

Causal Effect of Functional Treatment

Ruoxu Tan

RUOXUT@TONGJI.EDU.CN

*School of Mathematical Sciences and School of Economics and Management
Tongji University
Shanghai, China*

Wei Huang*

WEI.HUANG@UNIMELB.EDU.AU

*School of Mathematics and Statistics
University of Melbourne
Melbourne, VIC, Australia*

Zheng Zhang*

ZHENGZHANG@RUC.EDU.CN

*Center for Applied Statistics, Institute of Statistics & Big Data
Renmin University of China
Beijing, China*

Guosheng Yin

GYIN@HKU.HK

*Department of Statistics and Actuarial Science
University of Hong Kong
Hong Kong SAR, China*

Editor: Eric Laber

Abstract

We study the causal effect with a functional treatment variable, where practical applications often arise in neuroscience, biomedical sciences, etc. Previous research concerning the effect of a functional variable on an outcome is typically restricted to exploring correlation rather than causality. The generalized propensity score, which is often used to calibrate the selection bias, is not directly applicable to a functional treatment variable due to a lack of definition of probability density function for functional data. We propose three estimators for the average dose-response functional based on the functional linear model, namely, the functional stabilized weight estimator, the outcome regression estimator and the doubly robust estimator, each of which has its own merits. We study their theoretical properties, which are corroborated through extensive numerical experiments. A real data application on electroencephalography data and disease severity demonstrates the practical value of our methods.

Keywords: Causality, Double Robustness, Functional Data, Functional Linear Model, Treatment Effect

1. Introduction

In observational studies, a fundamental question is to investigate the causal effect of a treatment variable on an outcome. Much of the literature on estimating causal effects from observational data focus on a binary treatment variable (i.e., treatment versus control); see e.g., Rosenbaum and Rubin (1983); Hirano et al. (2003); Imai and Ratkovic (2014); Chan

*. Wei Huang and Zheng Zhang are corresponding authors.

et al. (2016); Athey et al. (2018); Ding et al. (2019); Tan (2020); Guo et al. (2021); Lin et al. (2023). Recently, there is a growing interest in studying more complex treatment variables, e.g., categorical (Schulte et al., 2014; Lopez and Gutman, 2017a,b; Chen et al., 2018; Li and Li, 2019; Luckett et al., 2019) or continuous treatment variables (Hirano and Imbens, 2004; Moodie et al., 2014; Galvao and Wang, 2015; Laber and Zhao, 2015; Kennedy et al., 2017; Fong et al., 2018; Kennedy, 2019; Ai et al., 2021; Bonvini and Kennedy, 2022).

Our interest is in estimating the average dose-response functional (ADRF) of a functional treatment variable. Functional data are collected repeatedly over a continuous domain, and are fundamentally different from scalar variables due to their infinite dimensionality (Wang et al., 2016). We focus on the causal relationship between a functional treatment variable and a scalar outcome variable in cross-sectional studies, where the treatment variable is real-valued over a continuous domain (e.g., space or frequency, rather than time) and the outcome variable does not vary with respect to that continuous domain. In particular, our work is motivated by an electroencephalography (EEG) dataset (Ciarleglio et al., 2022), which was collected via a standard headcap with multiple scalp electrodes. The main purpose there is to assess the EEG data as a potential moderator of treatment effect in a clinical trial. Here, we investigate possible causation between the EEG data and the severity of major depressive disorder. A random subsample of 20 frontal asymmetric curves, measuring the difference between the intensities of the human neuronal activities from two electrodes (right and left) placed on the scalp, is shown in Figure 1. For additional applications, see Morris (2015). As an illustration, we may be interested in the causal relationship between spectrometric data and fat content of certain types of food (Ferraty and Vieu, 2006) and the causal effect of body shape represented by the circumference over a given range on human visceral adipose tissue (Zhang et al., 2021). These problems are distinct from sequential decision problems in which multiple treatments are assigned longitudinally (Robins, 2000; Robins et al., 2000; Moodie and Stephens, 2012; Zhao et al., 2015; Kennedy, 2019). In addition, functional data have been used as one of the covariates in estimating optimal treatment regimes (McKeague and Qian, 2014; Ciarleglio et al., 2015, 2016, 2018; Laber and Staicu, 2018; Li and Kosorok, 2023; Park et al., 2023).

To identify the causal effect in an observational study, certain assumptions are required. In addition to standard assumptions such as no interference, consistency and positivity, a routinely used condition for identification is the unconfoundedness assumption (Rosenbaum and Rubin, 1983), i.e., the set of potential outcomes is conditionally independent of the treatment given the observed confounders. For a continuous scalar treatment, under the unconfoundedness assumption, the generalized propensity score or the stabilized weight (Hirano and Imbens, 2004; Imai and van Dyk, 2004; Galvao and Wang, 2015; Kennedy et al., 2017; Ai et al., 2021; Bonvini and Kennedy, 2022) has been widely used to calibrate the selection bias via estimating equations. It relies upon the (conditional) density of the treatment variable. However, the probability density function of functional data generally does not exist (Delaigle and Hall, 2010). Therefore, in our context, neither the generalized propensity score nor the stabilized weight are well defined, and thus traditional methods are not directly applicable. To circumvent this issue, Zhang et al. (2021) proposed to approximate the functional treatment variable based on functional principal component (PC) analysis. They then defined a stabilized weight function by the (conditional) probability density of the PC scores to adjust for the selection bias. However, as the number of PCs

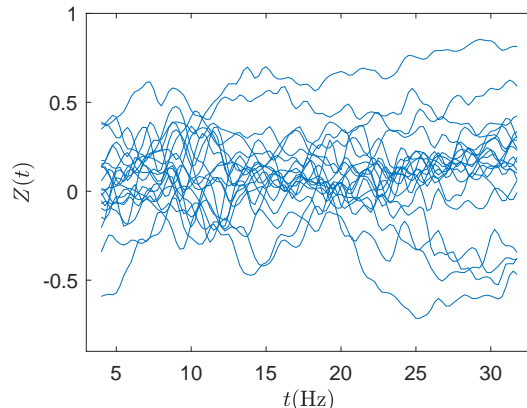


Figure 1: A random subsample of 20 frontal asymmetric curves $Z(\cdot)$ on a given frequency domain from the EEG dataset.

approaches infinity, this stabilized weight becomes ill-defined, and the consistency of their estimator is not guaranteed.

Under the unconfoundedness assumption in the context of a functional treatment variable, we propose three new estimation approaches to the ADRF: a functional stabilized weight (FSW) method, a partially linear outcome regression (OR) method, and a doubly robust (DR) approach combining the first two. The first method assumes a functional linear model for the ADRF and incorporates a new FSW. Our FSW is well-defined and guarantees the identification of the ADRF for any functional treatment variable. We propose a novel, consistent, nonparametric FSW estimator and estimate the ADRF using functional linear regression with a weighted outcome. The OR method assumes a more restrictive additive model that considers confounding variables. It quantifies the ADRF by the expectation of the parametric model with respect only to the confounding variables, which is computed through a backfitting algorithm. Our DR estimator is consistent if either of the first two estimators is consistent. If the partially linear model is correctly specified, the DR estimator converges at the same fast rate as the OR estimator. Our double robustness is different from those in the literature on a continuous scalar treatment variable (see e.g., Kennedy et al., 2017; Bonvini and Kennedy, 2022), where the estimator is consistent if one of two models for the conditional outcome and the generalized propensity score is correctly specified. In our case, the functional linear model for the ADRF is required for the consistency of all three estimators, while the FSW, serving the role of the generalized propensity score, is estimated nonparametrically.

More recently, Wang et al. (2023) proposed a so-called weight-modified kernel ridge regression (WMKRR) estimator for the ADRF. They assumed that the ADRF lies in a functional reproducing kernel Hilbert space (RKHS) and utilized the kernel ridge regression to estimate the ADRF. Their method also requires an estimation of the FSW. To this end, they proposed to minimize a novel uniform balancing error derived from their estimation of the ADRF. Although the kernel ridge regression is more flexible than the functional linear model that is used in our methods, it still faces the risk of model misspecification since the

model depends on the choice of the kernel. We will compare our methods with those from Zhang et al. (2021) and Wang et al. (2023) in simulation experiments.

The rest of the paper is organized as follows. We introduce the basic setup for the identification of ADRF in Section 2, followed by the development of our three estimators in Section 3. We investigate the convergence rates of the estimators in Section 4. The selection of tuning parameters is discussed in Section 5. Simulation experiments and the real data analysis are presented in Sections 6 and 7, respectively. We conclude in Section 8. Proofs and additional technical details are provided in the Appendix.

2. Identification of the Average Dose-Response Functional

We consider a functional treatment variable such that the observed treatment variable, denoted by $Z = Z(\cdot) : \mathcal{T} \rightarrow \mathbb{R}$, is a smooth real-valued random function defined on a compact interval \mathcal{T} , with $E\{\int_{\mathcal{T}} Z^2(t) dt\} < \infty$. Let $Y^*(z)$ be the potential outcome given the treatment $Z = z$ for $z \in L^2(\mathcal{T})$, the Hilbert space of squared integrable functions on \mathcal{T} . Let $X = (X_1, \dots, X_p)^\top \in \mathbb{R}^p$ be the p -dimensional observable confounding covariates that are related to both $Y^*(z)$ and Z .

For each individual, we only observe a confounding vector X , a random treatment Z and the corresponding outcome $Y = Y^*(Z) \in \mathbb{R}$. We further assume that the functional variable Z is fully observed. Essentially, our methodology can be applied when Z is only sparsely observed as long as the trajectory of Z can be well reconstructed, while the development of asymptotic properties in such cases would be much more challenging (Zhang and Wang, 2016; Zhou et al., 2023).

We are interested in estimating the average dose-response functional (ADRF), $E\{Y^*(z)\}$, for any $z \in L^2(\mathcal{T})$. Because we never observe $Y^*(z)$ simultaneously for all $z \in L^2(\mathcal{T})$, to identify $E\{Y^*(z)\}$ from the observed data, we impose the following identification assumptions.

Assumption 1

- (i) Unconfoundedness: Given X , Z is independent of $Y^*(z)$ for all $z \in L^2(\mathcal{T})$, i.e., $\{Y^*(z)\}_{z \in L^2(\mathcal{T})} \perp Z \mid X$.
- (ii) No interference: There is no interference among the units, i.e., each individual's outcome depends only on their own treatment.
- (iii) Consistency: $Y = Y^*(z)$ a.s. if $Z = z$.
- (iv) Positivity: The conditional density $f_{X|Z}$ satisfies $f_{X|Z}(X|Z) > 0$ a.s.

Assumption 1 is a natural extension of the classical identification assumption in the literature on the scalar treatment effect to the context of a functional treatment variable. In particular, we focus on cross-sectional studies where the interval \mathcal{T} does not represent a time interval and there is no dynamic confounding effect. To circumvent the problem of the nonexistent probability density function of functional data, the positivity condition (iv) is imposed on $f_{X|Z}$ instead of $f_{Z|X}$.

Under Assumption 1, the ADRF can be identified from the observable variables (X, Y, Z) in the three ways as follows. First, it is easy to verify

$$E\{Y^*(z)\} = E[E\{Y^*(z)|X\}] = E[E\{Y^*(z)|X, Z = z\}] = E\{E(Y|X, Z = z)\}. \quad (1)$$

Second, we can further write

$$\begin{aligned} E\{Y^*(z)\} &= E\{E(Y|X, Z = z)\} \\ &= \int_{\mathbb{R}^p} E(Y|X = x, Z = z) f_X(x) dx \\ &= \int_{\mathbb{R}^p} E(Y|X = x, Z = z) \frac{f_X(x)}{f_{X|Z}(x|z)} f_{X|Z}(x|z) dx \\ &= E\{\pi_0(Z, X) Y | Z = z\}, \end{aligned} \quad (2)$$

where $\pi_0(Z, X) = f_X(X)/f_{X|Z}(X|Z)$ is called the functional stabilized weight (FSW). If Z is a scalar random variable, we have $f_X(X)/f_{X|Z}(X|Z) = f_Z(Z)/f_{Z|X}(Z|X)$, where the right-hand side is the classical stabilized weight.

Third, let E_X denote the expectation with respect to X . Noting that

$$\begin{aligned} E[E_X\{E(Y|X, Z)\}|Z = z] &= E\{E(Y|X, Z = z)\}, \\ E\{E(Y|X, Z)\pi_0(X, Z)|Z = z\} &= E\{Y\pi_0(X, Z)|Z = z\}, \end{aligned}$$

together with (1) and (2), we can also identify $E\{Y^*(z)\}$ as

$$E\{Y^*(z)\} = E[\{Y - E(Y|X, Z)\}\pi_0(X, Z) + E_X\{E(Y|X, Z)\}|Z = z], \quad (3)$$

because the second and third terms above cancel out.

Based on the three identification strategies in (1), (2), and (3) respectively, we develop three estimators of $E\{Y^*(z)\}$. Each estimator is based on different model assumptions and has its own merits.

3. Models and Estimation Methods

For all three methods, we adopt a functional linear model on the ADRF,

$$E\{Y^*(z)\} = a + \int_{\mathcal{T}} b(t)z(t) dt, \quad (4)$$

where a is a scalar intercept, and $b \in L^2(\mathcal{T})$ is a slope function. The functional linear model is widely used in scalar-on-function regression, whereas the model here is assumed for the potential outcome $Y^*(z)$. Essentially, model (4) can be replaced by any other nonlinear regression models (e.g., Müller and Yao, 2008; Li et al., 2010; Wang et al., 2023) without affecting our main idea. We choose to use the functional linear model mainly due to its simplicity and interpretability. The slope function $b = b(\cdot)$ is a quantity of primary interest convenient for the causal inference because it measures the intensity of the functional treatment effect. For example, to compute the average treatment effect (ATE) between $z_1, z_2 \in L^2(\mathcal{T})$, $E\{Y^*(z_1) - Y^*(z_2)\}$, it suffices to compute $\int_{\mathcal{T}} b(t)\{z_1(t) - z_2(t)\} dt$. Additionally, the slope function provides useful causal information (e.g., a positive/negative causal relationship), even if the model is misspecified.

3.1 Functional Stabilized Weight Estimator

We construct our functional stabilized weight (FSW) estimator of the ADRF based on equation (2) and model (4). In this construction, we propose a nonparametric estimator for the FSW. In particular, equation (2) suggests that we can estimate $E\{Y^*(z)\}$ in (4) using the classical functional linear regression technique (e.g., Hall and Horowitz, 2007). Here, the response variable Y is replaced by $\hat{\pi}(Z, X)Y$, provided an estimator of π_0 , $\hat{\pi}$. To facilitate the development, suppose that an estimator $\hat{\pi}$ of π_0 is available for the time being. Let $G(t, s) = \text{Cov}\{Z(t), Z(s)\}$ be the covariance function of Z . Its eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq 0$ and the corresponding orthonormal eigenfunctions $\{\phi_j\}_{j=1}^\infty$ are defined as follows,

$$\int_{\mathcal{T}} G(t, s) \phi_j(s) ds = \lambda_j \phi_j(t), \text{ for all } t \in \mathcal{T}, \quad (5)$$

where $\{\phi_j\}_{j=1}^\infty$, also called the PC basis, is a complete basis of $L^2(\mathcal{T})$. According to the Karhunen–Loève expansion, we have

$$Z(t) = \mu_Z(t) + \sum_{j=1}^{\infty} \xi_j \phi_j(t), \quad (6)$$

where $\mu_Z(t) = E\{Z(t)\}$ and $\xi_j = \int_{\mathcal{T}} \{Z(t) - \mu_Z(t)\} \phi_j(t) dt$.

With a slight abuse of notation, let $G : L^2(\mathcal{T}) \rightarrow L^2(\mathcal{T})$ denote a linear operator such that $(G \circ \psi)(t) = \int_{\mathcal{T}} G(t, s) \psi(s) ds$, for a function $\psi \in L^2(\mathcal{T})$. If we let $\mu_{Y, \pi} = E\{\pi_0(Z, X)Y\} = a + \int_{\mathcal{T}} b(t) \mu_Z(t) dt$, by combining (2), (4) and (6), we can write

$$E\{\pi_0(Z, X)Y|Z = z\} - \mu_{Y, \pi} = \int_{\mathcal{T}} b(t) \{z(t) - \mu_Z(t)\} dt.$$

It follows that $(G \circ b)(t) = e(t)$, where $e(t) = E[\{\pi_0(Z, X)Y - \mu_{Y, \pi}\}\{Z(t) - \mu_Z(t)\}]$. Noting that $\{\phi_j\}_{j=1}^\infty$ is an orthonormal complete basis of $L^2(\mathcal{T})$, we can write $b(t) = \sum_{j=1}^\infty b_j \phi_j(t)$ and $e(t) = \sum_{j=1}^\infty e_j \phi_j(t)$, and combining them with (5), we obtain that, for all j ,

$$b_j = \int_{\mathcal{T}} b(t) \phi_j(t) dt, \text{ and } e_j = \int_{\mathcal{T}} e(t) \phi_j(t) dt = \int_{\mathcal{T}} (G \circ b)(t) \phi_j(t) dt = b_j \lambda_j.$$

The above arguments suggest we can estimate $b(t)$ by $\hat{b}_{\text{FSW}}(t) = \sum_{j=1}^q \hat{b}_{\text{FSW}, j} \hat{\phi}_j(t)$, where q is a truncation parameter and $\hat{b}_{\text{FSW}, j} = \hat{\lambda}_j^{-1} \hat{e}_{\text{FSW}, j}$ is an estimator of b_j . The $\hat{\lambda}_j$'s and $\hat{\phi}_j$'s are the eigenvalues and eigenfunctions of the empirical covariance function $\hat{G}(s, t) = \sum_{i=1}^n \{Z_i(s) - \hat{\mu}_Z(s)\} \{Z_i(t) - \hat{\mu}_Z(t)\} / n$, respectively, where $\hat{\mu}_Z(t) = \sum_{i=1}^n Z_i(t) / n$ is the empirical estimator of $\mu_Z(t)$. Moreover, by the definition of $e(t)$, we can estimate it by

$$\hat{e}_{\text{FSW}}(t) = \sum_{i=1}^n \frac{\{Y_i \hat{\pi}(Z_i, X_i) - \hat{\mu}_{Y, \pi}\} \{Z_i(t) - \hat{\mu}_Z(t)\}}{n}$$

with $\hat{\mu}_{Y, \pi} = \sum_{i=1}^n Y_i \hat{\pi}(Z_i, X_i) / n$, and take $\hat{e}_{\text{FSW}, j} = \int_{\mathcal{T}} \hat{e}_{\text{FSW}}(t) \hat{\phi}_j(t) dt$ as an estimator of e_j . Finally, the estimator of the intercept a is given by $\hat{a}_{\text{FSW}} = \hat{\mu}_{Y, \pi} - \int_{\mathcal{T}} \hat{b}_{\text{FSW}}(t) \hat{\mu}_Z(t) dt$.

