Novel Statistical Learning Methods for High-Dimensional Complex Biomedical Data Analysis

by

Daiwei Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics and Scientific Computing)
in The University of Michigan
2021

Doctoral Committee:

Professor Jian Kang, Co-Chair
Associate Professor Seunggeun Lee, Co-Chair
Professor Veera Baladandayuthapani
Professor Ji Zhu

Daiwei Zhang

daiweiz@umich.edu

ORCID ID: 0000-0002-5019-622X

For our grandmothers.

ACKNOWLEDGEMENTS

I am grateful for receiving my doctoral training from the Department of Biostatistics at the University of Michigan. It is a blessing to begin one's professional journey as a biostatistician in a friendly environment filled with experts from a diverse range of statistical and scientific backgrounds. The faculty members, with support from the efficient staff members, have provided a rigorous academic experience for me and my cohort and mentored us to become highly capable researchers. It has been a precious experience for me to work with this group of exceptionally intelligent professors and colleagues.

I would like to thank the people from whom I received valuable help during my residency in Ann Arbor. First and foremost, my gratitude goes to my co-advisors Drs. Seunggeun Lee and Jian Kang. Dr. Lee was my first mentor when I embarked on my degree program. He introduced me to the field of statistical genetics and helped me sharpen my computation skills. During the trial and error of my research, he patiently provided feedback on the obstacles I encountered and gave me the freedom to explore various potential solutions. He was also a gracious advisor in the ups and downs of my early years as a doctoral student. As his mentee, I have gained valuable experience from his expertise in genetics and grown significantly as a researcher in general. This dissertation would also have been impossible without my co-advisor Dr. Kang. Dr. Kang's passion for Bayesian methods and statistical learning has inspired me to dive deep in these directions. His mind is filled with myriads of brilliant ideas, and many of them have stimulated the development of my own novel approaches. His

devotion to work, as displayed in the emails replies he wrote after midnight and the manuscript comments he made during his vacation time, continues to motivate me to set a high standard for myself. I appreciate the time and effort that he spent on the training of my mathematical, statistical, computing, and writing skills, as well as the numerous research and collaboration opportunities he provided that have shaped my vision for the future of my career.

Furthermore, I would like to thank my committee members Drs. Veera Baladandayuthapani and Ji Zhu, for reviewing the dissertation and providing constructive criticism. The projects in this work were made possible with contributions from my collaborators, including Drs. Rounak Dey, Lexin Li, Chandra Sripada, and Tianci Liu. During my years as a doctoral student, I have also received countless help on projects outside my dissertation from numerous faculty mentors, including Drs. Xiang Zhou, Laura Scott, Tianwei Yu, Michael Boehnke, Lars Fritsche, as well as senior colleagues, including Emily Hector and Xianyong Yin. In addition, my research work would have been much less productive without the help from the supporting staff members, including computing cluster administrators Sean Caron and Daniel Barker, departmental coordinators Nicole Fenech, Fatma-Zohra Nedjari, and Kerry Sprague, Dr. Kirsten Herold from the writing lab, and Shelagh Saenz from the careers office.

Lastly and the most importantly, my greatest gratitude goes to my family. I am indebted to my mother- and father-in-law, who provided irreplaceable support for we two student parents during our most difficult days. As for our beloved daughter, I am blessed with the joy she has sprinkled in my life and the efficiency she has pushed me to achieve. Finally, the highest honor is reserved for my wife. She is a true helper of mine both at home and for work. Her wisdom, courage, perseverance, and self-sacrificial love have accompanied me through the mountains and valleys in our life.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

Over the past decades, biomedical data have grown rapidly both in dimension and in complexity. Traditional statistical models often lack the power of detecting the nonlinear associations underlying the complex high-dimensional biomedical data. Machine learning (ML) methods, on the other hand, have been shown to be successful for solving the challenging problems in some applications. However, because of a "black box" nature, standard ML neither elucidates the data-generation mechanism nor quantifies the model-fitting uncertainty, which have largely limited their usefulness in biomedical studies. Furthermore, the sample sizes required by sophisticated ML approaches, such as deep neural networks, for analyzing large-scale data, such as those commonly found in imaging genetics and spatial transcriptomics, are not widely affordable in typical medical studies. These difficulties have contributed to the relatively scant success of ML in biomedical applications. To address these challenges, this dissertation aims at developing several novel approaches that combine traditional statistical models with ML algorithms to efficiently and effectively analyze large-scale complex biomedical data.

In the first project, we develop a robust and fast method based on principal component analysis (PCA) for predicting population stratification (PS) from genotypes. PS is a major confounder in genome-wide association studies that can lead to false positive associations. Although PCA-based methods have been widely adopted for PS adjustment, existing methods are either biased toward the null or computationally expensive for large reference sets. In response, we propose two alternative approaches

that can estimate the asymptotic shrinkage bias using random matrix theory and reduce the computation cost with online SVD. The proposed methods are applied to extensive simulation studies and data in the UK Biobank and the 1000 Genomes Project. We show that compared with existing methods, our methods are unbiased and the computation cost is significantly lower.

In the second project, we propose a novel image-on-scalar regression (ISR) model to study the association between imaging measurements and scalar covariates. Statistical inferences on medical ISR is challenging due to the high imaging dimensionality, limited number of images, complex spatial correlations, and heterogeneous noises. To address these challenges, we utilize deep neural networks to model the spatially varying coefficient functions of the main effects, individual effects, and noise variance in the ISR model (NNISR). Compared to existing methods, NNISR is more flexible for capturing complex spatial patterns, more straightforward to interpret, and more accurate for small numbers of high-resolution images. We develop computationally efficient and scalable algorithms for parameter estimation and activation region selection. Theoretical analysis is conducted to establish estimation and selection consistency of the proposed method. The superiority of NNISR is further demonstrated through extensive simulations and analyses of brain fMRI data.

In the third project, we focus on modeling the conditional distribution of the response given predictors via deep neural networks. Standard neural network regression makes prediction on the response using the conditional mean and often assumes a simple homoscedastic error distribution. To better quantify prediction uncertainty, we develop a novel Bayesian hierarchical neural network model by introducing latent variables at each hidden layer, which induces high flexibility in modeling the predictive distribution of the response. In light of the special structure of the proposed model, we develop a scalable and accurate Gibbs sampling for posterior computation. We illustrate the proposed method via simulations and analysis of neuroimaging data.

CHAPTER I

Introduction

Machine learning (ML) methods have been successful in solving many artificial intelligence (AI) problems. ML models such as deep neural networks, support vector machines, and random forests are capable of detecting highly complex patterns and making accurate predictions. However, standard ML methods do not quantify the uncertainty involved in model fitting and data generation, which makes them unable to conduct the statistical inferences needed for scientific inquiries. Moreover, typical biomedical studies cannot afford the large sample sizes required for training ML models with high numbers of parameters. Furthermore, the "black box" characteristics of these models and their inability to provide insightful explanations have limited their usefulness in biomedical research and applications, as stakeholders often require a mechanistic (and ideally causal) understanding of the prediction-making procedure and how the model reacts to changes in the inputs [Wainberg et al., 2018]. Finally, it has been shown that even after achieving extraordinarily high training and testing accuracy, trained ML models can be sensitive to small perturbations in data [Su et al., 2019], which can potentially cause detrimental decision-making in safety-critical applications, such as precision health and clinical trials. This shortcoming is exacerbated by the heterogeneity of biomedical data, where the stream of upcoming samples can contain inputs that are drastically different from those used for model

training. For example, each cancer patient's genetic, imaging, and metabolomic profiles can be unique and share little similarity with other cancer patient's. Thus simply increasing the sample size will not guarantee a solution to this problem, since even a training set that contains all existing patients might not exhaust all the possible variations in biological systems of unfathomable complexity [Michael et al., 2018]. The lack of uncertainty quantification, training efficiency, interpretability, and robustness in ML methods may explain the skepticism of physicians and medical researchers, as well as ML's relatively scant success in biomedical fields, as compared to their wide adaptation in AI applications, such as computer vision and natural language processing.

In this dissertation, we aim to address these challenges by developing novel statistical machine learning methods for analyzing complex biomedical data. Before we present the projects in this dissertation, we will first introduce the relevant backgrounds in Sections 1.1 to 1.3. The dissertation outline is listed in Section 1.4.

1.1 Population stratification in genetic association studies

Since the first genome-wide association study (GWAS) was conducted in 2002 [Ozaki et al., 2002, Thomas et al., 2005, Balding et al., 2008, Ikegawa, 2012], many common single nucleotide polymorphisms (SNPs) have been discovered and verified to be associated with human diseases and traits. As the focus started to shift toward the research of rare variants, the need for larger sample sizes continued to grow, and analyses involving multiple study centers have become increasingly prevalent. In both single-center and multi-center genetic studies, adjusting for ancestry membership is a common practice to avoid spurious allelic associations caused by population stratification [Cardon and Palmer, 2003]. A widely used approach for detecting population structure from genotypes is principal component analysis (PCA). PCA utilizes singular value decomposition (SVD) to search for the linear direction with the greatest

sample variation. In its standard usage, PCA is applied to the study samples, and the resulting principal component (PC) scores serve as covariates to adjust for in association studies. However, when the study samples are composed of data from multiple sources, a consistent approach is needed to match the individual ancestry across datasets [Wang et al., 2015]. To this end, two-sample PCA methods have been developed to predict PC scores by using a reference panel that consists of samples outside the study set. However, existing methods of this type are either computationally costly or biased toward the null. Therefore, in this dissertation, we develop population stratification methods that offer fast and robust ancestry prediction. We apply our methods to predict the fine-scale (e.g. sub-European) ancestry of 488,366 genotyped samples collected from multiple study centers in the UK Biobank [Biobank, 2014], with 2,492 samples from the 1000 Genomes Project [Clarke et al., 2012] serving as the reference panel.

1.2 Association studies in functional neuroimaging

Decrease in the cost of collecting high-dimensional medical images has stimulated the availability of neuroimages in biomedical studies [Liu et al., 2017]. Brain images of different modalities, such as X-ray computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), single photon emission tomography (SPECT), functional MRI (fMRI), and magnetic resonance spectroscopy [Powers and Derdeyn, 2014], provide the physiological neuroimaging data needed for untangling the association of brain regions of interest (ROIs) with physiological, clinical, and demographic characteristics. On the other hand, the growing threat of neuropsychiatric disorders, such as Alzheimer's disease and autism spectrum disorder (ASD), presses for the discovery of ROIs that can potentially accelerate the detection of high-risk individuals and assist the development of intervention techniques [Chen et al., 2016]. To this end, image-on-scalar regression models (ISRs) become a

natural tool for finding the association between high-dimensional imaging data and scalar variables, such as cognitive score, genetic variants, and disease status. However, medical imaging analysis with ISRs is complicated by several difficulties, including the ultrahigh imaging dimensions, heterogeneous noises, limited numbers of training images, and complex spatial correlations. To address these challenges, we propose a deep learning-guided ISR that utilizes deep neural networks in the search of spatial patterns. We identify ROIs significantly associated with intellectual capacities by applying our methods to two neuroimaging data sets: the Autism Brain Imaging Data Exchange (ABIDE), a consortium aggregating the resting-state fMRI images, structural MRI images, and phenotypic information from 1,112 subjects [Di Martino et al., 2014], and the Adolescent Brain Cognitive Development study (ABCD), a study of over ten thousand 9- and 10-year-old children recruited from 21 sites in the United States [Jernigan et al., 2018].

1.3 Density learning with Bayesian neural networks

Although deep learning has achieved high testing prediction accuracy in many artificial intelligence applications, one of its major drawbacks is the lack of estimation on the model fitting error. This deficiency can be solved by treating the neural network as a Bayesian hierarchical model, also known as a Bayesian neural network (BNN), which provides a quantification of estimation uncertainty through the posterior distribution [Wang and Yeung, 2016]. However, in standard BNNs, variation in the noise is often simplified as a homoscedastic, zero-mean random variable. Such an approach may misrepresent or underestimate the deviation of the true outcome from the predicted value and cause detrimental decision-making [Huang et al., 2018], especially in the presence of multi-modality and heavy tails. To better quantify the unpredictable randomness in the data, we develop a generalized BNN model for learning not merely the conditional mean, but rather the whole conditional distribution of

the outcome. We show that our density learning model is capable of approximating a wide range of densities. Moreover, the usefulness of BNNs in real-life applications has been greatly limited by the lack of efficient and accurate posterior computation algorithms. Standard MCMC methods are inefficient for exploring the ultrahigh-dimensional parameters of BNNs [Izmailov et al., 2020], while variational inference methods tend to underestimate the posterior variances [Blei et al., 2017]. To address these challenges, we will propose a Bayesian hierarchical model with latent variables. The novel model is an extension of standard neural networks and is capable of representing complex noise structures. Moreover, in light of the model structure, we will develop an efficient posterior Gibbs sampler that utilizes the closed-form conditional distributions in our model. We will apply our approach to simulated data and neuroimaging data to evaluate its characterization of non-Gaussian noise distributions and assess its effectiveness against making overconfident predictions.

1.4 Dissertation outline

The remainder of the dissertation is organized as follows: In Chapter II, we develop a fast and robust approach (FRAPOSA) to predict the ancestry information of the genotypes in the UK Biobank. In Chapter III, we design a neural network-guided image-on-scalar regression model (NNISR) and apply it to functional magnetic resonance (fMRI) data. In Chapter IV, we propose the deep aleatoric neural network (DALEA) model and evaluate its performance on neuroimaging data. We conclude in Chapter V with a discussion and potential future directions.

CHAPTER II

Fast and Robust Ancestry Prediction Using Principal Component Analysis

2.1 Introduction

Population stratification (PS) is a major confounder for genetic association analysis [Price et al., 2006], and the adjustment of PS requires the estimation of the ancestry structure among study samples. Principal component analysis (PCA) is a multivariate statistical method which finds the direction of the maximal variability [Jolliffe, 2002]. By aggregating information across all the genetic markers, PCA has been effective for PS adjustment [Reich et al., 2008]. To adjust for PS, PCA can be applied to study data to calculate the principal component (PC) scores, which are regarded as variables of ancestry and can be used as covariates to adjust for. An alternative approach is predicting the PC scores of the study samples by using reference genotyped samples with detailed ancestry information. This prediction-based approach allows not only adjustment for PS but also inference of the ancestry memberships of the study samples. In addition, by using a common reference panel, predicted PC scores across different studies can be directly comparable, allowing to integrate and match the different study samples [Wang et al., 2015]. For example, using the predicted PC scores, Zhan et al. [2013] identified the ancestry-matched con-

trol samples from the publicly available NHLBI ESP sequencing data, which helped to identify rare variant associations.

The standard approach of predicting PC scores is to project the study samples onto the maximal variability directions, called PC loadings. In this paper, we call this approach simple projection (SP). However, when the number of features greatly exceeds the size of the reference samples, which is common for data in genome-wide association studies (GWAS), the PC scores predicted by SP are known to be systematically biased toward NULL [Dey and Lee, 2019]. This shrinkage bias can cause inaccurate prediction of the ancestry of each study sample and inappropriate adjustment of PS.

One way of addressing this shrinkage bias is presented by Wang et al. [2014, 2015]. Their solution is to combine one study sample with all the reference samples and find the PC scores of this augmented data set. The PC scores of the study individuals are then mapped to the reference sample PC space by a Procrustes transformation. We call this method "augmentation, decomposition, and Procrustes transformation" (ADP). This method has been shown to be effective in eliminating the shrinkage bias of study PC scores. However since ADP needs to run PCA separately for each of the augmented data sets, it is computationally expensive, especially with large reference samples. For example, the estimated computation time for predicting the ancestry of the UK Biobank data of 488,366 samples with 2,492 reference samples is 1,628 CPU hours. Since computation time is cubic to the reference sample size, the computation time will rapidly increase for larger reference samples.

To address the limitations of SP and ADP, we develop and propose two alternative methods for ancestry prediction and apply them to the UK Biobank data. The first approach removes the bias in SP by estimating the asymptotic bias factor, which is calculated based on random matrix theory [Dey and Lee, 2019]. The second approach improves the computational efficiency of ADP by using an online singular

value decomposition (SVD) algorithm [Halko et al., 2011], which obtains the SVD results of the augmented matrix by updating the SVD results of the reference matrix, since the latter only differs slightly from the former and many of the overlapping calculations can be avoided. We call the first approach "bias-adjusted projection" (AP) and the second approach "online augmentation, decomposition, and Procrustes transformation" (OADP).

In this paper, we evaluate the accuracy and computational efficiency of AP and OADP as compared to SP and ADP through extensive simulation studies and the analysis of the UK Biobank data. In the simulation studies, we show that AP and OADP have both achieved accuracy similar to or higher than that of ADP and computational efficiency close to that of SP. The UK Biobank data analysis shows that the proposed approaches are 80-2000 times faster than ADP. In addition, we have developed the open-source software FRAPOSA in Python that implements AP, OADP, SP, and ADP.

2.2 Methods

2.2.1 Model and PCA on the reference data

For PC score prediction, we have the reference samples and the study samples, which can be represented by two matrices. Let $\underline{\boldsymbol{X}}$ be a $p \times n$ matrix of reference genotypes and $\underline{\boldsymbol{Y}}$ be a $p \times m$ matrix of study genotypes, where p is the number of genetic markers, n is the number of reference samples, and m is the number of study samples. In our study, we only consider genotypes composed of biallelic single nucleotide polymorphisms (SNP), so each entry of $\underline{\boldsymbol{X}}$ and $\underline{\boldsymbol{Y}}$ is a minor allele count of 0, 1, or 2. For PCA, the reference data matrix is commonly standardized by subtracting the marker mean from each marker genotype and then dividing it by the marker standard deviation. The sample matrix $\underline{\boldsymbol{Y}}$ also can be standardized using marker means and

standard deviations calculated from the reference samples. Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ are the standardized reference and study data matrices, respectively. The sample covariance matrix is $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^\top/n$, and then by eigendecomposition,

$$n\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{U}\boldsymbol{D}^2\boldsymbol{U}^\top$$

where $\boldsymbol{D}^2 = \text{diag}(d_1^2, \ldots, d_n^2)$ is an $n \times n$ diagonal matrix of ordered sample eigenvalues and $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ is a $p \times n$ corresponding eigenvector matrix. The j^{th} PC score vector is $\boldsymbol{v}_j = \boldsymbol{X}^\top \boldsymbol{u}_j/d_j$, where $\boldsymbol{u}_j$ is the j^{th} sample eigenvector, which is also called the j^{th} PC loading. Alternatively, PC loadings and scores can be calculated using SVD, which is computationally more efficient when p is larger than n. By SVD

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \tag{2.1}$$

where $\boldsymbol{V} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_n)$ is the right singular vector matrix and $\boldsymbol{v}_j$ is the j^{th} PC scores. From (2.1),

$$\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top.$$

After calculating $\boldsymbol{v}_j$ and d_j from the eigendecomposition of $\boldsymbol{X}^\top\boldsymbol{X}$, the j^{th} loading, $\boldsymbol{u}_j$, can be calculated as $\boldsymbol{u}_j = \boldsymbol{X}\boldsymbol{v}_j/d_j$.

2.2.2 Predicting the PC scores of the study samples

Here we describe the existing approaches, SP and ADP, and the proposed approaches, AP and OADP, and their computation complexity to predict the top K PC scores. For practical purposes, we assume that $K \ll n \ll p$. Table 2.1 summarizes the computation complexity of the four methods.

Simple Projection (SP). SP directly uses the PC loadings of the reference sample PCA to predict the PC scores of the study samples. The SP algorithm of predicting

the top K PC scores and the computation complexity (CC) of each step is as follows:

1. Perform the reference sample PCA: $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top$. (CC: $\mathcal{O}[pn^2]$.)

2. Compute the PC loading matrix for the top K PCs: $\boldsymbol{U}_K = \boldsymbol{X}\boldsymbol{V}_K\boldsymbol{D}_K^{-1}$. Here $\boldsymbol{V}_K$ and $\boldsymbol{D}_K$ are the the first K columns of $\boldsymbol{V}$ and the upper-left $K \times K$ sub-matrix of $\boldsymbol{D}$, respectively. (CC: $\mathcal{O}[npK]$.)

3. Compute the predicted study PC scores for the top K PCs: $\boldsymbol{W}_K = \boldsymbol{Y}^\top \boldsymbol{U}_K$. (CC: $\mathcal{O}[mpK]$.)

The total computation complexity is $\mathcal{O}[pn^2 + mpK]$ (assuming $K \ll n \ll p$), which is the lowest among all the methods discussed in this paper. However, a major weakness of SP is the loss of accuracy when the number of makers, p, greatly exceeds the reference sample size, n, a situation that is common in GWAS. Lee et al. [2010] have shown that when $n < p$, the predicted PC scores can be shrunken toward NULL. This shrinkage bias limits the accuracy of SP for high-dimensional data.

Bias-Adjusted Projection (AP). AP calculates the asymptotic shrinkage bias of SP and adjusts the predicted PC scores using the estimated bias. The estimation of the bias requires all the eigenvalues of the the reference data matrix. The details for estimating the shrinkage factor are described in Dey and Lee [2019]. Suppose the population covariance matrix $\boldsymbol{\Sigma} = \mathrm{E}(\boldsymbol{X}\boldsymbol{X}^\top/n)$ has (population) eigenvalues $\lambda_1^2, \ldots, \lambda_p^2$, and the sample covariance matrix $\boldsymbol{S} = \boldsymbol{X}\boldsymbol{X}^\top/n$ has nonzero (sample) eigenvalues $d_1^2, \ldots, d_n^2$. First, the population eigenvalues are assumed to follow a generalized spiked population model (GSP), where only a few eigenvalues are large (which are called distant spikes) compared to the rest of them. The rest of the eigenvalues are relatively small but not necessarily all equal to each other. Then for the top few PCs that correspond to the distant spikes, the ratio of the variance of the reference PC scores and that of the study PC scores predicted by SP converges in

probability to the ratio of the corresponding population eigenvalues (distant spikes) and the sample eigenvalues as $p \to \infty,\ n \to \infty,\ p/n \to \gamma < \infty$. Formally, suppose $v_{kj} = \boldsymbol{x}_j^\top \boldsymbol{u}_k$ is the k-th PC score of the j-th subject in the standardized reference data $\boldsymbol{X}$, and $w_{kl} = \boldsymbol{y}_l^\top \boldsymbol{u}_k$ is the k-th PC score of the l-th subject in the standardized study data $\boldsymbol{Y}$. Then the shrinkage factor along the k-th PC score is defined as $\tau_k = \sqrt{\mathrm{Var}(w_{kl})/\mathrm{Var}(v_{kj})}$, and when λ_k is a distant spike with multiplicity one, $|\tau_k - d_k/\lambda_k| \xrightarrow{p} 0$. Dey and Lee [2019] provides two consistent estimators of λ_k for the distant spikes (i.e. $\hat{\lambda}_k$). The consistent estimator of τ_k can be obtained as $\hat{\tau}_k = d_k/\hat{\lambda}_k$. Among the two estimators of λ_k, we used the method called d-estimation, which is faster (CC: $\mathcal{O}[Kn]$) than the other l-estimation approach (CC: $\mathcal{O}[Kp]$).

The method for approximating the shrinkage factors has been implemented in the `hdpca` package in the R language [Dey and Lee, 2016]. The algorithm of AP is summarized below.

1. Perform the reference sample PCA: $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}^\top$. (CC: $\mathcal{O}[pn^2]$.)

2. Estimate the shrinkage factors $\hat{\tau}_1, \ldots, \hat{\tau}_K$ for the top K PCs, where $\hat{\tau}_k = d_k/\hat{\lambda}_k$ as defined above. (CC: $\mathcal{O}[Kn]$.)

3. Compute the PC loading matrix for the top K PCs with the adjustment for the shrinkage bias: $\boldsymbol{U}_K = \boldsymbol{X}\boldsymbol{V}_K\boldsymbol{D}_K^{-1}\boldsymbol{F}_K^{-1}$, where $\boldsymbol{F}_K = \mathrm{diag}(f_1, ..., f_K)$. (CC: $\mathcal{O}[pnK]$.)

4. Compute the predicted study PC scores for the top K PCs: $\boldsymbol{W}_K = \boldsymbol{Y}^\top \boldsymbol{U}_K$. (CC: $\mathcal{O}[mpK]$.)

The total computation complexity is $\mathcal{O}[pn^2 + mpK]$ (assuming $K \ll n \ll p$), which is the same as that of SP. This is because shrinkage factor estimation is asymptotic-based and can be computed rapidly with the sample eigenvalues. In addition, the shrinkage factor only needs to be calculated once for all the study samples.

Augmentation, Decomposition, and Procrustes Transformation (ADP). ADP, such as LASER and TRACE [Wang et al., 2014, 2015], predicts the study PC scores by using a different approach compared to SP and AP. ADP first augments the (standardized) reference matrix by appending a column vector of a (standardized) study sample. Then SVD is applied to the $p \times (n+1)$ augmented matrix $\tilde{\boldsymbol{X}}$. The resulted $(n+1) \times (n+1)$ right singular-vector matrix $\tilde{\boldsymbol{V}}$ can be divided into two parts: the first n rows $\tilde{\boldsymbol{V}}_{\text{ref}} = (\tilde{\boldsymbol{v}}_{\text{ref},1}^\top, \ldots, \tilde{\boldsymbol{v}}_{\text{ref},n}^\top)^\top$, which correspond to the reference samples, and the last row $\tilde{\boldsymbol{v}}_{\text{stu}}$, which corresponds to the one study sample. Since $\tilde{\boldsymbol{V}}_{\text{ref}}$ is different (though only slightly when n is large) from $\boldsymbol{V}$, the $n \times n$ right singular-vector matrix of the reference data, ADP uses the Procrustes transformation to map $\tilde{\boldsymbol{V}}_{\text{ref}}$ to $\boldsymbol{V}$ in the original reference PC space. That is, it finds a linear transformation of the form

$$f(\tilde{\boldsymbol{v}}_{\text{ref},i,K'}) = \rho \tilde{\boldsymbol{v}}_{\text{ref},i,K'} \boldsymbol{A} + \boldsymbol{c}$$

that minimizes the mean squared difference between $\boldsymbol{V}_K$ and the transformed $(f(\tilde{\boldsymbol{v}}_{\text{ref},1,K'})^\top, \ldots, f(\tilde{\boldsymbol{v}}_{\text{ref},n,K'})^\top)^\top$, where $\boldsymbol{V}_K$ is the first K columns of $\boldsymbol{V}$, $\tilde{\boldsymbol{v}}_{\text{ref},i,K'}$ is the first K' columns of $\tilde{\boldsymbol{v}}_{\text{ref},i}$, and $K \leq K'$. Here ρ is a non-negative scalar, $\boldsymbol{A}$ is an $K' \times K$ orthogonal matrix, and $\boldsymbol{c}$ is an $1 \times K$ row vector. We then apply this transformation to $\tilde{\boldsymbol{v}}_{\text{stu},K'}$, the first K' columns of $\tilde{\boldsymbol{v}}_{\text{stu}}$, to obtain the predicted PC score, $f(\tilde{\boldsymbol{v}}_{\text{stu},K'})$. The algorithm is summarized as follows.

1. Perform the reference sample PCA. $\boldsymbol{X}^\top \boldsymbol{X}$ is obtained in this process. (CC: $\mathcal{O}[pn^2]$.)

2. For a study sample $\boldsymbol{y}$, obtain $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ by computing $\boldsymbol{X}^\top \boldsymbol{y}$, $(\boldsymbol{X}^\top \boldsymbol{y})^\top$, and $\boldsymbol{y}^\top \boldsymbol{y}$ and appending them to the right edge, bottom edge, and bottom-right corner of $\boldsymbol{X}^\top \boldsymbol{X}$, respectively. (CC: $\mathcal{O}[pn]$.)

3. Apply eigendecomposition on $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}}$ to get $\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{X}} = \tilde{\boldsymbol{V}} \tilde{\boldsymbol{D}}^2 \tilde{\boldsymbol{V}}^\top$. (CC: $\mathcal{O}[n^3]$.)

4. Find the Procrustes transformation f from $\tilde{\boldsymbol{V}}_{\text{ref},K'}$, the first n rows and first K' columns of $\tilde{\boldsymbol{V}}$, to $\boldsymbol{V}_K$, the first K columns of $\boldsymbol{V}$. Note that $K' \geq K$. (CC: $\mathcal{O}[nK'^2]$)

5. Apply f to $\tilde{\boldsymbol{v}}_{\text{stu},K'}$, the last row and first K' columns of $\tilde{\boldsymbol{V}}$, to obtain the top K PC scores of the current study sample. (CC: $\mathcal{O}[KK']$)

6. Go to Step 2 for the next study sample unless all the study samples have been analyzed.

The total computation complexity is $\mathcal{O}[pn^2 + m(np + n^3)]$ given that $K' \ll n \ll p$. In our simulation studies and UK Biobank data analysis, setting $K = 4$ and $K' = 8$ was sufficient for separating the ancestry groups.

ADP is a nonparametric approach that does not require any assumption on the distribution of the eigenvalues and therefore can be more robust than AP. It does not suffer the shrinkage bias. A major disadvantage of ADP, however, is its high computation cost. In particular, as the reference size increases, the computation cost for a study sample increases cubicly.

Online Augmentation, Decomposition, and Procrustes Transformation (OADP). Since the augmented data matrix $\tilde{\boldsymbol{X}}$ differs in only one column from the reference matrix $\boldsymbol{X}$, the computational process for the SVD of $\tilde{\boldsymbol{X}}$ is numerically close to that for the SVD of $\boldsymbol{X}$. If we avoid the repeated computation and obtain the SVD of $\tilde{\boldsymbol{X}}$ by updating the SVD of $\boldsymbol{X}$, the computation cost can be greatly reduced. One of such "online" algorithms for SVD has been proposed for imaging processing [Brand, 2002]. This algorithm calculates SVD in an incremental manner and has the ability to rapidly update the top few singular values and vectors. Here we propose to use this online SVD algorithm to replace the standard SVD algorithm for ADP and call it "online augmentation, decomposition, and Procrustes transformation" (OADP). The algorithm for this method is as follows:

1. Perform the reference sample PCA. (CC: $\mathcal{O}[pn^2]$.)

2. Calculate the top K'' PC loadings: $\boldsymbol{U}_{K''} = \boldsymbol{X}\boldsymbol{V}_{K''}\boldsymbol{D}_{K''}^{-1}$. (CC: $\mathcal{O}[K''np]$.)

3. Calculate

$$\boldsymbol{b} = \boldsymbol{U}_{K''}^{\top}\boldsymbol{y} \quad \text{and} \quad g = \boldsymbol{y}^{\top}\boldsymbol{h},$$

   where $\boldsymbol{h}$ is the normalized $\boldsymbol{y} - \boldsymbol{U}_{K''}\boldsymbol{b}$. (CC: $\mathcal{O}[K''p]$.)

4. Calculate $\boldsymbol{Q}^{\top}\boldsymbol{Q}$, where

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{D}_{K''} & \boldsymbol{b} \\ \boldsymbol{0} & g \end{bmatrix}.$$

   (CC: $\mathcal{O}[K''^3]$.)

5. Apply eigendecomposition to $\boldsymbol{Q}^{\top}\boldsymbol{Q}$ to get $\boldsymbol{Q}^{\top}\boldsymbol{Q} = \ddot{\boldsymbol{V}}\ddot{\boldsymbol{D}}^2\ddot{\boldsymbol{V}}^{\top}$. (CC: $\mathcal{O}[K''^3]$.)

6. Calculate

$$\tilde{\boldsymbol{V}} = \begin{bmatrix} \boldsymbol{V}_{K''} & \boldsymbol{0} \\ \boldsymbol{0} & 1 \end{bmatrix} \ddot{\boldsymbol{V}}.$$

   (CC: $\mathcal{O}[nK''^2]$.)

7. Find the Procrustes transformation f from $\tilde{\boldsymbol{V}}_{\text{ref},K'}$, the first n rows and first K' columns of $\tilde{\boldsymbol{V}}$, to $\boldsymbol{V}_K$, the first K columns of $\boldsymbol{V}$. Note that $K'' \geq K' \geq K$. (CC: $\mathcal{O}[nK'^2]$)

8. Apply f to $\tilde{\boldsymbol{v}}_{\text{stu},K'}$, the last row and first K' columns of $\tilde{\boldsymbol{V}}$, to obtain the top K PC scores of the current study sample. (CC: $\mathcal{O}[KK']$)

9. Go to Step 3 for the next study sample unless all the study samples have been analyzed.

The total computation complexity is $\mathcal{O}[n^2p + m(K''p + K'^2n)]$ provided $K'' \ll n \ll p$. In our simulation studies and UK Biobank data analysis, setting $K = 4$, $K' = 8$,

and $K'' = 16$ was sufficient for the online SVD algorithm to approximate regular SVD well and separating the ancestry groups. The computation complexity of OADP for analyzing the study individuals increases linearly with respect to the reference sample size, which is much more efficient than ADP's cubicly increasing rate. The closeness between the results given by OADP and ADP is empirically shown in Section 2.3.

2.2.3 Simulation Studies

We simulated the genotype data using a coalescence-based grid simulation approach with population migration by Mathieson and McVean [2012]. In this approach, we simulated 4 different population groups in a 2×2 grid. In each population, we generated $(n + m)/2$ haploid genotypes with 100,000 biallelic genetic markers. Then we combined every two of the haploid genotypes to form $(n+m)/4$ diploid genotypes in each population. A large migration rate (M=100) was used to evaluate the performance of the proposed and existing methods in fine-scale population differentiation. Among the $(n + m)$ generated samples, we randomly selected reference and study samples. The reference sample size n ranged from 1000 to 3000, and the study sample size m was fixed to 200. The proportion of variants with minor allele frequency less than 0.05, 0.005, 0.0005 was 0.66, 0.37, 0.12, respectively.

After the individual genotypes were simulated, we applied SP, ADP, AP, and OADP to the data to predict the PC scores for the study samples. We only calculated the top 2 PCs, and for OADP and ADP, we calculated the top 8 PC scores (i.e. $K' = 8$) for the study samples and project them to the 2-dimensional reference PC score space through the Procrustes transformation. For OADP, we calculated the top 16 PC scores in the online SVD algorithm (i.e. $K'' = 16$) but used only the top 8 PCs for the Procrustes transformation (i.e. $K' = 8$). Finally, we used the 20-nearest-neighbor method to predict each study sample's population membership. It classified a study sample by the votes of the 20 nearest neighboring reference samples, where

the weight of each neighbor was inversely proportional to the distance in between.

To evaluate the accuracy of each method, we obtained the population means of the reference PC scores and calculated the scaled mean squared difference (MSD) between the reference population means and the corresponding study population means, that is,

$$\text{MSD} = \frac{\sum_{q=1}^{Q}\sum_{k=1}^{K}(D_{q,k} - C_{q,k})^2}{\sum_{q=1}^{Q}\sum_{k=1}^{K} C_{q,k}^2},$$

where $C_{q,k}$ and $D_{q,k}$ are population q's reference and study sample means, respectively, for the k^{th} PC.

To determine the proportion of the MSD that is caused by the prediction of the study samples rather than random variations, in each population we randomly selected some reference samples whose number is the same as that of the study samples. Then we calculated the MSD of these selected reference samples as if they are study samples. We repeated this procedure for 100 times to obtain an empirical null distribution of the MSD.

In addition, to directly compare different methods' predicted PC scores, we calculated their pairwise mean squared difference across all the samples and PCs.

For the comparison of computation cost, we applied each method 10 times for each experimental setting and obtained the mean of the study runtimes. Note that the study runtime did not include the time for running the reference sample PCA, reading and writing files, or predicting the population membership of the study samples from their predicted PC scores. For SP, AP, and OADP, we used our FRAPOSA software, which implements the methods using Python. For ADP, we used the TRACE software by Wang et al. [2015]. All the programs were run on a single-core CPU.

2.2.4 UK Biobank data analysis

We applied the proposed and existing methods to the UK Biobank data [Sudlow et al., 2015, Bycroft et al., 2018], which contained the genotypes of 488,366

individuals in the United Kingdom. The 1000 Genomes Project data served as our reference panel [Consortium et al., 2015]. We used the Phase 3 release of the 1000 Genomes data, which contained 84.4 million variants and 2,504 individuals from five super-populations: Africans, admixed Americans, East Asians, Europeans, and South Asians (Table 2.2). These populations were further divided into 26 sub-populations. By using the family structure information provided by the 1000 Genomes Project, we excluded all the individuals with at least one parent that was included in the data set, which resulted in 2,492 individuals for the reference panel. Furthermore, we intersected the 147,604 high-quality genotyped SNPs in the UK Biobank data with the 1000 Genomes SNPs, which gave us 145,282 SNPs in common.

After predicting PC scores, we further predicted the ancestry membership by using the 20-nearest-neighbors method, as in the simulation studies (Section 2.2.3). If a study sample's highest voted population had received less than or equal to 0.875 of the total weighted votes, we classified it as an admixed individual. Then, we investigated the finer-scale ancestry structures using the population-specific reference samples. For example, we used the 498 European 1000 Genomes samples, which consisted of Iberians, Britons, Finns, Toscani, and Utah resident with Northern and Western European ancestry, as the reference panel to predict the sub-population membership of the UK Biobank samples that had been predicted to be Europeans.

Since ADP was very slow for such large reference and study sample sizes, we did not apply ADP to all the study samples. Instead, we randomly selected 5000 study samples and used them to compare the performance of ADP against the other methods. The other three methods, SP, AP, and OADP, were applied to all the study samples. As in the simulation studies, accuracy was measured by MSD, and runtime excluded the time for PCA on the reference samples.

2.3 Results

2.3.1 Simulation studies

We applied the proposed (AP and OADP) and the existing methods (SP and ADP) to the grid-simulated genotypes with the reference sample sizes ranged from 1000 to 3000. Figure 2.1 shows the PC scores calculated by using 1000 reference samples. It shows that PCA has successfully clustered four different groups. As expected, SP showed systematic shrinkage, but AP, OADP, and ADP did not show the bias and had very similar predicted PC scores (Figure 2.2). As the reference sample size increased, the bias in SP was reduced, but it was still visible even when the reference sample size was 3000 (Figures A.1 and A.2).

Moreover, SP's MSD was more than 10 times higher than those of AP, OADP, and ADP when the reference sample size was 1000. SP's MSD was reduced as the number of reference samples increased, but even when the reference sample size was 3000, the MSD of SP was still at least 4 times higher than that of the other methods, which indicated a higher magnitude of shrinkage for SP. See Figure 2.3 and Table A.1. Among the proposed approaches, OADP generally had the smallest MSD.

When compared to the empirical null distribution, SP's MSD exceeded the mean of the empirical null distribution by 33 to 172 SDs. In comparison, the MSDs of AP, OADP, and ADP were only 6 to 12 SDs away from the mean of the empirical null distribution when the reference size was 1000, and 0.2 to 5 SDs away when the reference size was 1500 or greater. These observations indicated that the differences in MSD across different methods were mostly due to prediction error.

Figure 2.3 and Table A.1 report the computation time. For all the simulation settings, the runtime of ADP greatly exceeded those of the other methods and increased faster than linearly with the number of reference samples. In comparison, the runtime of of OADP only grew slightly, and the runtimes of AP and SP remained

almost unchanged, as the reference size increased. These observations were consistent with the $\mathcal{O}[n^3]$, $\mathcal{O}[n]$, $\mathcal{O}[1]$, and $\mathcal{O}[1]$ computation complexity of ADP, OADP, and AP and SP, respectively, with respect to reference size (for fixed data dimension and study size. See Table 2.1). When the reference size reached 3000, ADP's runtime for predicting 200 study samples was 3,369 seconds, which was more than 200 times of OADP's (16 seconds) and more than 16,000 times of AP's (0.20 seconds). In a study of 500,000 samples with a reference size of 3000, the projected computation time of ADP would be 2,340 CPU hours (97 CPU days), while OADP and AP would only require 11 and 0.14 CPU hours, respectively.

2.3.2 UK Biobank data analysis

To identify the ancestry structure of the UK Biobank data, we applied the proposed and existing approaches by using the 1000 Genomes data as references. The UK Biobank data contained 488,366 samples collected over multiple centers in the United Kingdom. The 2,492 independent samples from the 1000 Genomes data were used as the reference set. Sample sizes of the super-populations and sub-populations are given in Table 2.2. The predicted super-populations (by OADP) of the UK Biobank samples are shown in Table 2.3. Since ADP was computationally too expensive, we only applied ADP to 5000 randomly selected samples for method comparison. All the other methods were applied to all the 488,366 samples.

Figure 2.4 shows the top 4 PC scores of all the UK Biobank samples as predicted by SP, AP, and OADP. The super-populations (Africans, admixed Americans, East Asians, Europeans, and South Asians) were distinguishable by all these three methods. Even SP did not show strong shrinkage. The shrinkage factors for the top 4 PCs predicted by AP were 0.99, 0.99, 0.96, and 0.94.

To compare the PC score prediction of SP, AP, and OADP against ADP's, we applied each method to the 5000 randomly selected UK Biobank samples. The PC

scores are plotted in Figure A.3. All the methods gave similar predicted PC scores (Figure A.4), and the MSDs were also very close (Table 2.4).

Next, among the 461,807 UK Biobank samples that had been predicted to be Europeans by OADP, we further estimated their sub-population memberships. For the reference panel, we used the 498 European samples in the 1000 Genomes data, where each of them was Iberian, British, Finnish, Toscani, or a Utah resident with Northern and Western European ancestry. Each European UK Biobank study sample was predicted to be one of these sub-populations by using the 20-nearest-neighbor method on the PC scores in the same way as in the analysis of the global samples, except that the possibility of being identified as an admixed sample was not included. The top 4 reference and study PC scores of the European samples are shown in Figure 2.5. Compared to AP and OADP, SP clearly showed shrinkage in PC1 to PC4. The shrinkage factors for the top 4 PCs predicted by AP were 0.70, 0.40, 0.21, and 0.14.

Figure A.5 shows the PC scores predicted by SP, AP, OADP, and ADP of the 5000 randomly selected European UK Biobank study samples. The comparison of the PC scores is illustrated in Figure A.6. Compared to the other methods, PC scores predicted by SP were much closer to NULL. Unlike in the analysis of the global samples, SP had a much higher MSD between the population means for the European samples (Table 2.4).

In addition, we identified the African, East Asian, admixed American, South Asian, and admixed samples by using the OADP-predicted PC scores based on the global 1000 Genomes reference samples. Then SP, AP, and OADP were used to predict their finer-scale PC scores and ancestry memberships. The results are shown in Figures A.7 to A.11.

The computation cost is shown in Table 2.4. For the analysis of all the 488,366 UK Biobank samples, SP and AP both took 0.82 CPU hour, while OADP took 21

CPU hour. For ADP, because of its high computation cost, we only ran it on 500 study samples and then scaled its runtime to all the 488,366 samples. The projected runtime for ADP was 1682 hours, which was almost 80 times higher than OADP and 2000 times higher than SP and AP. For the computation cost of the analysis of the European samples, SP and AP both took 0.69 hour, and OADP took 17.75 hours. Because there were only 498 European reference samples, ADP was estimated to cost only 58.93 CPU hours when applied to the European samples.

2.4 Discussion

In this paper, we have compared two existing (SP and ADP) and two novel methods (AP and OADP) of predicting PC scores for the purpose of predicting population structure. The computation complexity calculation shows that our methods greatly exceed the speed of the existing ADP method when the reference sample size is large. Moreover, AP improves the accuracy of SP by adjusting for the shrinkage bias, which is asymptotically estimated from random matrix theory. Our simulation study and the analysis of the UK Biobank data have empirically demonstrate the efficiency and unbiasedness of our methods. AP and OADP have been shown to be 16 times to 16,000 times faster than ADP. They have also successfully separated the sub-populations in the UK Biobank data when SP shrinks most of the study samples toward NULL and is unable to cluster them.

In our simulation studies, we set the number of markers to 100,000. In studies focusing on specific regions in the genome, such as exome-chip or exome-sequencing studies, the number of variants available for ancestry prediction can be substantially smaller. To investigate the performance of the methods in such situations, we reduced the number of variants from 100,000 to 50,000 and 10,000. Figures A.12 and A.13 show that when the reference size was 1000, reducing the number of variants caused all the samples, reference and study, to be close to NULL. This would cause difficulties for

predicting the population membership of the study samples, as there were more study samples on the boundaries of the reference populations. ADP, OADP, and AP could still separate most of the samples from different populations. In comparison, SP's study PC scores clustered much more closely around NULL, although their population memberships were mostly distinguishable. On the other hand, Figure A.14 shows that the MSD remained almost unchanged as the number of variants was reduced. This was due to the fact that MSD was scaled with the scale of the reference PC scores and therefore would change little when the reference and study PC scores shrank by approximately the same magnitude.

In the UK Biobank data analysis, we observed that the PC scores predicted by SP had shrunken much more in the analysis of the European samples than in the analysis of the global samples. This difference could be caused by the sample size difference and the population diversity difference. To further investigate this issue, we randomly selected 498 global 1000 Genomes reference samples to analyze the 5000 randomly selected global UK Biobank samples. With the reference size the same as the European samples, the 5000 global samples' PC scores shrank more than when using all the 2,492 global reference samples, as shown in Figures A.15 and A.16. The shrinkage factors for the top 4 PCs predicted by AP were 0.96, 0.93, 0.80, and 0.70, which indicated stronger shrinkage effect compared to the analysis using all the 2,492 reference samples, especially in PC3 and PC4, though the shrinkage was not as strong as the shrinkage in the analysis of the European samples. For differences in population diversity, the global samples in the 1000 Genomes data had a fixation statistic F_{st} [Weir and Cockerham, 1984] of 0.087, while that of the Europeans samples was 0.005. Similarly, the proportion of the total variation explained by top 4 PCs was 0.090 for the global samples and 0.015 for the European UK Biobank samples. Both population diversity statistics show that the European populations did not differ as much as the global populations. We conclude that both the reference size difference

and the population diversity difference contributed to SP's large shrinkage in the European sample analysis as compared to the global sample analysis.

Throughout the paper, we estimate the ancestry membership of the study samples by predicting their PC scores with a reference panel. An alternative method would be combining the reference samples with the study samples and applying PCA to the combined data. However, a major drawback of this alternative is that when most of the study samples belong to one population, this population would dominate the analysis and cause inaccurate PC score prediction for samples in other populations. To illustrate this, we combined the European 1000 Genomes samples with the European UK Biobank samples and applied the FastPCA algorithm [Galinsky et al., 2016] to the combined data. The PC scores were then used to estimate the ancestry membership through the 20-nearest-neighbor method, as described in Section 2.2.3. Figure A.17 shows the PC scores estimated by FastPCA, and Table A.2 compares the ancestry membership estimated by FastPCA and OADP. The two methods estimated very similar numbers of samples to be British or Utah residents of Northern and Western European ancestry, which is what we would expect since the study data was dominated by these two populations. However, the two methods gave very different results for the other three European populations. In the most extreme case, the difference in the number of Finnish samples was more than ten-fold between the two methods' predictions. We note that, due to the lack of the fine-scale ancestry information, we cannot confirm that our method have provided more accurate results. However, considering the unsupervised nature of the alternative approach, it is reasonable to assume that the alternative approach would be less accurate. In addition, the alternative approach does not allow to compare samples in different studies, so it cannot be used for the sample matching in integrative analysis [Zhan et al., 2013].

An interesting phenomenon we have observed is that in most cases of the simulation studies and the UK Biobank analysis, OADP outperformed ADP in terms of

prediction accuracy as measured by MSD, even though OADP is an approximation method of ADP. One possible explanation for this phenomenon is that OADP only uses the first 16 PCs to update the top 8 PCs. OADP sacrifices the information of the lower-rank PCs in order to gain computation speed, but this might turn out to be an advantage for OADP's prediction accuracy, since it makes this method less vulnerable to outliers in the lower-rank PCs.

