Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Orthogonalized Kernel Debiased Machine Learning for Multimodal Data Analysis

Xiaowu Dai & Lexin Li

To cite this article: Xiaowu Dai & Lexin Li (2023) Orthogonalized Kernel Debiased Machine Learning for Multimodal Data Analysis, Journal of the American Statistical Association, 118:543, 1796-1810, DOI: 10.1080/01621459.2021.2013851

To link to this article: https://doi.org/10.1080/01621459.2021.2013851

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 118, NO. 543, 1796–1810: Theory and Methods
https://doi.org/10.1080/01621459.2021.2013851

Orthogonalized Kernel Debiased Machine Learning for Multimodal Data Analysis

Xiaowu Dai[a,b] and Lexin Li[c]

[a]Department of Economics, the University of California, Berkeley, CA; [b]Department of Electrical Engineering and Computer Sciences, the University of California, Berkeley, CA; [c]Department of Biostatistics and Epidemiology, the University of California, Berkeley, CA

ABSTRACT

Multimodal imaging has transformed neuroscience research. While it presents unprecedented opportunities, it also imposes serious challenges. Particularly, it is difficult to combine the merits of the interpretability attributed to a simple association model with the flexibility achieved by a highly adaptive nonlinear model. In this article, we propose an orthogonalized kernel debiased machine learning approach, which is built upon the Neyman orthogonality and a form of decomposition orthogonality, for multimodal data analysis. We target the setting that naturally arises in almost all multimodal studies, where there is a primary modality of interest, plus additional auxiliary modalities. We establish the root-N-consistency and asymptotic normality of the estimated primary parameter, the semi-parametric estimation efficiency, and the asymptotic validity of the confidence band of the predicted primary modality effect. Our proposal enjoys, to a good extent, both model interpretability and model flexibility. It is also considerably different from the existing statistical methods for multimodal data integration, as well as the orthogonality-based methods for high-dimensional inferences. We demonstrate the efficacy of our method through both simulations and an application to a multimodal neuroimaging study of Alzheimer's disease. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received February 2021
Accepted November 2021

KEYWORDS
Basis expansion;
High-dimensional inference;
Multimodal data integration;
Neuroimaging analysis;
Neyman orthogonality;
Reproducing kernel Hilbert space

1. Introduction

Multimodal neuroimaging, where different types of images are acquired for a common set of experimental subjects, is becoming a norm in neuroscience research. It uses different physical and physiological sensitivities of imaging scanners and technologies, and measures distinct brain characteristics including brain structures, functions and chemical constituents. Multimodal neuroimaging analysis aggregates such diverse but often complementary information, consolidates knowledge across different modalities, and produces improved understanding of neurological development or disorders (Uludağ and Roebroeck 2014). Multimodal data also frequently arise in many other scientific applications, for example, integrative genomics (Richardson, Tseng, and Sun 2016), multimodal healthcare (Cai et al. 2019), and audio-visual speech recognition (Baltrusaitis, Ahuja, C., and Morency 2019).

Our motivation is a multimodal neuroimaging study of Alzheimer's disease (AD). AD is an irreversible neurodegenerative disorder and the leading form of dementia in elderly subjects. The most notable AD imaging biomarker is the brain grey matter cortical atrophy measured by structural magnetic resonance imaging (MRI). Meanwhile, amyloid-β and tau are two hallmark pathological proteins that are believed to be part of the driving mechanism of AD, and both can be measured by positron emission tomography (PET) using different nuclear tracers. The current model of AD pathogenesis hypothesizes a sequence of biological cascade among different AD biomarkers (Jack et al. 2010). It is of great scientific interest to study how they interact with each other and how they affect the cognitive outcome. These questions are crucial for our understanding of AD pathophysiology, and also have important therapeutic implications.

While multimodal neuroimaging presents unprecedented opportunities, it also imposes numerous serious challenges. First, neuroimaging data are typically high-dimensional and highly correlated, with measurements of brain characteristics at hundreds of brain regions and millions of brain voxel locations, and those measurements are often spatially or temporally correlated. Besides, the associations between different imaging modalities, and between images and phenotypic outcomes, are complicated. A linear association model, despite its wide usage, is hardly adequate to capture such complex associations. Second, it is particularly challenging to balance between model interpretability and model flexibility. Breiman (2001) contrasted two modeling cultures: the "data modeling culture," which adopts parametric models that are easier to interpret and to perform inference but much less flexible, versus the "algorithmic modeling culture," also known as machine learning, which involves complex and sometimes black-box type models that are highly flexible and nonlinear but difficult to interpret and infer. Both approaches have been frequently adopted in neuroimaging analysis. Nevertheless, it is difficult to combine the merits of both. Most existing works on multimodal data integration either assume a simple parametric model for easy interpretation (e.g., Sperling et al. 2019; Li and Li 2021), or consider a flexible nonlinear model but sacrifice the interpretability or inference capability (e.g., Hinrichs et al. 2011; Alam et al. 2018). Finally, rigorously quantifying statistical significance of

CONTACT Lexin Li *lexinli@berkeley.edu* Division of Biostatistics, University of California Berkeley, Berkeley, CA 94720.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

the primary parameter of interest remains a fundamental question in scientific inquiries. There have been a large number of highly successful nonlinear modeling techniques, ranging from the more classical splines, reproducing kernels, and random forests, to more recent deep neural network models. However, it is notoriously difficult to carry out statistical inference when utilizing those flexible methods. Moreover, when it comes to inference, naively adding multiple modalities together may suffer from serious biases and produce misleading results, as we show later.

In this article, we propose an orthogonalized kernel debiased machine learning approach, built upon the Neyman orthogonality (Neyman 1959, 1979), and a form of decomposition orthogonality (Wahba 1990, chap. 3), for multimodal data analysis. The principal setting we target is that there is a *primary* modality of interest, plus additional *auxiliary* modalities. Such a setting naturally arises in almost all multimodal studies, and is particularly useful from the perspective of scientific inquiries. For instance, in AD pathophysiology modeling (Jack et al. 2010), it is often of interest to quantify the effect of brain structural atrophy on cognition after accounting for amyloid-β and tau accumulations. In this case, the structural atrophy can be treated as the primary modality, while amyloid-beta and tau are the auxiliary modalities. In imaging genetics studies (Zhu et al. 2014; Nathoo et al. 2019), brain imaging features often play the role of intermediate phenotype between the genetic variants and clinical outcome. In this case, the brain image can be taken as the primary modality, and the genetic variants as the auxiliary modality. Under this setting, we employ a basis expansion type model along with model error to characterize the association between the primary modality and the outcome, and develop rigorous inference methods for the main parameter of interest as well as the predicted primary modality effect. Meanwhile, we employ highly flexible machine learning methods to model the complex associations both between the auxiliary modalities and the outcome, and between the primary and auxiliary modalities. A key challenge that comes with flexible machine learning modeling is that its associated regularization bias and overfitting would introduce heavy bias in the estimation of the main parameter of interest. To remove such an impact, we employ two types of orthogonality formulations based on Neyman (1959, 1979), Chernozhukov et al. (2018), and Wahba (1990). We establish the $\sqrt{N}$-consistency and asymptotic normality of the estimated main parameter, the semi-parametric estimation efficiency, as well as the asymptotic validity of the confidence band of the predicted primary modality effect, where N is the sample size. Our proposed framework thus enjoys, to a good extent, both model interpretability and model flexibility.

Our proposal is considerably different from the existing statistical methods for multimodal data integration. Particularly, there have been a class of unsupervised multimodal analysis built on matrix or tensor factorization (Lock et al. 2013), or canonical correlation analysis (Mai and Zhang 2019; Shu, Wang, and Zhu 2020). By contrast, we aim at a supervised regression problem. Under the regression setting with multimodal predictors, Li, Liu, and Chen (2019) proposed an integrative reduced-rank regression. Xue and Qu (2021) developed an estimating equations approach to accommodate block missing patterns. Li and Li (2021) developed a factor analysis-based linear regression model. These methods are supervised, but all of them still assume linear type associations, and none utilizes any nonlinear machine learning modeling.

Relatedly, the Neyman orthogonality has played an important role in both statistics and econometrics. Early works date back to Newey (1990), Robins and Rotnitzky (1995), and van der Laan and Rubin (2006). Meanwhile, it has received revived interest in high-dimensional statistical inference in recent years, thanks to, most notably, Chernozhukov et al. (2018); see also many references therein. Our proposal can be viewed as an extension of the double/debiased machine learning framework developed by Chernozhukov et al. (2018). However, there are some fundamental differences. First and most importantly, we allow an additional model error for the primary modality, which has crucial implications in terms of model interpretation, estimation and theoretical analysis. In particular, Chernozhukov et al. (2018) focused on a low-dimensional primary parameter involving no additional error. Kozbur (2020) extended to a nonparametric primary function through basis expansion, but required that the function can be well approximated with a vanishing approximation error. By contrast, we do not impose a vanishing error, which distinguishes our proposal from Chernozhukov et al. (2018) and Kozbur (2020) and other double/debiased machine learning methods. This additional model error essentially offers improved inferential robustness. Depending on the scientific context, one may choose a simple and interpretable yet less accurate model for the primary modality, or one may choose a more accurate but perhaps less interpretable model, and our method works for both cases. On the other hand, this error imposes numerous new challenges. To address those challenges, we introduce a second form of orthogonality, similar to the perpendicularity in smoothing splines (Wahba 1990), to ensure the parameter identifiability. We construct a new reproducing kernel Hilbert space (RKHS) and employ residual learning to decouple and remove the impact of the model error in parameter estimation. We also develop new theoretical tools to establish the asymptotic guarantees of the estimated primary parameter under model error. Second, we establish the confidence band for the nonparametric primary regression function given the high-dimensional nonlinear nuisance function. This quantity is of key scientific interest, as it quantifies the predicted effect and the contribution of the primary modality. However, its inference is challenging, due to the nonparametric nature of the model, high dimensionality, and strong correlations between the modalities. The existing literature on high-dimensional nonparametric inference usually requires stronger conditions that are unlikely to hold in multimodal neuroimaging data. We extend the framework of Chernozhukov, Chetverikov, and Kato (2014), and approximate the supremum of high-dimensional empirical processes by a Gaussian multiplier process to obtain the asymptotically valid confidence band. Later, we further compare with a number of alternative solutions, both analytically and numerically.

The rest of the article is organized as follows. We introduce the model framework in Section 2, and develop an estimation procedure in Section 3. We derive the orthogonal statistical inference procedure and the theoretical guarantees in Section 4.

We analytically compare with the alternative methods in Section 5. We present the simulations in Section 6, and revisit the multimodal AD study in Section 7. We conclude the article with a further discussion on the innovation of our method in Section 8, and relegate all technical proofs to the supplementary appendix.

2. Model

Suppose there are $M+1$ modalities of predictors. Let $X = (X_{(1)}, \ldots, X_{(p)})^{\mathsf{T}} \in \mathcal{X}^p$ denote the p-dimensional vector of the primary modality, where $\mathcal{X} \subset \mathbb{R}$ is a compact domain and X follows the distribution P in $\mathcal{X}^p$. Let $Z_{(m)} \in \mathbb{R}^{p'_m}$ denote the p'_m-dimensional vector of the mth auxiliary modality, $m = 1, \ldots, M$, and let $Z = (Z_{(1)}^{\mathsf{T}}, \ldots, Z_{(M)}^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{p'}$ collect all auxiliary modalities, $p' = p'_1 + \cdots + p'_M$. Let $Y \in \mathbb{R}$ denote the response variable. We propose the following model framework:

$$Y = f_0(X) + g_0(Z) + U, \tag{1}$$

where $U \in \mathbb{R}$ is the measurement error that is independent of (X, Z) and $\mathbb{E}[U] = 0$ and $\mathbb{E}[U^2] = \sigma^2 < \infty$, f_0 is the regression function capturing the effect of the primary modality on the response, and g_0 is the function capturing the collective effects of the auxiliary modalities. We also note that we can extend (1) from a linear model form to a generalized linear model form, so that it works for a binary or count type of response variable.