It remains to estimate the FSW, $\pi_0(Z, X) = f_X(X)/f_{X|Z}(X|Z)$. A naive approach would first estimate the densities f_X and $f_{X|Z}$ and then take the ratio. However, this may lead to unstable estimates because an estimator of the denominator $f_{X|Z}$ is difficult to derive and may be too close to zero. Also, approximating Z by its PC scores would not alleviate the challenge but pose additional difficulty in theoretical justification. To circumvent these problems, we treat π_0 as a whole and develop a robust nonparametric estimator. We can find the conditional moments that identify $\pi_0(z, \cdot)$ for every fixed z ,

$$E\{\pi_0(z, X)v(X)|Z = z\} = \int_{\mathbb{R}^p} \frac{f_X(x)}{f_{X|Z}(x|z)} v(x) f_{X|Z}(x|z) dx = E\{v(X)\}, \quad (7)$$

which holds for any integrable function $v(X)$.

Proposition 1 *For any fixed $z \in L^2(\mathcal{T})$,*

$$E\{\pi(z, X)v(X)|Z = z\} = E\{v(X)\},$$

holds for all integrable functions $v(X)$ if and only if $\pi(z, X) = \pi_0(z, X)$ a.s.

The proof is given in Appendix A. This result indicates that π_0 is fully characterized by (7), and thus we can use it to define our estimator. Since (7) is defined for a fixed z , we need to estimate π_0 at all sample points Z_i , where (X_i, Y_i, Z_i) , for $i = 1, \dots, n$, are i.i.d. copies of (X, Y, Z) .

We consider the leave-one-out index set $\mathcal{S}_{-i} = \{j : j \neq i, j = 1, \dots, n\}$ to estimate π_0 at each Z_i . Specifically, for a given sample point $z = Z_i$, the right-hand side of (7) can be estimated by its empirical version $\sum_{j \in \mathcal{S}_{-i}} v(X_j)/(n-1)$. We propose to estimate the left-hand side of (7) by the Nadaraya–Watson estimator for a functional covariate (Ferraty and Vieu, 2006),

$$\frac{\sum_{j \in \mathcal{S}_{-i}} \pi_0(Z_i, X_j) v(X_j) K\{d(Z_j, Z_i)/h\}}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, Z_i)/h\}} = \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} v(X_j),$$

where $d(\cdot, \cdot)$ denotes a semi-metric in $L^2(\mathcal{T})$, $h > 0$ is a bandwidth and K is a kernel function quantifying the proximity of Z_j and Z_i . Commonly used choices for d include the L^2 norm and the projection metric $d(z_1, z_2) = [\int_{\mathcal{T}} \{\Pi(z_1) - \Pi(z_2)\}^2 dt]^{1/2}$, where Π denotes a projection operation to a subspace of $L^2(\mathcal{T})$.

As it is not possible to solve (7) for an infinite number of v 's from a given finite sample, we use a growing number of basis functions to approximate any suitable function v . Specifically, let $\nu_k(X) = (v_1(X), \dots, v_k(X))^{\top}$ be a set of known basis functions, which may serve as a finite-dimensional sieve approximation to the original function space of all integrable functions. We define our estimator of $\pi_0(Z_i, X_j)$, for $j \in \mathcal{S}_{-i}$, as the solution to the k equations,

$$\frac{\sum_{j \in \mathcal{S}_{-i}} \pi(Z_i, X_j) \nu_k(X_j) K\{d(Z_j, Z_i)/h\}}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, Z_i)/h\}} = \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j), \quad (8)$$

which asymptotically identifies π_0 as $k \rightarrow \infty$ and $h \rightarrow 0$. However, in practice, a finite number k of equations in (8) cannot identify a unique solution. Indeed, for any strictly

increasing and concave function ρ , replacing $\pi(Z_i, X_j)$ by $\rho'\{\hat{\eta}_{Z_i}^\top \nu_k(X_j)\}$ would satisfy (8). Here, $\rho'(\cdot)$ is the first derivative of $\rho(\cdot)$, and $\hat{\eta}_{Z_i}$ is a k -dimensional vector that maximizes the strictly concave function H_{hk, Z_i} , defined as

$$H_{hk, Z_i}(\eta) = \frac{\sum_{j \in \mathcal{S}_{-i}} \rho\{\eta^\top \nu_k(X_j)\} K\{d(Z_j, Z_i)/h\}}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, Z_i)/h\}} - \eta^\top \left\{ \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j) \right\}. \quad (9)$$

This can be verified by noting that the gradient $\nabla H_{hk, Z_i}(\hat{\eta}_{Z_i}) = 0$ corresponds to (8) with $\pi(Z_i, X_j)$ replaced by $\rho'\{\hat{\eta}_{Z_i}^\top \nu_k(X_j)\}$. Therefore, for a given strictly increasing and concave function ρ together with $\hat{\eta}_{Z_i}$ defined above, we define our estimator of $\pi_0(Z_i, X_j)$ as $\hat{\pi}_{hk}(Z_i, X_j) = \rho'\{\hat{\eta}_{Z_i}^\top \nu_k(X_j)\}$. This also induces the estimator $\hat{\pi}_{hk}(Z_i, X_i) = \rho'\{\hat{\eta}_{Z_i}^\top \nu_k(X_i)\}$ with X_j replaced by X_i .

Our estimator $\hat{\pi}_{hk}$ has a generalized empirical likelihood interpretation for a variety of choices for ρ . Specifically, we show in Appendix B that the estimator defined by (9) is the dual solution to minimizing some distance measure between the FSW and 1 locally for each Z_i , subject to the constraint (8) (i.e., the constraint (12) in Appendix B). That is, our estimator $\hat{\pi}_{hk}$ is the desired weight closest to the baseline uniform weight 1 locally in a small neighbourhood of each Z_i , subject to the finite sample version of the moment restriction in Proposition 1. The uniform weight can be considered as a baseline. This is because, when Z and $Y^*(z)$ are unconditionally independent for all $z \in L^2(\mathcal{T})$ (i.e., a randomized trial without confoundedness), the functional stabilized weights should be uniformly equal to 1. Different choices of ρ correspond to different distance measures (see Appendix B for other examples). We suggest to choose $\rho(v) = -\exp(-v - 1)$ which guarantees a positive weight estimator $\hat{\pi}_{hk}(Z_i, X_i)$.

3.2 Outcome Regression Estimator and Doubly Robust Estimator

To estimate the ADRF based on the outcome regression (OR) identification method in (1), an estimator of $E(Y|X, Z)$ is required. Thus, based on the functional linear model in (4), we further assume a partially linear additive model,

$$Y = a + \int_{\mathcal{T}} b(t)Z(t)dt + g(X; \theta) + \epsilon, \quad (10)$$

where a and $b(\cdot)$ are the same as those in (4), $g(\cdot; \theta)$ is a known function with an unknown parameter $\theta \in \mathbb{R}^p$ and ϵ is an error variable with $E(\epsilon|Z, X) = 0$ and $E(\epsilon^2|Z, X) < \infty$ uniformly for all Z and X . Note that X and Z are generally dependent. We mainly investigate the linear function $g(X; \theta) = \theta^\top X$, where X may involve some transformations, e.g., the log-transformation, polynomials, and exponential of some elements of X . Without loss of generality, we assume that $E(\theta^\top X) = 0$; otherwise the non-zero mean can be absorbed into a . Model (10) may be extended to include interactions between Z and X . However, research on functional models involving interactions between functional and multivariate continuous variables is scarce, even in the literature on non-causal functional regression. Such extension is beyond the scope of this paper and deserves future research.

Recalling $E\{Y^*(z)\} = E\{E(Y|X, Z = z)\}$ from (1), it is clear that (10) implies (4), and thus model (10) imposes a stronger structure than model (4). Specifically, the linear

part $\theta^\top X$ can be replaced by any nonparametric function of X whose mean is zero, which still implies model (4). However, model (10) gives a simpler estimator of $E\{Y^*(z)\}$ with a faster convergence rate and a better interpretability of the effects of the treatment and confounding variables on the outcome.

The OR estimators \hat{a}_{OR} , \hat{b}_{OR} and $\hat{\theta}_{\text{OR}}$ can be obtained by adapting the backfitting algorithm (Buja et al., 1989). In the first step, we set θ to be zero and apply the same method as in Section 3.1 to regress Y on Z , which gives the initial estimators of a and b in (10). Specifically, we estimate $b(t)$ initially by $\hat{b}_{\text{ini}}(t) = \sum_{j=1}^q \hat{b}_{\text{ini},j} \hat{\phi}_j(t)$ with $\hat{b}_{\text{ini},j} = \hat{\lambda}_j^{-1} \hat{e}_{\text{ini},j}$, where $\hat{\lambda}_j$'s and $\hat{\phi}_j$'s are the same as those in Section 3.1, while $\hat{e}_{\text{ini}}(t) = \sum_{i=1}^n (Y_i - \hat{\mu}_Y) \{Z_i(t) - \hat{\mu}_Z(t)\} / n$ with $\hat{\mu}_Y = \sum_{i=1}^n Y_i / n$. We then define $\hat{e}_{\text{ini},j} = \int_{\mathcal{T}} \hat{e}_{\text{ini}}(t) \hat{\phi}_j(t) dt$. Finally, the initial estimator \hat{a}_{ini} of a is given by $\hat{\mu}_Y - \int_{\mathcal{T}} \hat{b}_{\text{ini}}(t) \hat{\mu}_Z(t) dt$.

In the second step, we use the traditional least square method to regress the residual $Y_i - \hat{a}_{\text{ini}} - \int_{\mathcal{T}} \hat{b}_{\text{ini}}(t) Z_i(t) dt$ on X_i , for $i = 1, \dots, n$. Specifically, we estimate θ by $\hat{\theta}_{\text{ini}} = (\Xi^\top \Xi)^{-1} \Xi^\top \mathbb{Y}_{\text{res}}$, where Ξ is an $n \times p$ matrix with the (i, j) -th entry being X_{ij} , the j -th element of X_i , and

$$\mathbb{Y}_{\text{res}} = \left\{ Y_1 - \hat{a}_{\text{ini}} - \int_{\mathcal{T}} \hat{b}_{\text{ini}}(t) Z_1(t) dt, \dots, Y_n - \hat{a}_{\text{ini}} - \int_{\mathcal{T}} \hat{b}_{\text{ini}}(t) Z_n(t) dt \right\}^\top.$$

Subsequently, we can repeat the first step with Y_i replaced by $Y_i - \hat{\theta}_{\text{ini}}^\top X_i$ to update the estimators of \hat{a}_{ini} and \hat{b}_{ini} . This procedure is iterated until the outcome regression (OR) estimators of a , b and θ , denoted by \hat{a}_{OR} , \hat{b}_{OR} and $\hat{\theta}_{\text{OR}}$, are stabilized. More specifically, let $\hat{a}_{\text{OR}}^{(t)}$, $\hat{b}_{\text{OR}}^{(t)}$ and $\hat{\theta}_{\text{OR}}^{(t)}$ denote respectively the estimators of a, b and θ at the t -th iteration. We set our final estimators as $\hat{a}_{\text{OR}} = \hat{a}_{\text{OR}}^{(t+1)}$, $\hat{b}_{\text{OR}} = \hat{b}_{\text{OR}}^{(t+1)}$ and $\hat{\theta}_{\text{OR}} = \hat{\theta}_{\text{OR}}^{(t+1)}$ if $\{|\hat{a}_{\text{OR}}^{(t+1)} - \hat{a}_{\text{OR}}^{(t)}|^2 + \|\hat{b}_{\text{OR}}^{(t+1)} - \hat{b}_{\text{OR}}^{(t)}\|_{L^2}^2 + \|\hat{\theta}_{\text{OR}}^{(t+1)} - \hat{\theta}_{\text{OR}}^{(t)}\|_{\mathbb{R}^p}^2\}^{1/2} / \{|\hat{a}_{\text{OR}}^{(t)}|^2 + \|\hat{b}_{\text{OR}}^{(t)}\|_{L^2}^2 + \|\hat{\theta}_{\text{OR}}^{(t)}\|_{\mathbb{R}^p}^2\}^{1/2} < 0.05$ or t reaches 100 (a prespecified maximum number of iterations). Here, $\|\cdot\|_{L^2}$ and $\|\cdot\|_{\mathbb{R}^p}$ denote the L^2 norm in $L^2(\mathcal{T})$ and the Euclidean norm in \mathbb{R}^p , respectively.

The OR estimation is easy to implement but requires strong parametric assumptions, while the FSW estimator requires fewer modelling assumptions but is subject to a slow convergence rate. Thus, it is desirable to develop an estimator that possesses more attractive properties by combining these two. Note that (3) satisfies the so-called doubly robust (DR) property: for two generic functions $\tilde{E}(Y|X, Z)$ and $\tilde{\pi}(X, Z)$, it holds that

$$E\{Y^*(z)\} = E[\{Y - \tilde{E}(Y|X, Z)\} \tilde{\pi}(X, Z) + E_X\{\tilde{E}(Y|X, Z)\} | Z = z], \quad (11)$$

if either $\tilde{E}(Y|X, Z) = E(Y|X, Z)$ or $\tilde{\pi}(X, Z) = \pi_0(X, Z)$ but not necessarily both are satisfied; see Appendix C for the derivation.

Under model (4), (11) equals $a + \int_{\mathcal{T}} b(t) z(t) dt$. To estimate a and b , it suffices to conduct the functional linear regression as in Section 3.1 to regress an estimator of $\{Y - E(Y|X, Z)\} \pi(X, Z) + E_X\{E(Y|X, Z)\}$ on Z . Specifically, we estimate $E(Y|X, Z)$ by the OR estimator, $\hat{E}(Y|X, Z) = \hat{a}_{\text{OR}} + \int_{\mathcal{T}} \hat{b}_{\text{OR}}(t) Z(t) dt + \hat{\theta}_{\text{OR}}^\top X$, and $\pi_0(X, Z)$ by $\hat{\pi}_{hk}(X, Z)$. We define the DR estimator of b as $\hat{b}_{\text{DR}}(t) = \sum_{j=1}^q \hat{b}_{\text{DR},j} \hat{\phi}_j(t)$, where $\hat{b}_{\text{DR},j} = \hat{\lambda}_j^{-1} \hat{e}_{\text{DR},j}$ with $\hat{e}_{\text{DR},j} = \int_{\mathcal{T}} \hat{e}_{\text{DR}}(t) \hat{\phi}_j(t) dt$. The expression of $\hat{e}_{\text{DR}}(t)$ is

$$\frac{1}{n} \sum_{i=1}^n \left[\{Y_i - \hat{E}(Y|X_i, Z_i)\} \hat{\pi}_{hk}(X_i, Z_i) + \frac{1}{n} \sum_{j=1}^n \{\hat{E}(Y|X_j, Z_j)\} - \hat{\mu}_{Y, \text{DR}} \right] \{Z_i(t) - \hat{\mu}_Z(t)\},$$

where

$$\hat{\mu}_{Y,DR} = \frac{1}{n} \sum_{i=1}^n \left[\{Y_i - \hat{E}(Y|X_i, Z_i)\} \hat{\pi}_{hk}(X_i, Z_i) + \frac{1}{n} \sum_{j=1}^n \{\hat{E}(Y|X_j, Z_j)\} \right].$$

The DR estimator of a is defined as $\hat{a}_{DR} = \hat{\mu}_{Y,DR} - \int_{\mathcal{T}} \hat{b}_{DR} \hat{\mu}_Z(t) dt$.

Provided that model (4) is correctly specified, according to the DR property (11), the DR estimator $\hat{E}_{DR}\{Y^*(z)\} = \hat{a}_{DR} + \int_{\mathcal{T}} \hat{b}_{DR}(t) z(t) dt$ is consistent if either $\hat{E}(Y|X, Z)$ or $\hat{\pi}_{hk}(X, Z)$ is consistent. The consistency of $\hat{E}(Y|X, Z)$ mainly depends on the correct specification of the partially linear model (10), while the consistency of $\hat{\pi}_{hk}(X, Z)$, as a nonparametric estimator, relies on less restrictive (but more technical) assumptions.

4. Asymptotic Properties

We investigate the asymptotic convergence rates of the proposed three estimators, while assuming that Z is fully observed. Under the main model is $E\{Y^*(z)\} = a + \int_{\mathcal{T}} b(t) z(t) dt$ in (4), although the estimators of a are based on those of b , they are essentially empirical means, which can achieve the $n^{-1/2}$ convergence rate (Shin, 2009). Thus, our primary focus is to derive the convergence rates of our estimators of b , which, depending on the smoothness of Z and b , are slower than those in the finite-dimensional settings.