We have also noticed a limitation of AP. While the computation complexity and memory usage of SP can be further reduced by using some truncated SVD algorithm (such as the randomized SVD algorithm by Halko et al. [2011]) to compute the SVD for only the top K PCs of the reference matrix, AP requires all the eigenvalues and thus a full SVD or eigendecomposition of the reference matrix. This becomes especially important when the reference set is extremely large. In contrast, OADP needs only the top few singular values and vectors, which can be computed by randomized approaches even for large reference sets.

In addition, for concerns about relatedness in the samples, the proposed methods AP and OADP can in general be applied to high-dimensional genotype data as long as the reference samples are all unrelated. Relatedness among study samples would not affect PC score prediction accuracy.

As the cost of genotyping continues to decrease, larger genotype data sets will become available. High-dimensional large-sized data will be essential for identifying and adjusting for fine-scale population structure in GWAS, but they also creates a demand for computationally efficient algorithms. When the size of the reference samples increases, existing methods such as ADP would become impractical to use. But our methods will continue to operate within a reasonable computation time frame without losing accuracy and serve as useful tools for genetic studies. The SP, AP, OADP, and ADP methods have been implemented in the open source software FRAPOSA (`github.com/daviddaiweizhang/fraposa`).

Table 2.1: Computation complexity of SP, AP, ADP, and OADP.

Method	Reference Complexity	Study Complexity
SP	$\mathcal{O}[n^2p]$	$\mathcal{O}[mKp]$
AP	$\mathcal{O}[n^2p]$	$\mathcal{O}[mKp]$
ADP	$\mathcal{O}[n^2p]$	$\mathcal{O}[m(np+n^3)]$
OADP	$\mathcal{O}[n^2p]$	$\mathcal{O}[m(K''p+K'^2n)]$

Table 2.2: Super-population and sub-population sizes in the 1000 Genomes.

Super-Popu.	Size	Sub-Popu.	Size
Africans	657	ACB (African Caribbeans in Barbados)	96
		ASW (Americans of Afr. Ancestry in SW. USA)	61
		ESN (Esan in Nigeria)	99
		GWD (Gambian in W. Divisions in the Gambia)	113
		LWK (Luhya in Webuye, Kenya)	97
		MSL (Mende in Sierra Leone)	84
		YRI (Yoruba in Ibadan, Nigeria)	107
Americans	347	CLM (Colombians from Medellin, Colombia)	94
		MXL (Mexican Ancestry from Los Angeles USA)	64
		PEL (Peruvians from Lima, Peru)	85
		PUR (Puerto Ricans from Puerto Rico)	104
East Asians	503	CDX (Chinese Dai in Xishuangbanna, China)	92
		CHB (Han Chinese in Beijing, China)	103
		CHS (Southern Han Chinese)	105
		JPT (Japanese in Tokyo, Japan)	104
		KHV (Kinh in Ho Chi Minh City, Vietnam)	99
Europeans	498	CEU (Utah Residents with N. & W. Eur. Ancestry)	95
		FIN (Finnish in Finland)	99
		GBR (British in England and Scotland)	90
		IBS (Iberian Population in Spain)	107
		TSI (Toscani in Italia)	107
South Asians	487	BEB (Bengali from Bangladesh)	86
		GIH (Gujarati Indian from Houston, Texas)	102
		ITU (Indian Telugu from the UK)	102
		PJL (Punjabi from Lahore, Pakistan)	96
		STU (Sri Lankan Tamil from the UK)	101
Total	2492		2492

Note: The Americans are described as "admixed Americans" by the 1000 Genomes Project.

Table 2.3: Population memberships of the UK Biobank samples as predicted by OADP.

Predicted Population	Size
Africans	8169
Admixed Americans	2149
East Asians	2569
Europeans	461807
South Asians	10250
Admixed	3422
Total	488366

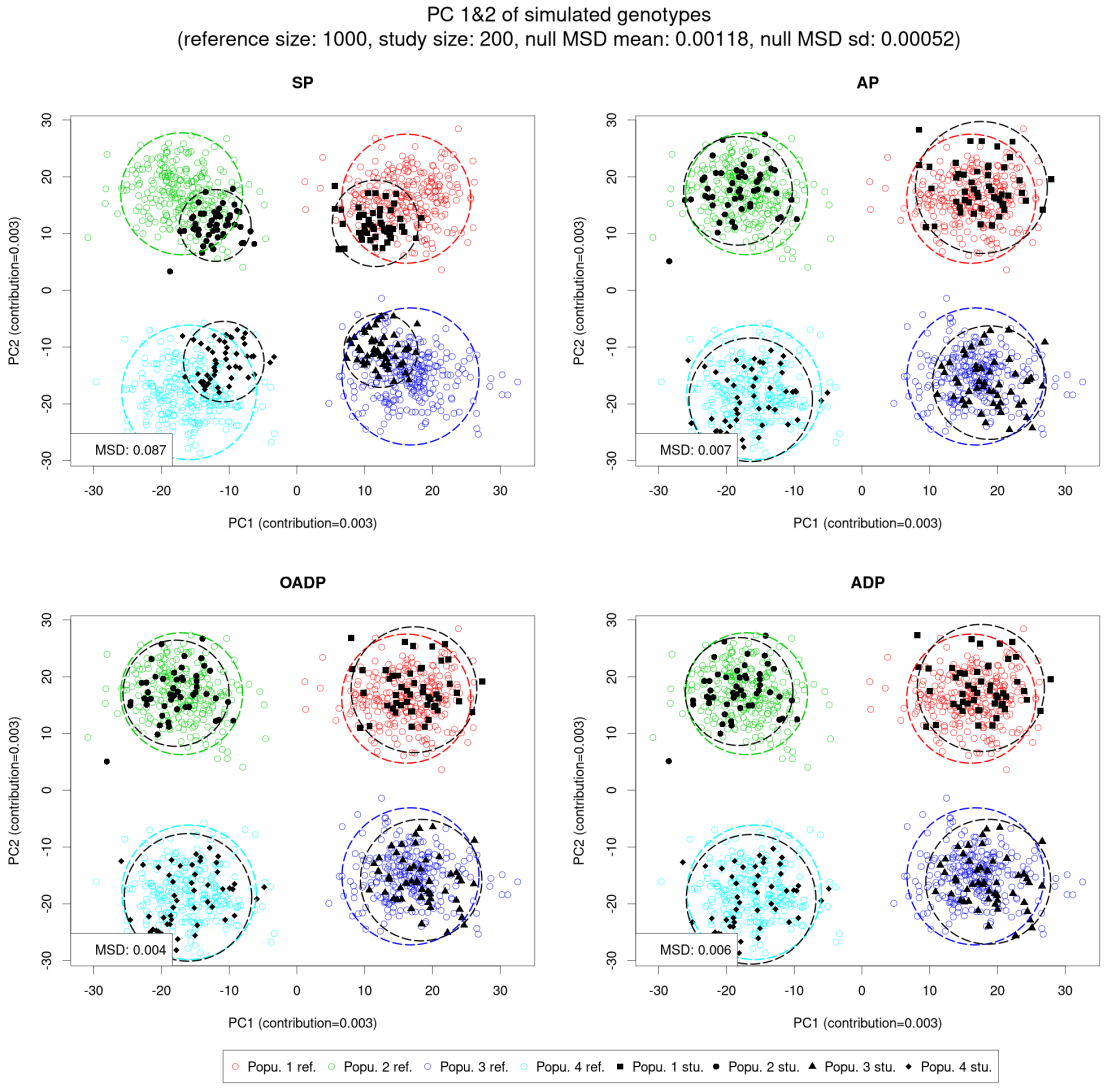
Note: Admixed samples are defined to be those whose highest vote is 0.875 or less of the total weighted votes, as determined by the 20-nearest-neighbor method.

Table 2.4: Estimated runtimes and MSDs of SP, AP, OADP, and ADP for the UK Biobank data analysis.

	Runtime (hr)		MSD	
Population	Global	European	Global	European
Ref. size	2492	498	2492	498
Study size	488,366	461,807	5000	5000
SP	0.82	0.69	0.156	0.360
AP	0.82	0.69	0.156	0.107
OADP	20.71	17.75	0.156	0.100
ADP	*1628.22	*58.93	0.153	0.102

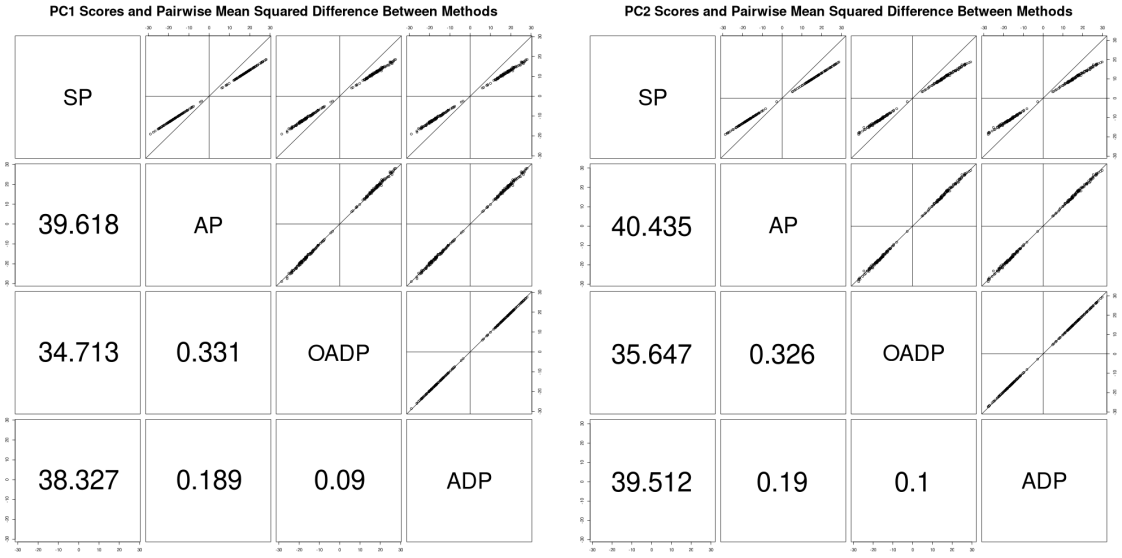
Note: Runtime was estimated from the 5000 randomly selected study samples.

Figure 2.1: PC scores of the simulated genotypes as predicted by SP, AP, OADP, and ADP.



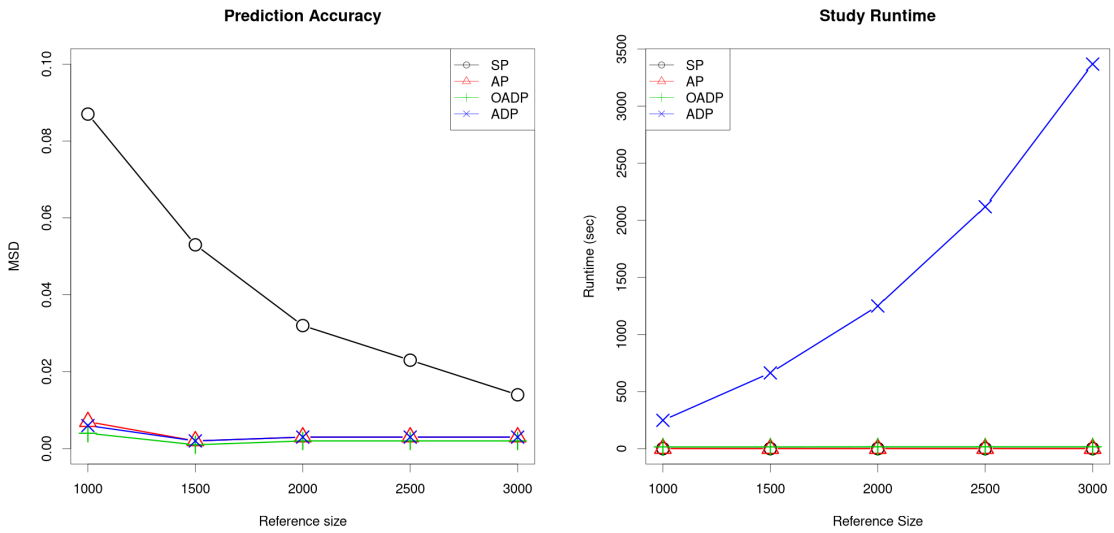
Note: The reference size was 1000. In each of the 4 populations, there were 250 reference samples and 50 study samples, where each sample contained 100,000 variants. The colored/black circle is centered at the reference/study sample mean and encloses 90% of the reference/study samples. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

Figure 2.2: Pairwise comparison of the simulated genotypes' PC scores as predicted by SP, AP, OADP, and ADP.



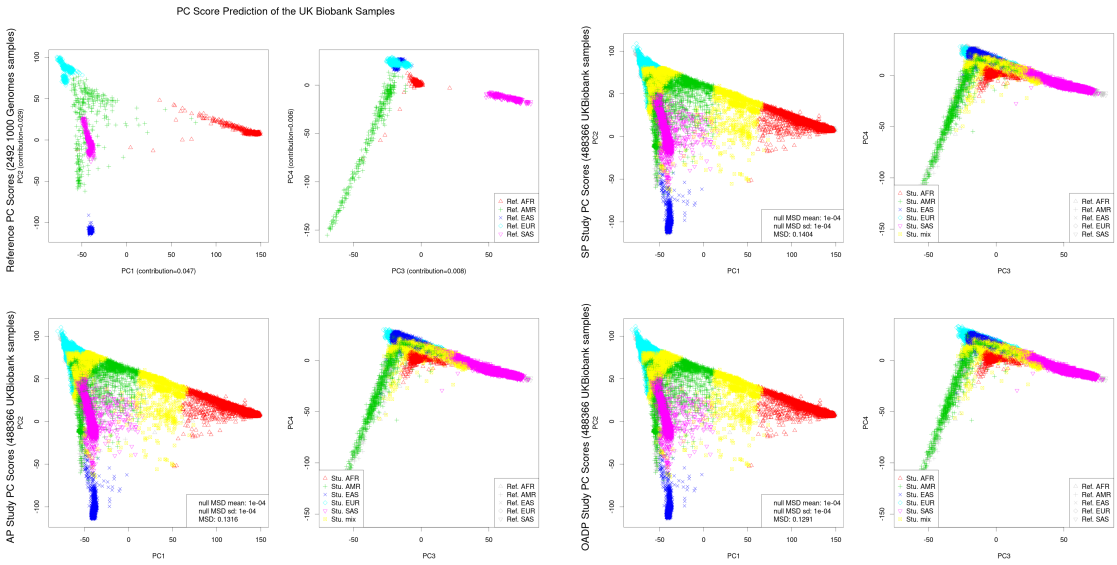
Note: The reference and study sizes were 1000 and 200, respectively. Each sample contained 100,000 variants. The upper panels show the PC scores, while the lower panels show the pairwise mean squared difference between the methods.

Figure 2.3: Comparison of the accuracy and runtimes of SP, AP, OADP, and ADP in the simulated data.



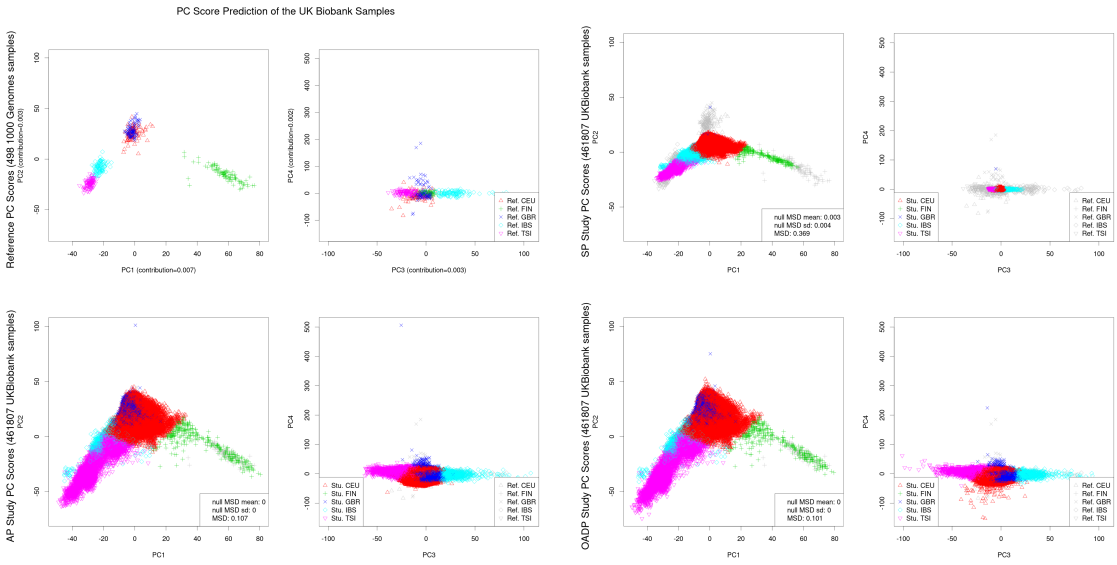
Note: Accuracy was measured by the MSD between the population means of the reference samples and the corresponding population means of the study samples, scaled by the average distance between the reference population means and the reference global mean. The runtimes only included the time for analyzing the study samples, and the computation cost for analyzing the reference samples was ignored. Each experimental setting's runtime was the average of 10 replications. A single-core CPU was used for all the cases. The study sample size was 200, and there were 100,000 variants. Only the top 2 PCs were calculated.

Figure 2.4: PC scores of all the UK Biobank samples, as predicted by SP, AP, and OADP.



Note: The reference panel consisted of all the 2,492 samples in the 1000 Genomes data. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD was scaled with the average distance between the reference population means and the reference global mean. The shrinkage factors for the top 4 PCs predicted by AP were 0.99, 0.99, 0.96, and 0.94. The F_{st} statistic was 0.10, and the total variation contributed from the top 4 PCs was 0.09.

Figure 2.5: PC scores of the European UK Biobank samples, as predicted by SP, AP, and OADP.



Note: European samples were identified by OADP using global 1000 Genome reference samples. The reference panel consisted of all the 498 European 1000 Genomes samples. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD was scaled with the average distance between the reference population means and the reference global mean. The shrinkage factors for the top 4 PCs predicted by AP were 0.70, 0.40, 0.21, and 0.14. The F_{st} statistic was 0.01, and the total variation contributed from the top 4 PCs was 0.02.

CHAPTER III

Image-on-Scalar Regression via Deep Neural Networks

3.1 Introduction

3.1.1 Background

With the rapid growth of medical imaging studies, it has become a scientific task of great importance to discover the patterns of the influence of potential factors on massive imaging data. Primary types of images include T1-weighted magnetic resonance imaging (MRI) data, contrast maps from task-based functional MRI (fMRI), and the local summary statistics of resting fMRI data, such as fractional amplitude of low-frequency fluctuations and weighted degree of network connectedness. A typical imaging dataset contains multiple individuals, with observations from each individual at spatial points (called voxels) in a large set (called a template) of predetermined locations inside a common volume (often three-dimensional). The statistical challenge is to develop a model that delineates the association of voxels or regions of interest (ROI) with a set of covariates of interests, such as demographic information, clinical characteristics, and other non-imaging measurements. We refer to this type of models as image-on-scalar regression models, where the images are regarded as a functional response variable whose mean value depends on a set of scalar predictive variables.

Several major challenges are associated with making inferences on the image-on-scalar regression models for analysis of medical imaging data. First, the spatial dependence between voxels can be highly complex. Due to biological and technological reasons, imaging signals are often contained in contiguous, sharp-edged regions that are sparsely distributed throughout the whole spatial volume [Chan and Shen, 2005, Tabelow et al., 2008, Chumbley and Friston, 2009]. For example, in neuroimaging studies, it has been found that the association with fear is substantially higher in the amygdala than in other brain regions [Whalen et al., 2001]. Second, the signals across individuals can be heterogeneous and might depend on unobserved variables. For example, underlying medical conditions and the psychological state at the time of scanning can affect the outputs of fMRI. Third, the number of individuals is often limited in imaging studies, while the noise level can be rather high due to machine artifacts and imperfect preprocessing, leading to a relatively low signal-to-noise ratio. This limitation makes it particularly difficult to apply traditional machine learning methods to image-on-scalar regression problems, since methods such as deep learning rely on a large sample size to train their highly flexible models. This might partially explain the relatively scant success of applying artificial neural networks to imaging studies with a small sample size, as compared to other fields such as computer vision and natural language processing.

3.1.2 Related work and our contributions

A straightforward method for fitting the image-on-scalar regression is the mass univariate analysis (MUA). This approach fits a general linear model (GLM) at each voxel and obtains voxel-wise test statistics to identify the brain regions that are significantly associated with the covariate of interest, after applying a multiple testing adjustment method such as the Bonferonni correction or false discovery rate control [Benjamini and Yekutieli, 2001]. A major limitation of MUA is that the spatial corre-

lation is not accounted for, which can result in low power for detecting significant brain regions and may potentially increase the false positive rate. To incorporate spatial correlation into MUA, one may smooth the imaging data through a kernel convolution before fitting the GLM. For example, statistical parametric mapping (SPM), utilizes random field theory to make classical inferences [Friston, 2003]. However, performing MUA on these pre-smoothed data can lead to low accuracy and low efficiency in estimating and testing the covariates' effects [Chumbley et al., 2009]. To improve the performance of noise reduction and feature selection, adaptive smoothing methods have been developed for data preprocessing [Yue et al., 2010] and parameter estimation [Polzehl and Spokoiny, 2000, Qiu, 2007], and those methods are especially powerful for detecting delicate patterns such as jump discontinuities.

Instead of modeling each voxel independently, one can consider the observed outcome image intensities over all the voxels along with the corresponding regression coefficients as tensors (i.e. multi-dimensional arrays) and impose certain sparsity structures for model fitting. For example, parsimonious tensor response regression [Li and Zhang, 2017], assumes the response tensor to be sparse after some linear transformation and aims to separate material and immaterial information. Another example is sparse tensor response regression (STOR) [Sun and Li, 2017], which embeds element-wise sparsity and low-rankness on the coefficient tensor and is designed to handle both symmetric and asymmetric responses. For analysis of medical imaging data involving a large number of voxels, this family of models has the difficulty in developing dimension reduction techniques that are both accurate and computationally efficient.

Alternative to treating the voxels as independent points or the indices of multi-dimensional arrays, we can also consider them as discrete grid points of the continuous spatially varying functions. In the image-on-scalar regression problem both the outcome image and the regression coefficients can be considered as the realizations of

spatially varying functions evaluated on voxels. Many methods have been developed for spatial data analysis in environmental health, epidemiology, and ecology [Cressie and Cassie, 1993, Diggle et al., 1998, Gelfand et al., 2003], where the spatially varying functions are typically assumed to be smooth or continuously differentiable up to certain degrees. Motivated by neuroimaging applications, the spatially varying coefficient model (SVCM) [Zhu et al., 2014] has been developed to systematically incorporate both spatial smoothness and jump discontinuities. The SVCM can also identify regions that are significantly associated with the covariates of interest by using a step-wise estimating procedure and the asymptotic Wald test. Recently, Chen et al. [2016] adopted a novel penalty function to detect the significant regions. In contrast, Li et al. [2020] and Yu et al. [2020] use bivariate splines over triangulation (BST) to approximate the coefficient function, while Gu et al. [2014] use spline smoothing to produce simultaneous confidence corridors. From the Bayesian perspective, Shi and Kang [2015] model the spatially varying functions as thresholded multiscale Gaussian processes, and Bussas et al. [2017] handle them as Gaussian processes with isotropic priors.

In the proposed NNISR model, NNs are adopted to approximate the spatially varying coefficient functions of the true effects. NNs are functions composed of multiple layers of linear transformations and nonlinear activation functions. NNs are very flexible to model nonlinear functions of multiple predictor variables as well as the interaction between them, which enables NNs to represent a wide variety of complex functions in various applications. In the recent years, NNs have been successful in artificial intelligence applications, such as visual object detection, natural language processing, and game playing [Goodfellow et al., 2016, LeCun et al., 2015, Silver et al., 2017, Fan et al., 2019]. For biomedical studies, NNs have been applied to drug activity prediction [Ma et al., 2015], brain circuit reconstruction [Helmstaedter et al., 2013], clinical radiology [Chartrand et al., 2017], regulatory genomics [Zou et al., 2019], and

cardiovascular medicine [Krittanawong et al., 2019].

Although deep NNs have numerous successful applications, it remains challenging to study the theoretical properties of NNs and related models. According to the universal approximation theory, any continuous function on a compact set can be approximated by a single layer NN with a sufficiently large number of nodes to an arbitrary degree of accuracy. For single-layer NNs, Mhaskar [1996] showed that the approximation errors can be bound by the number of nodes in the NN. Moreover, Ismailov [2014] proved that by using a specifically constructed activation function, a two-layer NN with a total of $3k + 2$ nodes are sufficient for approximating any k-dimensional multivariate continuous function arbitrarily well. For NNs with multiple hidden layers, Rolnick and Tegmark [2017] derived that the number of nodes required for approximating polynomials is proportional to the input dimension. Furthermore, Rotskoff and Vanden-Eijnden [2018] combined the approximation error with the training error and proved asymptotic properties by treating NNs as interacting particle systems. In addition, although shallower NNs are better understood than deeper NNs, the latter is known to be more efficient for representing functions than shallower NNs, in terms of the total number of nodes required [Telgarsky, 2016, Eldan and Shamir, 2016].

The expressiveness of single- and multi-layer NNs makes them promising estimators for nonparametric regression models. Many theoretical results have been developed for nonparametric regression with single-layer NNs. Their consistency, for example, was studied in Mielniczuk and Tyrcha [1993]. Furthermore, the rate of convergence (of the L_2 risk) could be derived by imposing restrictions on the regression function. For example, when the regression function has finite first moment in its Fourier representation, Barron [1991, 1993, 1994] provided a rate of convergence of $n^{-1/2}$ multiplied by a logarithmic term, where n is the sample size. Moreover, for p-smooth regression functions, it has been shown that the minimax rate of conver-

gence for any estimator is $n^{-\frac{2p}{2p+k}}$ [Stone, 1982], where k is the input dimension. In the case of NN estimators, McCaffrey and Gallant [1994] derived a sub-optimal rate of $n^{-\frac{2p}{2p+k+5}}$ for single-layer NNs, and Kohler and Krzyżak [2017] showed the minimax rate $n^{-\frac{2p}{2p+k'}}$ for two-layer NNs, where $k' \leq k$ represents the sparsity of the regression function. More recently, this sparse minimax rate, with the inclusion of a logarithmic term, was extended in Bauer et al. [2019] to multi-layer NNs with smooth activation functions, and Schmidt-Hieber et al. [2020] proved a similar result for NNs with rectified linear unit (ReLU) activation functions. From the Bayesian perspective, Polson and Rocková [2018] proved that the same convergence rate holds for the posterior probability concentration of deep Bayesian ReLU NNs with spike-and-slab priors on the weight parameters. For p-smooth regression functions, Liu et al. [2019] further improves the upper error bound by eliminating the logarithmic term.

In addition to nonparametric regression, the universal approximation ability of NNs has also motivated their usage in nonlinear variable selection. Traditional variable selection methods such as lasso [Tibshirani, 1996] and elastic nets [Zou and Hastie, 2005] have been successful in high-dimensional data regression. However, many of the existing variable selection methods are limited to linear models, which could be inadequate for capturing the nonlinear relations in complex systems, such as those in biological mechanisms [Janson, 2012]. To address this difficulty, NN-based methods have been proposed to model nonlinearity and detect interactions in model selections problems. Feng and Simon [2017] imposed a group lasso penalty on the weights of the first layer and showed the convergence of the weights for irrelevant features. Chen et al. [2020] proposed a NN model that consists a selection layer and and multiple approximation layers and provided a greedy algorithm for estimating the selection and approximation parameters. From a Bayesian perspective, Liang et al. [2018] treated variable selection as a sub-problem of NN structure selection and developed a Bayesian NN model in which a truncated binomial prior distribution is

assigned to on the number of non-zero weights in the NN to assure posterior selection consistency.

In this work, we propose a novel NN-based ISR model (NNISR) that takes advantage of NNs' universal approximation capability to estimate the associations between the images and the covariates Our model takes the FDA framework and constructs the spatially varying coefficient functions of the main effects, individual deviations, and noise variances through multi-layer NNs. The NNs are applied across voxels instead of across images, so that the spatial patterns are approximated by NNs but the associations between images and covariates are still assumed to be linear. This model structure relies on the high dimensionality of response images to provide ample training observations for the NNs esp At the same time, the imposed linear image-covariate relations ensures interpretability of the main effects and estimation efficiency when the number of images is small. For fitting the model, we provide an estimation algorithm based on stochastic gradient descent [Bottou, 2010] and hard-thresholding. Theoretically, we establish selection consistency by linking the convergence rate of the NNISR estimator to the convergence rate of the general NN regression model that has same architecture, where the sample size in the latter corresponds to the *product* of the number of voxels and the number of images in the former. The theoretical error bounds show that the estimation accuracy could be improved by increasing the number of voxels, even when the number of images remains fixed. We compare the performance of NNISR with existing methods through extensive simulation studies with complex spatial pattern designs skewed noise distributions. The advantage of NNISR is the most distinct in the settings with small image numbers and high imaging resolutions. The efficacy of NNISR is further evaluated by the analyses of brain fMRI data in the Autism Brain Imaging Data Exchange (ABIDE) [Di Martino et al., 2014] and the Adolescent Brain Cognitive Development (ABCD) study [Casey et al., 2018]. We conduct cross validation across experimental sites to evaluate the estima-

tion accuracy and compare the selected regions by each method to demonstrate their selection characteristics.

The remainder of the manuscript is organized as follows. We formulate the NNISR model and present the model fitting algorithm in Section 3.2.2. The theoretical properties of NNISR is established in Section 3.3. Next, we evaluate NNISR against existing methods through extensive simulation studies in Section 3.4 and apply them to brain fMRI data in Section 3.5. Finally, we conclude with a discussion in Section 3.6.

3.2 Image-on-Scalar Regression via Deep Neural Networks

3.2.1 Model specification

Suppose the imaging measurements are collected in a compact space $\mathcal{D} \subset \mathbb{R}^K$ along with Q covariate variables from M individuals. For each individual $m \in \{1, \ldots, M\}$, let $y_m(\boldsymbol{d}) \in \mathbb{R}$ be image measurements at each spatial location $\boldsymbol{d} \in \mathcal{D}$, which can be considered as a function of $\boldsymbol{d}$ in domain $\mathcal{D}$, and let $\boldsymbol{x}_m \in \mathbb{R}^Q$ be the covariate vector. The proposed image-on-scalar regression model is

$$y_m(\boldsymbol{d}) = \boldsymbol{x}_m^\top \boldsymbol{\beta}(\boldsymbol{d}) + \alpha_m(\boldsymbol{d}) + \epsilon_m(\boldsymbol{d}), \tag{3.1}$$

where $\boldsymbol{\beta}(\boldsymbol{d}) = \{\beta_1(\boldsymbol{d}), \ldots, \beta_Q(\boldsymbol{d})\}^\top \in \mathbb{R}^K$ is a vector of Q coefficient functions of $\boldsymbol{d}$, representing the main effects of covariates $\boldsymbol{x}_m$; $\alpha_m(\boldsymbol{d}) \in \mathbb{R}$ is a function of $\boldsymbol{d}$, characterizing the variation of the mth individual from the main effects $\boldsymbol{x}_m^\top \boldsymbol{\beta}(\boldsymbol{d})$; and $\epsilon_m(\boldsymbol{d})$ is the random noise at location $\boldsymbol{d}$, reflecting the imaging measurement errors. We assume that $\mathrm{E}\{\epsilon_m(\boldsymbol{d})\} = 0$, $\mathrm{Var}\{\epsilon_m(\boldsymbol{d})\} = \sigma^2(\boldsymbol{d})$, and $\epsilon_m(\boldsymbol{d})$ is independent from $\epsilon_{m'}(\boldsymbol{d}')$ for $m \neq m'$ or $\boldsymbol{d} \neq \boldsymbol{d}'$. Note that the noise variance $\sigma^2(\boldsymbol{d}) > 0$ is a function of $\boldsymbol{d}$ which has a flexibility to capture the spatial heterogeneity in the variation of measurement errors.

Of note, the individual effects $\boldsymbol{\alpha}(\boldsymbol{d}) = \{\alpha_1(\boldsymbol{d}), \ldots, \alpha_m(\boldsymbol{d})\}^\top$ play a similar role as

the functional random effects in SVCM [Zhu et al., 2014, Li et al., 2020], which are assumed to be identical copies of a stochastic process. However, in (3.1), for each individual, $\alpha_m(\boldsymbol{d})$ is a deterministic coefficient function which is unknown and needs to be estimated. To ensure the model identifiability and interpretability, we make a few assumptions on main effects, individual effects and noise variances. We list the key concepts here; see Section 3.3 for comprehensive and rigorous definitions.

1. (Piecewise smoothness): Functions $\{\beta_q(\boldsymbol{d})\}_{q=1}^{Q}, \{\alpha_m(\boldsymbol{d})\}_{m=1}^{M}$ and $\sigma^2(\boldsymbol{d})$ are piecewise smooth with a finite number of discontinuous jumps.

2. (Sparsity): For each covariate q, there exists a large region on which the main effects are equal to zero.

3. (Constant lower bound of nonzero effects): For each covariate q, the absolute values of the nonzero main effects have a positive constant lower bound.

In brain imaging application, the piecewise smoothness is introduced to model the spatial dependence of brain signals. The discontinuous jumps may reflect the brain activity differences among different types of brain tissues. The sparsity assumes the brain activation region is relatively small. The constant lower bound of nonzero effects models the sharp edges of brain activation regions.

To satisfy the aforementioned three assumptions, we adopt feed-forward neural networks (NNs) to model the spatially varying functions in (3.1). We define $\aleph(\boldsymbol{d} \mid \boldsymbol{\theta})$ as a general G-layer feed-forward NN with the input dimension R_0, the output dimension R_{G+1}, and the j^{th} layer having R_j hidden units for $j = 1, \ldots, G$, if $\aleph(\boldsymbol{d} \mid \boldsymbol{\theta})$ is a vector-value function of $\boldsymbol{d} \in \mathbb{R}^{R_0}$ and taking values in $\mathbb{R}^{R_{G+1}}$, which has the form

$$\aleph(\boldsymbol{d} \mid \boldsymbol{\theta}) = \boldsymbol{W}_G \boldsymbol{\phi}_G \left\{ \cdots \boldsymbol{W}_1 \boldsymbol{\phi}_1 \left(\boldsymbol{W}_0 \boldsymbol{d} + \boldsymbol{b}_0 \right) + \boldsymbol{b}_1 \cdots \right\} + \boldsymbol{b}_G,$$

where $\boldsymbol{W}_j \in \mathbb{R}^{R_{j+1} \times R_j}$ and $\boldsymbol{b}_j \in \mathbb{R}^{R_j}$ are the weight and bias parameters of the

j^{th} layer respectively, for $j \in \{0, 1, \ldots, G\}$. The activation function $\boldsymbol{\phi}_j(\boldsymbol{x}_j) = \{\phi_j(x_{j1}), \ldots, \phi_j(x_{jR_j})\}^\top$ for $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jR_j})^\top \in \mathbb{R}^{R_j}$, where $\phi_j(x) \in \mathbb{R}$ is a nonlinear function defined on $\mathbb{R}$. Common choices include the sigmoid function, i.e., $\phi_j(x) = \{1 + \exp(x)^{-1}\}^{-1}$ and the rectified linear unit (ReLU) function, i.e., $\phi_j(x) = \max(0, x)$. The parameter set $\boldsymbol{\theta} = \{\boldsymbol{W}_j, \boldsymbol{b}_j\}_{j=0}^G$ is a collection of weight and bias parameters. Thus, the NNs are specified by the network architecture $\{R_j\}_{j=0}^G$, activation functions $\{\phi_j(x)\}_{j=1}^G$, and the parameters $\boldsymbol{\theta}$.

We construct the spatially varying functions of main effects, individual effects and the noise variance in (3.1) by using three different NNs, respectively. In particular, we assume, for $\boldsymbol{d} \in \mathcal{D}$,

$$\begin{aligned} \boldsymbol{\beta}(\boldsymbol{d}) &= \aleph_\beta(\boldsymbol{d} \mid \boldsymbol{\theta}_\beta), \\ \boldsymbol{\alpha}(\boldsymbol{d}) &= \aleph_\alpha(\boldsymbol{d} \mid \boldsymbol{\theta}_\alpha), \\ \log\{\sigma^2(\boldsymbol{d})\} &= \aleph_\sigma(\boldsymbol{d} \mid \boldsymbol{\theta}_\sigma), \end{aligned} \tag{3.2}$$

where the input variables of the three NNs are all the spatial coordinate for $\boldsymbol{d} \in \mathcal{D}$. The output variables of $\aleph_\beta(\boldsymbol{d} \mid \boldsymbol{\theta}_\beta)$, $\aleph_\alpha(\boldsymbol{d} \mid \boldsymbol{\theta}_\alpha)$ and $\aleph_\sigma(\boldsymbol{d} \mid \boldsymbol{\theta}_\sigma)$ are of dimensions Q, M and 1 respectively. There are three key advantages of combining (3.1) and (3.2). First, the NNs have a very large flexibility to capture the complex patterns of the spatially-varying coefficient functions in terms of heterogeneous shapes and adaptive smoothness. Second, we do not need to directly interpret the NN parameters $\boldsymbol{\theta}_\beta$, $\boldsymbol{\theta}_\alpha$ and $\boldsymbol{\theta}_\sigma$ in (3.2); instead we focus on the explicit and straightforward interpretations on the outputs of the NNs, i.e., the spatially-varying functions $\boldsymbol{\beta}(\boldsymbol{d})$, $\boldsymbol{\alpha}(\boldsymbol{d})$ and $\sigma^2(\boldsymbol{d})$ in the ISR model (3.1). Moreover, in brain imaging applications, the spatial coordinates of voxels are considered as the input "data" for the NN models in (3.2). The number of voxels in the observed images becomes to the training sample size for the NN models. As high resolution brain images may contain hundreds of thousands or even

millions of voxels, which provides a sufficiently large sample size to train the deep NNs in (3.2).

3.2.2 Estimation method

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)^\top$ be the $M \times Q$ observed design matrix of covariates. Let $\boldsymbol{y}(\boldsymbol{d}) = \{y_1(\boldsymbol{d}), \ldots, y_M(\boldsymbol{d})\}^\top$ be a vector of imaging measurements at $\boldsymbol{d}$ for the M individuals. Suppose $\boldsymbol{y}(\boldsymbol{d})$ is only observed on V voxels, denoted as $\mathcal{D}_V = \{\boldsymbol{d}_v\}_{v=1}^V \subset \mathcal{D}$. Given the data $\{\boldsymbol{y}(\boldsymbol{d}_v)\}_{v=1}^V$ and $\boldsymbol{X}$, our goal is to estimate the spatially-varying functions $\boldsymbol{\beta}(\boldsymbol{d})$, $\boldsymbol{\alpha}(\boldsymbol{d})$ and $\sigma^2(\boldsymbol{d})$ in (3.1). Combining models (3.1) and (3.2), we convert the problem of interest to fitting NNs from the following model: for $v = 1, \ldots, V$,

$$\begin{aligned} \mathrm{E}\{\boldsymbol{y}(\boldsymbol{d}_v) \mid \boldsymbol{X}\} &= \boldsymbol{X}\aleph_\beta(\boldsymbol{d}_v \mid \boldsymbol{\theta}_\beta) + \aleph_\alpha(\boldsymbol{d}_v \mid \boldsymbol{\theta}_\alpha), \\ \mathrm{Var}\{\boldsymbol{y}(\boldsymbol{d}_v) \mid \boldsymbol{X}\} &= \boldsymbol{I}_M \exp\{\aleph_\sigma(\boldsymbol{d}_v \mid \boldsymbol{\theta}_\sigma)\}, \end{aligned} \tag{3.3}$$

where $\boldsymbol{I}_M$ is an $M \times M$ identity matrix. Our estimation procedure consists of three major steps.

Step One: Main effect estimation. The main effects are estimated by the following procedure.

1. Obtain a naive estimate of the noise variance

$$\tilde{\sigma}_v^2 = M^{-1} \left\| \left[\boldsymbol{I}_M - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top\right] \boldsymbol{y}(\boldsymbol{d}_v) \right\|_2^2$$

for $v = 1, \ldots, V$, which is the mean squared residuals of the MUA estimate.

2. Estimate the weight parameters of the main effect NN

$$\hat{\boldsymbol{\theta}}_\beta = \arg\min_{\boldsymbol{\theta}_\beta} \sum_{v=1}^{V} \left\| \boldsymbol{y}(\boldsymbol{d}_v) - \boldsymbol{X}\aleph_\beta(\boldsymbol{d}_v \mid \boldsymbol{\theta}_\beta) \right\|^2 \tilde{\sigma}_v^{-2} + \lambda \|\aleph(\boldsymbol{d}_v \mid \boldsymbol{\theta}_\beta)\|_1, \tag{3.4}$$

where $\lambda > 0$ is a tuning parameter. The loss function is minimized by mini-batch stochastic gradient descent (SGD) [Bottou et al., 1991, Bottou, 2010, Kingma and Ba, 2014]. Notice that in the context of our NNISR model, a mini-batch of samples in SGD corresponds to a subset of the voxels in $\mathcal{D}_V = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_V\}$.

3. Apply hard thresholding to the output of the main effect NN:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{d}) = \aleph_\beta(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}_\beta) \otimes \mathbb{I}\{|\aleph_\beta(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}_\beta)| > \boldsymbol{\rho}_\eta\} \tag{3.5}$$

where $\mathbb{I}$ is the element-wise indicator function, "$\otimes$" is the element-wise product operator, and $\boldsymbol{\rho}_\eta = (\rho_{\eta,1}, \ldots, \rho_{\eta,q})^\top$ is the covariate-dependent thresholding levels. For each covariate q, $\rho_{\eta,q}$ is determined by the η^{th} quantile of the absolute value of the main effect NN. Note that both hard thresholding and L_1 penalty are applied to the main effect NN to induce sparsity. This approach is similar to the procedure in thresholded LASSO [Zhou, 2010].

Step Two: Individual effect estimation. Obtain the estimate of the individual effects $\hat{\boldsymbol{\alpha}}(\boldsymbol{d}) = \aleph_\alpha(\boldsymbol{d}|\hat{\boldsymbol{\theta}}_\alpha)$ by estimating the weight parameters of the individual effect NN

$$\hat{\boldsymbol{\theta}}_\alpha = \arg\min_{\boldsymbol{\theta}_\alpha} \sum_{v=1}^{V} \left\| \boldsymbol{y}(\boldsymbol{d}_v) - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{d}_v) - \aleph_\alpha(\boldsymbol{d}_v \mid \boldsymbol{\theta}_\alpha) \right\|^2 \tilde{\sigma}_v^{-2}, \tag{3.6}$$

where the loss function is minimized by SGD.

Step Three: Noise variance estimation. Obtain the estimate of the noise variance by the following procedure.

1. Find the mean squared residuals of the NNISR estimate

$$\bar{\sigma}_v^2 = M^{-1} \left\| \boldsymbol{y}(\boldsymbol{d}_v) - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{d}_v) - \hat{\boldsymbol{\alpha}}(\boldsymbol{d}_v) \right\|_2^2$$

   for $v = 1, \ldots, V$.

2. Estimate the noise variance $\hat{\sigma}^2(\boldsymbol{d}) = \aleph_\sigma(\boldsymbol{d}|\hat{\boldsymbol{\theta}}_\sigma)$ by estimating the weight parameters of the noise variance NN:

$$\hat{\boldsymbol{\theta}}_\sigma = \arg\min_{\boldsymbol{\theta}_\sigma} \sum_{v=1}^{V} \left\| \bar{\sigma}_v^2 - \exp\{\aleph_\sigma(\boldsymbol{d}_v \mid \boldsymbol{\theta}_\sigma)\} \right\|_2^2, \tag{3.7}$$

   where the loss function is minimized by SGD.

Model tuning. The L_1 penalty weight λ for the main effects is selected by cross validation on the image recovery error. To reduce the computation cost, the full NNISR estimation procedure can be subsituted with voxel-wise LASSO [Tibshirani, 1996] in the cross validation for tuning λ. For the quantile selection threshold $\boldsymbol{\rho}_\eta$, it can be set equal to the proportion of voxels selected by MUA. The optimization of NN architectures have been discussed in many works [Bergstra et al., 2011, Feurer and Hutter, 2019, Zhang et al., 2019, Benardos and Vosniakos, 2007, Luo et al., 2018]. In the NNISR model, since the input dimension is often small (e.g. 2- or 3-dimensional), the NN architecture does not need to be as large and complex as those commonly used in applications such as computer vision and natural language processing. In our experiments, we found that the NN architecture does not have a major impact on the performance of NNISR (as long as the architecture is not extremely simple), and 4 hidden layers with 64 nodes in each layer was sufficient in every experimental setting.

3.3 Theoretical Properties

In this section, we perform a theoretical analysis of the NNISR estimator. In Lemma III.16, we provide an error bound on the L_2 estimation error of the main effects before thresholding, expressed in terms of the number of images, number of voxels, and number of nodes in the NNs, which is based on the theoretical results for nonparametric regression with NNs [Bauer et al., 2019, Schmidt-Hieber et al., 2020]. The error bound on the main effects provides an error bound on the individual effects (Corollary III.18) and the noise variance (Corollary III.19). Next, we demonstrate the selection consistency and prove the error bound for the L_0 sign error of the sparse main effect estimator (Theorem III.17). Finally, we present the optimal growth rates of the number of NN nodes and the number of images as a function of the number of voxels (Corollaries III.20 and III.21).

We start with the definitions used in our theoretical analysis of NNISR. As a common practice for theoretical work on nonparametric regression, we restrict the class of candidate functions in our nonparametric analysis, following the framework in Bauer et al. [2019] and Schmidt-Hieber et al. [2020]. We require the main effects, individual effects, and noise variance to satisfy piecewise smooth generalized hierarchical interaction models (piecewise smooth GHIMs), as defined in Definitions III.1 to III.3.

Definition III.1 (Hölder smoothness)**.** Let $K \in \mathbb{N}_+$ and $P \in \mathbb{R}_+$ with $P = P' + P''$, where $P' = \lceil P - 1 \rceil$ and $P'' = P - P'$. A function $f : \mathbb{R}^K \to \mathbb{R}$ is said to be smooth with Hölder index P (abbreviated as P-smooth) if there exist $c_{18}, c_{19} \in \mathbb{R}_+$ such that for every $(p_1, \ldots, p_K) \in \mathbb{N}_0^K$ with $\sum_{k=1}^K p_k = P'$, we have:

1. The partial derivative $\frac{\partial^{P'}}{\partial d_1^{p_1} \cdots \partial d_K^{p_K}} f$ exists.

2. For all $\boldsymbol{x} \in \mathbb{R}^K$,

$$\frac{\partial^{P'}}{\partial x_1^{p_1} \cdots \partial x_K^{p_K}} f(\boldsymbol{x}) < c_{18}.$$

3. For all $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^K$,

$$\left| \frac{\partial^{P'}}{\partial x_1^{p_1} \cdots \partial x_K^{p_K}} f(\boldsymbol{x}) - \frac{\partial^{P'}}{\partial x_1^{p_1} \cdots \partial x_K^{p_K}} f(\boldsymbol{x}') \right| \leq c_{19} \|\boldsymbol{x} - \boldsymbol{x}'\|_2^{P''}.$$

Definition III.2 (Smooth generalized hierarchical interaction model (smooth GHIM))**.** Let $K, K' \in \mathbb{N}_+$ and $P \in \mathbb{R}_+$.