Next, assuming that $f_0 : \mathcal{X}^p \to \mathbb{R}$ resides in an RKHS (Wahba 1990), we decompose f_0 as follows:

$$f_0(x) = \eta(x, \theta_0) + \delta_0(x), \tag{2}$$

where $\eta(\cdot, \theta_0)$ is a parametric component that preserves the interpretability of $f_0(\cdot)$, and δ_0 is a nonparametric component that accounts for model error. Together, they form a nonparametric model for $f_0(x)$. Despite the wide use of a simple linear model for $f_0(\cdot)$ in the literature, there has been ample evidence showing that the linear model is inadequate to capture the complex association between X and Y (e.g., Alam et al. 2018). This has motivated us to consider a more flexible model for $\eta(\cdot, \theta_0)$, meanwhile taking into account the model error $\delta_0(\cdot)$ as in (2).

Next, we employ a basis expansion type model for $\eta(x, \theta_0)$, due to its ease of interpretation, relative flexibility, as well as computational efficiency (Huang, Zhang, and Zhou 2007; Wang et al. 2014; Ma et al. 2015). Specifically, let $\{\phi_1, \ldots, \phi_s\}$ denote a collection of orthonormal and centered basis functions in $\mathcal{X}$, satisfying that $\mathbb{E}[\phi_k(X_{(j)})] = 0, j = 1, \ldots, p, k = 1, \ldots, s$, where s is the number of basis functions. There is a rich library of basis functions, including polynomial basis, Fourier basis, B-splines, among others. Denote $\mathcal{B}_s(x_{(j)}) = \text{Span}\{1, \phi_1(x_{(j)}), \ldots, \phi_s(x_{(j)})\}$ as the space spanned by these basis functions. Let the parametric component $\eta(x, \theta_0)$ be the projection of f_0 onto the space spanned by the tensor product of the basis functions, that is,

$$\eta(x, \theta_0) = \underset{f \in \otimes_{j=1}^p \mathcal{B}_s(x_{(j)})}{\arg\min} \int_{\mathcal{X}^p} [f(x) - f_0(x)]^2 dP(x) = \Phi(x)^{\mathsf{T}} \theta_0, \tag{3}$$

where $x = (x_{(1)}, \ldots, x_{(p)})^{\mathsf{T}} \in \mathcal{X}^p$, the basis vector $\Phi(x) = [1, \phi_1(x_{(1)}), \ldots, \phi_s(x_{(1)}), \ldots, \phi_1(x_{(p)}), \ldots, \phi_s(x_{(p)}), \ldots, \phi_1(x_{(1)}) \cdots \phi_p(x_{(p)})]^{\mathsf{T}} \in \mathbb{R}^d$, and $d = (s+1)^p$. Model (3) is a general model that includes main effects $\phi_i(x_{(j)})$, $i = 1, \ldots, s, j = 1, \ldots, p$, pairwise interactions $\phi_{i_1}(x_{(j_1)})\phi_{i_2}(x_{(j_2)})$, $i_1, i_2 = 1, \ldots, s, j_1, j_2 = 1, \ldots, p$, as well as higher-order interactions. It includes additive model (Hastie and Tibshirani 1990), linear model, and functional ANOVA model (Lin and Zhang 2006) as special cases. That is, when $\eta(x, \theta_0)$ is the projection of f_0 onto the space $\oplus_{j=1}^p \mathcal{B}_s(x_{(j)})$ spanned by the sum of the basis, then (3) is essentially an additive model. When $s = 1$ and $\phi_s(\cdot)$ is a centered linear basis function, (3) becomes a linear model. When $\eta(x, \theta_0)$ is the projection of f_0 onto the space spanned by the tensor product of the basis with pairwise or higher-order interactions, (3) becomes a functional ANOVA model.

Finally, we characterize the association between the primary modality X and the auxiliary modalities Z as,

$$\Phi(X) = r_0(Z) + V, \quad \mathbb{E}[V|Z] = 0, \tag{4}$$

where $V \in \mathbb{R}^d$ accounts for the part of the variation in $\Phi(X)$ that cannot be explained by Z, and r_0 captures the complicated association between Z and $\Phi(X)$.

Suppose the observed data $\{(X_i, Z_i, Y_i) : i = 1, \ldots, N\}$ are independent copies of (X, Z, Y) and satisfy the system of models (1) to (4). Our main goal is the statistical inference of θ_0, which reflects the interpretable effect of the primary modality X on the outcome Y, and of f_0, which reflects the predicted effect of the primary modality, and is also directly related to some causal effect and the quantification of the contribution of X. Meanwhile, we view $\{g_0, \delta_0, r_0\}$ as nuisance functions, and propose to use highly flexible machine learning methods, for example, random forests, reproducing kernels, or neural networks, to model them. The machine learning methods often use regularization to avoid overfitting, especially when X and Z are high-dimensional and highly nonlinear. However, regularization would introduce sizable bias, and would invalidate the subsequent inference on θ_0 and f_0. Actually, the naive estimator of θ_0 by simply plugging in the machine learning estimators of $\{g_0, \delta_0, r_0\}$ would fail to be $\sqrt{N}$-consistent; see Section 5. This has motivated us to develop an orthogonal statistical inference framework to correct the bias introduced by the flexible estimators of $\{g_0, \delta_0, r_0\}$, and to perform a valid inference for θ_0 and f_0.

3. Orthogonalized Kernel Debiased Machine Learning

We consider two orthogonality formulations that are essential for the construction of our estimator. We then present our estimation algorithm built on those orthogonal formulations.

3.1. Orthogonality

The first is the Neyman orthogonality (Neyman 1959, 1979; Chernozhukov et al. 2018), which allows the estimation of θ_0 to be locally insensitive to the values of nuisance functions, and thus one can plug in noisy estimates of the nuisance functions for the inference of θ_0. We consider the target parameter $\theta \in \mathbb{R}^d$, and the nuisance functions $r \in \mathcal{H}_r, g \in \mathcal{H}_g, \delta \in \mathcal{H}_\delta$, where $\mathcal{H}_r$ and $\mathcal{H}_g$ are functional spaces of finite mean squared functions, and $\mathcal{H}_\delta$ is an RKHS.

Definition 1 (Neyman orthogonality). A score function $\psi(\theta, r, g, \delta)$ is said to satisfy the Neyman orthogonality (Neyman 1959, 1979; Chernozhukov et al. 2018) if (i) The mean $\mathbb{E}[\psi(\theta_0, r_0, g_0, \delta_0)] = 0$ at $(\theta_0, r_0, g_0, \delta_0)$; (ii) The pathwise derivative map, $\partial_r\{\mathbb{E}[\psi(\theta_0, r_0 + t(r - r_0), g_0 + t(g - g_0), \delta_0 + t(\delta - \delta_0))]\}$, exists for all $t \in [0, 1)$, where r, g and δ lie in a neighborhood of $r_0 \in \mathcal{H}_r$, $g_0 \in \mathcal{H}_g$ and $\delta_0 \in \mathcal{H}_\delta$, respectively; (iii) The pathwise derivative vanishes at $t = 0$, in that $\partial_t\{\mathbb{E}[\psi(\theta_0, r_0+t(r-r_0), g_0+t(g-g_0), \delta_0+t(\delta-\delta_0))]\}|_{t=0} = 0$.

Proposition 1. Define the score function,

$$\psi(W; \theta, r, g, \delta) = [Y - \Phi(X)^\mathsf{T}\theta - g(Z) - \delta(X)][r(Z) - \Phi(X)],$$

where $W = (X, Y, Z)$. Then under the system of models (1) to (4), the score $\psi(W; \theta, m, \delta, g)$ is Neyman orthogonal at $(\theta_0, r_0, g_0, \delta_0)$.

We briefly comment that a similar idea to Neyman orthogonality is also used in targeted maximum likelihood estimation (van der Laan and Rubin 2006; Zheng and van der Laan 2011), which constructs an estimation equation for a target parameter and requires the score function to be in the orthogonal complement of the tangent space of the nuisance parameter.

In addition to the Neyman orthogonality, we also require the functions Φ and δ_0 in models (2) and (3) to satisfy a decomposition orthogonality, which is necessary for the identifiability of θ_0.

Definition 2 (Decomposition orthogonality). Suppose that $\Phi(\cdot)$ is bounded on $\mathcal{X}^p$. The functions Φ and δ_0 are said to satisfy the decomposition orthogonality if $\mathbb{E}_X[\Phi(X)\delta_0(X)] = 0$.

Proposition 2. Under models (2) and (3), θ_0 is identifiable only if Φ and δ_0 satisfy the decomposition orthogonality. Moreover, for any reproducing kernel $K(\cdot, \cdot)$ on $\mathcal{X}^p \times \mathcal{X}^p$, define

$$\begin{aligned} K_\delta(x, x') = K(x, x') &- \mathbb{E}_X[\Phi(X)^\mathsf{T}K(x, X)] \\ &\times \left(\mathbb{E}_X\{\mathbb{E}_{X'}[\Phi(X')K(X', X)]\Phi(X)^\mathsf{T}\}\right)^{-1} \\ &\mathbb{E}_{X'}[\Phi(X')K(x', X')], \end{aligned}$$

where X and X' are iid copies of the primary modality. Then $K_\delta(\cdot, \cdot) : \mathcal{X}^p \times \mathcal{X}^p \to \mathbb{R}$ is positive definite. Besides, for any $\widehat{\delta}(x) = \sum_{i=1}^m c_i K_\delta(x, x_i)$, with $c_i \in \mathbb{R}, x_i \in \mathcal{X}^p$ and $m \geq 1$, $\Phi(X)$ and $\widehat{\delta}(X)$ satisfy the decomposition orthogonality.

The decomposition orthogonality in Definition 2 is similar to the perpendicularity requirement in the smoothing splines literature (see, e.g., Wahba 1990, chap. 3), where the null space and the RKHS need to be perpendicular under certain norms in order to find a consistent estimator as the sample size diverges, while we use an ℓ_2-norm with respect to the distribution of X. Hereinafter, let $\mathcal{H}_\delta$ be the corresponding RKHS of the kernel $K_\delta(\cdot, \cdot)$. By the representer theorem (Wahba 1990), the M-estimator in RKHS $\mathcal{H}_\delta$ can be found in a finite-dimensional subspace of $\mathcal{H}_\delta$, that is, it can be written as $\widehat{\delta}(x) = \sum_{i=1}^m c_i K_\delta(x, x_i)$, with $c_i \in \mathbb{R}, x_i \in \mathcal{X}^p$ and $m \geq 1$. Proposition 2 shows that $\widehat{\delta}(X)$ and $\Phi(X)$ satisfy the decomposition orthogonality, which in turn ensures the identifiability of the primary parameter θ_0 we target.

Algorithm 1 Orthogonalized kernel debiased machine learning algorithm

1: Obtain the initial estimators $\widehat{\theta}^{(0)}, \widehat{g}^{(0)}, \widehat{\delta}^{(0)}$ by (5) using all the data.
2: Split the data randomly into Q non-overlapping chunks of equal size. For $q \in [Q]$, denote I_q as the corresponding set of data indices of the qth chunk, and $I_q^c = [N]\backslash I_q$.
3: **for** $q = 1$ to Q **do**
4: Obtain the estimator $\widehat{r}_0$ by (6) using the data in I_q^c.
5: **end for**
6: **repeat**
7: **for** $q = 1$ to Q **do**
8: Obtain the iterative estimators $\{\widehat{g}_q^{(t)}, \widehat{\delta}_q^{(t)}\}$ by (7) using the data in I_q^c.
9: Obtain the iterative estimator $\widetilde{\theta}_q^{(t)}$ by (8) using the data in I_q.
10: **end for**
11: Obtain the iterative estimator $\widehat{\theta}^{(t)}$ by (9).
12: **until** the stopping criterion is met.
13: Construct the final estimator $\widehat{\theta} \in \mathbb{R}^d$ by (10) using cross-fitting.

3.2. Iterative Cross-Fitting Procedure

We next present an estimation algorithm of θ_0 based on the orthogonality formulations in Propositions 1 and 2. The algorithm consists of five main steps. In the first step, we obtain the initial estimators of $\{\theta_0, g_0, \delta_0\}$. In the second step, we split the data into Q disjoint chunks. In the third step, we estimate r_0, and in the fourth step, we iteratively update the estimates of $\{g_0, \delta_0\}$ and θ_0. In these two steps, we obtain the estimates by leaving out some chunk of data in turn. In the fifth step, we construct the final estimator of θ_0, by first using only one chunk of data at a time, then averaging over all Q chunks. When estimating the nuisance functions $\{r_0, g_0, \delta_0\}$, we employ some penalized learning methods, where we denote $\text{PEN}_{\mathcal{H}_r}(r)$, $\text{PEN}_{\mathcal{H}_g}(g)$, $\text{PEN}_{\mathcal{H}_\delta}(\delta)$ as the penalty functionals in the candidate functional spaces $\mathcal{H}_r$, $\mathcal{H}_g$, and $\mathcal{H}_\delta$, respectively. Here, $\mathcal{H}_\delta$ is chosen to be the corresponding RKHS of $K_\delta(\cdot, \cdot)$ in Proposition 2, and $\text{PEN}_{\mathcal{H}_\delta}(\delta)$ is the penalty based on the squared RKHS-norm in $\mathcal{H}_\delta$. The choices of $\{\mathcal{H}_r, \mathcal{H}_g\}$ as well as the penalty functions depend on specific data applications, and the tuning follows the usual tuning procedures in penalized learning. We first summarize the procedure in Algorithm 1, then detail the main steps.