Let $\mathcal{X} \subset \mathbb{R}^p$ denote the support of X and $\mathcal{Z} \subset L^2(\mathcal{T})$ the support of Z . To derive the convergence rate of \hat{b}_{FSW} , we first provide the uniform convergence rate of the estimator $\hat{\pi}_{hk}$ of the FSW π_0 over \mathcal{X} and \mathcal{Z} , which requires the following conditions.

Condition A

- (A1) The set \mathcal{X} is compact. For all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$, $\pi_0(z, x)$ is strictly bounded away from zero and infinity.
- (A2) For all $z \in \mathcal{Z}$, $\exists \eta_z \in \mathbb{R}^k$ and a constant $\alpha > 0$ such that $\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\rho'^{(-1)}\{\pi_0(z, x)\} - \eta_z^\top \nu_k(x)| = O(k^{-\alpha})$, where $\rho'^{(-1)}$ is the inverse function of ρ' .
- (A3) The eigenvalues of $E\{\nu_k(X) \nu_k(X)^\top | Z = z\}$ are bounded away from zero and infinity uniformly in k and z . There exists a sequence of constants $\zeta(k)$ satisfying $\sup_{x \in \mathcal{X}} \|\nu_k(x)\| \leq \zeta(k)$.
- (A4) There exists a continuously differentiable function $\mu(h) > 0$, for $h > 0$, such that for all $z \in \mathcal{Z}$, $P\{d(Z, z) \leq h\} / \mu(h)$ is strictly bounded away from zero and infinity. There exists $h_0 > 0$ such that $\sup_{h \in (0, h_0)} \mu'(h)$ is bounded.
- (A5) The kernel $K : \mathbb{R} \rightarrow \mathbb{R}^+$ is bounded, Lipschitz and supported on $[0, 1]$ with $K(1) > 0$.
- (A6) For $g_{k_1, k_2, z}(X) = \rho'\{\eta_z^\top \nu_k(X)\} v_{k_1}(X)$ and $v_{k_1}(X) v_{k_2}(X)$ with $k_1, k_2 = 1, \dots, k$, there exists $\lambda > 0$ such that for all $m \geq 2$, $|E[g_{k_1, k_2, z_1}(X) | Z = z_1] - E[g_{k_1, k_2, z_2}(X) | Z = z_2]| / d^\lambda(z_1, z_2)$ and $E[|g_{k_1, k_2, z_1}(X)|^m | Z = z_1]$ are uniformly bounded over k, z_1 and z_2 .
- (A7) Let $\psi_{\mathcal{Z}}(\epsilon)$ be the Kolmogorov ϵ -entropy of \mathcal{Z} , i.e., $\psi_{\mathcal{Z}}(\epsilon) = \log\{N_\epsilon(\mathcal{Z})\}$, where $N_\epsilon(\mathcal{Z})$ is the minimal number of open balls in $L^2(\mathcal{T})$ of radius ϵ (with the semi-metric d)

covering \mathcal{Z} . It satisfies $\sum_{n=1}^{\infty} \exp[-(\delta - 1)\psi_{\mathcal{Z}}\{(\log n)/n\}] < \infty$ for some $\delta > 1$, and $(\log n)^2/\{n\mu(h)\} < \psi_{\mathcal{Z}}\{(\log n)/n\} < n\mu(h)/\log n$ for n large enough.

Condition (A1) requires that X is compactly supported, which can be relaxed by restricting the tail behaviour of the distribution of X at the cost of more tedious technical arguments. The boundedness of π_0 in Condition (A1) (or some equivalent condition) is also commonly required in the literature (Kennedy et al., 2017; Ai et al., 2021; D’Amour et al., 2021). This condition can be relaxed if other smoothness assumptions are made on the potential outcome distribution (Ma and Wang, 2020). Condition (A2) essentially assumes that the convergence rate of the sieve approximation $\eta_z^\top \nu_k(x)$ for $\rho^{(-1)}\{\pi_0(z, x)\}$ is polynomial. This can be satisfied, for example, with $\alpha = +\infty$ if X is discrete, and with $\alpha = s/r$ if X has r continuous components and ν_k is a power series, where s is the degree of smoothness of $\rho^{(-1)}\{\pi_0(z, \cdot)\}$ with respect to the continuous components in X for all $z \in L^2(\mathcal{T})$ (Chen, 2007). Further, Condition (A3) ensures that the sieve approximation conditional on a functional variable is not degenerate. Similar conditions are routinely assumed in the literature of sieve approximation (Newey, 1997). Conditions (A4) to (A6) are standard in functional nonparametric regression (Ferraty and Vieu, 2006; Ferraty et al., 2010). In particular, the function $\mu(h)$ in Condition (A4), also referred to as the small ball probability, has been studied extensively in the literature (Li and Shao, 2001; Ferraty and Vieu, 2006); Condition (A5) requires that K is compactly supported, which is not satisfied for the Gaussian kernel but convenient for technical arguments; the variable $g_{k_1, k_2, z}(X)$ in Condition (A6) serves as the role of the response variable in the traditional regression setting. Condition (A7) is less standard: it regularizes the support \mathcal{Z} by controlling its Kolmogorov entropy, which is used to establish the uniform convergence rate on \mathcal{Z} . This condition is satisfied, for example, if \mathcal{Z} is a compact set, $d(\cdot, \cdot)$ is a projection metric and $(\log n)^2 = O\{n\mu(h)\}$; see Ferraty et al. (2010) for other examples.

Theorem 1 *Under Conditions (A1) to (A7), assuming that $k \rightarrow \infty$ and $h \rightarrow 0$ as $n \rightarrow \infty$, we have*

$$\sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} |\hat{\pi}_{hk}(z, x) - \pi_0(z, x)| = O\left(\zeta(k) \left[k^{-\alpha} + h^\lambda \sqrt{k} + \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}k}{n\mu(h)}} \right]\right)$$

holds almost surely.

Theorem 1 provides the sup-norm convergence rate of $\hat{\pi}_{hk}$ over the support of X and Z . The proof is given in Appendix D. The term $\zeta(k)(k^{-\alpha} + h^\lambda \sqrt{k})$ and the term $\zeta(k)\sqrt{\psi_{\mathcal{Z}}\{(\log n)/n\}k/\{n\mu(h)\}}$ can be viewed as the convergence rates of the bias and standard deviation, respectively.

To provide the convergence rate of \hat{b}_{FSW} , we recall that $E\{\pi_0(Z, X)Y|Z = z\} = a + \int_{\mathcal{T}} b(t)z(t)dt$ under model (4). Let $\epsilon_\pi = \pi_0(Z, X)Y - a - \int_{\mathcal{T}} b(t)Z(t)dt$ be the residual variable. Let $\gamma > 1$ and $\beta > \gamma/2 + 1$ be two constants and the c_j ’s be some generic positive constants.

Condition B

(B1) The variable ϵ_π has a finite fourth moment, i.e., $E(\epsilon_\pi^4) < \infty$.

- (B2) The functional variable Z has a finite fourth moment, i.e., $\int_{\mathcal{T}} E\{Z^4(t)\} dt < \infty$. The PC score ξ_j defined in (6) satisfies $E(\xi_j^4) \leq c_1 \lambda_j^2$, and the eigenvalue λ_j defined in (5) satisfies $\lambda_j - \lambda_{j+1} \geq c_2^{-1} j^{-\gamma-1}$ for all $j \geq 1$.
- (B3) The coefficient b_j satisfies $|b_j| \leq c_3 j^{-\beta}$, for all $j \geq 1$.

Condition (B1) makes a mild restriction on the moment of ϵ_π . Condition (B2) imposes regularity conditions on the random process Z , which requires that the differences of the adjacent eigenvalues do not decay too fast. Condition (B3) assumes an upper bound for the decay rate of b_j , the coefficients of b projected on $\{\phi_j\}_{j=1}^\infty$. The latter two conditions are adopted from Hall and Horowitz (2007) for technical purposes.

Theorem 2 *Under Conditions (A1) to (A7), (B1) to (B3) and model (4), assuming that $k \rightarrow \infty$ and $h \rightarrow 0$ as $n \rightarrow \infty$ and choosing the truncation parameter $q \asymp n^{1/(\gamma+2\beta)}$, we have*

$$\int_{\mathcal{T}} \{\widehat{b}_{\text{FSW}}(t) - b(t)\}^2 dt = O_p \left(\left[k^{-2\alpha} + h^{2\lambda} k + \frac{\psi_Z\{(\log n)/n\}k}{n\mu(h)} \right] \zeta^2(k) n^{(\gamma+1)/(\gamma+2\beta)} \right).$$

The proof of Theorem 2 is given in Appendix E. The convergence rate of $\int_{\mathcal{T}} \{\widehat{b}_{\text{FSW}}(t) - b(t)\}^2 dt$ in Theorem 2 can be quite slow mainly because of the small ball probability function $\mu(h)$. In the case of Gaussian process endowed with a metric, $\mu(h)$ has an exponential decay rate of h (Li and Shao, 2001), which leads to a non-infinitesimal rate of $\int_{\mathcal{T}} \{\widehat{b}_{\text{FSW}}(t) - b(t)\}^2 dt$. However, $\mu(h)$ can be much larger by choosing a proper semi-metric d (Ferraty and Vieu, 2006; Ling and Vieu, 2018), so that the final convergence rate of $\int_{\mathcal{T}} \{\widehat{b}_{\text{FSW}}(t) - b(t)\}^2 dt$ reaches a polynomial decay of n . Another way to improve the convergence rate is to impose additional structure assumptions. For example, if we assume that the slope function b can be fully characterized by a finite number of PC basis functions, or the random process Z essentially lies on a finite dimensional (not necessarily Euclidean) space (Lin and Yao, 2021), then the truncation parameter q does not need to tend to infinity and the convergence rate of $\int_{\mathcal{T}} \{\widehat{b}_{\text{FSW}}(t) - b(t)\}^2 dt$ can be much faster. Without such assumptions, estimating b in an infinite dimensional space with a nonparametric estimator of π_0 leads to a cumbersome convergence rate.

The OR estimator \widehat{b}_{OR} relies on functional principal component analysis and the partially linear model (10). Additional conditions are needed to derive the convergence rate of \widehat{b}_{OR} .

Condition C

- (C1) The covariate X has finite fourth moment, i.e., $E(\|X\|^4) < \infty$. For $m = 1, \dots, p$ and $j \geq 1$, $|\text{Cov}\{X_m, \int_{\mathcal{T}} Z(t)\phi_j(t) dt\}| < c_4 j^{-\gamma-\beta}$.
- (C2) Let $g_m(t) = \sum_{j=1}^\infty \text{Cov}\{X_m, \int_{\mathcal{T}} Z(t)\phi_j(t) dt\} \phi_j(t) / \lambda_j$ and $u_{im} = X_{im} - E(X_m) - \int_{\mathcal{T}} g_m(t) \{Z_i(t) - \mu_Z(t)\} dt$. For $m = 1, \dots, p$, $E(u_{1m}^2 | Z_1) < \infty$, and $E(\mathbf{u}_1^\top \mathbf{u}_1)$ is positive definite with $\mathbf{u}_1 = (u_{11}, \dots, u_{1p})^\top$.

Conditions (C1) and (C2) are adopted from Shin (2009), which make mild assumptions on covariate X to ensure \sqrt{n} -consistency of the estimated coefficient $\widehat{\theta}_{\text{OR}}$.

Theorem 3 *Under Conditions (B2), (B3), (C1), (C2) and model (10), choosing the truncation parameter $q \asymp n^{1/(\gamma+2\beta)}$, we have*

$$\int_{\mathcal{T}} \{\widehat{b}_{\text{OR}}(t) - b(t)\}^2 dt = O_p(n^{(-2\beta+1)/(\gamma+2\beta)}).$$

Because \widehat{b}_{OR} is theoretically equivalent to the least square estimator proposed by Shin (2009) provided that the estimators of each component in the additive model are unique (Buja et al., 1989), the theorem above is a direct result from Theorem 3.2 in Shin (2009) and thus its proof is omitted. The convergence rate of \widehat{b}_{OR} is the same as that of the estimated slope function in the traditional functional linear regression (Hall and Horowitz, 2007), i.e., such rate remains unchanged despite of the additional estimation of θ . Compared with Theorem 2, the convergence rate here $O_p(n^{(-2\beta+1)/(\gamma+2\beta)}) = O_p(n^{-1}) \cdot O_p(n^{(\gamma+1)/(\gamma+2\beta)})$ is much faster. Specifically, the part of the convergence rate $O_p(\zeta^2(k)[k^{-2\alpha} + h^{2\lambda}k + \psi_{\mathcal{Z}}\{(\log n)/n\}k/\{n\mu(h)\}])$, which is the convergence rate of $\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\widehat{\pi}_{hk}(z, x) - \pi_0(z, x)|^2$, is replaced by a faster one, $O_p(n^{-1})$, because the nonparametric estimator $\widehat{\pi}_{hk}$ is not used.

According to the discussion at the end of Section 3.2, under model (4) and the assumptions of Theorem 1, \widehat{b}_{DR} is consistent. If assumptions of Theorem 2 hold, it has the same convergence rate as \widehat{b}_{FSW} . If the stronger model assumption (10) holds, then \widehat{b}_{DR} enjoys the same convergence rate as \widehat{b}_{OR} , provided that Condition C is satisfied. This differs from classical doubly robust estimators, whose consistency relies on correctly specifying one of the two parametric models.

5. Selection of Tuning Parameters

For the FSW estimator, our methodology requires a kernel function K with a metric d , basis functions ν_k and tuning parameters h , k and q . Using a compactly supported kernel, K would make the denominator of the first term in (9) too small for some Z_i and destabilize the maximization of H_{hk, Z_i} in (9). Therefore, we suggest using the Gaussian kernel, $K(u) = \exp(-u^2)$, with the semi-metric d being the L^2 norm. To develop a general estimation strategy for π_0 , we standardize the X_i 's and use a unified set of choices for ν_k . Letting $X_{\min} = (\min_i \{X_{i1}\}_{i=1}^n, \dots, \min_i \{X_{ip}\}_{i=1}^n)^\top$, $X_{\max} = (\max_i \{X_{i1}\}_{i=1}^n, \dots, \max_i \{X_{ip}\}_{i=1}^n)^\top$, we transform the X_i 's as follows,

$$X_{i,\text{st}} = \frac{2(X_i - X_{\min})}{X_{\max} - X_{\min}} - 1,$$

so that they have support on $[-1, 1]$. Let $X_{ij,\text{st}}$ be the j -th component of $X_{i,\text{st}}$ and \mathcal{P}_ℓ be the ℓ -th standard Legendre polynomial on $[-1, 1]$. If k is given, we define $v_1(X_i) = 1$ and $v_{p(\ell-1)+j+1}(X_i) = \mathcal{P}_\ell(X_{ij,\text{st}})$, for $i = 1, \dots, n$, $j = 1, \dots, p$ and $\ell = 1, \dots, (k-1)/p$. We restrict the range of k such that $(k-1)/p$ is a positive integer, and orthonormalize the resulting matrix $\{\nu_k(X_1), \dots, \nu_k(X_n)\}^\top$.

One may select h , k and q jointly by an L -fold cross-validation (CV), but this may be computationally intensive. We suggest a two-stage CV procedure as follows. In the first stage, we select the truncation parameter q using the OR estimator. Specifically, we randomly split the dataset into L parts, $\mathcal{S}_1, \dots, \mathcal{S}_L$. Let $\mathcal{S}_{-\ell}$ denote the remaining sample

with \mathcal{S}_ℓ excluded. We define the CV loss as

$$\text{CV}_L^{\text{OR}}(q) = \sum_{\ell=1}^L \sum_{i \in \mathcal{S}_\ell} \left[Y_i - \hat{a}_{-\ell, \text{OR}} - \sum_{j=1}^q \hat{b}_{j, -\ell, \text{OR}} \left\{ \int_{\mathcal{T}} \hat{\phi}_j(t) \hat{\mu}_Z(t) dt + \hat{\xi}_{ij} \right\} - \hat{\theta}_{-\ell, \text{OR}}^\top X_i \right]^2,$$

where $\hat{a}_{-\ell, \text{OR}}$, $\hat{b}_{j, -\ell, \text{OR}}$ and $\hat{\theta}_{-\ell, \text{OR}}$ are obtained by applying the method in Section 3.2 to $\mathcal{S}_{-\ell}$. We choose the truncation parameter as $\hat{q}_{\text{CV}} = \text{argmin}_{q \in \{1, \dots, q_{99}\}} \text{CV}_L^{\text{OR}}(q)$, where q_{99} is the number of principal components that can explain at least 99% of the variance of Z .

Noting that the rate requirements for q in Theorem 1 and Theorem 2 are the same, we thus use the same \hat{q}_{CV} in the FSW estimator. It remains to select h and k in the second stage. We consider

$$\begin{aligned} & \text{CV}_L^{\text{FSW}}(h, k) \\ &= \sum_{\ell=1}^L \sum_{i \in \mathcal{S}_\ell} \left[Y_i \hat{\pi}_{hk, -\ell}(Z_i, X_i) - \hat{a}_{-\ell, \text{FSW}} - \sum_{j=1}^{\hat{q}_{\text{CV}}} \hat{b}_{j, -\ell, \text{FSW}} \left\{ \int_{\mathcal{T}} \hat{\phi}_j(t) \hat{\mu}_Z(t) dt + \hat{\xi}_{ij} \right\} \right]^2, \end{aligned}$$

where $\hat{\xi}_{ij} = \int_{\mathcal{T}} \{Z_i(t) - \hat{\mu}_Z(t)\} \hat{\phi}_j(t) dt$ is the estimated PC score, $\hat{a}_{-\ell, \text{FSW}}$, $\hat{b}_{j, -\ell, \text{FSW}}$ and $\hat{\pi}_{hk, -\ell}$ are obtained by applying the method in Section 3.1 to $\mathcal{S}_{-\ell}$. We choose the one from a candidate set of $\{h, k\}$ such that $\text{CV}_L(h, k)$ is minimized.

