Biometrika (2016), *pp*. 1–14

Printed in Great Britain

Advance Access publication on xx xx xxxx

Kernel-based covariate functional balancing for observational studies

BY RAYMOND K. W. WONG

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.

raywong@stat.tamu.edu

AND KWUN CHUEN GARY CHAN

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

kcgchan@u.washington.edu

SUMMARY

Covariate balance is often advocated for objective causal inference since it mimics randomization in observational data. Unlike methods that balance specific moments of covariates, our proposal attains uniform approximate balance for covariate functions in a reproducing-kernel Hilbert space. The corresponding infinite-dimensional optimization problem is shown to have a finite-dimensional representation in terms of an eigenvalue optimization problem. Large-sample results are studied, and numerical examples show that the proposed method achieves better balance, with smaller sampling variability than existing methods.

Some key words: Average treatment effect; Eigenvalue optimization; Reproducing-kernel Hilbert space; Sobolev space.

1. INTRODUCTION

The estimation of average treatment effects is important in the evaluation of an intervention or a treatment, but is complicated by confounding in observational studies where the treatment is not randomly assigned. When treatment assignment is unconfounded conditional on observable covariates, two popular modeling strategies are based respectively on propensity score modeling (Rosenbaum & Rubin, 1983) and outcome regression modeling. Parametric approaches can suffer seriously from model misspecification, and there have been substantial recent efforts to construct more robust estimators within these modeling frameworks; see for example, Robins et al. (1994), Qin & Zhang (2007), Tan (2010), Graham et al. (2012), and Han & Wang (2013).

Since randomization is a gold standard to identify average treatment effects, Rubin (2007) advocated mimicking randomization, which balances the covariate distributions among the treated, the controls, and the combined sample, in the analysis of observational data. Based on these considerations, weighting-based covariate balancing methods have been proposed by Qin & Zhang (2007), Hainmueller (2012), Imai & Ratkovic (2014), Zubizarreta (2015) and Chan et al. (2016). A common feature of these methods is that a vector of user-specified functions of covariates is balanced. While balancing low-order moments of the covariates often yields good results, there is no guarantee that there will be sufficient balance over a large class of covariate functions. Matching is another general idea to attain covariate balance. Exact matching is not feasible for multiple continuous covariates, and a user-specified coarsening of the covariate space is needed (Iacus et al., 2011). In this paper, we shall focus on weighting-based methods.

Instead of balancing pre-specified moments of covariates, we propose a method to control the covariate functional balance over a reproducing-kernel Hilbert space (Aronszajn, 1950), which can be chosen large enough to contain any functions with mild smoothness constraints, including non-linearities and interactions. At a conceptual level, the comparison between covariate balancing with an increasing number of basis functions and kernel-based covariate functional balancing is analogous to the comparison of regression and smoothing splines in conditional mean estimation. Unlike regression splines, smoothing splines do not require pre-selection of the number of knots and their locations. Although achieving our goal involves a challenge due to an infinite-dimensional optimization problem, we show that it has a finite-dimensional representation and can be solved by eigenvalue optimization. Large sample properties are derived under minimal smoothness conditions on the outcome regression model. Consistent estimation of average treatment effects is then possible without first guessing or estimating the outcome regression function, and efficient estimation can be attained when the outcome regression function is estimated. Unlike weighting methods that require stringent smoothness conditions for the propensity score function, our method does not require smoothness of the propensity score.

2. KERNEL-BASED COVARIATE FUNCTIONAL BALANCING

2·1. *Preliminaries*

Let $Y(1)$ and $Y(0)$ be the potential outcomes when an individual is assigned to the treatment or control group respectively. We are interested in estimating the population average treatment effect $\tau = E\{Y(1) - Y(0)\}$. In practice, $Y(1)$ and $Y(0)$ are not both observed. With T the binary treatment indicator, we can represent the observed outcome as $Y = TY(1) + (1-T)Y(0)$. Moreover, we observe a vector of covariates $X \in \mathcal{X}$ for every individual, so the observed data are $\{(T_i, Y_i, X_i), i = 1, \ldots, N\}$ where N is the sample size. We assume that $[\{T_i, Y_i(1), Y_i(0), X_i\}, i = 1, \ldots, N]$ are independent and identically distributed, and that T is independent of $\{Y(1), Y(0)\}$ conditional on X.

Note that τ consists of two expectations, $E\{Y(1)\}$ and $E\{Y(0)\}$. In this work, we consider weighted estimation of these expectations. Without loss of generality, we focus on $E\{Y(1)\}$. In the following, we consider a weighting estimator of $E\{Y(1)\}$ that can be represented as $N^{-1}\sum_{i=1}^{N} T_i w_i Y_i$. Hence, for estimation of $E\{Y(1)\}$, we only need to specify weights $w_i (i : T_i = 1)$ for individuals in the treatment group.

Let $\pi(x) = \mathrm{pr}(T = 1 \mid X = x)$ be the propensity score. Assuming knowledge of $\pi(X_i)$ $(i : T_i = 1)$, w_i can be chosen as $\{\pi(X_i)\}^{-1}$ to obtain a consistent estimator of $E\{Y(1)\}$. In practice, propensity scores are usually unknown. In such scenarios, one can estimate the propensity score function to form a plug-in estimator for $E\{Y(1)\}$. However, estimation errors and model misspecification of the propensity score function can lead to significant error in the estimation of $E\{Y(1)\}$ due to the use of inverse probability weighting. Poor finite-sample performance of such estimators has been reported in the literature (Kang & Schafer, 2007).

Due to this unsatisfactory performance, some attention has been given to choosing $w_i (i : T_i = 1)$ via covariate balancing, which mimics randomization directly. To understand this, note that

$$E\left\{\frac{Tu(X)}{\pi(X)}\right\} = E\{u(X)\}, \tag{1}$$

for any measurable function $u : \mathcal{X} \to \mathbb{R}$ such that $E\{u(X)\}$ exists and is finite. Instead of modeling the propensity function, it is therefore natural to choose weights that ensure the validity of

the empirical finite-dimensional approximation of (1),

$$\frac{1}{N}\sum_{i=1}^{N} T_i w_i U(X_i) = \frac{1}{N}\sum_{i=1}^{N} U(X_i), \tag{2}$$

where $U(X) = (u_1(X),\ldots,u_L(X))^{\mathrm{T}}$ is a L-variate function of X. Here $\mathrm{span}\{u_1,\ldots,u_L\}$ can be viewed as a finite-dimensional approximation space of functions in which the balancing is enforced. Practical considerations may suggest a choice of $\{u_1,\ldots,u_L\}$. In this case, we call it parametric covariate balancing. Without assumptions on the outcome regression model, the balancing of fixed and finitely many component functions u_j in (1) may not lead to consistent estimation (Hellerstein & Imbens, 1999). To allow consistent estimation in a larger family of outcome regression functions, another direction is to allow L to increase with N (Chan et al., 2016). This has a nonparametric flavor similar to regression splines for which the number of knots grows with sample size. However, the choices of L and $\{u_1,\ldots,u_L\}$ are not obvious. In this work, we aim to balance covariate functionals nonparametrically via reproducing-kernel Hilbert space modeling of the approximation space.

