b) Let $\mathcal{H}^d$ be the d-dimensional Hausdorff measure on a closed set $\mathcal{S}$. We assume $\mathrm{supp}(u) \subseteq \mathcal{S}$.

(c) α and β have finite $(4+\delta)$th moments for some $\delta > 0$.

The assumption (a) is a standard condition in optimal transportation theory, which allows us to avoid many difficult discussions for irrelevant measure-theoretical scenarios. For most continuous probability distributions that are well defined on $\mathbb{R}^d$, the assumption (a) is naturally satisfied. The assumption (b) is required to prove the lower bound for the empirical Wasserstein distance . The assumption (b) can be satisfied in our problem setups as $\mathbb{R}^d$ with its usual topology is a locally compact d-dimensional Hausdorff space. The assumption (c) imposes mild finite moment conditions on α and β. We may relax this assumption to require only the finite 4th moments. We do not pursue that approach, as it is not the focus of this chapter.

Next, we present the lower bound for the 2-Wasserstein distance induced by an empirical OTM. This theorem clearly unveils the slow convergence issue caused by the "curse of dimensionality".

Theorem 3.4.1 (Lower bound for empirical Wasserstein distance)**.** *Under Assumption 3.4.1 (a) and (b), we have*

$$W_2(\boldsymbol{a}_n, \boldsymbol{b}_n) - W_2(\alpha, \beta) \gtrsim n^{-1/d}.$$

The following theorem provides not only $\sqrt{n}$-consistency but also asymptotic normality for the 2-Wasserstein distance induced by SMM. In practice, the asymptotic variances can be estimated by the random samples.

Theorem 3.4.2 (Asymptotic Normality for SMM)**.** *Under Assumption 3.4.1 (a) and (c), we have*

$$\begin{aligned}&\sqrt{n}\left(\widetilde{W_2}^2(\boldsymbol{a}_n,\boldsymbol{b}_n)-E\widetilde{W_2}^2(\boldsymbol{a}_n,\boldsymbol{b}_n)\right)\\&\to N(0,\frac{\sigma^2(\alpha,\beta)+\sigma^2(\beta,\alpha)}{2}),\end{aligned}$$

as $n\to\infty$. Denote ψ^* is the OTM from β to α, the asymptotic variance is the average of the following two terms

$$\begin{aligned}\sigma^2(\alpha,\beta)&=\int_{\mathbb{R}^d}\left(\|a\|^2-2\phi^*(a)\right)^2d\alpha(a)\\&\quad-\left(\int_{\mathbb{R}^d}\left(\|a\|^2-2\phi^*(a)\right)d\alpha(a)\right)^2,\\\textit{and}\quad\sigma^2(\beta,\alpha)&=\int_{\mathbb{R}^d}\left(\|b\|^2-2\psi^*(b)\right)^2d\beta(b)\\&\quad-\left(\int_{\mathbb{R}^d}\left(\|b\|^2-2\psi^*(b)\right)d\beta(b)\right)^2.\end{aligned}$$

Remark 3.4.1. *The results in Theorem 3.4.2 can be naturally extended to two random samples with unequal sizes. Suppose that we have two empirical measures $\boldsymbol{a}_n$ and $\boldsymbol{b}_m$ with $n\neq m$. The asymptotic normality for SMM follows*

$$\begin{aligned}&\sqrt{\frac{nm}{n+m}}\left(\widetilde{W_2}^2(\boldsymbol{a}_n,\boldsymbol{b}_m)-E\widetilde{W_2}^2(\boldsymbol{a}_n,\boldsymbol{b}_m)\right)\\&\to N(0,\frac{m}{n+m}\sigma^2(\alpha,\beta)+\frac{n}{n+m}\sigma^2(\beta,\alpha)).\end{aligned}$$

3.5 Experimental Studies

3.5.1 Simulation studies

In this subsection, we assess the performance of the proposed Smoothed Monge Map (SMM) with simulated examples. Let p_A and p_B be two continuous probability distributions that satisfy the following two probability models.

Uniform Model: p_A and p_B are two independent uniform distributions supported on $[2,3]\times[0,1]^{d-1}$;

Normal Model I: p_A and p_B are two independent d-dimensional multivariate normal distributions. Specifically, $p_A=\mathcal{N}_d(\boldsymbol{\mu}_A,\Sigma_A)$ and

$p_B = \mathcal{N}_d(\boldsymbol{\mu}_B, \Sigma_B)$, where we set $\boldsymbol{\mu}_A = \boldsymbol{\mu}_B = \mathbf{0}$, $\Sigma_A = I_d$, $\Sigma_B = 0.8^{|i-j|}$ $(i, j = 1, \ldots, d)$.

NORMAL MODEL II: $p_A = \mathcal{N}_d(\boldsymbol{\mu}_A, \Sigma_A)$ and $p_B = \mathcal{N}_d(\boldsymbol{\mu}_B, \Sigma_B)$, where we set $\boldsymbol{\mu}_A = \mathbf{0}$, $\boldsymbol{\mu}_B = \mathbf{1}$, $\Sigma_A = I_d$, $\Sigma_B = 0.8^{|i-j|}$ $(i, j = 1, \ldots, d)$.

In the UNIFORM MODEL, it is easy to check that $W_2^2(p_A, p_B) = 4$. In the NORMAL MODEL, $W_2^2(p_A, p_B)$ admits a closed form:

$$||\boldsymbol{\mu}_A - \boldsymbol{\mu}_B||_2^2 + tr\left(\Sigma_A + \Sigma_B - 2(\Sigma_A^{1/2}\Sigma_B\Sigma_A^{1/2})^{1/2}\right).$$

For each probability model, we draw two i.i.d. samples $\mathcal{A}_n = \{\boldsymbol{a}_i\}_{i=1}^n$ and $\mathbf{B}_n = \{\boldsymbol{b}_i\}_{i=1}^n$ from distributions p_A and p_B, respectively. Then, we estimate $W_2(p_A, p_B)$ by SMM and the other 5 mainstream competitors:

1. The empirical Monge estimator (MONGE);
2. The estimator yielded by network simplex (NS);
3. Sinkhorn divergence with $\epsilon = 0.1$ (SD(0.1));
4. Sinkhorn divergence with $\epsilon = 1$ (SD(1));
5. Sinkhorn divergence with $\epsilon = 10$ (SD(10)) [4];

[4] Monge and NS are implemented by the R package "transport" (Schuhmacher et al., 2019). Sinkhorn divergence methods are implemented by the Python package "POT" (Flamary & Courty, 2017).

The empirical Monge estimator is calculated as equation (3.3). The network simplex algorithm (Courty et al., 2014; Cuturi, 2013; Peyré, Cuturi, et al., 2019) is a widely used algorithm that solves the optimal transport problem with the Kantorovich formulation. The Sinkhorn divergence (Genevay et al., 2018) is calculated as

$$W_2(p_A, p_B; \epsilon) - \left(W_2(p_A, p_A; \epsilon) + W_2(p_B, p_B; \epsilon)\right)/2,$$

where $W_2(\cdot, \cdot; \epsilon)$ is the regularized 2-Wasserstein distance and ϵ is a regularization parameter that trades bias for "smoothness". The Sinkhorn divergence improves the computational cost of the empirical Monge map from $O(n^3 \log(n))$ to $O(n^2)$ (Altschuler et al., 2017; Cuturi, 2013; Peyré, Cuturi, et al., 2019). However, it is an inconsistent estimator of Wasserstein distance and suffers from the convergence issue in practice when d is large.

In this simulation , we set $n \in \{10^{1.5}, 10^2, 10^{2.5}, 10^3\}$, $d \in \{6, 9, 12\}$. For each scenario, we simulate 100 replications. The estimation performance is measured by the absolute deviation (AD):

$$AD = |\widehat{W}_2^2(\mathcal{A}_n, \mathbf{B}_n) - W_2^2(p_A, p_B)|,$$

where $\widehat{W}_2(\mathcal{A}_n, \mathbf{B}_n)$ is the 2-Wasserstein distance estimated by SMM or one of the other 5 competitors above. In Figure 3.3, Figure 3.4, and Figure 3.5, we plot the mean (solid lines) and standard deviation (vertical bars) of AD over 100 replications with respect to the sample size n in a in *log-log* scale.

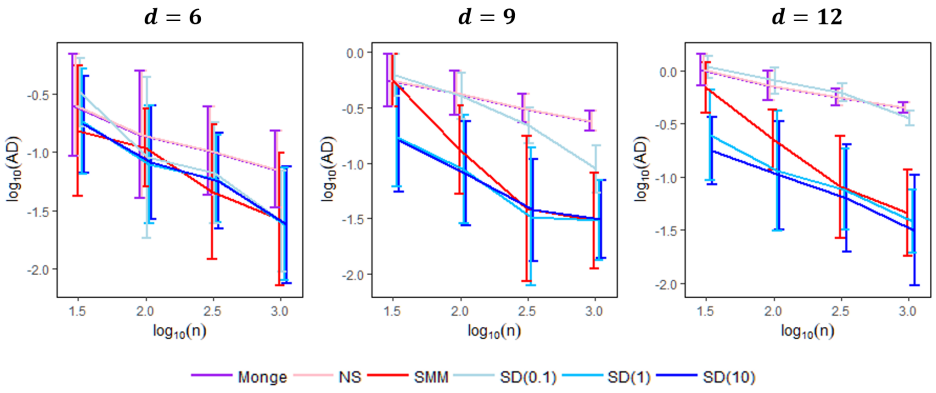

Figure 3.3: UNIFORM MODEL: the mean and standard deviation of the absolute deviation of 2-Wasserstein distance estimates with respect to n (in *log-log* scale).

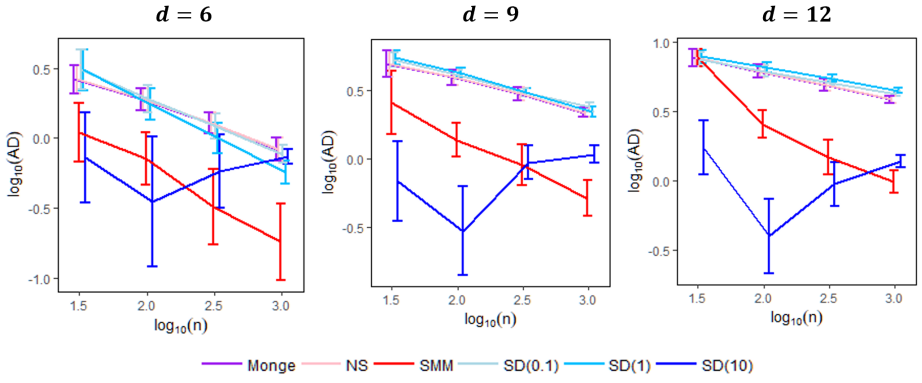

Figure 3.4: NORMAL MODEL I: the mean and standard deviation of the absolute deviation of 2-Wasserstein distance estimates with respect to n (in *log-log* scale).

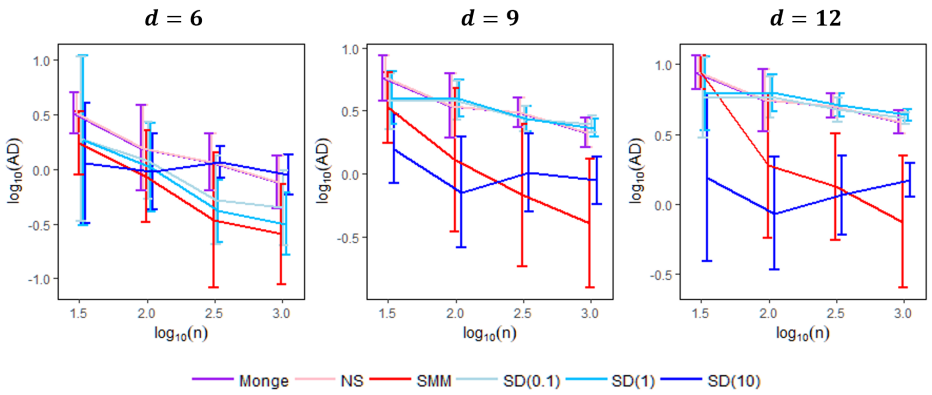

Figure 3.5: NORMAL MODEL II: the mean and standard deviation of the absolute deviation of 2-Wasserstein distance estimates with respect to n (in *log-log* scale).

As shown in Figure 3.3, Figure 3.4 and Figure 3.5, SMM (red line) performs well in all scenarios. The estimation error of SMM converges fast to zero regardless of the dimensionality d and the data generating process. MONGE (purple lines) and NS (pink lines) yield the same result in all scenarios, which is expected as the Brenier's theorem (Brenier, 1991) guarantees the equivalence of the Monge formulation and the Kantorovich formulation under our simulation setups. Besides, the estimation error of MONGE and NS decreases slowly as d increases, indicating that they are vulnerable to the "curse of dimensionality".

The three SD methods perform well in the UNIFORM MODEL; nevertheless, they do not perform well in the both NORMAL MODELS. For example, SD(0.1) and SD(1) perform even worse than MONGE, the naive Monge estimator. Also, we observe that the estimation errors of SD(10) fail to converge to zero as n increase. The inconsistent performance of SD methods shows that the bias term is largely affected by the true Wasserstein distance between two distributions and the choice of regularization parameter.

3.5.2 Hypothesis testing for distributional equivalence

Testing the equivalence of two probability distributions is a fundamental but challenging problem in statistics and machine learning. For example, in biostatistics, this two-sample test is essential for distinguishing the distributions of the treatment and control groups. For generative models, a problem of interest is to distinguish the real and synthesized populations.

This testing problem can be formulated as follows. Suppose that we observe two samples $\mathcal{A}_n = \{\boldsymbol{a}_i\}_{i=1}^n$ and $\mathbf{B}_m = \{\boldsymbol{b}_j\}_{j=1}^m$ from two underlying distri-

butions α and β, respectively. The null and alternative hypotheses are defined as

$$H_0 : \alpha = \beta \quad \text{versus} \quad H_A : \alpha \neq \beta.$$

Without any distributional assumption of α and β under the null, the above test is a non-parametric testing problem. As a measure of divergence between distributions, the Wasserstein distance has long been used for carrying out non-parametric two-sample tests (Del Barrio et al., 2000; Del Barrio et al., 1999; Munk & Czado, 1998; Ramdas et al., 2017). However, the existing literature of Wasserstein distance-based two-sample tests are mainly focused on the univariate case since the empirical Wasserstein distance suffers from the "curse of dimensionality". To address this issue, we propose to use the Smoothed Monge map (SMM) between $\mathcal{A}_n$ and $\mathbf{B}_n$ as a modified test statistic when d is moderate or large. We also compare the performance of SMM with the empirical Wasserstein distance-based test (WD), the MMD test (MMD) (Gretton et al., 2012), and the Sinkhorn divergence-based test ($SD(\epsilon)$) (Ramdas et al., 2017) with the regularization parameter ϵ being 0.1, 1, and 10.

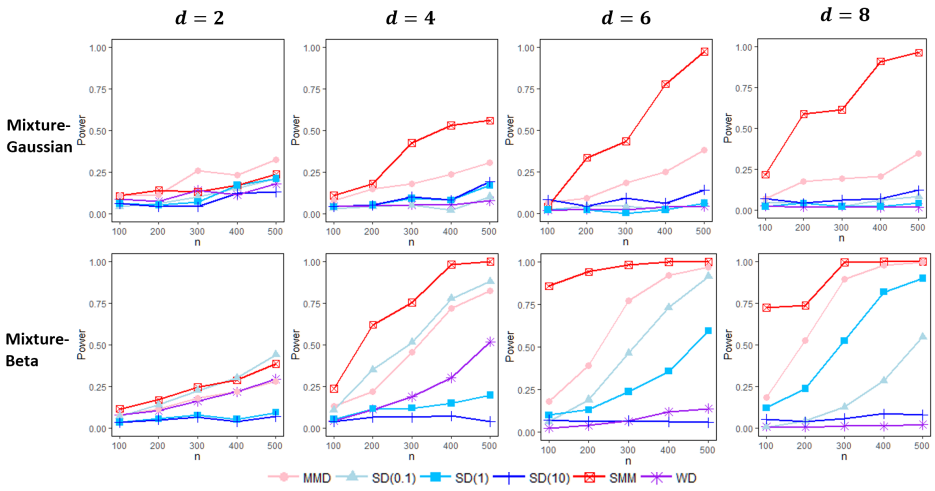

Figure 3.6: Power *vs.* sample size for different testing method. The upper row represent MIXTURE GAUSSIAN MODEL and the lower row represent MIXTURE BETA MODEL. Each column represent a different number of dimension d.