1. A function $f : \mathbb{R}^K \to \mathbb{R}$ is called a P-smooth $(K', 0)$-GHIM if there exist $\boldsymbol{a}_k \in \mathbb{R}^K$ $(k = 1, \ldots, K')$ and $g : \mathbb{R}^K \to \mathbb{R}$ such that

$$f(\boldsymbol{x}) = g(\boldsymbol{a}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{a}_{K'}^\top \boldsymbol{x}),$$

   where g is P-smooth.

2. Let $l \in \mathbb{N}_+$. A function $f : \mathbb{R}^K \to \mathbb{R}$ is called a P-smooth (K', l)-GHIM if there exist $\bar{R} \in \mathbb{N}_+$, $g_r : \mathbb{R}^{K'} \to \mathbb{R}$ $(r = 1, \ldots, \bar{R})$, $h_{r,k} : \mathbb{R}^K \to \mathbb{R}$ $(r = 1, \ldots, \bar{R};\ k = 1, \ldots, K')$ such that

$$f(\boldsymbol{x}) = \sum_{r=1}^{\bar{R}} g_r(h_{r,1}(\boldsymbol{x}), \ldots, h_{r,K'}(\boldsymbol{x})),$$

   where $g_r : \mathbb{R}^{K'} \to \mathbb{R}$ $(r = 1, \ldots, \bar{R})$ and $h_{r,k} : \mathbb{R}^K \to \mathbb{R}$ $(r = 1, \ldots, \bar{R};\ k = 1, \ldots, K')$ are P-smooth $(K', l-1)$-GHIMs.

Definition III.3 (Piecewise smooth GHIM)**.** Let $K, K' \in \mathbb{N}_+$ and $P \in \mathbb{R}_+$.

1. A function $f : \mathbb{R}^K \to \mathbb{R}$ is called a J-piecewise P-smooth $(K', 0)$-GHIM if there exist $g : \mathbb{R}^K \to \mathbb{R}$, $\boldsymbol{a}_k \in \mathbb{R}^K$ $(k = 1, \ldots, K')$ and a $\Omega \subset \mathbb{R}^K$ such that

$$f(\boldsymbol{x}) = g(\boldsymbol{a}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{a}_{K'}^\top \boldsymbol{x}) \cdot \mathbb{I}_\Omega(\boldsymbol{x})$$

   where g is P-smooth, and Ω is a J-side K-dimensional polytope.

2. Let $l \in \mathbb{N}_+$. A function $f : \mathbb{R}^K \to \mathbb{R}$ is called a J-piecewise P-smooth (K', l)-GHIM if there exist $g_r : \mathbb{R}^K \to \mathbb{R}$ $(r = 1, \ldots, \bar{R})$ and $\Omega_r \subset \mathbb{R}^K$ $(r = 1, \ldots, \bar{R})$ such that

$$f(\boldsymbol{x}) = \sum_{r=1}^{\bar{R}} g_r(h_{r,1}(\boldsymbol{x}), \ldots, h_{r,K'}(\boldsymbol{x})) \cdot \mathbb{I}_{\Omega_r}(\boldsymbol{x}),$$

where $g_r : \mathbb{R}^{K'} \to \mathbb{R}$ $(r = 1, \ldots, \bar{R})$ and $h_{r,k} : \mathbb{R}^K \to \mathbb{R}$ $(r = 1, \ldots, \bar{R};\ k = 1, \ldots, K')$ are P-smooth $(K', l-1)$-GHIMs, and $\Omega_r \subset \mathbb{R}^K$ $(r = 1, \ldots, \bar{R})$ are J-side K-dimensional polytopes.

Definition III.1 provides a general definition of function smoothness by bounding the derivatives up to a certain order. Definition III.2 describes a class of functions, GHIMs, that are constructed by composition and summation of multiple layers of smooth functions of the type defined in Definition III.1. GHIMs cover a wide range of functions. Both Definition III.1 and Definition III.2 are based on Bauer et al. [2019]. In Definition III.3, we generalize Definition III.2 by allowing the element functions to be piecewise smooth, where the boundary of the piecewise components are polytopes. The piecewise GHIMs in Definition III.3 covers a wide range of functions. For example, suppose the spatial volume is two-dimensional ($K = 2$). If the non-zero regions of $\boldsymbol{\beta}^*$ can be partitioned into a finite number of polygons, and $\boldsymbol{\beta}^*$ is Lipschitz continuous inside each partition, then $\boldsymbol{\beta}^*$ is a piecewise GHIM with a degree of smoothness equal to 1. In the special case of $J = 1$ and $\Omega_1 = \mathcal{D}$, the function is smooth over the whole spatial domain and satisfies the GHIMs in Definition III.2.

In searching for an optimal NN, we focus our theoretical analysis on a subset of multi-layer feed-forward NNs, where the number of layers, number of nodes, and the weight parameters are all bounded. This collection of NN functions are described in Definition III.4, following the framework in Bauer et al. [2019].

Definition III.4 (Candidate NNs)**.** For $\gamma > 0$ and $R, K'', K \in \mathbb{N}_+$, define $\mathcal{G}_{0,R,K'',K,\gamma}$

to be the collection of all functions $f : \mathbb{R}^K \to \mathbb{R}$ of the form

$$f(\boldsymbol{x}) = \sum_{r=1}^{R} \xi_r^{[3]} \phi\Big(\sum_{k'=1}^{4K''} \xi_{r,k'}^{[2]} \phi\Big(\sum_{k=1}^{K} \xi_{r,k',k}^{[1]} x_k + \xi_{r,k',0}^{[1]} \Big) + \xi_{r,0}^{[2]} \Big) + \xi_0^{[3]},$$

where $\phi(\cdot) = [1 + \exp(\cdot)^{-1}]^{-1}$, with $|\xi_{r,k',k}^{[1]}|, |\xi_{r,k'}^{[2]}|, |\xi_r^{[3]}| < \gamma$ $(r = 0, \ldots, R;\ k' = 0, \ldots, 4K'';\ k = 0, \ldots, K)$. Moreover, for $l \in \mathbb{N}_+$, define $\mathcal{G}_{l,R,K'',K,\gamma}$ to be the collection of all the functions $f : \mathbb{R}^K \to \mathbb{R}$ of the form

$$f(\boldsymbol{x}) = \sum_{r=1}^{\bar{R}} g_r(f_{r,1}(\boldsymbol{x}), \ldots, h_{r,K''}(\boldsymbol{x}))$$

for some $\bar{R} \in \mathbb{N}_+$, $g_r \in \mathcal{G}_{0,R,K'',K'',\gamma}$ $(r = 1, \ldots, \bar{R})$, and $h_{r,k'} \in \mathcal{G}_{l-1,R,K'',K,\gamma}$ $(r = 1, \ldots, \bar{R};\ k' = 1, \ldots, K'')$.

In Definition III.4, $\mathcal{G}_{l,R,K'',K,\gamma}$ is a collection of NN functions that are resulted from composing and taking the sum of l layers of two-layer feed-forward NNs, where R, K'', K specify the number of nodes in each element NN, and γ bounds the weight parameters in the NN. In addition, we use the logistic function as our activation function ϕ, as in Bauer et al. [2019]. Theoretical results on NN regression with ReLU activation can be found in Schmidt-Hieber et al. [2020].

Next, we ennumrate the conditions required in our derivation of the theoretical properties of NNISR. We first list the assumptions on the characteristics of the true model.

Assumption III.5. *For $q = 1, \ldots, Q$ and $m = 1, \ldots, M$, $\beta_q^*(\cdot)$, $\alpha_m^*(\cdot)$, and $\sigma_*^2(\cdot)$ are J-piecewise P-smooth (K', l)-GHIMs.*

Assumption III.6. *There exists a $c_{\mathcal{D}} \in \mathbb{R}_+$ such that $\mathcal{D} \subset [-c_{\mathcal{D}}, c_{\mathcal{D}}]^K$.*

Assumption III.7. *There exists a constant $c_{55} > 0$ such that for any $m \in \{1, \ldots, M\}$ and any $\boldsymbol{d} \in \mathcal{D}$, $|\alpha_m(\boldsymbol{d})| < c_{55}$.*

Assumption III.8. *There exist constants* $c_{56}, c_{59} > 0$ *such that* $\mathrm{E}[\boldsymbol{X}] = \mathbf{0}$, $\mathrm{E}[\|\boldsymbol{X}\|_2^4] < c_{56}$, *and* $\|\mathrm{Cov}[\boldsymbol{X}]\|_F^2 < c_{59}$. *In addition,* $\mathrm{Cov}[\boldsymbol{X}]$ *is positive-definite.*

Assumption III.9. *Let* $Z(\cdot) = \boldsymbol{X}^\top \boldsymbol{\beta}^*(\cdot) + \epsilon(\cdot)$. *There exist constants* $c_{52}, c_{53} \in \mathbb{R}_+$ *such that for all* $\boldsymbol{d} \in \mathcal{D}$, $\mathrm{E}\{\exp[c_{52} Z(\boldsymbol{d})^2]\} < c_{53}$.

Assumption III.5 constrains the smoothness of the spatially varying function of the main effects, individual effects, and noise variance by using the piecewise smooth GHIMs defined in Definition III.3. In comparison, similar smoothness conditions are required in [Zhu et al., 2014], which assumes the main effect coefficient function to be (piecewise) Lipschitz continuous. On the other hand, Li et al. [2020] imposes smoothness by assuming the main effects to be in a Sobolev space. Moreover, the domain of the SVFs is contained in a closed set (Assumption III.6), which is reasonable to assume for most imaging studies, as the measurement boundaries of the imaging machine is usually bounded, although the imaging resolution within this boundary could potentially be improved. In addition, the individual effects are globally bounded (Assumption III.7). For the covariates, Assumption III.8 enumerates the conditions on the moments and the covariance matrix of the covariates, while Assumption III.9 bound the expected exponential of the imaging values without the individual effects.

For the NN estimators, we require the following conditions.

Assumption III.10. *The NN parameters satisfy* $\aleph(\cdot \mid \boldsymbol{\theta}_\beta) \in \mathcal{G}^{Q}_{l,R,K'+J,K,\gamma}$, $\aleph(\cdot \mid \boldsymbol{\theta}_\alpha) \in \mathcal{G}^{M}_{l,R,K'+J,K,\gamma}$, *and* $\aleph(\cdot \mid \boldsymbol{\theta}_\sigma) \in \mathcal{G}_{l,R,K'+J,K,\gamma}$, *where* $\gamma = (MV)^{c_{40}}$ *for some sufficiently large constant* $c_{40} \in \mathbb{R}_+$.

Assumption III.11. *There exists a* $c_\aleph > 0$ *such that for any* $\boldsymbol{d} \in \mathcal{D}$, $\|\aleph(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}_\beta)\|_2^2 < c_\aleph$, $\|\aleph(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}_\alpha)\|_2^2 < c_\aleph$, *and* $\aleph(\boldsymbol{d} \mid \hat{\boldsymbol{\theta}}_\sigma)^2 < c_\aleph$.

Assumption III.12. *There exists a* $c_{67} > 0$ *such that* $\lambda \leq c_{67}(MV)^{-1}$.

Assumption III.10 defines the subset of feed-froward NNs from which we draw the candidate NNs for our NNISR estimator. In particular, the condition limits the

absolute values of the weight parameters. This framework is based on the theoretical analysis of NN regression in Bauer et al. [2019]. Assumption III.11 assumes that the NN functions are bounded in absolut evalues, which virtually always holds in practice. In addition, Assumption III.12 bounds the L_1 penalty for the main effect estimators. The rate in Assumption III.12 is analogous to the $\lambda_n/n \to 0$ upper bound for sign consistency of LASSO model selection [Zhao and Yu, 2006], although our estimator does not require a lower bound on the penalty weight, since hard-thresholding will be applied for model selection. The conditions for required for selection consistency are listed below.

Assumption III.13. *Let μ be the Lebesgue measure on $\mathcal{D} \subset \mathbb{R}^K$. For $q \in \{1, \ldots, Q\}$, let $\tilde{\mathcal{S}}_q^0 = \{\boldsymbol{d} \in \mathcal{D} : \beta_q^*(\boldsymbol{d}) = 0\}$. Then $\mu(\mathcal{D} \setminus \tilde{\mathcal{S}}_q^0)/\mu(\mathcal{D}) \ll 1$.*

Assumption III.14. *There exists a $\psi > 0$ such that for any $q \in \{1, \ldots, Q\}$,*

$$\inf_{\boldsymbol{d} \in \mathcal{D} \setminus \tilde{S}_q^0} |\beta_q^*(\boldsymbol{d})| > \psi.$$

Assumption III.15. *For any $q \in 1, \ldots, Q$, $\rho_q \to 0$, and there exists a $c_{80} > 0$ such that $\rho_q > c_{80} \log(M)^{-1}$.*

Assumption III.13 and Assumption III.14 assumes the main effects are sparse and there is a gap between zero and the the minimal signal levels. In addition, Assumption III.15 requires the selection threhsold to converge to zero but prevents it from decreasing too fast, which in practice can be realized by setting a minimum on the thresholding level below which the signals, if any, are negligible.

Our main theorem requires the following lemma, which provides an L_2 error bound of the penalized least-square estimator for the main effects. Recall that R is proportional to the total number of nodes in the NN (Assumption III.10), and P is the degree of smoothness of $\boldsymbol{\beta}$ (Assumption III.5).

Lemma III.16. *There exists a $c_{22} > 0$ such that for sufficiently large M and V,*

$$\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}(\boldsymbol{d}_v) - \aleph(\boldsymbol{d}_v \mid \hat{\boldsymbol{\theta}}_\beta)\right\|_2^2\right] \leq c_{22}[\log(MV)^3(M^{-1}V^{-1}R + R^{-\frac{2P}{K}}) + M^{-1}]. \tag{3.8}$$

Lemma III.16 decomposes the mean squared errors into a sum of three terms. The first term corresponds to the estimation error (i.e. the "variance" of the estimator). It decreases as the number of images or voxels increases, since a greater number of observations is provided, and increases as the number of nodes in the NNs increases, since more flexible models have higher estimation variation. The second term corresponds to the approximation error (i.e. the "bias" for the estimator) due to the imperfect approximation ability of neural networks. It decreases as the neural network incorporates more nodes, and the convergence is faster for regression functions with higher degree of smoothness. The third term corresponds to deviation from the true main effects caused by the error in estimating the individual effects (i.e. the difference between the target function and the expectation of the "observations"). Since the individual effects are also spatially correlated, the errors caused by them cannot be reduced by increasing the number of voxels, nor are they impacted by the complexity of the neural networks.

We now present our main theorem. In Lemma III.16, an error bound is established for the penalized least-square estimator. From here we apply hard thresholding to the estimate to induce sparsity. In Theorem III.17, we derive error bounds of the L_0 sign error and the L_2 error of the sparse main effects estimator. The L_0 sign error across all the voxels in a main effect is a weighted average between the false positive rate, false negative rate, and the false sign flipping rate, where the weights depend on the true proportion of the three signs in the main effect. It is equal to zero if and only if the signs on all the voxels of the main effects are estimated correctly. The error

bounds in Theorem III.17 are the same as that in Equation (3.8) up to a logarithmic factor. The result follows in a straightforward way from Lemma III.16 by Markov's inequality.

Theorem III.17. *There exists a $c_{79} \in \mathbb{R}_+$ such that for all sufficient large M and V,*

$$\begin{aligned}
&\mathrm{E}\left\{V^{-1}\sum_{v=1}^{V}\left\|\mathrm{sign}[\boldsymbol{\beta}(\boldsymbol{d}_v)]-\mathrm{sign}[\hat{\boldsymbol{\beta}}(\boldsymbol{d}_v)]\right\|_0\right\}\\
\leq& c_{79}[\log(MV)^5(M^{-1}V^{-1}R+R^{-\frac{2P}{K}})+\log(M)^2M^{-1}] \qquad (3.9)\\
&\mathrm{E}\left\{V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}(\boldsymbol{d}_v)-\hat{\boldsymbol{\beta}}(\boldsymbol{d}_v)\right\|_2^2\right\}\\
\leq& c_{80}[\log(MV)^5(M^{-1}V^{-1}R+R^{-\frac{2P}{K}})+\log(M)^2M^{-1}] \qquad (3.10)
\end{aligned}$$

Next, we derive the error bound of the individual effects and the noise variance. In Corollary III.18 and Corollary III.19, we show that the error bounds of the individual effects and the noise variance are similar to those of the main effects.

Corollary III.18. *There exists a $c_{26} > 0$ such that for sufficiently large M and V,*

$$\begin{aligned}
&\mathrm{E}\left[M^{-1}V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\alpha}(\boldsymbol{d}_v)-\hat{\boldsymbol{\alpha}}(\boldsymbol{d}_v)\|_2^2\right]\\
\leq& c_{26}[\log(MV)^5(M^{-1}V^{-1}R+R^{-\frac{2P}{K}})+\log(M)^2M^{-1}+\log(V)^3V^{-1}R].
\end{aligned}$$

Corollary III.19. *There exists a $c_{27} > 0$ such that for sufficiently large M and V,*

$$\begin{aligned}
&\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left\|\sigma^2(\boldsymbol{d}_v)-\hat{\sigma}^2(\boldsymbol{d}_v)\right\|_2^2\right]\\
\leq& c_{27}[\log(MV)^5(M^{-1}V^{-1}R+R^{-\frac{2P}{K}})+\log(M)^2M^{-1}+\log(V)^3V^{-1}R].
\end{aligned}$$

Unlike in Lemma III.16, the convergence of the error bound in Corollary III.18

and Corollary III.19 to zero relies on both the number of voxels and the number of images approaching infinity. As the number of images increases, the number of individual effects that need to be estimated also increases, and a sufficient number of voxels is necessary for estimating the individual effects, which is in turn required for determining the noise variance.

In Theorem III.17, the error bounds are a summation that involves the number of voxels, number of images, and NN complexity (as measured by the number of NN nodes). In practice, we would like to know the optimal NN complexity in relation to the numbers of voxels and images that results in the smallest error bound. Such optimization of the trade-off between the bias and the variance of the NNISR estimator is described in Corollary III.20.

Corollary III.20. *If* $R = c_{31}(MV)^{\frac{K}{2P+K}}$ *for some* $c_{31} > 0$*, then*

$$\begin{aligned}
&\mathrm{E}\left\{V^{-1}\sum_{v=1}^{V}\left\|\mathrm{sign}[\boldsymbol{\beta}(\boldsymbol{d}_v)] - \mathrm{sign}[\hat{\boldsymbol{\beta}}(\boldsymbol{d}_v)]\right\|_0\right\} \\
\leq & c_{32}[\log(MV)^5(MV)^{-\frac{2P}{2P+K}} + \log(M)^2M^{-1}] \\
&\mathrm{E}\left\{V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}(\boldsymbol{d}_v) - \hat{\boldsymbol{\beta}}(\boldsymbol{d}_v)\right\|_2^2\right\} \\
\leq & c_{33}[\log(MV)^5(MV)^{-\frac{2P}{2P+K}} + \log(M)^2M^{-1}]
\end{aligned}$$

for some constants $c_{32}, c_{33} \in \mathbb{R}_+$*, provided sufficiently large* M *and* V*.*

Corollary III.20 is an immediate consequence of Theorem III.17. By setting the NN flexibility R proportional to $(MV)^{\frac{K}{2P+K}}$, it optimizes the balance between the estimation error $M^{-1}V^{-1}R$ and the approximation error $R^{\frac{-2P}{K}}$ for the main effect. By setting $\nu = 2P/K$, the optimal network complexity can be rewritten as

$$(MV)^{\frac{K}{2P+K}} = (MV)^{\frac{1}{\nu+1}}.$$

The exponent here controls the optimal rate at which R should grow with respect to MV, and it converges to zero as ν approaches infinity. Recall that P is the Hölder degree of smoothness of the true main effect function, and K is its input dimension. Thus ν can be interpreted as a standardized degree of smoothness of the true spatially varying function of the main effects. Conceptually, when the regression function to be estimated is highly smooth, model flexibility is of low priority, as a simple model can already approximate the regression function fairly well, and a over-flexible model makes overfitting more likely. On the other hand, a highly non-smooth regression function can benefit greatly by a more flexible model. In this case, the model should grow fast in complexity as the numbers of images and voxels increase and provide more data points.

On top of the result in Corollary III.20, we can further set the number of images to be a function of the number of voxels, so that the optimal error bound depends on the latter only, as shown in Corollary III.21.

Corollary III.21. *Suppose $M \geq c_{35} V^{\frac{2P}{K}}$ for some $c_{35} > 0$. Then*

$$\mathrm{E}\left\{ V^{-1} \sum_{v=1}^{V} \left\| \mathrm{sign}[\boldsymbol{\beta}(\boldsymbol{d}_v)] - \mathrm{sign}[\hat{\boldsymbol{\beta}}(\boldsymbol{d}_v)] \right\|_0 \right\} \leq c_{37} \log(MV)^5 V^{-\frac{2P}{K}}$$

$$\mathrm{E}\left\{ V^{-1} \sum_{v=1}^{V} \left\| \boldsymbol{\beta}(\boldsymbol{d}_v) - \hat{\boldsymbol{\beta}}(\boldsymbol{d}_v) \right\|_2^2 \right\} \leq c_{34} \log(MV)^5 V^{-\frac{2P}{K}}$$

for some constant $c_{37}, c_{34} \in \mathbb{R}_+$, provided V is sufficiently large.

Corollary III.21 further optimizes the rate in Corollary III.20 by letting the number of images depends on the number of voxels. To achieve the most ideal error bound, the number of images should be no less than $V^{\frac{2P}{P}}$, up to a constant. Recall that in nonparametric regression in general, the minimax error bound is $N^{-\frac{2P}{2P+K}}$ [Stone, 1982]. The seemingly super-optimal performance of NNISR in Corollary III.21 (up to a logarithmic term) is due to the fact that the number of images grows along with

the number of voxels, which makes the number of observations increase faster than linearly with respect to V.

3.4 Simulation studies

3.4.1 Experiment setup

To assess the estimation and selection accuracy of NNISR, we conducted extensive simulation studies to compare it against baseline methods. We generated imaging data according to Equation (3.3). The covariates $\boldsymbol{x}_m\,(m = 1, \ldots, M)$ were 3-dimensional and were independently drawn from the standard Gaussian distribution. For the main effects $\beta_q^*(\cdot)\,(q = 1, \ldots Q)$, individual effects $\alpha_m^*(\cdot)\,(m = 1, \ldots M)$, and noise variance $\sigma_*^2(\cdot)$, we designed their spatially varying functions with diverse patterns and set their domain to a bounded 3-dimensional rectanguloid, as illustrated in Figure 3.1a. First, for each main effect $\beta_q^*(\cdot)$, the rectanguloid is further divided into a 2×2 grid among the first two axes of the spatial coordinates, with different spatial patterns in each of the 4 blocks. The top-left block contains no signals, and the value of the spatially varying function is set to zero in this area. In the top-right block, we generated two spherical regions that contains signals with strength smoothly diminishing to zero at the boundary. The size and location of the spherical regions are randomly and independently selected, and the two regions may be mutually exclusive or overlapping. In the latter case, a superposition of the two regions is produced, with the values inside the overlapping area equal to the sum of the values in the two regions. In the bottom-left block, the spatial pattern is similar to that in the top-right block, except that the regions are rectangular, with a constant level of signal, which gives rise to sharp edges. For the bottom-right block, the spatial pattern is a mixture of those in the top-right block and the bottom-left block. Moreover, for each individual effect $\alpha_m^*(\cdot)$, we design the spatial patterns as follows: a random spatial lo-

cation is selected as the center and assigned with a random value. Then the spatially varying function increases or decreases proportionally to the distance from the center. Finally, the variance of the noises varies in a periodic manner across the three axes in the spatial volume by using a randomly linearly transformed sine function. Slices of the main effects, noise variance, and an example of the individual effects are shown in Figure 3.1a. Notice that we have intentionally violated some of our theoretical assumptions listed in Section 3.3 to test the robustness of our method. For example, the non-zero regions in the main effects are not all polytopes (Assumption III.5), and some of the non-zero regions have continuous transitions toward zero on their boundaries (Assumption III.14).

Three experimental variables were adjusted in our studies. First, the noises followed either the standard Gaussian distribution or a $(\text{Chisq}_3 - 3)/\sqrt{6}$ distribution. Second, we set the imaging resolution to $16 \times 16 \times 8$, $32 \times 32 \times 8$, $64 \times 64 \times 8$, or $128 \times 128 \times 8$. Finally, each data set contained either 20 or 50 images. Each of the 12 experimental settings was replicated for 50 times. In all experiments, the ratio of the variances of the main effects, individual effects, and noises were set to $0.2 : 0.5 : 1.0$. The bottom-left panel of Figure 3.1a shows (a slice of) an example of the response image. With such a low signal-to-noise ratio, we intended to simulate the highly noisy imaging data common to biomedical imaging studies.

We compared NNISR against MUA, SPM, BST, STOR, and SVCM. The methods were assessed for their estimation and selection accuracy. The estimation accuracy was measured by the MSE between the true main effects and the estimated effects, while the selection accuracy was measured by the area under the operating characteristic curve (AUC), as well as the false positive and power. We report the median and the interquartile range (IQR) of these four summary statistics across the replications in each of the experimental settings.

For model selection in NNISR, the selection quantile was set to $\eta_q =$

$\min(\eta_{q,\text{MUA}}, \eta_{\max})$, where $\eta_{q,\text{MUA}}$ equals to the proportion of voxels selected by MUA for the q^{th} main effect, and $\eta_{\max} = 0.2$ to reflect the prior belief that in practice the proportion of significant voxels rarely exceeds 20%. In addition, the minimum selection level was set to $\rho_{\min} = 0.0001$, as effect levels below this are negligible. The test size was set to 0.05 for MUA, BST, SVCM, and SPM. Notice that the test size for SPM is automatically adjusted for the number of voxels by using its algorithm based on random field theory. Sparsity in STOR is not determined by a pre-specified test size but rather embedded in the algorithm and controlled by hyperparameters, which were tuned according to the authors' recommendations.

For the hyperparameters in NNISR, we used 4 hidden layers with 64 nodes in each layer. In our experiments, we noticed that the performance was not sensitive to the complexity of the NN architecture. Moreover, the weight of the L_1 penalty on the main effects were selected by cross-validation with LASSO on each voxel and then taking the geometric mean of the optimal weights across all the voxels. For the baseline methods, the hyperparameters were selected according to the recommendations in the original papers.

3.4.2 Experiment results

Table 3.1 shows the estimation accuracy, as measured by MSE, and selection accuracy, as measured by AUC, false positive rate, and power, of each method For estimation accuracy, NNISR has the lowest median MSE and the smallest IQR in all the experimental settings, which shows NNISR to be the uniformly most accurate and most stable method for estimating the main effects. The characteristics of NNISR's performance is demonstrated in several trends. First, the advantage of NNISR is the most prominent when the imaging resolution is high. In all the 4 combinations of noise distribution and number of images, there is a clear trend of decreasing MSE for NNISR as the number of voxels increases. Among the baseline methods, such trend is also

observed for SPM, BST, and STOR but with much slower speeds of improvement, while MUA and SVCM show no significant improvement with increasing imaging resolutions. For example, consider the experiment with Gaussian noise, 50 images, and $16 \times 16 \times 8$ voxels. Although the MSE of NNISR is the lowest and is less than 50% of those of MUA and SVCM, the MSEs of STOR, BST, and SPM are no more than 120% of NNISR's. However, as the imaging resolution is increased to $128 \times 128 \times 8$, the MSE of NNISR becomes less than 20% of MUA and SVCM and no more than 52% of those of STOR, BST, and SPM. In fact, NNISR's exploitation of high imaging resolutions is so efficient that its MSE for 20 images with $128 \times 128 \times 8$ voxels is less than its MSE for 50 images with $16 \times 16 \times 8$ voxels. In contrast, none of the baseline methods can overcome deficiency in the numbers of images with abundance in the number of voxels.

Moreover, the advantage of NNISR over the baseline methods is greater for small numbers of images than for large numbers of images. Compared to the second most accurate method in each setting, NNISR's reduction in MSE is 38% to 62% for 20 images and 7% to 48% for 50 images. This trend demonstrates NNISR's usefulness for datasets with deficiency in the number of images. Furthermore, NNISR is shown to be robust against skewed noise distribution. Compared to the Gaussian-generated noise, NNISR has lower MSE for chi-squared distribution in all the experimental settings. For the baseline methods, robustness against skewed noise also holds for MUA and SVCM, but SPM, BST, and STOR show clear increase in MSE for chi-squared noise. Overall, NNISR is the only method that exploits high imaging dimensions and robust against skewed distributions simultaneously. It is noticeable that for both the Gaussian and chi-squared noise, NNISR's estimation accuracy for 20 images with the highest resolution is better than any other method's accuracy for 50 images with any resolution. This result demonstrates the advantage of NNISR in imaging studies with limited numbers of images but high imaging resolutions.

To further illustrate the estimation characteristics of each method, we show two-dimensional slices of their three-dimensional estimates of the main effects in Figure 3.1. MUA has produced the most noisy estimate, which is expected due to its ignoring of spatial information. It estimation accuracy on a voxel highly depends on its local noise level, as shown by the contrast of the estimates in the noisy blobs versus the other imaging regions. SVCM, whose algorithm is based on smoothening the noise in the MUA estimate, has eliminated the majority of the estimation errors caused by lower noise, but most of the fluctuation in the high-noise regions have remained. In contrast, SPM and BST are less susceptible to white noise, but they tend to generate errors of wrinkle patterns, ignore the backdrop bias, (i.e. the gradual transition in the background from blue at the top to red at the bottom), and over-smoothen the true signals, as shown in the blurry boundaries of the rectangular activation regions in the estimates. Moreover, the characteristics of the estimate by STOR is very distinct from the other methods. It favors activation regions with rectangular boundaries, due to its treatment of images as tensors (multi-dimensional arrays) and imposing low-rankedness on them, and thus tend to produce spatially correlated error regions of rectangular shapes and overlook true activation regions with curvy boundaries. Finally, NNISR has successfully detected all the activation regions of various geometric shapes and adapted for both the smooth boundaries and the sharp boundaries. It has also generated the cleanest estimate, with the estimated values on most of the null voxels virtually equal to zero, eliminating most of the highly noisy blobs in MUA and SVCM as well as the smooth backdrop bias in SPM and BST, which is especially impressive considering the fact that model selection has yet been applied. The almost-sparse property of the NNISR estimate is caused by the L_1 penalty in the model, though it is extremely unlikely to shrink any voxel exactly to zero, unless all the weight parameters in the last layer of the NN is exactly zero, Overall, compared to the baseline methods, the estimate by NNISR resembles the true main effect more

closely, both quantitatively, as measured by the MSE, and visually, as shown by the plots. It has cleaned up many of the bias patterns common to the other methods. These results demonstrate the flexibility of NNISR and its robustness against not only white noise but also spatially correlated fluctuations in the images.

For selection accuracy, NNISR has higher median AUC than the other methods under all experimental settings except when the number of images and the number of voxels are both the smallest. The selection accuracy of NNISR increases as the number of voxels increases, and the performance of NNISR relative the the baseline methods is greater for small numbers of images with high resolutions than for large numbers of images with low resolutions. In addition, the AUCs of NNISR are similar in median for the Gaussian and chi-squared noises, with only a slight increase in IQR in the latter case.

The selection accuracy of each method is further demonstrated by their false positive rate and power. NNISR has successfully controls the false positive rate, which has a median of 0.05 or lower in all the experimental settings and is uniformly more stable than the other methods, as reflected in the IQR. For the baseline methods, MUA is more conservative than NNISR and has false positive rates between 0.03 and 0.04, while SPM is over-conservative with a uniform 0.00 false positive rate, due to its automatic adjustment for the number of voxels being tested and the low signal-to-noise ratio in the data design. In contrast, SVCM, BST, and STOR all have false positive rates above 0.05, although for STOR the false positive rate is not intended to be controlled. Finally, for the testing power, NNISR's performance reflects the reoccurring trends: improvements with increasing imaging resolution and robustness against skewed noise. The power is less comparable across methods, since SPM is over-conservative while SVCM, BST, and STOR are over-liberal. The most comparable method for NNISR in terms of power is MUA, since it has similar false negative rates. Compared to MUA, NNISR has lower power for small numbers of voxels and higher

power for large numbers of voxels, which again demonstrates NNISR's advantage on high-dimensional imaging data.

3.5 Analysis of fMRI data

3.5.1 Experiment setup

To evaluate the performance of NNISR in medical imaging studies, we applied it and baseline methods to two neuroimaging consortia: the Autism Brain Imaging Data Exchange (ABIDE) and the Adolescent Brain Cognitive Development study (ABCD). ABIDE and ABCD collected fMRI images from multiple experimental sites in the U. S. In addition, the cortia contain clinical characteristics such as cognitive ability (CA), disease status, and psychiatric diagnostics, as well as demographic information such as age and sex.

The ABIDE [Di Martino et al., 2014] project aims at improving the neurological understanding of autism and the associated cognitive behaviors. In our our analysis, we used the Phase I data of the study, which include 20 resting-state fMRI datasets from 19 experimental sites, with a total of 1,112 subjects. We employed a widely adopted fMRI processing pipeline [Craddock et al., 2013, He et al., 2019]. For the response image, we used the weighted degrees of network connectedness, which correspond to each voxel's number of direct connections to the other voxels. The covariate of primary interest is CA, which is measured by the full-scale intelligence quotient. In addition, our model adjusts for autism status, age, and sex. After removing the missing values, the dataset contains 821 subjects.

The ABCD [Casey et al., 2018] study focuses on studying the association between cognitive behaviors and brain development. This project collected the brain images and various CA assessment scores of more than 11,800 children of age 9 to 10 from 21 experimental sites. Our analysis used the minimally preprocessed 2-back task-based

fMRI data of 1991 subjects in 20 sites from the curated ABCD annual release 1.1 [Hagler Jr et al., 2019]. The response image is the contrast map of the 2-back task, which has been consistently found to engage brain regions for memory regulation processes and cognitive functions [Barch et al., 2013]. The covariate of interest, CA, is measured by the general CA component score [Sripada et al., 2019]. The psychiatric diagnostic score, age, and sex have also been adjusted in the model.

We compared the performance of NNISR with MUA, SPM, and SVCM. We were unable to evaluate STOR and BST, since their software sent out error messages when running on our fMRI data. To examine the estimation accuracy of each method, we conducted cross validation across the experimental sites. In each fold of the cross validation, one site was selected for training and the other sites were used for testing. Estimation accuracy was measured by the response image recovery MSE on the testing data, which is equivalent to the proportion of variation of the testing response images explained by the estimated main effects. Then we applied model selection to the estimated main effects, where the selection hyperparameters are the same as in the simulation studies, described in Section 3.4. As the true signals on the real data are unknown, we instead examine the selection results by their reproducibility, defined as follows. First, we combined all the experimental sites and applied the estimation and selection procedures to the combined dataset. To make the results comparable, in both the all-site analysis and each single-site analysis, we set all the baseline methods to select the same proportion of voxels as NNISR, according to the ranking of the signal strength defined by each method (e.g. the p-values for MUA). Then we compute the proportion of voxels selected for CA in the all-site analysis that are reproducible in the single-site analyses. A voxel is said to be reproducible if it is selected in at least 5 single-site analyses. We use this metric to measure the degree of selection reproducibility of each method on the fMRI data.

Furthermore, in the all-site analysis of the neuroimages, we summarized the selec-

tion results in various brain regions. We divided the brain volume into parcels based on the automated anatomical labeling (AAL) atlas [Tzourio-Mazoyer et al., 2002], and the parcels were further grouped into functional networks (FNs) [Power et al., 2011]. The voxels selected by different methods for CA were compared in each AAL region and FN. We reported the top ten regions and top five FNs as ranked by the proportion of voxels selected in each region or FN in the all-sample analysis. In addition, each region- and FN-based rate of reproducibility is also recorded. Finally, we investigated the top regions selected in the ABIDE and ABCD data in the literature and reported their biological significance in existing works.

3.5.2 Experiment results

Figure 3.2 shows the cross-validation testing MSE for response image recovery recovery across the experimental sites. To compare the relative performance of the methods, we report the testing MSE relative to that of MUA. Each point on the plot represents an experimental site, and the x-coordinate corresponds to the testing MSE of the MUA model trained on that site, and the y-coordinate corresponds to the proportion difference in testing MSE of each method compared to MUA. For the ABIDE data, the performance of MUA, SPM, and SVCM are similar when the MSE of MUA is low, although SPM tends to be less stable than the other methods. These are the cases where the estimation task is less challenging. As the MSE of MUA increases, the different methods are further differentiated. NNISR has the smaller relative MSE than the other methods on most sites. Moreover, the reduction of MSE by NNISR compared to MUA becomes greater when the MSE of MUA is greater, which corresponds to the experimental sites with data that are more difficult to generalize and on which MUA performs poorly. Similar performance patterns are observed on the ABCD data, with the aforementioned advantage of NNISR being even more prominent compared to MUA and the other baseline methods. These characteristics echos

the results in the simulation studies and demonstrate the usefulness of NNISR on difficult imaging datasets, which usually have limited numbers of images and high level of noise with complex spatially correlations.

The results for the reproducibility analysis is shown in Table 3.2. Reproduciblity is measured by the proportion of voxels selected in the all-sample analysis that are also selected in 5 or more single-site analyses. NNISR has achieved the highest rate of reproducibility, which is 217% and 153% of that of SVCM, the second most reproducible method. These results are visualized in Figures 3.3a and 3.3b. Only the voxels selected in all-site analysis are colored, with red and blue represents whether or not a voxel is reproducible. Beside the clear pattern that NNISR has consistently selected more voxels across the experimental sites, the regions it selected tend to form large, contiguous clusters, which are more interpretable in biomedical imaging studies than small, isolated patches sporadically scattered over the whole volume.

Moreover, Table 3.3 shows the region-based and FN-based selection results. For ABIDE, regions in the occipital lobe [Goriounova and Mansvelder, 2019, Yoon et al., 2017, Simard et al., 2015, Menary et al., 2013], the calcarine fissure and surrounding cortex [Yu et al., 2008], and the cuneus [Schnack et al., 2015, Haier et al., 2004, Song et al., 2008] have high proportions of voxels selected by most of the methods. This result is consistent with the known associations between the aforementioned regions and CA in the literature. In addition, these regions all belong to the visual FN, which has been found to be related to CA in existing works [Dubois et al., 2018, Hearne et al., 2016] and is the top FN selected by all the methods in ABIDE. The other FNs all have much lower low selection rates. Compared to the baseline methods, NNISR's reproducibility is higher in most of the top regions and FNs. Furthermore, the selection proportion by NNISR is more concentrated at the top regions in the all-sample analysis of ABIDE, This trend indicates that the voxels selected by NNISR are more closely aligned with biologically meaningful regions and networks in the

brain.

For ABCD, the regions with high selection rate for CA include those in the the parietal lobe [Haier et al., 2005, Woolgar et al., 2010, Jung and Haier, 2007], frontal lobe [Duncan et al., 1996, Roca et al., 2010], and the precuneus [Basten et al., 2015, Jauk et al., 2015]. For the FN-based selection results, the memory retrieval FN has the highest selection rate by all the methods, which is consistent with the fact that the ABCD response images used in our analysis are task-based contrast maps for activities designed to engage working memory. The selection rate for dorsal attention FN is also high, which has been found in the literature to be associated with both general CA [Hilger et al., 2020] and working memory capacity [Majerus et al., 2018, Broadway and Engle, 2010, Gray et al., 2017]. Other top FNs include frontal-parietal task control [Uddin et al., 2019, Zanto and Gazzaley, 2013] and salience [Hilger et al., 2017, Liang et al., 2016], which are consistent with findings in existing works regarding their associations with CA and working memory. Finally, the reproducibility rate of NNISR is much higher than the that of the other methods in most of the regions and networks, making NNISR the most stable and consistent method for model selection.

3.6 Discussion

In this work, we have presented a novel image-on-scalar regression model based on neural networks. From the perspective of functional data analysis, our model uses multi-layer feed-forward neural networks to approximate the spatially varying coefficient functions of the main effects, individual effects, and noise variance. Although conceptually straightforward, our NNISR model has been shown to be capable of adapting to a wide variety of spatial correlation patterns, including not only smooth transitions but also jump discontinuities across the spatial volume. We have provided an algorithm for model fitting and selection that takes advantage of the high-dimensionality of the imaging data. Such estimation procedure has been proved

to posses theoretically guaranteed convergence properties.

In our theoretical analysis, we have derived L_2 estimation error bounds for the main effects, individual effects, and noise variance. Our results are not only based on existing works on nonparametric regression with neural networks but also extend them to include not only globally smooth functions but also piecewise smooth functions. Moreover, for our model selection procedure, we have demonstrated their selection consistency and proved L_0 sign error bounds. We also have showed the optimal neural network complexity as a function of the number of images and voxels.

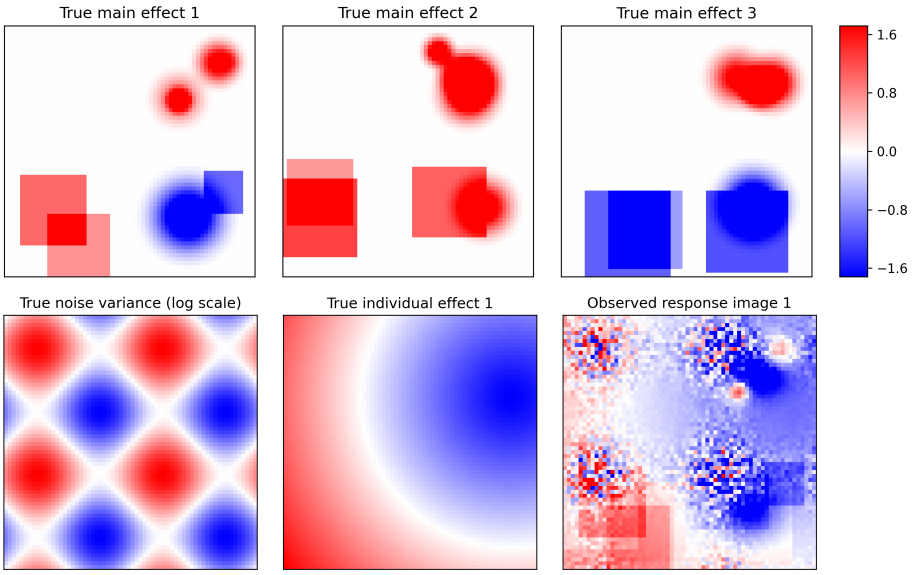
In our extensive simulation studies, we have designed complex spatial images to test NNISR against multiple existing image-on-scalar regression methods. NNISR has successfully eliminated a great proportion of the noises and learned most of the the underlying heterogeneous patterns in the main effects. Moreover, NNISR has been shown to effective in exploiting the increasing imaging dimensions compared to the existing methods, both in terms of estimation accuracy and selection accuracy. For the analysis of brain fMRI data, NNISR has produced more accurate estimates in the cross-validation across experimental sites. The advantage of NNISR over the baseline methods is more prominent when the estimation problem is more difficult. In addition, our model has also achieved the highest selection reproducibility, as measured by the proportion of voxels that are consistently selected in the single-site analyses. Finally, as we break down the results into AAL regions and functional networks, the top regions and networks selected by NNISR are supported by findings in existing works. The reproducibility rate in each region or network is also much higher for NNISR than the baseline methods.

We have envisioned multiple directions for future works. In our current setting, although the imaging resolution is high, the number of covariates is fixed at a constant value. We could extend the current theoretical results and modify the algorithm to accommodate an increasing number of covariates, which is common in applications

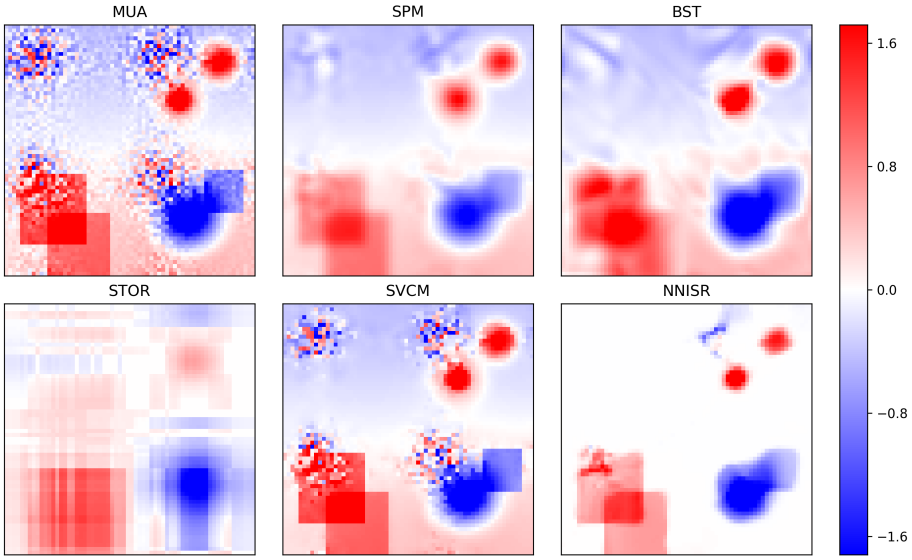
such as imaging genetics. In addition, the images are currently assumed to be uni-modal. This condition can be potentially generalized to multi-modal images, such as data generated from spatial transcriptomics.

3.7 Tables and Figures

Figure 3.1: Slices of the images for the true and estimated main effects, noise variance, individual effects, and observed response in the simulation studies.



(a) True parameters and observed responses.



(b) Estimates of main effect 1 by NNISR and the baseline methods.

Figure 3.2: Cross-site testing MSE for recovering imaging response on the fMRI data. Each point represents a single-site analysis, where the data from one experimental site is used for training and those from the other sites are used for testing. The x-coordinate equals to the testing MSE of MUA, which measures the overall difficulty of estimation and generalization for models trained on each site. The y-coordinate equals to the relative testing MSE of each method compared to MUA. The testing response recovery MSE is equivalent to the proportion of variance explained by the estimated main effects on the testing data, which is used as a metric to indirectly measure the estimation error.