Let $m(X) = E\{Y(1) \mid X\}$ and $Y_i(1) = m(X_i) + \varepsilon_i$ for $i = 1,\ldots,N$. Further assume that the ε_i are independent with $E(\varepsilon_i \mid X_i) = 0$ and $E(\varepsilon_i^2 \mid X_i) = \sigma_i^2 < \infty$. All weighting estimator of $E\{Y(1)\}$ admits the decomposition

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} T_i w_i Y_i &= \left\{\frac{1}{N}\sum_{i=1}^{N} T_i w_i Y_i - \frac{1}{N}\sum_{i=1}^{N} m(X_i)\right\} + \left[\frac{1}{N}\sum_{i=1}^{N} m(X_i) - E\{Y(1)\}\right] + E\{Y(1)\} \\ &= \frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)m(X_i) + \frac{1}{N}\sum_{i=1}^{N} T_i w_i \varepsilon_i + \left[\frac{1}{N}\sum_{i=1}^{N} m(X_i) - E\{Y(1)\}\right] + E\{Y(1)\},\end{aligned} \tag{3}$$

which allows a transparent understanding of the terms that have to be controlled. The first term on the right-hand side of (3) poses a challenge since the unknown outcome regression function m is intrinsically related to the outcome data, and could be complex and high-dimensional in general. To connect with covariate balancing, if $m \in \mathrm{span}\{u_1,\ldots,u_L\}$ in (2), we can control the first term. For the second term, the $\varepsilon_i (i = 1,\ldots,N)$ are independent of the choice of $w_i (i : T_i = 1)$ if the outcome data are not used to obtain the weights. Some control over the magnitude of w_i will lead to convergence of the second term. Corresponding details will be given in §2·4. The convergence of the third term is ensured by the law of large numbers.

2·2. *Construction of the method*

We consider the following empirical validity measure for any suitable function u,

$$S_N(w,u) = \left\{\frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)\,u(X_i)\right\}^2,$$

where $w = (w_1,\ldots,w_N)^{\mathrm{T}}$. In parametric covariate balancing, weights $w_i (i : T_i = 1)$ can be constructed to satisfy

$$\sup_{u\in\mathcal{U}_L} S_N(w,u) = 0,$$

where $\mathcal{U}_L = \mathrm{span}\{u_1,\ldots,u_L\}$ with $u_1,\ldots,u_L$ being suitable basis functions. In this case, the weights attain exact covariate balance as in (2) when the dimension of $\mathcal{U}_L$ is small.

Here the overall validity of (1) is instead controlled directly on an approximation space $\mathcal{H}$, a reproducing-kernel Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. Ideally, one would want to pick a large enough, possibly infinite-dimensional, space $\mathcal{H}$ to guarantee the control of $S_N(w,u)$ on a rich class of functions. Unlike sieve spaces, $\mathcal{H}$ is specified without reference to sample size. The matching of non-linear functions is also automatic if $\mathcal{H}$ is large enough to contain such functions, without the need to explicitly introduce particular non-linear basis functions in sieve spaces. For any Hilbert space $\mathcal{H}_1$ of functions of x_1 and any Hilbert space $\mathcal{H}_2$ of functions of x_2, the tensor product space $\mathcal{H}_1 \otimes \mathcal{H}_2$ is defined as the completion of the class $\{\sum_{k=1}^{\ell} f_1(x_1)f_2(x_2) : f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2, \ell = 1,2,\dots\}$ under the induced norm by $\mathcal{H}_1$ and $\mathcal{H}_2$. A popular choice of $\mathcal{H}$ is the tensor product reproducing-kernel Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_d$ with $\mathcal{H}_j$ being the reproducing-kernel Hilbert space of functions of the j-th component of X. Suppose the support of the covariate distribution is $[0,1]^d$ and $f^{(\ell)}$ is the ℓ-th derivative of a function f. Following Wahba (1990), one can pick $\mathcal{H}_j$ as the ℓ-th order Sobolev space $\mathcal{W}^{\ell,2}([0,1]) = \{f : f, f^{(1)}, \cdots, f^{(\ell-1)}$ are absolutely continuous, $f^{(\ell)} \in L^2[0,1]\}$ with norm

$$\|f\| = \left[\sum_{k=0}^{\ell-1}\left\{\int_0^1 f^{(k)}(t)dt\right\}^2 + \int_0^1 \left\{f^{(\ell)}(t)\right\}^2 dt\right]^{1/2}.$$

The second-order Sobolev space is one of the most common choices in practice and will be adopted in all of our numerical illustrations. Another common choice is the space generated by the Gaussian kernel, which will also be compared in numerical studies. If it is desirable to prioritize covariates based on prior beliefs, we can raise the components to different powers to reflect their relative importance. For Gaussian kernels, this is equivalent to using different bandwidth parameters for each covariate. In cases when there are binary or categorical covariates, one can choose the corresponding $\mathcal{H}_j$ as a reproducing-kernel Hilbert space with kernel $R(s,t) = I(s = t)$, for any levels s and t of such covariate, as suggested by Gu (2013); here I is an indicator function.

Ideally, we want to control $\sup_{u\in\mathcal{H}} S_N$. However, there are two issues. First, that $S_N(w,cu) = c^2 S_N(w,u)$ for any $c \geq 0$ suggests a scale issue of S_N with respect to u. Therefore, in order to use $S_N(w,u)$ to determine the weights w, the magnitude of u should be standardized. To cope with this, we notice

$$S_N(w,u) = \left\{\frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)u(X_i)\right\}^2 \leq \|u\|_N^2 \left\{\frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)^2\right\} \tag{4}$$

due to the Cauchy–Schwarz inequality, where $\|u\|_N^2 = N^{-1}\sum_{i=1}^{N} u(X_i)^2$. In view of (4), we restrict our focus to $\widetilde{\mathcal{H}}_N = \{u \in \mathcal{H} : \|u\|_N = 1\}$. Second, similar to many statistical and machine learning frameworks, the optimization of an unpenalized sample objective function will result in overfitting. In our case, the weights become highly unstable. To alleviate this, we control $\|\cdot\|_{\mathcal{H}}$ to emphasize the balance on smoother functions. Additionally, we penalize on $V_N(w) = N^{-1}\sum_{i=1}^{N} T_i w_i^2$ to control both the variabilities of w and of the second term in the right-hand side of (3). Overall, we consider the constrained minimization,

$$\min_{w\geq 1}\left[\sup_{u\in\widetilde{\mathcal{H}}_N}\left\{S_N(w,u) - \lambda_1\|u\|_{\mathcal{H}}^2\right\} + \lambda_2 V_N(w)\right], \tag{5}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters and the above minimization is only taken over $w_i (i : T_i = 1)$. The weights w_i are restricted to be greater than or equal to 1, as their counter-

parts, inverse propensities, satisfy $\{\pi(X_i)\}^{-1} \geq 1$. We denote the solution of (5) by $\widehat{w}$. Further discussion on these tuning parameters will be given in §2.4 and §2.5. In particular, we show that the convergence to zero of the first term of (3) can be ensured even when $\lambda_2 = 0$. This indicates that this extra tuning parameter is mostly needed for our justification of the convergence of the second term in (3).