As for the DR estimator, we still need to select the truncation parameter q . This can be achieved by minimizing the CV loss,

$$\begin{aligned} \text{CV}_L^{\text{DR}}(q) &= \sum_{\ell=1}^L \sum_{i \in \mathcal{S}_\ell} \left[\{Y_i - \hat{E}_{-\ell}(Y|X_i, Z_i)\} \hat{\pi}_{hk, -\ell}(X_i, Z_i) + \frac{1}{|\mathcal{S}_\ell|} \sum_{j \in \mathcal{S}_\ell} \{\hat{E}_{-\ell}(Y|X_j, Z_j)\} \right. \\ &\quad \left. - \hat{a}_{-\ell, \text{DR}} - \sum_{j=1}^q \hat{b}_{j, -\ell, \text{DR}} \left\{ \int_{\mathcal{T}} \hat{\phi}_j(t) \hat{\mu}_Z(t) dt + \hat{\xi}_{ij} \right\} \right]^2, \end{aligned}$$

where $|\{\cdot\}|$ denotes the number of elements in the set $\{\cdot\}$, $\hat{a}_{-\ell, \text{DR}}$, $\hat{E}_{-\ell}(Y|X_i, Z_i)$, $\hat{\pi}_{hk, -\ell}$ and $\hat{b}_{j, -\ell, \text{DR}}$ are obtained by applying the methods in Section 3.2 to $\mathcal{S}_{-\ell}$.

6. Simulation Study

In addition to the methods of functional stabilized weight (FSW), outcome regression (OR) and double robustness (DR) developed in Section 3, we also consider the method using the nonparametric principal component weight (PCW) proposed by Zhang et al. (2021), which assumes the same functional linear model for the ADRF; the weight-modified kernel ridge regression (WMKRR), a nonlinear method proposed by Wang et al. (2023); as well as the direct functional linear regression (FLR) and kernel ridge regression (KRR) of Y on Z . The FLR and KRR are expected to be biased if there are confounding effects. We select the tuning parameters h, k and q following the procedure in Section 5, and choose the number of PCs used in estimating the PCW so as to explain 95% of the variance of Z following Zhang et al. (2021). For a fair comparison, we use the truncation parameter \hat{q}_{CV} in Section 5 for estimating b in all methods that utilize the functional linear model.

Table 1: Mean (standard deviation) of the empirical mean squared error (MSE) of the ADRF with the best values highlighted in boldface.

$n = 200$	(i)	(ii)	(iii)	(iv)	(v)	(vi)
FSW	1.14 (0.66)	2.15 (1.36)	8.40 (8.60)	9.69 (10.08)	7.94 (12.86)	180.24 (42.74)
OR	1.11 (0.68)	6.96 (3.42)	6.85 (9.77)	11.55 (14.11)	29.52 (11.79)	192.79 (58.15)
DR	1.08 (0.66)	7.56 (3.65)	6.65 (9.56)	11.34 (14.36)	29.27 (12.12)	186.10 (52.82)
PCW	2.68 (1.90)	12.98 (2.93)	13.10 (7.74)	16.13 (8.31)	16.17 (9.72)	186.84 (46.58)
FLR	4.93 (1.63)	9.78 (2.29)	21.53 (7.90)	22.98 (7.72)	37.98 (14.08)	205.45 (61.24)
WMKRR	8.30 (3.23)	22.67 (4.72)	25.07 (10.54)	25.57 (10.67)	34.71 (14.18)	133.79 (40.30)
KRR	6.41 (1.83)	11.84 (2.42)	25.94 (6.59)	27.63 (6.60)	79.56 (21.87)	343.84 (158.34)
$n = 500$	(i)	(ii)	(iii)	(iv)	(v)	(vi)
FSW	0.50 (0.34)	1.36 (0.95)	5.82 (4.27)	4.57 (2.98)	3.60 (14.88)	173.47 (28.29)
OR	0.42 (0.33)	6.42 (2.58)	1.61 (2.52)	5.66 (2.53)	25.33 (6.13)	169.20 (41.27)
DR	0.38 (0.31)	6.86 (2.74)	1.52 (2.45)	5.64 (2.41)	25.39 (6.55)	166.22 (36.65)
PCW	1.68 (1.08)	12.16 (1.85)	8.42 (3.54)	11.50 (3.05)	12.09 (7.21)	169.64 (34.34)
FLR	4.31 (1.05)	9.00 (1.51)	17.74 (3.95)	18.83 (3.60)	33.14 (8.37)	174.16 (46.45)
WMKRR	6.62 (2.37)	19.62 (3.52)	23.89 (9.77)	24.80 (10.01)	35.26 (11.86)	125.12 (32.51)
KRR	5.14 (1.06)	11.10 (1.54)	25.06 (4.22)	26.38 (4.33)	79.08 (14.62)	302.76 (79.42)

The code to reproduce the simulation results for our methods and the FLR is available at <https://github.com/ruoxut/FunctionalTreatment>.

We consider the sample sizes $n = 200$ and 500 , and generate data for models (i) to (v) as follows. For $i = 1, \dots, n$, we generate the functional treatment $Z_i(t)$, for $t \in [0, 1]$, by $Z_i(t) = \sum_{j=1}^6 A_{ij} \phi_j(t)$, where $A_{i1} = 4U_{i1}$, $A_{i2} = 2\sqrt{3}U_{i2}$, $A_{i3} = 2\sqrt{2}U_{i3}$, $A_{i4} = 2U_{i4}$, $A_{i5} = U_{i5}$, $A_{i6} = U_{i6}/\sqrt{2}$ with the U_{ij} 's being the independent standard normal variables, $\phi_{2m-1}(t) = \sqrt{2} \sin(2m\pi t)$ and $\phi_{2m}(t) = \sqrt{2} \cos(2m\pi t)$ for $m = 1, 2, 3$. For the slope function b , we define $b(t) = 2\phi_1(t) + \phi_2(t) + \phi_3(t)/2 + \phi_4(t)/2$. The Z_i 's and b are the same as those in Zhang et al. (2021). We consider five models for the covariate X_i and the outcome variable Y_i :

- (i) $X_i = U_{i1} + \epsilon_{i1}$ and $Y_i = 1 + \int_0^1 b(t)Z_i(t) dt + 2X_i + \epsilon_{i2}$,
- (ii) $X_i = U_{i1} + 0.25\epsilon_{i1}$ and $Y_i = 1 + \int_0^1 b(t)Z_i(t) dt + 5 \sin(X_i) + \epsilon_{i2}$,
- (iii) $X_i = (X_{i1}, X_{i2})^\top = \{(U_{i1} + 1)^2 + \epsilon_{i1}, U_{i2}\}^\top$ and $Y_i = 1 + \int_0^1 b(t)Z_i(t) dt + 2X_{i1} + 2X_{i2} + \epsilon_{i2}$,
- (iv) $X_i = (X_{i1}, X_{i2})^\top = \{(U_{i1} + 1)^2 + \epsilon_{i1}, U_{i2}\}^\top$ and $Y_i = 1 + \int_0^1 b(t)Z_i(t) dt + 2X_{i1} + 2 \cos(X_{i1}) + 5.5 \sin(X_{i2}) + \epsilon_{i2}$,
- (v) $X_i = (X_{i1}, X_{i2})^\top = \{(U_{i1} + 1)^2 + \epsilon_{i1}, U_{i2}\}^\top$ and $Y_i = 1 + \int_0^1 b(t)Z_i(t) dt + \{\int_0^1 b(t)Z_i(t) dt\}^2/25 + 2X_{i1} + 5.5 \sin(X_{i2}) + \epsilon_{i2}$,

where $\epsilon_{i1} \sim N(0, 1)$ and $\epsilon_{i2} \sim N(0, 25)$ are generated independently. In addition, we also consider a nonlinear model from Wang et al. (2023),

- (vi) $X_i = (U_{i11}, U_{i21}, U_{i31}, U_{i41})^\top$, $Y_i = 10\{U_{i21}U_{i11}^2 + U_{i41}^2 \sin(2U_{i31})\} + 0.5A_{i1}^2 + 4 \sin(A_{i1})$, $Z_i(t) = \sum_{j=1}^4 A_{ij} \sqrt{2} \sin(2j\pi t)$, for $t \in [0, 1]$, where $A_{i1} = 4U_{i11} + U_{i12}$, $A_{i2} = 2\sqrt{3}U_{i21} +$

U_{i22} , $A_{i3} = 2\sqrt{2}U_{i31} + U_{i32}$, $A_{i4} = 2U_{i41} + U_{i42}$, with the U_{ijk} 's being the independent standard normal variables.

The PC scores A_{ij} of $Z_i(t)$ affect covariate X_i linearly in models (i) and (ii) and nonlinearly in models (iii) and (iv), and covariate X_i affects the outcome variable Y_i linearly in models (i) and (iii) and nonlinearly in models (ii) and (iv). The functional linear model (4) for the ADRF is correctly specified for models (i) to (iv), while the partially linear model (10) is only correctly specified for models (i) and (iii). Under models (v) and (vi), the functional linear model (4) for the ADRF is misspecified. Due to the confounding effect of X , the FLR and KRR ignoring X are expected to be biased. For each combination of model and sample size, we replicate 200 simulations and evaluate the results by the empirical mean squared error (MSE) of the ADRF following Wang et al. (2023), $\text{MSE} = \sum_{i=1}^n [\hat{E}\{Y^*(Z_i)\} - E\{Y^*(Z_i)\}]^2/n$, where $\hat{E}(\cdot)$ denotes a generic estimator of the ADRF.

Table 1 summarizes the mean and the standard deviation of the MSEs for all configurations. Our proposed estimators FSW, OR and DR outperform the other methods under models (i)–(iv). In particular, our three estimators are close to each other and are better than the others under models (i) and (iii), where covariate X affects the outcome variable Y linearly. Under models (ii) and (iv) where the partially linear model (10) is misspecified, the FSW performs the best, followed by the OR and DR, and it is worth noting that the latter two still perform better than the other methods. Under model (v) where the functional linear model is misspecified, the FSW remarkably performs the best, followed by the PCW, while the FLR and WMKRR perform similarly. Under model (v), the FSW is expected to be biased. The performance of PCW is better than FLR but inferior to our methods under models (i)–(iv). The WMRKK performs the best under model (vi), but its performance is not satisfactory under models (i)–(v).

To give visualization of representative estimated slope functions, we show the quartile curves of the samples corresponding to the first, second and third quartile values of 200 MSEs for model (i) in Figure 2. It can be seen that the estimated curves using the FSW, OR and DR are close to the truth when $n = 500$, while the PCW is more variable and the FLR is more biased. Figure 3 corresponds to the results under model (iv), i.e., a more challenging model structure. The differences among the estimators are more significant: FSW performs the best, followed by the DR, OR and PCW, while the FLR is more biased than the others.

We also conduct a small-scale running time comparison of our estimator FSW and the WMKRR proposed by Wang et al. (2023). Both estimators contain the same weight function π_0 estimated by different nonparametric methods. We compare the running time of computing the two estimators of π_0 , including selection of the tuning parameters. The experiments were carried out on a PC with an i7-12700 CPU and 16 GB RAM. As an illustration, model (i) is used as the data generation model with $n = 100, 500$ and 1000 . We repeat the experiments 10 times. The values of averaged running time in seconds are 6.19 ($n = 100$), 37.28 ($n = 500$) and 236.88 ($n = 1000$) for the FSW, while the corresponding values for the WMKRR are 36.27, 204.77 and 440.67. We see that, for the given model, computing the estimator of π_0 using the FSW is much faster than that using the WMKRR.

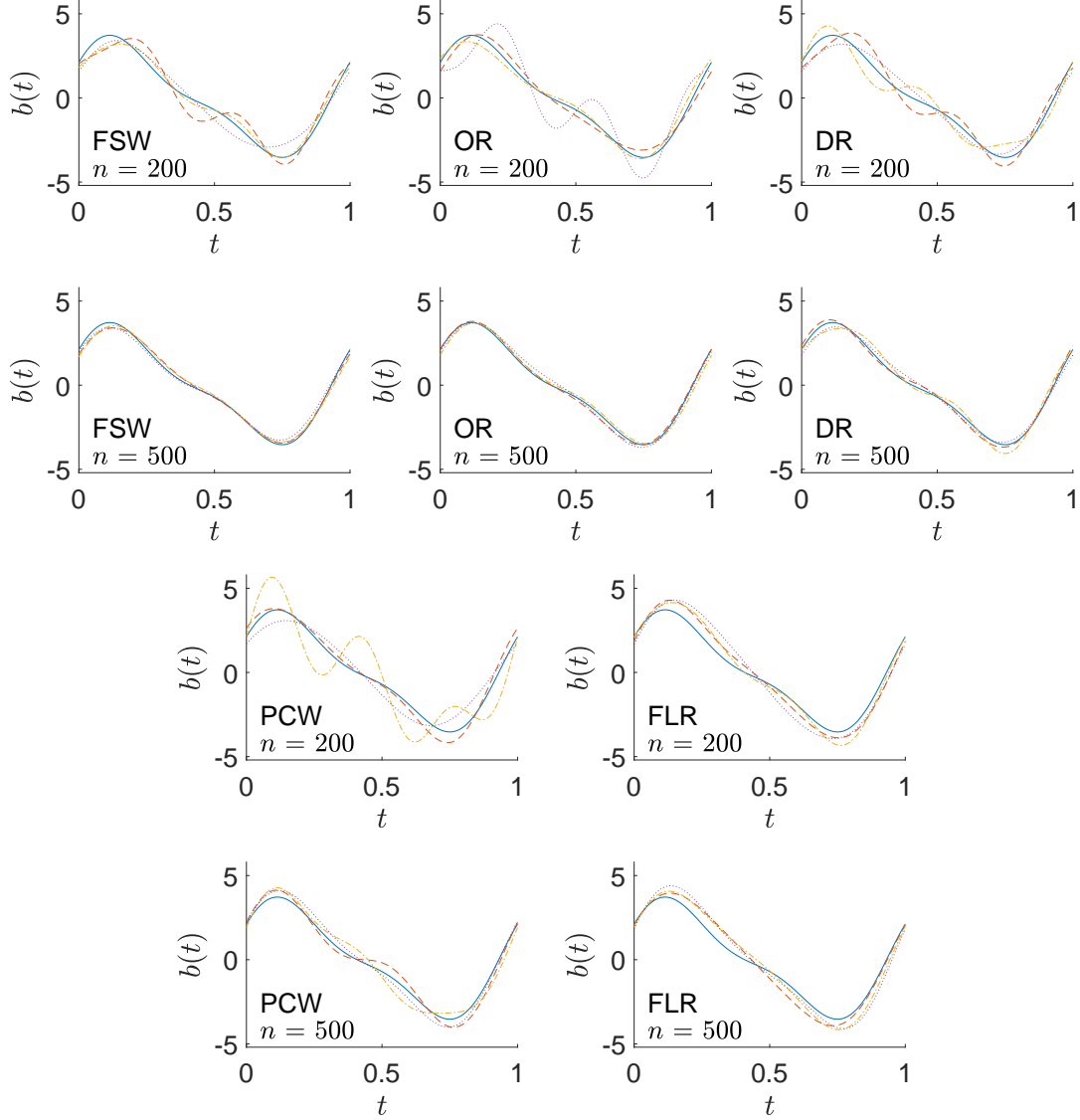


Figure 2: True curve (—), first (---), second (- · -) and third (· · ·) quartile estimated slope functions under model (i) with $n = 200$ and 500 for all methods.