For an overall sample size N, we generate a treatment group size n_t from a binomial distribution Binomial$(N, 0.5)$, and set the test group size as $n_c = N - n_t$. We consider the following two models

(I) MIXTURE GAUSSIAN MODEL: For the treatment group, we generate n_t i.i.d. observations from a d-dimensional mixture-Gaussian distri-

bution,

$$a_{ij} \sim \frac{1}{2}\text{Normal}(1,1) + \frac{1}{2}\text{Normal}(-1,1),$$
$$i = 1, \ldots, n_t, \quad j = 1, \ldots, d.$$

For the control group, we generate n_c i.i.d. observations from another d-dimensional mixture-Gaussian distribution

$$b_{ij} \sim \frac{1}{2}\text{Normal}(1,1) + \frac{1}{2}\text{Normal}(-1,0.8),$$
$$i = 1, \ldots, n_c, \quad j = 1, \ldots, d.$$

(II) MIXTURE BETA MODEL: For the treatment group, we generate n_t i.i.d. observations from a d-dimensional mixture-Beta distribution,

$$a_{ij} \sim \frac{1}{2}\text{Beta}(5,7) + \frac{1}{2}\text{Beta}(7,5),$$
$$i = 1, \ldots, n_t, \quad j = 1, \ldots, d.$$

For the control group, we generate n_c i.i.d. observations from another d-dimensional mixture-Beta distribution

$$b_{ij} \sim \frac{1}{2}\text{Beta}(4,5) + \frac{1}{2}\text{Beta}(5,4),$$
$$i = 1, \ldots, n_c, \quad j = 1, \ldots, d.$$

We set the overall sample size $N \in \{100, 200, \ldots, 500\}$ and the dimensionality $d \in \{2, 4, 6, 8\}$. The significant level is set to be 0.05 for all tests. For each two-sample test method, the asymptotic variances, as well as the critical values, are calculated with 500 replications. We empirically evaluate the power of a test as the percentage of replications that the null hypothesis was correctly rejected, based on 500 independent replications.

Figure 3.6 presents the power versus sample size for the two-sample tests based on different OTM estimators. It is no surprise that the performance of WD gradually deteriorates as the dimensionality d increases. This observation is in-line with the "curse of dimensionality" issue, as we have discussed. Sinkhorn divergence based methods have mediocre performance for all four values of d. SD(0.1) and SD(1) performs better than WD when $d = 6$ and $d = 8$. This justified the finding in Genevay et al., 2019 that the Sinkhorn divergence reduces over-fitting by imposing a regularization term. However, a large regularization value can create an overwhelming bias term that can hurt the performance of the

Sinkhorn method. As an example, SD(10) has the worst performance among all competitors. MMD method has a decent performance in all cases, as its power increases with the sample size. In general, SMM outperforms its competitors, and its power converges to one fast when $d \geq 4$. The testing results supported our claim that SMM is able to overcome the "curse of dimensionality" issue and induce a consistent estimator of Wasserstein distance.

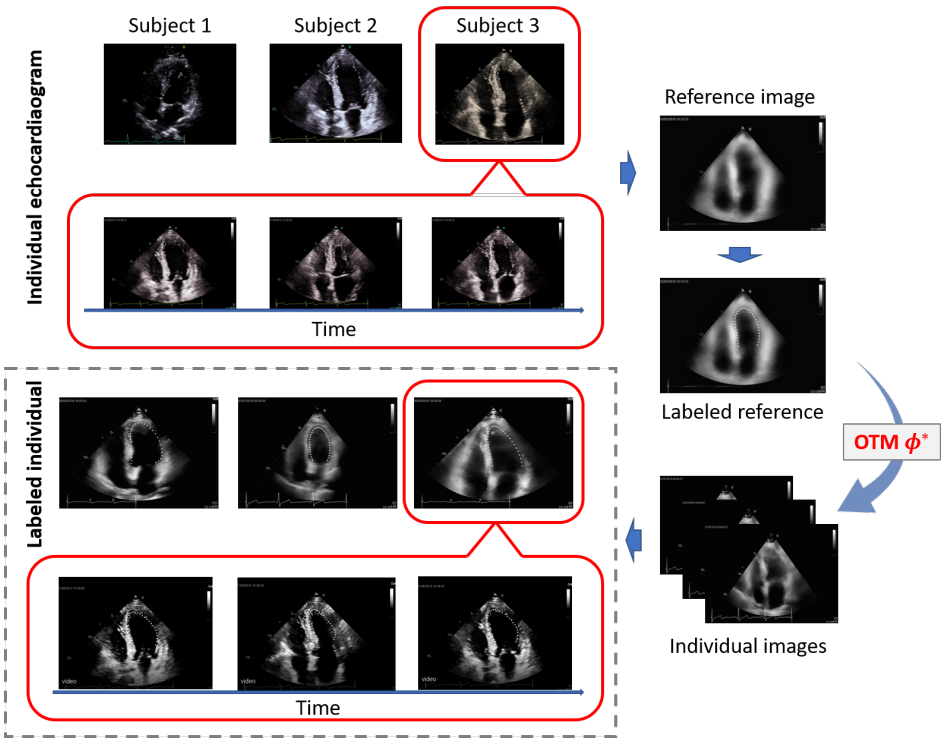

Figure 3.7: Illustration of echocardiogram tracing procedure and application results.

We apply the reference-based tracing method to trace the myocardial movement through SMM. We label on a reference image, which is the empirical Wasserstein barycenter of echocardiograms. Then we transport the labeled reference image to each individual image. The procedure is illustrated in figure 3.7. After the transportation, the transported label can capture LV-Endo in every frame of individuals. Combining all frames together, the transported labels trace the myocardial movement of LV-Endo. The tracing results are showed in figure 3.8.

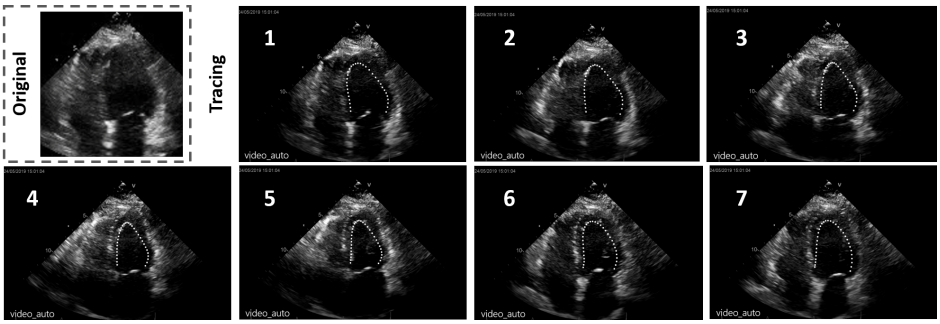

Figure 3.8: Tracing results of LV-Endo in echocardiogram through SMM.

We further evaluate our tracing result by calculating the movement features of certain segments of LV-Endo. We regard the feature values manually obtained by clinicians as the golden truth. When comparing the results of SMM and the manually obtained values, we claim that the tracing result is valid. The results are showed in table 3.1.

Table 3.1: Myocardial movement features for each segments in LV-Endo.

Method	Seg1	Seg2	Seg3	Seg4	Seg5	Seg6	Seg7
Manual	-11	-17	-28	-29	-29	-21	-17
SMM	-12	-16	-27	-31	-29	-19	-14

3.6 Concluding Remarks

We propose a novel optimal transport map estimator, named Smoothed Monge Map (SMM). SMM tackles the "curse of dimensionality" problem by applying univariate smoothing splines to each dimension of the empirical OTM. Hence, SMM can effectively alleviate the over-fitting issue and overcome the "curse of dimensionality." The 2-Wasserstein distance induced by SMM is $\sqrt{n}$-consistent and admits an optimal normality property. The superior performance of SMM over mainstream competitors has been justified by numerical experiments. Besides, we apply SMM to test the distributional equivalence of two high-dimensional random samples. In real data analysis, we use SMM to generate the reference image and transport the labeled points to each individual image. The transported labled points can help trace the myocardial movement of LV-Endo in echocardiogram. SMM also has the potentially to be employed in various interesting applications, including but not limited to image classification, domain adaptation, image registration, and generative model.

Chapter 4

Echocardiography Based Screening for Coronary Heart Disease Using An Ensemble Machine Learning Approach

Coronary heart disease is a global epidemic that leads to roughly 1/3 of deaths worldwide. Extensive clinical evidence suggests that a preventive screening of coronary heart disease at an earlier stage can greatly reduce mortality rate. The popular preventive procedures, e.g, the stress ECG test, are labor intensive and time consuming for clinicians. Moreover, these procedures may increase the risk of heart failure and are practically infeasible for senior subjects or subjects with disability. To address these issues, we present an echocardiography based screening method that only uses clinical records and myocardial movements as diagnosis features. We develop a ensemble machine learning approach to screen the patients. The entire practice only takes 30 seconds to complete. Based on a clinical trail that is conducted in the Beijing Hospital, the prediction accuracy of our method on testing data can reach 88% accuracy. Our work lays a foundation for the deployment of echocardiography-based screening tools for global improvements of cardio health.

4.1 Introduction

Coronary heart disease(CHD) is a global epidemic that led to 17.92 million (roughly 1/3 of) deaths worldwide in 2016 (Arnett et al., 2019; Lloyd-Jones

et al., 2010; Roth et al., 2017; Turco et al., 2018). It is reported that the ischemic heart disease, a late stage of CHD, killed 8.92 million people in 2015 and ranked as the number one killer among all diseases (Roth et al., 2017). The growing mortality rate of CHD not only causes significant loss on human resources but also causes many social problems. For example, the medical cost for CHD has increased exponentially in the past decades. Extensive clinical evidence suggests that a preventive screening of CHD at an earlier stage can greatly reduce mortality rate, improve prognosis, and more importantly, provide therapeutic guidance for patients (Thomas et al., 2018). A powerful screening method is highly desirable to curtail the global mortality burden and the social problems that come with CHD.

Despite the urgent needs, an efficient and clinically effective CHD screening procedure is still lacking because patients in very early stages of CHD usually have no visible clinical symptoms; as a consequence, CHD remains one of the leading causes of death even among developed countries. The majority of CHD diagnostic procedures are the radiology based approach such as computed tomography angiography (CTA) and coronary angiography (CA). These methods can directly visualize coronary artery and quantify the level of artery occlusion, and thus are considered as the gold standard for diagnosis. Though the radiology based methods are fairly effective in CHD diagnosis, their applications on preventive practice, especially on screening asymptomatic subjects are severely limited by the high operational cost, the requirement of expensive and high-maintenance equipment, the need for experienced medical researchers (Nicholls, 2019). More importantly, these procedures may bring the potential surgical risk and the radiology side effect on subjects.

A much less explored alternative are the echocardiography based diagnosis methods, which are commonly used to visualize the movements of the myocardium. Because CHD prevents patients' coronary artery from efficiently pumping blood to maintain healthy myocardial movements, echocardiograms provide a clue for CHD's diagnosis as it can be used to visualize myocardial movements. In fact, clinical practice suggests that some echo-cardiology based techniques, such as two dimensional speckle tracking (2D-STE) (Blessberger & Binder, 2010a), can indeed infer CHD with obvious symptoms such as severe coronary artery occlusion. Accumulating evidence shows that some dynamic features extracted by 2D-STE, such as global longitudinal strain (Skaarup et al., 2018) and time-to-peak strain change, are significantly different between the control group and patients with myocardial ischemia (Yang et al., 2013). These clinical observations suggests that echocardiography might be the new promise

for CHD screening as abnormal myocardial movement can be used to infer CHD (Blessberger & Binder, 2010b).

Although 2D-STE based screening is promising for its low operational cost, high practical convenience and high clinical safeness, how to effectively use it for individual assessment of cardio health is still an open question. There are no effective assessment models that can single out early-stage CHD patients with adequate sensitivity and specificity. It remains unknown which set of echocardiography based features can effectively quantify the significance of the myocardial change in response to minor myocardial anomaly. The current 2D-STE based research heavily relies on clinicians' experiences and domain-specific knowledge, and thus are highly subjective (Michel et al., 2017). Moreover, the requirement of laboratory-based practice as opposed to in-field and real-time analysis, limits their utility for large-scale population practice. Thus, in spite of the great promise, the preventive impact of 2D-STE technique still has not been achieved.

Coronary angiography is the gold-stand to confirm the stenosis. Although there are a huge number of CHD patients in the world, angiography is not recommended to all suspicious patients for its potential complication. Particularly, angiography is not appropriate for elder patients and those with severe renal failure or other end-stage organ failure. This echocardiogram AI model is almost applicable to in all those patients. Even more, it can help to rule out coronary heart disease, avoiding unnecessary coronary angiography. Coronary computed tomography angiography (CTA) requires contrast agents, so patients with renal and cardiac dysfunction are at greater risk. Moreover, the negative predictive value of coronary CTA was more sensitive. Tests like MRI and single-photon emission computed tomography (SPECT), which take too much time or have other side effects, are not commonly used.

Recently, the rapid development of computer vision and machine learning techniques has triggered a medical technology revolution. For example, the first clinical-grade computational pathology algorithm was proposed in (Campanella et al., 2019) for diagnosis of three types of cancers with an average accuracy of around 0.98. The deep learning algorithms have been developed for image processing for echocardiograms (Madani et al., 2018; J. Zhang et al., 2018). These unprecedented efforts changed our decision supporting system from experience-based decisions to quantitative-based decisions. Statistical models in place of highly skilled personnel plays a pivotal role in disease diagnosis. Unlike traditional health assessment methods which heavily relies on medical researchers' experience, the quantitative based methods rely on a series of quantitative, reliable, reproducible, multiplexed measurements that are ex-

tracted from large amounts of clinical practices. Their application requires no user intervention for field deployment and data capture, which can effectively bypass subjective errors.

In this article, we propose a novel ensemble machine learning method for 2D-STE based CHD assessment. In particular, we develop a classification stacking method to aggregate prediction power of 19 popular machine learning methods to provide best possible prediction outcomes. Using the new approach to learn a model from echocardiographic data, we can achieve much higher sensitivity and specificity than that using existing methods. The model obtained from our method can automatically trigger an early stage CHD warning, and thus can greatly save human efforts. The proposed methods integrate results of multiple popular machine learning models, each of which has around a 70% prediction accuracy for patients that need revascularization. By borrowing strengths from different machine learning models, our proposed model improves the classification accuracy from 70% to 88%, which supports our proposal of prototyping a tool for future population-based cardio health assessment.

4.2 Methods and Materials

4.2.1 Human Subjects

The study was approval by the Institutional review board of Beijing Hospital, and appropriate individual subject consent was provided for subjects.

The echocardiogram was performed by one experienced clinician on a GE Vivid E9 system (GE Medical Systems, Horten, Norway). Patients' images were stored in the same machine. Images were transported to offline system EchoPac version 201 (GE Healthcare, Horten, Norway), further analyzed by an experienced investigator.

4.2.2 Data and Feature

From March 2018 to October 2019, 836 subjects was enrolled in clinical trial (NCT03905200), of which 555 were diagnosed as CHD positive by coronary angiography (CA) or coronary computed angiography (CCA). Among the 555 CHD positive subjects, 424 of them were also examed by echocardiography one day before angiography was conducted. Patients with CAG level less than 3 are classified as CHD negative, others are classified as CHD positive.

We choose 67 features (numerical) in 2D-STE together with 5 clinic features (categorical) as our predictors to predict the risk of CHD.

4.2.3 Ensemble machine learning

We aim to build classification model that takes the features from echocardiography as input and predict whether the subject has CHD. Current classification methods rely on various underlying model assumptions, which hold the key to the success of the methods. When the data is highly heterogeneous and noisy such as the echocardiographic data that we are analyzing, it is not clear which model is adequate as the underlying assumptions are usually hard to validate. Single classification model does not provide satisfactory prediction results.

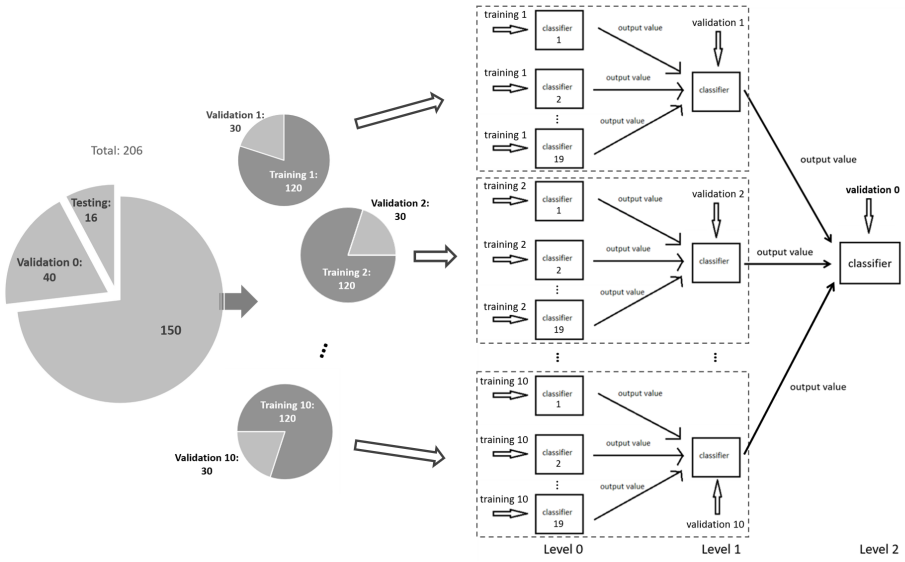

Figure 4.1: Flowchart of the two-level model stacking.