In the first step, we obtain the initial estimators of $\{\theta_0, g_0, \delta_0\}$ as follows:

$$\begin{aligned} \widehat{\theta}^{(0)} &= \arg\min_{\theta \in \mathbb{R}^d} \left\{ \frac{1}{N} \sum_{i=1}^N \left[Y_i - \Phi(X_i)^\mathsf{T}\theta - \widehat{g}^{(0)}(Z) \right]^2 \right\}, \\ \widehat{g}^{(0)} &= \arg\min_{g \in \mathcal{H}_g} \left\{ \frac{1}{N} \sum_{i=1}^N \left[Y_i - g(Z_i) \right]^2 + \lambda_N^g \text{PEN}_{\mathcal{H}_g}(g) \right\}, \end{aligned} \tag{5}$$

and $\widehat{\delta}^{(0)} = 0$. Here, $\lambda_N^g \geq 0$ is a tuning parameter, and we use all the N data samples.

In the second step, we randomly split the sample observations into $Q \geq 2$ nonoverlapping chunks of equal size $n = N/Q$. For notational simplicity, we assume N is divisible by Q. For each

$q \in [Q] = \{1, \ldots, Q\}$, we denote I_q as the set of indices in $[N] = \{1, \ldots, N\}$ corresponding to the data in the qth chunk, and denote $I_q^c = [N] \backslash I_q$ as the indices of the complementary data.

In the third step, we estimate the function r_0 by,

$$\widehat{r}_q = \underset{r \in \mathcal{H}_r}{\arg\min} \left\{ \frac{1}{n} \sum_{i \in I_q^c} [\Phi(X_i) - r(Z_i)]^2 + \lambda_n^r \mathrm{PEN}_{\mathcal{H}_r}(r) \right\}, \quad (6)$$

where $\lambda_n^r \geq 0$ is a tuning parameter. Note that we only use the data from I_q^c in (6). Besides, we estimate r_0 only once, without any iterations, for each $q \in [Q]$.

In the fourth step, we iteratively update the estimates of $\{g_0, \delta_0\}$ and θ_0. That is,

$$\left\{\widehat{g}_q^{(t)}, \widehat{\delta}_q^{(t)}\right\} = \underset{g \in \mathcal{H}_g, \delta \in \mathcal{H}_\delta}{\arg\min} \left\{ \frac{1}{n} \sum_{i \in I_q^c} \left[Y_i - \Phi(X_i)\widehat{\theta}^{(t-1)} - \delta(X_i) - g(Z_i) \right]^2 + \lambda_n^g \mathrm{PEN}_{\mathcal{H}_g}(g) + \lambda_n^\delta \mathrm{PEN}_{\mathcal{H}_\delta}(\delta) \right\}, \quad (7)$$

$$\widetilde{\theta}_q^{(t)} = \left\{ \frac{1}{n} \sum_{i \in I_q} \left[\Phi(X_i) - \widehat{r}_q(Z_i) \right] \Phi(X_i)^\mathsf{T} \right\}^{-1} \times \frac{1}{n} \sum_{i \in I_q} \left[\Phi(X_i) - \widehat{r}_q(Z_i) \right] \left[Y_i - \widehat{g}_q^{(t)}(Z_i) - \widehat{\delta}_q^{(t)}(X_i) \right], \quad (8)$$

$$\widehat{\theta}^{(t)} = \frac{1}{Q} \sum_{q=1}^{Q} \widetilde{\theta}_q^{(t)}, \quad (9)$$

where $\lambda_n^g, \lambda_n^\delta \geq 0$ are the tuning parameters. The estimation in (7) employs residual learning, since it is based on the residual $[Y - \Phi(X)\widehat{\theta}_q^{(t-1)}]$. The resulting estimator $\widehat{\delta}_q^{(t)}$ satisfies the decomposition orthogonality relative to Φ in Proposition 2. Besides, it involves only the complementary data in I_q^c. The estimation in (8) employs the Neyman orthogonality formulation in Proposition 1, and involves only the data in I_q. The estimation in (9) averages $\widetilde{\theta}_q^{(t)}$ from (8) across all $q = 1, \ldots Q$. Moreover, (8) and (9) together use the idea of centralized training with decentralized execution (Lowe et al. 2017), which greatly facilitates the convergence of the algorithm. We stop the iterations when some stopping criterion is met, for example, when the difference between two consecutive estimates of θ_0 is smaller than a threshold value. We also remark that, this step is essentially a Gauss-Seidel iterative algorithm that has been widely used in statistics (Buja, Hastie, and Tibshirani 1989). In our simulations, we find the algorithm converges fast, usually after only 3 to 5 iterations. We denote the final estimators for $\{g_0, \delta_0\}$ as $\{\widehat{g}_q, \widehat{\delta}_q\}, q \in [Q]$.

In the final step, we construct our orthogonal estimator for θ_0 using cross-fitting,

$$\widehat{\theta} = \left\{ \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{n} \sum_{i \in I_q} \left[\Phi(X_i) - \widehat{r}_q(Z_i) \right] \Phi(X_i)^\mathsf{T} \right\}^{-1} \times \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{n} \sum_{i \in I_q} \left[\Phi(X_i) - \widehat{r}_q(Z_i) \right] \left[Y_i - \widehat{g}_q(Z_i) - \widehat{\delta}_q(X_i) \right]. \quad (10)$$

That is, for each $q \in [Q]$, we use the chunk of data that is left out when estimating $\{r_0, g_0, \delta_0\}$ earlier, then average over all Q chunks. Cross-fitting has been commonly used in high-dimensional inferences in recent years; see, for example, Chernozhukov et al. (2018); Newey and Robins (2018). By swapping the roles of each chunk and the complementary chunks Q times, it ensures good statistical properties while regaining the efficiency of making use of all available data observations. Later, we show the estimator $\widehat{\theta}$ in (10) is actually semi-parametric efficient.

4. Statistical Inference

We aim at two key inference questions: inference for the primary parameter of interest θ_0, and inference for the primary regression function $f_0(\cdot)$. Both are crucial for scientific inquires. The former directly quantifies the relevance of the variables of the primary modality to the outcome. The latter captures the predicted effect and the contribution of the primary modality, and also has some causal interpretation under additional conditions.

4.1. Inference of the Primary Parameter θ_0

We begin with the study of the asymptotic behavior of the estimator $\widehat{\theta}$ in (10) as the sample size N tends to infinity. We establish the $\sqrt{N}$-convergence that $\|\widehat{\theta} - \theta_0\|_{\ell_2} = O_p(N^{-1/2})$, as well as the asymptotic normality that $\sqrt{N}(\widehat{\theta} - \theta_0)$ approaches a normal distribution. We note that this $\sqrt{N}$-convergence result is highly nontrivial, because the estimator $\widehat{\theta}$ in (10) involves the nuisance estimators $\{\widehat{r}_q, \widehat{g}_q, \widehat{\delta}_q\}$. When $\{r_0, g_0, \delta_0\}$ are estimated nonparametrically, the convergence rates of the estimators $\{\widehat{r}_q, \widehat{g}_q, \widehat{\delta}_q\}$ are generally slower than $O_p(N^{-1/2})$ (van der Vaart 1998). Later in Section 5, we show that many popular alternative methods cannot achieve the $\sqrt{N}$-consistency.

We first present a set of regularity conditions.

(C1) The basis vector $\Phi(\cdot)$ in (3) satisfies that $\mathbb{E}[\|\Phi(X)\|_{\ell_2}^2] < \infty$.

(C2) The error term $V \in \mathbb{R}^d$ in (4) satisfies that $\mathbb{E}(VV^\mathsf{T})$ is invertible and $\mathbb{E}(V^\mathsf{T}V) < \infty$.

(C3) The estimators $\widehat{r}_q$ as constructed in (6), and $\{\widehat{g}_q, \widehat{\delta}_q\}$ as constructed in (7) at the algorithmic convergence satisfy that $\mathbb{E}[\|\widehat{r}_q(Z) - r_0(Z)\|_{\ell_2}^2] = o(N^{-1/2})$, $\mathbb{E}\{[\widehat{g}_q(Z) - g_0(Z)]^2\} = o(N^{-1/2})$, and $\mathbb{E}\{[\widehat{\delta}_q(X) - \delta_0(X)]^2\} = o(N^{-1/2})$, for $q \in [Q]$ and Q is finite.

Condition (C1) is mild and holds for most practical choices of the basis functions. For example, (C1) holds with the continuous basis over the compact domain $\mathcal{X}^p$. Condition (C2) is a fairly standard regularity condition, and is needed for the asymptotic normality of parameter estimation in moment-based problems (Chernozhukov et al. 2018). Condition (C3) is different from requiring the estimators $\{\widehat{r}_q, \widehat{g}_q, \widehat{\delta}_q\}$ to be $\sqrt{N}$-consistent, which is difficult to satisfy for many nonparametric estimators. Instead, (C3) holds for a wide range of popular machine learning methods; for instance, it holds for the ℓ_1-penalized linear regression in a variety of sparse models (Bickel, Ritov, and Tsybakov 2009; Bühlmann and van de Geer 2011), a class of random forests (Biau 2012), a class of neural networks (Chen and White 1999), and numerous kernel methods in RKHS (Wahba 1990; van der Vaart 1998), among others. Moreover, we note that

(C3) is generally less restrictive than the Donsker conditions, which are commonly assumed in semi-parametric statistical analysis (Kosorok 2007). The Donsker conditions require the functional spaces $\{\mathcal{H}_r, \mathcal{H}_g, \mathcal{H}_\delta\}$ to have a bounded complexity, or more specifically, a bounded entropy integral. However, for multimodal data analysis where the dimension of the auxiliary modalities Z increases with the sample size, such a requirement fails even in the linear model setting with the parameter space specified by the Euclidean ball of unit radius (Raskutti, Wainwright, and Yu 2011). By contrast, (C3) holds in this example.

Under (C1)–(C3), we obtain the main theoretical result for our estimator $\widehat{\theta}$.

Theorem 1. Suppose the system of models (1) to (4), and the regularity conditions (C1) to (C3) hold. The orthogonalized kernel debiased machine learning estimator $\widehat{\theta}$ in (10) satisfies that,

$$\widehat{\theta} - \theta_0 = [\mathbb{E}(VV^{\mathsf{T}})]^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} V_i U_i \right) + o_p(N^{-1/2}).$$

where $\{(U_i, V_i) : i = 1, \ldots, N\}$ are independent copies of the error terms (U, V) in (1) and (4).

The proof of this theorem is given in Appendix S3. We make two remarks. First, a direct implication of Theorem 1 is the asymptotic normality of $\widehat{\theta}$, that is,

$$\sqrt{N}(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 [\mathbb{E}(VV^{\mathsf{T}})]^{-1}\right). \tag{11}$$

Second, the asymptotic normality in (11) further implies that we can construct the confidence interval for the primary parameter of interest θ_0 as,

$$\mathrm{CI}(\theta_0) = \widehat{\theta} \pm F_{\mathcal{N}}^{-1}(1 - \alpha/2) \sqrt{\sigma^2 (\mathbb{E}[VV^{\mathsf{T}}])^{-1}/N},$$

where $F_{\mathcal{N}}(\cdot)$ denotes the cumulative distribution function of the standard normal distribution. When the variance term $\sigma^2 \mathbb{E}[VV^{\mathsf{T}}]$ in (11) is unknown, we use a plug-in estimator,

$$\widehat{\Sigma}(\widehat{\theta}) = \widehat{J}^{-1} \Bigg\{ \frac{1}{nQ} \sum_{q=1}^{Q} \sum_{i \in I_q} \left[Y_i - \Phi(X_i)^{\mathsf{T}} \widehat{\theta} - \widehat{g}_q(Z_i) - \widehat{\delta}_q(X_i) \right]^2 [\widehat{r}_q(Z_i) - \Phi(X_i)][\widehat{r}_q(Z_i) - \Phi(X_i)]^{\mathsf{T}} \Bigg\} \widehat{J}^{-1},$$

where $\widehat{J} = (nQ)^{-1} \sum_{q=1}^{Q} \sum_{i \in I_q} [\Phi(X_i) - \widehat{r}_q(Z_i)] \Phi(X_i)^{\mathsf{T}}$. The next corollary shows that this plug-in estimator is consistent, and its proof is given in Appendix S4.