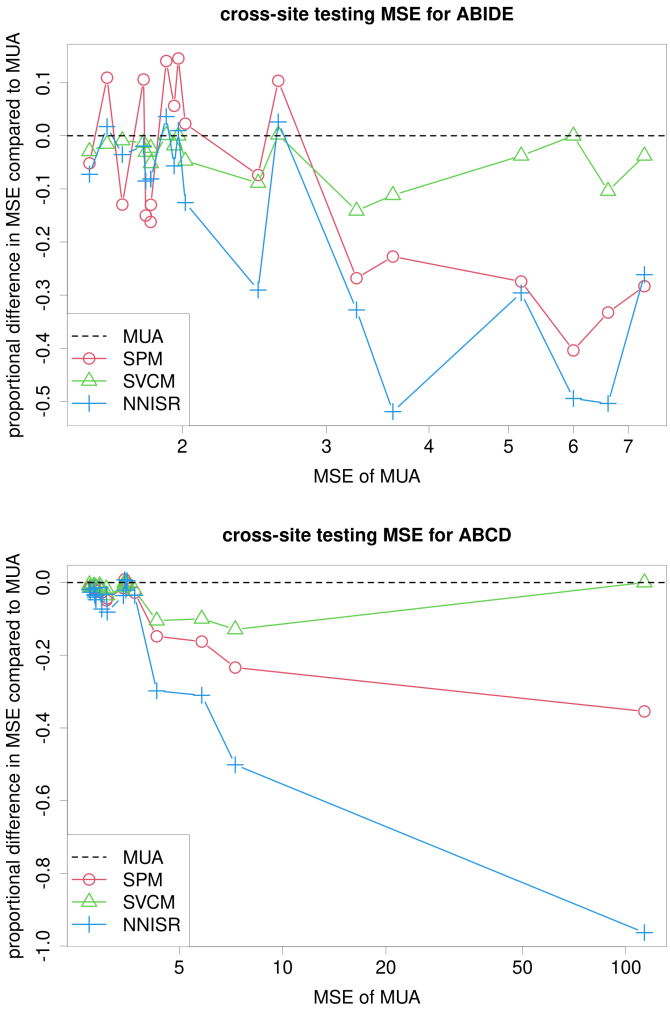
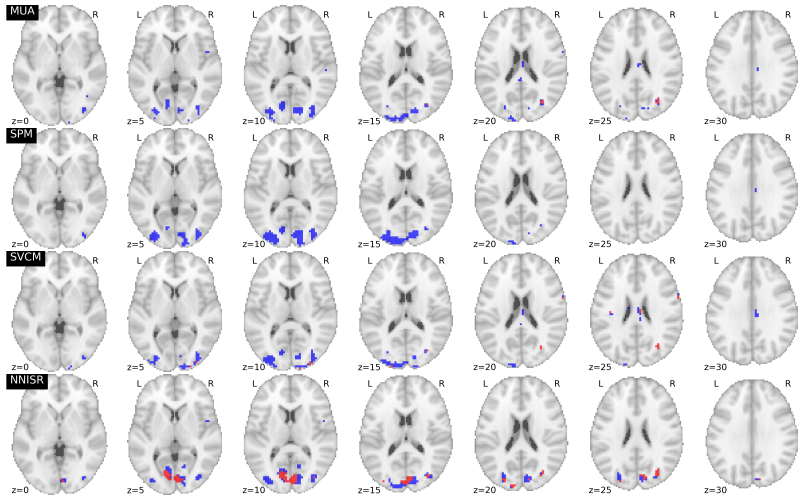
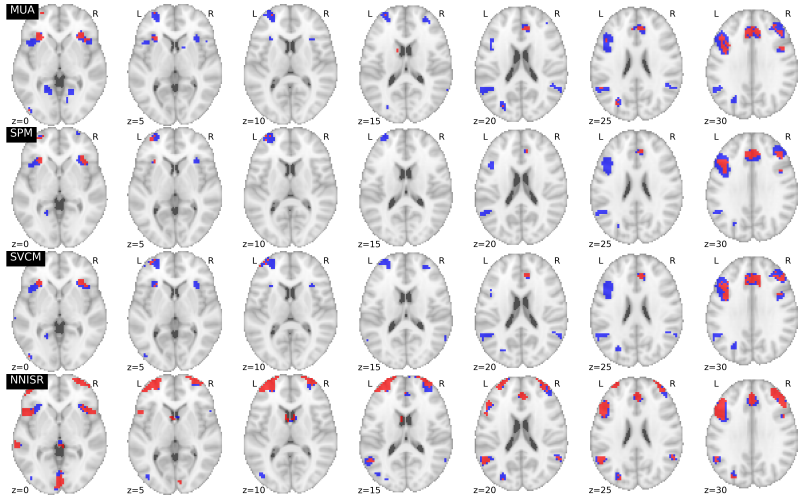


Figure 3.3: Selection results for cognitive ability (CA) on the fMRI data and the reproduciblity status of each voxel by NNISR and the baseline methods. Voxels selected for CA in the all-site analysis are shown in color. Red represents voxels selected in the all-site analysis that are reproducible in the single-site analyses, while blue represents voxels selected in the all-site analysis that are not reproducible in the single-site analyses. A voxel is said to be reproducible if it is selected in at least 5 single-site analyses.



(a) ABIDE



(b) ABCD

Table 3.1: Summary statistics for main effect estimation and selection in the simulation studies by NNISR and the baseline methods. The data are generated with the standard Gaussian or standardized chi-squared distribution, with the number of images M equal to 20 or 50 and the number of voxels set to $V' \times V' \times 8$, where V' varies from 16 to 128. Each setting is replicated for 50 times. The median and the interquartile range (displayed in parentheses) of the summary statistics are reported (in the unit of 0.01).

		$\mathcal{N}(0,1)$ noise						$(\chi^2_3 - 3)/\sqrt{6}$ noise					
M	V'	MUA	SPM	BST	STOR	SVCM	NNISR	MUA	SPM	BST	STOR	SVCM	NNISR
Mean squared error (MSE)													
20	16	191 (57)	61 (32)	63 (31)	63 (27)	157 (51)	38 (22)	187 (78)	66 (41)	69 (43)	59 (36)	153 (73)	35 (19)
20	32	190 (59)	59 (36)	59 (34)	51 (33)	161 (53)	28 (15)	185 (64)	63 (41)	64 (41)	52 (31)	155 (64)	25 (18)
20	64	192 (58)	58 (37)	58 (35)	47 (32)	160 (50)	23 (19)	185 (66)	62 (41)	62 (42)	53 (31)	158 (64)	20 (18)
20	128	193 (54)	58 (37)	56 (36)	44 (31)	161 (50)	22 (16)	186 (64)	62 (41)	61 (42)	50 (30)	157 (63)	19 (19)
50	16	66 (12)	30 (10)	27 (11)	28 (10)	52 (12)	25 (08)	65 (16)	32 (14)	26 (14)	30 (10)	51 (15)	24 (11)
50	32	65 (11)	26 (11)	24 (11)	24 (08)	52 (10)	16 (08)	66 (17)	27 (13)	24 (13)	26 (09)	52 (14)	15 (10)
50	64	65 (10)	24 (11)	22 (11)	23 (10)	52 (11)	12 (07)	66 (16)	25 (13)	22 (13)	22 (11)	53 (14)	13 (08)
50	128	64 (10)	24 (10)	21 (10)	22 (08)	51 (11)	11 (06)	67 (15)	24 (14)	21 (14)	21 (09)	53 (13)	11 (08)
Area under the operating characteristic curve (AUC)													
20	16	66 (08)	63 (13)	64 (11)	61 (12)	67 (09)	65 (10)	68 (07)	64 (12)	66 (08)	62 (14)	67 (08)	64 (10)
20	32	66 (08)	66 (11)	66 (11)	64 (15)	67 (10)	69 (08)	67 (07)	67 (10)	68 (07)	64 (12)	67 (08)	70 (10)
20	64	66 (07)	67 (11)	67 (10)	65 (12)	67 (09)	70 (06)	67 (07)	68 (10)	69 (07)	66 (09)	68 (08)	71 (09)
20	128	66 (08)	67 (11)	68 (11)	64 (09)	67 (09)	70 (06)	67 (07)	68 (10)	69 (07)	64 (11)	68 (08)	71 (09)
50	16	76 (08)	74 (11)	77 (09)	74 (14)	77 (08)	78 (09)	75 (07)	74 (12)	76 (08)	73 (09)	77 (08)	78 (09)
50	32	76 (08)	77 (10)	78 (09)	77 (14)	77 (08)	81 (06)	75 (07)	77 (10)	77 (09)	73 (10)	76 (08)	79 (08)
50	64	76 (08)	78 (09)	79 (09)	76 (16)	78 (08)	81 (05)	75 (06)	78 (09)	78 (08)	75 (09)	77 (07)	80 (08)
50	128	76 (08)	79 (08)	79 (09)	76 (15)	78 (08)	81 (06)	75 (07)	78 (09)	78 (09)	74 (12)	77 (07)	81 (08)
False positive rate													
20	16	04 (05)	00 (00)	23 (20)	45 (53)	08 (06)	05 (04)	04 (04)	00 (00)	23 (14)	51 (59)	07 (04)	04 (04)
20	32	04 (05)	00 (00)	23 (21)	80 (85)	08 (06)	04 (04)	03 (04)	00 (00)	23 (14)	75 (79)	07 (04)	03 (04)
20	64	04 (05)	00 (00)	23 (21)	82 (76)	08 (07)	03 (04)	03 (04)	00 (00)	24 (15)	85 (61)	08 (04)	03 (04)
20	128	04 (05)	00 (00)	23 (22)	89 (81)	08 (07)	03 (04)	03 (04)	00 (00)	24 (15)	84 (75)	08 (05)	03 (04)
50	16	04 (04)	00 (00)	38 (14)	41 (47)	07 (06)	05 (04)	05 (05)	00 (00)	41 (20)	56 (52)	08 (06)	05 (03)
50	32	03 (05)	00 (00)	39 (15)	86 (79)	08 (06)	03 (04)	04 (05)	00 (00)	42 (19)	63 (72)	08 (06)	04 (04)
50	64	03 (05)	00 (00)	39 (15)	85 (82)	07 (06)	03 (04)	04 (05)	00 (00)	42 (19)	90 (63)	08 (06)	04 (03)
50	128	04 (05)	00 (00)	39 (14)	69 (82)	07 (06)	03 (04)	04 (05)	00 (00)	43 (19)	79 (74)	08 (06)	03 (04)
Power													
20	16	21 (08)	00 (01)	45 (20)	69 (54)	28 (11)	19 (09)	23 (09)	00 (01)	45 (18)	71 (59)	29 (12)	21 (10)
20	32	22 (07)	01 (02)	47 (16)	85 (75)	27 (10)	24 (09)	24 (09)	01 (02)	49 (16)	81 (65)	29 (14)	25 (10)
20	64	22 (07)	01 (01)	48 (15)	91 (36)	28 (10)	25 (07)	23 (08)	01 (02)	51 (14)	95 (33)	30 (13)	26 (10)
20	128	22 (07)	01 (01)	49 (16)	93 (55)	28 (10)	25 (07)	24 (09)	01 (01)	52 (16)	92 (47)	30 (13)	26 (10)
50	16	42 (09)	02 (03)	73 (09)	79 (43)	49 (11)	36 (11)	43 (07)	03 (03)	73 (10)	82 (28)	50 (07)	39 (10)
50	32	42 (10)	05 (04)	75 (10)	97 (39)	41 (12)	42 (09)	43 (08)	05 (04)	75 (10)	89 (28)	50 (09)	44 (09)
50	64	42 (10)	05 (03)	76 (10)	97 (44)	50 (11)	43 (09)	43 (08)	06 (04)	76 (10)	97 (23)	51 (09)	45 (09)
50	128	42 (10)	04 (03)	78 (09)	94 (46)	50 (11)	45 (09)	43 (08)	05 (03)	77 (10)	95 (36)	51 (09)	46 (09)

Table 3.2: Selection reproducibility of voxels in the fMRI data by NNISR and the baseline methods. Reproducibility is measured by the proportion of voxels selected in the all-sample analysis that are also selected in at least 5 single-site analyses.

Dataset	MUA	SPM	SVCM	NNISR
ABIDE	0.021	0.000	0.168	0.365
ABCD	0.397	0.423	0.428	0.655

Table 3.3: Voxel selection and reproducibility of AAL regions and functional networks in the ABIDE and ABCD data. Inside each row for each method, the first column is the name of the region/network, the second column shows the proportion of voxels inside the region/network that are selected in the all-site analysis, and the third column (in parentheses) reports the proportion of these voxels that are reproducible in the single-site analyses, where reproducibility is defined as being selected in 5 or more single-site analyses. All the proportions are displayed in the unit of 0.01.

MUA		SPM		SVCM		NNISR	
AAL regions in ABIDE							
Occ.Mid.R	11 (10)	Occ.Mid.L	13 (00)	Rec.R	13 (00)	Cal.R	17 (40)
Cal.R	10 (00)	Cal.R	12 (00)	Rec.L	12 (25)	Cal.L	15 (67)
Rec.L	09 (00)	Occ.Sup.L	10 (00)	Occ.Mid.R	10 (23)	Cun.R	12 (27)
Occ.Mid.L	07 (00)	Occ.Mid.R	10 (00)	Occ.Mid.L	09 (00)	Occ.Mid.R	12 (46)
Occ.Sup.L	06 (04)	Cun.R	08 (00)	Occ.Sup.L	08 (04)	Occ.Mid.L	07 (14)
Sup.Mot.Are.R	06 (00)	Cal.L	05 (00)	Cal.R	06 (03)	Occ.Sup.L	06 (39)
Cal.L	05 (00)	Sup.Mot.Are.R	05 (00)	Fro.Med.Orb.L	05 (38)	Cun.L	05 (35)
Rec.R	04 (00)	Cun.L	04 (00)	Fro.Sup.Orb.R	05 (83)	Fus.L	03 (06)
Cun.L	04 (00)	Rec.L	03 (00)	Tem.Pol.Mid.L	04 (00)	Lin.L	03 (21)
Occ.Inf.R	04 (00)	Rec.R	02 (00)	Sup.Mot.Are.R	04 (00)	Sup.Mot.Are.R	02 (22)
Functional networks in ABIDE							
Vis	04 (03)	Vis	06 (00)	Vis	04 (07)	Vis	07 (38)
Ven.Att	01 (00)	Ven.Att	01 (00)	Sen.Som.Han	01 (24)	Def.Mod	01 (46)
Tas.Con	01 (00)	Cin.Ope.Tas.Con	01 (00)	Cin.Ope.Tas.Con	01 (04)	Ven.Att	00 (22)
Som.Han	01 (02)	Def.Mod	00 (00)	Ven.Att	01 (00)	Cin.Ope.Tas.Con	00 (20)
Def.Mod	01 (07)	Sen.Som.Han	00 (00)	Def.Mod	01 (21)	Sen.Som.Han	00 (22)
AAL regions in ABCD							
Par.Inf.L	49 (42)	Par.Inf.L	55 (40)	Par.Inf.L	51 (42)	Par.Inf.L	51 (74)
Par.Sup.L	40 (23)	Par.Sup.L	50 (30)	Par.Sup.R	47 (46)	Fro.Mid.Orb.L	44 (00)
Pre.L	35 (59)	Par.Sup.R	47 (35)	Par.Sup.L	46 (33)	Par.Sup.L	42 (70)
Par.Inf.R	34 (56)	Pre.L	40 (61)	Par.Inf.R	39 (56)	Fro.Mid.L	36 (79)
Inf.Ope.L	33 (40)	Par.Inf.R	38 (54)	Pre.L	35 (63)	Par.Sup.R	36 (25)
Par.Sup.R	32 (13)	Fro.Inf.Ope.L	33 (37)	Pre.R	32 (71)	Fro.Inf.Ope.L	33 (73)
Pre.R	31 (65)	Pre.R	33 (68)	Fro.Inf.Ope.L	31 (31)	Fro.Sup.Orb.L	32 (97)
Pre.L	29 (59)	Fro.Mid.L	28 (49)	Fro.Mid.L	29 (53)	Pre.L	31 (83)
Lin.L	25 (12)	Pre.L	27 (53)	Pre.L	28 (57)	Fro.Mid.Orb.R	31 (96)
Fro.Mid.L	24 (42)	Sup.Mot.Are.L	24 (82)	Lin.L	23 (03)	Fro.Mid.R	29 (75)
Functional networks in ABCD							
Mem.Ret	35 (59)	Mem.Ret	40 (61)	Mem.Ret	35 (63)	Mem.Ret	31 (83)
Dor.Att	22 (43)	Dor.Att	25 (49)	Dor.Att	23 (49)	Fro.Par.Tas.Con	24 (75)
Fro.Par.Tas.Con	19 (43)	Fro.Par.Tas.Con	19 (47)	Fro.Par.Tas.Con	20 (48)	Sal	23 (75)
Sal	18 (42)	Sal	17 (47)	Sal	19 (49)	Dor.Att	21 (77)
Cin.Ope.Tas.Con	17 (53)	Cin.Ope.Tas.Con	16 (62)	Cin.Ope.Tas.Con	16 (53)	Def.Mod	14 (72)

CHAPTER IV

Bayesian Deep Aleatoric Neural Networks

4.1 Introduction

Deep neural networks (DNNs) [Goodfellow et al., 2016, LeCun et al., 2015] have achieved state-of-the-art in numerous data analysis challenges [Pouyanfar et al., 2018]. It has been shown that the DNNs are highly successful and of great potential in a wide range of applications, ranging from computer vision [Voulodimos et al., 2018], natural language processing [Young et al., 2018], and autonomous driving [Grigorescu et al., 2020], to medical imaging [Ker et al., 2017], genomics [Zou et al., 2019], health management [Zemouri et al., 2019], astronomy [Meher and Panda, 2021], and agriculture [Kamilaris and Prenafeta-Boldú, 2018]. In some areas such as image classification, DNNs are able to produce a higher accuracy than human classifiers, and thus they have been given serious consideration in practice [Berner et al., 2021]. The rise of popularity for DNNs, in addition, is witnessed by an increasing number of conferences and workshops that exclusively focus on deep learning methods [Deng and Yu, 2014]

A key advantage of DNNs is their flexibility in fitting data with complex patterns. In theory, the universal approximation property ensures that DNNs have the ability to approximate any continuous function on a compact set up to arbitrary precision, provided that they have sufficient numbers of nodes and layers [Elbrächter et al., 2019]. With a large training set, DNNs have been shown in practice to be able to

learn the complex relations within high-dimensional data well and thus achieve a high prediction accuracy.

However, a major drawback of DNNs is their lack of quantification for uncertainty. Standard neural network regression methods aim at producing an optimal predictive function. but this is often done without accounting for estimation errors in model fitting (epistemic uncertainty) or unpredictable randomness in data (aleatoric uncertainty). To fit DNNs under the frequentist framework, a common approach is to minimize a loss function with respect to the weight parameters in the model. Gradient-based methods are often adopted for implementing the optimization procedure. The trained DNN is then used to generate a point prediction for each new sample. However, it remains unclear on how to accurately and appropriately estimate the standard errors of point predictions, since the weight parameters are not identifiable in a typical DNN model. Moreover, regularization methods are usually taken to obtain the point estimates, and these methods may pose challenges for statistical inferences (e.g. estimating the standard errors or constructing the confidence intervals) for weight parameters as well as point predictions in the DNN model. A practical solution is to add randomness in the optimization procedure (e.g. when initializing or updating the parameters) and independently train multiple versions of the model. An example is deep ensembles [Lakshminarayanan et al., 2016]. However, as a frequentist approach, methods in this category have no theoretical guarantee on the calibration of the confidence intervals [Guo et al., 2017]. Without accurate quantification of uncertainty in prediction, it is impossible to establish the reliability of DNNs even with a state-of-the-art architecture. The reliability of the prediction model is vitally important in many applications, especially the safety-critical ones. For example, if an assisted driving systems simply takes a point prediction of DNN to determine the next action on the road but fails to accurately measure the degree of confidence of the prediction, detrimental consequences can follow, including fatal

accidents [Huang et al., 2018].

In comparison, the Bayesian framework provides a natural way for measuring uncertainty in statistical models. Instead of searching for a best-fit model by minimizing a loss function, the Bayesian methods focus on the posterior distribution of the parameters of interests, which not only produces point estimates but also enables statistical inference. It is straightforward to use the posterior distributions to quantify the uncertainty in model fitting and outcome prediction. Bayesian modeling of DNNs can be traced back to as early as the 1990's [MacKay, 1995, Neal, 2012], and it has sustained an increasing interest in the recent years due to the rapid development in computation capacities in the past decade [Gal and Ghahramani, 2016, Wenzel et al., 2020, Wilson and Izmailov, 2020]. Bayesian DNNs (BDNNs) have been found to be not only practical for decision making under uncertainty but also boost the prediction accuracy of standard DNNs [Kendall and Gal, 2017, Izmailov et al., 2018].

Despite their natural representation of uncertainty, BDNNs face multiple statistical and computational difficulties. The posterior distributions of BDNNs not only lack closed-forms due to the nested nonlinearity across the hidden layers but also covers an ultrahigh number of parameters that need to be estimated. This makes posterior Markov chain Monte Carlo (MCMC) sampling extremely difficult for modern DNNs, which can have millions of weight parameters. Moreover, variational methods can provide fast approximations to the exact posterior distribution, but their unimodal nature makes it easy for them to be trapped at local modes, in which case the variance of the approximated posterior distribution could be significantly underestimated. Furthermore, due to their high degree of flexibility, many different DNNs can approximate the same function equally well, making it difficult to survey the posterior distribution among equivalent parametrization of the same target function. Finally, even though the model fitting uncertainty can be quantified by the posterior distribution of the parameters, the inherent randomness in data might still be over-

simplified if a homoscedastic noise is imposed in the model, since data density and noise level can vary greatly across the sample space. Moreover, the noise can contain multi-modality, heavy tails, and skewed outliers. In these cases, the uncertainty in the model will be underestimated by the posterior distribution, leading to over-confidence on predictions about the incoming data.

To address these challenges, we propose a novel model by introducing latent variables to the hidden nodes in the DNN. Instead of treating the whole DNN as a deterministic function and adding noise only to the last layer, we introduce aleatoric uncertainty to each hidden layer and assume all the intermediate values to be inherently noisy. We refer to this model a Bayesian deep aleatoric neural network (DALEA). Compared to standard BNNs, the incorporation of latent variables introduces greater degree of flexibility in learning the noise distribution. Although the latent variables are assumed to follow normal distributions, combining their effects with non-linear activation functions throughout the layers makes the final noise (i.e. deviation from the mean function, which does not have closed forms) no longer necessarily homoscedastic, uni-modal, or symmetric. Moreover, the proposed Bayesian hierarchical model with latent variables leads to closed forms for the full conditional distributions, based on which we develop a more efficient Gibbs sampler for posterior computation. We demonstrate the prediction accuracy and uncertainty quantification of DALEA via extensive simulations, comparing it against BDNNs with the Hamiltonian Monte Carlo sampler (HMC) [Neal et al., 2011], which has been shown to have a fast convergence rate and can survey the posterior distribution efficiently [Izmailov et al., 2021], and BDNNs with variational inference (VI) [Blei et al., 2017], which provides fast approximations of the posterior distribution. We investigate estimation and prediction accuracy, as well as the relation between widths of the credible intervals (which reflect the prediction confidence) and prediction errors. In addition, we apply DALEA to analysis of fMRI data in the the Adolescent Brain Cognitive

Development (ABCD) Study.

The rest of this Chapter is organized as follows. We first introduce the model in Section 4.2 by describing the details of its structure in Section 4.2.1. Then the conditional distributions are derived to develop a posterior sampling algorithm in Section 4.2.3. Next, we apply the methods to simulated data and show the experiment results in Section 4.3, followed by the experiments on neuroimaging data Section 4.4. We conclude this Chapter with a discussion in Section 4.5.

4.2 Bayesian Deep Aleatoric Neural Networks

We begin with basic notation. Let $\mathbb{R}^d$ represent a d-dimensional Euclidean vector space. Let $\mathbb{R}^{a\times b}$ represent the space of matrices with dimension a by b. All vectors are column vectors unless specified otherwise. Let $\mathbf{0}_d = (0,\ldots,0)^\top \in \mathbb{R}^d$, $\mathbf{1}_d = (1,\ldots,1)^\top \in \mathbb{R}^d$ and $\boldsymbol{I}_d$ be an $d \times d$ identity matrix. Let $\mathrm{N}(\mu, \Sigma)$ represent a normal distribution with mean $\mu \in \mathbb{R}^d$ and (co-)variance $\Sigma \in \mathbb{R}^{d\times d}$. Let $\mathrm{IG}(a, b)$ denote the inverse gamma distribution with shape a and rate b.

4.2.1 DNNs with latent variables

Suppose the observed data consist of vector-valued predictor variables $\boldsymbol{x}^{(n)} \in \mathbb{R}^P$ and vector-valued response variables $\boldsymbol{y}^{(n)} \in \mathbb{R}^Q$ for $n = 1,\ldots,N$. Our goal is to model the complex functional association between $\boldsymbol{y}^{(n)}$ and $\boldsymbol{x}^{(n)}$. For this purpose, we may first consider a standard DNN. Recall that in an L-layer feed-forward DNN, where the l^{th} hidden layer contains K_l $(l = 0,\ldots,L-1)$ units, the output layer has $K_L = Q$ units, and the input layer has $K_{-1} = P$ units, matching the dimensions of the response variable and the predictor variable respectively, the model can be

formulated recursively as follows:

$$\begin{aligned}
\boldsymbol{y}^{(n)} =& \boldsymbol{\epsilon}^{(n)} + \boldsymbol{\gamma}_L + \boldsymbol{\beta}_L \boldsymbol{u}_L^{(n)}, && \boldsymbol{\epsilon}^{(n)} \sim \mathrm{N}(\mathbf{0}_Q, \tau_L^2 \boldsymbol{I}_Q) \\
\boldsymbol{u}_{l+1}^{(n)} =& h\{\boldsymbol{\gamma}_l + \boldsymbol{\beta}_l \boldsymbol{u}_l^{(n)}\}, && \text{for } l = 0, \ldots, L-1 \\
\boldsymbol{u}_0^{(n)} =& \boldsymbol{x}^{(n)}
\end{aligned}$$

where $h(\cdot)$ is a nonlinear activation function, such as the rectified linear unit (ReLU) function $\max(0, \cdot)$ and the logistic function $[1 + \exp(\cdot)^{-1}]^{-1}$. The lth hidden layer consists of two sets of parameters: the weight parameter $\boldsymbol{\beta}_l = (\beta_{l,k,k'}) \in \mathbb{R}^{K_l \times K_{l-1}}$ and the bias parameter $\boldsymbol{\gamma}_l = (\gamma_{l,k}) \in \mathbb{R}^{K_l}$.

To quantify the uncertainty of the functional association between $\boldsymbol{y}^{(n)}$ and $\boldsymbol{x}^{(n)}$, we propose a deep aleatoric neural network (DALEA) model by introducing two sets of latent variables in the DNNs. In particular, we assume

$$\begin{aligned}
\boldsymbol{y}^{(n)} =& \boldsymbol{\epsilon}_L^{(n)} + \boldsymbol{\gamma}_L + \boldsymbol{\beta}_L \boldsymbol{u}_L^{(n)}, && \boldsymbol{\epsilon}_L^{(n)} \sim \mathrm{N}(\mathbf{0}_Q, \tau_L^2 \boldsymbol{I}_Q) \\
\boldsymbol{u}_{l+1}^{(n)} =& \boldsymbol{\delta}_l^{(n)} + h\{\boldsymbol{\epsilon}_l^{(n)} + \boldsymbol{\gamma}_l + \boldsymbol{\beta}_l \boldsymbol{u}_l^{(n)}\}, && \text{for } l = 0, \ldots, L-1 \\
\boldsymbol{\delta}_l^{(n)} \overset{\text{iid}}{\sim}& \mathrm{N}(\mathbf{0}_{K_l}, \sigma_l^2 \boldsymbol{I}_{K_l}), && \boldsymbol{\epsilon}_l^{(n)} \overset{\text{iid}}{\sim} \mathrm{N}(\mathbf{0}_{K_l}, \tau_l^2 \boldsymbol{I}_{K_l}) \\
\boldsymbol{u}_0^{(n)} =& \boldsymbol{x}^{(n)}.
\end{aligned} \tag{4.1}$$

In the l^{th} layer, latent variables $\boldsymbol{\delta}_l^{(n)} = (\delta_{l,k}^{(n)}) \in \mathbb{R}^{K_l}$ and $\boldsymbol{\epsilon}_l^{(n)} = (\epsilon_{l,k}^{(n)}) \in \mathbb{R}^{K_l}$ are independently and identically distributed across $n = 1, \ldots, N$. The introduction of the latent variables enables the model to represent complex noise distributions that cannot be fully characterized by the conditional mean and the conditional variance. An example of the flexibility of DALEA's conditional distribution is illustrated in Figure 4.1. DALEA is able to generate not only non-linear mean functions but also a variety of noise patterns. The conditional distribution is virtually Gaussian at ± 3. Near ± 2, the conditional distribution starts to become skewed, which gradually

evolves into bi-modal around ± 1 and eventually to tri-modal at 0. In this example, for any given value of the input variable, the conditional distribution is a Gaussian mixture. Intuitively, as Gaussian mixtures are universal approximators of densities (Plataniotis and Hatzinakos [2017], Calcaterra and Boldt [2008], Goodfellow et al. [2016, Sec. 3.9.6]) and DNNs are universal approximators of functions [Scarselli and Tsoi, 1998, Yarotsky, 2017, Lu and Lu, 2020], DALEA has the potential of being an universal approximator of conditional densities. In contrast, if a model only fits the conditional mean and the conditional variance (Figure 4.1, center panel), the noise structure will be over-simplified, which can results in inaccurate and and inefficient uncertainty quantification, especially when the true conditional distribution is heavy-tailed, skewed, or multi-modal. This problem is exacerbated if the conditional variance is further assumed to be constant and only the conditinal mean is learned (Figure 4.1, bottom panel), which is a practice common in standard applications of DNNs.

DALEA can be viewed as a frequentist model or a Bayesian model. To make statistical inferences on DALEA under the Bayesian framework, we assign the normal priors on the weight and bias parameters in the l^{th} layer, for $l = 0, \ldots, L$:

$$\beta_{l,k,k'} \overset{\text{iid}}{\sim} \mathrm{N}(0, \rho_l^2), \qquad \gamma_{l,k} \overset{\text{iid}}{\sim} \mathrm{N}(0, \xi_l^2).$$

We assign inverse gamma priors for the variance parameters, for $l = 0, \ldots, L$:

$$\rho_l^2 \overset{\text{iid}}{\sim} \mathrm{IG}(a_\rho, b_\rho), \quad \xi_l^2 \overset{\text{iid}}{\sim} \mathrm{IG}(a_\xi, b_\xi), \quad \tau_l^2 \overset{\text{iid}}{\sim} \mathrm{IG}(a_\tau, b_\tau), \quad \sigma_l^2 \overset{\text{iid}}{\sim} \mathrm{IG}(a_\sigma, b_\sigma).$$

4.2.2 Model representation

To develop efficient posterior computation algorithms, we consider an equivalent model representation of DALEA. Let $\boldsymbol{u}_0 = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}) \in \mathbb{R}^{P \times N}$ and

$\boldsymbol{v}_L = (\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(N)}) \in \mathbb{R}^{Q \times N}$. Moreover, let $\boldsymbol{\beta}_{l,k} = (\beta_{l,k,1}, \ldots, \beta_{l,k,K_{L-1}}) \in \mathbb{R}^{1 \times K_{l-1}}$ for $k = 1, \ldots, K_l$ and $l = 0, \ldots, L$. Then the DALEA model (4.1) can be equivalently represented as

$$\boldsymbol{v}_{l,k} \overset{\text{iid}}{\sim} \mathrm{N}(\gamma_{l,k}\mathbf{1}_N + \boldsymbol{\beta}_{l,k}\boldsymbol{u}_l, \tau_l^2 \boldsymbol{I}_N) \in \mathbb{R}^N, \quad \text{for } k = 1, \ldots, K_l, \qquad l = 0, \ldots, L \tag{4.2}$$

$$\boldsymbol{u}_{l+1,k} \overset{\text{iid}}{\sim} \mathrm{N}(h(\boldsymbol{v}_{l,k}), \sigma_l^2 \boldsymbol{I}_N) \in \mathbb{R}^N, \qquad \text{for } k = 1, \ldots, K_{l-1}, \quad l = 0, \ldots, L-1 \tag{4.3}$$

where $\boldsymbol{u}_l = (\boldsymbol{u}_{l,1}^\top, \ldots, \boldsymbol{u}_{l,K_{l-1}}^\top)^\top \in \mathbb{R}^{K_{l-1} \times N}$.

4.2.3 Posterior Computation

Compared to standard multi-layer feed-forward neural networks, including latent variables in DALEA not only allows for more model flexibility but also gives rise to conditional distributions with closed-forms, which enables us to develop a posterior sampling algorithm based on Gibbs sampling. In addition, the computation of the conditional distributions is parallelizable across samples and layers. The conditional distributions can be demonstrated by using Equations (4.2) and (4.3), which provides the full probability density function for all the parameters and data in the DALEA model. Before we start, we need to introduce the heterogeneous normal distribution, which will be used in the subsequent derivations.

Definition IV.1 (Heterogeneous normal distribution)**.** Let $J \geq 1$ and

$$-\infty = c_0 < c_1 < c_{J-1} \ldots < c_J = \infty$$

$$\mu_1, \ldots, \mu_J \in \mathbb{R}$$

$$\tau_1, \ldots, \tau_J \in \mathbb{R}.$$

Define ϕ and Φ to be the PDF and CDF of the standard normal distribution, respectively, and let ψ be the PDF of the truncated normal distribution with mean μ and

variance τ^2 inside the interval $[c', c'']$:

$$\psi(x|\mu, \tau^2, c', c'') = \mathbb{I}\{x \in [c', c'']\} \frac{\tau^{-1}\phi(\frac{x-\mu}{\tau})}{\Phi(\frac{c''-\mu}{\tau}) - \Phi(\frac{c'-\mu}{\tau})}$$

Then

$$x \sim \text{HN}[(c_1, \mu_1, \tau_1^2), \ldots, (c_J, \mu_J, \tau_J^2)]$$

if and only if the PDF of x equals

$$f[x|(c_1, \mu_1, \tau_1^2), \ldots, (c_J, \mu_J, \tau_J^2)]] = \sum_{j=1}^{J} \mathbb{I}\{x \in (c_{j-1}, c_j]\} \pi_j \psi(x|\mu_j, \tau_j^2, c_{j-1}, c_j),$$

where

$$\pi_j = \frac{\zeta_j}{\sum_{j=1}^{J} \zeta_j}, \quad \zeta_j = \begin{cases} 1, & j = 1 \\ \nu_j \zeta_{j-1}, & 1 < j \leq J \end{cases}, \quad \nu_j = \frac{\psi(c_{j-1}|\mu_{j-1}, \tau_{j-1}^2, c_{j-2}, c_{j-1})}{\psi(c_{j-1}|\mu_j, \tau_j^2, c_{j-1}, c_j)}.$$

According to Definition IV.1, a heterogeneous normal distribution by definition is a mixture of a finite number of truncated normal distributions whose supports form a partition of $\mathbb{R}$. Moreover, it is straightforward to show that the PDF of a heterogeneous normal distribution is continuous, though not necessarily differentiable at the border points $c_1, \ldots, c_{J-1}$.

Moreover, to derive the full conditional distributions of the parameters in DALEA, we need the the following assumption on the activation function.

Assumption IV.2. *The activation function $h(\cdot)$ is a continuous piecewise linear function with a finite number (call it J) of linear components:*

$$h(x) = \sum_{j=1}^{J} (b_j x + b_j') \cdot \mathbb{I}\{x \in [c_{j-1}, c_j]\}$$

for some coefficients $b_j, b_j' \in \mathbb{R}$ and $-\infty < c_1 < \ldots < c_{J-1} < \infty$ that satisfy $b_j c_j + b_j' =$

$b_{j+1}c_j + b'_{j+1}$ *for* $j = 1, \ldots, J-1$.

The family of functions defined in Assumption IV.2 includes many common activation functions, such as ReLU ($\max[0, x]$), leaky ReLU ($\max[0, x] + \min[0, rx]$ for some constant $r > 0$), truncated ReLU ($\min[1, \max[0, x]]$), and hard sigmoid ($\min[1, \max[0, \frac{1}{2} + rx)]]$ for some constant $r > 0$). Smooth activation functions are not piecewise linear, but many of them they can be approximated by one of such. For example, the logistic function $[1 + \exp(-x)]^{-1}$ can be approximated by the hard sigmoid function, while the softplus function $\log[1 + \exp(x)]$ can be approximated by the ReLU function.

Given Assumption IV.2, the pre-activation latent variables $\boldsymbol{v}_{l,k}^{(n)}$ conditioned on the other variables, denoted as "rest", independently follow heterogeneous normal distributions. For brevity, we only present the case for the truncated ReLU activation function:

$$
\begin{aligned}
v_{l,k}^{(n)} \;\Big|\; \text{rest} \sim \text{HN}[&(c_1, \mu_{1,v,l,k,n}, \upsilon_{1,v,l,k,n}), (c_2, \mu_{2,v,l,k,n}, \upsilon_{2,v,l,k,n}), (c_3, \mu_{3,v,l,k,n}, \upsilon_{3,v,l,k,n})] \\
\upsilon_{1,v,l,k,n}^2 =& \upsilon_{3,v,l,k,n}^2 = \tau_l^2 \\
\upsilon_{2,v,l,k,n}^2 =& \tau_l^{-2} + \sigma_l^{-2} \\
\mu_{1,v,l,k,n} =& \mu_{3,v,l,k,n} = \gamma_{l,k} + \boldsymbol{\beta}_{l,k}\boldsymbol{u}_l^{(n)} \\
\mu_{2,v,l,k,n} =& \frac{\tau_l^{-2}}{\tau_l^{-2} + \sigma_l^{-2}}(\gamma_{l,k} + \boldsymbol{\beta}_{l,k}\boldsymbol{u}_l^{(n)}) + \frac{\sigma_l^{-2}}{\tau_l^{-2} + \sigma_l^{-2}} u_{l+1,k}^{(n)} \\
c_1 =& 0,\ c_2 = 1,\ c_3 = \infty.
\end{aligned}
$$

Next, we show that the weight parameters $\boldsymbol{\beta}_l, \boldsymbol{\gamma}_l$ and the post-activation latent variable $\boldsymbol{u}_l$ both have conditional distribution being a normal distribution. Recall that in Bayesian linear regression, if the data $\boldsymbol{X} \in \mathbb{R}^{K \times N}$, $\boldsymbol{y} \in \mathbb{R}^{1 \times N}$ and the weight

parameter $\boldsymbol{\beta} \in \mathbb{R}^{1\times K}$ satisfy

$$\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\beta}\boldsymbol{X}, \tau^2\boldsymbol{I}_N), \qquad \boldsymbol{\beta}|\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \qquad \boldsymbol{\Sigma}_0 = \rho^2\boldsymbol{I}_K$$

for some known noise variance τ^2 and prior parameter parameter ρ^2, then the posterior distribution of $\boldsymbol{\beta}$ has conjugate form

$$\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X} \sim& \mathrm{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
\boldsymbol{\Sigma}_N =& (\tau^{-2}\boldsymbol{X}\boldsymbol{X}^\top + \rho^{-2}\boldsymbol{I}_K)^{-1} \\
\boldsymbol{\mu}_N =& (\boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{\Sigma}_0^{-1})^{-1}(\boldsymbol{\mu}_0\boldsymbol{\Sigma}_0^{-1} + \boldsymbol{y}\boldsymbol{X}^\top).
\end{aligned}$$

Then the weight parameters $\boldsymbol{\beta}_l, \boldsymbol{\gamma}_l$ in DALEA have conditional distribution

$$\begin{aligned}
(\boldsymbol{\beta}_{l,k}, \gamma_{l,k}) \,\Big|\, \text{rest} \sim& \mathrm{N}(\boldsymbol{\mu}_{\beta,\gamma,l,k}, \boldsymbol{\Sigma}_{\beta,\gamma,l,k}) \\
\boldsymbol{\Sigma}_{\beta,\gamma,l,k} =& (\tau_l^{-2}\bar{\boldsymbol{u}}_l\bar{\boldsymbol{u}}_l^\top + \boldsymbol{\Sigma}_{0,\beta,\gamma,l,k}^{-1})^{-1} \\
\boldsymbol{\mu}_{\beta,\gamma,l,k} =& (\bar{\boldsymbol{u}}_l\bar{\boldsymbol{u}}_l^\top + \boldsymbol{\Sigma}_{0,\beta,\gamma,l,k}^{-1})^{-1}\boldsymbol{v}_{l,k}\bar{\boldsymbol{u}}_l^\top \\
\boldsymbol{\Sigma}_{0,\beta,\gamma,l,k} =& \mathrm{diag}(\rho_l^2, \ldots, \rho_l^2, \xi_l^2)
\end{aligned}$$

where $\bar{\boldsymbol{u}}_l = (\boldsymbol{u}_l, \boldsymbol{1})$. Moreover, by swapping and transposing the input variables with the weight parameters in the Bayesian linear regression model with conjugate priors, the latent variable $\boldsymbol{u}_l$ conditioned on the other variables follows

$$\begin{aligned}
\boldsymbol{u}_l^{(n)} \,\Big|\, \text{rest} \sim& \mathrm{N}(\mu_{u,l,n}, \boldsymbol{\Sigma}_{u,l,n}) \\
\boldsymbol{\Sigma}_{u,l,n} =& (\tau_l^{-2}\boldsymbol{\beta}_l^\top\boldsymbol{\beta}_l + \boldsymbol{\Sigma}_{0,u,l,n}^{-1})^{-1} \\
\boldsymbol{\mu}_{\beta,\gamma,l,k} =& (\boldsymbol{\beta}_l\boldsymbol{\beta}_l^\top + \boldsymbol{\Sigma}_{0,u,l,n}^{-1})^{-1}(\boldsymbol{\Sigma}_{0,u,l,n}^{-1}\boldsymbol{w}_l^{(n)} + \boldsymbol{\beta}_l^\top\bar{\boldsymbol{v}}_l^{(n)}) \\
\boldsymbol{\Sigma}_{0,u,l,n} =& \sigma_{l-1}^2\boldsymbol{I}_{K_{l-1}}, \quad \boldsymbol{w}_l^{(n)} = h(\boldsymbol{v}_{l-1}^{(n)}), \quad \bar{\boldsymbol{v}}_l^{(n)} = \boldsymbol{v}_l^{(n)} - \boldsymbol{\gamma}_l.
\end{aligned}$$

Finally, the full conditionals of the variance parameters are given by

$$\begin{aligned}
\tau_l^2 \;\Big|\; \text{rest} &\sim \text{IG}\left(a_\tau + \frac{1}{2}K_l N, b_\tau + \frac{1}{2}\|\boldsymbol{v}_l - \boldsymbol{\beta}\boldsymbol{u}_l\|_F^2\right) \\
\sigma_l^2 \;\Big|\; \text{rest} &\sim \text{IG}\left(a_\sigma + \frac{1}{2}K_l N, b_\sigma + \frac{1}{2}\|\boldsymbol{u}_{l+1} - h(\boldsymbol{v}_l)\|_F^2\right) \\
\rho_l^2 \;\Big|\; \text{rest} &\sim \text{IG}\left(a_\rho + \frac{1}{2}K_l K_{l-1}, b_\rho + \frac{1}{2}\|\boldsymbol{\beta}_l\|_F^2\right) \\
\xi_l^2 \;\Big|\; \text{rest} &\sim \text{IG}\left(a_\xi + \frac{1}{2}K_l, b_\xi + \frac{1}{2}\|\boldsymbol{\gamma}_l\|_2^2\right)
\end{aligned}$$

Based on the conditional distributions, we obtain the posterior Gibbs sampling algorithm in Algorithm 1. The computation complexity (per posterior sample) is $\mathcal{O}[NLK^2]$, where N is the training sample size, L is the number of layers, and K is the average number of nodes in each layer. By using mini-batch methods, this rate can be reduced to $\mathcal{O}[N_{\text{batch}}LK^2]$, where N_{batch} is the size of the mini-batches. Notice that the algorithm outputs M many posterior samples of the predictive distribution. For any $\tilde{M} \in \mathbb{Z}_+$, each predictive distribution $f^{[m]}(\cdot|\cdot)$ $(1, \ldots, M)$ can be used as a sampler based on a new sample $\boldsymbol{x}_{\text{new}}$ to draw $\tilde{M}$ many predicted samples $\boldsymbol{y}_{\text{new},m,1}, \ldots, \boldsymbol{y}_{\text{new},m,\tilde{M}} \overset{\text{iid}}{\sim} f^{[m]}(\boldsymbol{y}_{\text{new}}|\boldsymbol{x}_{\text{new}})$, which in total samples $M\tilde{M}$ many predictions. The sample dimension $1, \ldots, M$ corresponds to variation due to uncertainty in model fitting (epistemic), which can be reduced by increasing the sample size, while the dimension $1, \ldots, \tilde{M}$ corresponds to uncertain inherent to the model (aleatoric), which cannot be reduced by increasing the sample size.

In addition, although our work primarily focuses on continuous outcomes, DALEA is also able to analyze categorical data, where the softmax function is approximated by continuous piecewise linear functions. A detailed description of the categorical DALEA model is presented in Appendix C.1.

4.3 Simulations

4.3.1 Experiment setup

We simulated data with nonlinear mean functions and heteroscedastic, non-Gaussian noise distributions. To better visualize the analysis, both the predictors and outcomes are one-dimensional. The predictors $x^{(n)}$ $(1, \ldots, N)$ are randomly drawn between -4 and 4 with different sample density to emulate uneven distribution of samples in real data. We sampled 200 training samples, with 10% of them uniformly sampled in (-4, -2), 40% in (-2, 0), 40% in (0, 2), and 10% in (2, 4). For the testing set, we drew 2000 samples uniformly in (-4, 4). The outcome is set to $y^{(n)} = g(x^{(n)}) + \epsilon^{(n)}$, with mean function being $g(x) = \sin(\pi x) + 2x$ and noise sampled from Gaussian, chi-squared, and Gaussian mixture distributions. In the first setting, the noise are sampled from zero-mean Gaussian distributions. The standard deviation is set to 0.2 for $x \in (-4, 0)$ and 1.0 for $x \in (0, 4)$. In the second setting, the noise distribution is the same as in the first setting except that the zero-mean Gaussian distribution is replaced with centered and rescaled chi-squared distribution with one degree of freedom $(\chi_1^2 - 1)/\sqrt{2}$. In the third setting, we make the noise follow mixture of Gaussian distributions. For $x \in (-4, 0)$, the noise is sampled from $\mathrm{N}(0, 0.1)$. For $x \in (0, 4)$, it follows a mixture of $\mathrm{N}(-1, 0.1)$ and $\mathrm{N}(1, 0.1)$ with equal weights. Examples of the simulation data design are illustrated in Figure 4.2.

We compared the performance of DALEA with standard BDNN optimized by Hamiltonian Monte Carlo (HMC) and variational inference (VI). DNN ensembles (DNNE) and deterministic DNN with optimizers based on mini-batch stochastic gradient descent (SGD) was also included in our experiments. Although deterministic DNNs cannot quantify uncertainty, we used them as a reference for the point estimates (e.g. posterior means) of DALEA and the other alternative methods. Each experimental setting was replicated for 25 times. For the neural network architecture,

we used one hidden layer with 32 nodes, activated by the bounded ReLU function. We used weak prior distribution InverseGamma(0.001, 0.001) for the parameter weights, as well as for the latent variables in DALEA. We took 18000 burn-in samples took another 6000 samples and thinned them into 600 samples. This procedure was repeated on 5 independent chains. To reduce the computation cost, we initialize the weight parameters by fitting the model with SGD. For HMC, we set the number of leap frog steps to 10 and initialized the step size to 0.001 multiplied by the standard deviation of the parameters across the independent chains. In the burn-in process, we dynamically adjust the step size to make the acceptance rate close to 0.6. For DNNE, we set the number of independent networks to 5, as recommended in Lakshminarayanan et al. [2016]. For SGD, we used the ADAM [Kingma and Ba, 2014] optimizer.

For performance evaluation, we first access each methods' prediction accuracy by the MSE of the point estimates (posterior mean for DALEA, HMC, and VI, ensemble mean for DNNE, and model output for SGD) for the true mean function on the testing set. Then we compare their uncertainty quantification by examining the correlation between the CI width and the estimation error. Correlations are computed with Kendall correlation coefficient [Kendall, 1945] in order to ensure robustness against outliers. Finally, we divide the testing samples into 5 strata based on the quantile of the CI and show the stratified MSE between the posterior mean and the true mean function, which demonstrates the relation between the estimation confidence and the estimation error.

4.3.2 Experiment results

Figure 4.3 shows the estimation MSE of the posterior mean. In all the three settings, DALEA achieved the best estimation accuracy, as its posterior mean had the uniformly lowest MSE. DNNE had the highest estimation errors, followed by SGD and then by VI, whose MSEs were at least three times higher than those of DALEA.