A small number of recent papers have also considered kernel-based methods for covariate balancing. An unpublished paper by Zhao (arXiv:1601.05890) considered a dual formulation of the method of Imai & Ratkovic (2014) for the estimation of $\pi(x)$ under a logistic regression model, and generalized the linear predictor into a non-linear one using the kernel trick. Since this method aims at estimating $\pi(x)$, it requires smoothness conditions on π and penalizes on smoothness of the resulting estimate. Our method does not require smoothness of π and penalizes the smoothness of the balancing functions. An unpublished paper by Kallus (arXiv:1606.05188) considered weights that minimize the dual norm of a balancing error. Given a reproducing-kernel Hilbert space, this method does not have the ability to adapt to a relevant subset of functions. An external parameter is required to index the function space, such as the dispersion parameter of a Gaussian kernel, which needs to be specified in an ad-hoc manner. Due to the lack of an explicit tuning parameter, this method will not work well for Sobolev space which does not have extra indexing parameters. Our method works for a given reproducing-kernel Hilbert space by using a data-adaptive tuning to promote balancing of smoother functions within the given space. An unpublished paper of Hazlett (arXiv: 1605.00155) proposed an extension of the moment-based balancing method of Hainmueller (2012) to balance the columns of the Gram matrix. Since the Gram matrix is $N \times N$, exact balancing of N moment conditions under additional constraints on the weights are often computationally infeasible. Balancing a low-rank approximation of the Gram matrix may be an ad-hoc solution but the theoretical properties have not been studied.

2·3. *Finite-dimensional representation*

Many common choices of reproducing-kernel Hilbert space, including Sobolev Hilbert space, are infinite-dimensional and therefore, the inner optimization in (5) is essentially an infinite-dimensional optimization which is seemingly impractical. Fortunately, we shall show that the solution of (5) enjoys a finite-dimensional representation. First, the inner optimization of (5) can be expressed as

$$\sup_{u \in \mathcal{H}} \left\{ \frac{S_N(w,u)}{\|u\|_N^2} - \lambda_1 \frac{\|u\|_{\mathcal{H}}^2}{\|u\|_N^2} \right\}.$$

Let K be the reproducing kernel of $\mathcal{H}$. By the representer theorem (Wahba, 1990), the solution lies in a finite-dimensional subspace $\mathrm{span}\{K(X_j,\cdot) : j = 1,\ldots,N\}$. Now this optimization is equivalent to:

$$\sup_{\alpha=(\alpha_1,\ldots,\alpha_N)^{\mathrm{T}} \in \mathbb{R}^N} \left[\frac{S_N\left\{w, \sum_{j=1}^N \alpha_j K(X_j,\cdot)\right\}}{\alpha^{\mathrm{T}} M^2 \alpha / N} - \lambda_1 \frac{\alpha^{\mathrm{T}} M \alpha}{\alpha^{\mathrm{T}} M^2 \alpha / N} \right], \tag{6}$$

where M is a $N \times N$ matrix with (i,j)-th element $K(X_i, X_j)$. This matrix is positive semi-definite and is commonly known as the Gram matrix. Let the eigen-decomposition of M be

$$M = \begin{pmatrix} P_1 & P_2 \end{pmatrix} \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix} \begin{pmatrix} P_1^{\mathrm{T}} \\ P_2^{\mathrm{T}} \end{pmatrix},$$

where Q_1 and Q_2 are diagonal matrices. In particular, $Q_2 = 0$. Let r be the rank of Q_1. We remark that P_2 and Q_2 do not exist if $r = N$, but the following derivation still holds. Moreover,

$$S_N\left\{w, \sum_{j=1}^{N} \alpha_j K(X_j, \cdot)\right\} = \frac{1}{N^2}\alpha^{\mathrm{T}} M A(w) M \alpha, \tag{7}$$

where $A(w) = a(w)a(w)^{\mathrm{T}}$ with $a(w) = (T_1 w_1 - 1, T_2 w_2 - 1, \ldots, T_N w_N - 1)^{\mathrm{T}}$. Let $\beta = Q_1 P_1^{\mathrm{T}} \alpha / N^{1/2}$. The constrained optimization (6) is then equivalent to

$$\sup_{\beta \in \mathbb{R}^r : \|\beta\| \le 1} \beta^{\mathrm{T}} \left\{\frac{1}{N} P_1^{\mathrm{T}} A(w) P_1 - N\lambda_1 Q_1^{-1}\right\} \beta.$$

Therefore, the target optimization becomes

$$\min_{w \ge 1} \left[\sigma_{\max}\left\{\frac{1}{N} P_1^{\mathrm{T}} A(w) P_1 - N\lambda_1 Q_1^{-1}\right\} + \lambda_2 V_N(w)\right], \tag{8}$$

where $\sigma_{\max}(M)$ represents the maximum eigenvalue of a matrix M. Again, the above minimization is only taken over $w_i (i : T_i = 1)$. Since $P_1^{\mathrm{T}} a(w)$ is an affine transformation of w and V_N is a convex function, the objective function of this minimization is convex with respect to w, due to Proposition 1, whose proof is given in the Supplementary Material. Due to convexity and Slater's condition of strict feasibility, a necessary and sufficient condition for a global minimizer of (8) is the corresponding Karush–Kuhn–Tucker condition using subdifferentials.

PROPOSITION 1. *Let $B \in \mathbb{R}^{r \times r}$ be a symmetric matrix. The function $\sigma_{\max}(vv^{\mathrm{T}} + B)$ is convex with respect to $v \in \mathbb{R}^r$.*

As for the computation, we note that the maximum eigenvalue is evaluated at a rank-one modification of a diagonal matrix, which can be computed efficiently by solving the secular equation (O'leary & Stewart, 1990) in a common linear algebra package such as LAPACK. The objective function is second-order differentiable with respect to the w_i when the maximum eigenvalue of $P_1^{\mathrm{T}} A(w) P_1 / N - N\lambda_1 Q_1^{-1}$ has multiplicity 1. Moreover, the corresponding gradient has a closed-form expression. In this case, a common and fast nonlinear optimization method such as the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bound constraints can be applied. Non-differentiability exists when the largest two eigenvalues of $P_1^{\mathrm{T}} A(w) P_1 / N - N\lambda_1 Q_1^{-1}$ coincide. To ensure validity, one could employ the following two-part computational strategy. First, one applies the Broyden–Fletcher–Goldfarb–Shanno algorithm and checks numerically whether the maximum eigenvalue evaluated at the resulting solution is repeated. If not, the objective function is differentiable at this solution and the Karush–Kuhn–Tucker condition is satisfied. Thus, the minimizer is obtained. Otherwise, the nonlinear eigenvalue optimization method of Overton (1992, Section 5), which is applicable to the scenario of repeated eigenvalues, is initialized by the former estimate and then applied. In our practical experience, the second step is seldom needed and has negligible effect to the final solution. Therefore, for fast computation, we only apply the first part in our numerical illustrations.

2·4. *Theoretical properties*

For notational simplicity, we shall study the theoretical properties of the proposed estimator for $\mathcal{H}$ being the tensor product of ℓ-th order Sobolev spaces, as studied extensively in smoothing splines (Wahba, 1990; Gu, 2013). Our results can be extended to other choices of $\mathcal{H}$ if an entropy result and a uniform boundedness of the unit ball $\{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \le 1\}$ are supplied; see the Supplementary Material. For instance, the respective entropy result of Gaussian reproducing-kernel

Hilbert space can be obtained from Zhou (2002). As mentioned, we concentrate on $E\{Y(1)\}$. Similar conditions are required for $E\{Y(0)\}$ to obtain results on the average treatment effect $\tau = E\{Y(1) - Y(0)\}$.