7. Real Data Analysis

We illustrate the estimation of the ADRF using the five methods, FSW, OR, DR, PCW and FLR, on the electroencephalography (EEG) dataset from Ciarleglio et al. (2022). The EEG is a relatively low-cost tool to measure human neuronal activities. The measurements are taken from multiple electrodes placed on the scalps of subjects, and they are then processed and transformed to produce current source density curves on the frequency domain, which can provide information on the intensity of neuronal activities. In particular, the

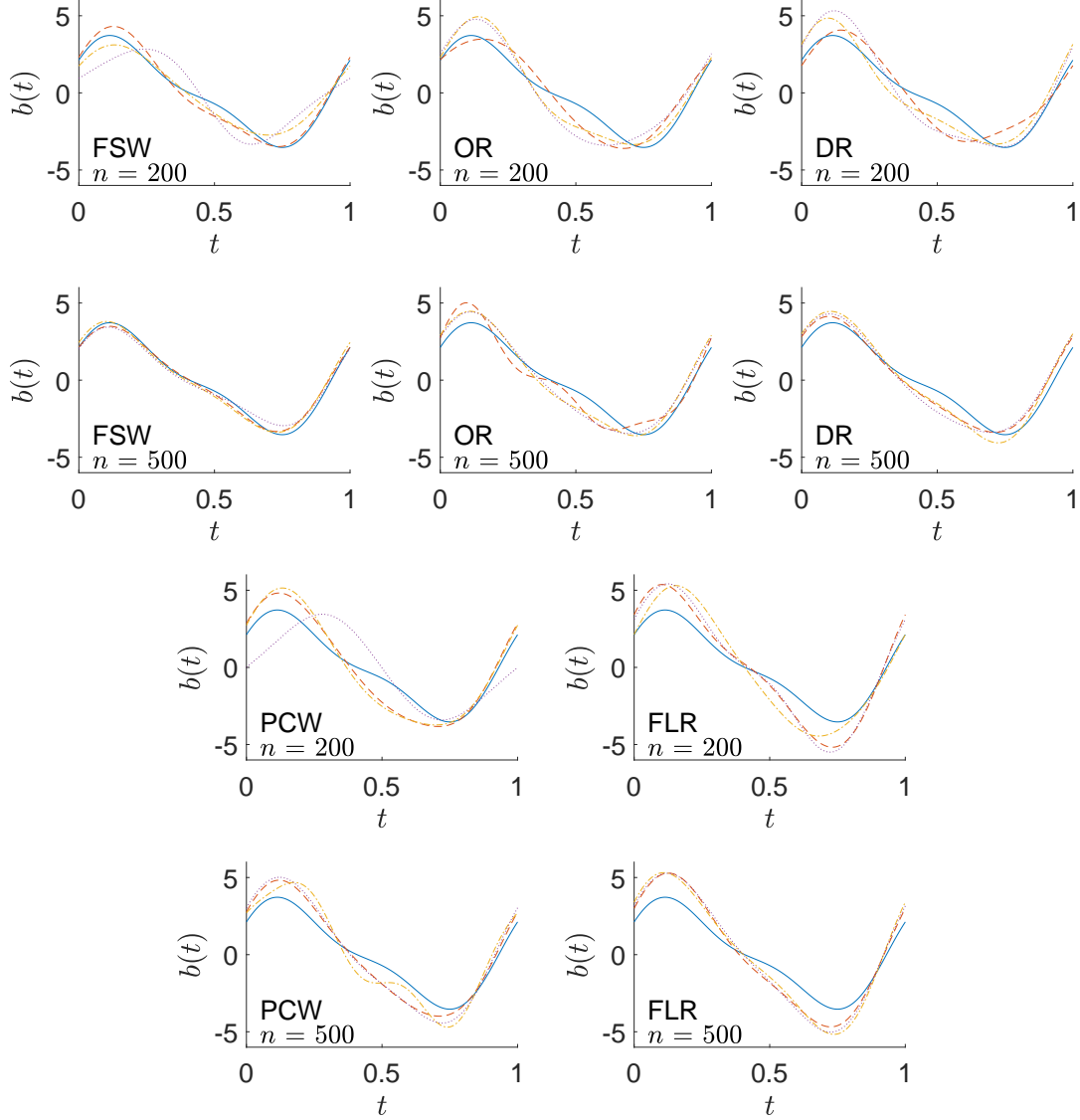


Figure 3: True curve (—), first (---), second (- · -) and third (· · ·) quartile estimated slope functions under model (iv) with $n = 200$ and 500 for all methods.

frontal asymmetry curves, considered potential biomarkers for major depressive disorder, are treated as the functional treatment variable $Z(t)$, for $t \in [4, 31.75]$ Hz. The outcome variable Y is defined as $\log(\tilde{Y} + 1)$, where \tilde{Y} is the quick inventory of depressive symptomatology (QIDS) score measuring the severity of depressive symptomatology. A larger value of Y indicates a more severe depressive disorder. We investigate the causal effect of neuronal activities represented by the frontal asymmetry curves on the severity of major depressive disorder. Potential confounding covariates $X = (X_1, X_2, X_3)^\top$ include age X_1 , sex X_2 (1 for female and 0 for male) and Edinburgh Handedness Inventory score X_3 (rang-

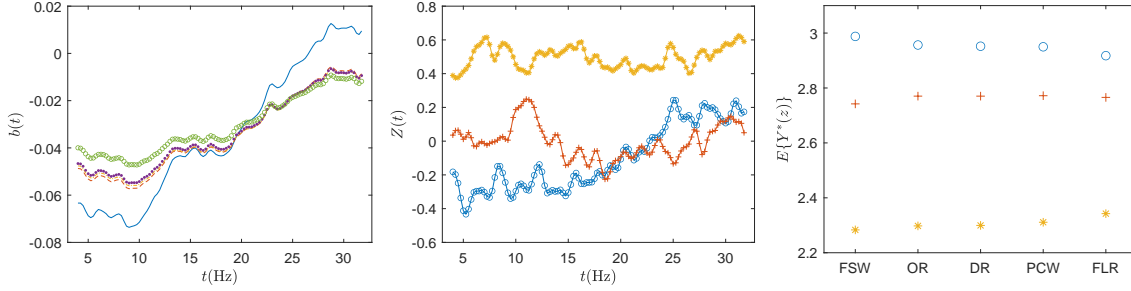


Figure 4: Left: the estimated slope functions using FSW (—), OR (---), DR (— · —), PCW (···) and FLR (○); middle: the frontal asymmetry curves corresponding to the smallest (○), median (+) and largest (*) average values over 4 to 15 Hz; right: the estimated $E\{Y^*(Z)\}$ of the three selected curves with marks corresponding to the middle panel.

ing from -100 to 100 corresponding to completely left to right handedness). Individuals with missing Z (29.6% of the total sample) are removed, which results in a sample of size 85 males and 151 females. The means (standard deviations) of Y , X_1 and X_3 are 2.73 (0.72), 35.97 (13.07) and 72.69 (48.76), respectively. For comparison of the estimated ADRF using different methods, the confounding variables X are centralized.

We use the number of PCs explaining 95% of the variance of Z (equal to 11 for this dataset) in estimating the PCW. We choose the truncation parameter q by \hat{q}_{CV} to estimate the slope function b for all methods and choose h and k used in FSW by minimizing CV_L^{FSW} as in Section 5. As a result, $\hat{q}_{CV} = 2$ is used, which can explain about 80% of the variance of Z .

The left panel of Figure 4 shows the estimated slope functions. All the estimated slope functions have the same increasing trend over frequencies, while they have a negative effect on the outcome variable for low frequencies. In other words, subjects with higher frontal asymmetry values on the low-frequency domain tend to be healthier. This is consistent with the observations in Ciarleglio et al. (2022), where it was found a negative association between the frontal asymmetry curves and the major depressive disorder status using their functional regression model. More specifically, in the left panel of Figure 4, the estimated slope function of FSW is the steepest and shows the largest effect in the low-frequency domain. In contrast, the estimated slope function of FLR has the overall smallest absolute values; the other three curves are quite close to each other.

To visualize the above negative effect in the estimated slope function, we compute the estimated ADRF $E\{Y^*(z)\}$ of three curves, which correspond to the smallest, median and largest average frontal asymmetry values over low frequencies 4 to 15 Hz. The middle panel of Figure 4 exhibits the selected curves and the right panel shows the corresponding estimated $E\{Y^*(z)\}$ using all methods, which are around 2.9, 2.8 and 2.3 for the three curves, respectively.

8. Discussion

In the case of a functional treatment variable, we propose three estimators of the ADRF, namely, the FSW, OR and DR estimators, based on the functional linear model for the ADRF. The consistency of the FSW estimator relies on the functional linear model of $E\{Y^*(z)\}$ by developing a nonparametric estimator of the weight; while the OR estimator requires a more restrictive linear model for Y regressed on (Z, X) . The DR estimator is consistent if either of the first two estimators is consistent.

It is of interest to construct the confidence band for the slope function to better quantify the ADRF or ATE, which, however, is difficult even in the simpler context of functional linear regression. As shown in Cardot et al. (2007), it is impossible for an estimator of the slope function to converge in distribution to a nondegenerate random element under the norm topology. For the OR estimator, it is possible to adapt the method proposed by Imaizumi and Kato (2019) to our context to construct an approximate confidence band. For the other two estimators, the construction of the confidence band is challenging due to the less restrictive modelling assumption and the nonparametric estimator of the FSW, which warrants further investigation.

Acknowledgements

We thank the Action Editor and two referees for their careful reviews that greatly improved our work. We also thank Dr. Jiayi Wang for sharing her code. The research of Zheng Zhang is supported by the funds from the National Key R&D Program of China [grant number 2022YFA1008300] and the National Natural Science Foundation of China [grant number 12371284]. Tan's research was supported by NSFC (No. 12401363 and 12471263), the Fundamental Research Funds for the Central Universities, and Key Laboratory of Intelligent Computing and Applications (Ministry of Education). Yin's research was supported by the Research Grants Council of Hong Kong (17308321), the Theme-based Research Scheme (TRS) Project T45-401/22-N, and the Patrick SC Poon Endowment Fund.

Appendix A. Proof of Proposition 1

Proof The ‘if’ part follows from (7), and thus we show the ‘only if’ part below. Suppose

$$E\{\pi(z, X)v(X)|Z = z\} = E\{v(X)\},$$

for all integrable functions v . Since $E\{v(X)\} = E\{\pi_0(z, X)v(X)|Z = z\}$ for all integrable functions v , we have

$$E[\{\pi(z, X) - \pi_0(z, X)\}v(X)|Z = z] = 0.$$

Taking $v(X) = \exp(ia^\top X)$ for all $a \in \mathbb{R}^p$, we have

$$E[\{\pi(z, X) - \pi_0(z, X)\} \exp(ia^\top X)|Z = z] = 0,$$

According to the uniqueness of the Fourier transformation, we conclude that $\pi(z, \cdot) = \pi_0(z, \cdot)$ a.s. ■

Appendix B. Generalized empirical likelihood interpretation of $\hat{\pi}_{hk}$

To gain more insight on the function ρ , we investigate the generalized empirical likelihood interpretation of our estimator. We show that the estimator defined by (9) is the dual solution to the following local generalized empirical likelihood maximization problem. For each Z_i , $i = 1, \dots, n$,

$$\begin{aligned} & \max_{\{\pi(Z_i, X_j)\}_{j \in \mathcal{S}_{-i}}} - \frac{\sum_{j \in \mathcal{S}_{-i}} D\{\pi(Z_i, X_j)\} K\{d(Z_j, Z_i)/h\}}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, Z_i)/h\}}, \\ \text{s.t. } & \frac{\sum_{j \in \mathcal{S}_{-i}} \pi(Z_i, X_j) \nu_k(X_j) K\{d(Z_j, Z_i)/h\}}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, Z_i)/h\}} = \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j), \end{aligned} \quad (12)$$

where $D(w)$ is a distance measure between w and 1 for $w \in \mathbb{R}$. The function $D(v)$ is continuously differentiable satisfying $D(1) = 0$ and $\rho(-u) = D\{D'^{(-1)}(u)\} - u \cdot D'^{(-1)}(u)$, where $D'^{(-1)}$ is the inverse function of D' , the first derivative of D . Different choices of ρ correspond to different distance measures. For example, if we take $\rho(v) = -\exp(-v - 1)$, then $D(v) = -v \log(v)$ is the information entropy and the weights correspond to the exponential tilting (Imbens et al., 1998). Other choices include $\rho(v) = \log(v) + 1$ with $D(v) = -\log(v)$, the empirical likelihood (Owen, 1988), and $\rho(v) = -(1 - v)^2/2$ with $D(v) = (1 - v)^2/2$, yielding the implied weights of the continuous updating of the generalized method of moments (Hansen et al., 1996).

Following Tseng and Bertsekas (1991), we show that the dual problem of (12) is to maximize H_{hk, Z_i} in (9). For each $i = 1, \dots, n$, we let $\bar{\nu}_k = \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j)/(n-1)$, $K_{ij} = K\{d(Z_j, Z_i)/h\}/[\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, Z_i)/h\}]$, $\pi_{ij} = \pi(Z_i, X_j)$ and $w_{ij} = K_{ij}\pi_{ij}$, for $j \in \mathcal{S}_{-i}$. Moreover, let $D_{ij}(\nu) = D(\nu/K_{ij})$, $V_k = (\nu_k(X_j))_{j \in \mathcal{S}_{-i}} \in \mathbb{R}^{k \times (n-1)}$, $w_i = (w_{ij})_{j \in \mathcal{S}_{-i}}^\top$, and $F(w_i) = \sum_{j \in \mathcal{S}_{-i}} D_{ij}(w_{ij})K_{ij}$. Then, (12) can be written as

$$\min_{w_i} F(w_i), \text{ s.t. } V_k w_i = \bar{\nu}_k. \quad (13)$$

We define the conjugate convex function of F as

$$\begin{aligned}
F^*(u) &= \sup_{w_i} \sum_{j \in \mathcal{S}_{-i}} \{u_j w_{ij} - K_{ij} D_{ij}(w_{ij})\} \\
&= \sup_{\{\pi_{ij}\}_{j \in \mathcal{S}_{-i}}} \sum_{j \in \mathcal{S}_{-i}} K_{ij} \{u_j \pi_{ij} - D(\pi_{ij})\} \\
&= \sum_{j \in \mathcal{S}_{-i}} K_{ij} \{u_j \pi_{ij}^* - D(\pi_{ij}^*)\},
\end{aligned}$$

where the π_{ij}^* 's satisfy

$$u_j = D'(\pi_{ij}^*) \Rightarrow \pi_{ij}^* = D'^{(-1)}(u_j), \quad j = 1, \dots, n,$$

by taking the first order condition. It follows that

$$F^*(u) = - \sum_{j \in \mathcal{S}_{-i}} K_{ij} \rho(-u_j), \quad (14)$$

where $\rho(-u) = D\{D'^{(-1)}(u)\} - u \cdot D'^{(-1)}(u)$.

Following Tseng and Bertsekas (1991) and using (14), we conclude that the dual problem of (12) is

$$\begin{aligned}
\max_{\eta \in \mathbb{R}^k} \{-F^*(\eta^\top V_k) + \eta^\top \bar{\nu}_k\} &= \max_{\eta \in \mathbb{R}^k} \sum_{j \in \mathcal{S}_{-i}} \left[K_{ij} \rho\{-\eta^\top \nu_k(X_j)\} + \frac{\eta^\top \nu_k(X_j)}{n-1} \right] \\
&= \max_{\eta \in \mathbb{R}^k} \sum_{j \in \mathcal{S}_{-i}} \left[K_{ij} \rho\{\eta^\top \nu_k(X_j)\} - \frac{\eta^\top \nu_k(X_j)}{n-1} \right] = \max_{\eta \in \mathbb{R}^k} H_{hk, Z_i}(\eta).
\end{aligned}$$

Appendix C. Proof of Double Robustness in (11)

Recall from Section 2 that $E\{Y^*(z)\} = E\{Y\pi_0(X, Z)|Z = z\} = E\{E(Y|X, Z = z)\}$. First assuming that $\tilde{E}(Y|X, Z) = E(Y|X, Z)$, then we have

$$\begin{aligned}
&E[\{Y - \tilde{E}(Y|X, Z)\}\tilde{\pi}(X, Z) + E_X\{\tilde{E}(Y|X, Z)\}|Z = z] \\
&= E[\{Y - E(Y|X, Z)\}\tilde{\pi}(X, Z)|Z = z] + E\{E(Y|X, Z = z)\} \\
&= E\{Y^*(z)\},
\end{aligned}$$

where the second equality follows from the law of total expectation.

Assuming that $\tilde{\pi}(X, Z) = \pi_0(X, Z) = f_X(X)/f_{X|Z}(X|Z)$, we have

$$\begin{aligned}
&E[\{Y - \tilde{E}(Y|X, Z)\}\tilde{\pi}(X, Z) + E_X\{\tilde{E}(Y|X, Z)\}|Z = z] \\
&= E\{Y\pi_0(X, Z)|Z = z\} - E\{\tilde{E}(Y|X, Z)\pi_0(X, Z)|Z = z\} + E_X\{\tilde{E}(Y|X, Z)\}|Z = z] \\
&= E\{Y^*(z)\} - \int \tilde{E}(Y|X = x, Z = z) f_X(x) dx + \int \tilde{E}(Y|X = x, Z = z) f_X(x) dx \\
&= E\{Y^*(z)\}.
\end{aligned}$$

Therefore, the doubly robust property is proved.

Appendix D. Proof of Theorem 1

Proof Recall from Section 3.1 that

$$\widehat{\pi}_{hk}(Z_i, X_i) = \rho' \{ \widehat{\eta}_{Z_i}^\top \nu_k(X_i) \},$$

where $\widehat{\eta}_{Z_i}$ is a k -dimensional vector maximizing the strictly concave function H_{hk, Z_i} defined by

$$H_{hk, Z_i}(\eta) = \frac{\sum_{j \in \mathcal{S}_{-i}} \rho \{ \eta^\top \nu_k(X_j) \} K \{ d(Z_j, Z_i)/h \}}{\sum_{j \in \mathcal{S}_{-i}} K \{ d(Z_j, Z_i)/h \}} - \eta^\top \left\{ \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j) \right\}.$$

For a fixed $z \in L^2(\mathcal{T})$, we have

$$\widehat{\pi}_{hk}(z, X_i) = \rho' \{ \widehat{\eta}_z^\top \nu_k(X_i) \},$$

where $\widehat{\eta}_z$ is a k -dimensional vector maximizing $H_{hk, z}$ defined by

$$H_{hk, z}(\eta) = \frac{\sum_{j \in \mathcal{S}_{-i}} \rho \{ \eta^\top \nu_k(X_j) \} K \{ d(Z_j, z)/h \}}{\sum_{j \in \mathcal{S}_{-i}} K \{ d(Z_j, z)/h \}} - \eta^\top \left\{ \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j) \right\}. \quad (15)$$

Here, \mathcal{S}_{-i} is independent of Z_i for $i = 1, \dots, n$.

Let $H_{hk, z}^*(\eta)$ be the theoretical counterpart of $H_{hk, z}(\eta)$,

$$H_{hk, z}^*(\eta) = \frac{E[\rho \{ \eta^\top \nu_k(X) \} K \{ d(Z, z)/h \}]}{E\{K \{ d(Z, z)/h \}\}} - \eta^\top E\{\nu_k(X)\},$$

$\eta_z^* = \operatorname{argmax}_{\eta \in \mathbb{R}^p} H_{hk, z}^*(\eta)$ and $\pi_{hk}^*(z, x) = \rho' \{ \eta_z^{*\top} \nu_k(x) \}$.