Compared VI and SGD, HMC's performance was closer to DALEA, but its MSE was still significantly higher. Moreover, the IQR of DALEA is overall lower than that of the baseline methods, (although this pattern is less visible on the box plots since they are log-scaled), reflecting a higher estimation stability. Across the experimental settings, the performance of the different methods had similar trends. All the methods achieved the lowest median MSE on the Gaussian noise, which is expected due to its simpler structure compared to the other noise distributions. In comparison, the chi-square noise elevated the median MSE slightly and widened the IQR, a sign for decreased estimation stability, potentially caused by the high proportion of skewed outliers. The highest MSE was observed in the Gaussian mixture noise data. These results suggest that the multi-modal characteristics of the Gaussian mixture distributions making the estimation task challenging for all methods. Within each type of noise, DALEA was the most accurate and overall the most stable compared to the alternative methods. In addition, the Bayesian methods achieved lower estimation MSE than the frequentist methods. This trend shows that the extra computation effort spent on uncertainty quantification not only provides information on the degree of confidence but also improves the accuracy of the estimates.

Moreover, for uncertainty quantification, DALEA had the most positive correlation between estimation confidence and estimation error for the Gaussian and chi-squared noises. In the case of Gaussian mixture noise, the median correlation for DALEA is about the same as that of DNNE, though the latter had narrower IQR. Furthermore, we stratified the MSE of the posterior mean by the width of the CIs to examine the usefulness of the latter as a predictor of the former. The results are shown in Section 4.6. There was a clear trend for DALEA between the quantile of CI width and the estimation MSE. For the data with Gaussian chi-squared noise, the median MSE of the posterior mean increased monotonically across the strata of the CI widths. This relation shows a close and positive correspondence between estima-

tion confidence and estimation error for DALEA. Similar trend was observed for the Gaussian mixture noise, although the the last two strata exhibited more fluxtuation. In contrast, the CIs of HMC was not as monotonic as DALEA in all the settings. DNNE had overall positive trends across the strata, but the increment was almost binary, with the MSE having two distinc groups of levels for CI quantiles less than 0.4 and those greater than 0.4. In addition, there was no clear trend for VI between the CI width and estimation MSE. Overall, compared to the baseline methods, the degree of estimation confidence by DALEA corresponds more closely to the actual estimation error, which shows its higher reliability in uncertainty quantification.

Finally, we illustrate examples of the posterior distributions by DALEA and HMC in Figures D.1 to D.3. For the experiments with Gaussian noise, it is clear that DALEA's CI width is closely related to noise level and data density. The training samples in (-4, 0) is much less noisy than those in (0, 4), and this difference is distinctively reflected in the CI width. Moreover, inside the interval of (-4, 0), the training data points are denser in (-2, 0) than in (-4, -2). As a result, CI is narrower in the former than in the latter. The same trend is observed for the (0, 2) interval vs (2, 4) interval. In addition, regions with wider CI also tend to have less curvature in the posterior mean, showing the model's preference for simpler structure in the absence of abundant data. In contrast, the CI band for HMC is much more constant. For example, the samples are less noisy in (-4, 0) than in (0, 4), but the CI is about the same width in the two intervals. Inside (-4, 0), the predictive CI over-covers the testing samples by a large margin, wasting the extra width of the CI. On the other hand, the true mean function in (0, 4) is more difficult to estimate, as shown by the greater deviation of the posterior mean from the true mean. However, instead of increasing the CI width in this region, HMC attempted to overfit the data, which is indicated by the close alignment between the erroneous spikes in the posterior mean and the outliers in the training data Figure 4.2. As a result, the posterior

distribution of HMC is more bumpy. Compared to HMC, DALEA is less sensitive to outliers, since it is able to accommodate samples far away from the mean by adjusting the corresponding latent variables without drifting the mean function too much. Moreover, VI completely missed the periodic fluctuations of the true mean function and only fit the broader increasing trend. Thus its CI varied little across the input space, since deviation of its posterior mean from the true mean was dominated by the periodic error. DNNE had similar patterns as VI, although the estiamted mean and CI width are more linear. and the CI width less constant. Overall, compared to the alternative methods, DALEA produced more informative CI and more robust posterior mean, which resulted in higher estimation accuracy and better uncertainty quantification.

Similar patterns were also observed on the datasets with chi-squared and Gaussian mixture noise. In the chi-squared case (Figure D.2), HMC was even more prune to outliers, which occurred at a higher frequency in comparison to the Gaussian noise data, while DALEA still produced accurate posterior mean and CI width, without being too influenced by the outlying training samples. VI, on the other hand, omitted the sinusoidal fluctuation in the true mean function as in the case of the Gaussian noise, which caused low estimation accuracy and almost constant CI width. The performance of DNNE was similar to that for Gaussian noise. Finally, in the case of Gaussian mixture noise, all the methods faced greater estimation difficulty, but the posterior distributions by different methods had very different characteristics. However, for example in the interval of (0, 4) with bimodal outcomes, HMC still tired to fit the non-normally distributed training samples by increasing the curvature of the estimated function, but DALEA, in the contrary, made the mean function smoother and increased the width of the CI to accommodate for the greater uncertainty in the data, which resulted in a CIs with more constant width in (0, 4), although there were still uniformly wider than in the less noisy (-4, 0) interval. For VI and DNNE, the

CI width mostly depended on the modality of the data, and the estimation error was dominated by the periodic pattern in the true mean function and had little relation with the CI with. Altogether, these characteristics of the posterior distributions illustrate that DALEA is able to not only fit highly nonlinear mean functions but also adjust the confidence level appropriately inside highly noisy regions.

4.4 Analysis of neuroimaging data

We further evaluate the performance of DALEA by applying it to the fMRI data in the Adolescent Brain Cognitive Development (ABCD) Study. The ABCD data includes the task fMRI images and other clinical characteristics of 11,800 children from multiple study sites in the U. S. In our analysis, we used the minimally processed contrast map for the 2-back task, which is designed to engage brain regions for cognitive functions including memory regulation. Moreover, we divided the whole brain volume into 90 regions according to the automated anatomical labeling (AAL) atlas and take the mean of the voxels inside each region, which resulted in 90 model features. We aimed at using these input variables, as well as age, sex, and the psychiatric diagnostic score, to predict cognitive ability (CA), which is measured by the general CA component score. Both the model features and the outcome variable were inverse-normal transformed. After removing missing values, our data contained 1,911 subjects.

We replicated the experiment for 25 times. For each replicate, 75% of the samples were selected to form the training set, while the other 25% were used for testing. DALEA and the baseline methods were applied to the ABCD data, with the hyperparameters being the same as those in the simulation studies (Section 4.3.1). We accessed prediction accuracy by the Kendall correlation between the posterior mean and the testing samples. To evaluate uncertainty quantification, we first computed the MSE between the posterior mean on the testing samples. Then we calculated the

Kendall correlation between the MSE and the CI width.

Figure 4.6a shows the correlation of each method's posterior mean with the testing samples. Neither method achieved a very high testing r^2, which indicated a high noise level. The prediction correlation of DALEA was higher than all the other methods except DNNE. However, DALEA had the most positive correlation between the CI width and the testing MSE, as shown in Figure 4.6b. Since the prediction difficulty was already high, the task of uncertainty quantification was no simpler, which caused the low MSE-CI correlations for all the methods. But even so, DALEA still achieved MSE-CI correlations that are significantly higher than all the other methods. In fact, the MSE-CI correlation for DALEA was positive in all of the 25 replicates, while the baseline methods had negative MSE-CI correlations in at least 6 replicates. Overall, in the analysis of the ABCD data, the esimation accuracy of DALEA was higher than standard DNN and baseline Bayesian DNN methods, while the correlation bewteen accuracy and confidence for DALEA was higher than that of all the existing methods for DNN uncertainty quantification.

4.5 Discussion

In this work, we have presented a novel Bayesian deep neural network model with the capacity of representing complex noise structure in the data. By adding normal latent variables to the intermediate values both before and after activation, our DALEA model redistributes the noise from the last layer in the DNN to every layer in the DNN, making it possible to adapt for heteroscedasticity and other intricate relations in the aleatoric uncertainty. In light of the model structure of DALEA, we have developed a Gibbs sampling algorithm that allows us to sample from the posterior distribution without using computationally expensive Metropolis-Hastings-based MCMC methods or uni-modal approximations by variational inference methods. Moreover, we have demonstrated the effectiveness of DALEA both in terms of estimation accuracy and

uncertainty quantification by comparing it against BDNN with Hamiltonian Monte Carlo, BDNN with variational inference, DNN ensembles, and standard DNNs on simulated data and neuroimaging data. In the extensive simulation studies, DALEA has been shown to be able to adjust the CI width efficiently to represent the degree of estimation uncertainty at each local point. Compared to standard DNNs and DNN ensembles DALEA has achieved higher estimation and prediction accuracy. Compare to BNN with HMC and VI, DALEA has not only produced more accurate posterior means but also generated credibility intervals that are more closely correlated with the true estimation error, even in the presence of skewed or multimodal noises. Furthermore, DALEA has been shown to be robust against outliers. Instead of shifting the whole estimated mean function to accommodate for the outlying observations, a trend that is common for HMC, DALEA can adjust the CI width to reflect the higher degree of uncertainty caused by the outliers. Finally, in the analysis of the fMRI data in ABCD, DALEA has achieved higher prediction accuracy than standard DNNs and outperformed all the alternative methods in uncertainty quantification.

For furture work, one limitation of DALEA that we have noticed is the unidentifiability of the weight parameters, which is also a difficulty in standard DNN models. Imposing more stringent structures on the weights and reducing the number of possible equivalent parametrizations can potentially improve the efficiency and performance of the model. In addition, we plan on applying our model to biomedical datasets with various sample sizes and data dimensions to further evaluate its estimation accuracy and uncertainty quantification. We look forward to the future development of DALEA that will provide more effective and efficient tools for analyzing high-dimensional complex biomedical data.

4.6 Tables and Figures

Algorithm 1: Posterior Gibbs sampler for DALEA

Input: data $\boldsymbol{x}$, $\boldsymbol{y}$, and hyper parameters

Output: Posterior samples of the predictive distribution $\mathcal{F} = \left\{ f^{[1]}(\cdot|\cdot), \ldots, f^{[M]}(\cdot|\cdot) \right\}$

Initialize $\mathcal{F} \leftarrow \{\}$, $\boldsymbol{u}_{-1} \leftarrow \boldsymbol{x}$, $\boldsymbol{v}_L \leftarrow \boldsymbol{y}$;
for $l \leftarrow 0, \ldots, L$ **do**
  Sample $\rho_l^2 \overset{\text{iid}}{\sim} \text{IG}(a_\rho, b_\rho)$, $\xi_l^2 \overset{\text{iid}}{\sim} \text{IG}(a_\xi, b_\xi)$, $\tau_l^2 \overset{\text{iid}}{\sim} \text{IG}(a_\tau, b_\tau)$;
  Sample $\boldsymbol{\beta}_l \overset{\text{iid}}{\sim} \text{N}(0, \rho_l^2)$, $\boldsymbol{\gamma}_l \overset{\text{iid}}{\sim} \text{N}(0, \xi_l^2)$, $\boldsymbol{v}_l \sim \text{N}(\boldsymbol{\gamma}_l + \boldsymbol{\beta}_l \boldsymbol{u}_l, \tau_l^2)$;
  if $l > 0$ **then**
    Sample $\sigma_l^2 \overset{\text{iid}}{\sim} \text{IG}(a_\sigma, b_\sigma)$, $\boldsymbol{u}_{l+1} \sim \text{N}[h(\boldsymbol{v}_l), \sigma_l^2]$;
  end
end
for $m \leftarrow -m_{\text{burnin}} + 1, \ldots, -1, 0, 1, \ldots, M$ **do**
  for $m' \leftarrow 1, \ldots, m_{\text{thinnning}}$ **do**
    for $l \leftarrow 0, \ldots, L$ **do**
      Sample $\boldsymbol{\beta}_l, \boldsymbol{\gamma}_l \sim f(\boldsymbol{\beta}_l, \boldsymbol{\gamma}_l | \boldsymbol{u}_l, \boldsymbol{v}_l, \tau_l, \rho_l)$;
      Sample $\boldsymbol{v}_l \sim f(\boldsymbol{v}_l | \boldsymbol{\beta}_l, \boldsymbol{\gamma}_l, \boldsymbol{u}_l, \boldsymbol{u}_{l+1}, \tau_l, \sigma_l)$;
      Sample $\boldsymbol{u}_l \sim f(\boldsymbol{u}_l | \boldsymbol{\beta}_l, \boldsymbol{\gamma}_l, \boldsymbol{v}_l, \boldsymbol{v}_{l-1}, \tau_l, \sigma_{l-1})$;
      Sample $\tau_l^2 \sim f(\tau_l^2 | \boldsymbol{u}_l, \boldsymbol{v}_l, \boldsymbol{\beta}_l, \boldsymbol{\gamma}_l)$;
      Sample $\sigma_l^2 \sim f(\sigma_l^2 | \boldsymbol{u}_{l+1}, \boldsymbol{v}_l)$;
      Sample $\rho_l^2 \sim f(\rho_l^2 | \boldsymbol{\beta}_l)$;
      Sample $\xi_l^2 \sim f(\xi_l^2 | \boldsymbol{\gamma}_l)$;
    end
  end
  if $m > 0$ **then**
    $\boldsymbol{\beta}^{[m]} \leftarrow \boldsymbol{\beta}_0, \ldots, \boldsymbol{\beta}_L$;
    $\boldsymbol{\gamma}^{[m]} \leftarrow \boldsymbol{\gamma}_0, \ldots, \boldsymbol{\gamma}_L$;
    $\boldsymbol{\tau}^{[m]} \leftarrow \tau_0, \ldots, \tau_L$;
    $\boldsymbol{\sigma}^{[m]} \leftarrow \sigma_0, \ldots, \sigma_{L-1}$;
    Obtain the predictive distribution from the parameters:
      $f^{[m]}(\cdot|\cdot) \leftarrow f_{\boldsymbol{\beta}^{[m]}, \boldsymbol{\gamma}^{[m]}, \boldsymbol{\sigma}^{[m]}, \boldsymbol{\tau}^{[m]}}(\cdot|\cdot)$;
    Save the distribution: $\mathcal{F} \leftarrow \mathcal{F} \cup \{ f^{[m]}(\cdot|\cdot) \}$;
  end
end

Figure 4.1: Comparison of the conditional distribution of different models. Colors in the heatmap represent the conditional density. The red solid line corresponds the conditional mean, while the orange dashed lines correspond the 0.025-0.975 conditoinal quantiles. The top panel illustrates the conditional distribution of a DALEA model. The center and bottom panels show the conditional distributions with homoscedastic and heteroscedastic Gaussian noise, respectively, that best approximate that of the DALEA model.

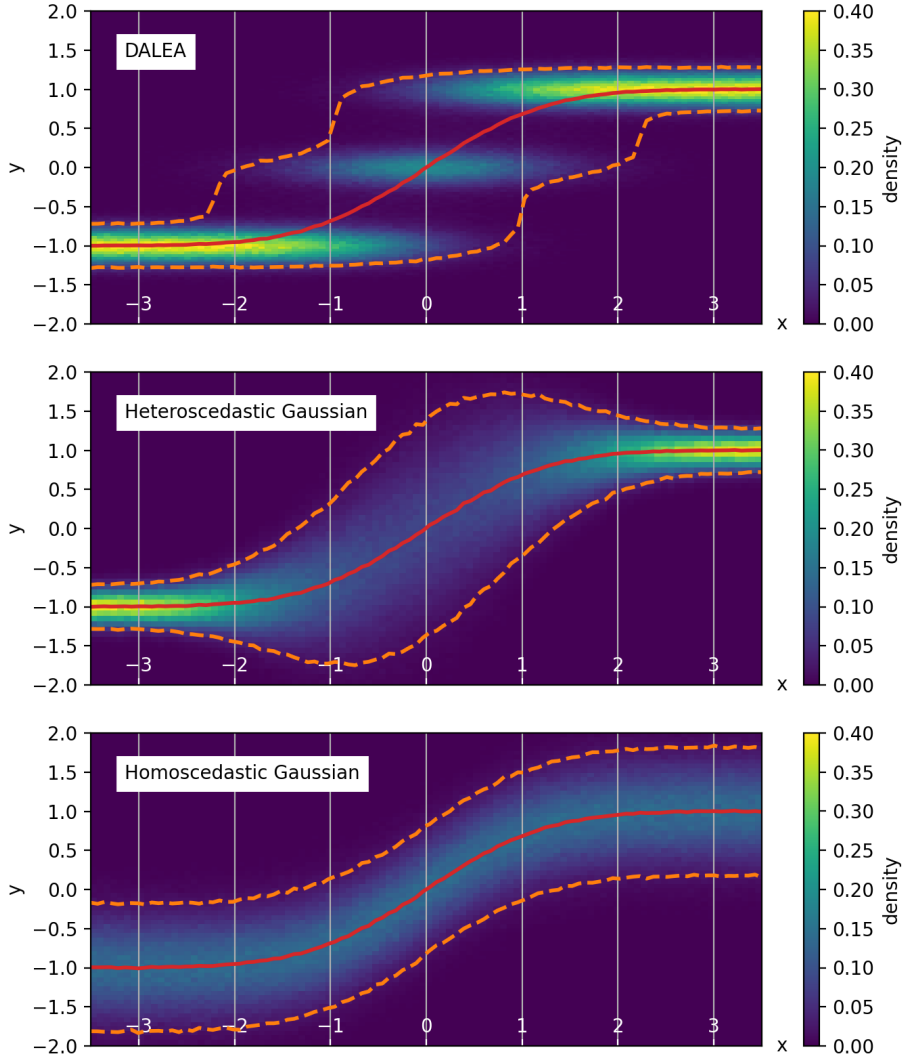
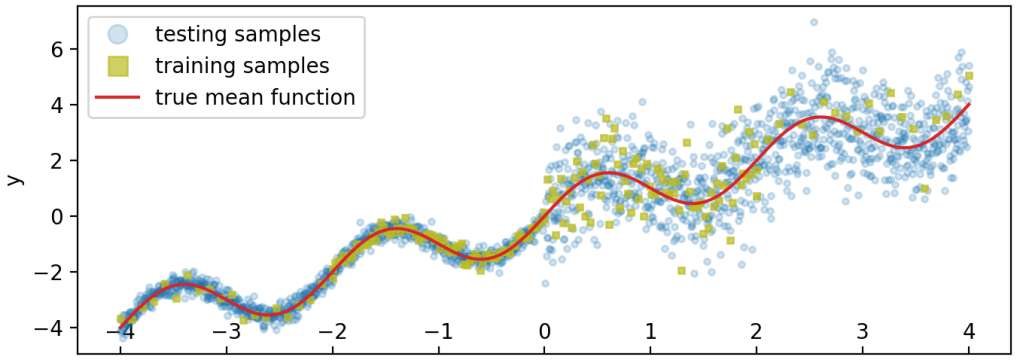
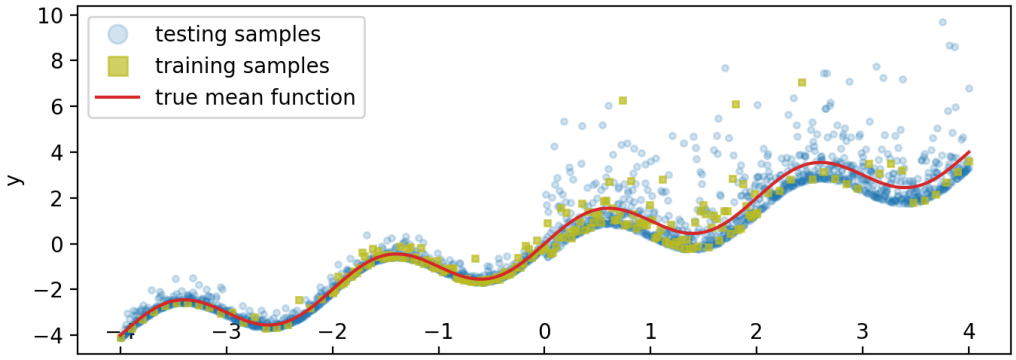


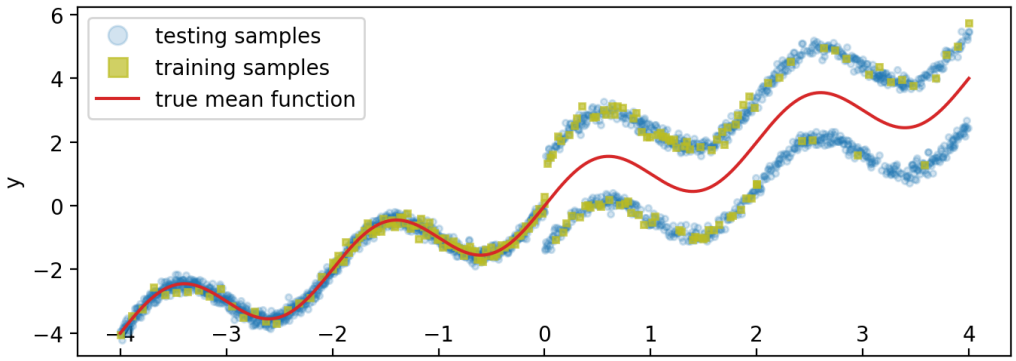
Figure 4.2: Data design in simulation studies.



(a) Gaussian noise



(b) Chi-squared noise



(c) Gaussian mixture noise

Figure 4.3: Estimation accuracy in simulated data. Accuracy is measured by the MSE (shown in $\log_{10}$ scale) between the true mean function (on the testing samples) and the point estimates for them. Point estimates are posterior mean for DALEA, HMC, and VI, ensemble mean for DNNE, and the output of the trained DNN for SGD.

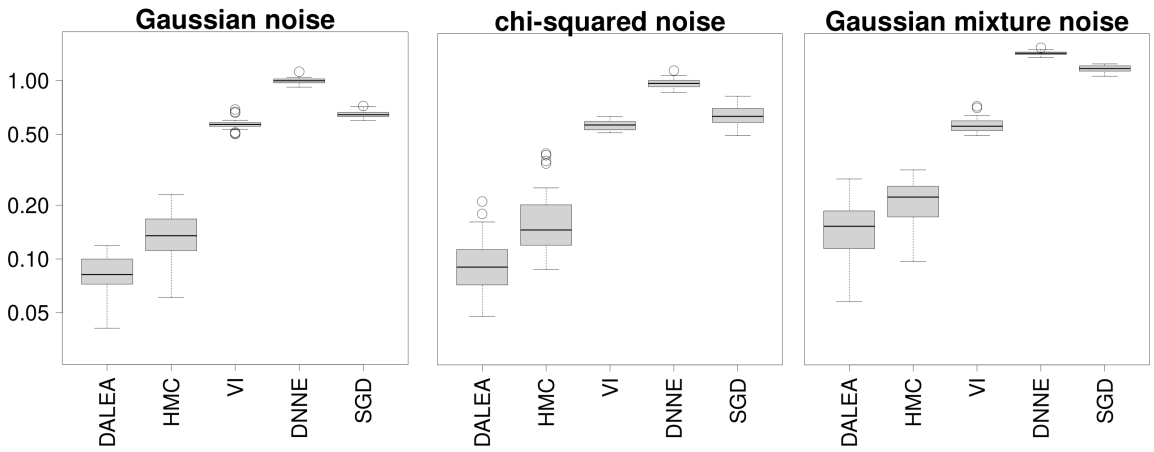


Figure 4.4: Correlation between CI width and estimation error of the posterior mean.

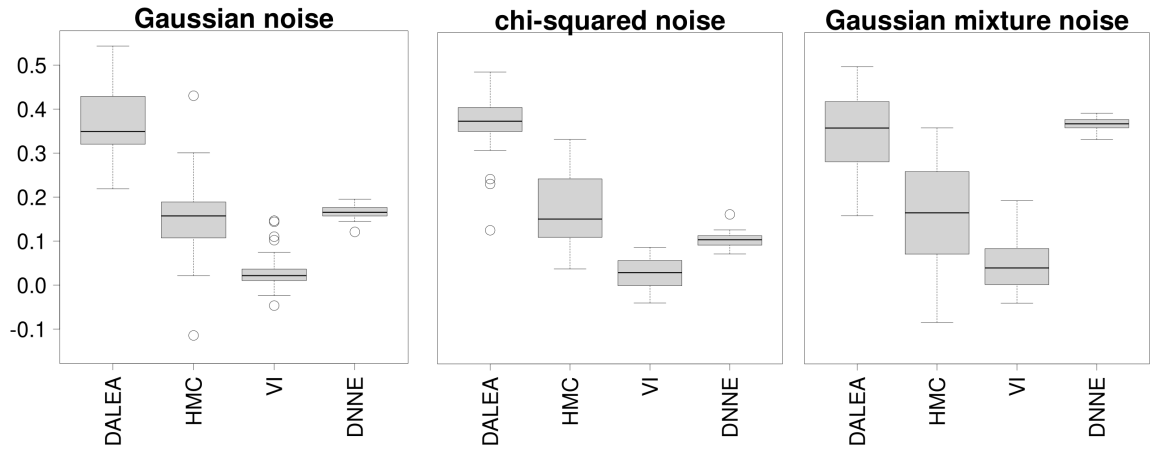
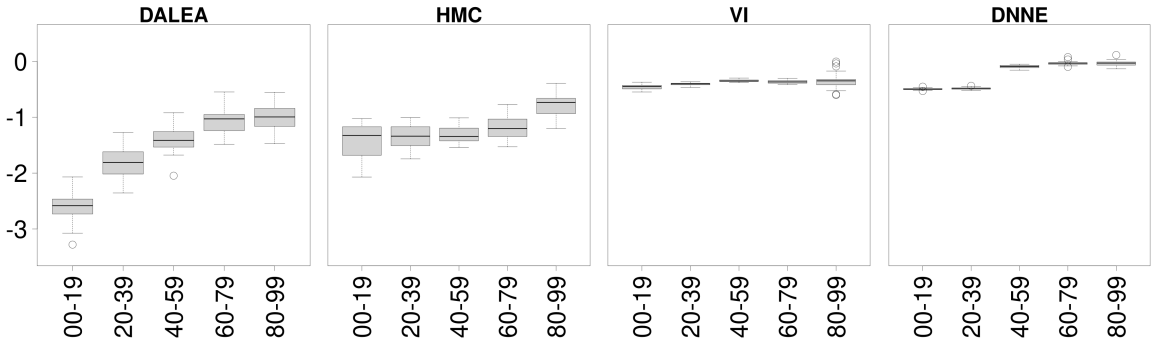
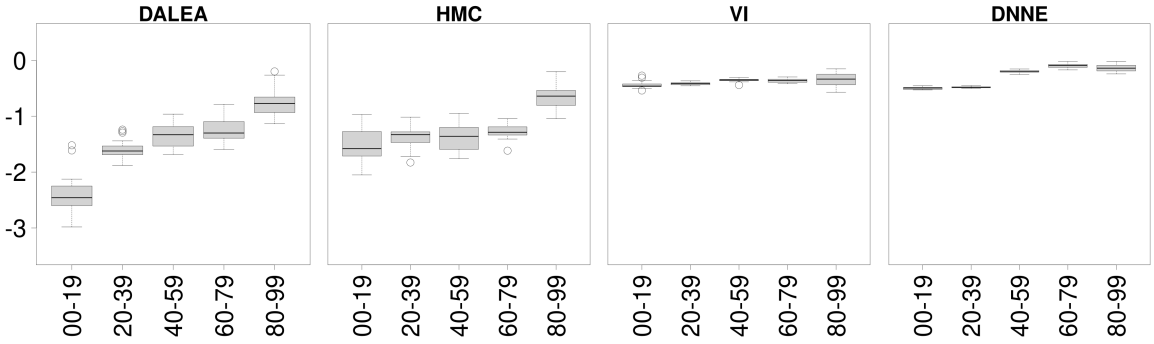


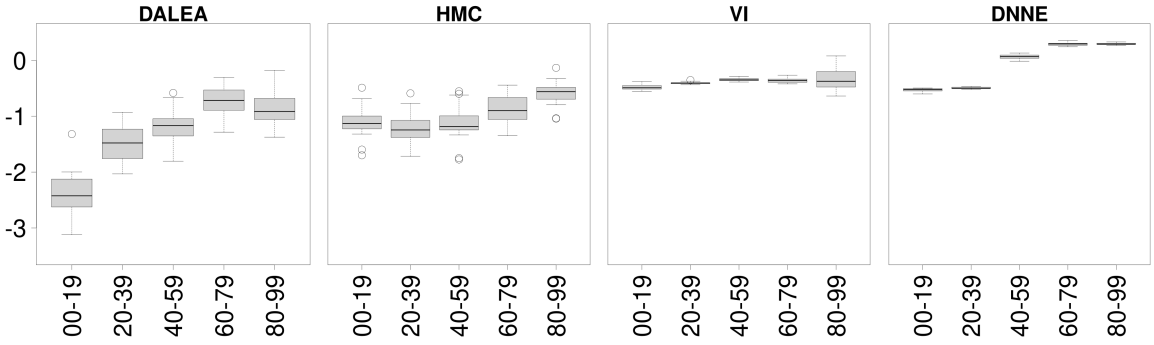
Figure 4.5: Accuracy as a function of confidence. x-axis: Strata of CI width percentile. y-axis: MSE ($\log_{10}$ scale) between posterior mean and true mean function on the testing data.



(a) Gaussian noise

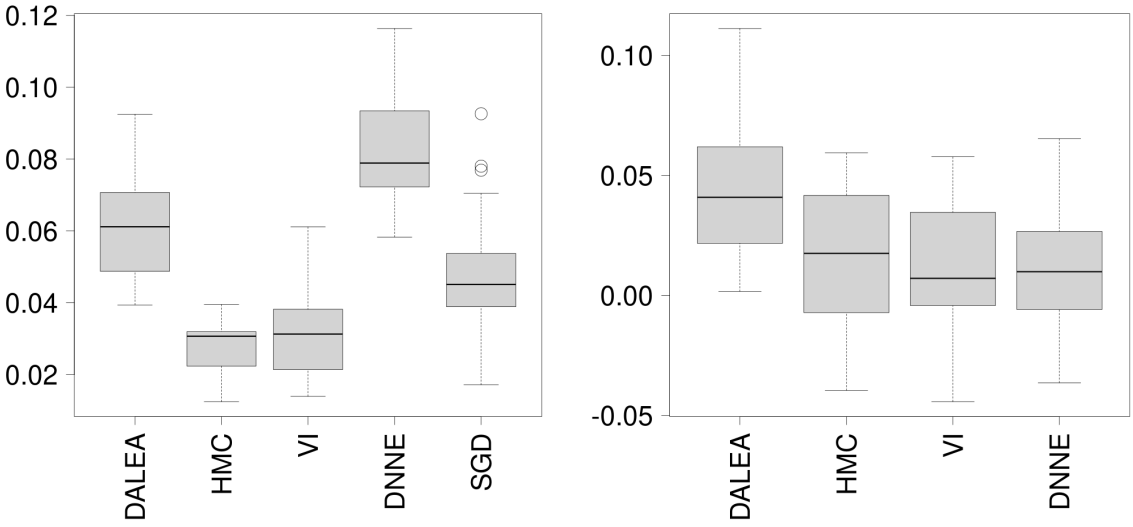


(b) Chi-squared noise



(c) Gaussian mixture noise

Figure 4.6: Analysis results for estimation accuracy and uncertainty quantification in the ABCD data.



(a) Testing prediction r^2.

(b) Testing correlation between CI width and squared prediction error.

CHAPTER V

Conclusion

This work has been motivated by the rapid growth of biomedical data. As the data dimension and complexity increase much faster than the sample size in biomedical studies, traditional statistical methods have limited power in discovering the diverse patterns present in data. On the other hand, blind application of machine learning methods without insight into the data-generation mechanism has limited usefulness in the biomedical setting. To address these challenges, in this dissertation I developed several novel methods for analyzing biomedical data from a variety of application areas. In Chapter II, I proposed two methods for predicting population structures from genotypes. The proposed methods are robust against the shrinkage effect caused by the data dimension growing at a higher rate than the training sample size. From another angle, the proposed methods avoid overfitting the high-dimensional PCA by updating the training set and decomposing the data matrix for every new sample. In Chapter III, I developed a neural network-based method for regressing high-dimensional imaging data on scalar variables. In light of the spatial correlation among the ultrahigh number of voxels, I used neural networks to take the voxels as samples to learn the complex spatial patterns and provide more accurate estimates of the association coefficients. This solution interprets the dimension of the data as another dimension of the sample size, which in effect converts the

curse of dimensionality into a blessing of dimensionality. In Chapter IV, I proposed a novel Bayesian neural network model with the capacity to represent complex noise distributions in order to better quantify uncertainty. The proposed method is able to avoid oversimplifying randomness in model fitting and data generation and provide predictions with confidence even under heterogeneous information density and noise structure. As a whole, this dissertation focuses on developing more effective and robust methods for analyzing biomedical data with growing dimension and complexity but limited sample sizes. The flexibility of sophisticated machine learning algorithms has been combined with the theoretical properties of traditional statistical frameworks to produce methods that are not only accurate but also scientifically interpretable and sample-efficient. It is my hope that the novel methods proposed in this dissertation will provide more useful tool sets for analyzing the next-generation biomedical data.

Several directions have been envisioned for future research. The methods proposed in this work focus on datasets with relatively small sample sizes compared to data dimension and complexity, which is common in typical biomedical studies currently. However, with the rapid development of biotechnology, the sample size is catching up, which is causing an increasing interest among scientists in analyzing datasets that are large in both sample size and data dimension. To address this trend, the proposed methods can utilize randomized algorithms to process high volumes of training samples. For example, the population structure prediction method in Chapter II can be extended by replacing standard SVD with randomized SVD, accommodating larger training sets for ancestry prediction at finer levels. On the other hand, the posterior sampler proposed in Chapter IV can incorporate more advanced techniques to accelerate the sampling speed as the training size grows. For the image-on-scalar regression method in Chapter III, not only can the model be extended to handle large number of images, but the number of covariates and the number of outcomes

(imaging channels) can also grow together with the number of voxels, which would require more efficient training algorithms. By developing statistical learning methods that can handle growth in different dimensions in data, information of higher quality and quantity can be retrieved from big and complex biomedical datasets to better facilitate the understanding of biological processes and assist the discovery of novel therapies.

APPENDICES

APPENDIX A

Supplementary Tables and Figures for Experiments on FRAPOSA

A.1 Supplementary Tables and Figures of FRAPOSA Experiments

Table A.1: The study runtimes, MSDs, and the pairwise mean squared differences between methods, as the reference size varied for the simulated genotypes. The runtimes were the averages of running each setting for 10 times. "MSD" is the mean squared difference between the means of the reference populations and the means of the study populations, scaled by the average distance between the reference population means and the reference global mean. "Pairwise mean squared difference between methods" measures the distance between the PC scores predicted by the two methods. F_{st} is the fixation index of the reference samples, and the proportional eigenvalue is the ratio of the sum of the top 2 eigenvalues to the sum of all the eigenvalues for the reference PCA. The number of variants was 100,000, and the study sample size was 200. Only the top 2 PCs were calculated.

Reference Size	1000	1500	2000	2500	3000
Runtime (sec)					
SP	0.26	0.28	0.25	0.22	0.22
AP	0.27	0.25	0.25	0.21	0.20
OADP	15.12	15.21	15.95	15.99	15.89
ADP	247.93	663.23	1249.75	2119.08	3368.69
MSD (10^{-3})					
Null mean	1.2	1.1	1.1	1.1	1.0
Null SD	0.5	0.5	0.5	0.5	0.4
SP	87	53	32	23	14
AP	7	2	3	3	3
OADP	4	1	2	2	2
ADP	6	2	3	3	3
Pairwise mean squared differences between methods (10^{-3})					
ADP-OADP	0.10	0.03	0.02	0.01	0.01
ADP-AP	0.19	0.12	0.09	0.06	0.05
ADP-SP	39	19	12	8	6
OADP-AP	0.33	0.16	0.11	0.07	0.05
OADP-SP	35	17	11	7	5
AP-SP	40	19	12	8	6
Population diversity statistics (10^{-3})					
F_{st}	4.01	4.29	3.98	4.27	4.06
Proportional eigenvalue	6.0	5.3	4.8	4.7	4.4

Table A.2: Number of European UK Biobank samples predicted by OADP and FastPCA to belong to each ancestry group. FastPCA was applied to the combined samples of the European samples in 1000 Genomes and UK Biobank data. European UK Biobank samples were identified by OADP using global 1000 Genome reference samples. The PC scores of each of the the UK Biobank samples were then used to predict its ancestry membership by using the 20-nearest-neighbor method.

Population	OADP	FastPCA
CEU & GBR	450465	450336
FIN	317	4348
IBS	1816	4175
TSI	9209	2948
Total	461807	461807

Figure A.1: The PC 1 and PC2 scores of the simulated genotypes as predicted by SP, AP, OADP, and ADP when the number of variants was 100,000, and the reference size was 2000. In each of the 4 populations, there were 250 reference samples and 50 study samples, where each sample contained 100,000 variants. The colored/black circle is centered at the reference/study sample mean and encloses 90% of the reference/study samples. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

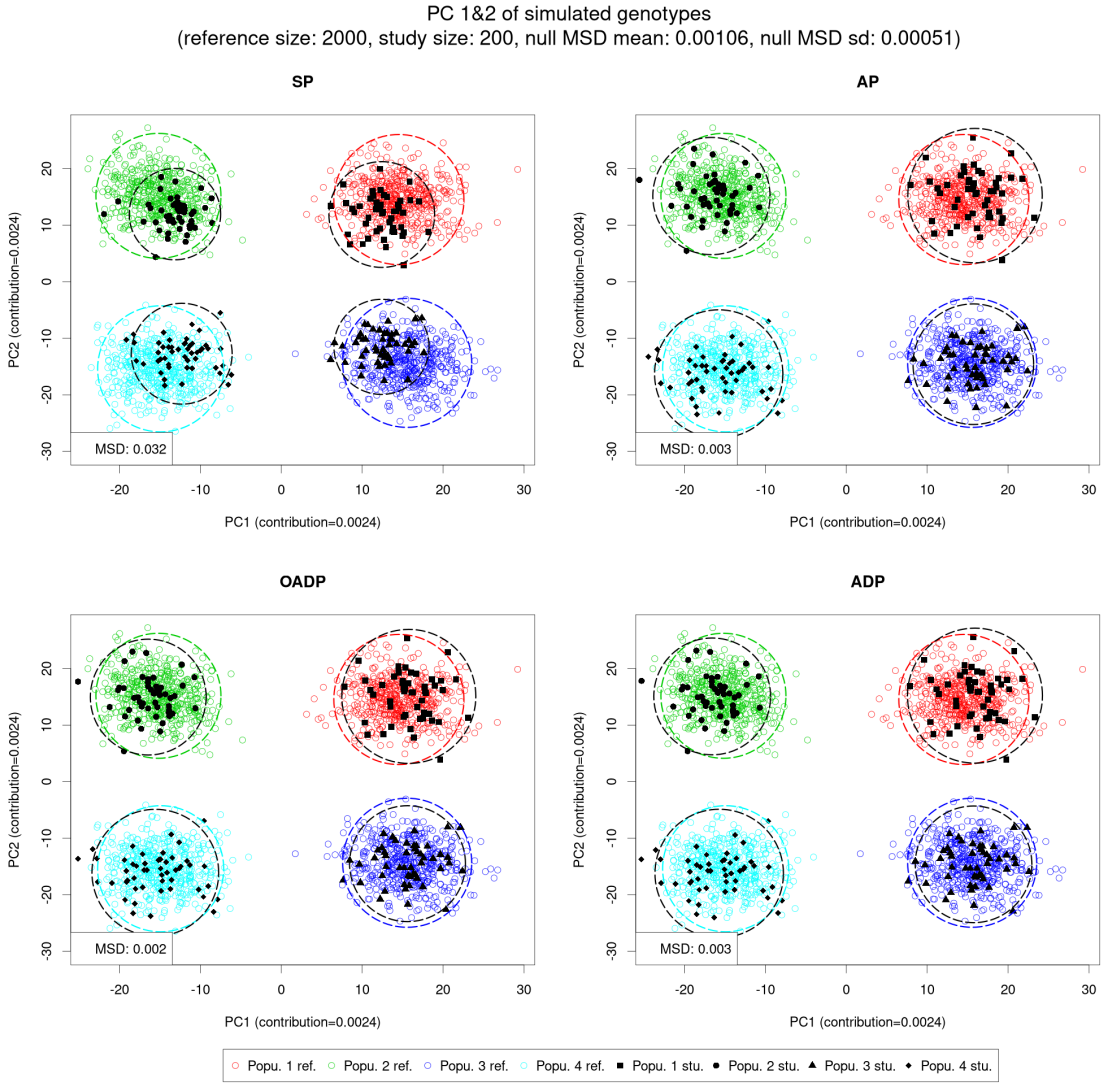


Figure A.2: The PC1 and PC2 scores of the simulated genotypes as predicted by SP, AP, OADP, and ADP when the number of variants was 100,000, and the reference size was 3000. In each of the 4 populations, there were 250 reference samples and 50 study samples, where each sample contained 100,000 variants. The colored/black circle is centered at the reference/study sample mean and encloses 90% of the reference/study samples. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

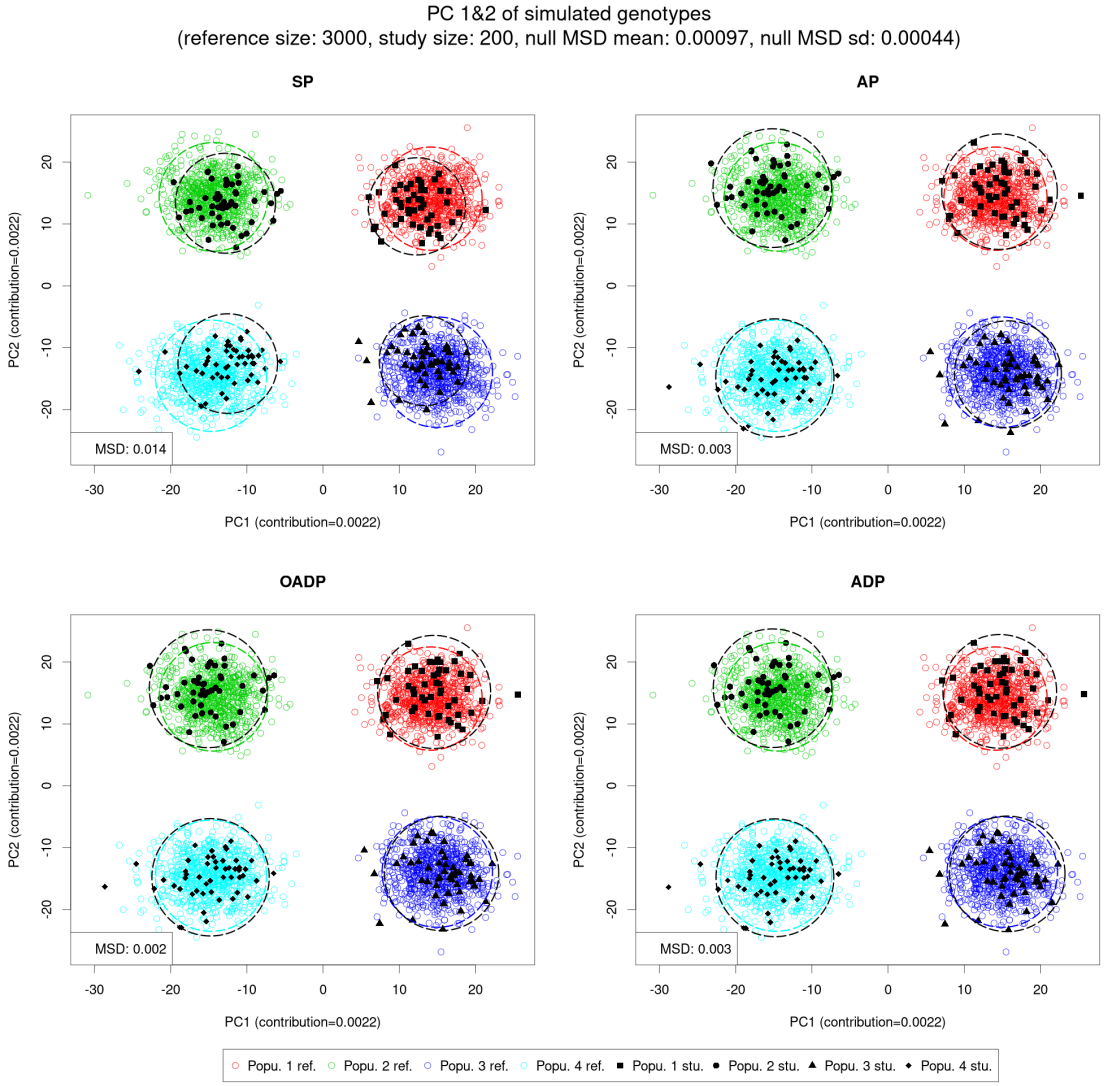


Figure A.3: PC scores of 5000 randomly selected UK Biobank samples, as predicted by SP, AP, OADP, and ADP. The reference panel consisted of all the 2492 samples in the 1000 Genomes data. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

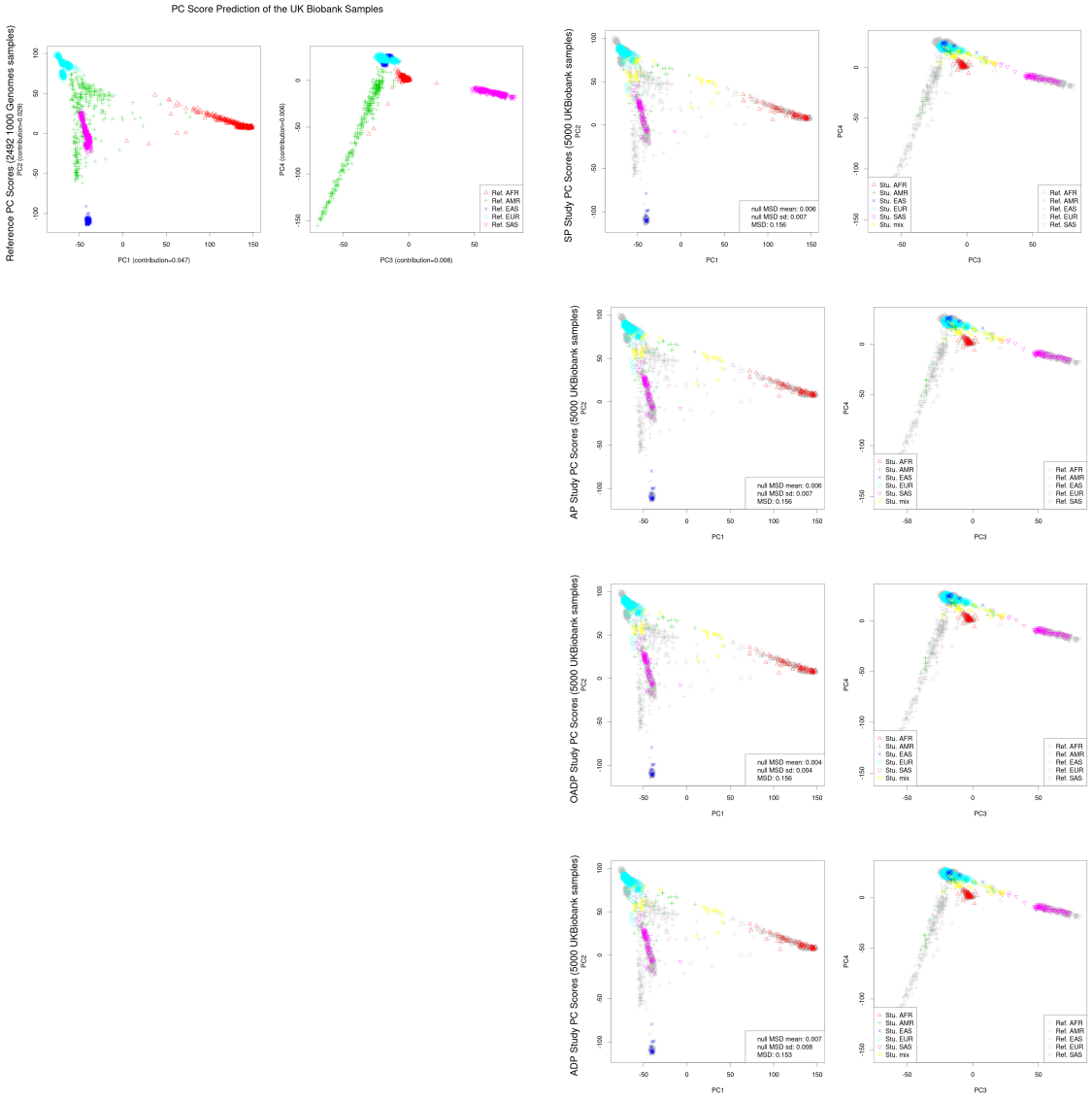


Figure A.4: Pairwise method-to-method comparison of the top 4 PC scores predicted by SP, AP, OADP, and ADP for the 5000 randomly selected UK Biobank study samples. The differences across methods were measured by the mean squared difference between the two methods' PC score prediction. The 2492 samples in 1000 Genomes were used as the reference set. The coloring of the samples represents the populations predicted by OADP. The upper panels show the PC scores, while the lower panels show the pairwise mean squared difference between the methods.