Corollary 1. Suppose the conditions of Theorem 1 hold. If U in (1) and the elements of V in (4) have bounded fourth moment, then the plug-in estimator $\widehat{\Sigma}(\widehat{\theta})$ is consistent, in that

$$\widehat{\Sigma}(\widehat{\theta}) \xrightarrow{p} \sigma^2 \left(\mathbb{E}[VV^{\mathsf{T}}] \right)^{-1}.$$

Next, we discuss the efficiency of the estimator $\widehat{\theta}$. We first note that the estimation problem for θ_0 under the system of models (1) to (4) is semi-parametric. This is because the parameter of interest $\theta_0 \in \mathbb{R}^d$ is finite-dimensional as specified in (3), while the parameter space of models (1) and (2) contains high-dimensional, or infinite-dimensional functional spaces as $\{g_0, \delta_0\} \in \mathcal{H}_g \otimes \mathcal{H}_\delta$. We also allow the dimensions of g_0 and δ_0 to grow with the sample size N. The next theorem shows that $\widehat{\theta}$ in (10) is semiparametric efficient (Kosorok 2007), in that it achieves the highest possible efficiency, if the measurement error U follows a normal distribution. The proof of this theorem is given in Appendix S5, along with a brief review of the background on semiparametric estimation efficiency.

Theorem 2. Suppose the conditions of Theorem 1 hold. If the measurement error U in (1) follows a normal distribution, then the estimator $\widehat{\theta}$ in (10) is semiparametric efficient.

4.2. Inference of the Primary Function f_0

We next consider inference of the primary regression function $f_0(\cdot)$, which is of particular interest for several reasons. First of all, it quantifies the predicted effect of the primary modality X on the outcome Y. In addition, it also captures the amount of contribution of the primary modality, in terms of the percentage of variation explained, given all other modalities in the model. Finally, under some additional assumptions, f_0 is directly related to the notions of the partial dependence of Y on X, as well as the total effect of X on Y in a causal inference sense.

Given the orthogonal estimator $\widehat{\theta}$ in (10), a natural estimator for f_0 is $\widehat{f}(x) = \Phi(x)^{\mathsf{T}} \widehat{\theta}$. We seek the confidence band for f_0. A confidence band $\mathcal{C}_N$ is a set of confidence intervals, $\mathcal{C}_N = \left\{ \mathcal{C}_N(x) = [c_L(x), c_U(x)] \;\middle|\; x \in \mathcal{X}^p \right\}$. Consider the empirical process $\sup_{x \in \mathcal{X}^p} \sqrt{N}[\widehat{f}(x) - f_0(x)]$, whose distribution can be approximated by a Gaussian multiplier process,

$$\widehat{\mathbb{H}}_N(x) = \sqrt{N} \Phi(x)^{\mathsf{T}} \left\{ \frac{1}{nQ} \sum_{q=1}^{Q} \sum_{i \in I_q} [\Phi(X_i) - \widehat{r}_q(Z_i)] \Phi(X_i)^{\mathsf{T}} \right\}^{-1} \frac{1}{nQ} \sum_{q=1}^{Q} \sum_{i \in I_q} [\Phi(X_i) - \widehat{r}_q(Z_i)] \widehat{\sigma}(\widehat{\theta}) \xi_i,$$

where the estimator $\widehat{\sigma}^2(\widehat{\theta}) = (nQ)^{-1} \sum_{q=1}^{Q} \sum_{i \in I_q} [Y_i - \Phi(X_i)^{\mathsf{T}} \widehat{\theta} - \widehat{g}_q(Z_i) - \widehat{\delta}_q(X_i)]^2$, and $\xi = (\xi_1, \ldots, \xi_N)^{\mathsf{T}} \in \mathbb{R}^N$ are independent $\mathcal{N}(0, 1)$ random variables. Let $\widehat{c}_N(\alpha/2)$ be the $(1 - \alpha/2)$th quantile of $\sup_{x \in \mathcal{X}^p} \widehat{\mathbb{H}}_N(x)$. We construct the $100 \times (1 - \alpha)\%$ confidence band for f_0 as,

$$\mathcal{C}_N = \left\{ \mathcal{C}_N(x) = \left[\widehat{f}(x) - \frac{\widehat{c}_N(\alpha/2)}{\sqrt{N}}, \widehat{f}(x) + \frac{\widehat{c}_N(\alpha/2)}{\sqrt{N}} \right] \;\middle|\; x \in \mathcal{X}^p \right\}. \tag{12}$$

To establish the asymptotic validity of (12), we first present a modified version of the regularity condition (C3), and an additional condition regarding the function f_0.

(C3′) The estimators $\widehat{r}_q$ as constructed in (6), and $\{\widehat{g}_q, \widehat{\delta}_q\}$ as constructed in (7) at the algorithmic convergence satisfy that $\mathbb{E}[\|\widehat{r}_q(Z) - r_0(Z)\|_{\ell_2}^2] = O(N^{-1/2-c_r})$, $\mathbb{E}[(\widehat{g}_q(Z) - g_0(Z))^2] = O(N^{-1/2-c_g})$, and $\mathbb{E}[(\widehat{\delta}_q(X) - \delta_0(X))^2] = O(N^{-1/2-c_\delta})$, for some constants $c_r, c_g, c_\delta \in (0, 1/2]$, $q \in [Q]$, and Q is finite.

(C4) The function $f_0 : \mathcal{X}^p \to \mathbb{R}$ resides in the kth-order Sobolev space, $k > p$, in that f_0 and the derivatives $f_0^{(v)}$

are absolutely continuous for any vector of nonnegative integers $\nu \in \mathbb{N}_0^p$ with $\|\nu\|_{\ell_1} \le k-1$, and $\mathbb{E}\{[f_0^{(\nu)}(X)]^2\} < \infty$ for any $\nu \in \mathbb{N}_0^p$ with $\|\nu\|_{\ell_1} = k$.

Condition (C3′) is slightly stronger than (C3), which is necessary to obtain the asymptotic validity of the confidence band $\mathcal{C}_N$ in (12). Nevertheless, (C3′) continues to hold for a wide range of commonly used machine learning methods, including all the aforementioned ones where (C3) holds. Condition (C4) is a standard regularity condition in the literature on nonparametric estimations (Wahba 1990; van der Vaart 1998).

The next theorem shows that the confidence band $\mathcal{C}_N$ in (12) is asymptotically valid, in the sense that the coverage holds uniformly for all $x \in \mathcal{X}^p$ under a fixed f_0,

$$\liminf_{N\to\infty} \mathbb{P}\big[f_0(x) \in \mathcal{C}_N(x), \text{ for all } x \in \mathcal{X}^p\big] \ge 1-\alpha.$$

Theorem 3. Suppose the system of models (1) to (4), and the regularity conditions (C1), (C2), (C3′), and (C4) hold. Let s be the number of bases for each function component in (3), and $c_{\min} = \min\{c_r, c_g, c_\delta\} > 0$. Suppose the measurement error U in (1) follows a normal distribution, and the number of basis functions $s = \lceil N^{(1+2c)/2k} \rceil$ for a constant $c \in \big(0, (k-p)/2(k+p)\big]$. Then, there exist a constant $C > 0$, such that the coverage of the confidence band $\mathcal{C}_N$ in (12) satisfies,

$$\mathbb{P}\big[f_0(x) \in \mathcal{C}_N(x), \quad \text{for all} \quad x \in \mathcal{X}^p\big] \ge 1-\alpha-CN^{-c}, \quad \text{for any} \quad 0<\alpha<1.$$

Consequently, the confidence band $\mathcal{C}_N$ in (12) is asymptotically valid.

The proof of this theorem is given in Appendix S6, and is built upon the framework of using the Gaussian multiplier process to approximate the distribution of the supremum of empirical processes (Chernozhukov, Chetverikov, and Kato 2014). We first note that, for the inference of f_0, we require the number of basis functions s to diverge with the sample size, but for the inference of θ_0, we do *not* require a diverging s. When s diverges, the error term $V \in \mathbb{R}^{(s+1)^p}$ in (4) has a diverging dimension too. Nevertheless, Theorem 3 continues to hold. We next compare Theorem 3 with Lu, Kolar, and Liu (2020) and Kozbur (2020). Lu, Kolar, and Liu (2020) studied the inference of nonparametric additive models, but required there only exists a weak dependency between the covariates, for example, between X and Z, in that the difference between the joint distribution and the product of marginal distributions is small under a certain norm. Multimodal data, however, are typically highly correlated (Uludağ and Roebroeck 2014), and as such, the requirement of Lu, Kolar, and Liu (2020) may not always hold. By contrast, we allow a strong dependency between X and Z, and employ (4) to model potentially complex dependency between X and Z. Kozbur (2020) considered a nonparametric primary function f_0 through basis expansion, but required the approximation error to vanish at a rate faster than $\sqrt{N}$, which can be rather restrictive. By contrast, we do not require a vanishing approximation error for our method. This has a crucial implication, because it essentially allows one to use a simple and interpretable model to characterize the parametric component of f_0, for example, a linear model, which itself can be inaccurate and may induce a nonnegligible approximation error. Finally, we briefly comment that, to establish an honest confidence band with a uniform coverage for all $f_0 \in \mathcal{H}_f$ and data-generating functions, one needs to fully characterize $\mathcal{H}_f$ and to extend the classical Smirnov-Bickel-Rosenblatt condition (Giné and Nickl 2009) to the multimodal setting. We leave a full investigation as future research.

In addition to the predicted effect, the function f_0 also captures the amount of contribution of the primary modality given other modalities. Recall that in the classical linear regression model, the coefficient of determination R^2 measures the percentage of total variation in the response that has been explained by the predictors. We next show that f_0 is directly related to R^2, then derive the confidence interval for the R^2 measure. Consider the population version of R^2,

$$R^2 = 1 - \frac{\mathbb{E}(\mathrm{RSS})}{\mathbb{E}(\mathrm{TSS})}, \quad \text{where} \quad \mathbb{E}(\mathrm{RSS}) = \mathbb{E}\left[\{Y - f_0(X)\}^2\right], \quad \mathbb{E}(\mathrm{TSS}) = \mathbb{E}\left[(Y-\bar{Y})^2\right], \tag{13}$$

$\bar{Y} = N^{-1}\sum_{i=1}^N Y_i$, and RSS and TSS denote the residual sum of squares and total sum of squares, respectively. Define $\widehat{f}_{(1)}(x) = \widehat{f}(x) - N^{-1/2}\widehat{c}_N(\alpha/2)$, and $\widehat{f}_{(2)}(x) = \widehat{f}(x) + N^{-1/2}\widehat{c}_N(\alpha/2)$. Then denote $R^2_{(1)} = 1 - \sum_{i=1}^N [Y_i - \widehat{f}_{(1)}(X_i)]^2 / \sum_{i=1}^N (Y_i - \bar{Y})^2$, and $R^2_{(2)} = 1 - \sum_{i=1}^N [Y_i - \widehat{f}_{(2)}(X_i)]^2 / \sum_{i=1}^N (Y_i - \bar{Y})^2$. We construct the $100 \times (1-\alpha)\%$ confidence interval for R^2 as follows:

$$\mathrm{CI}(R^2) = \Big(\min(R^2_{(1)}, R^2_{(2)}),\ \max(R^2_{(1)}, R^2_{(2)})\Big).$$

The next corollary, following directly from Theorem 3, shows this is a valid confidence interval.

Corollary 2. Suppose the conditions of Theorem 3 hold. The confidence interval $\mathrm{CI}(R^2)$ is valid, in that $\liminf_{N\to\infty} \mathbb{P}\left[R^2 \in \mathrm{CI}(R^2)\right] \ge 1-\alpha$.