Assumption 1. The propensity $\pi(\cdot)$ is uniformly bounded away from 0. That is, there exists a constant C such that $1/\pi(x) \leq C < \infty$ for all $x \in \mathcal{X}$.

Assumption 2. The ratio d/ℓ is less than 2.

Assumption 3. The regression function $m(\cdot)$ belongs to $\mathcal{H}$.

Assumption 4. The errors $\{\varepsilon_i\}$ are uncorrelated where $E(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma_i^2 \leq \sigma^2$ for all $i = 1, \ldots, N$. Further $\{\varepsilon_i\}$ are independent of $\{T_i\}$ and $\{X_i\}$.

The above assumptions are very mild. Assumption 1 is the usual overlap condition required for identification. There are no additional smoothness assumptions on $\pi(\cdot)$ which would typically be required in propensity score or covariate balancing methods (Hirano et al., 2003; Chan et al., 2016). Assumption 2 corresponds to the weakest smoothness assumption on $m(\cdot)$ in smoothing spline regression. We use the notation $A_N \asymp B_N$ to represent $A_n = O(B_N)$ and $B_N = O(A_N)$ for some sequences A_N and B_N.

THEOREM 1. *Suppose Assumptions* 1-3 *hold. If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 = O(N^{-1})$, *then* $S_N(\widehat{w}, m) = O_p(N^{-1})\|m\|_N^2$. *If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 \asymp N^{-1}$, *then* $V_N(\widehat{w}) = O_p(1)$ *and there exist constants* $W > 0$ *and* $S^2 > 0$ *such that* $E\{V_N(\widehat{w})\} \leq W$ *and* $E\{NS_N(\widehat{w}, m)\} \leq S^2$.

Theorem 1 supplies the rate of convergence of the first term in (3), and boundedness of the expectation of the second term in (3). Convergence of $S_N(\widehat{w}, m)$ is guaranteed even if λ_2 is chosen as 0. However, to ensure the boundedness of $E\{V_N(\widehat{w})\}$, additional regularization is needed and hence $\lambda_2 > 0$ is proposed. The following theorem establishes the $N^{1/2}$-consistency of the weighting estimator. Moreover, we show that the asymptotic distribution has a finite variance.

THEOREM 2. *Suppose Assumptions* 1–4 *hold and* $m \in \mathcal{H}$. *If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 \asymp N^{-1}$,

$$\frac{1}{N}\sum_{i=1}^{N} T_i \widehat{w}_i Y_i - E\{Y(1)\} = O_p(N^{-1/2}).$$

Moreover, $N^{1/2}[\sum_{i=1}^{N} T_i \widehat{w}_i Y_i/N - E\{Y(1)\}]$ *has finite asymptotic variance.*

Although Theorem 2 only gives the rate of convergence of the estimator, it is stronger than recent results for other kernel-based methods for the estimation of average treatment effects. Zhao (arXiv:1601.05890) and Hazlett (arXiv:1605.00155) do not provide the rate of convergences of their estimators. To our knowledge, the only paper that contains a rate of convergence for kernel-based methods is Kallus (arXiv:1612.08321), who showed a root-N convergence rate under a strong assumption that $m(X)$ is linear in X and did not develop the asymptotic distribution. In fact, when linear assumptions hold, parametric covariate balancing is sufficient for estimating the average treatment effects (Qin & Zhang, 2007). When $m(\cdot)$ is a general function, the difficulty in theoretical development lies in the first term of (3), which is shown to attain the same rate of convergence as the other two terms of (3), but its asymptotic distribution is not available. For the sieve-based method (Chan et al., 2016), the growth rate of the sieve approximation space can be carefully chosen in a range such that terms analogous to the first term of (3) have a faster convergence rate than the dominating terms. In our case, similar to nonparametric regression, there is only a particular growth rate of λ_1 such that the bias and variance of the first term of (3)

are balanced. In fact, it is possible that the term has an asymptotic bias of order $N^{-1/2}$. In §2·6, a modified estimator is studied by debiasing the first term of (3), so that its rate of convergence is faster than $N^{-1/2}$ and is dominated by the other terms. In that case, the asymptotic distribution can be derived. Further discussion of the relationship between Theorem 2 and the literature is given in Remark 3.

2·5. *Tuning parameter selection*

In Theorems 1 and 2, λ_1 and λ_2 are required to decrease at the same order N^{-1}, so as to achieve the desired asymptotic results. To reduce the amount of tuning, we choose $\lambda_2 = \zeta\lambda_1$ where $\zeta > 0$ is fixed. As explained above, λ_2 is chosen to be positive mostly to ensure the boundedness of $E\{V_N(\widehat{w})\}$. From our practical experience, the term $V_N(\widehat{w})$ is usually stable and does not take large values even if λ_2 is small. Therefore, we are inclined to choose a small ζ. In all of our numerical illustrations, ζ is fixed at 0.01. Now we focus on the choice of λ_1. Note that the tuning of λ_1 is similar to choosing the dimension of the sieve space in Chan et al. (2016), which is a difficult and mostly unsolved problem. In this paper, we do not attempt to solve this problem rigorously, but to provide a reasonable solution.

By Lagrange multipliers, the optimization $\sup_{u\in\widetilde{\mathcal{H}}_N}\{S_N(w,u) - \lambda_1\|u\|^2_{\mathcal{H}}\}$ is equivalent to $\sup_{\{u\in\widetilde{\mathcal{H}}_N:\|u\|_{\mathcal{H}}\le\gamma\}} S_N(w,u)$ for some γ, where there exists a correspondence between γ and λ_1. Since a larger regularization parameter corresponds to a stricter constraint, γ decreases with λ_1. We use

$$B_N(w) = \sup_{\{u\in\widetilde{\mathcal{H}}_N:\|u\|_{\mathcal{H}}\le\gamma\}} S_N(w,u), \tag{9}$$

as a measure of the balancing error over $\{u \in \widetilde{\mathcal{H}}_N : \|u\|_{\mathcal{H}} \le \gamma\}$ with respect to the weights w. Due to the large subset of functions to balance, $B_N(\widehat{w})$ is large when γ is large, or equivalently, when λ_1 is small. When γ decreases, or equivalently, λ_1 increases, $B_N(\widehat{w})$ typically decreases to approximately zero, as the resulting weight $\widehat{w}$ approximately balances the whole subset $\{u \in \widetilde{\mathcal{H}}_N : \|u\|_{\mathcal{H}} \le \gamma\}$. An example is given in Fig. 1 which will be discussed in §3.2. When this happens, a further decrease of γ would not lead to any significant decrease in $B_N(\widehat{w})$. The key idea is to choose the smallest λ_1 that achieves such approximate balancing, to ensure the largest subset of functions being well-balanced. In practice, we compute our estimator with respect to a grid of λ_1: $\lambda_1^{(1)} < \cdots < \lambda_1^{(J)}$. Write $\widehat{w}^{(j)}$ as the estimator with respect to $\lambda_1^{(j)}$. We select $\lambda_1^{(j^*)}$ as our choice of λ_1 if j^* is the smallest j such that

$$\frac{B_N(\widehat{w}^{(j+1)}) - B_N(\widehat{w}^{(j)})}{\lambda_1^{(j+1)} - \lambda_1^{(j)}} \ge e,$$

where e is chosen as a negative constant of small magnitude. In the numerical illustrations, we set $e = -10^{-6}$.