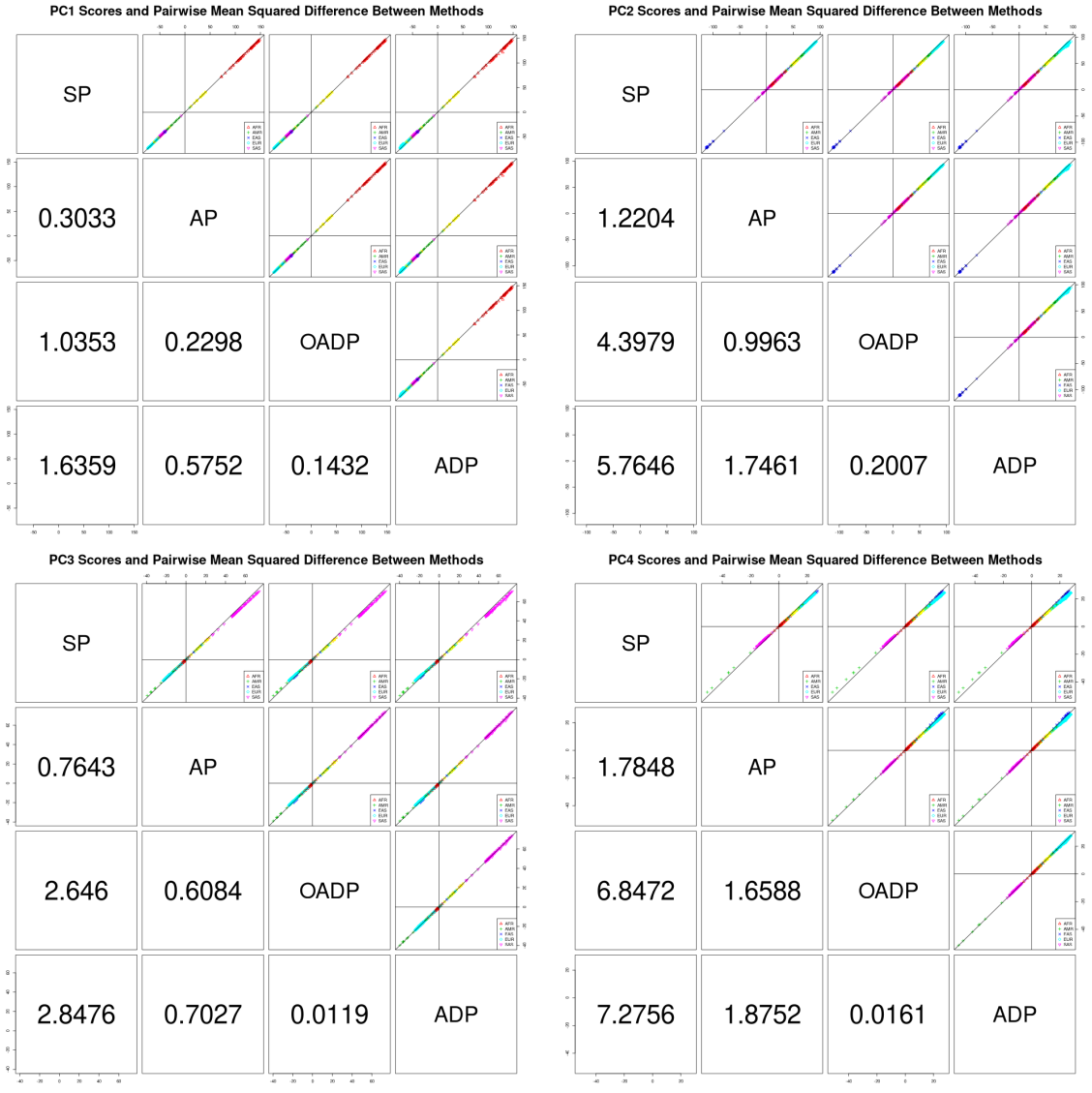


Figure A.5: PC scores of the 5000 randomly selected European UK Biobank samples, as predicted by SP, AP, OADP, and ADP. European samples were identified by OADP using global 1000 Genome reference samples. The reference panel consisted of all the 498 European 1000 Genomes samples. The population membership of each study sample was predicted by the popular votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

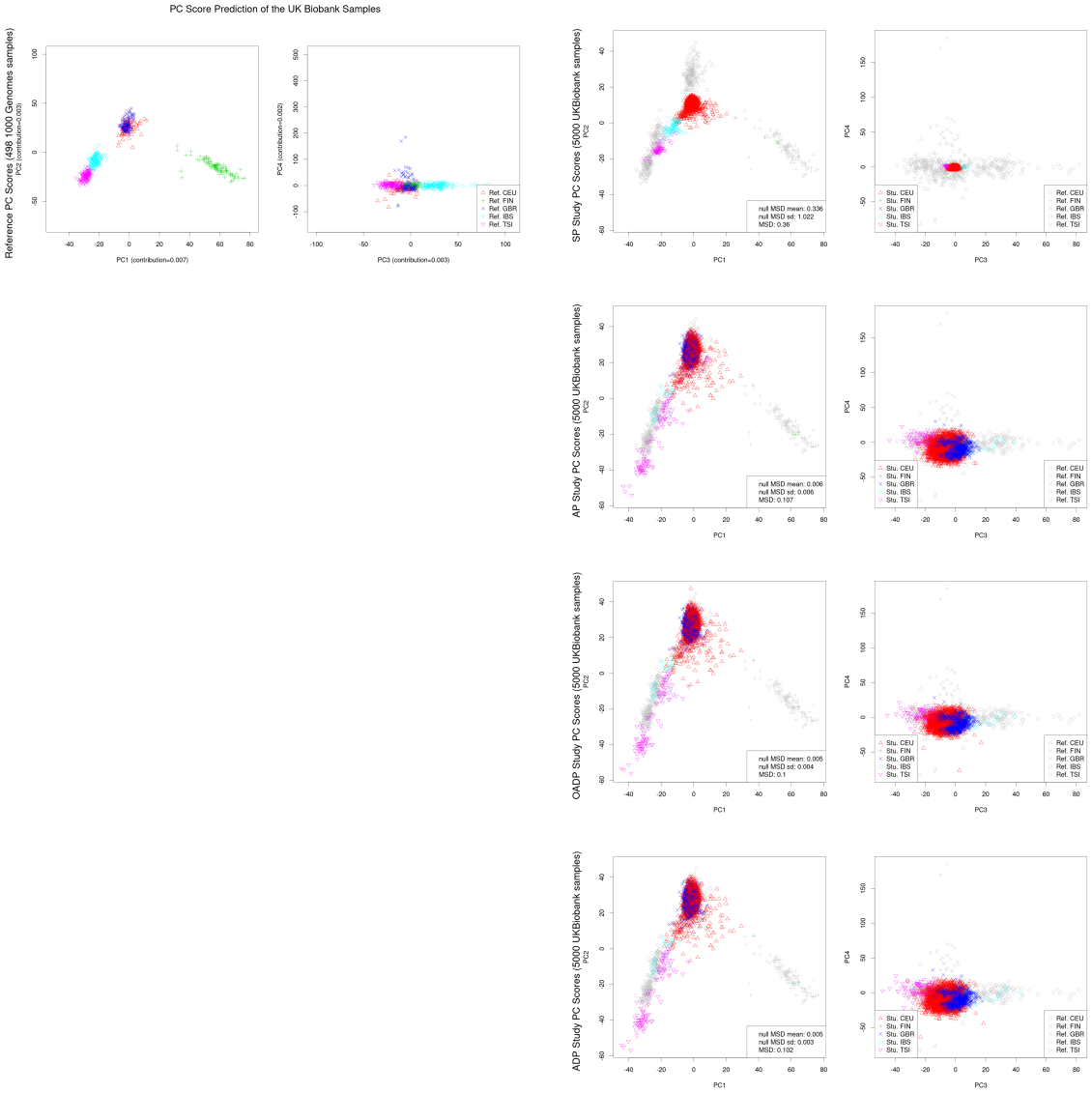


Figure A.6: Pairwise method-to-method comparison of the top 4 PC scores predicted by SP, AP, OADP, and ADP for the 5000 randomly selected European UK Biobank study samples. The differences across methods were measured by the mean squared difference between the two methods' PC score prediction. The 498 European 1000 Genomes samples were used as the reference set. The coloring of the samples represents the populations predicted by OADP. The upper panels show the PC scores, while the lower panels show the pairwise mean squared difference between the methods.

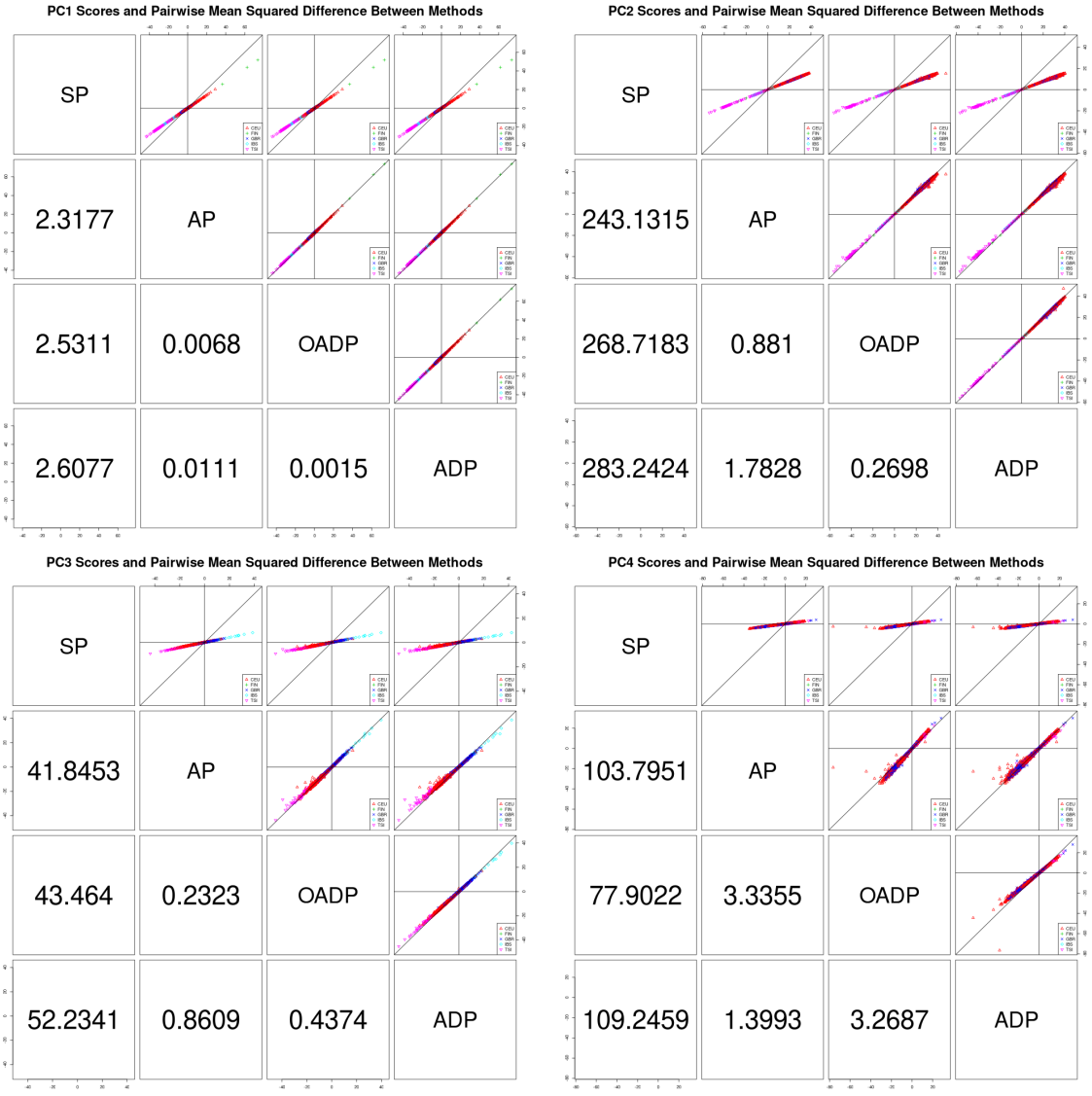


Figure A.7: PC scores of the African UK Biobank samples, as predicted by SP, AP, and OADP. African samples were identified by OADP using global 1000 Genome reference samples. The reference panel consisted of all the 657 African 1000 Genomes samples. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

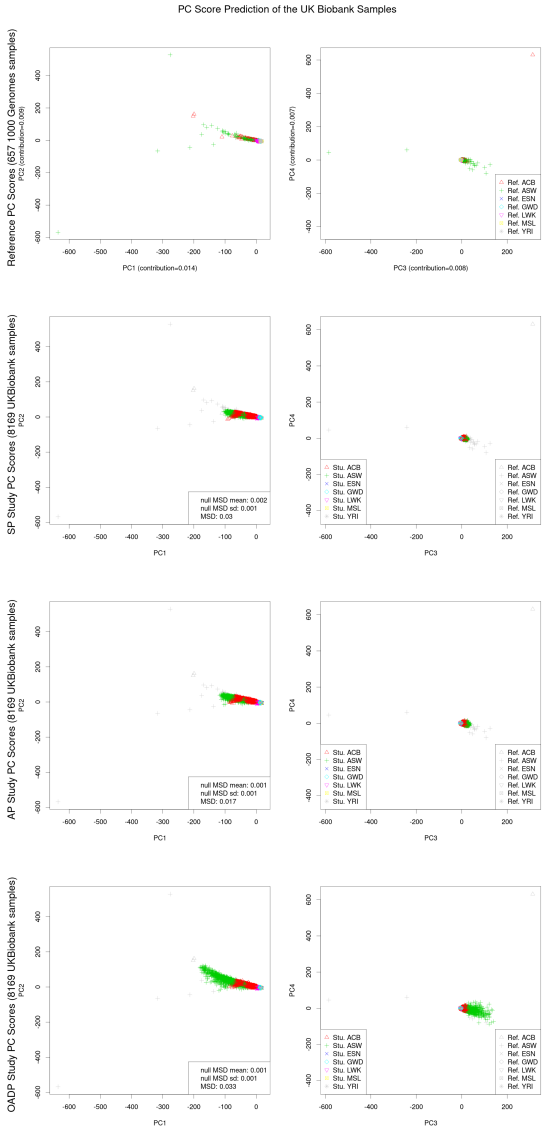


Figure A.8: PC scores of the admixed American UK Biobank samples, as predicted by SP, AP, and OADP. Admixed American samples were identified by OADP using global 1000 Genome reference samples. The reference panel consisted of all the 347 admixed American 1000 Genomes samples. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

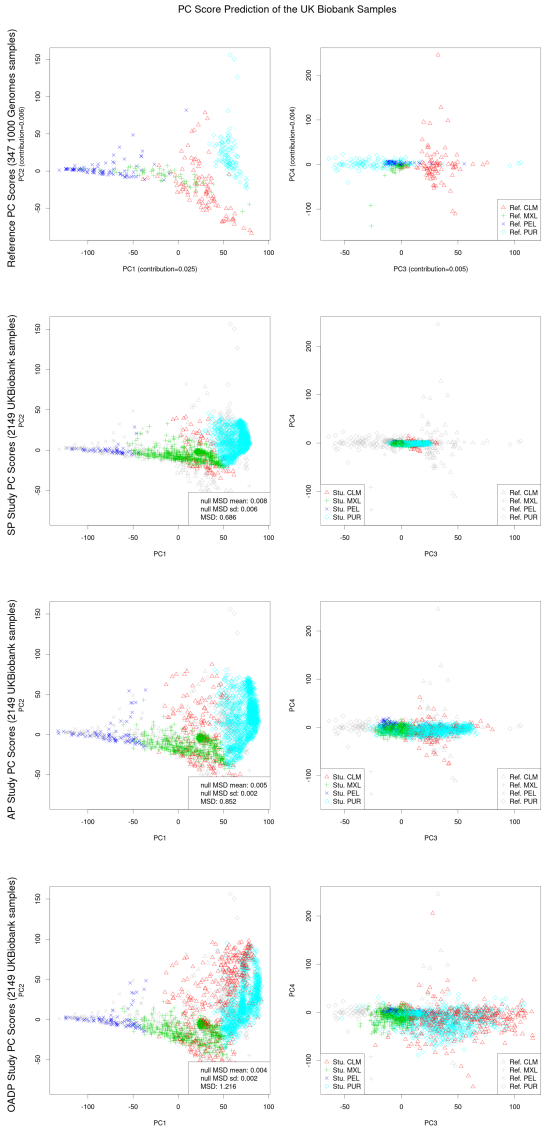


Figure A.9: PC scores of the East Asian UK Biobank samples, as predicted by SP, AP, and OADP. East Asian samples were identified by OADP using global 1000 Genome reference samples. The reference panel consisted of all the 503 East Asian 1000 Genomes samples. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

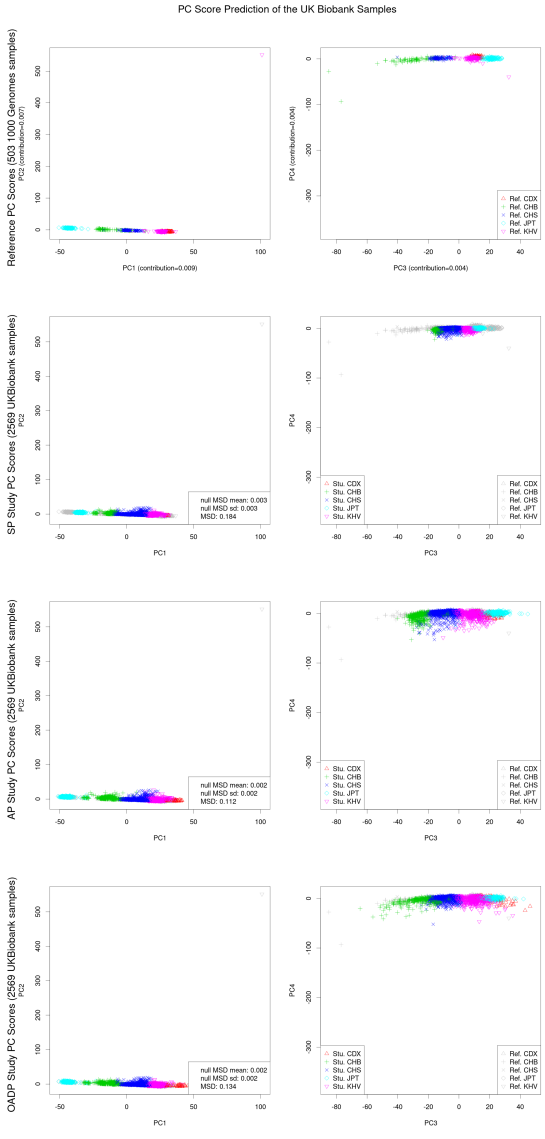


Figure A.10: PC scores of the South Asian UK Biobank samples, as predicted by SP, AP, and OADP. South Asian samples were identified by OADP using global 1000 Genome reference samples. The reference panel consisted of all the 487 South Asian 1000 Genomes samples. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

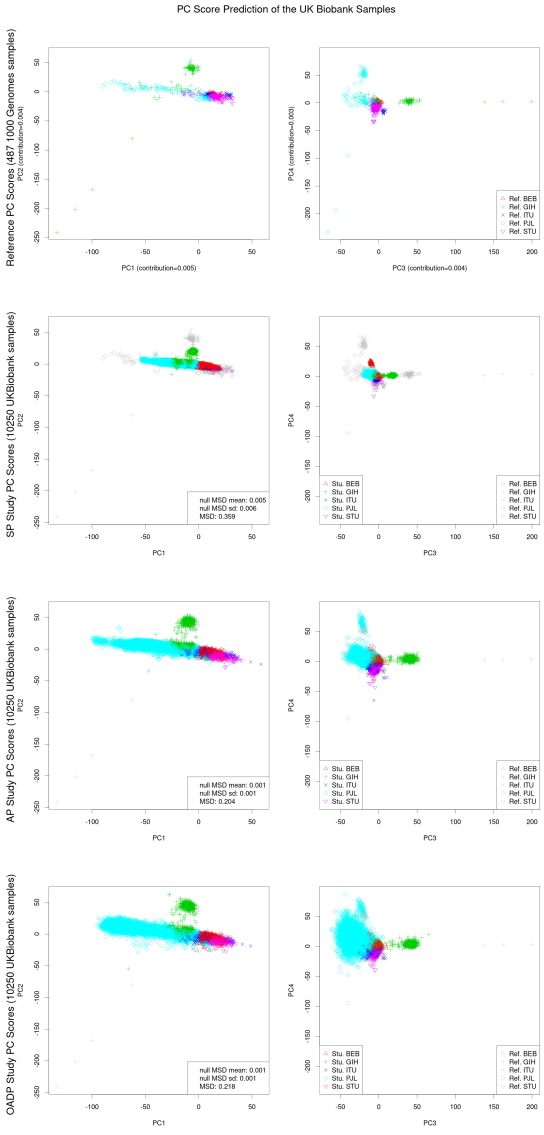


Figure A.11: PC scores of the admixed UK Biobank samples, as predicted by SP, AP, and OADP. Admixed samples were identified by OADP using global 1000 Genome reference samples. The reference panel consisted of all the 2492 samples in the 1000 Genomes data. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. Admixed samples are defined to be those whose highest-voted population received 0.875 or less of the total weighted votes by the 20-nearest-neighbor method. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

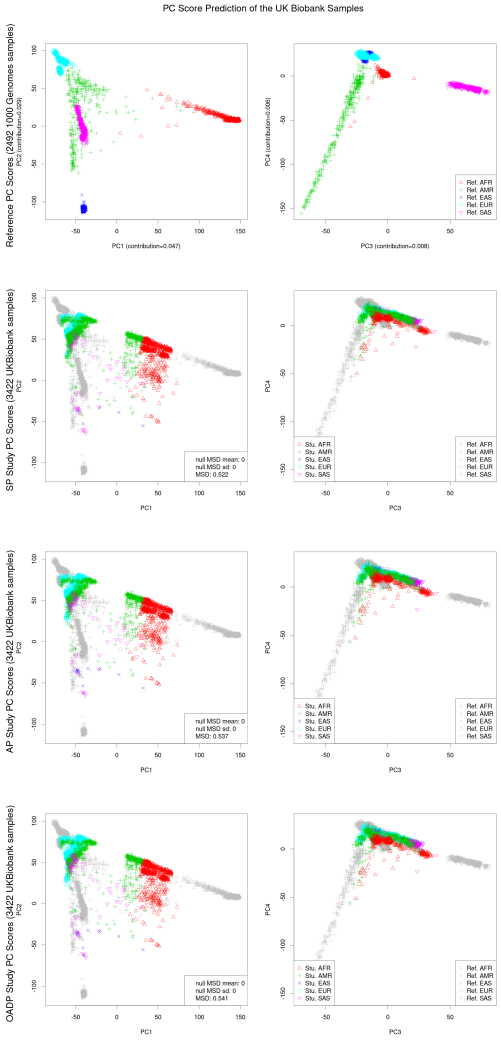


Figure A.12: The PC1 and PC2 scores of the simulated genotypes as predicted by SP, AP, OADP, and ADP when the number of variants was 50,000 and the reference size was 1000. In each of the 4 populations, there were 250 reference samples and 50 study samples. The colored/black circle is centered at the reference/study sample mean and encloses 90% of the reference/study samples. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

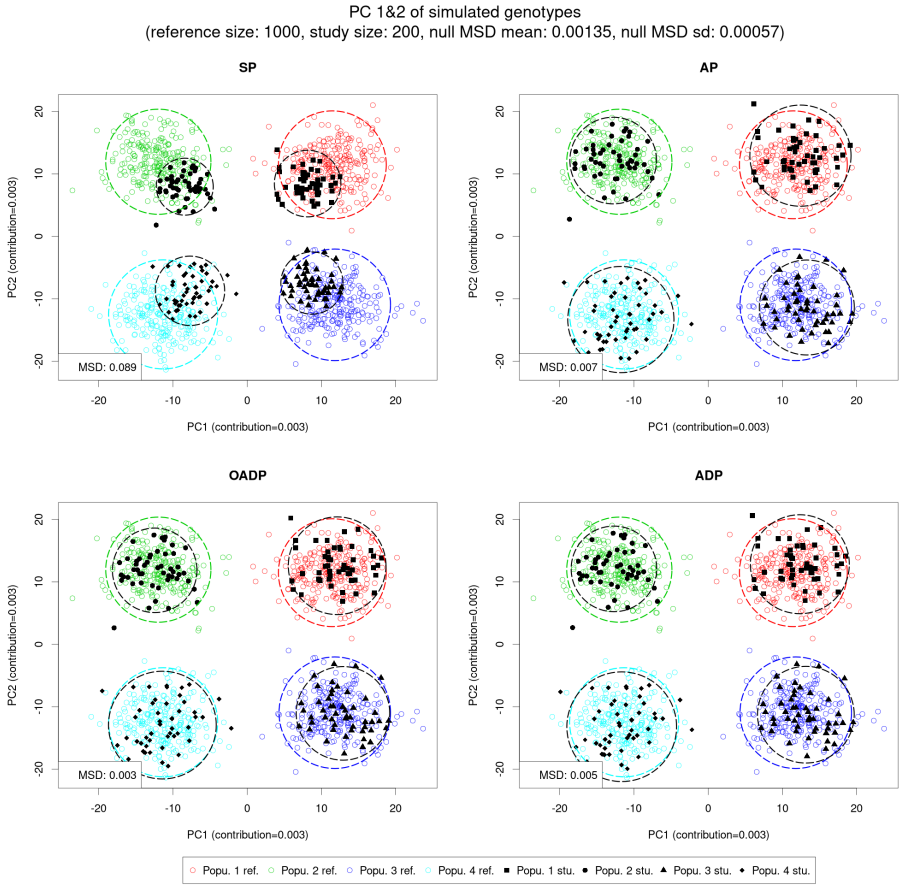


Figure A.13: The PC1 and PC2 scores of the simulated genotypes as predicted by SP, AP, OADP, and ADP when the number of variants was 10,000 and the reference size was 1000. In each of the 4 populations, there were 250 reference samples and 50 study samples. The colored/black circle is centered at the reference/study sample mean and encloses 90% of the reference/study samples. The MSD has been scaled with the average distance between the reference population means and the reference global mean.

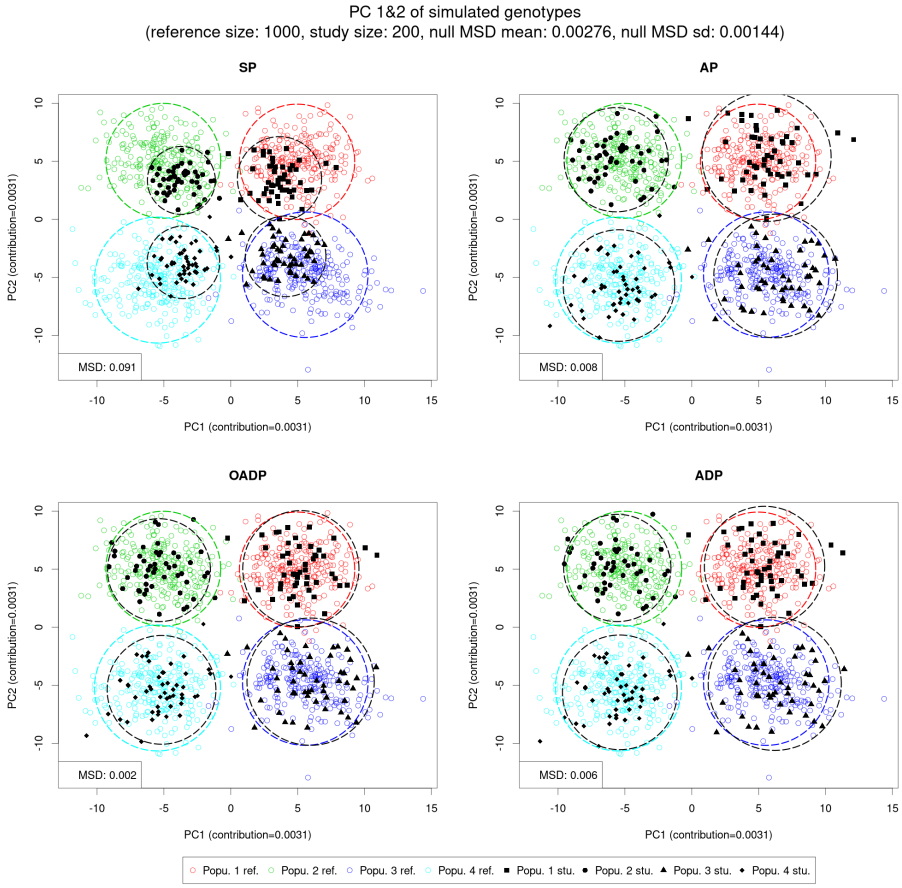


Figure A.14: Comparison of the accuracy of SP, AP, OADP, and ADP when applied to the simulated genotype data. Accuracy was measured by the MSD between the population means of the reference samples and the corresponding population means of the study samples, scaled by the average distance between the reference population means and the reference global mean. Only the top 2 PCs were calculated.

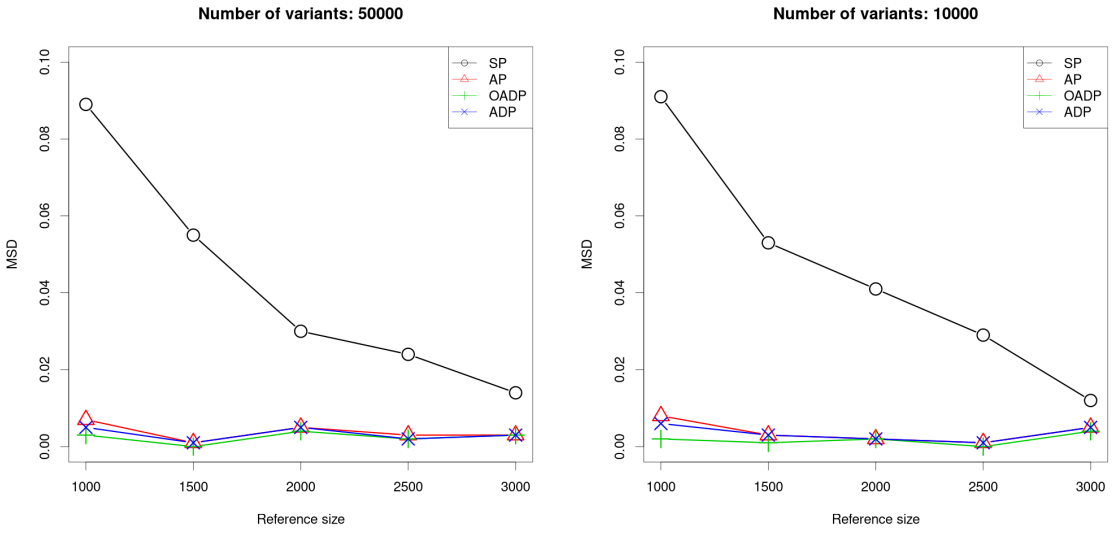


Figure A.15: PC scores of 5000 randomly selected UK Biobank samples, as predicted by SP, AP, OADP, and ADP. The reference panel consisted of 498 randomly selected samples in the 1000 Genomes data, so that the reference size was the same as that in the analysis of the European samples. The population membership of each study sample was predicted by the votes of the 20 nearest reference samples with weights inversely proportional to the distance in between. The MSD has been scaled with the average distance between the reference population means and the reference global mean. The shrinkage factors for the top 4 PCs predicted by AP were 0.96, 0.93, 0.80, and 0.70.

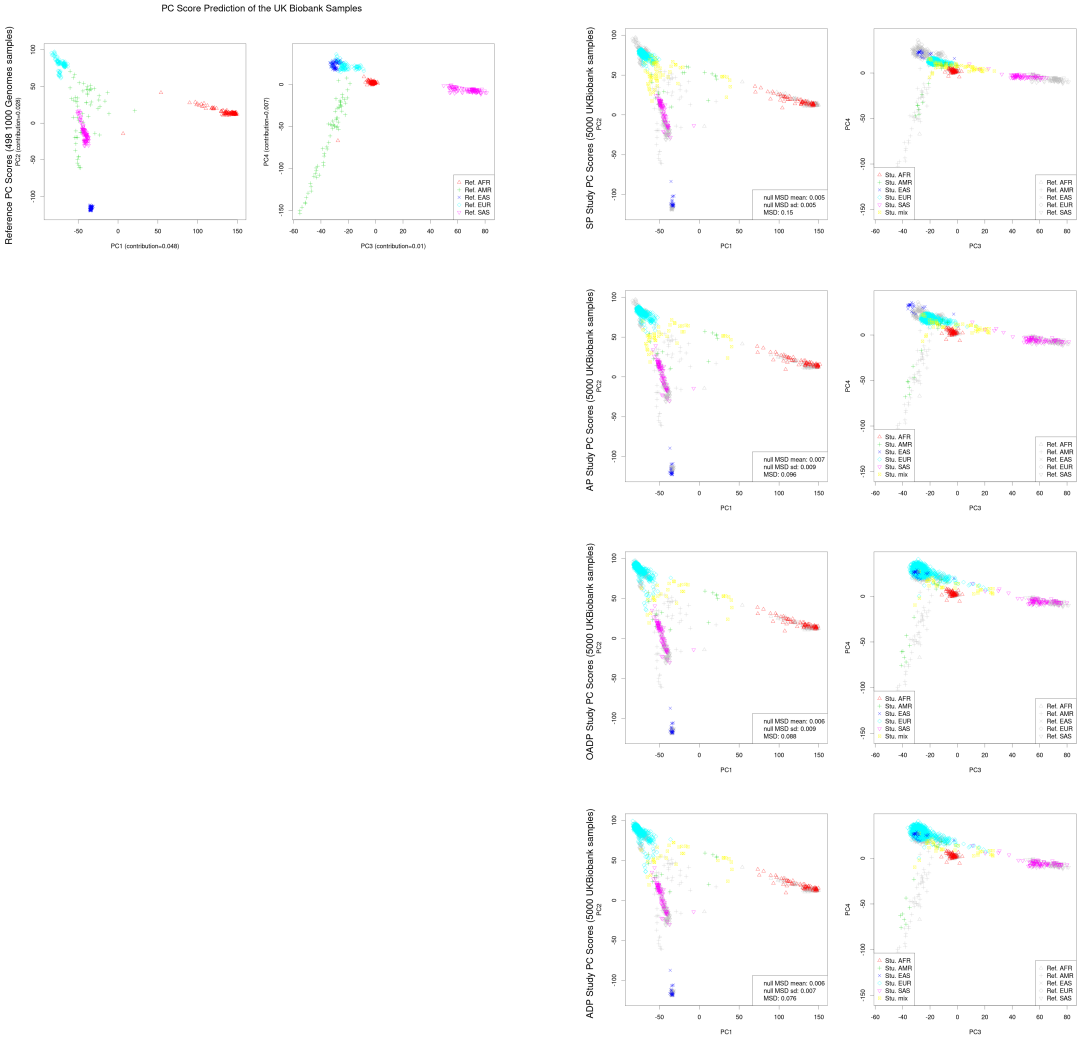


Figure A.16: Pairwise method-to-method comparison of the top 4 PC scores predicted by SP, AP, OADP, and ADP for the 5000 randomly selected UK Biobank study samples. The differences across methods were measured by the mean squared difference between the two methods' PC score prediction. The 2492 samples in the 1000 Genomes data were used as the reference set. The coloring of the samples represents the populations predicted by OADP. The upper panels show the PC scores, while the lower panels show the pairwise mean squared difference between the methods.

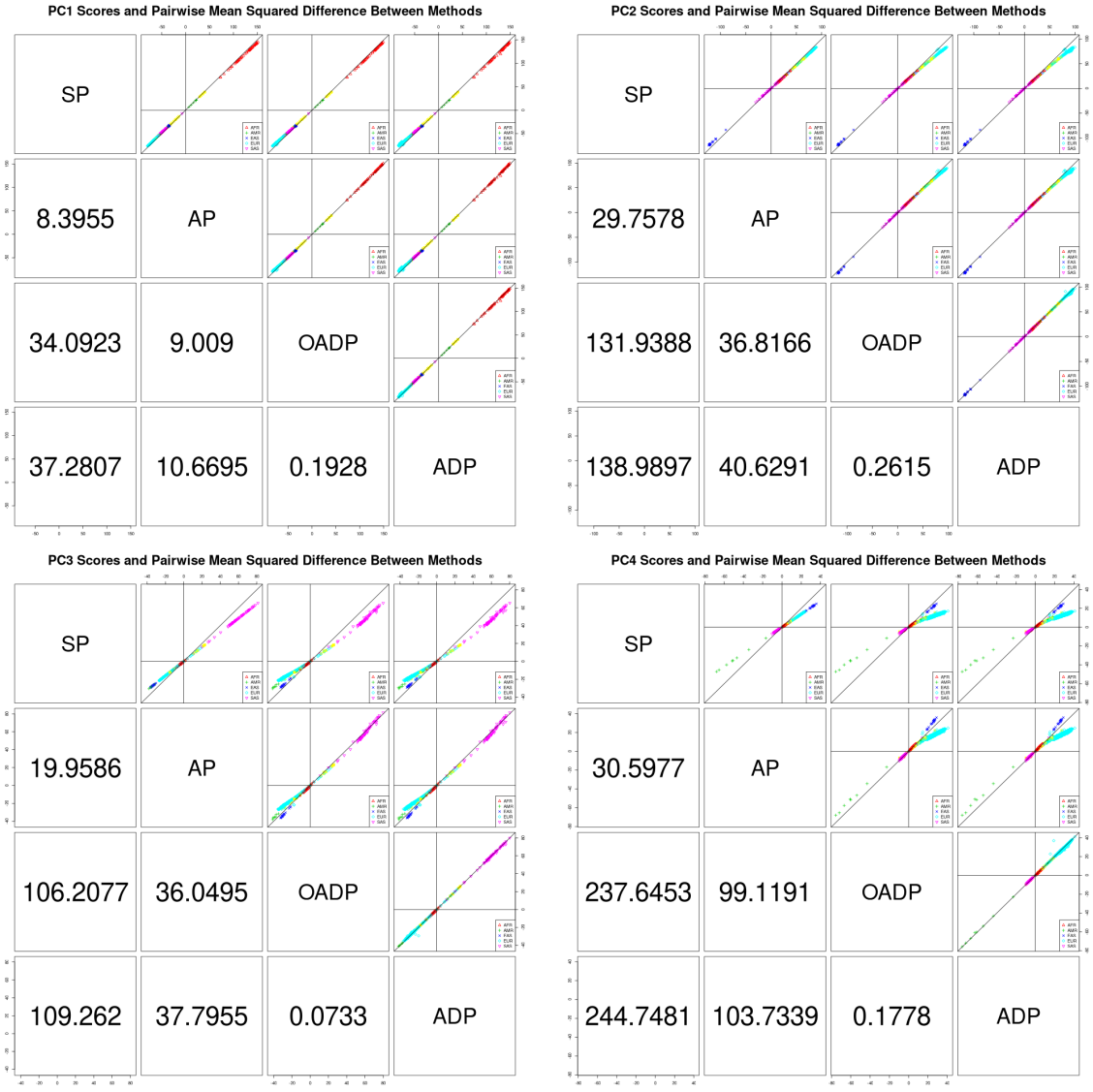
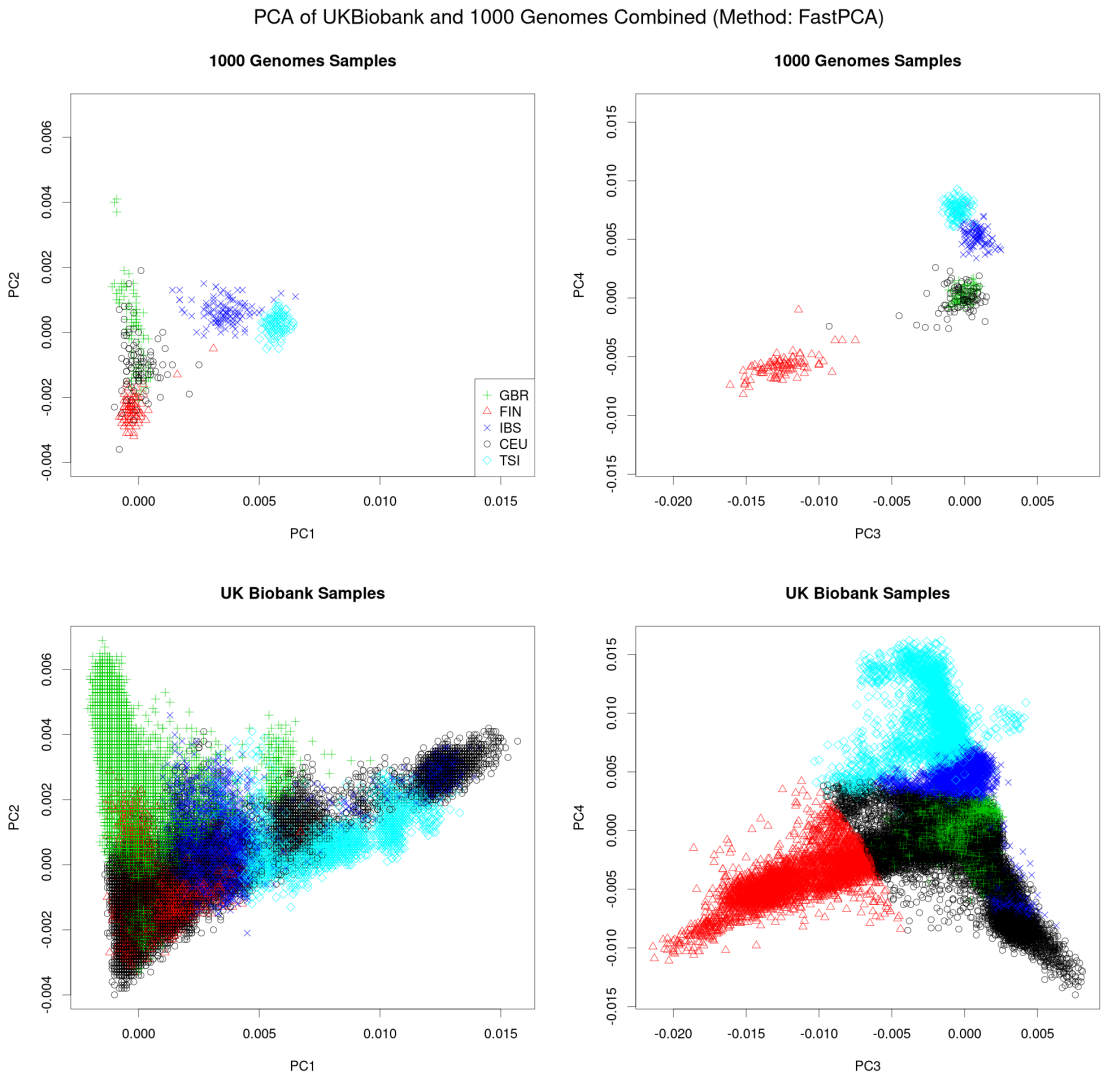


Figure A.17: PCA of the combined data of 498 European 1000 Genomes samples and 461,807 European UK Biobank samples. The total sample size was 462,305. The Europeans in the UK Biobank data were identified by using OADP with all the 2492 samples in the 1000 Genomes data as the reference panel and then applying the 20-nearest-neighbor method. The analysis used the FastPCA algorithm implemented in the Eigensoft software.



APPENDIX B

Proofs for Theoretical Properties of NNISR

B.1 Proofs

For clarity, let $\boldsymbol{\beta}^*$, $\boldsymbol{\alpha}^*$, and σ_*^2 be the true main effects, individual effects, and noise variance, respectively, in this section.

Definition B.1 (Covering number)**.** Let $\mathcal{F}$ be a collection of functions $f : \mathbb{R}^Q \to \mathbb{R}$, Moreover, let $A \subset \mathbb{R}^Q$ and $\delta > 0$. A finite collection $f_1, \ldots, f_L : \mathbb{R}^Q \to \mathbb{R}$ is called a $(\delta, \|\cdot\|_{\infty,A})$-cover of $\mathcal{F}$ if for any $f \in \mathcal{F}$, there exists an $l \in \{1, \ldots, L\}$ such that $\sup_{\boldsymbol{x}\in A} |f(\boldsymbol{x}) - f_l(\boldsymbol{x})| < \delta$. Furthermore, the $(\delta, \|\cdot\|_{\infty,A})$-covering number of $\mathcal{F}$ is the cardinality of the smallest $(\delta, \|\cdot\|_{\infty,A})$-cover of $\mathcal{F}$, denoted as $\mathfrak{N}(\mathcal{F}, \delta, \|\cdot\|_{\infty,A})$

Remark B.2. Definition B.1 is based on Definitions 9.1 and 9.2 in Györfi et al. [2002].