To improve the classification performance, we apply a number of popular classification methods to our data and employ the ensemble machine learning approach borrow the strength of all these classifiers and build the final prediction model. The ensemble machine learning methods can be divided into three classes: bagging, boosting, and stacking (Z.-H. Zhou, 2012). In particular, bagging aims to reduce variance; boosting decreases bias, and stacking improves the prediction. We opt to the use the stacking (Breiman, 1996; Wolpert, 1992). This approach is particularly popular when the signal-to-noise ratio of the data is low. We propose here a two-step classifiers stacking method, under which we first train individual classifier using randomly sampled training data and then train the stacked classifier model using the rest of training data (validation data). By randomly partition the training set multiple times, we can reduce the classification error that caused by wrong model aggregation. In particular, we divide the 424 subjects into a training set which contains 360 subjects and a testing set

which contains 64 subjects. The training set was further divided into a first step training set with 288 randomly selected subjects and a first step validation set with 72 subjects. For first step training set, we repeatedly sample 58 individuals randomly from it as a validation set for the first level model ensemble, and use the rest of 230 subjects to train the candidate models. In this chapter, we build 19 candidate classifiers using 19 most popular machine learning approaches and select the the best of them by majority votes. As shown in Figure 4.1, by sampling 10 times the validation error is starting to saturate although further improvement can be made with more patients enrolled.

4.3 Results

4.3.1 Feature Extraction

For each patient, the recorded echocardiography consists of three parasternal short-axis standard sections: mitral valve section, papillary section and apical section, as well as three apical standard sections: 4-chamber view section, 2-chamber view section, and the longitudinal long-axis view section. Left ventricular wall (LVW) was divided into 17 segments, each of which was analyzed individually. Peak systolic longitudinal and radial strains were assessed in all 17 segments. The epicardium and endocardium of the left ventricle (LV) were traced automatically and adjusted manually when required at the end-systole. Mid-myocardial border was determined at the midpoints between the endocardial and epicardial borders. The regions of interest (ROI) covers the endocardium, myocardium, and epicardium. ROI was adjusted locally if it is off-track. In 2D-STE echocardiography, the most important parameter is strain, which quantifies the deformation of myocardium by recording the contractions. Layer-specific analysis of endocardial, mid-myocardial, and epicardial strains were performed in the six views for the radial strain. Global longitudinal strain (GLS) is obtained by averaging the values of all segments. Myocardium usually consists of three heterogeneous layers of muscle fibers (Vendelin et al., 2002).

To assess the deformation of certain myocardial regions such as the left ventricular wall (LVW), we first divide the entire LVW into 17 functional homogeneous segments, and then generate a strain tensor to record the deformation of all sub-regions. The tensor strain is simply a regional extension of strains that are commonly used to quantify the shortening, thickening and lengthening of each sub-region's myocardium in both longitudinal and radial directions and in a short period.

As the longitudinally orientated myocardial fibres is the most susceptible to ischemia (Skaarup et al., 2018), we use global longitudinal strain (GLS) to form features that can help predict CHD. It was shown in (Delgado et al., 2008) that the GLS can successfully predict CHD (AUC=0.92) for patients with non–ST-segment elevation acute coronary syndromes (NSTE-ACS).

Ventricular contractive dysfunction occurs prior to ECG change in sub-endocardium. As the left ventricular wall which consists three layers of muscle fibres with heterogeneous strains (Vendelin et al., 2002). Layer-specific strain is associated with coronary artery disease independently(L. Zhang et al., 2016). In coronary disease, layer-specific strain is quite helpful because longitudinally orientated myocardial fibres located in the sub-endocardium is known to be most susceptible to ischemia (Reimer et al., 1977). The diagnostic accuracy tends to be higher than ECG, troponin, and GRACE that is computed using tomography (Caspar et al., 2017). Recently, global longitudinal strain has been recommended as the top priority index in diagnosing diverse cardiac diseases (Nagueh et al., 2016; Nauta et al., 2018).

In myocardium, micro-vascular communications are network structured which can form some dual arterial perfusion zones. Simply relying on single index might be inaccurate to decide the etiology. In our model, assessment of myocardium ischemia was measured by longitudinal strain, strain rate, time to peak, and a specific layer strain. Such multiple indices integration strategy employed by our model can reduce the prediction error by each index (Gjesdal et al., 2007; L. Zhang et al., 2016).

Table 4.1: P-values of 2D-STE features.

	GLPS (**p-value: 0.002**)			PSD		
	Epi	Mid	Endo			
p-value	.024	.049	.076	.731		
	Peak strain	Strain rate	Time to peak	SAX-AP (**p-value: 0.876**)		
				Epi	Mid	Endo
p-value	.024	.041	.179	.982	.952	.598
	SAX-PM (**p-value: 0.503**)			SAX-MV (**p-value: 0.277**)		
	Epi	Mid	Endo	Epi	Mid	Endo
p-value	.663	.654	.682	.247	.175	.516

We chose 64 features (numerical) in 2D-STE together with 5 clinic features (categorical) as our predictors to build the classification model to predict whether the subject has CHD, as shown in table 4.2.

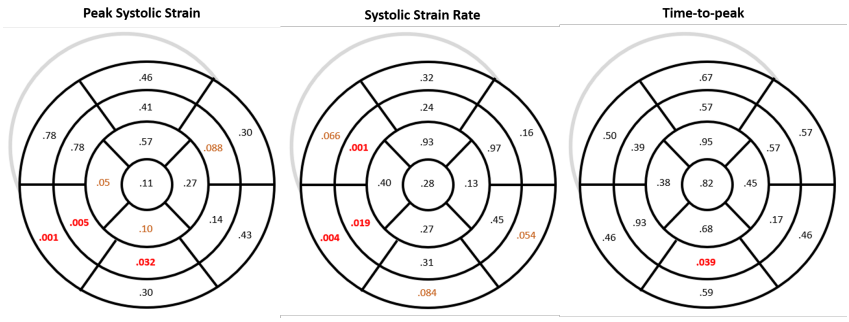

Figure 4.2: P-values measuring differences on PSS, SSR and TP of 17 segments between case and control groups.

Table 4.2: Features chosen to be predictors in CHD prediction model.

2D-STE features		
Longitudinal strain (mid-layer)	Peak systolic strain (PSS)	17 segments
	Rate of systolic strain (SSR)	17 segments
	Time-to-peak (TP)	17 segments
Global strain for radio (GS)	Mitral valve level (MV)	3 layers (ENDO/MID/EPI)
	Papillary muscle level (PM)	3 layers
	Apical level (AP)	3 layers
Global longitudinal peak strain (GLPS)		3 layers (ENDO/MID/EPI)
Peak standard deviation (PSD)		
Clinic features		
Age (integer)		
Gender (M/F)		
Hypertension (Y/N)		
Diabetes (Y/N)		
Hyperlipemia (Y/N)		
Smoke (Y/N)		
Family history (Y/N)		

4.3.2 Principle component analysis

To reduce the dimension of features, we applied principle component analysis (PCA) on the 17 segments of peak systolic strain (PSS), rate of systolic strain (SSR) and time-to-peak (TP).

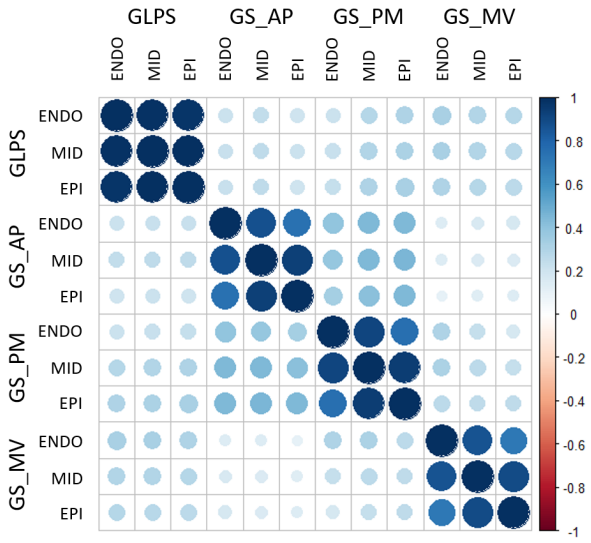

Figure 4.3: Correlation matrix of global longitudinal strains and radial strains of apical level, papillary muscle level and mitral valve level.

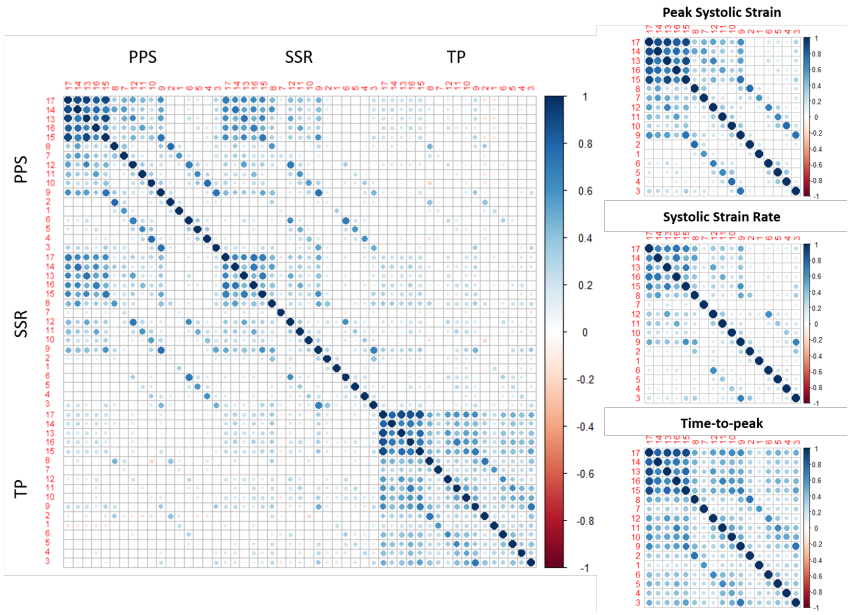

Figure 4.4: Correlation matrix of 17 segments on PSS, SSR and TP. The column in the left panel represents the correlationship of 17 segments for PSS, SSR and TP respectively.

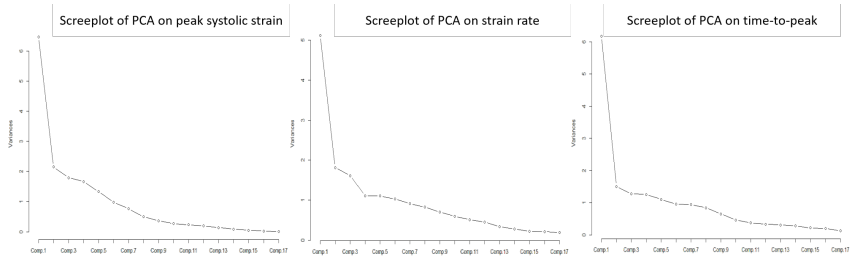

Figure 4.5: Screeplot of PCA on peak systolic strain, systolic strain rate and time-to-peak.

We first study the correlations among the numerical features. Figure 4.3 shows the correlation matrix of global longitudinal strains and radial strains. We can see that longitudinal strains are poorly correlated with radial strains, and for radial strains, each level is poorly correlated with each other. Figure 4.4 shows the correlation matrix of 17 segments on PSS, SSR and TP. From the plot, we can see that PSS is correlated with SSR, while TP is poorly correlated with PSS and SSR. When examing the correlationship of all 17 segments for PSS, SSR and TP respectively, we can see that for PSS, SSR and TP, the apex and apical layer are highly correlated; for PSS, six segments in the middle layer are highly correlated to their neighboring segments in the basal layer; for SSR, middle layer and basal layer are poorly correlated; and for TP, the correlation of all 17 segments are higher. Based on the results of the correlation study, we choose to conduct PCA on PSS, SSR and TP respectively. Figure 4.5 shows the scree-plots of PCs for these three types of features. We can find obvious ankles in each plot, which can lead us choosing the number of PCs. Figure 4.6 is the heatmaps of the first 3 PC loadings for PSS, SSR and TP. From Figure 4.6, we can see that 1) for PSS, the first PC represents the apex, the apical layer and the basal/mid anteroseptal; the second PC represents the basal/mid inferoseptal, the basal/mid inferior, and the basal/mid inferolateral; the third PC represents the basal/mid anterior and the basal/mid anterolaterial. 2) for SSR, the first PC represents the apex and the apical layer; the second PC represents the basal/mid anteroseptal and the basal/mid inferolateral; the third PC represents the basal layer. 3) for TP, the first PC represents the overall average of the 17 segments; the second PC represents the basal/mid anterior, the basal/mid anterolateral, and the basal/mid inferolateral; the third PC is similar as the second PC. Thus we choose the first 3 PCs for PSS and SSR, and the first 2 PCs for TP.

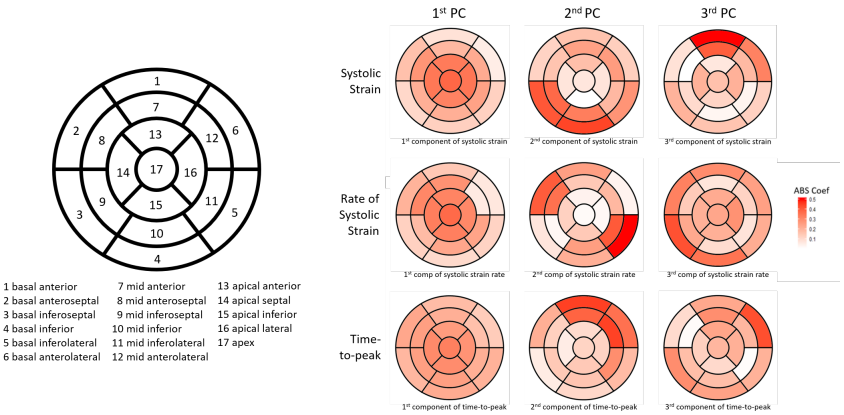

Figure 4.6: Heatmaps of contributions of 17 segments in first three PCs of peak systolic strain, systolic strain rate and time-to-peak. Column from left to right represents the first PC to the third PC respectively, and the top row represents PSS, the middle row represents SSR and the bottom row represents TP.

4.3.3 Two-step classifier stacking

Our implementation achieved an accuracy of 0.877 on the test set, with sensitivity 0.903 and specificity 0.843. The accuracy is significantly better than 0.71 the highest accuracy achieved by an individual model and constituting a 23.5% improve compared with its performance on the test set (Tab. 4.3). Figure 4.7 shows ROC curves for each individual model and our stacking model. The solid black line represents our 2-level stacking model and other colored lines represent each individual model in table 4.3 respectively. The result shows the AUC for the individual models range from 0.503 to 0.773, while the AUC for the stacking model is 0.904, which is significantly higher than the individuals.

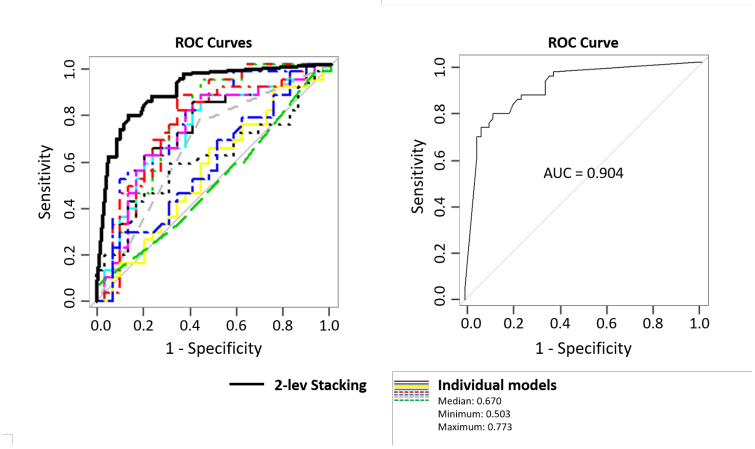

Figure 4.7: ROC curves on 2-level stacking model and individual models.

Table 4.3: Testing accuracy of individual classification model.

Model	Accuracy
logistic regression	.68
penalized logistic regression	.71
cumulative probability model	.69
random forest	.59
weighted subspace random forest	.59
SVM with class weight	.70
SVM with polynomial kernel	.66
SVM with radial kernel	.64
K-nearest neighbor	.58
LDA	.70
sparsed LDA	.59
naive Bayes	.64
Bayes generalized linear model	.68
Gaussian process with polynomial kernel	.70
Gaussian process with radial kernel	.65
Neural network	.63
Monotone multi-layer perceptron neural network	.69
model average neural network	.65
stochastic gradient boosting	.58

4.4 Discussion

The contribution of our 2D-STE echocardiography screening method is twofold. First, its can effectively use the information obtained from echocardiography for CHD screening reduce the mortality and morbidity. Second, it is the first automatic and clinically applicable screening method which can greatly save medical efforts. Some imaging technologies have been applied in the prevention work to reduce the morbidity and mortality (Gomez-Pardo et al., 2016). However, echocardiogram is one of the most promising modalities in the cardiovascular field.