Note that

$$\begin{aligned} & \sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} |\widehat{\pi}_{hk}(z, x) - \pi_0(z, x)| \\ & \leq \sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} |\widehat{\pi}_{hk}(z, x) - \pi_{hk}^*(z, x)| + \sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} |\pi_{hk}^*(z, x) - \pi_0(z, x)| \end{aligned}$$

To prove Theorem 1, it suffices to prove the convergence rates for the two terms on the right-hand side of the above inequality. Lemmas 2 and 3 provide such results. Lemma 1 is needed to prove Lemmas 2 and 3. ■

Lemma 1 *Let*

$$\widehat{e}_{k_1, k_2}(z) = \frac{\sum_{j \in \mathcal{S}_{-i}} g_{k_1, k_2, z}(X_j) K \{ d(Z_j, z)/h \}}{(n-1)E[K \{ d(Z_j, z)/h \}]},$$

where $g_{k_1, k_2, z}$ is defined in (A6), and

$$\widehat{f}(z) = \frac{\sum_{j \in \mathcal{S}_{-i}} K \{ d(Z_j, z)/h \}}{(n-1)E[K \{ d(Z_j, z)/h \}]}$$

Under Conditions (A4) to (A7), we have

$$\begin{aligned} \sup_{z \in \mathcal{Z}} |\hat{f}(z) - 1| &= O_{a.c.} \left\{ \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}}{n\mu(h)}} \right\}, \\ \sum_{n=1}^{\infty} P \left(\inf_{z \in \mathcal{Z}} \hat{f}(z) < \frac{1}{2} \right) &< \infty, \\ \sup_{z \in \mathcal{Z}} |E\{\hat{e}_{k_1, k_2}(z)\} - E[g_{k_1, k_2, z}(X)|Z = z]| &= O(h^\lambda), \quad \forall k_1, k_2 \in \{1, \dots, k\}, \\ \sup_{z \in \mathcal{Z}} |\hat{e}_{k_1, k_2}(z) - E\{\hat{e}_{k_1, k_2}(z)\}| &= O_{a.c.} \left\{ \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}}{n\mu(h)}} \right\}, \quad \forall k_1, k_2 \in \{1, \dots, k\}, \end{aligned}$$

hold uniformly over k .

Proof

Note that for all $k_1, k_2 = 1, \dots, k$, $g_{k_1, k_2, z}$ is a scalar function, and Condition (A6) is assumed uniformly over k . Therefore, Lemmas 8 to 11 in Ferraty et al. (2010) can be used to derive the lemma above. Also, note the $O_{a.c.}$ denotes the almost complete convergence, which implies the almost sure convergence and convergence in probability. \blacksquare

Lemma 2 Under Conditions (A1) to (A6), we have

$$\sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} |\pi_{hk}^*(z, x) - \pi_0(z, x)| = O\{\zeta(k)(k^{-\alpha} + h^\lambda \sqrt{k})\}$$

Proof We first prove that

$$\sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} |\pi_0(z, x) - \rho' \{\eta_z^\top \nu_k(x)\}| = O(k^{-\alpha}), \quad (16)$$

where η_z is given in Condition (A2).

Recall that ρ is a strictly increasing and concave function, so $\rho'^{(-1)}$ is strictly decreasing. Thus, we have

$$\rho'^{(-1)}(c_2) \leq \inf_{(z, x) \in \mathcal{Z} \times \mathcal{X}} \rho'^{(-1)}\{\pi_0(z, x)\} \leq \sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} \rho'^{(-1)}\{\pi_0(z, x)\} \leq \rho'^{(-1)}(c_1),$$

for some positive constants c_1, c_2 , because $\pi_0(z, x)$ is bounded away from zero and infinity by Condition (A1). According to Condition (A2), there exists a constant $C_1 > 0$ such that

$$\sup_{(z, x) \in \mathcal{Z} \times \mathcal{X}} |\rho'^{(-1)}\{\pi_0(z, x)\} - \eta_z^\top \nu_k(x)| \leq C_1 k^{-\alpha}.$$

Thus, we have, $\forall z \in \mathcal{Z}$ and $\forall x \in \mathcal{X}$,

$$\begin{aligned} \eta_z^\top \nu_k(x) &\in [\rho'^{(-1)}\{\pi_0(z, x)\} - C_1 k^{-\alpha}, \rho'^{(-1)}\{\pi_0(z, x)\} + C_1 k^{-\alpha}] \\ &\subset [\rho'^{(-1)}(c_2) - C_1 k^{-\alpha}, \rho'^{(-1)}(c_1) + C_1 k^{-\alpha}] \end{aligned} \quad (17)$$

and

$$\begin{aligned}
 & \rho' \{ \eta_z^\top \nu_k(x) + C_1 k^{-\alpha} \} - \rho' \{ \eta_z^\top \nu_k(x) \} \\
 & < \pi_0(z, x) - \rho' \{ \eta_z^\top \nu_k(x) \} \\
 & < \rho' \{ \eta_z^\top \nu_k(x) - C_1 k^{-\alpha} \} - \rho' \{ \eta_z^\top \nu_k(x) \}.
 \end{aligned} \tag{18}$$

By the mean value theorem, for k large enough, there exist

$$\begin{aligned}
 \xi_1(z, x) & \in (\eta_z^\top \nu_k(x), \eta_z^\top \nu_k(x) + C_1 k^{-\alpha}) \subset [\rho'^{(-1)}(c_2) - C_1 k^{-\alpha}, \rho'^{(-1)}(c_1) + 2C_1 k^{-\alpha}] \subset \Gamma_1, \\
 \xi_2(z, x) & \in (\eta_z^\top \nu_k(x) - C_1 k^{-\alpha}, \eta_z^\top \nu_k(x)) \subset [\rho'^{(-1)}(c_2) - 2C_1 k^{-\alpha}, \rho'^{(-1)}(c_1) + C_1 k^{-\alpha}] \subset \Gamma_1,
 \end{aligned}$$

such that

$$\begin{aligned}
 \rho' \{ \eta_z^\top \nu_k(x) + C_1 k^{-\alpha} \} - \rho' \{ \eta_z^\top \nu_k(x) \} & = \rho'' \{ \xi_1(z, x) \} C_1 k^{-\alpha} \geq \inf_{u \in \Gamma_1} \rho''(u) C_1 k^{-\alpha}, \\
 \rho' \{ \eta_z^\top \nu_k(x) - C_1 k^{-\alpha} \} - \rho' \{ \eta_z^\top \nu_k(x) \} & = -\rho'' \{ \xi_2(z, x) \} C_1 k^{-\alpha} \leq \sup_{u \in \Gamma_1} \{-\rho''(u)\} C_1 k^{-\alpha},
 \end{aligned} \tag{19}$$

where

$$\Gamma_1 = [\rho'^{(-1)}(c_2) - 1, \rho'^{(-1)}(c_1) + 1].$$

Note that $\max\{-\inf_{u \in \Gamma_1} \rho''(u), \sup_{u \in \Gamma_1} \{-\rho''(u)\}\}$ is a positive and finite constant, because Γ_1 is compact and ρ'' is continuous and strictly negative. Thus, the claim (16) is proved using (18) and (19).

Next, we compute a bound for $\sup_{z \in \mathcal{Z}} \|\nabla H_{hk,z}^*(\eta_z)\|$. We have

$$\begin{aligned}
 & \sup_{z \in \mathcal{Z}} \|\nabla H_{hk,z}^*(\eta_z)\| \\
 & = \sup_{z \in \mathcal{Z}} \left\| \frac{E[\rho' \{ \eta_z^\top \nu_k(X) \} \nu_k(X) K \{ d(Z, z)/h \}]}{E[K \{ d(Z, z)/h \}]} - E\{\nu_k(X)\} \right\| \\
 & \leq \sup_{z \in \mathcal{Z}} \left\| \frac{E[\rho' \{ \eta_z^\top \nu_k(X) \} \nu_k(X) K \{ d(Z, z)/h \}]}{E[K \{ d(Z, z)/h \}]} - E[\rho' \{ \eta_z^\top \nu_k(X) \} \nu_k(X) | Z = z] \right\| \\
 & \quad + \sup_{z \in \mathcal{Z}} \|E([\rho' \{ \eta_z^\top \nu_k(X) \} - \pi_0(z, X)] \nu_k(X) | Z = z)\| \\
 & \leq \sqrt{\sum_{k'=1}^k \sup_{z \in \mathcal{Z}} \left| \frac{E[\rho' \{ \eta_z^\top \nu_k(X) \} \nu_{k'}(X) K \{ d(Z, z)/h \}]}{E[K \{ d(Z, z)/h \}]} - E[\rho' \{ \eta_z^\top \nu_k(X) \} \nu_{k'}(X) | Z = z] \right|^2} \\
 & \quad + \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\rho' \{ \eta_z^\top \nu_k(x) \} - \pi_0(z, x)| = O(\sqrt{k} h^\lambda + k^{-\alpha}),
 \end{aligned} \tag{20}$$

where Lemma 1 and equation (16) are used for the last inequality.

For all $z \in \mathcal{Z}$ and for some constant $C_2 > 0$ (to be chosen later), define the set

$$\Lambda_z = \{ \eta \in \mathbb{R}^k : \|\eta - \eta_z\| \leq C_2(k^{-\alpha} + h^\lambda \sqrt{k}) \}.$$

By Condition (A3), we have $\forall \eta \in \Lambda_z$,

$$\sup_{x \in \mathcal{X}} |\eta^\top \nu_k(x) - \eta_z^\top \nu_k(x)| \leq \|\eta - \eta_z\| \sup_{x \in \mathcal{X}} \|\nu_k(x)\| < C_2(k^{-\alpha} + h^\lambda \sqrt{k}) \zeta(k),$$

so that $\forall x \in \mathcal{X}$ and for sufficiently large k , by (17),

$$\begin{aligned} & \eta^\top \nu_k(x) \\ & \in [\rho'^{(-1)}(c_2) - C_1 k^{-\alpha} - C_2(k^{-\alpha} + h^\lambda \sqrt{k})\zeta(k), \rho'^{(-1)}(c_1) + C_1 k^{-\alpha} + C_2(k^{-\alpha} + h^\lambda \sqrt{k})\zeta(k)] \subset \Gamma_1. \end{aligned} \quad (21)$$

Based on (20), there exists a $C_4 > 0$ that is independent of z such that $\sup_{z \in \mathcal{Z}} \|\nabla H_{hk,z}^*(\eta_z)\| < C_4(k^{-\alpha} + h^\lambda \sqrt{k})$. For any $\eta \in \partial\Lambda_z$, i.e. $\|\eta - \eta_z\| = C_2(k^{-\alpha} + h^\lambda \sqrt{k})$, by the mean value theorem we have

$$\begin{aligned} & H_{hk,z}^*(\eta) - H_{hk,z}^*(\eta_z) \\ &= (\eta - \eta_z)^\top \nabla H_{hk,z}^*(\eta_z) + \frac{1}{2}(\eta - \eta_z)^\top \nabla^2 H_{hk,z}^*(\bar{\eta}_z)(\eta - \eta_z) \\ &\leq \|\eta - \eta_z\| \|\nabla H_{hk,z}^*(\eta_z)\| \\ &\quad + \frac{1}{2}(\eta - \eta_z)^\top \frac{E[\rho''\{\bar{\eta}_z^\top \nu_k(X)\} K\{d(Z, z)/h\} \nu_k(X) \nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} (\eta - \eta_z) \\ &\leq \|\eta - \eta_z\| \|\nabla H_{hk,z}^*(\eta_z)\| - \frac{a_1}{2}(\eta - \eta_z)^\top [E\{\nu_k(X) \nu_k(X)^\top | Z = z\} + O(h^\lambda) J_{k \times k}](\eta - \eta_z) \\ &\leq \|\eta - \eta_z\| \left\{ \|\nabla H_{hk,z}^*(\eta_z)\| - \frac{a_1 \{\lambda_{\min} + O(h^\lambda k)\}}{2} \|\eta - \eta_z\| \right\} \\ &\leq \|\eta - \eta_z\| \{C_4(k^{-\alpha} + h^\lambda \sqrt{k}) - a_1 \{\lambda_{\min} + o(1)\} C_2(k^{-\alpha} + h^\lambda \sqrt{k})/2\}, \end{aligned}$$

where $J_{k \times k}$ is the $k \times k$ matrix of ones, $\bar{\eta}_z$ lies between η and η_z on $\partial\Lambda_z$, $a_1 = \inf_{u \in \Gamma_1} \{-\rho''(u)\} > 0$ uniformly in $z \in \mathcal{Z}$, $\lambda_{\min} = \inf_{z \in \mathcal{Z}} \lambda_{z, \min}$ with $\lambda_{z, \min}$ the smallest eigenvalue of $E\{\nu_k(X) \nu_k(X)^\top | Z = z\}$, and we used $O(h^\lambda k) = o(1)$, Lemma 1 and Condition (A3). Therefore, we choose

$$C_2 > \frac{2C_4}{a_1 \lambda_{\min}},$$

so that $H_{hk,z}^*(\eta) < H_{hk,z}^*(\eta_z)$, for $\eta \in \partial\Lambda_z$. Since $H_{hk,z}^*$ is continuous, there is a local maximum of $H_{hk,z}^*$ in the interior of Λ_z . On the other hand, $H_{hk,z}^*$ is a strictly concave function with the unique maximum at η_z^* . Therefore, we conclude that $\eta_z^* \in \Lambda_z^\circ$, i.e. $\|\eta_z^* - \eta_z\| < C_2(k^{-\alpha} + h^\lambda \sqrt{k})$. Note that C_2 is independent of z so that

$$\sup_{z \in \mathcal{Z}} \|\eta_z^* - \eta_z\| < C_2(k^{-\alpha} + h^\lambda \sqrt{k}). \quad (22)$$

Recalling (21) and using (22) and Condition (A3), for large enough k , we have

$$\begin{aligned} & \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\rho'\{\eta_z^{*\top} \nu_k(x)\} - \rho'\{\eta_z^\top \nu_k(x)\}| \\ &= \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\rho''\{\xi^*(z, x)\}| \|\eta_z^{*\top} \nu_k(x) - \eta_z^\top \nu_k(x)\| \\ &\leq a_2 \sup_{z \in \mathcal{Z}} \|\eta_z^* - \eta_z\| \sup_{x \in \mathcal{X}} \|\nu_k(x)\| \leq a_2 C_2(k^{-\alpha} + h^\lambda \sqrt{k})\zeta(k), \end{aligned} \quad (23)$$

where $\xi^*(z, x) \in \Gamma_1$ lies between $\eta_z^{*\top} \nu_k(x)$ and $\eta_z^\top \nu_k(x)$ and $a_2 = \sup_{u \in \Gamma_1} |\rho''(u)|$. Finally, using (16) and (23), we conclude that

$$\begin{aligned} & \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\pi_{hk}^*(z, x) - \pi_0(z, x)| \\ & \leq \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\pi_{hk}^*(z, x) - \rho' \{ \eta_z^\top \nu_k(x) \}| + \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\rho' \{ \eta_z^\top \nu_k(x) \} - \pi_0(z, x)| \\ & = O\{\zeta(k)(k^{-\alpha} + h^\lambda \sqrt{k})\}. \end{aligned}$$

■

Lemma 3 *Under Conditions (A1) to (A7), we have*

$$\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\widehat{\pi}_{hk}(z, x) - \pi_{hk}^*(z, x)| = O\left\{ \zeta(k) \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}k}{n\mu(h)}} \right\}.$$

Proof The key step is to provide the convergence rate of $\sup_{z \in \mathcal{Z}} \|\widehat{\eta}_z - \eta_z^*\|$, for which we need to provide the convergence rate of $\sup_{z \in \mathcal{Z}} \|\nabla H_{hk,z}(\eta_z^*)\|$. For all $z \in \mathcal{Z}$, we write

$$\begin{aligned} \widehat{S}_n &= \frac{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, z)/h\} \nu_k(X_j) \nu_k(X_j)^\top}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, z)/h\}} \\ &= \frac{E[K\{d(Z, z)/h\} \nu_k(X) \nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} + \frac{(n-1)E[K\{d(Z, z)/h\}]}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, z)/h\}} \\ &\quad \times \left[\frac{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, z)/h\} \nu_k(X_j) \nu_k(X_j)^\top}{(n-1)E[K\{d(Z, z)/h\}]} - \frac{E[K\{d(Z, z)/h\} \nu_k(X) \nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} \right] \\ &\quad - \left(1 - \frac{(n-1)E[K\{d(Z, z)/h\}]}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, z)/h\}} \right) \frac{E[K\{d(Z, z)/h\} \nu_k(X) \nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} \\ &\equiv \frac{E[K\{d(Z, z)/h\} \nu_k(X) \nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} + \frac{(n-1)E[K\{d(Z, z)/h\}]}{\sum_{j \in \mathcal{S}_{-i}} K\{d(Z_j, z)/h\}} \times \mathcal{A}_{1,z,k} \\ &\quad - \mathcal{A}_{2,z} \times \frac{E[K\{d(Z, z)/h\} \nu_k(X) \nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} \end{aligned} \tag{24}$$

It follows from Lemma 1 that

$$\sup_{z \in \mathcal{Z}} \mathcal{A}_{1,z,k} = O\left\{ \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}}{n\mu(h)}} \right\} J_{k \times k} \text{ and } \sup_{z \in \mathcal{Z}} \mathcal{A}_{2,z} = O\left\{ \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}}{n\mu(h)}} \right\}.$$

Using Lemma 1 again, we have

$$\sup_{z \in \mathcal{Z}} \frac{(n-1)E[K\{d(Z, z)/h\}]}{\sum_{i \in \mathcal{S}_{-i}} K\{d(Z_i, z)/h\}} \rightarrow 1,$$

and for any fixed k and all $z \in \mathcal{Z}$

$$\frac{E[K\{d(Z, z)/h\}\nu_k(X)\nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} \rightarrow E[\nu_k(X)\nu_k(X)^\top | Z = z],$$

as $n \rightarrow \infty$, whose eigenvalues are bounded away from zero and infinity uniformly in k and z by Condition (A3).