Lemma B.3. *Let $\boldsymbol{U} \in \mathbb{R}^Q$ be a random vector with $\|\operatorname{E}(\boldsymbol{U})\|_2 < \infty$ and $\|\operatorname{Cov}(\boldsymbol{U})\|_F < \infty$. Moreover, assume $\operatorname{Cov}(\boldsymbol{U})$ is positive-definite. Then for any fixed vector $\boldsymbol{a} \in \mathbb{R}^Q$,*

$$[\operatorname{E}(\boldsymbol{U})^\top \boldsymbol{a}]^2 + \operatorname{eigmin}[\operatorname{Cov}(\boldsymbol{U})]\|\boldsymbol{a}\|_2^2 \leq \operatorname{E}[(\boldsymbol{U}^\top \boldsymbol{a})^2] \leq [\operatorname{E}(\boldsymbol{U})^\top \boldsymbol{a}]^2 + \operatorname{eigmax}[\operatorname{Cov}(\boldsymbol{U})]\|\boldsymbol{a}\|_2^2.$$

Proof. We have

$$\mathrm{E}[(\boldsymbol{U}^\top \boldsymbol{a})^2] = \mathrm{E}[\boldsymbol{U}^\top \boldsymbol{a}]^2 + \mathrm{Var}[\boldsymbol{U}^\top \boldsymbol{a}] = [\mathrm{E}(\boldsymbol{U})^\top \boldsymbol{a}]^2 + \boldsymbol{a}^\top \mathrm{Cov}[\boldsymbol{U}]\boldsymbol{a}$$

The eigen-decomposition of $\mathrm{Cov}(\boldsymbol{U})$ implies

$$\mathrm{eigmin}[\mathrm{Cov}(\boldsymbol{U})]\boldsymbol{a}^\top \boldsymbol{a} \leq \boldsymbol{a}^\top \mathrm{Cov}[\boldsymbol{U}]\boldsymbol{a} \leq \mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{U})]\boldsymbol{a}^\top \boldsymbol{a}.$$

The proof is complete. □

Corollary B.4. *For any* $\hat{\boldsymbol{\beta}} : \mathcal{D} \to \mathbb{R}$,

$$\begin{aligned}
&V^{-1} \sum_{v=1}^{V} \int_{\boldsymbol{x}} \left[\boldsymbol{x}^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \mu_{\boldsymbol{X}}(d\boldsymbol{x}) \\
\geq &\, \mathrm{eigmin}[\mathrm{Cov}(\boldsymbol{X})] V^{-1} \sum_{v=1}^{V} \left\| \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \hat{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right\|_2^2 \\
&V^{-1} \sum_{v=1}^{V} \int_{\boldsymbol{x}} \left[\boldsymbol{x}^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{x}^\top \hat{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \mu_{\boldsymbol{X}}(d\boldsymbol{x}) \\
\leq &\, \mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{X})] V^{-1} \sum_{v=1}^{V} \left\| \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \hat{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right\|_2^2 .
\end{aligned}$$

Proof. This result is a consequence of applying Lemma B.3 to $\boldsymbol{X}$ and $\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \hat{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)$, with $E[\boldsymbol{X}] = \boldsymbol{0}$ and $\mathrm{Cov}(\boldsymbol{X})$ being positive-definite by Assumption III.8.

□

Lemma B.5. *Let*

$$\bar{\boldsymbol{\beta}}'_{M,V} = \underset{\boldsymbol{\beta} \in \mathcal{F}_{M,V}}{\arg\min} \frac{1}{MV} \sum_{v=1}^{V} \sum_{m=1}^{M} [Y_m - \boldsymbol{x}_m^\top \boldsymbol{\beta}(\boldsymbol{d}_v)]^2$$

be the least square estimator based on a collection of functions $\mathcal{F}_{M,V}$*. Moreover, for some constant* $c_{23} > 0$*, truncate* $\bar{\boldsymbol{\beta}}'_{M,V}$ *at level* $b_{M,V} = c_{23} \log(MV)$ *to obtain*

$\bar{\boldsymbol{\beta}}_{M,V} = \mathbb{T}_{b_{M,V}} \circ \bar{\boldsymbol{\beta}}'_{M,V}$. *Define*

$$\Delta^2_{M,V} = V^{-1} \sum_{v=1}^{V} \int_{\boldsymbol{x}} \left[\boldsymbol{x}^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{x}^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \mu_{\boldsymbol{X}}(d\boldsymbol{x}).$$

Then there exists a constant $c_{58} > 0$ *such that*

$$\begin{aligned} \mathrm{E}\left[\Delta^2_{M,V}\right] \leq & c_{58} \{1 + \log \mathfrak{N}[\mathcal{F}_{M,V}, (MV)^{-1} b^{-1}_{M,V}, \|\cdot\|_{\infty,\mathcal{D}}]\} \log(MV)^2 (MV)^{-1} \\ & + c_{58} \inf_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} V^{-1} \sum_{v=1}^{V} \|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|_2^2 + c_{58} M^{-1} \end{aligned}$$

Proof. First, for any fixed dataset

$$\Gamma_{M,V} = \{\boldsymbol{x}_m, \boldsymbol{y}_m(\boldsymbol{d}_v) : m \in \{1, \ldots, M\}, v \in \{1, \ldots, V\}\}$$

and the corresponding truncated least square estimator $\bar{\boldsymbol{\beta}}_{M,V}(\cdot)$, the error can be decomposed into

$$\begin{aligned} \Delta^2_{M,V} = & V^{-1} \sum_{v=1}^{V} \mathrm{E}_{\boldsymbol{X}|\Gamma_{M,V}} \left\{ \left[\boldsymbol{X}^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{X}^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \right\} \\ = & T_{1,M,V} + T_{2,M,V} + T_{3,M,V} + T_{4,M,V} \end{aligned}$$

where

$$
\begin{aligned}
T_{1,M,V} =& V^{-1} \sum_{v=1}^{V} \mathrm{E}_{\boldsymbol{X}, Z(\boldsymbol{d}_v)|\Gamma_{M,V}} \left\{ \left[Z(\boldsymbol{d}_v) - \boldsymbol{X}^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \right\} \\
& - V^{-1} \sum_{v=1}^{V} \mathrm{E}_{\boldsymbol{X}, Z(\boldsymbol{d}_v)} \left\{ \left[Z(\boldsymbol{d}_v) - \boldsymbol{X}^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \right\} \\
& - V^{-1} \sum_{v=1}^{V} \mathrm{E}_{\boldsymbol{X}, Z(\boldsymbol{d}_v)|\Gamma_{M,V}} \left\{ \left[\mathbb{T}_{b_{M,V}} \circ Z(\boldsymbol{d}_v) - \boldsymbol{X}^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \right\} \\
& + V^{-1} \sum_{v=1}^{V} \mathrm{E}_{\boldsymbol{X}, Z(\boldsymbol{d}_v)} \left\{ \left[\mathbb{T}_{b_{M,V}} \circ Z(\boldsymbol{d}_v) - \zeta_{b_{M,V}}(\boldsymbol{d}_v) \right]^2 \right\} \\
T_{2,M,V} =& V^{-1} \sum_{v=1}^{V} \mathrm{E}_{\boldsymbol{X}, Z(\boldsymbol{d}_v)|\Gamma_{M,V}} \left\{ \left[\mathbb{T}_{b_{M,V}} \circ Z(\boldsymbol{d}_v) - \boldsymbol{X}^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \right\} \\
& - V^{-1} \sum_{v=1}^{V} \mathrm{E}_{\boldsymbol{X}, Z(\boldsymbol{d}_v)} \left\{ \left[\mathbb{T}_{b_{M,V}} \circ Z(\boldsymbol{d}_v) - \zeta_{b_{M,V}}(\boldsymbol{d}_v) \right]^2 \right\} \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[\mathbb{T}_{b_{M,V}} \circ Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \\
& + V^{-1} \sum_{v=1}^{V} \left[\mathbb{T}_{b_{M,V}} \circ Z_m(\boldsymbol{d}_v) - \zeta_{b_{M,V}}(\boldsymbol{d}_v) \right]^2 \\
T_{3,M,V} =& 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[\mathbb{T}_{b_{M,V}} \circ Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[\mathbb{T}_{b_{M,V}} \circ Z_m(\boldsymbol{d}_v) - \zeta_{b_{M,V}}(\boldsymbol{d}_v) \right]^2 \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 - \left[Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \\
T_{4,M,V} =& 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 - \left[Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2,
\end{aligned}
$$

where $\mathbb{T}_{b_{M,V}} \circ Z(\cdot)$ is the truncated version of $Z(\cdot)$ at level $b_{M,V}$, and $\zeta_{b_{M,V}}(\cdot) = \mathrm{E}[\mathbb{T}_{b_{M,V}} \circ Z(\cdot)]$. We bound each term separately. For $T_{1,M,V}$, $T_{2,M,V}$, and $T_{3,M,V}$, the proof of their bounds is identical to in the proof of Lemma 1 in Bauer et al. [2019],

which gives

$$
\begin{aligned}
T_{1,M,V} \leq & c_{47} \log(MV)(MV)^{-1} \\
\mathrm{E}[T_{2,M,V}] \leq & c_{48} \log(MV)^2 \log\{\mathfrak{N}[\mathcal{F}_{M,V}, (MV)^{-1} b_{M,V}^{-1}], \|\cdot\|_{\infty,\mathcal{D}}\}(MV)^{-1} \\
\mathrm{E}[T_{3,M,V}] \leq & c_{49} \log(MV)(MV)^{-1}.
\end{aligned}
$$

For $T_{4,M,V}$, we further decompose it into $T_{4,M,V} = T_{41,M,V} + T_{42,M,V}$, where

$$
\begin{aligned}
T_{41,M,V} =& 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2 - \left[Z_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2 \\
& -2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2 - \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2 \\
T_{42,M,V} =& 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2 - \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2
\end{aligned}
$$

We can bound $T_{41,M,V}$ as

$$\begin{aligned}
\frac{1}{2}\,\mathrm{E}[T_{41,M,V}] =& (MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left[Z_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\\
&-(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left[Z_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2\\
&-(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left[Z_m(\boldsymbol{d}_v)+\alpha_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\\
&+(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left[Z_m(\boldsymbol{d}_v)+\alpha_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2\\
=& (MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\alpha_m(\boldsymbol{d}_v)\boldsymbol{X}_m^\top[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)]\right\}\\
=& V^{-1}\sum_{v=1}^{V}\mathrm{E}\left\{\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]\left[M^{-1}\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)\boldsymbol{X}_m^\top\right]\right\}\\
\le& \sqrt{\mathrm{E}\left\{V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right\|_2^2\right\}}\\
&\cdot\sqrt{\mathrm{E}\left\{V^{-1}\sum_{v=1}^{V}\left\|M^{-1}\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)\boldsymbol{X}_m\right\|_2^2\right\}}
\end{aligned}$$

Notice that the first term is bounded by $\sqrt{\mathrm{eigmin}[\mathrm{Cov}(\boldsymbol{X})]^{-1}\,\mathrm{E}[\Delta_{M,V}^2]}$ by Corol-

lary B.4. For the second term, for any $v \in \{1, \ldots, V\}$,

$$
\begin{aligned}
&\mathrm{E}\left\{\left\|M^{-1}\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)\boldsymbol{X}_m\right\|_2^2\right\} \\
=&\mathrm{E}\left\{\sum_{q=1}^{Q}\left[M^{-1}\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)X_{m,q}\right]^2\right\} \\
=&\sum_{q=1}^{Q}\left\{\mathrm{E}\left[M^{-1}\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)X_{m,q}\right]\right\}^2 + \mathrm{Var}\left\{M^{-1}\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)X_{m,q}\right\} \\
=&\sum_{q=1}^{Q}\left\{M^{-1}\left[\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)\right]\mathrm{E}[X_q]\right\}^2 + M^{-2}\left[\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)^2\right]\mathrm{Var}\left\{X_q\right\} \\
=&0 + \left[\sum_{m=1}^{M}\alpha_m(\boldsymbol{d}_v)^2\right]M^{-2}\sum_{q=1}^{Q}\mathrm{Var}\left\{X_q\right\} \\
\leq&\left[\sum_{m=1}^{M}c_{55}^2\right]M^{-2}\|\operatorname{Cov}(\boldsymbol{X})\|_F^2 = c_{55}^2c_{59}M^{-1}
\end{aligned}
$$

by Assumption III.8. Then $T_{41,M,V}$ is bounded by

$$
\mathrm{E}[T_{41,M,V}] \leq 2c_{55}^2c_{59}M^{-\frac{1}{2}}\sqrt{\operatorname{eigmin}[\operatorname{Cov}(\boldsymbol{X})]^{-1}\,\mathrm{E}[\Delta_{M,V}^2]}.
$$

For $T_{42,M,V}$, define $A_{M,V}$ to be the event that

$$
\max\{|Y_m(\boldsymbol{d}_v)| : 1 \leq m \leq M, 1 \leq v \leq V\} > b_{M,V}.
$$

Then $T_{42,M,V} = T_{7,M,V} + T_{8,M,V}$, where

$$
\begin{aligned}
T_{7,M,V} =&2(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\mathbb{I}[A_{M,V}] \\
T_{8,M,V} =&2(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\mathbb{I}[A_{M,V}^c] \\
&-2(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2.
\end{aligned}
$$

For $T_{7,M,V}$, we have

$$\begin{aligned}
&\frac{1}{2}\,\mathrm{E}[T_{7,M,V}] \\
=&\,\mathrm{E}\left\{\mathbb{I}[A_{M,V}](MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\right\} \\
\leq&\sqrt{\mathrm{E}\left\{\mathbb{I}[A_{M,V}]\right\}}\sqrt{(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^4\right\}} \\
\leq&\sqrt{\Pr[A_{M,V}]}\sqrt{(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{8Y_m(\boldsymbol{d}_v)^4+8[\boldsymbol{X}^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)]^4\right\}} \\
\leq&\sqrt{\Pr[A_{M,V}]}\sqrt{(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{8Z_m(\boldsymbol{d}_v)^4+8\alpha_m(\boldsymbol{d}_v)^4+8[\boldsymbol{X}^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)]^4\right\}}
\end{aligned}$$

Notice that because of $x < \exp(x)$ for all $x \in \mathbb{R}$, we have

$$\begin{aligned}
\mathrm{E}[Z_m(\boldsymbol{d})^4] =&\,\mathrm{E}\{[Z_m(\boldsymbol{d})^2]^2\} \leq \mathrm{E}\left\{\left[\frac{2}{c_{52}}\exp\left(\frac{c_{52}}{2}Z_m(\boldsymbol{d})^2\right)\right]^2\right\} \\
=&\frac{4}{c_{52}^2}\,\mathrm{E}[\exp(c_{52}Z_m(\boldsymbol{d})^2)] \leq \frac{4c_{53}}{c_{52}^2}
\end{aligned}$$

by Assumption III.9. Moreover, $|\alpha_m(\boldsymbol{d}_v)| < c_{55}$ by Assumption III.7, and

$$\begin{aligned}
\mathrm{E}\left\{\left|\boldsymbol{X}^\top\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right|^4\right\} \leq&\,\mathrm{E}\left\{\left[\|\boldsymbol{X}\|\|\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|\right]^4\right\} = \|\bar{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|^4\,\mathrm{E}\left\{\|\boldsymbol{X}\|^4\right\} \\
\leq&c_{56}b_{M,V}^4 = c_{56}c_{23}^4\log(MV)^4
\end{aligned}$$

by Assumption III.8. Thus the first term is bounded by $c_{57}\log(MV)^2$ for some constant $c_{57} > 0$. For the second term, by using the inequality

$$\mathbb{I}[|U| > b_{M,V}] \leq \frac{\exp(c_{52}U^2)}{\exp(c_{52}b_{M,V}^2)},$$

we have

$$\begin{aligned}\sqrt{\Pr[A_{M,V}]} \leq & \sqrt{\sum_{v=1}^{V}\sum_{m=1}^{M}\Pr[|Y_m(\boldsymbol{d}_v)| > b_{M,V}]} \\ \leq & \sqrt{\sum_{v=1}^{V}\sum_{m=1}^{M}\Pr[|Z_m(\boldsymbol{d}_v)| + |\alpha_m(\boldsymbol{d}_v)| > b_{M,V}]} \\ \leq & \sqrt{\sum_{v=1}^{V}\sum_{m=1}^{M}\Pr[|Z_m(\boldsymbol{d}_v)| > b_{M,V} - c_{55}]} \\ \leq & \sqrt{\frac{\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}[\exp[c_{52}Z_m(\boldsymbol{d}_v)^2]]}{\exp[c_{52}(b_{M,V} - c_{55})^2]}} \leq c_{43}\frac{\sqrt{MV}}{\exp[c_{44}\log(MV)^2]}.\end{aligned}$$

Since for any $c > 0$,

$$\exp(-c\log(n)^2) = n^{-c\log(n)} < n^{-2}$$

for $n > \exp(2/c)$, we get

$$\frac{1}{2}\mathrm{E}[T_{7,M,V}] \leq c_{45}\frac{\log(MV)^2\sqrt{MV}}{M^2V^2} \leq c_{46}\frac{1}{MV}$$

for MV sufficiently large. For $T_{8,M,V}$, because $|\mathbb{T}_b(z) - y| < |z - y|$ for all $|y| < b$, we

have

$$
\begin{aligned}
T_{8,M,V} \leq & 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}'_{M,V}(\boldsymbol{d}_v) \right]^2 \mathbb{I}[A_{M,V}^c] \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \\
\leq & 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \bar{\boldsymbol{\beta}}'_{M,V}(\boldsymbol{d}_v) \right]^2 - \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \\
= & 2(MV)^{-1} \inf_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}(\boldsymbol{d}_v) \right]^2 \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \\
\leq & 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 - \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2
\end{aligned}
$$

for any $\ddot{\boldsymbol{\beta}}_{M,V} \in \mathcal{F}_{M,V}$. Choose $\ddot{\boldsymbol{\beta}}_{M,V}$ such that

$$
\begin{aligned}
& V^{-1} \sum_{v=1}^{V} \left\| \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right\|^2 \\
\leq & \inf_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} V^{-1} \sum_{v=1}^{V} \left\| \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v) \right\|^2 + (MV)^{-1}.
\end{aligned}
$$

Then

$$
\begin{aligned}
&\frac{1}{2}\,\mathrm{E}[T_{8,M,V}] \\
\leq&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\right\} \\
-&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2\right\} \\
=&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2\right\} \\
+&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\right\} \\
+&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{2\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]\left[\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]\right\} \\
-&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]^2\right\} \\
=&T_{81,M,V}+T_{82,M,V},
\end{aligned}
$$

where

$$
\begin{aligned}
T_{81,M,V}=&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\right\} \\
T_{82,M,V}=&(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{2\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]\left[\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]\right\}.
\end{aligned}
$$

For $T_{81,M,V}$, we have

$$\begin{aligned}
T_{81,M,V} =& V^{-1}\sum_{v=1}^{V}\mathrm{E}\left\{\left[\boldsymbol{X}^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{X}^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\right\}\\
=& V^{-1}\sum_{v=1}^{V}\int_{\boldsymbol{x}}\left[\boldsymbol{x}^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{x}^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\mu_{\boldsymbol{X}}(d\boldsymbol{x})\\
\leq& \mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{X})]V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right\|^2\\
\leq& \mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{X})]\inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}}V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\right\|^2\\
&+\mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{X})](MV)^{-1}
\end{aligned}$$

by Lemma B.3. For $T_{82,M,V}$, we have

$$\begin{aligned}
&T_{82,M,V}\\
=&2(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[Y_m(\boldsymbol{d}_v)-\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]\boldsymbol{X}_m\right\}^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]\\
=&2(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[\epsilon_m(\boldsymbol{d}_v)+\alpha_m(\boldsymbol{d}_v)\right]\boldsymbol{X}_m\right\}^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]\\
=&2(MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\epsilon_m(\boldsymbol{d}_v)+\alpha_m(\boldsymbol{d}_v)\right\}\mathrm{E}\left\{\boldsymbol{X}_m\right\}^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]
\end{aligned}$$

by the independence between ϵ_m and $\boldsymbol{X}_m$. Then $T_{82,M,V}=0$ because $E[\boldsymbol{X}_m]=\boldsymbol{0}$. Therefore, $\frac{1}{2}\mathrm{E}[T_{8,M,V}]\leq T_{81,M,V}$, which leads to

$$\begin{aligned}
\mathrm{E}[T_{4,M,V}]\leq& c_{50}\left\{\inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}}V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\right\|^2\right\}\\
&+c_{50}\left\{+(MV)^{-1}+M^{-\frac{1}{2}}\sqrt{\mathrm{eigmin}[\mathrm{Cov}(\boldsymbol{X})]^{-1}\,\mathrm{E}[\Delta_{M,V}^2]}\right\}.
\end{aligned}$$

Altogether, we have

$$\begin{aligned}\mathrm{E}[\Delta_{M,V}^2] \leq c_{50}\Bigg\{ & \inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}} V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|^2 \\ & +\log(MV)^2(1+\log\{\mathfrak{N}[\mathcal{F}_{M,V},(MV)^{-1}b_{M,V}^{-1}],\|\cdot\|_{\infty,\mathcal{D}}\})(MV)^{-1} \\ & +M^{-\frac{1}{2}}\sqrt{\mathrm{eigmin}[\mathrm{Cov}(\boldsymbol{X})]^{-1}\,\mathrm{E}[\Delta_{M,V}^2]}\Bigg\},\end{aligned}$$

which implies

$$\begin{aligned} & \left(\sqrt{\mathrm{E}[\Delta_{M,V}^2]}-\frac{1}{2}c_{50}M^{-\frac{1}{2}}\right)^2 \\ =& \,\mathrm{E}[\Delta_{M,V}^2]-c_{50}M^{-\frac{1}{2}}\sqrt{\mathrm{E}[\Delta_{M,V}^2]}+\frac{1}{4}c_{50}^2M^{-1} \\ \leq & c_{50}\inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}} V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|^2 \\ & +c_{50}\log(MV)^2(1+\log\{\mathfrak{N}[\mathcal{F}_{M,V},(MV)^{-1}b_{M,V}^{-1}],\|\cdot\|_{\infty,\mathcal{D}}\})(MV)^{-1} \\ & +\frac{1}{4}c_{50}^2M^{-1}.\end{aligned}$$

Thus

$$\begin{aligned} & \sqrt{\mathrm{E}[\Delta_{M,V}^2]} \\ \leq & c_{20}\Bigg\{\inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}} V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|^2 \\ & +\log(MV)^2(1+\log\{\mathfrak{N}[\mathcal{F}_{M,V},(MV)^{-1}b_{M,V}^{-1}],\|\cdot\|_{\infty,\mathcal{D}}\})(MV)^{-1}+M^{-1}\Bigg\}^{\frac{1}{2}} \\ & +c_{20}\sqrt{M^{-1}} \\ \leq & c_{28}\Bigg\{\inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}} V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|^2 \\ & +\log(MV)^2(1+\log\{\mathfrak{N}[\mathcal{F}_{M,V},(MV)^{-1}b_{M,V}^{-1}],\|\cdot\|_{\infty,\mathcal{D}}\})(MV)^{-1}+M^{-1}\Bigg\}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{E}[\Delta_{M,V}^2] \leq & c_{58} \inf_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} V^{-1} \sum_{v=1}^{V} \|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|^2 \\ & + c_{58} \log(MV)^2 (1 + \log\{\mathfrak{N}[\mathcal{F}_{M,V}, (MV)^{-1} b_{M,V}^{-1}], \|\cdot\|_{\infty,\mathcal{D}}\})(MV)^{-1} \\ & + c_{58} M^{-1}.\end{aligned}$$

□

Lemma B.6. *For $\sigma^2 : \mathcal{D} \to \mathbb{R}_+$, let*

$$\tilde{\boldsymbol{\beta}}'_{M,V} = \operatorname*{arg\,min}_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} \frac{1}{MV} \sum_{v=1}^{V} \sum_{m=1}^{M} [Y_m - \boldsymbol{x}_m^\top \boldsymbol{\beta}(\boldsymbol{d}_v)]^2 \sigma^{-2}(\boldsymbol{d}_v).$$

be the weighted least square estimator based on a collection of functions $\mathcal{F}_{M,V}$. Moreover, for some constant $c_{23} > 0$, truncate $\tilde{\boldsymbol{\beta}}'_{M,V}$ at level $b_{M,V} = c_{23} \log(MV)$ to obtain $\tilde{\boldsymbol{\beta}}_{M,V} = \mathbb{T}_{b_{M,V}} \circ \tilde{\boldsymbol{\beta}}'_{M,V}$. Then there exists a constant $c_{59} > 0$ such that

$$\begin{aligned}& V^{-1} \sum_{v=1}^{V} \int_{\boldsymbol{x}} \left[\boldsymbol{x}^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2 \mu_{\boldsymbol{X}}(d\boldsymbol{x}) \\ \leq & c_{59} \{1 + \log \mathfrak{N}[\mathcal{F}_{M,V}, (MV)^{-1} b_{M,V}^{-1}, \|\cdot\|_{\infty,\mathcal{D}}]\} \log(MV)^2 (MV)^{-1} \\ & + c_{59} \inf_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} V^{-1} \sum_{v=1}^{V} \|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|_2^2 + c_{59} M^{-1}\end{aligned}$$

Proof. This result is obtained by applying the same argument for Lemma B.5, which does not require assumptions on the distribution of $\boldsymbol{d}$ (i.e. regardless of whether the L_2 errors at the voxels $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_V$ are weighted uniformly or weighted by $\sigma^{-2}(\boldsymbol{d}_1), \ldots, \sigma^{-2}(\boldsymbol{d}_V)$). □

Lemma B.7. *For $\sigma^2 : \mathcal{D} \to \mathbb{R}_+$ and $\lambda_{M,V} > 0$ with $\lambda_{M,V} \leq c_{67}(MV)^{-1}$ for some*

constant $c_{67} > 0$*, let*

$$\hat{\boldsymbol{\beta}}'_{M,V} = \underset{\boldsymbol{\beta}\in\mathcal{F}_{M,V}}{\arg\min} \frac{1}{MV}\sum_{v=1}^{V}\sum_{m=1}^{M}[Y_m - \boldsymbol{x}_m^\top\boldsymbol{\beta}(\boldsymbol{d}_v)]^2\sigma^{-2}(\boldsymbol{d}_v) + \lambda_{M,V}\frac{1}{V}\sum_{v=1}^{V}\|\boldsymbol{\beta}(\boldsymbol{d}_v)\|_1$$

be the L_1*-penalized weighted least square estimator based on a collection of functions* $\mathcal{F}_{M,V}$*, which satisfies* $\mathcal{F}_{M_1,V_1} \subset \mathcal{F}_{M_2,V_2}$ *for all* $M_1 \leq M_2$ *and* $V_1 \leq V_2$*. Moreover, for some constant* $c_{23} > 0$*, truncate* $\hat{\boldsymbol{\beta}}'_{M,V}$ *at level* $b_{M,V} = c_{23}\log(MV)$ *to obtain* $\hat{\boldsymbol{\beta}}_{M,V} = \mathbb{T}_{b_{M,V}} \circ \hat{\boldsymbol{\beta}}'_{M,V}$*. There exists a constant* $c_{59} > 0$ *such that*

$$\begin{aligned}
&V^{-1}\sum_{v=1}^{V}\int_{\boldsymbol{x}}\left[\boldsymbol{x}^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{x}^\top\hat{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\mu_{\boldsymbol{X}}(d\boldsymbol{x}) \\
\leq& c_{59}\{1 + \log\mathfrak{N}[\mathcal{F}_{M,V}, (MV)^{-1}b_{M,V}^{-1}, \|\cdot\|_{\infty,\mathcal{D}}]\}\log(MV)^2(MV)^{-1} \\
+& c_{59}\inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}} V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|_2^2 + c_{59}M^{-1}
\end{aligned}$$

Proof. The proof is almost identical to those of Lemma B.5 and Lemma B.6. We only

need to change the arguments for $T_{8,M,V}$ and after:

$$
\begin{aligned}
T_{8,M,V} \leq & 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \hat{\boldsymbol{\beta}}'_{M,V}(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v) \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v) \\
\leq & 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \hat{\boldsymbol{\beta}}'_{M,V}(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v) \\
& + 2\lambda_{M,V} V^{-1} \sum_{v=1}^{V} \|\boldsymbol{\beta}(\boldsymbol{d}_v)\|_1 \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v) \\
= & \inf_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v) \\
& \qquad + 2\lambda_{M,V} V^{-1} \sum_{v=1}^{V} \|\boldsymbol{\beta}(\boldsymbol{d}_v)\|_1 \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v) \\
\leq & 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v) \\
& + 2\lambda_{M,V} V^{-1} \sum_{v=1}^{V} \|\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|_1 \\
& - 2(MV)^{-1} \sum_{v=1}^{V} \sum_{m=1}^{M} \left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top \boldsymbol{\beta}^*(\boldsymbol{d}_v) \right]^2 \sigma^{-2}(\boldsymbol{d}_v)
\end{aligned}
$$

for any $\ddot{\boldsymbol{\beta}}_{M,V} \in \mathcal{F}_{M,V}$. Choose $\ddot{\boldsymbol{\beta}}_{M,V}$ such that

$$
\begin{aligned}
& V^{-1} \sum_{v=1}^{V} \left\| \boldsymbol{\beta}^*(\boldsymbol{d}_v) - \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v) \right\|^2 \\
\leq & \inf_{\boldsymbol{\beta} \in \mathcal{F}_{M,V}} V^{-1} \sum_{v=1}^{V} \|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\|^2 + (MV)^{-1}.
\end{aligned}
$$

Then by the same argument as in Lemma B.5 and Lemma B.6

$$\frac{1}{2}\,\mathrm{E}[T_{8,M,V}] \leq T_{81,M,V} + T_{82,M,V} + T_{83,M,V},$$

where

$$\begin{aligned}
T_{81,M,V} =& (MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{\left[\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\right\}\sigma^{-2}(\boldsymbol{d}_v)\\
=& V^{-1}\sum_{v=1}^{V}\mathrm{E}\left\{\left[\boldsymbol{X}^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{X}^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]^2\right\}\sigma^{-2}(\boldsymbol{d}_v)\\
\leq& \mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{X})]V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right\|^2\\
\leq& \mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{X})]\inf_{\boldsymbol{\beta}\in\mathcal{F}_{M,V}} V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{\beta}_{M,V}(\boldsymbol{d}_v)\right\|^2\\
&+ \mathrm{eigmax}[\mathrm{Cov}(\boldsymbol{X})](MV)^{-1}\\
T_{82,M,V} =& (MV)^{-1}\sum_{v=1}^{V}\sum_{m=1}^{M}\mathrm{E}\left\{2\left[Y_m(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v)\right]\left[\boldsymbol{X}_m^\top\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \boldsymbol{X}_m^\top\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\right]\right\}\\
=& 0\\
T_{83,M,V} =& \lambda_{M,V}V^{-1}\sum_{v=1}^{V}\|\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|_1 \leq \lambda_{M,V}Q^{\frac{1}{2}}V^{-1}\sum_{v=1}^{V}\|\ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|_2\\
\leq& \lambda_{M,V}Q^{\frac{1}{2}}V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v)\|_2 + \lambda_{M,V}Q^{\frac{1}{2}}V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|_2\\
\leq& \lambda_{M,V}Q^{\frac{1}{2}}V^{-1}\sum_{v=1}^{V}Q^{\frac{1}{2}}c_{51} + \lambda_{M,V}Q^{\frac{1}{2}}\sqrt{V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|_2^2}\\
\leq& c_{51}\lambda_{M,V}Q + \lambda_{M,V}Q^{\frac{1}{2}}c_{66} \leq c_{68}(MV)^{-1}
\end{aligned}$$

for some constant c_{66} such that $V^{-1}\sum_{v=1}^{V}\|\boldsymbol{\beta}^*(\boldsymbol{d}_v) - \ddot{\boldsymbol{\beta}}_{M,V}(\boldsymbol{d}_v)\|_2^2 \leq c_{66}^2$ for M and V sufficiently large, and $c_{68} = (c_{51}Q + c_{66}Q^{\frac{1}{2}})c_{67}$. Then

$$\frac{1}{2}\,\mathrm{E}[T_{8,M,V}] \leq T_{81,M,V} + T_{83,M,V} \leq c_{69}T_{81,M,V}$$

for some constant $c_{69} > 0$. The rest of the proof is the same as in Lemma B.5 and Lemma B.6. □

Lemma B.8. *Let $N \in \mathbb{N}_+$. Suppose random vector $(\boldsymbol{X}_N, Y_N) \in \mathbb{R}^K \times \mathbb{R}$ satisfies $\mathrm{E}[Y_N^2] \leq \infty$ and has conditional mean function $f(\boldsymbol{x}) = \mathrm{E}[Y_N | \boldsymbol{X}_N = \boldsymbol{x}]$ for all N. Moreover, suppose f is a P-smooth (K', l)-GHIM with $P > 1$, and define $P' = \lfloor P-1 \rfloor$. Let $S_N, T \in \mathbb{N}_+$ with $T \geq P'$, and let $a_N \in \mathbb{R}_+$ satisfy $1 \leq a_N \leq S_N$ and $a_N^{T+P'+3} \leq S_N^P$ for N sufficiently large. Set $R = \binom{K'+T}{K'}(T+1)(S_N+1)^{K'}$. For arbitrary $c_{29} \in \mathbb{R}_+$ and $\kappa_N \in (0, 1]$, let $\gamma = \log(N) S_N^{K'+P(2T+3)+1} \kappa_N^{-1}$. There exist $\hat{f}_N \in \mathcal{G}_{l,R,K',K,\gamma}$, $\chi_n \subset \mathbb{R}^K$, and $c_{30}, c_{36} \in \mathbb{R}_+$ such that for N sufficiently large,*

$$|f(\boldsymbol{x}) - \hat{f}_N(\boldsymbol{x})| \leq c_{30} a_N^{T+P'+3} S_N^{-P} \qquad \text{for all } \boldsymbol{x} \in [-a_N, a_N]^K \setminus \chi_N,$$
$$|\hat{f}_N(\boldsymbol{x})| \leq c_{36} a_N^{P'} S_N^{K'+TP} \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^K$$

where $\Pr_{\boldsymbol{X}_N}(\chi_N) < c_{29} \kappa_N$.

Proof. Lemma B.8 is based on Theorem 3 in Bauer et al. [2019]. In addition to simplifying the conditions to suit the purpose of our theoretical analysis, we generalizes the result by extending the constant probability measure $\Pr_{\boldsymbol{X}}$ to the probability measure $\Pr_{\boldsymbol{X}_N}$ that depends on N, which make the result applicable to the fixed-design model in NNISR, since the empirical mean of the samples in a fixed-design model can be interpreted as the theoretical mean of a discrete distribution with the support having a cardinality equal to the sample size.

To grant this generalized result, the proof in Bauer et al. [2019] only needs minimal adjustment. In the original proof, $[-a_N, a_N]^K$ is divided into S_N^K many K-dimensional hypercubes, and it is shown that for each hypercube, a neural network exists such that inside the hypercube (except near the boundary), the neural network approximates the Taylor polynomial of f arbitrarily well, while outside the hypercube (except near the boundary), the neural network is arbitrarily close to zero. As for all the transi-

tion zones near the boundaries, the union of all of them can be made to have $\Pr_{\boldsymbol{X}}$ measure less than $c_{29}\kappa_N$ by appropriately shifting the grid that divides $[-a_N, a_N]^K$ into hypercubes. However, this argument does not require the probability measure to be independent of N. In fact, since a different neural network is fitted for each N and the mean conditional function f does not depend on N, even if the probability measure $\Pr_{\boldsymbol{X}_N}$ is allowed to vary with respect to N, the same grid-shifting argument is still valid, and thus the bound on the exception set χ_N remains unchanged. See Supplement A in Bauer et al. [2019] for the original proof. □

Lemma B.9. *Let $N \in \mathbb{N}_+$. Suppose random vector $(\boldsymbol{X}_N, Y_N) \in \mathbb{R}^K \times \mathbb{R}$ satisfies $\mathrm{E}[Y_N^2] \leq \infty$ and has conditional mean function $f(\boldsymbol{x}) = \mathrm{E}[Y_N|\boldsymbol{X}_N = \boldsymbol{x}]$ for all N. Moreover, suppose f is a J-piecewise P-smooth (K', l)-GHIM with $P > 1$, and define $P' = \lfloor P - 1 \rfloor$. Let $S_N, T \in \mathbb{N}_+$ with $T \geq P'$, and let $a_N \in \mathbb{R}_+$ satisfy $1 \leq a_N \leq S_N$ and $a_N^{T+P'+3} \leq S_N^P$ for N sufficiently large. Set $R = \binom{K'+T}{K'}(T+1)(S_N+1)^{K'}$. For arbitrary $c_{29} \in \mathbb{R}_+$ and $\kappa_N \in (0, 1]$, let $\gamma = \log(N) S_N^{K'+P(2T+3)+1} \kappa_N^{-1}$. There exist $\hat{f}_N \in \mathcal{G}_{l,R,K'+J,K,\gamma}$, $\chi_n \subset \mathbb{R}^K$, and $c_{30}, c_{36} \in \mathbb{R}_+$ such that for N sufficiently large,*

$$|f(\boldsymbol{x}) - \hat{f}_N(\boldsymbol{x})| \leq c_{30} a_N^{T+P'+3} S_N^{-P} \qquad \text{for all } \boldsymbol{x} \in [-a_N, a_N]^K \setminus \chi_N,$$
$$|\hat{f}_N(\boldsymbol{x})| \leq c_{36} a_N^{P'} S_N^{K'+TP} \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^K$$

where $\Pr_{\boldsymbol{X}_N}(\chi_N) < c_{29}\kappa_N$.

Proof. We first consider the case when $l = 0$. Since f is a J-piecewise P-smooth $(K', 0)$-GHIM, there exists a P-smooth $g : \mathbb{R}^K \to \mathbb{R}$ and a J-side polytope $\Omega \subset \mathbb{R}^K$ such that

$$f(\boldsymbol{x}) = g(\boldsymbol{a}_1^\top \boldsymbol{x}, \ldots, \boldsymbol{a}_{K'}^\top \boldsymbol{x}) \cdot \mathbb{I}_\Omega(\boldsymbol{x})$$

according to Definition III.3. By Lemma B.8, there exists a $\tilde{f}_N \in \mathcal{G}_{l,R,K',K,\gamma}$ of the

form

$$\tilde{f}_N(\boldsymbol{x}) = \sum_{r=1}^{R} \xi_r^{[3]} \phi\Big(\sum_{k'=1}^{4K'} \xi_{r,k'}^{[2]} \phi\Big(\sum_{k=1}^{K} \xi_{r,k',k}^{[1]} x_k + \xi_{r,k',0}^{[1]} \Big) + \xi_{r,0}^{[2]} \Big) + \xi_0^{[3]},$$

such that $\|\tilde{f}_N\|_\infty \leq c_{64} a_N^{P'} S_N^{K'+TP}$ and $|f(\boldsymbol{x}) - \tilde{f}_N(\boldsymbol{x})| \leq c_{65} a_N^{T+P'+3} S_N^{-P}$ for all $\boldsymbol{x} \in [-a_N, a_N]^K$ except for a set with $\Pr_{\boldsymbol{X}_N}$ measure less than $\frac{1}{2} c_{29} \kappa_N$. We insert a neural network of the type defined in Lemma 6 of Bauer et al. [2019] to make $\tilde{f}_N$ vanish outside Ω:

$$\begin{aligned} \hat{f}_N(\boldsymbol{x}) = & \sum_{r=1}^{R} \xi_r^{[3]} \phi\Bigg(\sum_{k'=1}^{4K'} \xi_{r,k'}^{[2]} \phi\Big(\sum_{k=1}^{K} \xi_{r,k',k}^{[1]} x_k + \xi_{r,k',0}^{[1]} \Big) \\ & + \sum_{k'=4K'+1}^{4K'+J} \xi_{r,k'}^{[2]} \phi\Big(\sum_{k=1}^{K} \xi_{r,k',k}^{[1]} x_k + \xi_{r,k',0}^{[1]} \Big) + \xi_{r,0}^{[2]} \Bigg) + \xi_0^{[3]} \end{aligned}$$

such that $|\xi_{r,k'}^{[2]}|, |\xi_{r,k',k}^{[1]}| < \gamma$ $(r = 1, \ldots, R;\ k' = 4K'+1, \ldots, 4K'+J;\ k = 1, \ldots, K)$ and for some constant $c_{38}, c_{39} \in \mathbb{R}_+$,

$$\begin{aligned} |\tilde{f}_N(\boldsymbol{x}) - \hat{f}_N(\boldsymbol{x})| < & c_{38} |f(\boldsymbol{x}) - \tilde{f}_N(\boldsymbol{x})| \qquad \text{for all } \boldsymbol{x} \in \Omega \setminus \partial_{\delta_N} \Omega \\ |\hat{f}_N(\boldsymbol{x})| < & c_{39} |f(\boldsymbol{x}) - \tilde{f}_N(\boldsymbol{x})| \qquad \text{for all } \boldsymbol{x} \in \Omega^c \setminus \partial_{\delta_N} \Omega \end{aligned}$$

for all N sufficiently large, where $\partial_{\delta_N} \Omega \subset \mathbb{R}^K$ is the collection of all the points within a distance of δ_N from the boundary of Ω. By setting δ_N to be the same as the width of the hypercubes in the proof for Theorem 2 in Bauer et al. [2019], $\partial_{\delta_N} \Omega$ has $\Pr_{\boldsymbol{X}_N}$ measure less than the union of all the hypercube boundaries for N sufficiently large, since the number of hypercubes increases with N. Thus $\Pr_{\boldsymbol{X}_N}(\partial_{\delta_N} \Omega) < \frac{1}{2} c_{29} \kappa_N$ for N sufficiently large. Therefore, for some constants $c_{30}, c_{36} \in \mathbb{R}_+$,

$$\begin{aligned} |f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| \leq & c_{30} a_N^{T+P'+3} S_N^{-P} \qquad \text{for all } \boldsymbol{x} \in [-a_N, a_N]^K \setminus \chi_N, \\ |\hat{f}(\boldsymbol{x})| \leq & c_{36} a_N^{P'} S_N^{K'+TP} \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^K \end{aligned}$$

where $\Pr_{\boldsymbol{X}_N}(\chi_N) < c_{29}\kappa_N$. Finally, observe that $\hat{f}_N \in \mathcal{G}_{0,R,K'+\lceil J/4\rceil,K,\gamma} \subset \mathcal{G}_{0,R,K'+J,K,\gamma}$, which completes the proof for the case of $l = 0$.

For $l > 0$, we have

$$f(\boldsymbol{x}) = \sum_{r=1}^{\bar{R}} f_r(\boldsymbol{x}) \cdot \mathbb{I}_{\Omega_r}(\boldsymbol{x}) = \sum_{r=1}^{\bar{R}} g_r(h_{r,1}(\boldsymbol{x}), \ldots, h_{r,K'}(\boldsymbol{x})) \cdot \mathbb{I}_{\Omega_r}(\boldsymbol{x})$$

for some P-smooth $(K', l-1)$-GHIMs $g_r : \mathbb{R}^{K'} \to \mathbb{R}$ $(r = 1, \ldots, \bar{R})$ and $h_{r,k} : \mathbb{R}^K \to \mathbb{R}$ $(r = 1, \ldots, \bar{R};\ k = 1, \ldots, K')$, and J-side polytopes $\Omega_r \subset \mathbb{R}^K$ $(r = 1, \ldots, \bar{R})$. By Lemma B.8, $f_r(\boldsymbol{x})$ can be approximated by $\tilde{f}_{r,N}$ that is a composition of $\tilde{g}_{r,N} \in \mathcal{G}_{0,R,K',K',\gamma}$ with $\hat{h}_{r,k,N} \in \mathcal{G}_{l-1,R,K',K,\gamma}$:

$$\tilde{f}_{r,N}(\boldsymbol{x}) = \sum_{r=1}^{R} \xi_r^{[3]} \phi\Big(\sum_{k'=1}^{4K'} \xi_{r,k'}^{[2]} \phi\Big(\sum_{k=1}^{K'} \xi_{r,k',k}^{[1]} \hat{h}_{r,k,N}(\boldsymbol{x}) + \xi_{r,k',0}^{[1]} \Big) + \xi_{r,0}^{[2]} \Big) + \xi_0^{[3]}.$$

Let Ω_r be bounded by hyperplanes $\tilde{\boldsymbol{a}}_{r,j}\boldsymbol{x} + \bar{a}_{r,j} \leq 0$ $(j = 1, \ldots, J)$ with $\tilde{\boldsymbol{a}}_{r,j} \in \mathbb{R}^K$ and $\bar{\boldsymbol{a}}_{r,j} \in \mathbb{R}$. Define $h_{r,K'+j}(\boldsymbol{x}) = \tilde{\boldsymbol{a}}_{r,j}\boldsymbol{x} + \bar{a}_{r,j}$, which is P-smooth, since it is a linear function. Then $h_{r,K'+j}$ $(j = 1, \ldots, J)$ can be approximated by some $\hat{h}_{r,K'+j,N} \in \mathcal{G}_{l-1,R,K',K,\gamma}$ with the approximation error no greater than that of $h_{r,k}$ and $\hat{h}_{r,k,N}$ for $k = 1, \ldots K'$. Then by applying the same argument as in the case for $l = 0$, we can insert a neural network of the type defined in Lemma 6 of Bauer et al. [2019] that results in

$$\hat{f}_{r,N}(\boldsymbol{x}) = \sum_{r=1}^{R} \xi_r^{[3]} \phi\Big(\sum_{k'=1}^{4K'+J} \xi_{r,k'}^{[2]} \phi\Big(\sum_{k=1}^{K'+J} \xi_{r,k',k}^{[1]} \hat{h}_{r,k,N}(\boldsymbol{x}) + \xi_{r,k',0}^{[1]} \Big) + \xi_{r,0}^{[2]} \Big) + \xi_0^{[3]}$$

such that the weights are all bounded by γ, and for some constants $c_{30}, c_{36} \in \mathbb{R}_+$,

$$|f(\boldsymbol{x}) - \hat{f}(\boldsymbol{x})| \leq c_{30} a_N^{T+P'+3} S_N^{-P} \qquad \text{for all } \boldsymbol{x} \in [-a_N, a_N]^K \setminus \chi_N,$$
$$|\hat{f}(\boldsymbol{x})| \leq c_{36} a_N^{P'} S_N^{K'+TP} \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^K$$

where $\Pr_{\boldsymbol{X}_N}(\chi_N) < c_{29}\kappa_N$. In addition, $\hat{f}_N \in \mathcal{G}_{l,R,K'+\lceil J/4\rceil,K'+J,\gamma} \subset \mathcal{G}_{l,R,K'+J,K'+J,\gamma}$, which completes the proof for the case of $l > 0$. □

Proof of Lemma III.16. Lemma B.7 decomposes the total L_2 error of $\boldsymbol{\beta}^*$ into estimation error, approximation error, and bias due to individual effects (which is of order M^{-1}. The estimation error depends on the covering number of the collection of neural networks, which is given by Lemma 2 in Bauer et al. [2019] by setting the sample size to MV. The approximation error is given by Lemma B.9 by setting $N = V$ and $\Pr_{\boldsymbol{X}}(A) = \sum_{v=1}^{V} \mathbb{I}[\boldsymbol{d}_v \in A]$. The rest of the proofs is the same as the proof for Theorem 1 in Bauer et al. [2019], except that here we do not require the neural network complexity (which is controlled by R) to be a function of M and V, which makes the error bound in Equation (3.8) dependent on R.