Comparing to other modalities, 2D-STE echocardiography has its unique advantages. The sub-endocardial myocardial fibres are oriented longitudinally, so the longitudinal myocardial function is affected primarily when ischemia is onset. The global longitudinal strain presenting the ventricular contractive dysfunction occurs prior to ECG change. Therefore, the machine learning model is able to offer more information than the ECG. Conventional echocardiographic parameters are mainly based on a visual assessment of the ventricular

wall motion. Subtle abnormalities might be unresolved by human eyes (Caspar et al., 2017). The ability of conventional echocardiogram is limited in diagnosing CHD, especially in the early stage. 2D-STE image is able to detect minimal abnormalities of systolic function (Delgado et al., 2008; Di Bella et al., 2014).

Coronary angiography is the gold-stand to confirm the stenosis. Although there are a huge number of CHD patients in the world, angiography is not recommended to all suspicious patients for its potential complication. Particularly, angiography is not appropriate for elder patients and those with severe renal failure or other end-stage organ failure. This echocardiogram AI model is almost applicable to in all those patients. Even more, it can help to rule out coronary heart disease, avoiding unnecessary coronary angiography. Coronary computed tomography angiography (CTA) requires contrast agents, so patients with renal and cardiac dysfunction are at greater risk. Moreover, the negative predictive value of coronary CTA was more sensitive. Tests like MRI and single-photon emission computed tomography (SPECT), which take too much time or have other side effects, are not commonly used.

The potential clinical applications of this echocardiogram machine learning model are enormous. STE was applied clinically as a supplementary diagnostic method before, and now practical to those suspicious coronary patients. We clinicians have the responsibility to apply the safest and most effective means to the high risk populations. Study has demonstrated that early intervention can reduce the mortality and major events in coronary heart disease patients (Gaye et al., 2017). We are sure that this machine learning tool in the study would have the revolutionary impact in the diagnosis modality strategic. With this novel implement, speckle tracking image is not a supplementary, but a practical and ideal non-invasive method for the early diagnosis of coronary heart disease for clinical physicians. In fact, this machine learning model of 2D-speckle tracking will also be helpful in re-evaluating the recovery from ischemia events after initial hospitalization. Also, it can be recommended as routine in the routine physical examination. This machine learning echo-model is now ready to be applied in daily clinical practice. The global rates of coronary heart disease and its injurious effects would decrease largely with this novel application. It would be a giant leap in the public health control work.

Our study is a single center study. The data derived from the same vendor machine. The differences of echo-cardiographic inter-vendors and post-processing algorithms were not taken into account. During processing, if the images were not clear enough, the software can not accurately recognize the epicardial or endocardial border. Therefore, it may have some bias to the results. Another limitation is that the Speckle tracking analysis process was not

automatically. The individual difference between physicians might also have some impact to the interpretation.

Chapter 5

Conclusion

In the big data era, observations in different modes are generated from vastly different sources. Data fusion that combines multi-source data into fused data is widely used to extract information from a large variety of data and may dictate the ultimate performance in big data enterprise. In this thesis, we illustrate data fusion through three levels, data level fusion, feature level fusion, and decision level fusion. We propose a set of statistical methods under each level of data fusion. The methods are shown to be effective in information integration. In Chapter II, we introduce the feature space fusion method. Such a method can integrate common features of different datasets while retaining data heterogeneity. In Chapter III, we combine optimal transport with smoothing splines to improve the estimation of Wasserstein distance . With better estimation, we can generate effective references and provide a more accurate image tracing process. In Chapter IV, we discuss how to use ensemble learning to improve predicting power in CHD screening problem. Data methods can also be applied to image classification, domain adaptation, and generative model. Future studies might explore data fusion methods under network structures, or consider data fusion under certain practical concerns such as data privacy and communication limitations.

Appendix A

Proof for Chapter 2

A.1 Proofs of main theoretical results

Let $\mu_B(u) = E\left(X|B^TX = u\right)$ and $w_B(u) = E\left\{XX^T|B^TX = u\right\}$. following conditions 2.4.1 - 2.4.5, we first show the convergence of algorithm 2. To simplify the notation, we consider only two-node case, i.e. $S = 2$, and define the weight matrix

$$W = (w_1, w_2)^T \otimes I_p, \tag{A.1}$$

where $w_1 + w_2 = 1$. Recall that the estimate of B within node s in the t-th iteration is $B_{(t)}$. Let $\rho_{j,t}^{(s)} = \frac{1}{n_s}\sum_{i=1}^{n_s} K_{h(t)}(B_{(t)}^T \boldsymbol{x}_{ij,t}^{(s)})$. It follows from step II in section 2 that

$$\begin{aligned}\Gamma^{(t+1)} =& \ell\left(B\right) + \left\{\sum_{j,i}^{n_s} \rho_{j,t}^{(s)} K_{h_{(t)}}\left(B_{(t)}^T \tilde{\boldsymbol{x}}_{ij,t}^{(s)}\right) \tilde{\boldsymbol{x}}_{ij,t}^{(s)} \left(\tilde{\boldsymbol{x}}_{ij,t}^{(s)}\right)^T\right\}^{-1} \\ & \times \sum_{j,i}^{n_s} \rho_{j,t}^{(s)} K_{h(t)}\left(B_{(t)}^T \tilde{\boldsymbol{x}}_{ij,t}^{(s)}\right) \tilde{\boldsymbol{x}}_{ij,t}^{(s)} \left\{Y_i - a_{j,t}^{(s)} - \ell\left(B\right)^T \tilde{\boldsymbol{x}}_{ij,t}^{(s)}\right\},\end{aligned} \tag{A.2}$$

where $\tilde{\boldsymbol{x}}_{ij,t}^{(s)} = \mathbf{b}_{j,t}^{(s)} \otimes \boldsymbol{x}_{ij}^{(s)}$ as defined in section 2. Following Step III, we then estimate $B_{(t+1)}^{(s)}$ in node s. By fusing neighboring $B_{(t+1)}^{(s)}$'s, here are $B_{(t+1)}^{(1)}$ and $B_{(t+1)}^{(2)}$, we obtain the estimate $B_{(t+1)}$ in the next iteration. By Lemmas 1-5 below, we can establish the following recurring relationship

$$\ell\left(B_{(t+1)}\right) - \ell\left(B\right) = \Theta\left\{\ell\left(B_{(t)}\right) - \ell\left(B\right)\right\} + e_{(t)} \tag{A.3}$$

with $|\Theta| < 1$ and $|e_{(t)}| = o(1)$ almost surely. Here $|\cdot|$ is the spectral norm of matrix. Such a recurring relationship implies the convergence of the algorithm.

Before the proof, we need to introduce some notations. Let

$$G(u) = E(Y|B^T\boldsymbol{x} = u), \delta_{kh} = \left\{\log n/\left(nh^k\right)\right\}^{1/2},$$

$$\delta_n = (\log n/n)^{1/2}, r_{kh} = h^2 + \delta_{kh},$$

and $\boldsymbol{x}_{ix} = \boldsymbol{x}_i - x$.

Let $\mu_{kp} = \int K\left(v_1, \cdots, v_p\right) v_1^k dv_1 \cdots dv_p$. Let $f_B(u)$ be the density function of $B^T\boldsymbol{x}$, $\nu_B(x) = \mu_B\left(B^T x\right) - x$, and

$$\overline{w}_B(x) = w_B\left(B^T x\right) - \mu_B\left(B^T x\right)\mu_B^T\left(B^T x\right).$$

For simplicity, we denote $f_B(B^T\boldsymbol{x})$ by $f_B(\boldsymbol{x})$, and $\mu_B(B^T\boldsymbol{x})$ by $\mu_B(\boldsymbol{x})$. For any square matrix A, A^{-1} denotes the inverse if exists, and A^+ denotes the Moore-Penrose inverse. Let $\mathcal{D}_x$ be any impact set of $\mathbb{R}^p$.

In order to prove the convergence of the algorithm, we introduce a set of lemmas (Lemma A.1.1 - A.1.4), whose proofs are similar to those of the lemmas in Xia et al., 2007.

Lemma A.1.1. *Let*

$$\begin{pmatrix} a_x \\ \mathbf{b}_x h \end{pmatrix} = \left\{\sum_{i=1}^n K_h\left(\boldsymbol{x}_{ix}\right)\begin{pmatrix} 1 \\ \boldsymbol{x}_{ix}/h \end{pmatrix}\begin{pmatrix} 1 \\ \boldsymbol{x}_{ix}/h \end{pmatrix}^\top\right\}^{-1} \times \sum_{i=1}^n K_h\left(X_{ix}\right)\begin{pmatrix} 1 \\ \boldsymbol{x}_{ix}/h \end{pmatrix} Y_i.$$

Under the assumptions, if $h \propto n^{-\zeta}$ with $0 < \zeta < 1/p$, then we have

$$a_x = G(B^T x) + \frac{1}{2}\sum_{\kappa=1}^q \nabla^2_{\kappa,\kappa} G(B^T x)h^2 + \mathcal{O}\left(h^3 + \delta_{ph}|x \in \mathcal{D}_x\right),$$

$$\begin{aligned}\mathbf{b}_x =& \nabla G(B^T x) + \left\{\mu_{2p} nh f(x)\right\}^{-1}\sum_{i=1}^n K_h\left(\boldsymbol{x}_{ix}\right)\left(\boldsymbol{x}_{ix}/h\right)\epsilon_i \\ &+ \mathcal{O}\left(r_{ph}|x \in \mathcal{D}_x\right).\end{aligned}$$

Lemma A.1.2. *Let*

$$\Sigma_n^{B'}(x) = n^{-1}\sum_{i=1}^n K_h\left(B'^T\boldsymbol{x}_{ix}\right)\begin{pmatrix} 1 \\ B'^T\boldsymbol{x}_{ix}/h \end{pmatrix}\begin{pmatrix} 1 \\ B'^T\boldsymbol{x}_{ix}/h \end{pmatrix}^T,$$

$$\begin{pmatrix} a_x^{B'} \\ \mathbf{b}_x^{B'} h \end{pmatrix} = \left\{n\Sigma_n^{B'}(x)\right\}^{-1}\sum_{i=1}^n K_h\left(B'^T\boldsymbol{x}_{ix}\right)\begin{pmatrix} 1 \\ B'^T\boldsymbol{x}_{ix}/h \end{pmatrix} Y_i.$$

Under the assumptions, if $h \propto n^{-\zeta}$ *with* $0 < \zeta < 1/p$*, then we have*

$$\begin{aligned} a_x^{B'} =& G\left(B^T x\right) + \nabla^T G\left(B^T x\right)\left(B - B'\right)^T \nu_{B'}(x) \\ &+ \frac{1}{2}\sum_{\kappa=1}^{d_0} \nabla^2_{\kappa,\kappa} G\left(B^T x\right) h^2 \\ &+ \mathcal{V}_{1n}^{B'}(x) + \mathcal{O}\left(\Delta_n^{B'} | x \in \mathcal{D}_x, B' \in \mathcal{B}\right), \end{aligned}$$

$$\begin{aligned} \mathbf{b}_x^{B'} h =& \nabla G\left(B^T x\right) h + Q_1^{B'}(x) h^3 + \mathcal{V}_{2n}^{B'}(x) \\ &+ \mathcal{O}\left(\Delta_n^{B'} | x \in \mathcal{D}_x, B' \in \mathcal{B}\right), \end{aligned}$$

where $\mathcal{B} = \{B : B^T B = I_q\}$, $\Delta_n^{B'} = h^4 + \delta_{d_0 h}^2 + h\delta_{B'} + \delta_{B'}^2$ *with* $\delta_{B'} = |B' - B|$,

$$\begin{aligned} Q_1^{B'}(x) =& \frac{1}{2} f_{B'}^{-1}(x) \nabla^2 G\left(B^T x\right) \nabla f_{B'}(x) \\ &+ \frac{1}{6}\mu_{4q}\left\{\nabla^3_{1,1,1} G\left(B^T x\right), \ldots, \nabla^3_{q,q,q} G\left(B^T x\right)\right\}^T, \end{aligned}$$

$$\begin{aligned} \mathcal{V}_{1n}^{B'}(x) &= \mathcal{E}_{n,1}^{B'}(x) - h\nabla^\top f_{B'}(x) \mathcal{E}_{n,2}^{B'}(x), \\ \mathcal{V}_{2n}^{B'}(x) &= \mathcal{E}_{n,2}^{B'}(x) - h\nabla f_{B'}(x) \mathcal{E}_{n,1}^{B'}(x), \\ \mathcal{E}_{n,1}^{B'}(x) &= \left\{n f_{B'}(x)\right\}^{-1} \sum_{i=1}^n K_h\left(B'^T \boldsymbol{x}_{ix}\right) \epsilon_i, \end{aligned}$$

and

$$\mathcal{E}_{n,2}^{B'}(x) = \left\{n f_{B'}(x)\right\}^{-1} \sum_{i=1}^n K_h\left(B'^T \boldsymbol{x}_{ix}\right)\left(B'^T \boldsymbol{x}_{ix}/h\right) \epsilon_i.$$

Lemma A.1.3. *Let* $\boldsymbol{x}_{ij}^{B'} = \mathbf{b}_j^{B'} \otimes \boldsymbol{x}_{ij}$ *where* $\mathbf{b}_j^{B'} = \mathbf{b}_{\boldsymbol{x}_j}^{B'}$*. Suppose conditions 2.4.1 - 2.4.4 hold and* $h \propto n^{-\zeta}$ *with* $0 < \zeta < 1/q$ *and* $\delta_{B'}/h \to 0$*. We have*

$$\begin{aligned} \left\{n^{-2}\sum_{j,i=1}^n K_h\left(B'^T \boldsymbol{x}_{ij}\right) \boldsymbol{x}_{ij}^{B'} \boldsymbol{x}_{ij}^{B'}\right\}^{-1} =& \left(I_q \otimes B\right) M_0^{-1}\left(I_q \otimes B^T\right) h^{-2} \\ &+ \left(I_q \otimes B\right) L_0 + L_0^T\left(I_q \otimes B^T\right) \\ &+ \frac{1}{2}\tilde{D}^+ \\ &+ \mathcal{O}\left\{\left(\tilde{r}_{qh} + \delta_{B'}\right)/h | B' \in \mathcal{B}\right\}, \end{aligned}$$

where $M_0 = E\left\{f_B(\boldsymbol{x}_i)\nabla G\left(B^T\boldsymbol{x}_i\right)\nabla^T G\left(B^T\boldsymbol{x}_i\right)\right\}$, $\tilde{r}_{qh} = h^2 + \delta_{qh} + \delta_{qh}^2/h^2$, L_0 *is a constant matrix, and*

$$\tilde{D} = E\left\{f_B(\boldsymbol{x}_i)\nabla G\left(B^T\boldsymbol{x}_i\right)\nabla^T G\left(B^T\boldsymbol{x}_i\right)\otimes\overline{w}_B(\boldsymbol{x}_i)\right\}.$$

Lemma A.1.4. *Suppose conditions 2.4.1 - 2.4.4 hold and* $h \propto n^{-\zeta}$ *with* $0 < \zeta < 1/q$ *and* $\delta_{B'}/h \to 0$. *We have*

$$\begin{aligned}
&n^{-2}\sum_{j,i=1}^{n} K_h\left(B'^T\boldsymbol{x}_{ij}\right)\mathbf{b}_j^{B'}\otimes\boldsymbol{x}_{ij}\left\{Y_i - a_j^{B'} - \ell(B)^\top\boldsymbol{x}_{ij}^{B'}\right\}\\
&= D\ell\left(B'-B\right) + \Phi_n + \mathcal{O}\left\{\tilde{\Delta}_n^{B'}|B'\in\mathcal{B}\right\},
\end{aligned}$$

where

$$\tilde{\Delta}_n^{B'} = h^4 + \delta_{qh}^2 + \delta_{B'}r_{qh}/h + \delta_{B'}^2 + \delta_n h,$$

$$\Phi_n = n^{-1}\sum_{i=1}^{n} f_B(\boldsymbol{x}_i)\nabla G\left(B^T\boldsymbol{x}_i\right)\otimes\nu_B(\boldsymbol{x}_i)\epsilon_i,$$

and

$$D = E\left\{f_B(\boldsymbol{x}_i)\nabla G\left(B^T\boldsymbol{x}_i\right)\otimes\nu_B(\boldsymbol{x}_i)\left[\nabla G\left(B^T\boldsymbol{x}_i\right)\otimes\nu_B(\boldsymbol{x}_i)\right]^T\right\}.$$