2·6. *An efficient modified estimator*

Since the outcome regression function $m(\cdot)$ is assumed to be in a reproducing-kernel Hilbert space $\mathcal{H}$, a kernel-based estimator $\widehat{m}(\cdot)$, such as smoothing splines (Gu, 2013), can be employed, and $N^{-1}\sum_{i=1}^N \widehat{m}(X_i)$ is a natural estimator of $E\{Y(1)\} = E[E\{Y(1) \mid X\}] = E\{m(X)\}$. However, since randomization is administered before collecting any outcome data, Rubin (2007) advocated the estimation of treatment effects without using outcome data to avoid data snooping. On the other hand, Chernozhukov et al. (arXiv:1608.00060) and Athey et al. (arXiv:1604.07125) advocate the use of an estimated outcome regression function to improve the theoretical results

in high-dimensional settings. Inspired by these results, we modify the weighting estimator by subtracting $N^{-1}\sum_{i=1}^{N}(T_i w_i - 1)\widehat{m}(X_i)$ from both sides of (3), so that the first term in the decomposition becomes $N^{-1}\sum_{i=1}^{N}(T_i w_i - 1)\{m(X_i) - \widehat{m}(X_i)\}$, while the remaining two terms are unchanged. It can then be shown that the first term has a rate of convergence faster than $N^{-1/2}$ under mild assumptions, and the asymptotic distribution of the resulting estimator will be derived.

The estimator takes the form

$$\frac{1}{N}\sum_{i=1}^{N}\{T_i w_i Y_i - (T_i w_i - 1)\widehat{m}(X_i)\} = \frac{1}{N}\sum_{i=1}^{N}[T_i w_i\{Y_i - \widehat{m}(X_i)\} + \widehat{m}(X_i)]$$

which has the same form as the residual balancing estimator proposed in Athey et al. (arXiv:1604.07125). They consider a different setting of high-dimensional linear regression model with sparsity assumptions, and showed that their estimator attains the semiparametric efficiency bound.

Our analysis requires the additional technical assumption such that $\widehat{w}$ is $o_p(N^{1/2})$. To achieve this, we adopt an assumption as in Athey et al. (arXiv:1604.07125) that $\widehat{w} \le BN^{1/3}$ for a prespecified large positive constant B. This can be enforced in the optimization (8) easily together with the constraint $\widehat{w} \ge 1$. For clarity, we call this estimator $\widetilde{w} = (\widetilde{w}_1, \ldots, \widetilde{w}_N)^{\mathrm{T}}$.

THEOREM 3. *Suppose Assumptions* 1, 2 *and* 4 *hold with* $\sigma_i^2 = \sigma^2$ *for all* i. *Also, assume* $\max_i E|\varepsilon_i|^3 < \infty$. *Let* $h = m - \widehat{m} \in \mathcal{H}$ *such that* $\|h\|_N = o_p(1)$ *and* $\|h\|_{\mathcal{H}} = O_p(1)$. *Further, assume* $\lambda_1 = o(N^{-1})$, $\lambda_2\|h\|_N^2 = o_p(N^{-1})$, *and* $\lambda_1^{-1} = o(\lambda_2^{(2\ell-d)/d}N^{2\ell/d})$. *Write*

$$J_N = N^{1/2}\left(\left[\frac{1}{N}\sum_{i=1}^{N}T_i\widetilde{w}_i\{Y_i - \widehat{m}(X_i)\} + \frac{1}{N}\sum_{i=1}^{N}\widehat{m}(X_i)\right] - E\{Y(1)\}\right),$$

$$J_N^* = [\mathrm{var}\{m(X_1)\}]^{1/2}F + \sigma N^{-1/2}\sum_{j=1}^{N}T_j\widetilde{w}_j G_j,$$

where $F, G_1, \ldots, G_N$ *are independent and identically distributed standard normal random variables independent of* $X_1, \ldots, X_N$, $T_1, \ldots, T_N$ *and* $\varepsilon_1, \ldots, \varepsilon_N$. *Let* ψ_N *and* ψ_N^* *be the corresponding characteristic function of* J_N *and* J_N^* *respectively. Then*

$$|\psi_N(t) - \psi_N^*(t)| \to 0, \quad t \in \mathbb{R},$$

where ψ_N^* *is twice differentiable, and*

$$\limsup_N \mathrm{var}(J_N) \le \mathrm{var}\{m(X_1)\} + \sigma^2 V, \tag{10}$$

where $V = E\{1/\pi(X_1)\}$.

COROLLARY 1. *Under the assumptions of Theorem* 3,

$$N^{1/2}\{\sigma^2 V_N(\widetilde{w})\}^{-1/2}\left[\frac{1}{N}\sum_{i=1}^{N}T_i\widetilde{w}_i\{Y_i - \widehat{m}(X_i)\} + \frac{1}{N}\sum_{i=1}^{N}\{\widehat{m}(X_i) - m(X_i)\}\right]$$

converges in distribution to a standard normal distribution as $N \to \infty$.

Remark 1. In Theorem 3, the estimand is $E\{Y(1)\}$, whereas in Corollary 1, the estimand is a finite-sample conditional average, $N^{-1}\sum_{i=1}^{N}E(Y_i(1) \mid X_i) = N^{-1}\sum_{i=1}^{N}m_1(X_i)$. Athey et al.

(arXiv: 1604.07125) considered a finite-sample conditional average treatment effect and obtained a result similar to Corollary 1. Normalization by $V_N(\widetilde{w})$ is possible in Corollary 1 following a conditional central limit theorem, since $\widetilde{w}$ depends only on (T_i, X_i) $(i = 1,\ldots,N)$ and can be treated as constants upon conditioning. To derive the limiting distribution of J_N in Theorem 3, one cannot use a similar normalization because the handling of extra term $N^{-1}\sum_{i=1}^{N}\{\widehat{m}(X_i) - m(X_i)\}$ requires averaging across the X distribution. If $V_N(\widetilde{w})$ converges to a constant in probability, one could use Slutsky's theorem to claim asymptotic normality of J_N. Theorem 3 requires a partially conditional central limit theorem which is proven in the Supplementary Material and the distribution of J_N can be approximated by a weighted sum of independent standard normal random variables. The asymptotic variance is bounded above by the right-hand side of (10), which is the semiparametric efficiency bound (Robins et al., 1994; Hahn, 1998).

Remark 2. Compared to Theorem 2, Theorem 3 requires different conditions on the orders of λ_1 and λ_2. These order specifications, together with a diminishing $\|h\|_N$, allow a direct asymptotic comparison between $V_N(\widetilde{w})$ and V, which leads to $V_N(\widetilde{w}) \le V\{1 + o_p(1)\}$. This is essential for achieving (10) in our proof. To make sense of the theorem, the conditions $\lambda_1 = o(N^{-1})$ and $\lambda_1^{-1} = o\{\lambda_2^{(2\ell-d)/d} N^{2\ell/d}\}$ should not lead to a null set of λ_1. As an illustration, suppose $\widehat{m}$ achieves the optimal rate $\|h\|_N \asymp N^{-\ell/(2\ell+d)}$, then one can take $\lambda_2 = o\{N^{-d/(2\ell+d)}\}$, which suggests $\lambda_2^{(2\ell-d)/d} N^{2\ell/d} = o\{N^{(d^2+4\ell^2)/(d^2+2\ell d)}\}$. Due to Assumption 2, $(d^2+4\ell^2)(d^2+2\ell d)^{-1} > 1$. Therefore, there exist choices of λ_1 and λ_2 that fulfill the assumption of Theorem 3. We found in simulations that the practical performance of the modified estimator is not sensitive to λ_1 and λ_2, and we thus use the method described in §2.5 to obtain these tuning parameters.