Proof of Theorem III.17. Let $\beta^{\circ}_{M,V,R,q}(\boldsymbol{d}_v)$ be the thresholded estimator. Let MV be sufficiently large so that $c_{81}\log(MV)^{-1} < \psi/2$. For $q \in \{1,\ldots,Q\}$ and $v \in \{1,\ldots,V\}$, notice that

$$\begin{aligned}
&\mathrm{E}\left[\left|\operatorname{sign}[\beta^{*}_{M,V,R,q}(\boldsymbol{d}_v)] - \operatorname{sign}[\beta^{\circ}_{M,V,R,q}(\boldsymbol{d}_v)]\right|\right] \\
&= \Pr\left[\left|\operatorname{sign}[\beta^{*}_{M,V,R,q}(\boldsymbol{d}_v)]\right| \neq \left|\operatorname{sign}[\beta^{\circ}_{M,V,R,q}]\right|\right] \\
&+ \Pr\left[\operatorname{sign}[\beta^{*}_{M,V,R,q}(\boldsymbol{d}_v)] = -\operatorname{sign}[\beta^{\circ}_{M,V,R,q}]\right].
\end{aligned}$$

Moreover,

$$
\begin{aligned}
&\Pr\left[|\operatorname{sign}[\beta^*_{M,V,R,q}(\boldsymbol{d}_v)]| \neq |\operatorname{sign}[\beta^\circ_{M,V,R,q}]|\right] \\
=&\Pr\left[|\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v)| \geq \rho_{M,V,R,q}\right]\mathbb{I}\left[\beta^*_{M,V,R,q}(\boldsymbol{d}_v) = 0\right] \\
+&\Pr\left[|\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v)| < \rho_{M,V,R,q}\right]\mathbb{I}\left[\beta^*_{M,V,R,q}(\boldsymbol{d}_v) \neq 0\right] \\
\leq&\Pr\left[|\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v) - \beta^*_{M,V,R,q}(\boldsymbol{d}_v)| \geq c_{80}\log(MV)^{-1}\right]\mathbb{I}\left[\beta^*_{M,V,R,q}(\boldsymbol{d}_v) = 0\right] \\
+&\Pr\left[|\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v) - \beta^*_{M,V,R,q}(\boldsymbol{d}_v)| > \psi/2\right]\mathbb{I}\left[\beta^*_{M,V,R,q}(\boldsymbol{d}_v) \neq 0\right] \\
\leq&\Pr\left[|\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v) - \beta^*_{M,V,R,q}(\boldsymbol{d}_v)| \geq c_{80}\log(MV)^{-1}\right] \\
\leq&c_{80}^{-2}\log(MV)^2\,\mathrm{E}\left[\left|\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v) - \beta^*_{M,V,R,q}(\boldsymbol{d}_v)\right|^2\right]
\end{aligned}
$$

by Markov's Inequality. By the same argument,

$$
\begin{aligned}
&\Pr\left[\operatorname{sign}[\beta^*_{M,V,R,q}(\boldsymbol{d}_v)] = -\operatorname{sign}[\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v)]\right] \\
\leq&c_{80}^{-2}\log(MV)^2\,\mathrm{E}\left[\left|\beta^\circ_{M,V,R,q}(\boldsymbol{d}_v) - \beta^*_{M,V,R,q}(\boldsymbol{d}_v)\right|^2\right].
\end{aligned}
$$

Thus

$$
\begin{aligned}
&\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left\|\operatorname{sign}[\boldsymbol{\beta}^*_{M,V,R}(\boldsymbol{d}_v)] - \operatorname{sign}[\boldsymbol{\beta}^\circ_{M,V,R}(\boldsymbol{d}_v)]\right\|_0\right] \\
\leq&2Qc_{80}^{-2}\log(MV)^2\,\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{\beta}^\circ_{M,V,R,q}(\boldsymbol{d}_v) - \boldsymbol{\beta}^*_{M,V,R,q}(\boldsymbol{d}_v)\right\|^2\right] \\
\leq&c_{79}[\log(MV)^5(M^{-1}V^{-1}R + R^{-\frac{2b}{K}}) + \log(MV)^2M^{-1}]
\end{aligned}
$$

for some constant $c_{79} > 0$.

Proof of Corollary III.18. Consider the conditional loss function for estimating the individual effects:

$$
\begin{aligned}
&\ell_{\alpha|\beta,\sigma^2}(\boldsymbol{\alpha}|\hat{\boldsymbol{\beta}}_{M,V,R},\sigma^2) \\
&= M^{-1}V^{-1}\sum_{v=1}^{V}\left\|\boldsymbol{y}(\boldsymbol{d}_v)-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{M,V,R}(\boldsymbol{d}_v)-\boldsymbol{\alpha}(\boldsymbol{d}_v)\right\|_2^2\sigma^{-2}(\boldsymbol{d}_v). \\
=&M^{-1}V^{-1}\sum_{m=1}^{M}\sum_{v=1}^{V}\left|y_m(\boldsymbol{d}_v)-\boldsymbol{x}_m^\top\hat{\boldsymbol{\beta}}_{M,V,R}(\boldsymbol{d}_v)-\alpha_m(\boldsymbol{d}_v)\right|^2\sigma^{-2}(\boldsymbol{d}_v) \\
=&M^{-1}V^{-1}\sum_{m=1}^{M}\sum_{v=1}^{V}\left|\alpha_m^*(\boldsymbol{d}_v)+\boldsymbol{x}_m^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\hat{\boldsymbol{\beta}}_{M,V,R}(\boldsymbol{d}_v)\right]+\epsilon_m(\boldsymbol{d}_v)-\alpha_m(\boldsymbol{d}_v)\right|^2\sigma^{-2}(\boldsymbol{d}_v).
\end{aligned}
$$

Since each α_m is fitted independently by using samples on V voxels, it has error bound

$$
\begin{aligned}
&\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left|\alpha_m^*(\boldsymbol{d}_v)+\boldsymbol{x}_m^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\hat{\boldsymbol{\beta}}_{M,V,R}(\boldsymbol{d}_v)\right]-\hat{\alpha}_{M,V,R,m}(\boldsymbol{d}_v)\right|^2\right] \\
\leq& c_{82}[\log(V)^3(V^{-1}R+R^{-\frac{2b}{K}})]
\end{aligned}
$$

for some constant $c_{82}\in\mathbb{R}_+$. Moreover,

$$
\begin{aligned}
&\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left|\boldsymbol{x}_m^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\hat{\boldsymbol{\beta}}_{M,V,R}(\boldsymbol{d}_v)\right]\right|^2\right] \\
\leq& c_{83}[\log(MV)^5((MV)^{-1}R+R^{-\frac{2b}{K}})+\log(M)^2M^{-1}]
\end{aligned}
$$

for some constants $c_{83} \in \mathbb{R}_+$ by Lemma III.16. Then

$$
\begin{aligned}
&\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left|\alpha_m^*(\boldsymbol{d}_v)-\hat{\alpha}_{M,V,R,m}(\boldsymbol{d}_v)\right|^2\right]^{1/2}\\
\leq&\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left|\alpha_m^*(\boldsymbol{d}_v)+\boldsymbol{x}_m^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\hat{\boldsymbol{\beta}}_{M,V,R}(\boldsymbol{d}_v)\right]-\hat{\alpha}_{M,V,R,m}(\boldsymbol{d}_v)\right|^2\right]^{1/2}\\
+&\mathrm{E}\left[V^{-1}\sum_{v=1}^{V}\left|\boldsymbol{x}_m^\top\left[\boldsymbol{\beta}^*(\boldsymbol{d}_v)-\hat{\boldsymbol{\beta}}_{M,V,R}(\boldsymbol{d}_v)\right]\right|^2\right]^{1/2}\\
\leq&c_{26}[\log(MV)^5((MV)^{-1}R+R^{-\frac{2b}{K}})+\log(M)^2M^{-1}+\log(V)^3(V^{-1}R)]^{1/2}
\end{aligned}
$$

for some constant $c_{26} \in \mathbb{R}_+$.

Proof of Corollary III.19. The proof follows the same argument as in the proof for Corollary III.18.

APPENDIX C

DALEA for Categorical Outcomes

C.1 DALEA for categorical outcomes

For classification problems, we use the softmax function to map the values of the nodes in the last layer to a probability vector that sums to one:

$$\begin{aligned}&\text{softmax}^{-1}\left\{\text{E}\left[\boldsymbol{y}^{(n)}\right]\right\} = \\ &\boldsymbol{\beta}_{J+1}\boldsymbol{\phi}\left[\cdots\boldsymbol{\beta}_2\boldsymbol{\phi}\left[\boldsymbol{\beta}_1\boldsymbol{x}^{(n)}+\boldsymbol{\alpha}_1+\boldsymbol{\delta}_1^{(n)}\right]+\boldsymbol{\alpha}_2+\boldsymbol{\delta}_2^{(n)}\cdots\right]+\boldsymbol{\alpha}_{J+1}+\boldsymbol{\delta}_{J+1}^{(n)},\end{aligned} \tag{C.1}$$

where

$$\text{softmax}(\boldsymbol{z}) = \frac{[\exp(\boldsymbol{z}), 1]^\top}{\mathbf{1}^\top \exp(\boldsymbol{z}) + 1}$$

Suppose there are K outcome categories, and each outcome $\boldsymbol{y}^{(n)}$ belongs to category $\bar{k}^{(n)}$. We represent the outcome as a one-hot vector $\boldsymbol{y}^{(n)} = [y_1^{(n)}, \ldots, y_K^{(n)}]$, where

$$y_k^{(n)} = \begin{cases} 1, & \text{if } k = \bar{k}^{(n)}, \\ 0, & \text{otherwise,} \end{cases}$$

for $k = 1, \ldots, K$. The problem is to sample $\boldsymbol{z}^{(n)}|\boldsymbol{y}^{(n)}, \boldsymbol{\mu}^{(n)}, \tau^2$ from the model

$$\begin{aligned}\boldsymbol{z}^{(n)} &\overset{\text{iid}}{\sim} \text{N}(\boldsymbol{\mu}^{(n)}, \tau^2) \\ \boldsymbol{\pi}^{(n)} &= \text{softmax}(\boldsymbol{z}^{(n)}) \\ \boldsymbol{y}^{(n)} &\sim \text{Multinoulli}(\boldsymbol{\pi}^{(n)}),\end{aligned}$$

where $\boldsymbol{z}^{(n)}, \boldsymbol{\mu}^{(n)} \in \mathbb{R}^{K-1}$ and $\tau^2 \in \mathbb{R}$. To do this, we update one element of $\boldsymbol{z}^{(n)}$ at a time while fixing the other $K-2$ elements. Then the joint log density for element k

is

$$
\begin{aligned}
&-\log f\left(z_k^{(n)}, \boldsymbol{y}^{(n)} \middle| \boldsymbol{z}_{-k}^{(n)}, \mu_k^{(n)}, \tau^2\right) \\
&= -\log f\left(z_k^{(n)}, k^{(n)} \middle| \boldsymbol{z}_{-k}^{(n)}, \mu_k^{(n)}, \tau^2\right) \\
&= \frac{1}{2\tau^2}\left(z_k^{(n)} - \mu_k^{(n)}\right)^2 + \log\left[\exp\left(z_k^{(n)}\right) + \sum \exp\left(\boldsymbol{z}_{-k}^{(n)}\right)\right] - z_{\bar{k}^{(n)}}^{(n)} + C_0 \\
&= \frac{1}{2\tau^2}\left(z_k^{(n)} - \mu_k^{(n)}\right)^2 + \log\left[\frac{\exp\left(z_k^{(n)}\right)}{\sum \exp\left(\boldsymbol{z}_{-k}^{(n)}\right) + 1} + 1\right] - z_{\bar{k}^{(n)}}^{(n)} + C_1 \\
&= \frac{1}{2\tau^2}\left(z_k^{(n)} - \mu_k^{(n)}\right)^2 + \log\left[\exp\left(z_k^{(n)} - a_k^{(n)}\right) + 1\right] - z_{\bar{k}^{(n)}}^{(n)} + C_1 \\
&= \frac{1}{2\tau^2}\left(z_k^{(n)} - \mu_k^{(n)}\right)^2 + \log\left[\exp\left\{s_k^{(n)}\left(z_k^{(n)} - a_k^{(n)}\right)\right\} + 1\right] + C_2
\end{aligned}
$$

where

$$
\begin{aligned}
a_k^{(n)} &= \log\left[\sum \exp\left(\boldsymbol{z}_{-k}^{(n)}\right) + 1\right] \\
s_k^{(n)} &= \begin{cases} -1, & \text{if } k = \bar{k}^{(n)}, \\ 1, & \text{otherwise,} \end{cases}
\end{aligned}
$$

Notice that $\log\left[\exp\left(z_k^{(n)} - a_k^{(n)}\right) + 1\right]$ is the softplus function with respect to $z_k^{(n)}$ centered at $a_k^{(n)}$, which approaches the ReLU function $\max(0, \cdot)$ when $z_k^{(n)} \to \pm\infty$, and is convex around $a_k^{(n)}$. Thus we can approximate it by breaking its domain into

three parts:

$$\begin{aligned}
\log\{\exp[s(z-a)]+1\} &= \phi[s(z-a)] \\
&\approx \psi[s(z-a)] \\
&= \begin{cases} 0 & \text{if } s(z-a) < -\frac{1}{2c} \\ \frac{c}{2}(z-a+\frac{s}{2c})^2 & \text{if } -\frac{1}{2c} \le s(z-a) \le \frac{1}{2c} \\ s(z-a) & \text{if } s(z-a) > \frac{1}{2c} \end{cases} \\
&= \begin{cases} \frac{s-1}{2}(z-a) & \text{if } z \in (-\infty,\, a-\frac{1}{2c}) \\ \frac{c}{2}(z-a+\frac{s}{2c})^2 & \text{if } z \in [a-\frac{1}{2c},\, a+\frac{1}{2c}] \\ \frac{s+1}{2}(z-a) & \text{if } z \in (a+\frac{1}{2c},\, \infty) \end{cases}
\end{aligned}$$

where $c > 0$ is a constant for approximating the logistic function with a hard sigmoid function

$$\{\exp[s(z-a)]^{-1}+1\}^{-1} = \phi'[s(z-a)] \approx \psi'[s(z-a)] = \min[\max[0.5+sc(z-a),0],1].$$

For example, the first-order Taylor polynomial of ϕ' at 0 sets $c = 0.25$, while TensorFlow and Theano sets $c = 0.2$. (For the middle part, we may be tempted to use the Taylor polynomial of ϕ centered at a (i.e. $\log(2) + 0.5(z-a) + 0.125(z-a)^2$) or centered at one of the two boundary points, but that does not guarantee the overall

function to be continuous.) Then the density function is broken into three cases:

$$
\begin{aligned}
&-\log f\left(z_k^{(n)}, \boldsymbol{y}^{(n)} \middle| \boldsymbol{z}_{-k}^{(n)}, \mu_k^{(n)}, \tau^2\right) \\
=&\frac{1}{2\tau^2}\left(z_k^{(n)}-\mu_k^{(n)}\right)^2+\log\left[\exp\left\{s_k^{(n)}\left(z_k^{(n)}-a_k^{(n)}\right)\right\}+1\right]+C_2 \\
\approx&\frac{1}{2\tau^2}\left(z_k^{(n)}-\mu_k^{(n)}\right)^2+\psi\left[s_k^{(n)}\left(z_k^{(n)}-a_k^{(n)}\right)\right]+C_2 \\
=&\begin{cases}
\frac{1}{2}\tau^{-2}\left(z_k^{(n)}-\mu_k^{(n)}\right)^2+C_2 & \text{if } s_k^{(n)}\left(z_k^{(n)}-a_k^{(n)}\right)<-\frac{1}{2c} \\
\frac{1}{2}(\tau^{-2}+c)\left\{z-\left[\frac{\tau^{-2}}{\tau^{-2}+c}\mu+\frac{c}{\tau^{-2}+c}\left(a-\frac{s_k^{(n)}}{2c}\right)\right]\right\}+C_3 & \text{if } -\frac{1}{2c}\le s_k^{(n)}\left(z_k^{(n)}-a_k^{(n)}\right)\le\frac{1}{2c} \\
\frac{1}{2}\tau^{-2}\left[z_k^{(n)}-\left(\mu_k^{(n)}-s_k^{(n)}\tau^2\right)\right]^2+C_4 & \text{if } s_k^{(n)}\left(z_k^{(n)}-a_k^{(n)}\right)>\frac{1}{2c}
\end{cases} \\
=&\begin{cases}
\frac{1}{2}\tau^{-2}\left[z_k^{(n)}-\left(\mu_k^{(n)}-\frac{s_k^{(n)}-1}{2}\tau^2\right)\right]^2+C_4 & \text{if } z_k^{(n)}\in\left(-\infty,\, a_k^{(n)}-\frac{1}{2c}\right) \\
\frac{1}{2}(\tau^{-2}+c)\left\{z-\left[\frac{\tau^{-2}}{\tau^{-2}+c}\mu+\frac{c}{\tau^{-2}+c}\left(a-\frac{s_k^{(n)}}{2c}\right)\right]\right\}+C_3 & \text{if } z_k^{(n)}\in\left[a_k^{(n)}-\frac{1}{2c},\, a_k^{(n)}+\frac{1}{2c}\right] \\
\frac{1}{2}\tau^{-2}\left[z_k^{(n)}-\left(\mu_k^{(n)}-\frac{s_k^{(n)}+1}{2}\tau^2\right)\right]^2+C_4 & \text{if } z_k^{(n)}\in\left(a_k^{(n)}+\frac{1}{2c},\, \infty\right)
\end{cases}
\end{aligned}
$$

In all the cases the density has a quadratic form, and the density overall is continuous, which implies that the distribution is a three-part heterogeneous normal distribution.

APPENDIX D

Supplementary Tables and Figures for DALEA

Figure D.1: Posterior distributions for data generated with Gaussian noise.

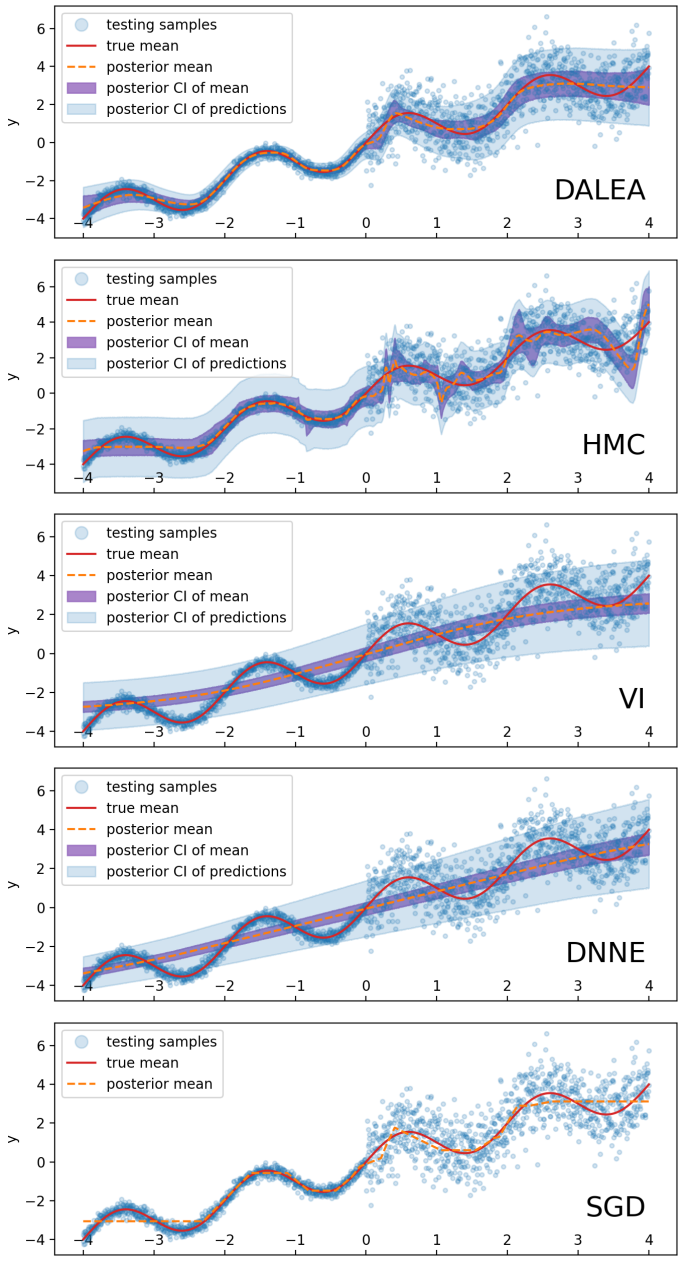


Figure D.2: Posterior distributions for data generated with centered chi-squared noise.

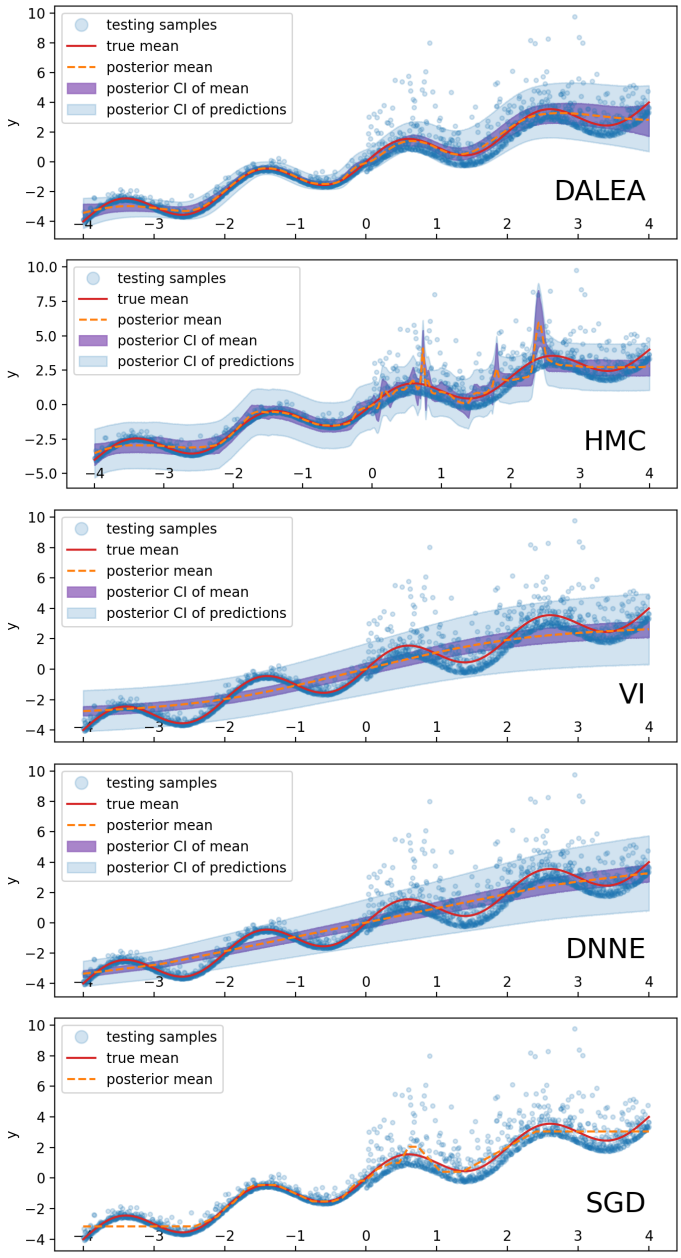
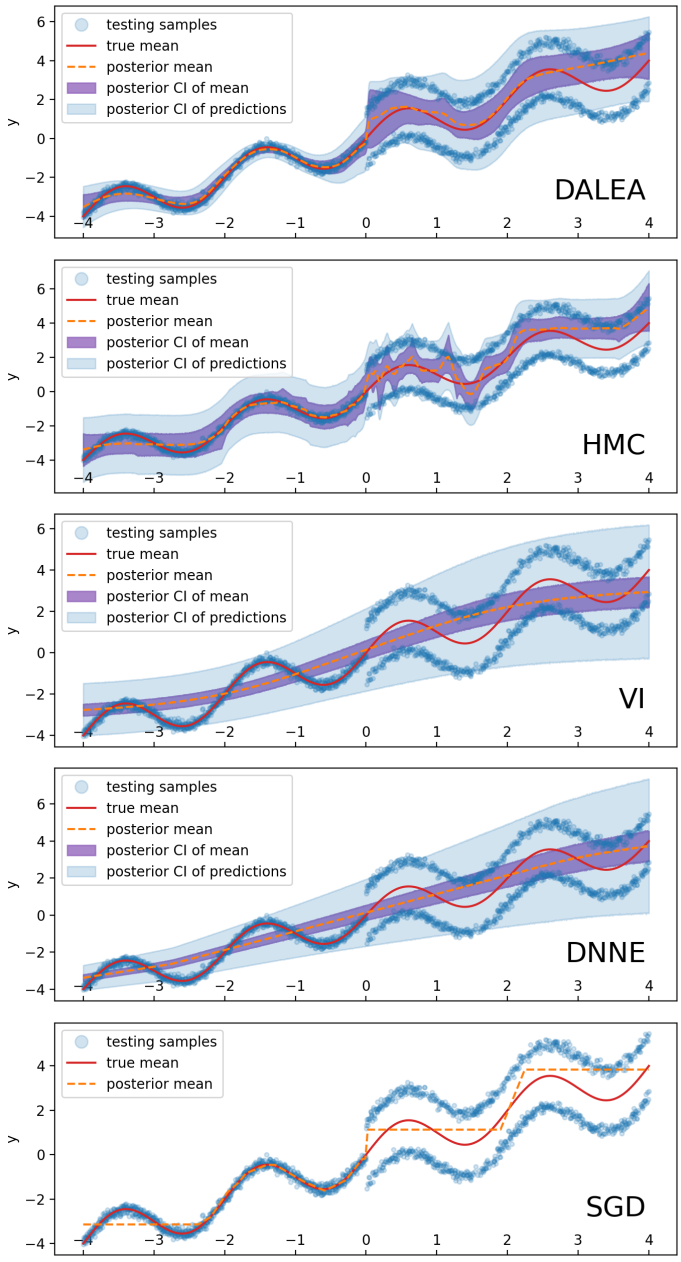


Figure D.3: Posterior distributions for data generated with Gaussian mixture noise.



BIBLIOGRAPHY

BIBLIOGRAPHY

David J Balding, Martin Bishop, and Chris Cannings. *Handbook of statistical genetics*. John Wiley & Sons, 2008.

Deanna M Barch, Gregory C Burgess, Michael P Harms, Steven E Petersen, Bradley L Schlaggar, Maurizio Corbetta, Matthew F Glasser, Sandra Curtiss, Sachin Dixit, Cindy Feldt, et al. Function in the human connectome: task-fmri and individual differences in behavior. *Neuroimage*, 80:169–189, 2013.

Andrew R Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, pages 561–576. Springer, 1991.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.

Ulrike Basten, Kirsten Hilger, and Christian J Fiebach. Where smart brains are different: A quantitative meta-analysis of functional and structural brain imaging studies on intelligence. *Intelligence*, 51:10–27, 2015.

Benedikt Bauer, Michael Kohler, et al. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Annals of Statistics*, 47(4):2261–2285, 2019.

PG Benardos and G-C Vosniakos. Optimizing feedforward artificial neural network architecture. *Engineering applications of artificial intelligence*, 20(3):365–382, 2007.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, pages 1165–1188, 2001.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *25th annual conference on neural information processing systems (NIPS 2011)*, volume 24. Neural Information Processing Systems Foundation, 2011.

Julius Berner, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. The modern mathematics of deep learning. *arXiv preprint arXiv:2105.04026*, 2021.

UK Biobank. About uk biobank. *Available at h ttps://www. ukbiobank. ac. uk/a bout-biobank-uk*, 2014.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

Léon Bottou et al. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991.

Matthew Brand. Incremental singular value decomposition of uncertain data with missing values. In *European Conference on Computer Vision*, pages 707–720. Springer, 2002.

James M Broadway and Randall W Engle. Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, 42(2):563–570, 2010.

Matthias Bussas, Christoph Sawade, Nicolas Kühn, Tobias Scheffer, and Niels Landwehr. Varying-coefficient models for geospatial transfer learning. *Machine Learning*, 106(9-10):1419–1440, 2017.

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562 (7726):203, 2018.

Craig Calcaterra and Axel Boldt. Approximating with gaussians. *arXiv preprint arXiv:0805.3795*, 2008.

Lon R Cardon and Lyle J Palmer. Population stratification and spurious allelic association. *The Lancet*, 361(9357):598–604, 2003.

BJ Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, 2018.

Tony F Chan and Jianhong Shen. *Image processing and analysis: variational, PDE, wavelet, and stochastic methods*. SIAM, 2005.

Gabriel Chartrand, Phillip M Cheng, Eugene Vorontsov, Michal Drozdzal, Simon Turcotte, Christopher J Pal, Samuel Kadoury, and An Tang. Deep learning: a primer for radiologists. *Radiographics*, 37(7):2113–2131, 2017.

Yao Chen, Xiao Wang, Linglong Kong, and Hongtu Zhu. Local region sparse learning for image-on-scalar regression. *arXiv preprint arXiv:1605.08501*, 2016.

Yao Chen, Qingyi Gao, Faming Liang, and Xiao Wang. Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics*, pages 1–9, 2020.

J. Chumbley, K. J. Worsley, G. Flandin, and K. J. Friston. False discovery rate revisited: Fdr and topological inference using gaussian random fields. *NeuroImage*, 44:62–70, 2009.

Justin R Chumbley and Karl J Friston. False discovery rate revisited: Fdr and topological inference using gaussian random fields. *Neuroimage*, 44(1):62–70, 2009.

Laura Clarke, Xiangqun Zheng-Bradley, Richard Smith, Eugene Kulesha, Chunlin Xiao, Iliana Toneva, Brendan Vaughan, Don Preuss, Rasko Leinonen, Martin Shumway, et al. The 1000 genomes project: data management and community access. *Nature methods*, 9(5):459–462, 2012.

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

Cameron Craddock, Sharad Sikka, Brian Cheung, Ranjeet Khanuja, Satrajit S Ghosh, Chaogan Yan, Qingyang Li, Daniel Lurie, Joshua Vogelstein, Randal Burns, Stanley Colcombe, Maarten Mennes, Clare Kelly, Adriana Di Martino, Francisco Xavier Castellanos, and Michael Milham. Towards automated analysis of connectomes: The configurable pipeline for the analysis of connectomes (c-pac). *Frontiers in Neuroinformatics*, (42), 2013. ISSN 1662-5196. doi: 10.3389/conf.fninf.2013.09.00042. URL `http://www.frontiersin.org/neuroinformatics/10.3389/conf.fninf.2013.09.00042/full`.

Noel AC Cressie and Noel A Cassie. *Statistics for Spatial Data*, volume 900. Wiley New York, 1993.

Li Deng and Dong Yu. Deep learning: methods and applications. *Foundations and trends in signal processing*, 7(3–4):197–387, 2014.

Rounak Dey and Seunggeun Lee. *hdpca: Principal Component Analysis in High-Dimensional Data*, 2016. URL `https://CRAN.R-project.org/package=hdpca`. R package version 1.0.0.

Rounak Dey and Seunggeun Lee. Asymptotic properties of principal component analysis and shrinkage-bias adjustment under the generalized spiked population model. *Journal of Multivariate Analysis*, 173:145–164, 2019.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X Castellanos, Kaat Alaerts, Jeffrey S Anderson, Michal Assaf, Susan Y Bookheimer,

Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014.

Peter J Diggle, JA Tawn, and RA Moyeed. Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(3):299–350, 1998.

Julien Dubois, Paola Galdi, Lynn K Paul, and Ralph Adolphs. A distributed brain network predicts general intelligence from resting-state human neuroimaging data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1756): 20170284, 2018.

John Duncan, Hazel Emslie, Phyllis Williams, Roger Johnson, and Charles Freer. Intelligence and the frontal lobe: The organization of goal-directed behavior. *Cognitive psychology*, 30(3):257–303, 1996.

Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. Deep neural network approximation theory. *arXiv preprint arXiv:1901.02220*, 2019.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

Jianqing Fan, Cong Ma, and Yiqiao Zhong. A selective overview of deep learning. *arXiv preprint arXiv:1904.05526*, 2019.

Jean Feng and Noah Simon. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*, 2017.

Matthias Feurer and Frank Hutter. Hyperparameter optimization. In *Automated Machine Learning*, pages 3–33. Springer, Cham, 2019.

Karl J Friston. Statistical parametric mapping. In *Neuroscience databases*, pages 237–250. Springer, 2003.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

Kevin J Galinsky, Gaurav Bhatia, Po-Ru Loh, Stoyan Georgiev, Sayan Mukherjee, Nick J Patterson, and Alkes L Price. Fast principal-component analysis reveals convergent evolution of adh1b in europe and east asia. *The American Journal of Human Genetics*, 98(3):456–472, 2016.

Alan E Gelfand, Hyon-Jung Kim, CF Sirmans, and Sudipto Banerjee. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396, 2003.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Natalia A Goriounova and Huibert D Mansvelder. Genes, cells and brain areas of intelligence. *Frontiers in human neuroscience*, 13:44, 2019.

Shelley Gray, Samuel Green, Mary Alt, T Hogan, Trudy Kuo, Shara Brinkley, and Nelson Cowan. The structure of working memory in young children and its relation to intelligence. *Journal of Memory and Language*, 92:183–201, 2017.

Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3): 362–386, 2020.

Lijie Gu, Li Wang, Wolfgang K Härdle, and Lijian Yang. A simultaneous confidence corridor for varying coefficient regression with sparse functional data. *TEST*, 23 (4):806–843, 2014.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

Donald J Hagler Jr, SeanN Hatton, M Daniela Cornejo, Carolina Makowski, Damien A Fair, Anthony Steven Dick, Matthew T Sutherland, BJ Casey, Deanna M Barch, Michael P Harms, et al. Image processing and analysis methods for the adolescent brain cognitive development study. *NeuroImage*, 202:116091, 2019.

Richard J Haier, Rex E Jung, Ronald A Yeo, Kevin Head, and Michael T Alkire. Structural brain variation and general intelligence. *Neuroimage*, 23(1):425–433, 2004.

Richard J Haier, Rex E Jung, Ronald A Yeo, Kevin Head, and Michael T Alkire. The neuroanatomy of general intelligence: sex matters. *NeuroImage*, 25(1):320–327, 2005.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

Kevin He, Han Xu, and Jian Kang. A selective overview of feature screening methods with applications to neuroimaging data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(2):e1454, 2019.

Luke J Hearne, Jason B Mattingley, and Luca Cocchi. Functional brain networks related to individual differences in human intelligence at rest. *Scientific reports*, 6 (1):1–8, 2016.

Moritz Helmstaedter, Kevin L Briggman, Srinivas C Turaga, Viren Jain, H Sebastian Seung, and Winfried Denk. Connectomic reconstruction of the inner plexiform layer in the mouse retina. *Nature*, 500(7461):168–174, 2013.

Kirsten Hilger, Matthias Ekman, Christian J Fiebach, and Ulrike Basten. Efficient hubs in the intelligent brain: Nodal efficiency of hub regions in the salience network is associated with general intelligence. *Intelligence*, 60:10–25, 2017.

Kirsten Hilger, Makoto Fukushima, Olaf Sporns, and Christian J Fiebach. Temporal stability of functional brain modules associated with human intelligence. *Human brain mapping*, 41(2):362–372, 2020.

Xiaowei Huang, Daniel Kroening, Marta Kwiatkowska, Wenjie Ruan, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. Safety and trustworthiness of deep neural networks: A survey. *arXiv preprint arXiv:1812.08342*, 2018.

Shiro Ikegawa. A short history of the genome-wide association study: where we were and where we are going. *Genomics & informatics*, 10(4):220, 2012.

Vugar E Ismailov. On the approximation by neural networks with bounded number of neurons in hidden layers. *Journal of Mathematical Analysis and Applications*, 417(2):963–969, 2014.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Pavel Izmailov, Wesley J Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are bayesian neural network posteriors really like? *arXiv preprint arXiv:2104.14421*, 2021.

Natalia B Janson. Non-linear dynamics of biological systems. *Contemporary Physics*, 53(2):137–168, 2012.

Emanuel Jauk, Aljoscha C Neubauer, Beate Dunst, Andreas Fink, and Mathias Benedek. Gray matter correlates of creative potential: A latent variable voxel-based morphometry study. *NeuroImage*, 111:312–320, 2015.

Terry L Jernigan, Sandra A Brown, and Gayathri J Dowling. The adolescent brain cognitive development study. *Journal of research on adolescence: the official journal of the Society for Research on Adolescence*, 28(1):154–156, 2018.

I. T. Jolliffe. *Principal component analysis*. Springer, New York, 2002. ISBN 978-0-387-95442-4.

Rex E Jung and Richard J Haier. The parieto-frontal integration theory (p-fit) of intelligence: converging neuroimaging evidence. *Behavioral and Brain Sciences*, 30 (2):135, 2007.

Andreas Kamilaris and Francesc X Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and electronics in agriculture*, 147:70–90, 2018.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *arXiv preprint arXiv:1703.04977*, 2017.

Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3): 239–251, 1945.

Justin Ker, Lipo Wang, Jai Rao, and Tchoyoson Lim. Deep learning applications in medical image analysis. *Ieee Access*, 6:9375–9389, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Michael Kohler and Adam Krzyżak. Nonparametric regression based on hierarchical interaction models. *IEEE Transactions on Information Theory*, 63(3):1620–1630, 2017.

Chayakrit Krittanawong, Kipp W Johnson, Robert S Rosenson, Zhen Wang, Mehmet Aydar, Usman Baber, James K Min, WH Wilson Tang, Jonathan L Halperin, and Sanjiv M Narayan. Deep learning for cardiovascular medicine: a practical primer. *European heart journal*, 40(25):2058–2073, 2019.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Seunggeun Lee, Fei Zou, and Fred A Wright. Convergence and prediction of principal component scores in high-dimensional settings. *Annals of statistics*, 38(6):3605, 2010.

Lexin Li and Xin Zhang. Parsimonious tensor response regression. *Journal of the American Statistical Association*, 112(519):1131–1146, 2017.

Xinyi Li, Li Wang, Huixia Judy Wang, and Alzheimer's Disease Neuroimaging Initiative. Sparse learning and structure identification for ultrahigh-dimensional image-on-scalar regression. *Journal of the American Statistical Association*, pages 1–15, 2020.

Faming Liang, Qizhai Li, and Lei Zhou. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523): 955–972, 2018.

Xia Liang, Qihong Zou, Yong He, and Yihong Yang. Topologically reorganized connectivity architecture of default-mode, executive-control, and salience networks across working memory task loads. *Cerebral cortex*, 26(4):1501–1511, 2016.

Ruiqi Liu, Ben Boukai, and Zuofeng Shang. Optimal nonparametric inference via deep neural network. *arXiv preprint arXiv:1902.01687*, 2019.

Ying Liu, Bowei Yan, Kathleen Merikangas, and Haochang Shou. Total variation regularized tensor-on-scalar regression. *arXiv preprint arXiv:1703.05264*, 2017.

Yulong Lu and Jianfeng Lu. A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in Neural Information Processing Systems*, 33, 2020.

Renqian Luo, Fei Tian, Tao Qin, Enhong Chen, and Tie-Yan Liu. Neural architecture optimization. *arXiv preprint arXiv:1808.07233*, 2018.

Junshui Ma, Robert P Sheridan, Andy Liaw, George E Dahl, and Vladimir Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. *Journal of chemical information and modeling*, 55(2):263–274, 2015.

David JC MacKay. Probable networks and plausible predictions-a review of practical bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469, 1995.

Steve Majerus, Frédéric Péters, Marion Bouffier, Nelson Cowan, and Christophe Phillips. The dorsal attention network reflects both encoding load and top–down control during working memory. *Journal of Cognitive Neuroscience*, 30(2):144–159, 2018.

Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3):243, 2012.

Daniel F McCaffrey and A Ronald Gallant. Convergence rates for single hidden layer feedforward networks. *Neural Networks*, 7(1):147–158, 1994.

Saroj K Meher and Ganapati Panda. Deep learning in astronomy: a tutorial perspective. *The European Physical Journal Special Topics*, pages 1–33, 2021.

Kyle Menary, Paul F Collins, James N Porter, Ryan Muetzel, Elizabeth A Olson, Vipin Kumar, Michael Steinbach, Kelvin O Lim, and Monica Luciana. Associations between cortical thickness and general intelligence in children, adolescents and young adults. *Intelligence*, 41(5):597–606, 2013.

Hrushikesh N Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, 8(1):164–177, 1996.

K Yu Michael, Jianzhu Ma, Jasmin Fisher, Jason F Kreisberg, Benjamin J Raphael, and Trey Ideker. Visible machine learning for biomedicine. *Cell*, 173(7):1562–1565, 2018.

Jan Mielniczuk and Joanna Tyrcha. Consistency of multilayer perceptron regression estimators. *Neural Networks*, 6(7):1019–1022, 1993.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

Kouichi Ozaki, Yozo Ohnishi, Aritoshi Iida, Akihiko Sekine, Ryo Yamada, Tatsuhiko Tsunoda, Hiroshi Sato, Hideyuki Sato, Masatsugu Hori, Yusuke Nakamura, et al. Functional snps in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nature genetics*, 32(4):650–654, 2002.

Kostantinos N Plataniotis and Dimitris Hatzinakos. Gaussian mixtures and their applications to signal processing. In *Advanced signal processing handbook*, pages 89–124. CRC Press, 2017.

Nicholas Polson and Veronika Rocková. Posterior concentration for sparse deep learning. *arXiv preprint arXiv:1803.09138*, 2018.

Jörg Polzehl and Vladimir G Spokoiny. Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):335–354, 2000.

Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and Sundaraja S Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, 51(5):1–36, 2018.

Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

W.J. Powers and C.P. Derdeyn. Neuroimaging, overview. pages 398 – 399, 2014. doi: https://doi.org/10.1016/B978-0-12-385157-4.00200-1. URL `http://www.sciencedirect.com/science/article/pii/B9780123851574002001`.

Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904, 2006.

Peihua Qiu. Jump surface estimation, edge detection, and image restoration. *Journal of the American Statistical Association*, 102(478):745–756, 2007.

David Reich, Alkes L Price, and Nick Patterson. Principal component analysis of genetic data. *Nature genetics*, 40(5):491, 2008.

María Roca, Alice Parr, Russell Thompson, Alexandra Woolgar, Teresa Torralva, Nagui Antoun, Facundo Manes, and John Duncan. Executive function and fluid intelligence after frontal lobe lesions. *Brain*, 133(1):234–247, 2010.

David Rolnick and Max Tegmark. The power of deeper networks for expressing natural functions. *arXiv preprint arXiv:1705.05502*, 2017.

Grant M Rotskoff and Eric Vanden-Eijnden. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *stat*, 1050:22, 2018.

Franco Scarselli and Ah Chung Tsoi. Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results. *Neural networks*, 11(1):15–37, 1998.

Johannes Schmidt-Hieber et al. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

Hugo G Schnack, Neeltje EM Van Haren, Rachel M Brouwer, Alan Evans, Sarah Durston, Dorret I Boomsma, René S Kahn, and Hilleke E Hulshoff Pol. Changes in thickness and surface area of the human cortex and their relationship with intelligence. *Cerebral cortex*, 25(6):1608–1617, 2015.

Ran Shi and Jian Kang. Thresholded multiscale gaussian processes with application to bayesian feature selection for massive neuroimaging data. *arXiv preprint arXiv:1504.06074*, 2015.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

Isabelle Simard, David Luck, Laurent Mottron, Thomas A Zeffiro, and Isabelle Soulières. Autistic fluid intelligence: Increased reliance on visual functional connectivity with diminished modulation of coupling by task difficulty. *NeuroImage: Clinical*, 9:467–478, 2015.

Ming Song, Yuan Zhou, Jun Li, Yong Liu, Lixia Tian, Chunshui Yu, and Tianzi Jiang. Brain spontaneous functional connectivity and intelligence. *Neuroimage*, 41 (3):1168–1176, 2008.

Chandra Sripada, Saige Rutherford, Mike Angstadt, Wesley K Thompson, Monica Luciana, Alexander Weigard, Luke H Hyde, and Mary Heitzeg. Prediction of neurocognition in youth from resting state fmri. *Molecular Psychiatry*, pages 1–9, 2019.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

Will Wei Sun and Lexin Li. Store: sparse tensor response regression and neuroimaging analysis. *The Journal of Machine Learning Research*, 18(1):4908–4944, 2017.

Karsten Tabelow, Jörg Polzehl, Vladimir Spokoiny, and Henning U Voss. Diffusion tensor imaging: structural adaptive smoothing. *NeuroImage*, 39(4):1763–1773, 2008.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

Duncan C Thomas, Robert W Haile, and David Duggan. Recent developments in genomewide association scans: a workshop summary and review. *The American Journal of Human Genetics*, 77(3):337–345, 2005.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Nathalie Tzourio-Mazoyer, Brigitte Landeau, Dimitri Papathanassiou, Fabrice Crivello, Olivier Etard, Nicolas Delcroix, Bernard Mazoyer, and Marc Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *Neuroimage*, 15(1):273–289, 2002.

Lucina Q Uddin, BT Thomas Yeo, and R Nathan Spreng. Towards a universal taxonomy of macro-scale functional human brain networks. *Brain topography*, 32 (6):926–942, 2019.

Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

Michael Wainberg, Daniele Merico, Andrew Delong, and Brendan J Frey. Deep learning in biomedicine. *Nature biotechnology*, 36(9):829–838, 2018.

Chaolong Wang, Xiaowei Zhan, Jennifer Bragg-Gresham, Hyun Min Kang, Dwight Stambolian, Emily Y Chew, Kari E Branham, John Heckenlively, Robert Fulton, Richard K Wilson, et al. Ancestry estimation and control of population stratification for sequence-based association studies. *Nature genetics*, 46(4):409, 2014.

Chaolong Wang, Xiaowei Zhan, Liming Liang, Gonçalo R Abecasis, and Xihong Lin. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *The American Journal of Human Genetics*, 96(6):926–937, 2015.

Hao Wang and Dit-Yan Yeung. Towards bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3395–3408, 2016.

Bruce S Weir and C Clark Cockerham. Estimating f-statistics for the analysis of population structure. *evolution*, 38(6):1358–1370, 1984.

Florian Wenzel, Kevin Roth, Bastiaan S Veeling, Jakub Świątkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.

Paul J Whalen, Lisa M Shin, Sean C McInerney, Håkan Fischer, Christopher I Wright, and Scott L Rauch. A functional mri study of human amygdala responses to facial expressions of fear versus anger. *Emotion*, 1(1):70, 2001.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020.

Alexandra Woolgar, Alice Parr, Rhodri Cusack, Russell Thompson, Ian Nimmo-Smith, Teresa Torralva, Maria Roca, Nagui Antoun, Facundo Manes, and John Duncan. Fluid intelligence loss linked to restricted regions of damage within frontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 107(33): 14899–14902, 2010.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

Youngwoo Bryan Yoon, Won-Gyo Shin, Tae Young Lee, Ji-Won Hur, Kang Ik K Cho, William Seunghyun Sohn, Seung-Goo Kim, Kwang-Hyuk Lee, and Jun Soo Kwon. Brain structural networks associated with intelligence and visuomotor ability. *Scientific reports*, 7(1):1–9, 2017.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

Chunshui Yu, Jun Li, Yong Liu, Wen Qin, Yonghui Li, Ni Shu, Tianzi Jiang, and Kuncheng Li. White matter tract integrity and intelligence in patients with mental retardation and healthy adults. *Neuroimage*, 40(4):1533–1541, 2008.

Shan Yu, Guannan Wang, Li Wang, and Lijian Yang. Multivariate spline estimation and inference for image-on-scalar regression. *Statistica Sinica*, 2020.

Yu Yue, Ji Meng Loh, and Martin A Lindquist. Adaptive spatial smoothing of fmri images. *Statistics and its Interface*, 3:3–13, 2010.

Theodore P Zanto and Adam Gazzaley. Fronto-parietal network: flexible hub of cognitive control. *Trends in cognitive sciences*, 17(12):602–603, 2013.

Ryad Zemouri, Noureddine Zerhouni, and Daniel Racoceanu. Deep learning in the biomedical applications: Recent and future status. *Applied Sciences*, 9(8):1526, 2019.

Xiaowei Zhan, David E Larson, Chaolong Wang, Daniel C Koboldt, Yuri V Sergeev, Robert S Fulton, Lucinda L Fulton, Catrina C Fronick, Kari E Branham, Jennifer Bragg-Gresham, et al. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. *Nature genetics*, 45(11):1375, 2013.

Xiang Zhang, Xiaocong Chen, Lina Yao, Chang Ge, and Manqing Dong. Deep neural network hyperparameter optimization with orthogonal array tuning. In *International Conference on Neural Information Processing*, pages 287–295. Springer, 2019.

Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

Shuheng Zhou. Thresholded lasso for high dimensional variable selection and statistical estimation. *arXiv preprint arXiv:1002.1583*, 2010.

Hongtu Zhu, Jianqing Fan, and Linglong Kong. Spatially varying coefficient model for neuroimaging data with jump discontinuities. *Journal of the American Statistical Association*, 109(507):1084–1098, 2014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2): 301–320, 2005.

James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature genetics*, 51(1): 12–18, 2019.