Let $\delta_{(t)} = \delta_{B'_{(t)}}$ denote the estimation error in the t-th iteration. By (A.2), lemmas A.1.3 and A.1.4 and the facts that $(I_q\otimes B^T)D = 0$ and $(I_q\otimes B^T)\Phi_n = 0$ if $\delta_{(t)}\log n/h_{(t)} = o(1)$ and $\delta_n/h_{(t)}^2 = o(1)$, we have

$$\begin{aligned}
\Gamma^{(t+1)} =& \ell(B_0) + \tilde{D}_0^+ D_0\ell\left(B_{(t)} - B_0\right) + \tilde{D}_0^+\Phi_n + (I_{d_0}\otimes B_0)L_0\mathcal{O}\left(c_n^{(t)}\right)\\
&+ \mathcal{O}\left\{\tilde{\Delta}_n^{B_{(t)}} + \left(\delta_{(t)}+\delta_n\right)\left(r_{d_0h_{(t)}} + \delta_{(t)}\right)/h_{(t)}\right\}\\
=& (I_{d_0}\otimes B_0)\left\{\ell(I_{d_0}) + \mathcal{O}\left(c_n^{(t)}\right)\right\} + \tilde{D}_0^+ D_0\ell\left(B_{(t)} - B_0\right) + \tilde{D}_0^+\Phi_n\\
&+ \mathcal{O}\left\{\tilde{\Delta}_n^{B_{(t)}} + \left(\delta_{(t)}+\delta_n\right)\left(r_{d_0h_{(t)}} + \delta_{(t)}\right)/h_{(t)}\right\},
\end{aligned}$$

where $c_n^{(t)} = \tilde{\Delta}_n^{B_{(t)}}/h_{(t)}^2 + \delta_{(t)} + \delta_n$. Since $\delta_{qh_{(t)}}/h_{(t)} = o(1)$, we have

$$\mathcal{M}(\Gamma^{(t+1)}) = B\Lambda_n^{(t)} + \mathcal{O}\left\{\delta_n + \delta_{(t)} + \tilde{\Delta}_n^{B_{(t)}}\right\},$$

where $\Lambda_n^{(t)} = I_q + \mathcal{O}\left(c_n^{(t)}\right)$ and $\mathcal{M}(\cdot)$ is defined in section 2. Note that

$$\begin{aligned}\tilde{\Lambda}_n^{(t+1)} &= \left\{\mathcal{M}\left(\Gamma^{(t+1)}\right)\right\}^\top \mathcal{M}\left(\Gamma^{(t+1)}\right) \\ &= \left(\Lambda_n^{(t)}\right)^2 + \mathcal{O}\left\{\delta_n + \delta_{(t)} + \tilde{\Delta}_n^{B_{(t)}}\right\}.\end{aligned}$$

If $c_n^{(t)} = o(1)$ almost surely, then by step III in section 2,

$$B_{(t+1)} = \mathcal{M}\left(\Gamma^{(t+1)}\right)\left\{\tilde{\Lambda}_n^{(t+1)}\right\}^{-1} = B + \mathcal{O}\left\{\delta_n + \delta_{(t)} + \tilde{\Delta}_n^{B_{(t)}}\right\}.$$

It follows that

$$\begin{aligned}\ell\left(B_{(t+1)}\right) =& \ell\left(B\right) + \tilde{D}^+ D\ell\left(B_{(t)} - B\right) + \tilde{D}^+\Phi_n \\ &+ \mathcal{O}\left\{\tilde{c}_n^{(t)}\left(\delta_{(t)} + \delta_n\right) + \tilde{\Delta}_n^{B_{(t)}}\right\},\end{aligned}$$

where $\tilde{c}_n^{(t)} = c_n^{(t)} + \left(r_{qh_{(t)}} + \delta_{(t)}\right)/h_{(t)}$. Thus for node s,

$$\begin{aligned}\ell\left(B_{(t+1)}^{(s)} - B\right) =& \tilde{D}_s^+ D_s \ell\left(B_{(t)} - B\right) + \tilde{D}_s^+\Phi_{n_s} \\ &+ \mathcal{O}\left\{\tilde{c}_{n_s}^{(t)}\left(\delta_{(t)} + \delta_{n_s}\right) + \tilde{\Delta}_{n_s}^{B_{(t)}}\right\}, s = 1, 2.\end{aligned}$$

From algorithm 1, we have $B_{(t+1)} = (B_{(t+1)}^{(1)}, B_{(t+1)}^{(2)})W = B_{(t+1)}^{(1)}W_1 + B_{(t+1)}^{(2)}W_2$, where $W = (W_1, W_2)$ is the given weight matrix following the form (A.1). Note that $\ell(B_{(t+1)}^{(s)}W_s) = W_s^T \otimes I_p \ell(B_{(t+1)}^{(s)})$ for $s = 1, 2$. Denote $\tilde{D}_s^+\Phi_{n_s} + \mathcal{O}\left\{\tilde{c}_{n_s}^{(t)}\left(\delta_{(t)} + \delta_{n_s}\right) + \tilde{\Delta}_{n_s}^{B_{(t)}}\right\} = e_{(t)}^{(s)}$. H. Wang and Xia, 2008 has showed that $|e_{(t)}^{(s)}| = o(1)$ for $s = 1, 2$. By the fact that $(W_1 + W_2)^T \otimes I_p = I_{pq}$ and $|\tilde{D}_s^+ D_s| = |\tilde{B}\tilde{B}^T|/2$, where $\tilde{B}$ is the orthogonal complement of B, i.e $(B, \tilde{(B)}) = I_p$ (H. Wang & Xia, 2008), we have

$$\begin{aligned}\ell(B_{(t+1)}) &= W_1^T \otimes I_p \ell(B_{(t+1)}^{(1)}) + W_2^T \otimes I_p \ell(B_{(t+1)}^{(2)}) \\ &= \ell(B) + \left(W_1^T \otimes I_p \tilde{D}_1^+ D_1 + W_2^T \otimes I_p \tilde{D}_2^+ D_2\right)\ell(B_{(t)} - B) \\ &+ W_1^T \otimes I_p e_{(t)}^{(1)} + W_2^T \otimes I_p e_{(t)}^{(2)}.\end{aligned}$$

Denote $\Theta = W_1^T \otimes I_p \tilde{D}_1^+ D_1 + W_2^T \otimes I_p \tilde{D}_2^+ D_2$ and $e_{(t+1)} = W_1^T \otimes I_p e_{(t)}^{(1)} + W_2^T \otimes I_p e_{(t)}^{(2)}$. To prove the convergence, we need to show (i) $|\Theta| < 1$; and (ii) $|e_{(t+1)}| = o(1)$.

(i)

$$\begin{aligned}|\Theta| &\leq |W_1^T \otimes I_p \tilde{D}_1^+ D_1| + |W_2^T \otimes I_p \tilde{D}_2^+ D_2| \\ &\leq |W_1^T \otimes I_p||\tilde{D}_1^+ D_1| + |W_2^T \otimes I_p||\tilde{D}_2^+ D_2| \\ &= w_1|\tilde{D}_1^+ D_1| + w_2|\tilde{D}_2^+ D_2| = |\tilde{B}\tilde{B}^T|/2 < 1.\end{aligned}$$

(ii)

$$\begin{aligned}|e_{(t+1)}| &\leq |W_1^T \otimes I_p||e_{(t)}^{(1)}| + |W_2^T \otimes I_p||e_{(t)}^{(2)}| \\ &= w_1|e_{(t)}^{(1)}| + w_2|e_{(t)}^{(2)}| = o(1).\end{aligned}$$

Thus the convergence of the algorithm is proved. Next we will show the consistency and efficiency.

Lemma A.1.5. *Suppose a set of random variables X_n are bounded, i.e. there exists an $b \in \mathbb{R}^+$, s.t. $|X_n| \leq b$. If $X_n = o_p(1)$, then*

$$E(|X_n|) = o(1). \tag{A.4}$$

Proof. Recall that $X_n = o_p(1)$ means that $\forall \epsilon > 0$, $\lim_{n\to\infty} P(|X_n| \geq \epsilon) = 0$.

Follow the definition, $\forall \epsilon > 0$, let $f_n(x)$ be the density function of X_n, we have

$$\begin{aligned}E(|X_n|) &= \int_{\mathcal{X}} |x_n| f_n(x) dx \\ &= \int_{|X_n|\geq\epsilon_0} |X_n|\, f_n(x) dx + \int_{|X_n|<\epsilon} |X_n|\, f_n(x) dx \\ &< b \int_{|X_n|\geq\epsilon} f_n(x) dx + \epsilon \int_{|X_n|<\epsilon} f_n(x) dx. \end{aligned} \tag{A.5}$$

Notice that as $n \to \infty$,

$$\int_{|X_n|\geq\epsilon} f_n(x) dx = P(|X_n| \geq \epsilon) \to 0,$$

and

$$\int_{|X_n|<\epsilon} f_n(x) dx = P(|X_n| < \epsilon) \leq 1,$$

thus,

$$\lim_{n\to\infty} E(|X_n|) < \epsilon.$$

Thus, $E(|X_n|) = o(1)$.

□

Lemma A.1.6. *Suppose* $\{X_n\}$ *is a sequence of random variables, as* $n \to \infty$, $E(X_n) \to 0$ *and* $\text{var}(X_n) \to 0$, *then* $\lim_{n\to\infty} P(X_n > 0) = 0$, *i.e. for all* $\epsilon, \delta > 0$, *there exists* $N > 0$, *when* $n \geq N$,

$$P(X_n > \delta) < \epsilon. \tag{A.6}$$

Proof. Let $\mu_n = E(X_n), \sigma_n = \text{var}(X_n)$. For all $\delta > 0$, notice that $|X_n - \mu_n| \geq |X_n| - |\mu_n|$, so

$$P(|X_n - \mu_n| > \frac{\delta}{2}) \geq P(|X_n| > \frac{\delta}{2} + |\mu_n|). \tag{A.7}$$

Since $\mu_n \to 0, n \to \infty$, then there exists $N_0 > 0$, for $n \geq N_0$, $|\mu_n| < \frac{\delta}{2}$. Thus

$$P(|X_n| > \frac{\delta}{2} + |\mu_n|) \geq P(|X_n| > \delta). \tag{A.8}$$

In addition, $\sigma_n \to 0$ as $n \to \infty$, then for all $\epsilon > 0$, and for the previous $\delta > 0$, there exists $N_1 > 0$, for $n \geq N1$, $\sigma_n < \epsilon \times \frac{\delta}{2}$, let $k = \frac{1}{\sqrt{\epsilon}}$, we have $k\sigma_n < \frac{\delta}{2}$. Thus

$$P(|X_n - \mu_n| > \frac{\delta}{2}) \leq P(|X_n - \mu_n| > k\sigma_n). \tag{A.9}$$

With (A.7), (A.8) and (A.9), together with Chebyshev's inequality, we have, for all $\epsilon, \delta > 0$, there exists $N = \max N_0, N_1$, when $n \geq N$,

$$P(X_n > \delta) \leq P(|X_n| > \delta) < P(|X_n - \mu_n| > \frac{1}{\sqrt{\epsilon}}\sigma_n) \leq \epsilon.$$

□

A.2 Proof of Theorem 2.4.1

Proof. Suppose the process is within two nodes. From Theorem 2, as $n_i \to \infty$, we have $\hat{d} \to q$. Thus we consider

$$m^2(\hat{B}, B) = \text{Tr}\{\hat{B}^T(I - P_B)\hat{B}\},$$

where $P_B = B(B^TB)^{-1}B^T$. Under the constrain $B^TB = I$, $P_B = BB^T$. Denote $P = I - P_B$, notice P is also a symmetric projection matrix and is fixed.

Denote the estimates obtained in a single node is $\hat{B}_k, k = 1, 2$. Then $\hat{B} = (\hat{B}_1, \hat{B}_2)W$, where $W^T = (W_1^T, W_2^T)$ is the $2p \times p$ coefficient matrix, i.e. $\hat{B} = \hat{B}_1W_1 + \hat{B}_2W_2$. Without lose of generality, we assume that $\hat{B}$ is

orthogonal. In fact, if $\hat{B}^T\hat{B} \neq I$, we further make the QR-decomposition $\hat{B} = Q_B R_B$ and use Q_B as our estimation. Then we can see that $Q_B = (\hat{B}_1, \hat{B}_2)WR^{-1}$.

For simplify, we denote $m = m(\hat{B}, B)$, $m_1 = m(\hat{B}_1, B)$ and $m_2 = m(\hat{B}_2, B)$. Thus

$$\begin{aligned} m^2 =& \operatorname{Tr}\{(\hat{B}_1W_1 + \hat{B}_2W_2)^T P(\hat{B}_1W_1 + \hat{B}_2W_2)\} \\ =& \operatorname{Tr}(W_1^T\hat{B}_1^T\hat{B}_1W_1) + \operatorname{Tr}(W_2^T\hat{B}_2^T\hat{B}_2W_2) + 2\operatorname{Tr}(W_1^T\hat{B}_1^TP\hat{B}_2W_2) \\ \leq& m_1^2\operatorname{Tr}(W_1W_1^T) + m_2^2\operatorname{Tr}(W_2W_2^T) \\ &+ 2\operatorname{Tr}(\hat{B}_1^TP\hat{B}_2)\operatorname{Tr}(W_2W_1^T) \end{aligned} \tag{A.10}$$

Denote $\omega_1 = \operatorname{Tr}(W_1W_1^T)$, $\omega_2 = \operatorname{Tr}(W_2W_2^T)$ and $\omega_{12} = \operatorname{Tr}(W_2W_1^T)$, then from (A.10), we have

$$E(m^2) \leq \omega_1 E(m_1^2) + \omega_2 E(m_2^2) + 2\omega_{12} E\left(\operatorname{Tr}(\hat{B}_1^TP\hat{B}_2)\right). \tag{A.11}$$

Now we look at the term $E\left(\operatorname{Tr}(\hat{B}_1^TP\hat{B}_2)\right)$ in (A.11). Let $B_k^* = P\hat{B}_k, k = 1, 2$, since P is a symmetric projection matrix, $\hat{B}_1^TP\hat{B}_2 = B_1^{*T}B_2^*$. Then we rewrite the term as $E\left(\operatorname{Tr}(B_1^{*T}B_2^*)\right) = \operatorname{Tr}\left(E(B_1^{*T}B_2^*)\right)$.

Because $(B_1^* - B_2^*)^T(B_1^* - B_2^*)$ is semi-positive definite. In another word, we have

$$\operatorname{Tr}\{E\left((B_1^* - B_2^*)^T(B_1^* - B_2^*)\right)\} \geq 0,$$

then

$$\begin{aligned} 2\operatorname{Tr}\left(E(B_1^{*T}B_2^*)\right) =& \operatorname{Tr}\left(E(B_1^{*T}B_2^*)\right) + \operatorname{Tr}\left(E(B_2^{*T}B_1^*)\right) \\ \leq& \operatorname{Tr}\left(E(B_1^{*T}B_1^*)\right) + \operatorname{Tr}\left(E(B_2^{*T}B_2^*)\right) \\ =& E(m_1^2) + E(m_2^2). \end{aligned} \tag{A.12}$$

Plug (A.12) back into (A.11), we have

$$E(m^2) \leq (\omega_1 + \omega_{12})E(m_1^2) + (\omega_2 + \omega_{12})E(m_2^2). \tag{A.13}$$

In particular, consider $d = 1$. Let $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ be the estimate of $\boldsymbol{\beta}$ in node 1 and node 2 respectively. Denote

$$\tilde{\boldsymbol{\beta}} = w_1\hat{\boldsymbol{\beta}}_1 + w_2\hat{\boldsymbol{\beta}}_2,$$

where w_1 and w_2 are some coefficients. Then the estimate of $\boldsymbol{\beta}$ in our method is $\hat{\boldsymbol{\beta}}_d = \frac{\tilde{\boldsymbol{\beta}}}{||\tilde{\boldsymbol{\beta}}||}$. By definition,

$$m^2 = \boldsymbol{\beta}^T(I - \hat{\boldsymbol{\beta}}_d\hat{\boldsymbol{\beta}}_d^T)\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_d^T(I - \boldsymbol{\beta}\boldsymbol{\beta}^T)\hat{\boldsymbol{\beta}}_d = \frac{1}{||\tilde{\boldsymbol{\beta}}||^2}\tilde{\boldsymbol{\beta}}^T(I - \boldsymbol{\beta}\boldsymbol{\beta}^T)\tilde{\boldsymbol{\beta}}.$$

Denote $A = I - \boldsymbol{\beta}\boldsymbol{\beta}^T$, notice that A is a projection matrix, then

$$\begin{aligned} E(m^2) &= \frac{E\left((w_1\hat{\boldsymbol{\beta}}_1 + w_2\hat{\boldsymbol{\beta}}_2)^T A(w_1\hat{\boldsymbol{\beta}}_1 + w_2\hat{\boldsymbol{\beta}}_2)\right)}{||\tilde{\boldsymbol{\beta}}||^2} \\ &= \frac{w_1^2 E(\hat{\boldsymbol{\beta}}_1^T A\hat{\boldsymbol{\beta}}_1) + w_2^2 E(\hat{\boldsymbol{\beta}}_2^T A\hat{\boldsymbol{\beta}}_2) + 2w_1w_2E(\hat{\boldsymbol{\beta}}_1^T A\hat{\boldsymbol{\beta}}_2)}{||\tilde{\boldsymbol{\beta}}||^2}. \end{aligned} \quad \text{(A.14)}$$