Remark 3. Most existing efficient methods require explicit or implicit estimation of both $\pi(\cdot)$ and $m(\cdot)$. Chernozhukov et al. (arXiv:1608.00060) gave a general result on the convergence rate required on both $\pi(\cdot)$ and $m(\cdot)$ for efficient estimation. Even though weighting methods do not explicitly estimate $m(\cdot)$, estimating-equation-based methods would give rise to implicit estimators of $m(\cdot)$ that attain good rates of convergence (Hirano et al., 2003; Chan et al., 2016). However, weights constructed based on complex optimization problems may not even converge to the true inverse propensities, see Athey et al. (arXiv: 1604.07125) who, under a sparse linear model assumption, proposed an efficient estimator by controlling the balancing error of linear functions and the estimation error for $m(\cdot)$. Although our modified estimator is not a direct kernel-based extension of their method, we have arrived at a similar conclusion. Our method only requires $\|\widehat{m} - m\|_N = o_p(1)$ and does not require the smoothness of $\pi(\cdot)$ or linearity of $m(\cdot)$, and is therefore less vulnerable to the curse of dimensionality. Note that the weighting estimator as described in Theorem 2 corresponds to $\widehat{m} = 0$, and therefore $\|\widehat{m}-m\|_N$ is not $o_p(1)$ when $m(\cdot)$ is not the zero function. In view of the current literature, Theorem 2 is an interesting result because convergence to neither $\pi(\cdot)$ nor $m(\cdot)$ is established.

3. NUMERICAL EXAMPLES

3·1. *Simulation study*

Simulation studies were conducted to evaluate the finite sample performance of the proposed estimator. We considered simulation settings where the propensity score and outcome regression models are non-linear functions of the observed covariates, with possibly non-smooth propensity score functions. For each observation, we generated a ten-dimensional multivariate standard Gaussian random vector $Z = (Z_1,\ldots,Z_{10})^{\mathrm{T}}$. The observed covariates are

$X = (X_1, \ldots, X_{10})^{\mathrm{T}}$ where $X_1 = \exp(Z_1/2), X_2 = Z_2/\{1 + \exp(Z_1)\}, X_3 = (Z_1 Z_3/25 + 0.6)^3$, $X_4 = (Z_2 + Z_4 + 20)^2$ and $X_j = Z_j$ $(j = 5, \ldots, 10)$. Three propensity score models are studied; model 1 is $\mathrm{pr}(T = 1 \mid Z) = \exp(-Z_1 - 0.1Z_4)/(1 + \exp(-Z_1 - 0.1Z_4)$, model 2 is $\mathrm{pr}(T = 1 \mid Z) = \exp\{-Z_1 - 0.1Z_4 + \eta_2(\widetilde{Z})\}/[1 + \exp\{-Z_1 - 0.1Z_4 + \eta_2(\widetilde{Z})\}]$, and model 3 is $\mathrm{pr}(T = 1 \mid Z) = \exp\{-Z_1 - 0.1Z_4 + \eta_3(\widetilde{Z})\}/[1 + \exp\{-Z_1 - 0.1Z_4 + \eta_3(\widetilde{Z})\}]$, where $\widetilde{Z} = (Z_2 + Z_4 + Z_6 + Z_8 + Z_{10})/5$, η_2 is the scaling function of the Daubechies 4-tap wavelet (Daubechies, 1992), and η_3 is the Weierstrass function with parameters $a = 2$ and $b = 13$. The functions η_2 and η_3 are chosen such that the propensity functions in models 2 and 3 are non-smooth. Two outcome regression models are studied: model A is $Y = 210 + (1.5T - 0.5)(27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4) + \epsilon$, and model B is $Y = Z_1 Z_2^3 Z_3^2 Z_4 + Z_4 |Z_1|^{0.5} + \epsilon$, where ϵ has standard normal distribution. For each scenario, we compared the proposed weighting and modified estimators using two commonly employed kernels: the second-order Sobolev kernel and the Gaussian kernel with bandwidth parameter chosen via the median heuristics (Gretton et al., 2005). We also compared the Horvitz–Thompson estimator where the weights are the inverse of propensity scores estimated by maximum likelihood under a working logistic regression model with X being the predictors, the Hájek estimator which is a normalized version of the Horvitz–Thompson estimator with weights summing up to N, the inverse probability weighting estimator using covariate balancing propensity score of Imai & Ratkovic (2014), the stable balancing weights of Zubizarreta (2015), and the nonparametric covariate balancing estimator of Chan et al. (2016) with exponential weights. The first moment of X was balanced explicitly for these methods. We compared the bias, root mean squared error and covariate balance of the methods, where covariate balance is evaluated at the true conditional mean function. In particular, we calculate $S_N(\widehat{w}, m)$ to evaluate the covariate balance of the treatment and the combined groups, also its counterpart for the covariate balance of the controls and the combined groups, and report the sum of these two measures. The reason for comparing the covariate balance at the true conditional mean function is that it is the optimal function to balance but is unknown in practice. For each scenario, 1000 independent data sets are generated, and the results for outcome models A and B with sample size $N = 200$ are given in Tables 1 and 2 respectively.

The results show that the empirical performance of the estimators are related to the degree of covariate balancing. Without any explicit covariate balancing, the Horvitz–Thompson estimator can be highly unstable. The Hájek estimator balances the constant function, the Imai–Ratkovic estimator balances X, the estimators of Zubizarreta and Chan et al. balance both the constant function and X. For outcome model A, the balance of both constant and X is important and the omission of either constraints can lead to a poor performance. For outcome model B, the balance of X often implies approximate balance of the constant and therefore the estimators of Imai and Ratkovic, as was Zubizarreta and Chan et al. had similar performance. However, in both cases, the proposed method outperformed the other estimators because it can also control the balance of nonlinear and higher-order moments. We attempted to compute a Horvitz–Thompson estimator using a smoothing spline logistic regression model with the same kernel as the proposed method using the R package `gss`, but the program did not converge in reasonable time. We also tried to exactly balance the second moments in addition to the first moments of ten baseline covariates in the existing methods, but the algorithms did not converge in a substantial fraction of simulations. This shortcoming of the existing methods can be circumvented by the proposed methods.

3·2. *Data analysis*

We compare the proposed methods with others using a study of the impact of child abduction by a militant group on the future income of abductees who escape later (Blattman & Annan, 2010). The data contain 741 males in Uganda collected during 2005–2006, of which 462 had

Table 1. *Biases, root mean squared errors and overall covariate balancing measures of various weighting estimators for outcome model A; the reported numbers are averages obtained from 1000 simulated datasets*

	PS1			PS2			PS3		
	Bias	RMSE	Bal	Bias	RMSE	Bal	Bias	RMSE	Bal
Proposed weighting (S)	−192	470	28	−114	422	20	−86	420	21
Proposed weighting (G)	−362	599	86	−232	505	58	−201	504	67
Proposed modified (S)	150	512	28	243	486	20	255	480	21
Proposed modified (G)	−91	378	86	45	368	58	−28	370	67
Horvitz–Thompson	>9999	>9999	>9999	>9999	>9999	>9999	7298	>9999	>9999
Hájek	837	1792	275	693	1429	165	662	1480	175
Imai–Ratkovic	−527	1720	331	−187	1523	274	224	1516	253
Zubizarreta	444	715	28	389	658	23	381	630	21
Chan et al.	262	594	21	226	549	17	244	533	16

The sample sizes were $N = 200$. The values of Bias and RMSE were multiplied by 100, RMSE represents the root mean squared error, and Bal represents an overall covariate balancing measure. S: Sobolev kernel; G: Gaussian kernel.