Finally, we note that f_0, under some additional conditions, has a causal interpretation, and is directly related to the notions of partial dependence and total effect. Consequently, our proposed orthogonal inference procedure for f_0 may be useful for inferring causal effect.

Specifically, following Friedman (2001), the partial dependence of the response Y on the primary modality $X = x_0 \in \mathcal{X}^p$ is defined as,

$$\mathbb{E}_Z\left[\mathbb{E}_U(Y)\right] = \mathbb{E}_Z\left[f_0(x_0) + g_0(z)\right] = f_0(x_0) + c, \quad c \in \mathbb{R}, \tag{14}$$

where (X, Z, Y) follows model (1). That is, the partial dependence is the expectation of Y over the marginal distribution of all modalities other than X. It is different from the conditional expectation, $\mathbb{E}_Z[\mathbb{E}_U(Y)|X = x_0] = \mathbb{E}_{Z|X=x_0}[f_0(x_0) + g_0(z)]$, where the expectation is taken over the conditional distribution of Z given $X = x_0$. By (14), we see that the partial dependence is equal to $f_0(x_0)$ up to an additive constant c. This property does not hold for the conditional expectation.

Next, following Pearl (2009) and Zhao and Hastie (2021), the partial dependence measure in (14) coincides with the backdoor adjustment formula for identifying the causal effect of X on Y given the observational data. More specifically, view (1) as a structural equation model, where each of the $(M+1)$ modalities

$\{X, Z_{(1)}, \ldots, Z_{(M)}\}$ corresponds to one of the $(M+1)$ nodes in a directed acyclic graph (Pearl 2009). Let a path be a consecutive sequence of edges of the directed graph, and a back-door path be a path that contains an arrow into X. If the following back-door criteria are satisfied, such that none of $\{Z_{(1)}, \ldots, Z_{(M)}\}$ is a descendant of X, and $\{Z_{(1)}, \ldots, Z_{(M)}\}$ blocks all back-door paths between X and Y, then the partial dependence measure in (14), or equivalently $f_0(\cdot)$, can be interpreted as the total effect of the primary modality X affecting the outcome Y.

5. Comparison With Alternative Methods

We next analytically compare our method with a number of important alternative solutions, and carefully evaluate the asymptotic behavior of each estimator.

5.1. Uni-Modality Regression

A common solution in practice is to focus on a single data modality and exclude all other modalities from the analysis. This approach is simple, and shares a similar spirit as the marginal regression (Fan and Lv 2008). We term it as the *uni-modality regression*. Specifically, it regresses the outcome on the primary modality, and estimate the primary parameter θ_0 by,

$$\widehat{\theta}_{\mathrm{UR}} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[Y_i - \Phi(X_i)^{\mathsf{T}} \theta \right]^2 \right\}.$$

Proposition 3 characterizes the asymptotic behavior of the uni-modality estimator $\widehat{\theta}_{\mathrm{UR}}$.

Proposition 3. Suppose the system of models (1)–(4) hold. Suppose $\mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}]$ is invertible. Then, the uni-modality regression estimator $\widehat{\theta}_{\mathrm{UR}}$ satisfies that,

$$\widehat{\theta}_{\mathrm{UR}} - \theta_0 = \left\{ \mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}] \right\}^{-1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \Phi(X_i) \left[\delta_0(X_i) + g_0(Z_i) + U_i \right] \right\} + o_p(N^{-1/2}).$$

The proof of this proposition is given in Appendix S7. We next compare the behavior of $\widehat{\theta}_{\mathrm{UR}}$ with our orthogonal estimator $\widehat{\theta}$ in (10) in terms of the asymptotic bias and variance, respectively.

In terms of the bias, we note that $\widehat{\theta}_{\mathrm{UR}}$ may suffer from a severe bias, because

$$\mathbb{E}(\widehat{\theta}_{\mathrm{UR}}) - \theta_0 = \left\{ \mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}] \right\}^{-1} \mathbb{E}\{\Phi(X)[\delta_0(X) + g_0(Z)]\} + o(N^{-1/2}),$$

which can be arbitrarily large, due to both the model error δ_0 in (2), and the effect of the auxiliary modality reflected by g_0 in (1). In multimodal analysis, however, both δ_0 and g_0 can be substantial. Because of this bias, we have $\sqrt{N}(\widehat{\theta}_{\mathrm{UR}} - \theta_0) = O_p(\sqrt{N})$, which diverges as N tends to infinity. Consequently, $\widehat{\theta}_{\mathrm{UR}}$ is unsuitable for statistical inference tasks. By contrast, the proposed orthogonal estimator $\widehat{\theta}$ is asymptotically unbiased.

In terms of the variance, we note that $\widehat{\theta}_{\mathrm{UR}}$ achieves a variance that is no larger than that of $\widehat{\theta}$. Specifically, the asymptotic variance of $\widehat{\theta}_{\mathrm{UR}}$ is $\mathrm{var}(\widehat{\theta}_{\mathrm{UR}}) = N^{-1}\sigma^2\{\mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}]\}^{-1}$. Compared to the asymptotic variance of our orthogonal estimator $\widehat{\theta}$ as given in (11), we have

$$\mathrm{var}(\widehat{\theta}) - \mathrm{var}(\widehat{\theta}_{\mathrm{UR}}) \geq 0, \text{ as } N \to \infty,$$

in the sense that the difference of the two covariance matrices is semi-positive definite. The two asymptotic variances are equal only when $r_0 = 0$ in (4), that is, when the primary and auxiliary modalities are completely independent of each other. The inflated variance of $\widehat{\theta}$ compared to that of $\widehat{\theta}_{\mathrm{UR}}$ is due to the intrinsic correlation between X and Z that is modeled by r_0. It can be viewed as a generalization of the well-known variance inflation phenomenon in the classical linear regression model due to the collinearity. For instance, consider the linear model $Y = X\theta_0 + Z^{\mathsf{T}}\beta_0 + U$, with $\mathbb{E}(X) = \mathbb{E}(Y) = 0$. The variance of the least-square estimator becomes $\mathbb{E}(U^2)/[\mathbb{E}(X^2)(1-\kappa)]$ after incorporating the auxiliary modality Z, where $\kappa = \mathbb{E}(XZ^{\mathsf{T}})[\mathbb{E}(ZZ^{\mathsf{T}})]^{-1}\mathbb{E}(ZX)/\mathbb{E}(X^2)$ characterizes the correlation between X and Z. This variance increases compared to the case when there is no Z in the model. On the other hand, we also note that, the orthogonal estimator $\widehat{\theta}$ actually attains the smallest possible variance when Z is incorporated, as shown in Theorem 2.

5.2. Debiased Uni-Modality Regression

We next consider a debiased version of the uni-modality regression. Numerous debiasing strategies have been successfully developed in high-dimensional regression modeling in recent years (see, e.g., Zhang and Zhang 2014; van de Geer et al. 2014; Cai and Guo 2017, among others). The debiased estimator is obtained in two stages. First, the model error δ_0 is estimated based on the uni-modality regression estimator $\widehat{\theta}_{\mathrm{UR}}$ and some machine learning method as in (7),

$$\widehat{\delta}_{\mathrm{DUR}} = \underset{\delta \in \mathcal{H}_\delta}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[Y_i - \Phi(X_i)\widehat{\theta}_{\mathrm{UR}} - \delta(X_i) \right]^2 + \lambda_N^{\delta} \mathrm{PEN}_{\mathcal{H}_\delta}(\delta) \right\},$$

where $\lambda_N^{\delta} \geq 0$ is a tuning parameter. Then the debiased estimator of θ_0 is obtained by explicitly taking the model error into account

$$\widehat{\theta}_{\mathrm{DUR}} = \underset{\theta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left[Y_i - \Phi(X_i)^{\mathsf{T}}\theta - \widehat{\delta}_{\mathrm{DUR}}(X_i) \right]^2 \right\}.$$

Proposition 4 characterizes the asymptotic behavior of the debiased uni-modality estimator $\widehat{\theta}_{\mathrm{DUR}}$.

Proposition 4. Suppose the conditions of Proposition 3 hold. Suppose the regularity condition (C1) holds. Then, the debiased uni-modality regression estimator $\widehat{\theta}_{\mathrm{DUR}}$ satisfies that,

$$\widehat{\theta}_{\mathrm{DUR}} - \theta_0 = \{\mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}]\}^{-1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \Phi(X_i)[g_0(Z_i) + U_i] \right\} + O_p[(\mathbb{E}\{[\widehat{\delta}_{\mathrm{DUR}}(X) - \delta_0(X)]^2\})^{1/2}] + o_p(N^{-1/2}).$$

The proof of this proposition is given in Appendix S8. We make two observations regarding the asymptotic bias of $\widehat{\theta}_{\mathrm{DUR}}$. First, $\widehat{\theta}_{\mathrm{DUR}}$ indeed achieves a reduced bias compared to the uni-modality estimator $\widehat{\theta}_{\mathrm{UR}}$. This is because under the regularity condition (C3), the bias of $\widehat{\theta}_{\mathrm{DUR}}$ is

$$\mathbb{E}(\widehat{\theta}_{\mathrm{DUR}})-\theta_0 = \{\mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}]\}^{-1}\mathbb{E}[\Phi(X)g_0(Z)]+o(N^{-1/4}).$$

Comparing this bias with that of $\widehat{\theta}_{\mathrm{UR}}$, we see that $\widehat{\theta}_{\mathrm{DUR}}$ removes the bias term due to the model error δ_0 as $N \to \infty$, but $\widehat{\theta}_{\mathrm{UR}}$ does not. On the other hand, $\widehat{\theta}_{\mathrm{DUR}}$ is still an inconsistent and biased estimator of θ_0, because $\widehat{\theta}_{\mathrm{DUR}}$ does not remove the bias due to the effect of the auxiliary modality g_0. Consequently, $\widehat{\theta}_{\mathrm{DUR}}$ is unsuitable for statistical inference neither.

5.3. Simple Joint Regression

Another common solution in multimodal analysis is to incorporate multiple data modalities in a simple additive fashion into a single regression model. This strategy is intuitive, and we term it as the *simple joint regression*. Specifically, it obtains the joint estimator for $\{\theta_0, g_0\}$ as,

$$\{\widehat{\theta}_{\mathrm{SJR}}, \widehat{g}_{\mathrm{SJR}}\} = \underset{\theta\in\mathbb{R}^d, g\in\mathcal{H}_g}{\arg\min} \left\{ \frac{1}{N}\sum_{i=1}^{N}\left[Y_i - \Phi(X_i)^{\mathsf{T}}\theta - g(Z)\right]^2 + \lambda_N^g \mathrm{PEN}_{\mathcal{H}_g}(g)\right\},$$

where $\lambda_N^g \geq 0$ is a tuning parameter, and $\widehat{g}_{\mathrm{SJR}}$ is obtained by a machine learning method as in (7).

Proposition 5 characterizes the asymptotic behavior of the simple joint estimator $\widehat{\theta}_{\mathrm{SJR}}$.

Proposition 5. Suppose the conditions of Proposition 3 hold. Suppose the regularity condition (C1) holds. Then, the simple joint regression estimator $\widehat{\theta}_{\mathrm{SJR}}$ satisfies that

$$\widehat{\theta}_{\mathrm{SJR}} - \theta_0 = \{\mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}]\}^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\Phi(X_i)[\delta_0(X_i)+U_i]\right\} + O_p((\mathbb{E}\{[\widehat{g}_{\mathrm{SJR}}(Z)-g_0(Z)]^2\})^{1/2}) + o_p(N^{-1/2}).$$

The proof of this proposition is given in Appendix S9. We again study the asymptotic behavior of $\widehat{\theta}_{\mathrm{SJR}}$. Under the regularity condition (C3), the asymptotic bias of $\widehat{\theta}_{\mathrm{SJR}}$ is,

$$\mathbb{E}(\widehat{\theta}_{\mathrm{SJR}}) - \theta_0 = \{\mathbb{E}[\Phi(X)\Phi(X)^{\mathsf{T}}]\}^{-1}\mathbb{E}[\Phi(X)\delta_0(X)] + o(N^{-1/4}),$$

which is not vanishing due to the non-zero model error δ_0. The mean squared error of $\widehat{\theta}_{\mathrm{SJR}}$ is,

$$\mathbb{E}\left[(\widehat{\theta}_{\mathrm{SJR}} - \theta_0)^2\right] = O(\mathbb{E}\{[\widehat{g}_{\mathrm{SJR}}(Z) - g_0(Z)]^2 + \delta_0^2(X)\}),$$

which does not converge at the rate of N^{-1} if $\widehat{g}_{\mathrm{SJR}}$ is estimated using machine learning methods, or if δ_0 is not negligible. Consequently, $\widehat{\theta}_{\mathrm{SJR}}$ is generally an inefficient and biased estimator of θ_0.