Combing the results above, we deduce that for all $z \in \mathcal{Z}$,

$$\hat{S}_n = \frac{E[K\{d(Z, z)/h\}\nu_k(X)\nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} + O\left\{\sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}}{n\mu(h)}}\right\} J_{k \times k}. \quad (25)$$

By Conditions (A3) and (A5), for h sufficiently small, there exist two positive constants s_1 and s_2 that are independent of z such that

$$\begin{aligned} 0 < s_1 &\leq \lambda_{\min}\left(\frac{E[K\{d(Z, z)/h\}\nu_k(X)\nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]}\right) \\ &\leq \lambda_{\max}\left(\frac{E[K\{d(Z, z)/h\}\nu_k(X)\nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]}\right) \leq s_2 < \infty, \end{aligned} \quad (26)$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of A , respectively. Consider the event set for all $z \in \mathcal{Z}$,

$$\begin{aligned} \mathcal{E}_n(z) &= \left\{(\eta - \eta_z^*)^\top \hat{S}_n (\eta - \eta_z^*) \right. \\ &\quad \left. > (\eta - \eta_z^*)^\top \left(\frac{E[K\{d(Z, z)/h\}\nu_k(X)\nu_k(X)^\top]}{E[K\{d(Z, z)/h\}]} - \frac{s_1}{2} I_{k \times k} \right) (\eta - \eta_z^*), \eta \neq \eta_z^* \right\}. \end{aligned}$$

Using (25) and (26), we deduce that for h sufficiently small,

$$P(\mathcal{E}_n(z)) = 1. \quad (27)$$

On the other hand, noting that by defition of η_z^* ,

$$\frac{E[\rho' \{\eta_z^{*\top} \nu_k(X)\} K\{d(Z, z)/h\} \nu_k(X)]}{E[K\{d(Z, z)/h\}]} = E\{\nu_k(X)\},$$

we have

$$\begin{aligned}
 & \sup_{z \in \mathcal{Z}} \|\nabla H_{hk,z}(\eta_z^*)\| \\
 &= \sup_{z \in \mathcal{Z}} \left\| \frac{\sum_{j \in \mathcal{S}_{-i}} \rho' \{ \eta_z^{*\top} \nu_k(X_j) \} K \{ d(Z_j, z)/h \} \nu_k(X_j)}{\sum_{j \in \mathcal{S}_{-i}} K \{ d(Z_j, z)/h \}} - \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j) \right\| \\
 &\leq \sup_{z \in \mathcal{Z}} \left\| \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \frac{\rho' \{ \eta_z^{*\top} \nu_k(X_j) \} K \{ d(Z_j, z)/h \} \nu_k(X_j)}{E[K \{ d(Z, z)/h \}]} \right. \\
 &\quad \left. - \frac{E[\rho' \{ \eta_z^{*\top} \nu_k(X) \} K \{ d(Z, z)/h \} \nu_k(X)]}{E[K \{ d(Z, z)/h \}]} \right\| + \left\| E\{\nu_k(X)\} - \frac{1}{n-1} \sum_{j \in \mathcal{S}_{-i}} \nu_k(X_j) \right\| \\
 &\quad + \sup_{z \in \mathcal{Z}} \left\| \frac{\sum_{j \in \mathcal{S}_{-i}} \rho' \{ \eta_z^{*\top} \nu_k(X_j) \} K \{ d(Z_j, z)/h \} \nu_k(X_j)}{\sum_{j \in \mathcal{S}_{-i}} K \{ d(Z_j, z)/h \}} \left(\frac{\sum_{j \in \mathcal{S}_{-i}} K \{ d(Z_j, z)/h \}}{(n-1)E[K \{ d(Z, z)/h \}]} - 1 \right) \right\| \\
 &= O \left\{ \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \right\} + O \left(\sqrt{\frac{k}{n}} \right) + O \left\{ \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \right\} = O \left\{ \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \right\},
 \end{aligned}$$

where Lemma 1 is used for the second equality. This means that there exists a constant $C_3 > 0$ that is independent of z such that

$$\sup_{z \in \mathcal{Z}} \|\nabla H_{hk,z}(\eta_z^*)\| < C_3 \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}}. \quad (28)$$

For some $C_4 > 0$ (to be chosen later), define the set

$$\hat{\Lambda}_z = \left\{ \eta \in \mathbb{R}^k : \|\eta - \eta_z^*\| \leq C_4 C_3 \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \right\}.$$

For all $\eta \in \hat{\Lambda}_z$, $x \in \mathcal{X}$, and for sufficiently large n and k , by Condition (A3), (17) and (22), we have

$$\begin{aligned}
 & \|\eta^\top \nu_k(x) - \eta_z^{*\top} \nu_k(x)\| \leq \|\eta - \eta_z^*\| \|\nu_k(x)\| \leq C_4 C_3 \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \zeta(k) \\
 & \Rightarrow \eta^\top \nu_k(x) \\
 & \in \left[\eta_z^{*\top} \nu_k(x) - C_4 C_3 \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \zeta(k), \eta_z^{*\top} \nu_k(x) + C_4 C_3 \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \zeta(k) \right] \\
 & \subset \left[\rho'^{(-1)}(c_2) - C_1 k^{-\alpha} - C_2(k^{-\alpha} + h^\lambda \sqrt{k}) - C_4 C_3 \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \zeta(k), \right. \\
 & \quad \left. \rho'^{(-1)}(c_1) + C_1 k^{-\alpha} + C_2(k^{-\alpha} + h^\lambda \sqrt{k}) + C_4 C_3 \sqrt{\frac{\psi_{\mathcal{Z}} \{ (\log n)/n \} k}{n\mu(h)}} \zeta(k) \right] \\
 & \subset \Gamma_1 \equiv [\rho'^{(-1)}(c_2) - 1, \rho'^{(-1)}(c_1) + 1]. \quad (29)
 \end{aligned}$$

By the mean value theorem, for any $\eta \in \partial\widehat{\Lambda}_z$, we have, based on the fact that $P(\mathcal{E}_n(z)) = 1$ from (27),

$$\begin{aligned}
& H_{hk,z}(\eta) - H_{hk,z}(\eta_z^*) \\
&= (\eta - \eta_z^*)^\top \nabla H_{hk,z}(\eta_z^*) + \frac{1}{2}(\eta - \eta_z^*)^\top \nabla^2 H_{hk,z}(\bar{\eta})(\eta - \eta_z^*) \\
&\leq \|\eta - \eta_z^*\| \|\nabla H_{hk,z}(\eta_z^*)\| - \frac{a}{2}(\eta - \eta_z^*)^\top \widehat{S}_n(\eta - \eta_z^*) \\
&< \|\eta - \eta_z^*\| \|\nabla H_{hk,z}(\eta_z^*)\| - \frac{a}{2}(\eta - \eta_z^*)^\top \left[\frac{E[K\{d(Z,z)/h\}\nu_k(X)\nu_k(X)^\top]}{E[K\{d(Z,z)/h\}]} - \frac{s_1}{2}I_{k \times k} \right] (\eta - \eta_z^*) \\
&\leq \|\eta - \eta_z^*\| \|\nabla H_{hk,z}(\eta_z^*)\| - \frac{a}{2}(\eta - \eta_z^*)^\top \left(s_1 I_{k \times k} - \frac{s_1}{2} I_{k \times k} \right) (\eta - \eta_z^*) \\
&\leq \|\eta - \eta_z^*\| \left(\|\nabla H_{hk,z}(\eta_z^*)\| - \frac{a}{4}s_1 \|\eta - \eta_z^*\| \right), \tag{30}
\end{aligned}$$

where $\bar{\eta} \in \partial\widehat{\Lambda}_z$ lies between η and η_z^* , $a = \inf_{u \in \Gamma_1} \{-\rho''(u)\} > 0$ and the third inequality follows from (26). Therefore, we choose

$$C_4 > \frac{4}{as_1},$$

so that $H_{hk,z}(\eta) < H_{hk,z}(\eta_z)$, for $\eta \in \partial\widehat{\Lambda}_z$. On the other hand, we know that $\widehat{\eta}_z$ is the unique maximum point of $H_{hk,z}(\eta)$. It follows that $\widehat{\eta}_z \in \widehat{\Lambda}_z^o(\epsilon)$, i.e. $\|\widehat{\eta}_z - \eta_z^*\| \leq C_4 C_3 \sqrt{\psi_{\mathcal{Z}}\{(\log n)/n\}k/n\mu(h)}$. Noting that C_4 and C_3 are independent of z , we conclude that

$$\sup_{z \in \mathcal{Z}} \|\widehat{\eta}_z - \eta_z^*\| = O\left\{ \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}k}{n\mu(h)}} \right\}. \tag{31}$$

Finally, we are able to prove the convergence rate for $\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\widehat{\pi}_{hk}(z,x) - \pi_{hk}^*(z,x)|$. By the mean value theorem, we have

$$\widehat{\pi}_{hk}(z,x) - \pi_{hk}^*(z,x) = \rho'\{\widehat{\eta}_z^\top \nu_k(x)\} - \rho'\{\eta_z^{*\top} \nu_k(x)\} = \rho''\{\bar{\eta}_z^\top \nu_k(x)\}(\widehat{\eta}_z - \eta_z^*)^\top \nu_k(x),$$

where $\bar{\eta}_z^\top$ lies between $\widehat{\eta}_z$ and η_z^* . From (29) and (31), we have

$$\sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\rho''\{\bar{\eta}_z^\top \nu_k(x)\}| = O(1). \tag{32}$$

It follows that by combining (31), (32) and Condition (A3),

$$\begin{aligned}
& \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\widehat{\pi}_{hk}(z,x) - \pi_{hk}^*(z,x)| \\
&\leq \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\rho''\{\bar{\eta}_z^\top \nu_k(x)\}| \sup_{z \in \mathcal{Z}} \|\widehat{\eta}_z - \eta_z^*\| \sup_{x \in \mathcal{X}} \|\nu_k(x)\| = O\left\{ \zeta(k) \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\}k}{n\mu(h)}} \right\}.
\end{aligned}$$

■

Appendix E. Proof of Theorem 2

To aid the proof of Theorem 2, we need the following lemma.

Lemma 4 *Let $\mathcal{E}_d = \mathcal{E}_d(n) = \{\lambda_d/2 \geq \|\widehat{G} - G\|_O\}$, where $\|M\|_O = \int \int M^2(s, t) ds dt$ for a symmetric bivariate function M . Under Conditions (B1) to (B3), we have*

$$P(\mathcal{E}_q) \rightarrow 1,$$

as $n \rightarrow \infty$.

Proof This follows from Section 5.1 of Hall and Horowitz (2007). ■

We give the formal proof of Theorem 2.

Proof From (5.2) of Hall and Horowitz (2007), we know that $\sup_j |\widehat{\lambda}_j - \lambda_j| \leq \|\widehat{G} - G\|_O$, where $\|\cdot\|_O$ is defined in Lemma 4. Let $\mathcal{E}_q = \{\lambda_q/2 \geq \|\widehat{G} - G\|_O\}$. We conclude that

$$\frac{1}{2}\lambda_j \leq \widehat{\lambda}_j \leq \frac{3}{2}\lambda_j, \quad (33)$$

for $j = 1, \dots, q$, provided that the event \mathcal{E}_q holds. We see from Lemma 4 that $P(\mathcal{E}_q) \rightarrow 1$ as $n \rightarrow \infty$. Therefore, since our result is probabilistic, we can argue under the assumption that the event \mathcal{E}_q holds.

Recall that $\widehat{b}_{\text{FSW}} = \sum_{j=1}^q \widehat{b}_{\text{FSW},j} \widehat{\phi}_j$, where $\widehat{b}_{\text{FSW},j} = \widehat{\lambda}_j^{-1} \widehat{e}_{\text{FSW},j}$. Write $\widehat{b}_{\text{FSW},j} = \check{b}_j + \widehat{\lambda}_j^{-1}(S_{j1} + S_{j2} + S_{j3})$, where $\check{b}_j = \widehat{\lambda}_j^{-1} \int_{\mathcal{T}} e(t) \phi_j(t) dt$, $S_{j1} = \int_{\mathcal{T}} \{\widehat{e}_{\text{FSW}}(t) - e(t)\} \phi_j(t) dt$, $S_{j2} = \int_{\mathcal{T}} e(t) \{\widehat{\phi}_j(t) - \phi_j(t)\} dt$ and $S_{j3} = \int_{\mathcal{T}} \{\widehat{e}_{\text{FSW}}(t) - e(t)\} \{\widehat{\phi}_j(t) - \phi_j(t)\} dt$. Using (33) and the fact that $S_{j3}^2 \leq \|\widehat{e}_{\text{FSW}} - e\|^2 \|\widehat{\phi}_j - \phi_j\|^2$, we have

$$\begin{aligned} \sum_{j=1}^q (\widehat{b}_{\text{FSW},j} - \check{b}_j)^2 &\leq 3 \sum_{j=1}^q \widehat{\lambda}_j^{-2} (S_{j1}^2 + S_{j2}^2 + S_{j3}^2) \\ &\leq 12 \sum_{j=1}^q \lambda_j^{-2} (S_{j1}^2 + S_{j2}^2 + S_{j3}^2) \\ &\leq 12 \sum_{j=1}^q \lambda_j^{-2} (S_{j1}^2 + S_{j2}^2) + 12 \sum_{j=1}^q \lambda_j^{-2} \|\widehat{e}_{\text{FSW}} - e\|^2 \|\widehat{\phi}_j - \phi_j\|^2, \end{aligned} \quad (34)$$

Also, we have

$$\begin{aligned} \int_{\mathcal{T}} \left\{ \sum_{j=1}^q b_j \widehat{\phi}_j(t) - b(t) \right\}^2 dt &\leq 2 \int_{\mathcal{T}} \left[\sum_{j=1}^q b_j \{\widehat{\phi}_j(t) - \phi_j(t)\} \right]^2 dt + 2 \sum_{j=q+1}^{\infty} b_j^2 \\ &\leq 2q \sum_{j=1}^q b_j^2 \|\widehat{\phi}_j - \phi_j\|^2 + 2 \sum_{j=q+1}^{\infty} b_j^2. \end{aligned} \quad (35)$$

Using (34) and (35), we have

$$\begin{aligned}
& \int_{\mathcal{T}} \{\widehat{b}_{\text{FSW}}(t) - b(t)\}^2 dt \\
&= \int_{\mathcal{T}} \left\{ \sum_{j=1}^q \widehat{b}_{\text{FSW},j} \widehat{\phi}_j(t) - \sum_{j=1}^q \check{b}_j \widehat{\phi}_j(t) + \sum_{j=1}^q \check{b}_j \widehat{\phi}_j(t) - \sum_{j=1}^q b_j \widehat{\phi}_j(t) + \sum_{j=1}^q b_j \widehat{\phi}_j(t) - b(t) \right\}^2 dt \\
&\leq 3 \sum_{j=1}^q (\widehat{b}_{\text{FSW},j} - \check{b}_j)^2 + 3 \sum_{j=1}^q (\check{b}_j - b_j)^2 + 3 \int_{\mathcal{T}} \left\{ \sum_{j=1}^q b_j \widehat{\phi}_j(t) - b(t) \right\}^2 dt \\
&\leq 36 \sum_{j=1}^q \lambda_j^{-2} (S_{j1}^2 + S_{j2}^2) + 36 \sum_{j=1}^q (\lambda_j^{-2} \|\widehat{e}_{\text{FSW}} - e\|^2 + q b_j^2) \|\widehat{\phi}_j - \phi_j\|^2 + 6 \sum_{j=q+1}^{\infty} b_j^2 \\
&\quad + o_p(n^{-(2\beta-1)/(\gamma+2\beta)}), \tag{36}
\end{aligned}$$

where $\sum_{j=1}^q (\check{b}_j - b_j)^2 = o_p(n^{-(2\beta-1)/(\gamma+2\beta)})$ is used for the last inequality. This can be proved following the same argument of proving (5.11) in Hall and Horowitz (2007) with the response variable Y replaced by $Y\pi_0(Z, X)$. Also, using $q \asymp n^{1/(\gamma+2\beta)}$ and Condition (B3), we have $\sum_{j=q+1}^{\infty} b_j^2 \asymp \int_q^{\infty} t^{-2\beta} dt = O(n^{-(2\beta-1)/(\gamma+2\beta)})$. From Condition (B2), we know that there exists a constant $C_1 > 1$ such that $\lambda_j \geq C_1 j^{-\gamma}$. From (5.21) and (5.22) of Hall and Horowitz (2007), we know that $\|\widehat{\phi}_j - \phi_j\|^2 = O_p(j^2 n^{-1})$. Finally, using (5.30) and (5.31) of Hall and Horowitz (2007), we have $\sum_{j=1}^q \lambda_j^{-2} S_{j2}^2 = O_p(n^{-(2\beta-1)/(\gamma+2\beta)})$. Combining these results with (36), we deduce that there exists a constant $C_2 > 0$ such that

$$\int_{\mathcal{T}} \{\widehat{b}_{\text{FSW}}(t) - b(t)\}^2 dt \leq C_2 \sum_{j=1}^q \lambda_j^{-2} (S_{j1}^2 + \|\widehat{e}_{\text{FSW}} - e\|^2 j^2 n^{-1}) + O_p(n^{-(2\beta-1)/(\gamma+2\beta)}) \tag{37}$$

It remains to provide the convergence rates of S_{j1}^2 and $\|\widehat{e}_{\text{FSW}} - e\|^2$.