Let $A\hat{\boldsymbol{\beta}}_i = (\beta_{i1}^*, \ldots, \beta_{ip}^*), i = 1, 2$. Then

$$E(\hat{\boldsymbol{\beta}}_i^T A\hat{\boldsymbol{\beta}}_i) = \sum_{j=1}^{p} E\left((\beta_{ij}^*)^2\right), i = 1, 2,$$

and

$$E(\hat{\boldsymbol{\beta}}_1^T A\hat{\boldsymbol{\beta}}_2) = \sum_{j=1}^{p} E(\beta_{1j}^*\beta_{2j}^*).$$

By Cauchy–Schwarz inequality,

$$E(\beta_{1j}^*\beta_{2j}^*) \leq \sqrt{E\left((\beta_{1j}^*)^2\right)E\left((\beta_{2j}^*)^2\right)} \leq \frac{E\left((\beta_{1j}^*)^2\right) + E\left((\beta_{2j}^*)^2\right)}{2}.$$

Thus (A.14) can be rewritten as

$$
\begin{aligned}
& E(m^2) && \text{(A.15)} \\
=& \frac{1}{||\tilde{\boldsymbol{\beta}}||^2} \sum_{j=1}^{p} \left\{ w_1^2 E\left((\beta_{1j}^*)^2\right) + w_2^2 E\left((\beta_{2j}^*)^2\right) + 2w_1 w_2 E(\beta_{1j}^* \beta_{2j}^*) \right\} \\
\leq& \frac{1}{||\tilde{\boldsymbol{\beta}}||^2} \sum_{j=1}^{p} \left\{ w_1^2 E\left((\beta_{1j}^*)^2\right) + w_2^2 E\left((\beta_{2j}^*)^2\right) \right\} \\
& + \frac{1}{||\tilde{\boldsymbol{\beta}}||^2} \sum_{j=1}^{p} \left\{ w_1 w_2 \left[E\left((\beta_{1j}^*)^2\right) + E\left((\beta_{2j}^*)^2\right) \right] \right\} \\
=& \frac{1}{||\tilde{\boldsymbol{\beta}}||^2} \sum_{j=1}^{p} \left\{ w_1 (w_1 + w_2) E\left((\beta_{1j}^*)^2\right) \right\} \\
& + \frac{1}{||\tilde{\boldsymbol{\beta}}||^2} \sum_{j=1}^{p} \left\{ + w_2 (w_1 + w_2) E\left((\beta_{2j}^*)^2\right) \right\} \\
=& \frac{1}{||\tilde{\boldsymbol{\beta}}||^2} \left\{ w_1(w_1 + w_2) E(\hat{\boldsymbol{\beta}}_1^T A \hat{\boldsymbol{\beta}}_1) + w_2(w_1 + w_2) E(\hat{\boldsymbol{\beta}}_2^T A \hat{\boldsymbol{\beta}}_2) \right\} \\
=& \frac{1}{||\tilde{\boldsymbol{\beta}}||^2} \left\{ w_1(w_1 + w_2) E(m_1^2) + w_2(w_1 + w_2) E(m_2^2) \right\}. && \text{(A.16)}
\end{aligned}
$$

Notice that $||\tilde{\boldsymbol{\beta}}||^2 = w_1^2||\hat{\boldsymbol{\beta}}_1||^2 + w_2^2||\hat{\boldsymbol{\beta}}_2||^2 + 2w_1w_2\hat{\boldsymbol{\beta}}_1^T\hat{\boldsymbol{\beta}}_2$, and $\hat{\boldsymbol{\beta}}_1^T\hat{\boldsymbol{\beta}}_2 \leq ||\hat{\boldsymbol{\beta}}_1||^2||\hat{\boldsymbol{\beta}}_2||^2$. By the constraint of $\hat{\boldsymbol{\beta}}$, we know that $||\hat{\boldsymbol{\beta}}_1|| = ||\hat{\boldsymbol{\beta}}_2|| = 1$, and without loss of generality, we can assume that $\hat{\boldsymbol{\beta}}_1^T\hat{\boldsymbol{\beta}}_2 \geq 0$, thus $||\tilde{\boldsymbol{\beta}}||^2 \geq w_1^2 + w_2^2$. Together with (A.15), we can get

$$E(m^2) \leq uE(m_1^2) + vE(m_2^2), \tag{A.17}$$

where $u = \frac{w_1(w_1+w_2)}{w_1^2+w_2^2}$ and $v = \frac{w_2(w_1+w_2)}{w_1^2+w_2^2}$ only depend on w_1 and w_2 we set.

Since $m_1^2 = o_p(1)$ and $m_2^2 = o_p(1)$. From (A.13) and $lemma A.1.5$

$$E(m^2) = o(1),$$

Thus we have $E(m) = o(1)$ and var$(m) = o(1)$.

Let $Y_{n_1,n_2} = m - m_k$, for $k = 1, 2$. Easily we can prove $E(Y_{n_1,n_2}) \to 0$ and var$(Y_{n_1,n_2}) \to 0$, as $n_1, n_2 \to \infty$. By Lemma A.1.6, we have

$$\lim_{n_1,n_2\to\infty} P(Y_{n_1,n_2} > 0) = 0.$$

□

Lemma A.2.1. *Suppose* $\{X_n\}$ *and* $\{Y_n\}$ *are two sets of non-negative random variables, and* $\{a_n\}$ *is a series converges to 0, i.e.* $\lim_{n\to\infty} a_n = 0$*. If we have* $X_n = O_p(a_n)$ *and* $\lim_{n\to\infty} P(Y_n > X_n) = 0$*, then*

$$Y_n = O_p(a_n).$$

Proof. By the definition of "O_p", we have that for all $\epsilon > 0$, there exists an $N^* > 0$ and $M > 0$, such that when $n \geq N^*$, $P(X_n > Ma_n) < \epsilon/2$.

Since $\lim_{n\to\infty} P(Y_n > X_n) = 0$, then for this ϵ, there exists an $N^{**} > 0$, such that when $n \geq N^{**}$, we have $P(Y_n > X_n) < \epsilon/2$.

Let $N = \max\{N^*, N^{**}\}$, then we have

$$\begin{aligned} P(Y_n > Ma_n) &= P(Y_n > X_n > Ma_n) + P(X_n \geq Y_n > Ma_n) \\ &\leq P(Y_n > X_n) + P(X_n > Ma_n) \\ &< \epsilon. \end{aligned} \tag{A.18}$$

From (A.18), we then have $Y_n = O_p(a_n)$.

□

A.3 Proof of Theorem 2.4.3

Proof. Suppose the bandwidth $h_s \sim n_s^{-1/(p+4)}$ for $s = 1, 2$, and the weight matrix $W = (\frac{n_1}{n_1+n_2}, \frac{n_2}{n_1+n_2})^T \otimes I_p$. By (A.13), we have $E(m^2) \leq (\omega_1 + \omega_{12})E(m_1^2) + (\omega_2 + \omega_{12})E(m_2^2)$ with $\omega_s = \mathrm{Tr}(W_s W_s^T) = pw_s^2, s = 1, 2$ and $\omega_{12} = \mathrm{Tr}(W_2 W_1^T) = pw_1 w_2$, where $w_1 = \frac{n_1}{n_1+n_2}$ and $w_2 = \frac{n_2}{n_1+n_2}$. By the fact that $E(m_s) = 0$ and $m_s = O_p\left(n_s^{-\frac{3}{p+4}} \log n_s\right)$ for $s = 1, 2$ (Xia et al., 2007; Xia et al., 2002), we have that $E(m_s^2) = \mathrm{var}(m_s) = O\left(n_s^{-\frac{6}{p+4}} (\log n_s)^2\right)$ for $s = 1, 2$. Let $n = n_1 + n_2$, we have

$$\begin{aligned} w_s E(m_s^2) &= O\left(\tfrac{n_s}{n} n_s^{-\frac{6}{p+4}} (\log n_s)^2\right) \\ &= O\left(n^{-\frac{6}{p+4}} (\log n)^2 \left[\left(\tfrac{n_s}{n}\right)^{1-\frac{6}{p+4}} \left(\tfrac{\log n_s}{\log n}\right)^2\right]\right). \end{aligned}$$

Notice that when $p \geq 2$, the term $\left(\frac{n_s}{n}\right)^{1-\frac{6}{p+4}} \left(\frac{\log n_s}{\log n}\right)^2 < 1$, thus $w_s E(m_s^2) < O(n^{-\frac{6}{p+4}} (\log n)^2)$. In the end, we have $E(m^2) \leq p\left(w_1 E(m_1^2) + w_2 E(m_2^2)\right) =$

$O(n^{-\frac{6}{p+4}}(\log n)^2)$, i.e. $m = O_p(n^{-\frac{3}{p+4}} \log n)$. Then the theorem is proved.

□

Appendix B

Proof of Chapter 3

B.1 Proof of Theorem 3.4.1

First, we presents some definitions and lemmas to facilitate the proof of Theorem 1.

Definition B.1.1. *Let M be a compact metric space. Given a set $S \subseteq M$, the ϵ-covering number of S, denoted $\mathcal{N}_\epsilon(S)$, is the minimum k such that there exists k closed balls $B_1, \ldots, B_k$ of diameter ϵ, and $S \subseteq \cup_{1 \leq i \leq k} B_i$. The ϵ-dimension of S is the quantity*

$$d_\epsilon(S) := \frac{\log \mathcal{N}_\epsilon(S)}{-\log \epsilon}.$$

In the following theoretical discussions, it is convenient to work with measures instead of sets. The following definition (Dudley, 1969; Weed & Bach, 2019) extends the ϵ-covering number to the language of entropy (Posner et al., 1967).

Definition B.1.2. *Given a measure μ on M, the (ϵ, τ)-covering number is*

$$\mathcal{N}_\epsilon(\mu, \tau) := \inf \left\{ \mathcal{N}_\epsilon(S) : \mu(S) \geq 1 - \tau \right\},$$

and the (ϵ, τ)-dimension is

$$d_\epsilon(\mu, \tau) := \frac{\log \mathcal{N}_\epsilon(\mu, \tau)}{-\log \epsilon}.$$

Note that the (ϵ, τ)-covering number and (ϵ, τ)-dimension are monotonic decreasing as τ increases. This implies that we can define the following upper and lower limits of the (ϵ, τ)-dimension.

Definition B.1.3. *The upper and lower Wasserstein dimensions are defined respectively*

$$d_p^*(\mu) := \inf\left\{ s \in (2p, \infty) : \limsup_{\epsilon \to 0} d_\epsilon(\mu, \epsilon^{\frac{sp}{s-2p}}) \leq s \right\},$$

$$d_*(\mu) := \lim_{\tau \to 0} \liminf_{\epsilon \to 0} d_\epsilon(\mu, \tau).$$

Lemma B.1.4 (c.f. Theorem 1 in Weed and Bach, 2019)**.** *Let μ_n be the empirical measure of μ summarized from a random sample of size n, and $p \in [1, \infty)$. If $s > d_p^*(\mu)$ and $t < d_*(\mu)$, then*

$$E[W_p(\mu, \mu_n)] \lesssim n^{-1/s} \quad \textit{and} \quad W_p(\mu, \mu_n) \gtrsim n^{-1/t}.$$

Proof of Theorem 3.4.1. First, we extend the lower bound result in Lemma B.1.4 to the case of two empirical measures $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$ that are supported on n observations drawn from α and β, respectively. Suppose that there exists a triple of positive constants e, τ and t such that

$$\mathcal{N}_\epsilon(\boldsymbol{a}_n, \tau) \geq \epsilon^{-t} \tag{B.1}$$

for all $e \leq \epsilon$.

Given e and n, if we choose a small enough t such that $\epsilon = n^{-1/t}/2 \geq e$. Let $S = \cup_{\boldsymbol{b} \in \text{supp}(\boldsymbol{b}_n)} B(\boldsymbol{b}, \epsilon/2)$. Then, we have

$$\mathcal{N}_\epsilon(\boldsymbol{a}_n, \tau) \geq \epsilon^{-t} > n,$$

and hence $\boldsymbol{a}_n(S) < 1 - \tau$. This is equivalent to show that, for any $\boldsymbol{a} \sim \boldsymbol{a}_n$, the probability of the event $\mathcal{E} = \{\|\boldsymbol{a}, \text{supp}(\boldsymbol{b}_n)\| \geq \epsilon/2\}$ is at least τ.

Then, if $(\boldsymbol{a}_i, \boldsymbol{b}_i)$ are i.i.d. observations of the coupling of $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$, we have the following inequality holds

$$\begin{aligned} W_p^p(\boldsymbol{a}_n, \boldsymbol{b}_n) &= E\left[\|\boldsymbol{a}_i, \boldsymbol{b}_i\|^p\right] \geq E\left[\|\boldsymbol{a}_i, \text{supp}(\boldsymbol{b}_n)\|^p\right] \\ &\geq \tau(\epsilon/2)^p = \tau 4^{-p} n^{-p/t}. \end{aligned}$$

This immediately yields the following lower bound for the empirical Wasserstein 2 distance

$$W_2(\boldsymbol{a}_n, \boldsymbol{b}_n) \geq \tau 4^{-1} n^{-1/t}. \tag{B.2}$$

By the definition of $d_*(\mu)$, we can show that the condition in (B.1) can be satisfied if we choose $t < d_*(\mu)$.

Next, we show under Assumption 1 (b), for any $p \in [1, d/2]$, $d \leq d_*(\alpha)$. When Assumption 1 (b) holds, for all $\tau > 0$, there exists a $\sigma > 0$ such that any set T for which $\alpha(T) \geq 1 - \tau$ satisfies $\mathcal{H}^d(T) \geq \sigma$.

Then, for any covering$\{B(\boldsymbol{x}_i, \epsilon)\}$ of T by balls of radius ϵ, we must have $\sum_i \epsilon^d \geq \sigma$.

Therefore, such a covering contains at least $\sigma\epsilon^{-d}$ balls. Then, we have

$$\frac{\log \mathcal{N}_\epsilon(\alpha, \tau)}{-\log \epsilon} \geq d + \frac{\log \sigma}{-\log \epsilon}. \tag{B.3}$$

By taking limits to both sides of (B.3), we arrive at $d_*(\alpha) \geq d$.

Combing the results in (B.2) and (B.3), we finish the proof of Lemma 1. □

B.2 Proof of Theorem 3.4.2

First, we present a useful lemma to bound the asymptotic variance of the 2-Wasserstein distance induced by a Smoothed Monge Map. Then, the proof of Theorem 2 should immediately follow the results of this lemma.

Lemma B.2.1. *Let α and β be two densities with finite $4 + \delta$ moments for some $\delta > 0$. Let $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_j\}_{j=1}^m$ be two i.i.d. random sample drawn from α and β, respectively. Further, we use $\boldsymbol{a}_n$ and $\boldsymbol{b}_m$ to denote the empirical measures of $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_j\}_{j=1}^m$. Then, we have*

$$\text{var}\left(\widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_m)\right) \leq \frac{C(\alpha, \beta)}{n} + \frac{C(\beta, \alpha)}{m}, \tag{B.4}$$

where

$$\begin{aligned} C(\alpha, \beta) = {} & 8\left(E(\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|^2 \|\boldsymbol{a}_1\|)\right) \\ & + 8\left((E\|\boldsymbol{a}_1 - \boldsymbol{a}_2\|^4)^{1/2}\Big(\int_{\mathbb{R}^d} \|\boldsymbol{b}\|^4 d\beta(\boldsymbol{b})\Big)^{1/2}\right), \end{aligned}$$

and $C(\beta, \alpha)$ is defined in a symmetric way.

Proof of Lemma B.2.1. According to the definition of the Smoothed Monge Map, $\widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_m)$ is invariant with respect to any permutations within $\{\boldsymbol{a}_i\}_{i=1}^n$ or any permutations within $\{\boldsymbol{b}_j\}_{j=1}^m$.