5.4. Double/Debiased Machine Learning

The seminal work of Chernozhukov et al. (2018) developed the framework of double/debiased machine learning (DML), which lays the foundation for the inference of the primary parameter of interest in the presence of high-dimensional nuisance parameters. Our proposal extends the DML framework to incorporate the additional model error δ_0. More specifically, DML randomly splits the data into Q disjoint chunks, and estimates g_0 by

$$\widehat{g}_{\mathrm{DML},q} = \underset{g\in\mathcal{H}_g}{\arg\min}\left\{\frac{1}{n}\sum_{i\in I_q^c}[Y_i - g(Z_i)]^2 + \lambda_n^g \mathrm{PEN}_{\mathcal{H}_g}(g)\right\}.$$

where $\lambda_n^g \geq 0$ is a tuning parameter. It then estimates θ_0 by

$$\widehat{\theta}_{\mathrm{DML}} = \left\{\frac{1}{nQ}\sum_{q=1}^{Q}\sum_{i\in I_q}[\Phi(X_i) - \widehat{r}_q(Z_i)]\Phi(X_i)^{\mathsf{T}}\right\}^{-1} \frac{1}{nQ}\sum_{q=1}^{Q}\sum_{i\in I_q}[\Phi(X_i) - \widehat{r}_q(Z_i)][Y_i - \widehat{g}_{\mathrm{DML},q}(Z_i)].$$

Proposition 6 characterizes the asymptotic behavior of DML estimator $\widehat{\theta}_{\mathrm{DML}}$.

Proposition 6. Suppose the conditions of Proposition 3 hold. Suppose the regularity conditions (C1)–(C3) hold. Then the DML estimator $\widehat{\theta}_{\mathrm{DML}}$ satisfies that,

$$\widehat{\theta}_{\mathrm{DML}} - \theta_0 = (\mathbb{E}[VV^{\mathsf{T}}])^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}V_iU_i\right) + O_p(\{\mathbb{E}[\delta_0^2(X)]\}^{1/2}) + o_p(N^{-1/2}).$$

The proof is given in Appendix S10. The mean squared error of $\widehat{\theta}_{\mathrm{DML}}$ is,

$$\mathbb{E}\left[(\widehat{\theta}_{\mathrm{DML}}-\theta_0)^2\right] = \frac{1}{N}\sigma^2(\mathbb{E}[VV^{\mathsf{T}}])^{-1} + O(\mathbb{E}[\delta_0^2(X)]) + o(N^{-1}).$$

Compared to our estimator $\widehat{\theta}$, whose mean squared error is $N^{-1}\sigma^2(\mathbb{E}[VV^{\mathsf{T}}])^{-1} + o(N^{-1})$, $\widehat{\theta}_{\mathrm{DML}}$ has an inflated mean squared error at the order of $\mathbb{E}[\delta_0^2(X)]$. Consequently, it cannot achieve the $\sqrt{N}$-consistency if the model error δ_0 is not negligible.

6. Simulations

We next study the finite-sample performance of the proposed orthogonalized kernel debiased machine learning (OKDML) method. We first evaluate the performance of inferring θ_0 in an additive model setting. We also numerically compare with the alternative methods of uni-modality regression (UR), debiased uni-modality regression (DUR), simple joint regression (SJR), and double machine learning (DML) that ignores δ_0. We next evaluate the performance of inferring f_0 in a high-dimensional additive setting. We also study the sensitivity of using different machine learning methods for nuisance function estimation when inferring θ_0, and report the results in Section S11 of the appendix. In all these examples, the model error δ_0 is estimated in the RKHS constructed as in Proposition 2. We use the Matérn

kernel $K(x,x') = (1+\sqrt{5}\|x-x'\|+5\|x-x'\|^2/3)\exp(-\sqrt{5}\|x-x'\|)$, where the corresponding RKHS contains twice differentiable functions. The tuning parameter λ_n^δ in (7) is selected by generalized cross-validation (Wahba 1990). We set $Q = 2$ in Algorithm 1.

6.1. Empirical Performance of Inference on θ_0

We begin with an additive model, $Y_i = f_0(X_i) + g_{01}(Z_{i1}) + g_{02}(Z_{i2}) + g_{03}(Z_{i3}) + U_i$, where

$$f_0(x) = 5x - [\cos(2\pi x) + \sin(2\pi x)],$$

$$g_{01}(z_1) = 6\big[0.1\sin(2\pi z_1) + 0.2\cos(2\pi z_1) + 0.3\sin^2(2\pi z_1) + 0.4\cos^3(2\pi z_1) + 0.5\sin^3(2\pi z_1))\big],$$

$$g_{02}(z_2) = 3(2z_2-1)^2, \quad g_{03}(z_3) = \frac{4\sin(2\pi z_3)}{2-\sin(2\pi z_3)}.$$

We generate random variables $E_1,\ldots,E_5$ independently from Uniform[0, 1], and set the primary and auxiliary modalities as $X = (E_1+\rho E_5)/(1+\rho) \in \mathcal{X} = [0,1]$, and $Z_j = (E_{j+1}+\rho E_5)/(1+\rho)$, for some $\rho > 0$ and $j = 1,2,3$. The correlation between any two variables in X and Z is thus $\rho^2/(1+\rho^2)$. We generate iid copies $(X_i, Z_{i1}, Z_{i2}, Z_{i3})$ of (X, Z_1, Z_2, Z_3), and generate the error U_i from $\mathcal{N}(0,\sigma^2)$. We set the sample size $N = 500$. We set $\eta(x,\theta_0) = \theta_0 x$, and apply the random forests averaged over 500 trees to estimate the nuisance functions $\{r_0, g_0\}$.

Figure 1 shows the histograms of the competing estimators, $\widehat{\theta}_{\mathrm{UR}}, \widehat{\theta}_{\mathrm{DUR}}, \widehat{\theta}_{\mathrm{SJR}}, \widehat{\theta}_{\mathrm{DML}}$, and our proposed OKDML estimator $\widehat{\theta}_{\mathrm{OKDML}}$, under $\rho = 1$ and $\sigma = 1$, based on 500 data replications. It is clearly seen that all four competing estimators are biased, whereas the histogram of the OKDML estimator $\widehat{\theta}_{\mathrm{OKDML}}$ matches that of the normal distribution. Figure 2 further reports the empirical mean squared error of different estimators under various combinations of the noise level σ and the correlation level ρ. When σ^{-1} increases, the signal-to-noise ratio increases. However, the mean squared errors of the four competing methods do not decrease much due to the estimation bias, whereas the mean squared error of our OKDML estimator continuously decreases.

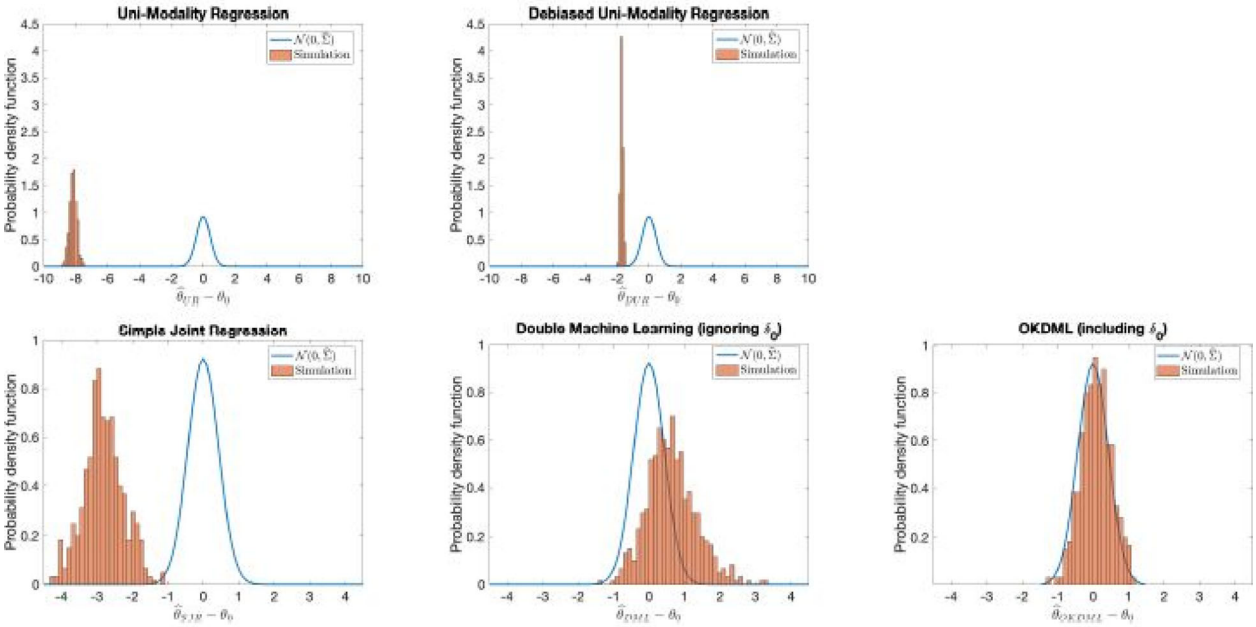

Figure 1. Empirical distribution of the estimator of θ_0 based on 500 data replications. The bell-shape curve denotes the oracle normal distribution.

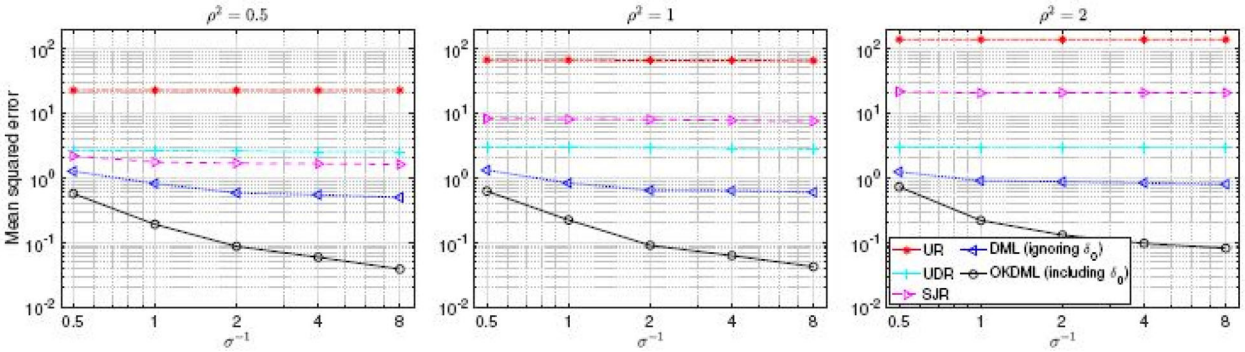

Figure 2. Mean squared error of the estimator of θ_0 with varying noise level σ and correlation level ρ. Both axes are in the log scale.

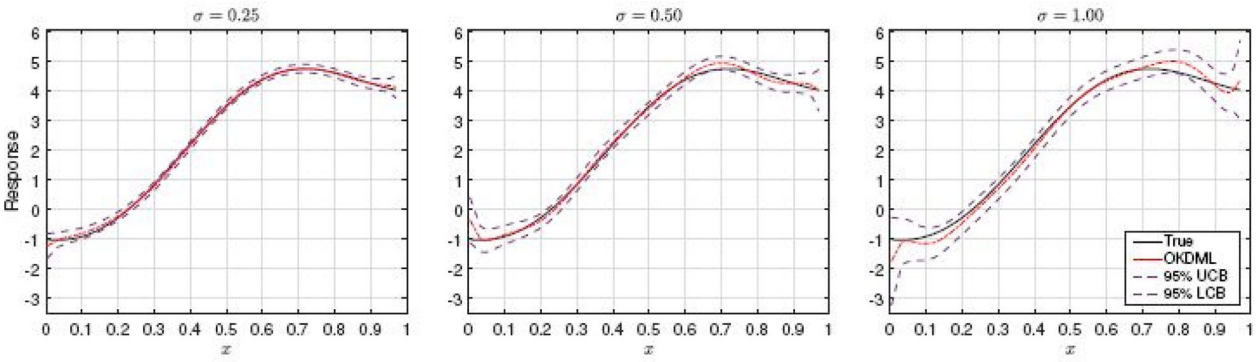

Figure 3. The true and estimated primary function $f_0(x)$, with the 95% upper and lower confidence bounds, of the OKDML method, under varying noise level σ.