Table 2. *Biases, root mean squared errors and overall covariate balancing measures of various weighting estimators for outcome model B; the reported numbers are averages obtained from 1000 simulated datasets*

	PS1			PS2			PS3		
	Bias	RMSE	Bal	Bias	RMSE	Bal	Bias	RMSE	Bal
Proposed weighting (S)	−10	85	1	−7	81	0	−8	85	0
Proposed weighting (G)	−7	92	1	−3	88	0	−5	94	0
Proposed modified (S)	3	82	1	−8	82	0	−9	85	0
Proposed modified (G)	−6	89	1	−2	85	0	−4	92	0
Horvitz–Thompson	131	3629	1151	−1	881	66	−35	2092	451
Hájek	4	392	14	−3	221	4	−2	367	12
Imai–Ratkovic	−7	110	1	−6	97	1	−6	108	1
Zubizarreta	−8	115	1	−9	98	1	−10	108	1
Chan et al.	− 8	123	1	−9	99	1	−8	111	1

The sample sizes were $N = 200$. The values of Bias and RMSE were multiplied by 100, RMSE represents the root mean squared error, and Bal represents an overall covariate balancing measure. S: Sobolev kernel; G: Gaussian kernel.

been abducted by militant groups before 2005 but had escaped by the time of the study. Covariates include geographical region, age at 1996, father's education, mother's education, whether the parents had died during or before 1996, whether the father is a farmer and household size in 1996. The investigators chose to collect covariate values in 1996 because it predates most abductions and is also easily recalled as the year of the first election since 1980. The authors discuss the plausibility of the unconfounded treatment assignment since abduction is mostly due to random night raids on rural homes. The outcome of interest here is the daily wage of the study participants in Ugandan shillings in 2005. We compared the estimators as in the simulation studies in §3.1. Table 3 shows the point estimates and 95% confidence intervals based on 1000 bootstrap samples. All methods comparing the abducted to the non-abducted group give a small but non-significant decrease in income. However, a small difference is noted between the proposed method and other methods, indicating that a mild non-linear effect is possibly present, especially in the non-abducted group. To further illustrate this point, we compared the maximal balancing error $B_N(w)$ as a function of λ_1, which, as defined by (9), measures the balancing error as a func-

Table 3. *The effect of child abduction to income in Uganda.*

	$E\{Y(1)\}$	$E\{Y(0)\}$	τ
Proposed weighting (S)	1530 (1219, 1886)	1851 (1247, 2354)	−321 (−893, 431)
Proposed weighting (G)	1516 (1212, 1822)	1671 (1231, 2204)	−156 (−809, 382)
Proposed modified (S)	1532 (1239, 1945)	1867 (1256, 2489)	−355 (−993, 444)
Proposed modified (G)	1536 (1238, 1845)	1689 (1269, 2249)	−153 (−819, 380)
Horvitz–Thompson	1573 (1234, 2033)	2135 (1478, 3075)	−562 (−1667, 242)
Hájek	1573 (1234, 2027)	2131 (1471, 3064)	−558 (−1614, 241)
Imai and Ratkovic	1599 (1256, 2062)	1998 (1381, 2857)	−399 (−1312, 365)
Zubizarreta	1591 (1253, 2073)	2165 (1492, 3046)	−574 (−1613, 229)
Chan et al.	1580 (1246, 2060)	2144 (1485, 3034)	−564 (−1576, 238)

Numbers in parentheses are 95% confidence intervals. S: Sobolev kernel; G: Gaussian kernel.

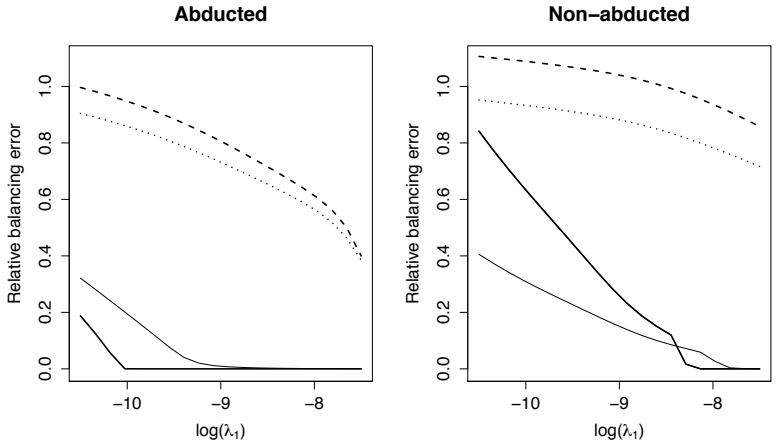

Fig. 1. Supremum balancing error for the child abduction data. Solid bold curves correspond to the proposed estimator using a Sobolev kernel, solid unbold curves correspond to the proposed estimator using a Gaussian kernel, dashed curves correspond to the Horvitz–Thompson estimator, dotted curves correspond to the Imai–Ratkovic estimator.

tion of the size of nested subsets of $\widetilde{\mathcal{H}}_N$, which is chosen as the Sobolev space. The subspace is smaller with an increasing λ_1, containing smoother functions. We standardize the comparisons by dividing the balancing error of constant weights that are used in unweighted comparisons. As seen in Fig. 1, the proposed estimator had approximately no balancing error after reaching a data-dependent threshold, so that any smoother functions can be approximately balanced. This is not the case for other estimators, since there will be residual imbalance for non-linear functions of a given smoothness. We note that the Imai–Ratkovic estimator has less balancing error than the Horvitz–Thompson estimator with maximum likelihood weights, because the former explicitly balances more moments than the latter, including linear and some non-linear covariate functionals. The Imai–Ratkovic estimator gives the closest result to the proposed estimators, but their confidence intervals are consistently narrower than other methods.

ACKNOWLEDGEMENT

The authors thank the editor, an associate editor and a reviewer for their helpful comments and suggestions. The work of the first author was partially supported by the U.S. National Science

Foundation. In addition, most of this work was conducted while the first author was affiliated with Iowa State University. The work of the second author was partially supported by the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health and the U.S. National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of Proposition 1 and Theorems 1–3.

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.* **68**, 337–404.

BLATTMAN, C. & ANNAN, J. (2010). The consequences of child soldiering. *Rev. Econ. Stat.* **92**, 882–898.

CHAN, K. C. G., YAM, S. C. P. & ZHANG, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Statist. Soc. B* **78**, 673–700.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.

GRAHAM, B. S., PINTO, C. C. D. X. & EGEL, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* **79**, 1053–1079.

GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. & SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6**, 2075–2129.

GU, C. (2013). *Smoothing Spline ANOVA Models*. New York: Springer.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–332.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**, 25–46.