Recall that

$$\widehat{e}_{\text{FSW}}(t) = \frac{1}{n} \sum_{i=1}^n \{\widehat{\pi}_{hk}(Z_i, X_i) Y_i - \widehat{\mu}_{Y,\pi}\} \{Z_i(t) - \widehat{\mu}(t)\},$$

where $\widehat{\mu}_{Y,\pi} = \sum_{i=1}^n \widehat{\pi}_{hk}(Z_i, X_i) Y_i / n$. Let

$$\widetilde{e}(t) = \frac{1}{n} \sum_{i=1}^n \{\pi_0(Z_i, X_i) Y_i - \widetilde{\mu}_{Y,\pi}\} \{Z_i(t) - \widehat{\mu}(t)\},$$

where $\widetilde{\mu}_{Y,\pi} = \sum_{i=1}^n \pi_0(Z_i, X_i) Y_i / n$. Also, recall from the Karhunen-Loève expansion that $Z_i(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_{ij} \phi_j(t)$, where the PC scores ξ_{ij} satisfy $E(\xi_{ij}) = 0$, $E(\xi_{ij}^2) = \lambda_j$ and $E(\xi_{ij} \xi_{ij'}) = 0$ for $j \neq j'$.

Defining $\bar{\epsilon} = \sum_{i=1}^n \epsilon_i/n$ and $\bar{\xi}_j = \sum_{i=1}^n \xi_{ij}/n$, we can write

$$\begin{aligned}
 S_{j1} &= \int_{\mathcal{T}} \{\widehat{e}_{\text{FSW}}(t) - \widetilde{e}(t) + \widetilde{e}(t) - e(t)\} \phi_j(t) dt \\
 &= \frac{1}{n} \sum_{i=1}^n \{Y_i \widehat{\pi}_{hk}(Z_i, X_i) - Y_i \pi_0(Z_i, X_i) - \widehat{\mu}_{Y,\pi} + \widetilde{\mu}_{Y,\pi}\} \int_{\mathcal{T}} \{Z_i(t) - \widehat{\mu}(t)\} \phi_j(t) dt \\
 &\quad + \frac{1}{n} \sum_{i=1}^n \{Y_i \pi_0(Z_i, X_i) - \widetilde{\mu}_{Y,\pi}\} \int_{\mathcal{T}} \{Z_i(t) - \widehat{\mu}(t)\} \phi_j(t) dt - E[\{Y \pi_0(Z, X) - \mu_{Y,\pi}\} \xi_j] \\
 &= \frac{1}{n} \sum_{i=1}^n Y_i \{\widehat{\pi}_{hk}(Z_i, X_i) - \pi_0(Z_i, X_i)\} (\xi_{ij} - \bar{\xi}_j) \\
 &\quad + \frac{1}{n} \sum_{i=1}^n \left[\int_{\mathcal{T}} b(t) \{Z_i(t) - \widehat{\mu}(t)\} dt + \epsilon_i - \bar{\epsilon} \right] (\xi_{ij} - \bar{\xi}_j) - E[\{Y \pi_0(Z, X) - \mu_{Y,\pi}\} \xi_j] \\
 &= A_{j1} + A_{j2} + A_{j3} + A_{j4}, \tag{38}
 \end{aligned}$$

where

$$\begin{aligned}
 A_{j1} &= \frac{1}{n} \sum_{i=1}^n Y_i \{\widehat{\pi}_{hk}(Z_i, X_i) - \pi_0(Z_i, X_i)\} (\xi_{ij} - \bar{\xi}_j), \\
 A_{j2} &= \frac{1}{n} \sum_{i=1}^n \left[\xi_{ij} \int_{\mathcal{T}} b(t) Z_i(t) dt - E \left\{ \xi_{ij} \int_{\mathcal{T}} b(t) Z_i(t) dt \right\} \right], \\
 A_{j3} &= \frac{1}{n} \sum_{i=1}^n \{\xi_{ij} \epsilon_i - E(\xi_{ij} \epsilon_i)\}, \\
 A_{j4} &= -\bar{\xi}_j \int_{\mathcal{T}} b(t) \widehat{\mu}(t) dt - \bar{\xi}_j \bar{\epsilon}.
 \end{aligned}$$

We next provide the convergence rates for the A_{jk} 's sequentially.

First note that $E(\xi_{ij} - \bar{\xi}_j)^2 = (n-1)E(\xi_j^2)/n = O(\lambda_j)$ by Condition (B2). It follows that

$$\frac{1}{n} \sum_{i=1}^n (\xi_{ij} - \bar{\xi}_j)^2 = O_p(\lambda_j). \tag{39}$$

Using the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
 |A_{j1}| &\leq \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\widehat{\pi}_{hk}(z, x) - \pi_0(z, x)| \left(\frac{1}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^n (\xi_{ij} - \bar{\xi}_j)^2 \right)^{1/2} \\
 &= O \left[\zeta(k) \left\{ k^{-\alpha} + h^\lambda \sqrt{k} + \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\} k}{n\mu(h)}} \right\} \right] \cdot O_p(1) \cdot O_p(\lambda_j^{1/2}) \\
 &= O_p \left[\lambda_j^{1/2} \zeta(k) \left\{ k^{-\alpha} + h^\lambda \sqrt{k} + \sqrt{\frac{\psi_{\mathcal{Z}}\{(\log n)/n\} k}{n\mu(h)}} \right\} \right], \tag{40}
 \end{aligned}$$

based on Theorem 1 and (39).

For A_{j2} , we have

$$\text{var}(A_{j2}) \leq \frac{1}{n} E \left\{ \xi_{ij}^2 \int_{\mathcal{T}} b^2(t) Z_i^2(t) dt \right\} \leq \frac{1}{n} \{E(\xi_j^4)\}^{1/2} \left[E \left\{ \int_{\mathcal{T}} b^4(t) Z^4(t) dt \right\} \right]^{1/2} = O\left(\frac{\lambda_j}{n}\right),$$

where the last equality follows from Condition (B2). Similarly, we have

$$\text{var}(A_{j3}) \leq \frac{1}{n} E(\xi_j^2 \epsilon^2) \leq \frac{1}{n} \sqrt{E(\xi_j^4) E(\epsilon^4)} = O\left(\frac{\lambda_j}{n}\right),$$

where the last equality follows from Conditions (B2) and (C1). It follows that

$$A_{j2} = O_p\left(\sqrt{\frac{\lambda_j}{n}}\right), \quad A_{j3} = O_p\left(\sqrt{\frac{\lambda_j}{n}}\right). \quad (41)$$

Note that $\hat{\mu} \rightarrow \mu$ and $\bar{\epsilon} \rightarrow E(\epsilon)$ a.s. as $n \rightarrow \infty$ by the law of large numbers, and μ and $E(\epsilon)$ are bounded. It follows that $\int_{\mathcal{T}} b(t) \hat{\mu}(t) dt$ and $\bar{\epsilon}$ are bounded for large n . Thus, there exists a constant $C_3 > 0$ such that for n sufficiently large,

$$|A_{j4}| \leq C_3 \bar{\xi}_j = O_p\left\{\sqrt{E(\bar{\xi}_j^2)}\right\} = O_p\left(\sqrt{\frac{\lambda_j}{n}}\right). \quad (42)$$

Combing (38), (39), (41) and (42), we conclude that

$$S_{j1}^2 = O_p\left[\lambda_j \zeta^2(k) \left\{k^{-2\alpha} + h^{2\lambda} k + \frac{\psi_{\mathcal{Z}}\{(\log n)/n\} k}{n\mu(h)}\right\}\right]. \quad (43)$$

As for $\|\hat{e}_{\text{FSW}} - e\|^2$, we have

$$\|\hat{e}_{\text{FSW}} - e\|^2 \leq 2\|\hat{e}_{\text{FSW}} - \tilde{e}\|^2 + 2\|\tilde{e} - e\|^2,$$

where

$$\begin{aligned} & |\hat{e}_{\text{FSW}}(t) - \tilde{e}(t)| \\ &= \left| \frac{1}{n} \sum_{i=1}^n Y_i \{\hat{\pi}_{hk}(Z_i, X_i) - \pi_0(Z_i, X_i)\} \{Z_i(t) - \hat{\mu}(t)\} \right| \\ &\leq \sup_{(z,x) \in \mathcal{Z} \times \mathcal{X}} |\hat{\pi}_{hk}(z, x) - \pi_0(z, x)| \left(\frac{1}{n} \sum_{i=1}^n Y_i^2 \right)^{1/2} \left[\frac{1}{n} \sum_{i=1}^n \{Z_i(t) - \hat{\mu}(t)\}^2 \right]^{1/2}. \end{aligned}$$

It follows by using Theorem 1 and the fact that $\sum_{i=1}^n \int_{\mathcal{T}} \{Z_i(t) - \hat{\mu}(t)\}^2 dt/n \rightarrow \int_{\mathcal{T}} \text{var}\{Z(t)\} dt = O(1)$ as $n \rightarrow \infty$ that

$$\|\hat{e}_{\text{FSW}} - \tilde{e}\|^2 = O_p\left[\zeta^2(k) \left\{k^{-2\alpha} + h^{2\lambda} k + \frac{\psi_{\mathcal{Z}}\{(\log n)/n\} k}{n\mu(h)}\right\}\right].$$

Also, the standard result on the convergence rate of the empirical mean estimator gives $\|\tilde{e} - e\|^2 = O_p(n^{-1})$. Therefore, we conclude that

$$\|\hat{e}_{\text{FSW}} - e\|^2 = O_p\left[\zeta^2(k) \left\{k^{-2\alpha} + h^{2\lambda} k + \frac{\psi_{\mathcal{Z}}\{(\log n)/n\} k}{n\mu(h)}\right\}\right]. \quad (44)$$

The proof is completed by combining (37), (43) and (44). ■

References

- Chunrong Ai, Oliver Linton, Kaiji Motegi, and Zheng Zhang. A unified framework for efficient estimation of general treatment models. *Quantitative Economics*, 12(3):779–816, 2021.
- Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: de-biased inference of average treatment effects in high dimensions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):597–623, 2018.
- Matteo Bonvini and Edward H Kennedy. Fast convergence rates for dose-response estimation. *arXiv preprint arXiv:2207.11825*, 2022.
- Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.
- Hervé Cardot, André Mas, and Pascal Sarda. Clt in functional linear regression models. *Probability Theory and Related Fields*, 138(3):325–361, 2007.
- Kwun Chuen Gary Chan, Sheung Chi Phillip Yam, and Zheng Zhang. Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3):673–700, 2016.
- Jingxiang Chen, Haoda Fu, Xuanyao He, Michael R Kosorok, and Yufeng Liu. Estimating individualized treatment rules for ordinal treatments. *Biometrics*, 74(3):924–933, 2018.
- Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. *Handbook of Econometrics*, 6:5549–5632, 2007.
- Adam Ciarleglio, Eva Petkova, R Todd Ogden, and Thaddeus Tarpey. Treatment decisions based on scalar and functional baseline covariates. *Biometrics*, 71(4):884–894, 2015.
- Adam Ciarleglio, Eva Petkova, Thaddeus Tarpey, and R Todd Ogden. Flexible functional regression methods for estimating individualized treatment rules. *Stat*, 5(1):185–199, 2016.
- Adam Ciarleglio, Eva Petkova, Todd Ogden, and Thaddeus Tarpey. Constructing treatment decision rules based on scalar and functional predictors when moderators of treatment effect are unknown. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 67(5):1331–1356, 2018.
- Adam Ciarleglio, Eva Petkova, and Ofer Harel. Elucidating age and sex-dependent association between frontal eeg asymmetry and depression: An application of multiple imputation in functional regression. *Journal of the American Statistical Association*, 117(537):12–26, 2022.
- Aurore Delaigle and Peter Hall. Defining probability density for a distribution of random functions. *The Annals of Statistics*, 38(2):1171–1193, 2010.

- Peng Ding, Avi Feller, and Luke Miratrix. Decomposing treatment effect variation. *Journal of the American Statistical Association*, 114(525):304–317, 2019.
- Alexander D’Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2): 644–654, 2021.
- Frédéric Ferraty and Philippe Vieu. *Nonparametric functional data analysis: theory and practice*, volume 76. Springer, 2006.
- Frédéric Ferraty, Ali Laksaci, Amel Tadj, and Philippe Vieu. Rate of uniform consistency for nonparametric estimates with functional variables. *Journal of Statistical planning and inference*, 140(2):335–352, 2010.
- C. Fong, C. Hazlett, and K. Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12:156–177, 2018.
- Antonio F Galvao and Liang Wang. Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment. *Journal of the American Statistical Association*, 110(512):1528–1542, 2015.
- Wenchuan Guo, Xiao-Hua Zhou, and Shujie Ma. Estimation of optimal individualized treatment rules using a covariate-specific treatment effect curve with high-dimensional covariates. *Journal of the American Statistical Association*, 116(533):309–321, 2021.
- Peter Hall and Joel L Horowitz. Methodology and convergence rates for functional linear regression. *The Annals of Statistics*, 35(1):70–91, 2007.
- Lars Peter Hansen, John Heaton, and Amir Yaron. Finite-sample properties of some alternative gmm estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.
- Keisuke Hirano and Guido W Imbens. The propensity score with continuous treatments. *Applied Bayesian Modeling and Causal Inference from Incomplete-data Perspectives*, 226164: 73–84, 2004.
- Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.
- K. Imai and D. A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99:854–866, 2004.
- Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.
- Masaaki Imaizumi and Kengo Kato. A simple method to construct confidence bands in functional linear regression. *Statistica Sinica*, 29(4):2055–2081, 2019.
- Guido W Imbens, Richard H Spady, and Phillip Johnson. Information theoretic approaches to inference in moment condition models. *Econometrica*, 66(2):333, 1998.

- Edward H Kennedy. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 114(526):645–656, 2019.
- Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4):1229–1245, 2017.
- Eric B Laber and Ana-Maria Staicu. Functional feature construction for individualized treatment regimes. *Journal of the American Statistical Association*, 113(523):1219–1227, 2018.
- Eric B Laber and Ying-Qi Zhao. Tree-based methods for individualized treatment regimes. *Biometrika*, 102(3):501–514, 2015.
- Fan Li and Fan Li. Propensity score weighting for causal inference with multiple treatments. *The Annals of Applied Statistics*, 13(4):2389–2415, 2019.
- Wenbo V Li and Q-M Shao. Gaussian processes: inequalities, small ball probabilities and applications. *Handbook of Statistics*, 19:533–597, 2001.
- Xinyi Li and Michael R Kosorok. Functional individualized treatment regimes with imaging features. *arXiv preprint arXiv:2304.13003*, 2023.
- Yehua Li, Naisyin Wang, and Raymond J Carroll. Generalized functional linear models with semiparametric single-index interactions. *Journal of the American Statistical Association*, 105(490):621–633, 2010.
- Zhenhua Lin and Fang Yao. Functional regression on the manifold with contamination. *Biometrika*, 108(1):167–181, 2021.
- Zhexiao Lin, Peng Ding, and Fang Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217, 2023.
- Nengxiang Ling and Philippe Vieu. Nonparametric modelling for functional data: selected survey and tracks for future. *Statistics*, 52(4):934–949, 2018.
- Michael J Lopez and Roe Gutman. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454, 2017a.
- Michael J Lopez and Roe Gutman. Estimating the average treatment effects of nutritional label use using subclassification with regression adjustment. *Statistical methods in medical research*, 26(2):839–864, 2017b.
- Daniel J Lockett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.
- Xinwei Ma and Jingshen Wang. Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115(532):1851–1860, 2020.

- Ian W McKeague and Min Qian. Estimation of treatment policies based on functional predictors. *Statistica Sinica*, 24(3):1461, 2014.
- Erica EM Moodie and David A Stephens. Estimation of dose–response functions for longitudinal data using the generalised propensity score. *Statistical methods in medical research*, 21(2):149–166, 2012.
- Erica EM Moodie, Nema Dean, and Yue Ru Sun. Q-learning: Flexible learning about useful utilities. *Statistics in Biosciences*, 6:223–243, 2014.
- Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2: 321–359, 2015.
- Hans-Georg Müller and Fang Yao. Functional additive models. *Journal of the American Statistical Association*, 103(484):1534–1544, 2008.
- Whitney K Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.
- Art B Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249, 1988.
- Hyung Park, Eva Petkova, Thaddeus Tarpey, and R Todd Ogden. Functional additive models for optimizing individualized treatment rules. *Biometrics*, 79(1):113–126, 2023.
- James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 95–133. Springer, 2000.
- James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.
- P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:45–55, 1983.
- Phillip J Schulte, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science: a Review Journal of the Institute of Mathematical Statistics*, 29(4):640, 2014.
- Hyejin Shin. Partial functional linear regression. *Journal of Statistical Planning and Inference*, 139(10):3405–3418, 2009.
- Zhiqiang Tan. Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *The Annals of Statistics*, 48(2):811–837, 2020.
- Paul Tseng and Dimitri P Bertsekas. Relaxation methods for problems with strictly convex costs and linear constraints. *Mathematics of Operations Research*, 16(3):462–481, 1991.
- Jane-Ling Wang, Jeng-Min Chiou, and Hans-Georg Müller. Functional data analysis. *Annual Review of Statistics and its Application*, 3:257–295, 2016.

- Jiayi Wang, Raymond KW Wong, Xiaoke Zhang, and Kwun Chuen Gary Chan. Flexible functional treatment effect estimation. *arXiv preprint arXiv:2309.08039*, 2023.
- Xiaoke Zhang and Jane-Ling Wang. From sparse to dense functional data and beyond. *The Annals of Statistics*, 44(5):2281–2321, 2016.
- Xiaoke Zhang, Wu Xue, and Qiyue Wang. Covariate balancing functional propensity score for functional treatments in cross-sectional observational studies. *Computational Statistics & Data Analysis*, 163:107303, 2021.
- Ying-Qi Zhao, Donglin Zeng, Eric B Laber, and Michael R Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.
- Hang Zhou, Fang Yao, and Huiming Zhang. Functional linear regression for discretely observed data: from ideal to reality. *Biometrika*, 110(2):381–393, 2023.