6.2. Empirical Performance of Inference on f_0

We next consider a high-dimensional additive model, $Y_i = f_0(X_i) + \sum_{j=1}^{600} g_{0j}(Z_{ij}) + U_i$, where $f_0(x), g_{01}(z_1), g_{02}(z_2), g_{03}(z_3)$ are the same as the first example, and

$$g_{0j}(z_j) = z_j, \quad \text{for } j \in \{4, \ldots, 100\},$$
$$g_{0j}(z_j) = 0, \quad \text{for } j \in \{101, \ldots, 600\}.$$

We generate random variables $E_1, \ldots, E_{602}$ independently from Uniform[0, 1], and set the primary and auxiliary modalities as $X = (E_1 + \rho E_{602})/(1 + \rho)$, and $Z_j = (E_{j+1} + \rho E_{602})/(1 + \rho)$, for $\rho = 1$ and $j = 1, \ldots, 600$. We generate i.i.d. copies $(X_i, Z_{i1}, Z_{i2}, \ldots, Z_{i600})$ of $(X, Z_1, Z_2, \ldots, Z_{600})$, and generate the error U_i from $\mathcal{N}(0, \sigma^2)$ with $\sigma \in \{0.25, 0.5, 1\}$. We set the sample size $N = 500$.

We construct both the confidence band (12) for the primary effect $f_0(x)$, and the confidence interval (13) for the coefficient of determination R^2. We use polynomial basis functions with $s = 5$ following Theorem 3, while we estimate δ_0 in a similar way as in the first example. We employ the Lasso to estimate the nuisance functions $\{r_0, g_0\}$ due to the high-dimensionality of this example, and tune the Lasso parameter using tenfold cross-validation. We compute the quantile estimator $\widehat{c}_N(\alpha/2)$ in (12) by bootstrap with 500 replications.

Figure 3 shows the true and estimated primary function $f_0(x)$, along with the 95% upper and lower confidence bounds, of the proposed orthogonal method with the varying noise level σ. We also compute the empirical coverage probability of the confidence band $\mathcal{C}_N$ at the significance level 95%, by discretizing the interval $\mathcal{X} = [0, 1]$ into 1000 grids, then calculating the percentage that the confidence band covers the truth on the 1000 grid points in 500 data replications. The resulting coverage probability is 0.968, 0.958, and 0.946, when $\sigma = 0.25, 0.50$, and 1.00, respectively. Moreover, we compute the empirical coverage probability of CI(R^2) as the percentage that the confidence interval covers the true R^2. The resulting coverage probability is 0.990, 0.972, and 0.964, when $\sigma = 0.25, 0.50$, and 1.00, respectively. It is seen from both the estimated function and the coverage probability that our proposed method works well.

7. Multimodal Neuroimaging Study for AD

We revisit the motivating example of multimodal neuroimaging analysis for AD. The data are part of the Berkeley Aging Cohort Study, and consists of 697 subjects. For each subject, the imaging data includes the anatomical MRI scan, which measures brain cortical thickness and is summarized as a 68-dimensional vector that corresponds to 68 predefined brain regions-of-interest (ROIs), and the PET scan, which measures tau deposition and is summarized as a 70-dimensional vector that corresponds to 70 ROIs. In addition, the subject's age, gender, education, and a scalar measure of the total amyloid-β accumulation are collected. The response is a composite cognition score that combines assessments of episodic memory, timed executive function, and global cognition. We study two scientific questions given this data, first, the effect of brain atrophy on cognition after controlling for demographic variables and amyloid-β, tau depositions, and second, the cascade of AD biomarkers as suggested by Jack et al. (2010).

For the first problem, we take the brain MRI cortical thickness as the primary modality, with $p = 68$, and take the PET tau deposition along with the demographic variables and the total amyloid-β as the auxiliary modalities, resulting in $p' = 74$. We apply the proposed OKDML method to infer the effect of cortical thickness of individual brain regions on the cognitive outcome. We adopt a similar implementation as used in our first simulation example, and set $\eta(x, \theta_0) = \theta_0^{\mathsf{T}} x$. Table 1 reports the estimated effects of the brain regions where the cortical thickness is found to be significantly correlated with the cognitive outcome after controlling for amyloid-β, tau and other covariates, with the corresponding p-values under the FDR control at the 5% level (Benjamini and Hochberg 1995). These findings agree well with the AD literature. Particularly, the entorhinal cortex is a brain area located in the medial temporal lobe, and functions as a hub in a widespread network for memory, navigation and the perception of time. Atrophy in the entorhinal cortex has been consistently reported in AD (Pini et al. 2016). The parahippocampal gyrus is a grey matter

Table 1. Multimodal study of AD: the identified significant brain regions.

	Estimate	SD	p-value
Entorhinal cortex, left	3.214	0.709	6.957×10^{-6}
Entorhinal cortex, right	2.853	0.671	2.454×10^{-5}
Superior temporal cortex, left	10.42	2.444	2.321×10^{-5}
Superior temporal cortex, right	5.061	1.451	5.213×10^{-4}
Parahippocampal gyrus, left	1.076	0.362	3.112×10^{-3}
Parahippocampal gyrus, right	1.366	0.474	4.098×10^{-3}

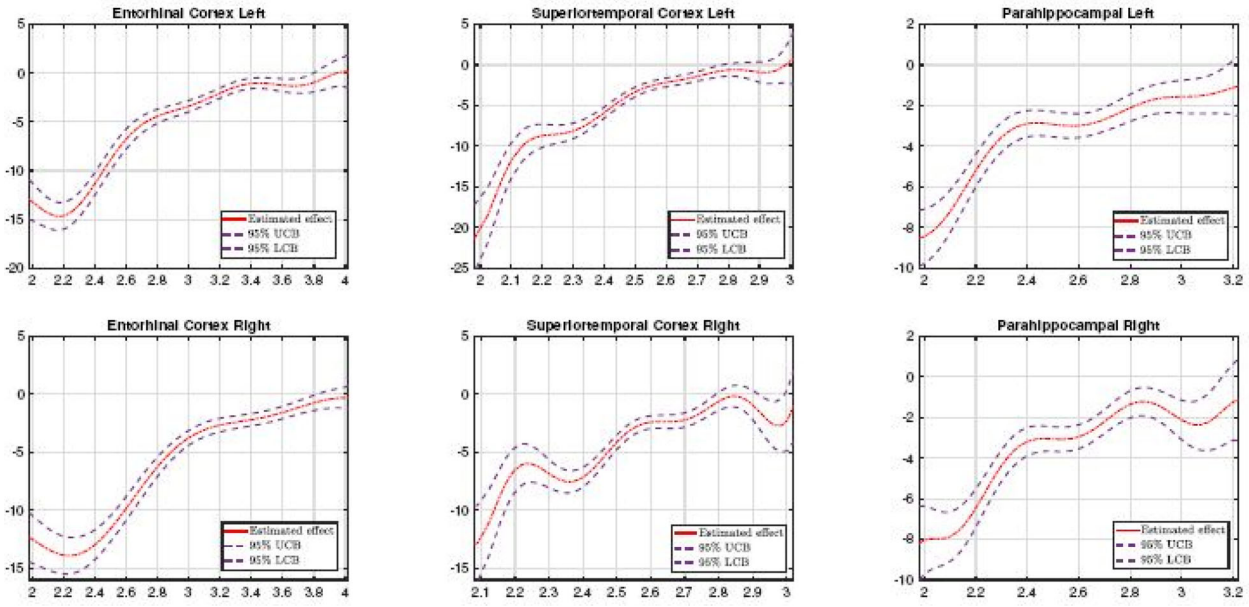

Figure 4. The estimated individual effect of the significant brain regions.

cortical region of the brain that surrounds the hippocampus, and plays an important role in memory encoding and retrieval. It is among the first to suffer damage from AD (Jack et al. 2010). The superior temporal gyrus locates in the temporal lobe, and contains the Wernicke's area responsible for processing of speech. Its connection with AD needs further verification. Moreover, Figure 4 shows the confidence band for the estimated individual effect of each significant brain region. Besides, the 95% confidence interval for the R^2 measure is $(0.402, 0.437)$, which supports the common belief that brain structural atrophy is closely related to the cognition outcome.

For the second problem, Jack et al. (2010) suggested that tau deposition precedes structural atrophy in AD pathogenesis. To help verify this theory, we take the PET tau deposition as the primary modality, with $p = 70$, then compare two model fits, one with the MRI cortical thickness as part of the auxiliary modalities, and the other without. In both models, we include age, gender, education and amyloid-β as the auxiliary modalities. This yields $p' = 72$ when the cortical thickness is included, and $p' = 4$ if not. We obtain the 95% confidence interval for the total effect of tau, which is $(-1.724, 0.702)$ when the cortical thickness is included, and $(-5.212, -3.945)$ when it is not. These results suggest that, not including structural atrophy as the auxiliary modality would result in a much larger effect of tau on cognition outcome, which in turn implies structural atrophy likely occurs after tau deposition, and thus lends some support to the existing theory.

8. Discussion

We conclude the article by reiterating and further elaborating the innovation of our proposal and its difference from Chernozhukov et al. (2018). We divide our discussion in two parts: the inference for the primary parameter θ_0, and the inference for the primary function f_0. For each part, we first discuss why the question is important, what are the challenges, and why the existing solutions are not directly applicable. We then detail our methodological and theoretical contributions.

(A) Inference for θ_0: A key innovation of our proposal is that we allow an explicit and non-vanishing model error δ_0 for the primary modality effect f_0 in (2), whereas Chernozhukov et al. (2018) did not consider δ_0. This difference has profound implications in model interpretation, estimation approach, and theoretical analysis, which in turn differentiates our proposal from the existing DML solutions such as Chernozhukov et al. (2018) and Kozbur (2020).

1. In scientific studies such as multimodal analysis, it is crucial to balance model interpretability and model flexibility, which is also the main motivation for this article. In numerous applications, it is not uncommon for scientists to employ some relatively simple models, for example, linear models, for the primary modality. Such models are easy to interpret, but may not be accurate, and can induce a nonnegligible approximation error. In other applications, it is likely to employ more advanced and accurate but less interpretable models. It is thus pivotal to offer inferential robustness for both cases, and to achieve a balanced trade-off between model interpretability and model flexibility.
2. Chernozhukov et al. (2018) focused on a low-dimensional primary parameter involving no additional error. Kozbur (2020) extended to a nonparametric primary function through basis expansion, but imposed that the error must be negligible, in that the squared approximation error is $o(N^{-1})$. However, this condition requires either the working model to be sufficiently close to the truth, or the number of basis functions to diverge to infinity with the sample size, which in effect excludes the use of simple yet inaccurate models in characterizing the effect of the primary modality. We also utilize basis expansion to approximate the primary modality effect, but we do not require a vanishing approximation error, nor a diverging number of basis functions, when we establish the asymptotic guarantees of the estimated θ_0.

3. To decouple the primary parameter θ_0 and the non-negligible model error δ_0, we introduce the second form of orthogonality, the decomposition orthogonality, in addition to the Neyman orthogonality, into the framework of double/debiased machine learning. The new orthogonality is similar to the perpendicularity property in smoothing splines (Wahba 1990). We show in Proposition 2 that, this decomposition orthogonality between the expanded basis functions and the model error ensures the identifiability of the primary parameter θ_0. This is a new result, and is potentially useful for obtaining improved inferential robustness in other settings too when there exist non-negligible model error.
4. Methodologically, the new decomposition orthogonality leads to the construction of a new RKHS, and a residual learning approach in our estimation algorithm, which helps decouple and remove the impact of the model error in parameter estimation.
5. Theoretically, we successfully establish the $\sqrt{N}$-consistency and asymptotic normality of the estimated main parameter under model error. Compared to the existing semi-parametric inferential analysis, our proof relies on the score function that is Neyman orthogonal with respect to the model error δ_0, and as such requires a weaker regularity condition (C3) than the Donsker conditions that are common but would often fail in multimodal analysis. Compared to the alternative multimodal solutions, including uni-modality regression, debiased uni-modality regression, simple joint regression, and double/debiased machine learning without taking into account δ_0, we show in Section 5 that our estimator is unbiased, but the alternative ones all suffer from a non-vanishing estimation bias when there is model error.
6. We also show that our estimator is semi-parametric efficient, in that it achieves the highest possible efficiency, when the measurement error U follows a normal distribution. This is also a new result, and its proof is based on constructing an oracle estimator from an ideal finite-dimensional parameter space that achieves the same asymptotic variance as our estimator from an infinite-dimensional parameter space.