We draw $\boldsymbol{a}_1'$ and $\boldsymbol{b}_1'$ from α and β as two copies of $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ which are independent of $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_j\}_{j=1}^m$. Then, we use $\boldsymbol{a}_n'$ and $\boldsymbol{b}_m'$ to denote the empirical measures of $\{\boldsymbol{a}_1', \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$ and $\{\boldsymbol{b}_1', \boldsymbol{b}_2, \ldots, \boldsymbol{b}_m\}$, where we replace $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ with $\boldsymbol{a}_1'$ and $\boldsymbol{b}_1'$. Further, we introduce the following notations

for the (squared) 2-Wasserstein distances induced by Smoothed Monge Maps.

$$Z_1 = \widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_m), \quad Z_2 = \widetilde{W}_2^2(\boldsymbol{a}'_n, \boldsymbol{b}_m), \quad \text{and} \quad Z_3 = \widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}'_m).$$

Follow the Efron-Stein inequality, one can see that

$$\text{var}\left(\widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_m)\right) \leq \frac{n}{4} E(Z_1 - Z_2)^2 + \frac{m}{4} E(Z_1 - Z_3)^2. \tag{B.5}$$

First, we bound $E(Z_1 - Z_2)^2$. Denote ψ and ψ' the Smoothed Monge Maps from $\boldsymbol{a}_n$ to $\boldsymbol{b}_m$ and from $\boldsymbol{a}'_n$ to $\boldsymbol{b}_m$, respectively. We define $\pi_{i,j}$ as the probability that ψ assigns to the pair $(\boldsymbol{a}_i, \boldsymbol{b}_j)$, and $c_{i,j} = \|\boldsymbol{a}_i - \boldsymbol{b}_j\|^2$. Also, we define $\pi'_{i,j}$ and $c'_{i,j}$ in a similar fashion for samples $\{\boldsymbol{a}'_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$ and $\{\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_m\}$. With the notations above, we have

$$Z_2 = \sum_{i=1}^{n}\sum_{j=1}^{m} c'_{i,j}\pi'_{i,j} \quad \text{and} \quad Z_1 \leq \sum_{i=1}^{n}\sum_{j=1}^{m} c_{i,j}\pi'_{i,j}.$$

Since the Smoothed Monge Map does not force a one-to-one map, replacing $\boldsymbol{a}_1$ with $\boldsymbol{a}'_1$ will not affect the cost functions for $i \geq 2$. In other words, we have $c_{i,j} = c'_{i,j}$ for $i \geq 2$. Then we have

$$Z_1 - Z_2 \leq \sum_{j=1}^{m} \pi'_{1,j}(c_{1,j} - c'_{1,j}) \leq \|\boldsymbol{a}_1 - \boldsymbol{a}'_1\| \tag{B.6}$$

$$\times \sum_{j=1}^{m} \pi'_{1,j} \left(\|\boldsymbol{a}_1\| + \|\boldsymbol{a}'_1\| + 2\|\boldsymbol{b}_j\|\right). \tag{B.7}$$

As we apply equal weights to observations, we have $\sum_{j=1}^{m} \pi'_{1,j} = \frac{1}{n}$. Thus, we can simplify (B.7) by

$$Z_1 - Z_2 \leq \|\boldsymbol{a}_1 - \boldsymbol{a}'_1\| \left\{ \frac{1}{n}\left(\|\boldsymbol{a}_1\| + \|\boldsymbol{a}'_1\|\right) + 2\sum_{j=1}^{m} \|\boldsymbol{b}_j\| \right\}.$$

Then, follow Theorem 3.1 in Del Barrio and Loubes, 2019, we have

$$\frac{n}{4} E(Z_1 - Z_2)^2 \leq \frac{C(\alpha, \beta)}{n}. \tag{B.8}$$

To bound the second term on the right hand side of (B.5), we utilize the following exchangeability

$$E\left(\sum_{j=1}^{m}\pi'_{1,j}\|\boldsymbol{b}_j\|^4\right) = \frac{1}{n}E\left(\sum_{i=1}^{n}\sum_{j=1}^{m}\pi'_{i,j}\|\boldsymbol{b}_j\|^4\right)$$
$$= \frac{1}{n}E\left(\frac{1}{m}\sum_{j=1}^{m}\pi'_{i,j}\|\boldsymbol{b}_j\|^4\right) = \frac{1}{n}E\left(\|\boldsymbol{b}_j\|^4\right).$$

Again, follow Theorem 3.1 in Del Barrio and Loubes, 2019, we have

$$\frac{m}{4}E(Z_1 - Z_3)^2 \leq \frac{C(\beta,\alpha)}{m}. \tag{B.9}$$

By plugging (B.8) and (B.9) back to (B.5), we finish the proof of Lemma B.2.1. □

Proof of Theorem 3.4.2. Suppose that we have two densities α and β that satisfy Assumption 1. Let $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_j\}_{j=1}^n$ be two i.i.d. random samples drawn from α and β, respectively. Then, according to Lemma B.2.1, the variance of the (squared) 2-Wasserstein distance induced by the Smoothed Monge Map between $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_j\}_{j=1}^n$ is upper bounded by

$$\text{var}\left(\widetilde{W}_2^2(\boldsymbol{a}_n,\boldsymbol{b}_n)\right) \leq \frac{C(\alpha,\beta)+C(\alpha,\beta)}{n}.$$

Given $n\text{var}\left(\widetilde{W}_2^2(\boldsymbol{a}_n,\boldsymbol{b}_n)\right)$ is bounded, the central limit theorem naturally yields the following statement

$$\sqrt{n}\left(\widetilde{W}_2^2(\boldsymbol{a}_n,\boldsymbol{b}_n) - E[\widetilde{W}_2^2(\boldsymbol{a}_n,\boldsymbol{b}_n)]\right) \to N(0,\sigma_W^2), \tag{B.10}$$

where σ_W^2 is the limit of $n\text{var}\left(\widetilde{W}_2^2(\boldsymbol{a}_n,\boldsymbol{b}_n)\right)$ as $n \to \infty$.

Let $\phi(\cdot)$ and $\psi(\cdot)$ be the the optimal transport maps from α to β and from β to α, respectively. We define the following two variances

$$\begin{aligned}
\sigma^2(\alpha,\beta) &= \int_{\mathbb{R}^d} \big(\|a\|^2 - 2\phi^*(a)\big)^2 \, d\alpha(a) \\
&\quad - \left(\int_{\mathbb{R}^d} \big(\|a\|^2 - 2\phi^*(a)\big) d\alpha(a)\right)^2, \\
\text{and} \quad \sigma^2(\beta,\alpha) &= \int_{\mathbb{R}^d} \big(\|b\|^2 - 2\psi^*(b)\big)^2 \, d\beta(b) \\
&\quad - \left(\int_{\mathbb{R}^d} \big(\|b\|^2 - 2\psi^*(b)\big) d\beta(b)\right)^2.
\end{aligned}$$

Next, we would like to show that the asymptotic variance satisfies

$$\sigma_W^2 = \frac{\sigma^2(\alpha,\beta) + \sigma^2(\beta,\alpha)}{2}. \tag{B.11}$$

Similar as the notations used in the proof of Lemma B.2.1, we draw $\boldsymbol{a}_1'$ and $\boldsymbol{b}_1'$ from α and β as two copies of $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ which are independent of $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{b}_j\}_{j=1}^n$. We use $\boldsymbol{a}_n'$ and $\boldsymbol{b}_n'$ to denote the empirical measures of $\{\boldsymbol{a}_1', \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n\}$ and $\{\boldsymbol{b}_1', \boldsymbol{b}_2, \ldots, \boldsymbol{b}_n\}$, where we replace $\boldsymbol{a}_1$ and $\boldsymbol{b}_1$ with $\boldsymbol{a}_1'$ and $\boldsymbol{b}_1'$. Then, we define the following three residual terms

$$\begin{aligned}
R_1 = \widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_n) &- \int_{\mathbb{R}^d} \big(\|a\|^2 - 2\phi(a)\big) \, d\boldsymbol{a}_n(a) \\
&- \int_{\mathbb{R}^d} \big(\|b\|^2 - 2\psi(b)\big) \, d\boldsymbol{b}_n(b), \\
R_2 = \widetilde{W}_2^2(\boldsymbol{a}_n', \boldsymbol{b}_n) &- \int_{\mathbb{R}^d} \big(\|a\|^2 - 2\phi(a)\big) \, d\boldsymbol{a}_n'(a) \\
&- \int_{\mathbb{R}^d} \big(\|b\|^2 - 2\psi(b)\big) \, d\boldsymbol{b}_n(b), \\
R_3 = \widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_n') &- \int_{\mathbb{R}^d} \big(\|a\|^2 - 2\phi(a)\big) \, d\boldsymbol{a}_n(a) \\
&- \int_{\mathbb{R}^d} \big(\|b\|^2 - 2\psi(b)\big) \, d\boldsymbol{b}_n'(b).
\end{aligned}$$

To prove (B.11) is equivalent to show that

$$\frac{n}{2}\text{var}(R_1) \to 0, \quad \text{as} \quad n \to \infty. \tag{B.12}$$

To prove (B.12), according to the Efron-Stein inequality, it is suffice to show that

$$n^2 E(R_1 - R_2)^2 \to 0 \quad \text{and} \quad n^2 E(R_1 - R_2)^2 \to 0 \quad \text{as} \quad n \to \infty.$$

We will prove the first claim and the second one naturally follows by the symmetry. Denote $\widetilde{\phi}$ and $\widetilde{\psi}$ the Smoothed Monge Maps from $\boldsymbol{a}_n$ to $\boldsymbol{b}_n$ and from $\boldsymbol{b}_n$ to $\boldsymbol{a}_n$, respectively. The almost sure convergence of the empirical Monge map and the consistency of the smoothing spline estimator together imply that $\widetilde{\phi} \to \phi$ and $\widetilde{\psi} \to \psi$ a.s..

Then, we can write out $\widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_n)$ as

$$\begin{aligned}\widetilde{W}_2^2(\boldsymbol{a}_n, \boldsymbol{b}_n) = &\int_{\mathbb{R}^d} \left(\|a\|^2 - 2\widetilde{\phi}(a) \right) d\boldsymbol{a}_n(a) \\ &+ \int_{\mathbb{R}^d} \left(\|b\|^2 - 2\widetilde{\psi}(b) \right) d\boldsymbol{b}_n(b). \end{aligned} \tag{B.13}$$

Also, we have

$$\begin{aligned}\widetilde{W}_2^2(\boldsymbol{a}'_n, \boldsymbol{b}_n) \geq &\int_{\mathbb{R}^d} \left(\|a\|^2 - 2\widetilde{\phi}(a) \right) d\boldsymbol{a}'_n(a) \\ &+ \int_{\mathbb{R}^d} \left(\|b\|^2 - 2\widetilde{\psi}(b) \right) d\boldsymbol{b}_n(b). \end{aligned} \tag{B.14}$$

With (B.13) and (B.14), we can upper bound the difference between R_1 and R_2 by

$$\begin{aligned} R_1 - R_2 &\leq 2 \int_{\mathbb{R}^d} \left(\phi(a) - \widetilde{\phi}(a) \right) d\boldsymbol{a}_n(a) - 2 \int_{\mathbb{R}^d} \left(\phi(a) - \widetilde{\phi}(a) \right) d\boldsymbol{a}'_n(a) \\ &= \frac{2}{n} \left\{ \left[\phi(\boldsymbol{a}_1) - \widetilde{\phi}(\boldsymbol{a}_1) \right] - \left[\phi(\boldsymbol{a}'_1) - \widetilde{\phi}(\boldsymbol{a}'_1) \right] \right\}. \end{aligned} \tag{B.15}$$

With the almost sure convergence of the Smoothed Monge Map, the results in (B.15) implies the almost sure convergence of $n(R_1 - R_2)$, i.e. $n(R_1 - R_2) \to 0$ a.s..

Besides, according to the Theorem 3.2 in Del Barrio and Loubes, 2019, we know that $n^2(R_1 - R_2)$ is uniformly integrable. Therefore, we finish the proof by concluding that $n^2 E(R_1 - R_2)^2 \to 0$ as $n \to \infty$. □

Bibliography

Altschuler, J., Weed, J., & Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration, In *Advances in neural information processing systems*.

Alvarez-Melis, D., Jaakkola, T., & Jegelka, S. (2018). Structured optimal transport, In *International conference on artificial intelligence and statistics*.

Andoni, A., & Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions, In *2006 47th annual ieee symposium on foundations of computer science (focs'06)*. IEEE.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks, In *International conference on machine learning*.

Arnett, D. K., Blumenthal, R. S., Albert, M. A., Buroker, A. B., Goldberger, Z. D., Hahn, E. J., Himmelfarb, C. D., Khera, A., Lloyd-Jones, D., McEvoy, J. W., Michos, E. D., Miedema, M. D., Munoz, D., Smith, J., S. C., Virani, S. S., Williams, S., K. A., Yeboah, J., & Ziaeian, B. (2019). 2019 acc/aha guideline on the primary prevention of cardiovascular disease: A report of the american college of cardiology/american heart association task force on clinical practice guidelines. *J Am Coll Cardiol*. https://doi.org/10.1016/j.jacc.2019.03.010

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

Bellman, R. E. (2015). *Adaptive control processes: A guided tour* (Vol. 2045). Princeton university press.

Bernheim, A. M., Kittipovanonth, M., Takahashi, P. Y., Gharacholou, S. M., Scott, C. G., & Pellikka, P. A. (2011). Does the prognostic value of dobutamine stress echocardiography differ among different age groups? *American heart journal*, *161*(4), 740–745.

Bertsekas, D. P. (1981). A new algorithm for the assignment problem. *Mathematical Programming*, *21*(1), 152–171.

Bertsekas, D. P. (1992). Auction algorithms for network flow problems: A tutorial introduction. *Computational optimization and applications*, *1*(1), 7–66.

Bertsimas, D., & Tsitsiklis, J. N. (1997). *Introduction to linear optimization* (Vol. 6). Athena Scientific Belmont, MA.

Bhushan Damodaran, B., Kellenberger, B., Flamary, R., Tuia, D., & Courty, N. (2018). Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation, In *Proceedings of the european conference on computer vision (eccv)*.

Blanchard, G., & Mücke, N. (2016). Parallelizing spectral algorithms for kernel learning. *arXiv preprint arXiv:1610.07487*.

Blessberger, H., & Binder, T. (2010a). Two dimensional speckle tracking echocardiography: Basic principles. *Heart*, *96*(9), 716–722.

Blessberger, H., & Binder, T. (2010b). Two dimensional speckle tracking echocardiography: Clinical applications. *Heart*, *96*(24), 2032–2040.

Breiman, L. (1996). Stacked regressions. *Machine Learning*, *24*(1), 49–64.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, *44*(4), 375–417.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, *25*(8), 1301–1309.

Canas, G., & Rosasco, L. (2012). Learning probability measures with respect to optimal transport metrics, In *Advances in neural information processing systems*.

Caspar, T., Samet, H., Ohana, M., Germain, P., El Ghannudi, S., Talha, S., Morel, O., & Ohlmann, P. (2017). Longitudinal 2d strain can help diagnose coronary artery disease in patients with suspected non-st-elevation acute coronary syndrome but apparent normal global and segmental systolic function. *Int J Cardiol*, *236*, 91–94. https://doi.org/10.1016/j.ijcard.2017.02.068

Cazelles, E., Seguy, V., Bigot, J., Cuturi, M., & Papadakis, N. (2018). Geodesic pca versus log-pca of histograms in the wasserstein space. *SIAM Journal on Scientific Computing*, *40*(2), B429–B456.

Chen, S., Ma, K., & Zheng, Y. (2019). Tan: Temporal affine network for real-time left ventricle anatomical structure analysis based on 2d ultrasound videos. *arXiv preprint arXiv:1904.00631*.

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

Courty, N., Flamary, R., Habrard, A., & Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation, In *Advances in neural information processing systems*.

Courty, N., Flamary, R., & Tuia, D. (2014). Domain adaptation with regularized optimal transport, In *Joint european conference on machine learning and knowledge discovery in databases*. Springer.

Courty, N., Flamary, R., Tuia, D., & Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(9), 1853–1865.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport, In *Advances in neural information processing systems*.

Cuturi, M., & Doucet, A. (2014). Fast computation of wasserstein barycenters.

Dalgard, F. J., Gieler, U., Tomas-Aragones, L., Lien, L., Poot, F., Jemec, G. B., Misery, L., Szabo, C., Linder, D., Sampogna, F., Et al. (2015). The psychological burden of skin diseases: A cross-sectional multicenter study among dermatological out-patients in 13 european countries. *Journal of Investigative Dermatology*, *135*(4), 984–991.

Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., Csörgö, S., Cuadras, C. M., de Wet, T., Giné, E., Lockhart, R., Munk, A., & Stute, W. (2000). Contributions of empirical and quantile processes to the asymptotic theory of goodness-of-fit tests. *Test*, *9*(1), 1–96.

Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C., & Rodrıguez-Rodrıguez, J. M. (1999). Tests of goodness of fit based on the l2-wasserstein distance. *Annals of Statistics*, 1230–1239.

Del Barrio, E., Gordaliza, P., Lescornel, H., & Loubes, J.-M. (2019). Central limit theorem and bootstrap procedure for wasserstein's variations with an application to structural relationships between distributions. *Journal of Multivariate Analysis*, *169*, 341–362.

Del Barrio, E., & Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, *47*(2), 926–951.

Delgado, V., Ypenburg, C., van Bommel, R. J., Tops, L. F., Mollema, S. A., Marsan, N. A., Bleeker, G. B., Schalij, M. J., & Bax, J. J. (2008). Assessment of left ventricular dyssynchrony by speckle tracking strain imaging comparison between longitudinal, circumferential, and radial strain in cardiac resynchronization therapy. *J Am Coll Cardiol*, *51*(20), 1944–52. https://doi.org/10.1016/j.jacc.2008.02.040

Di Bella, G., Pizzino, F., Minutoli, F., Zito, C., Donato, R., Dattilo, G., Oreto, G., Baldari, S., Vita, G., Khandheria, B. K., & Carerj, S. (2014). The

mosaic of the cardiac amyloidosis diagnosis: Role of imaging in subtypes and stages of the disease. *Eur Heart J Cardiovasc Imaging*, *15*(12), 1307–15. https://doi.org/10.1093/ehjci/jeu158

Dudley, R. M. (1969). The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, *40*(1), 40–50.

Elton, E. J., Gruber, M. J., & Padberg, M. W. (1977). Simple rules for optimal portfolio selection: The multi group case. *Journal of Financial and Quantitative Analysis*, *12*(3), 329–345.

Evans, D., Jones, A. J., & Schmidt, W. M. (2002). Asymptotic moments of near–neighbour distance distributions. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *458*(2028), 2839–2849.

Fan, J. (2018). *Local polynomial modelling and its applications: Monographs on statistics and applied probability 66*. Routledge.

Fan, J., Wang, D., Wang, K., & Zhu, Z. (2017). Distributed estimation of principal eigenspaces. *arXiv preprint arXiv:1702.06488*.

Ferradans, S., Papadakis, N., Peyré, G., & Aujol, J.-F. (2014). Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, *7*(3), 1853–1882.

Flamary, R., & Courty, N. (2017). Pot python optimal transport library. https://github.com/rflamary/POT

Flamary, R., Cuturi, M., Courty, N., & Rakotomamonjy, A. (2018). Wasserstein discriminant analysis. *Machine Learning*, *107*(12), 1923–1945.

Forrow, A., Hütter, J.-C., Nitzan, M., Rigollet, P., Schiebinger, G., & Weed, J. (2019). Statistical optimal transport via factored couplings, In *The 22nd international conference on artificial intelligence and statistics*.

Fournier, N., & Guillin, A. (2015). On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, *162*(3-4), 707–738.

Gaye, B., Canonico, M., Perier, M. C., Samieri, C., Berr, C., Dartigues, J. F., Tzourio, C., Elbaz, A., & Empana, J. P. (2017). Ideal cardiovascular health, mortality, and vascular events in elderly subjects: The three-city study. *J Am Coll Cardiol*, *69*(25), 3015–3026. https://doi.org/10.1016/j.jacc.2017.05.011

Genevay, A., Chizat, L., Bach, F., Cuturi, M., & Peyré, G. (2019). Sample complexity of sinkhorn divergences, In *The 22nd international conference on artificial intelligence and statistics*.

Genevay, A., Peyre, G., & Cuturi, M. (2018). Learning generative models with sinkhorn divergences, In *International conference on artificial intelligence and statistics*.

Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J. H., Harrington, R. A., Liang, D. H., Ashley, E. A., & Zou, J. Y. (2020). Deep learning interpretation of echocardiograms. *npj Digital Medicine*, *3*(1), 1–10.

Gjesdal, O., Hopp, E., Vartdal, T., Lunde, K., Helle-Valle, T., Aakhus, S., Smith, H. J., Ihlen, H., & Edvardsen, T. (2007). Global longitudinal strain measured by two-dimensional speckle tracking echocardiography is closely related to myocardial infarct size in chronic ischaemic heart disease. *Clin Sci (Lond)*, *113*(6), 287–96. https://doi.org/10.1042/CS20070066

Gomez-Pardo, E., Fernandez-Alvira, J. M., Vilanova, M., Haro, D., Martinez, R., Carvajal, I., Carral, V., Rodriguez, C., de Miguel, M., Bodega, P., Santos-Beneit, G., Penalvo, J. L., Marina, I., Perez-Farinos, N., Dal Re, M., Villar, C., Robledo, T., Vedanthan, R., Bansilal, S., & Fuster, V. (2016). A comprehensive lifestyle peer group-based intervention on cardiovascular risk factors: The randomized controlled fifty-fifty program. *J Am Coll Cardiol*, *67*(5), 476–85. https://doi.org/10.1016/j.jacc.2015.10.033

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(Mar), 723–773.

Gu, C. (2013). *Smoothing spline anova models*. Springer Science & Business Media.

Guo, Z.-C., Lin, S.-B., & Zhou, D.-X. (2017). Learning theory of distributed spectral algorithms. *Inverse Problems*, *33*(7), 74009.

Györfi, L., Kohler, M., Krzyzak, A., & Walk, H. (2006). *A distribution-free theory of nonparametric regression*. Springer Science & Business Media.

Haghighat, M., Abdel-Mottaleb, M., & Alhalabi, W. (2016). Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Transactions on Information Forensics and Security*, *11*(9), 1984–1996.

Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. *Proceedings of the IEEE*, *85*(1), 6–23.

Hall, D. L., & McMullen, S. A. (2004). *Mathematical techniques in multisensor data fusion*. Artech House.

Hinneburg, A., Aggarwal, C. C., & Keim, D. A. (2000). What is the nearest neighbor in high dimensional spaces?, In *26th internat. conference on very large databases*.

Ichimura, H., & Todd, P. E. (2007). Implementing nonparametric and semiparametric estimators. *Handbook of Econometrics*, *6*, 5369–5468.

Kantorovich, L. V. (2006). On a problem of monge. *Journal of Mathematical Sciences*, *133*(4), 1383–1383.

Kantorovitch, L. (1958). On the translocation of masses. *Management Science*, *5*(1), 1–4.

Kim, H.-H., Han, S.-U., Kim, M.-C., Hyung, W. J., Kim, W., Lee, H.-J., Ryu, S. W., Cho, G. S., Song, K. Y., & Ryu, S. Y. (2014). Long-term results of laparoscopic gastrectomy for gastric cancer: A large-scale case-control and case-matched korean multicenter study. *Journal of Clinical Oncology*, *32*(7), 627–633.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, *2*(1-2), 83–97.

Lee, J. D., Liu, Q., Sun, Y., & Taylor, J. E. (2017). Communication-efficient sparse regression. *The Journal of Machine Learning Research*, *18*(1), 115–144.

Li, B. (2018). *Sufficient dimension reduction: Methods and applications with r*. Chapman; Hall/CRC.

Li, K.-C. (2000). High dimensional data analysis via the SIR/PHD approach.

Lloyd-Jones, D. M., Hong, Y., Labarthe, D., Mozaffarian, D., Appel, L. J., Van Horn, L., Greenlund, K., Daniels, S., Nichol, G., Tomaselli, G. F., Arnett, D. K., Fonarow, G. C., Ho, P. M., Lauer, M. S., Masoudi, F. A., Robertson, R. M., Roger, V., Schwamm, L. H., Sorlie, P., … Statistics, C. (2010). Defining and setting national goals for cardiovascular health promotion and disease reduction: The american heart association's strategic impact goal through 2020 and beyond. *Circulation*, *121*(4), 586–613. https://doi.org/10.1161/CIRCULATIONAHA.109.192703

Ma, P., Huang, J. Z., & Zhang, N. (2015). Efficient computation of smoothing splines via adaptive basis sampling. *Biometrika*, *102*(3), 631–645.

Madani, A., Arnaout, R., Mofrad, M., & Arnaout, R. (2018). Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digital Medicine*, *1*(1), 6.

Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W., & Ma, P. (2019). Large-scale optimal transport map estimation using projection pursuit, In *Advances in neural information processing systems*.

Michel, J. B., Sangha, D. M., & Erwin III, J. P. (2017). Burnout among cardiologists. *American Journal of Cardiology*, *119*, 938–940.

Muja, M., & Lowe, D. G. (2014). Scalable nearest neighbor algorithms for high dimensional data. *IEEE transactions on pattern analysis and machine intelligence*, *36*(11), 2227–2240.

Munk, A., & Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(1), 223–241.

Nagueh, S. F., Smiseth, O. A., Appleton, C. P., Byrd, 3., B. F., Dokainish, H., Edvardsen, T., Flachskampf, F. A., Gillebert, T. C., Klein, A. L., Lancellotti, P., Marino, P., Oh, J. K., Alexandru Popescu, B., Waggoner, A. D., Houston, T., Oslo, N., Phoenix, A., Nashville, T., Hamilton, O. C., … St. Louis, M. (2016). Recommendations for the evaluation of left ventricular diastolic function by echocardiography: An update from the american society of echocardiography and the european association of cardiovascular imaging. *Eur Heart J Cardiovasc Imaging*, *17*(12), 1321–1360. https://doi.org/10.1093/ehjci/jew082

Nauta, J. F., Hummel, Y. M., van der Meer, P., Lam, C. S. P., Voors, A. A., & van Melle, J. P. (2018). Correlation with invasive left ventricular filling pressures and prognostic relevance of the echocardiographic diastolic parameters used in the 2016 esc heart failure guidelines and in the 2016 ase/eacvi recommendations: A systematic review in patients with heart failure with preserved ejection fraction. *Eur J Heart Fail*, *20*(9), 1303–1311. https://doi.org/10.1002/ejhf.1220

Nicholls, M. (2019). Cardiologists and the burnout scenario. *European Heart Journal*, *40*, 5–6.

Ouyang, D., He, B., Ghorbani, A., Langlotz, C., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A., & Zou, J. Y. (2019). Interpretable ai for beat-to-beat cardiac function assessment. *medRxiv*, 19012419.

Panaretos, V. M., & Zemel, Y. (2019). Statistical aspects of wasserstein distances. *Annual review of statistics and its application*, *6*, 405–431.

Percus, A. G., & Martin, O. C. (1998). Scaling universalities ofkth-nearest neighbor distances on closed manifolds. *advances in applied mathematics*, *21*(3), 424–436.

Perrot, M., Courty, N., Flamary, R., & Habrard, A. (2016). Mapping estimation for discrete optimal transport, In *Advances in neural information processing systems*.

Peyré, G., Cuturi, M. Et al. (2019). Computational optimal transport. *Foundations and Trends® in Machine Learning*, *11*(5-6), 355–607.

Posner, E. C., Rodemich, E. R., Rumsey, H., Et al. (1967). Epsilon entropy of stochastic processes. *The Annals of Mathematical Statistics*, *38*(4), 1000–1020.

Powell, J. L., Stock, J. H., Stoker, T. M., Et al. (1986). Semiparametric estimation of weighted average derivatives.

Rabin, J., Ferradans, S., & Papadakis, N. (2014). Adaptive color transfer with relaxed optimal transport, In *2014 ieee international conference on image processing (icip)*. IEEE.

Rahbar, K., Ahmadzadehfar, H., Kratochwil, C., Haberkorn, U., Schäfers, M., Essler, M., Baum, R. P., Kulkarni, H. R., Schmidt, M., Drzezga, A., Et al. (2017). German multicenter study investigating 177lu-psma-617 radioligand therapy in advanced prostate cancer patients. *Journal of Nuclear Medicine*, *58*(1), 85–90.

Ramdas, A., Trillos, N., & Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, *19*(2), 47.

Reich, S. (2013). A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, *35*(4), A2013–A2024.

Reimer, K. A., Lowe, J. E., Rasmussen, M. M., & Jennings, R. B. (1977). The wavefront phenomenon of ischemic cell death. 1. myocardial infarct size vs duration of coronary occlusion in dogs. *Circulation*, *56*(5), 786–94. https://doi.org/10.1161/01.cir.56.5.786

Roth, G. A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S. F., Abyu, G., Ahmed, M., Aksut, B., Alam, T., Alam, K., Alla, F., Alvis-Guzman, N., Amrock, S., Ansari, H., Arnlov, J., Asayesh, H., Atey, T. M., Avila-Burgos, L., Awasthi, A., … El Razek, H. M. A., Et al. (2017). Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *J Am Coll Cardiol*, *70*(1), 1–25. https://doi.org/10.1016/j.jacc.2017.04.052

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of. *Reading: Addison-Wesley*, *169*.

Schuhmacher, D., Bähre, B., Gottschlich, C., Hartmann, V., Heinemann, F., & Schmitzer, B. (2019). *Transport: Computation of optimal transport plans and wasserstein distances* [R package version 0.12-1]. R package version 0.12-1. https://cran.r-project.org/package=transport

Schwartz, B. (1994). A computational analysis of the auction algorithm. *European journal of operational research*, *74*(1), 161–169.

Skaarup, K. G., Iversen, A., Jorgensen, P. G., Olsen, F. J., Grove, G. L., Jensen, J. S., & Biering-Sorensen, T. (2018). Association between layer-specific global longitudinal strain and adverse outcomes following acute coro-

nary syndrome. *Eur Heart J Cardiovasc Imaging*, *19*(12), 1334–1342. https://doi.org/10.1093/ehjci/jey004

Su, Z., Wang, Y., Shi, R., Zeng, W., Sun, J., Luo, F., & Gu, X. (2015). Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, *37*(11), 2246–2259.

Thomas, H., Diamond, J., Vieco, A., Chaudhuri, S., Shinnar, E., Cromer, S., Perel, P., Mensah, G. A., Narula, J., Johnson, C. O., Roth, G. A., & Moran, A. E. (2018). Global atlas of cardiovascular disease 2000-2016: The path to prevention and control. *Glob Heart*, *13*(3), 143–163. https://doi.org/10.1016/j.gheart.2018.09.511

Turco, J. V., Inal-Veith, A., & Fuster, V. (2018). Cardiovascular health promotion: An issue that can no longer wait. *J Am Coll Cardiol*, *72*(8), 908–913. https://doi.org/10.1016/j.jacc.2018.07.007

Vendelin, M., Bovendeerd, P. H., Engelbrecht, J., & Arts, T. (2002). Optimizing ventricular fibers: Uniform strain or stress, but not atp consumption, leads to high efficiency. *Am J Physiol Heart Circ Physiol*, *283*(3), H1072–81. https://doi.org/10.1152/ajpheart.00874.2001

Villani, C. (2008). *Optimal transport: Old and new*. Springer Science & Business Media.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wahba, G., & Craven, P. (1978). Smoothing noisy data with spline functions. estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, *31*, 377–404.

Wang, H., & Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, *103*(482), 811–821.

Wang, X., Shen, J., & Ruppert, D. (2011). On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics*, *5*, 1–17.

Wasserman, L. (2013). *All of statistics: A concise course in statistical inference*. Springer Science & Business Media.

Weed, J., & Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *Bernoulli*, *25*(4A), 2620–2648.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, *5*(2), 241–259.

Xia, Y. Et al. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, *35*(6), 2654–2690.

Xia, Y., Tong, H., Li, W., & Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(3), 363–410.

Yang, B., Daimon, M., Ishii, K., Kawata, T., Miyazaki, S., Hirose, K., Ichikawa, R., Chiang, S. J., Suzuki, H., Miyauchi, K., & Daida, H. (2013). Prediction of coronary artery stenosis at rest in patients with normal left ventricular wall motion. segmental analyses using strain imaging diastolic index. *Int Heart J*, *54*(5), 266–72. https://www.ncbi.nlm.nih.gov/pubmed/24097214

Zhang, J., Gajjala, S., Agrawal, P., Tison, G. H., Hallock, L. A., Beussink-Nelson, L., Lassen, M. H., Fan, E., Aras, M. A., Jordan, C., Fleischmann, K. E., Melisko, M., Qasim, A., Shah, S. J., Bajcsy, R., & Deo, R. C. (2018). Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy. *Circulation*, *138*(16), 1623–1635.

Zhang, L., Wu, W. C., Ma, H., & Wang, H. (2016). Usefulness of layer-specific strain for identifying complex cad and predicting the severity of coronary lesions in patients with non-st-segment elevation acute coronary syndrome: Compared with syntax score. *Int J Cardiol*, *223*, 1045–1052. https://doi.org/10.1016/j.ijcard.2016.08.277

Zhang, T., Zhu, D., Jiang, X., Zhang, S., Kou, Z., Guo, L., & Liu, T. (2016). Group-wise consistent cortical parcellation based on connectional profiles. *Medical image analysis*, *32*, 32–45.

Zhang, Y., Duchi, J., & Wainwright, M. (2015). Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, *16*(1), 3299–3340.

Zhou, H. H., Singh, V., Johnson, S. C., Wahba, G., Initiative, A. D. N., Et al. (2018). Statistical tests and identifiability conditions for pooling and analyzing multisite datasets. *Proceedings of the National Academy of Sciences*, *115*(7), 1481–1486.

Zhou, H. H., Zhang, Y., Ithapu, V. K., Johnson, S. C., Wahba, G., & Singh, V. (2017). When can multi-site datasets be pooled for regression? hypothesis tests, ℓ_2-consistency and neuroscience applications. *arXiv preprint arXiv:1709.00640*.

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. Chapman; Hall/CRC.