(B) Inference for f_0: Another key innovation of our proposal is that we establish the confidence band for the nonparametric primary function f_0 in the presence of high-dimensional nonlinear nuisance function, whereas Chernozhukov et al. (2018) considered a low-dimensional primary parameter involving no nonparametric f_0.

1. The function f_0 captures the predicted effect of the primary modality, quantifies the amount of contribution of the primary modality in terms of the percentage of variation explained, and also has some causal interpretation under additional conditions. It is thus of great scientific interest to perform rigorous inference on f_0.
2. The high-dimensional nonparametric inference of f_0 is challenging. Construction of confidence intervals in such a setting is often intertwined with penalized model estimation and selection, giving rise to post-regularization inference. There has been pioneering research on high-dimensional inference for parametric models such as linear and generalized linear models (Zhang and Zhang 2014; van de Geer et al. 2014; Cai and Guo 2017, among others). Early nonparametric inference usually focused on a fixed dimensionality (e.g., Wahba 1983; Fan and Jiang 2005). More recently, Lu, Kolar, and Liu (2020) and Kozbur (2020) studied high-dimensional inference for nonparametric models. However, as we point out after Theorem 3, Lu, Kolar, and Liu (2020) required the variables to be only weakly correlated, which is unlikely to hold for multimodal data, whereas Kozbur (2020) required a fast vanishing approximation error, which sacrifices model interpretability.
3. Our inference on f_0 is different from the existing literature, as it targets a high-dimensional nonparametric regression setting, allows the primary and auxiliary modalities to be strongly correlated, and also takes into account a non-negligible approximation error when modeling the primary modality effect.
4. Technically, we extend the inferential framework of Chernozhukov, Chetverikov, and Kato (2014) to our system of models for multimodal data analysis. We construct the supremum of high-dimensional empirical processes arising from our OKDML estimator, which enables us to control the supreme norm rate of our estimator, while allowing a diverging dimensionality. We then approximate the supremum with a Gaussian multiplier process to derive the corresponding quantiles and to obtain the asymptotically valid confidence band.

In summary, our proposal integrates reproducing kernel learning (Wahba 1990) with double/debiased machine learning (Chernozhukov et al. 2018). We believe it makes a useful addition to and also extends the scope of the general methodology and theory for multimodal data analysis, high-dimensional nonparametric inference, as well as double/debiased machine learning. Meanwhile, such an extension is far from simple and straightforward.

Supplementary Appendix

The online Supplementary Appendix collects all technical proofs and additional simulation results.

Acknowledgments

The authors thank to the editor, the associate editor, and two referees for their constructive comments and suggestions.

Funding

Dai's research was partially supported by CDAR, Department of Economics, the University of California, Berkeley, and this work was done while Dai was visiting the Simons Institute for the Theory of Computing. Li's research was partially supported by NSF grant CIF-2102227, and NIH grants R01AG061303, R01AG062542, and R01AG034570.

References

Alam, M. A., Lin, H.-Y., Deng, H.-W., Calhoun, V. D., and Wang, Y.-P. (2018), "A Kernel Machine Method for Detecting Higher Order Interactions in Multimodal Datasets: Application to Schizophrenia," *Journal of Neuroscience Methods*, 309, 161–174. [1796,1798]

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. (2019), "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41, 423–443. [1796]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [1806]

Biau, G. (2012), "Analysis of a Random Forests Model," *Journal of Machine Learning Research*, 13, 1063–1095. [1800]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [1800]

Breiman, L. (2001), "Statistical Modeling: The Two Cultures," *Statistical Science*, 16, 199–231. [1796]

Bühlmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data: Methods, Theory and Applications*, Heidelberg: Springer [1800]

Buja, A., Hastie, T., and Tibshirani, R. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics*, 17, 453–510. [1800]

Cai, Q., Wang, H., Li, Z., and Liu, X. (2019), "A Survey on Multimodal Data-Driven Smart Healthcare Systems: Approaches and Applications," *IEEE Access*, 7, 133583–133599. [1796]

Cai, T. T., and Guo, Z. (2017), "Confidence Intervals for High-Dimensional Linear Regression: Minimax Rates and Adaptivity," *The Annals of Statistics*, 45, 615–646. [1803,1808]

Chen, X., and White, H. (1999), "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions on Information Theory*, 45, 682–691. [1800]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters: Double/Debiased Machine Learning," *The Econometrics Journal*, 21, C1–C68. [1797,1798,1799,1800,1804,1807,1808]

Chernozhukov, V., Chetverikov, D., and Kato, K. (2014), "Anti-Concentration and Honest, Adaptive Confidence Bands," *The Annals of Statistics*, 42, 1787–1818. [1797,1802,1808]

Fan, J., and Jiang, J. (2005), "Nonparametric Inferences for Additive Models," *Journal of the American Statistical Association*, 100, 890–907. [1808]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [1803]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, 29, 1189–1232. [1802]

Giné, E., and Nickl, R. (2009), "An Exponential Inequality for the Distribution Function of the Kernel Density Estimator, With Applications to Adaptive Estimation," *Probability Theory and Related Fields*, 143, 569–596. [1802]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, Boca Raton, FL: CRC Press. [1798]

Hinrichs, C., Singh, V., Xu, G., Johnson, S. C., and Initiative, A. D. N. (2011), "Predictive Markers for ad in a Multi-Modality Framework: An Analysis of MCI Progression in the ADNI Population," *Neuroimage*, 55, 574–589. [1796]

Huang, J. Z., Zhang, L., and Zhou, L. (2007), "Efficient Estimation in Marginal Partially Linear Models for Longitudinal/Clustered Data Using Splines," *Scandinavian Journal of Statistics*, 34, 451–477. [1798]

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., and Trojanowski, J. Q. (2010), "Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological Cascade," *The Lancet Neurology*, 9, 119 – 128. [1796,1797,1806,1807]

Kosorok, M. R. (2007), *Introduction to Empirical Processes and Semiparametric Inference*, Springer Science & Business Media, New York: Springer. [1801]

Kozbur, D. (2020), "Inference in Additively Separable Models With a High-Dimensional Set of Conditioning Variables," *Journal of Business & Economic Statistics*, 39, 1–17. [1797,1802,1807,1808]

Li, G., Liu, X., and Chen, K. (2019), "Integrative Multi-View Reduced-Rank Regression: Bridging Group-Sparse and Low-Rank Models," *Biometrics*, 75, 593–602. [1797]

Li, Q., and Li, L. (2021), "Integrative Factor Regression and Its Inference for Multimodal Data Analysis," *Journal of the American Statistical Association*, DOI: 10.1080/01621459.2021.1914635. [1796,1797]

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 34, 2272–2297. [1798]

Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2013), "Joint and Individual Variation Explained (JIVE) for Integrated Analysis of Multiple Data Types," *The Annals of Applied Statistics*, 7, 523. [1797]

Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., and Mordatch, I. (2017), "Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6382–6393. Curran Associates. [1800]

Lu, J., Kolar, M., and Liu, H. (2020), "Kernel Meets Sieve: Post-Regularization Confidence Bands for Sparse Additive Model," *Journal of the American Statistical Association*, 115, 2084–2099 [1802,1808]

Ma, S., Carroll, R. J., Liang, H., and Xu, S. (2015), "Estimation and Inference in Generalized Additive Coefficient Models for Nonlinear Interactions With High-dimensional Covariates," *Annals of Statistics*, 43, 2102. [1798]

Mai, Q., and Zhang, X. (2019), "An Iterative Penalized Least Squares Approach to Sparse Canonical Correlation Analysis," *Biometrics*, 75, 734–744. [1797]

Nathoo, F. S., Kong, L., Zhu, H., and for the Alzheimer's Disease Neuroimaging Initiative. (2019), "A Review of Statistical Methods in Imaging Genetics," *Canadian Journal of Statistics*, 47, 108–131. [1797]

Newey, W. K. (1990), "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5, 99–135. [1797]

Newey, W. K., and Robins, J. R. (2018), "Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation," arXiv:1801.09138. [1800]

Neyman, J. (1959), "Optimal Asymptotic Tests of Composite Statistical Hypotheses," in *Probability and Statistics*, ed. U. Grenander, New York: Wiley, pp. 416–444. [1797,1798,1799]

Neyman, J. (1979), "$c(\alpha)$ Tests and Their Use," *Sankhya*, 1–21. [1797,1798,1799]

Pearl, J. (2009), *Causality*, Cambridge University Press, Cambridge. [1802,1803]

Pini, L., Pievani, M., Bocchetta, M., Altomare, D., Bosco, P., Cavedo, E., Galluzzi, S., Marizzoni, M., and Frisoni, G. B. (2016), "Brain Atrophy in Alzheimer's Disease and Aging," *Ageing Research Reviews*, 30, 25–48. [1806]

Raskutti, G., Wainwright, M. J., and Yu, B. (2011), "Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q-Balls," *IEEE Transactions on Information Theory*, 57, 6976–6994. [1801]

Richardson, S., Tseng, G. C., and Sun, W. (2016), "Statistical Methods in Integrative Genomics," *Annual Reviews of Statistics and Its Applications*, 3, 181–209. [1796]

Robins, J. M., and Rotnitzky, A. (1995), "Semiparametric Efficiency in Multivariate Regression Models With Missing Data," *Journal of the American Statistical Association*, 90, 122–129. [1797]

Shu, H., Wang, X., and Zhu, H. (2020), "D-cca: A Decomposition-Based Canonical Correlation Analysis for High-Dimensional Datasets," *Journal of the American Statistical Association*, 115, 292–306. [1797]

Sperling, R. A., Mormino, E. C., and Schultz, A. P., Betensky, R. A., Papp, K. V., Amariglio, R. E., Hanseeuw, B. J., Buckley, R., Chhatwal, J., Hedden, T., Marshall, G. A., Quiroz, Y. T., Donovan, N. J., Jackson, J., Gatchel, J. R., Rabin, J. S., Jacobs, H., Yang, H.-S., Properzi, M., Kirn, D. R., Rentz, D. M., Johnson, K. A. (2019), "The Impact of Amyloid-Beta and Tau on Prospective Cognitive Decline in Older Individuals," *Annals of Neurology*, 85, 181–193. [1796]

Uludağ, K., and Roebroeck, A. (2014), "General Overview on the Merits of Multimodal Neuroimaging Data Fusion," *Neuroimage*, 102, 3–10. [1796,1802]

van de Geer, S., Bühlmann, P., Ritov, Y. A., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [1803,1808]

van der Laan, M. J., and Rubin, D. (2006), "Targeted Maximum Likelihood Learning," *The International Journal of Biostatistics*, 2, 1–38. [1797,1799]

van der Vaart, A. W. (1998), *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics*, Cambridge: Cambridge University Press. [1800,1802]

Wahba, G. (1983), "Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Series B, 45, 133–150. [1808]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia, PA: SIAM. [1797,1798,1799,1800,1802,1805,1808]

Wang, L., Xue, L., Qu, A., and Liang, H. (2014), "Estimation and Model Selection in Generalized Additive Partial Linear Models for Correlated Data With Diverging Number of Covariates," *The Annals of Statistics*, 42, 592–624. [1798]

Xue, F., and Qu, A. (2021), "Integrating Multi-Source Block-Wise Missing Data in Model Selection," *Journal of the American Statistical Association*, 116, 1914–1927. [1797]

Zhang, C., and Zhang, S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1803,1808]

Zhao, Q., and Hastie, T. (2021), "Causal Interpretations of Black-Box Models," *Journal of Business & Economic Statistics*, 39, 272–281. [1802]

Zheng, W., and van der Laan, M. J. (2011), "Cross-Validated Targeted Minimum-Loss-Based Estimation," in *Targeted Learning*, eds. M. J. van der Laan, S. Rose, New York: Springer, pp. 459–474. Springer. [1799]

Zhu, H., Khondker, Z., Lu, Z., and Ibrahim, J. G. (2014), "Bayesian Generalized Low Rank Regression Models for Neuroimaging Phenotypes and Genetic Markers," *Journal of the American Statistical Association*, 109, 977–990. [1797]