HAN, P. & WANG, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* **100**, 417–430.

HELLERSTEIN, J. K. & IMBENS, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Stat.* **81**, 1–14.

HIRANO, K., IMBENS, G. & RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

IACUS, S. M., KING, G. & PORRO, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *J. Am. Statist. Ass.* **106**, 345–361.

IMAI, K. & RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc. B* **76**, 243–263.

KANG, J. D. & SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statist. Sci.* **22**, 523–539.

O'LEARY, D. & STEWART, G. (1990). Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices. *J. Comput. Phys.* **90**, 497–505.

OVERTON, M. L. (1992). Large-scale optimization of eigenvalues. *SIAM J. Optim.* **2**, 88–120.

QIN, J. & ZHANG, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Statist. Soc. B* **69**, 101–122.

ROBINS, J., ROTNITZKY, A. & ZHAO, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.* **89**, 846–866.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* **26**, 20–36.

TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–682.

WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

ZHOU, D.-X. (2002). The covering number in learning theory. *J. Complexity* **18**, 739–767.

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Ass.* **110**, 910–922.

[*Received x x. Revised x x*]

Biometrika (2016), *pp*. 1–14

Printed in Great Britain

Advance Access publication on xx xx xxxx

Kernel-based covariate functional balancing for observational studies

BY RAYMOND K. W. WONG

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.

raywong@stat.tamu.edu

AND KWUN CHUEN GARY CHAN

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

kcgchan@u.washington.edu

SUMMARY

Covariate balance is often advocated for objective causal inference since it mimics randomization in observational data. Unlike methods that balance specific moments of covariates, our proposal attains uniform approximate balance for covariate functions in a reproducing-kernel Hilbert space. The corresponding infinite-dimensional optimization problem is shown to have a finite-dimensional representation in terms of an eigenvalue optimization problem. Large-sample results are studied, and numerical examples show that the proposed method achieves better balance, with smaller sampling variability than existing methods.

Some key words: Average treatment effect; Eigenvalue optimization; Reproducing-kernel Hilbert space; Sobolev space.

1. INTRODUCTION

The estimation of average treatment effects is important in the evaluation of an intervention or a treatment, but is complicated by confounding in observational studies where the treatment is not randomly assigned. When treatment assignment is unconfounded conditional on observable covariates, two popular modeling strategies are based respectively on propensity score modeling (Rosenbaum & Rubin, 1983) and outcome regression modeling. Parametric approaches can suffer seriously from model misspecification, and there have been substantial recent efforts to construct more robust estimators within these modeling frameworks; see for example, Robins et al. (1994), Qin & Zhang (2007), Tan (2010), Graham et al. (2012), and Han & Wang (2013).

Since randomization is a gold standard to identify average treatment effects, Rubin (2007) advocated mimicking randomization, which balances the covariate distributions among the treated, the controls, and the combined sample, in the analysis of observational data. Based on these considerations, weighting-based covariate balancing methods have been proposed by Qin & Zhang (2007), Hainmueller (2012), Imai & Ratkovic (2014), Zubizarreta (2015) and Chan et al. (2016). A common feature of these methods is that a vector of user-specified functions of covariates is balanced. While balancing low-order moments of the covariates often yields good results, there is no guarantee that there will be sufficient balance over a large class of covariate functions. Matching is another general idea to attain covariate balance. Exact matching is not feasible for multiple continuous covariates, and a user-specified coarsening of the covariate space is needed (Iacus et al., 2011). In this paper, we shall focus on weighting-based methods.

Instead of balancing pre-specified moments of covariates, we propose a method to control the covariate functional balance over a reproducing-kernel Hilbert space (Aronszajn, 1950), which can be chosen large enough to contain any functions with mild smoothness constraints, including non-linearities and interactions. At a conceptual level, the comparison between covariate balancing with an increasing number of basis functions and kernel-based covariate functional balancing is analogous to the comparison of regression and smoothing splines in conditional mean estimation. Unlike regression splines, smoothing splines do not require pre-selection of the number of knots and their locations. Although achieving our goal involves a challenge due to an infinite-dimensional optimization problem, we show that it has a finite-dimensional representation and can be solved by eigenvalue optimization. Large sample properties are derived under minimal smoothness conditions on the outcome regression model. Consistent estimation of average treatment effects is then possible without first guessing or estimating the outcome regression function, and efficient estimation can be attained when the outcome regression function is estimated. Unlike weighting methods that require stringent smoothness conditions for the propensity score function, our method does not require smoothness of the propensity score.

2. KERNEL-BASED COVARIATE FUNCTIONAL BALANCING

2·1. *Preliminaries*

Let $Y(1)$ and $Y(0)$ be the potential outcomes when an individual is assigned to the treatment or control group respectively. We are interested in estimating the population average treatment effect $\tau = E\{Y(1) - Y(0)\}$. In practice, $Y(1)$ and $Y(0)$ are not both observed. With T the binary treatment indicator, we can represent the observed outcome as $Y = TY(1) + (1-T)Y(0)$. Moreover, we observe a vector of covariates $X \in \mathcal{X}$ for every individual, so the observed data are $\{(T_i, Y_i, X_i), i = 1, \ldots, N\}$ where N is the sample size. We assume that $[\{T_i, Y_i(1), Y_i(0), X_i\}, i = 1, \ldots, N]$ are independent and identically distributed, and that T is independent of $\{Y(1), Y(0)\}$ conditional on X.

Note that τ consists of two expectations, $E\{Y(1)\}$ and $E\{Y(0)\}$. In this work, we consider weighted estimation of these expectations. Without loss of generality, we focus on $E\{Y(1)\}$. In the following, we consider a weighting estimator of $E\{Y(1)\}$ that can be represented as $N^{-1}\sum_{i=1}^{N} T_i w_i Y_i$. Hence, for estimation of $E\{Y(1)\}$, we only need to specify weights $w_i (i : T_i = 1)$ for individuals in the treatment group.

Let $\pi(x) = \mathrm{pr}(T = 1 \mid X = x)$ be the propensity score. Assuming knowledge of $\pi(X_i)$ $(i : T_i = 1)$, w_i can be chosen as $\{\pi(X_i)\}^{-1}$ to obtain a consistent estimator of $E\{Y(1)\}$. In practice, propensity scores are usually unknown. In such scenarios, one can estimate the propensity score function to form a plug-in estimator for $E\{Y(1)\}$. However, estimation errors and model misspecification of the propensity score function can lead to significant error in the estimation of $E\{Y(1)\}$ due to the use of inverse probability weighting. Poor finite-sample performance of such estimators has been reported in the literature (Kang & Schafer, 2007).

Due to this unsatisfactory performance, some attention has been given to choosing $w_i (i : T_i = 1)$ via covariate balancing, which mimics randomization directly. To understand this, note that

$$E\left\{\frac{Tu(X)}{\pi(X)}\right\} = E\{u(X)\}, \tag{1}$$

for any measurable function $u : \mathcal{X} \to \mathbb{R}$ such that $E\{u(X)\}$ exists and is finite. Instead of modeling the propensity function, it is therefore natural to choose weights that ensure the validity of

the empirical finite-dimensional approximation of (1),

$$\frac{1}{N}\sum_{i=1}^{N} T_i w_i U(X_i) = \frac{1}{N}\sum_{i=1}^{N} U(X_i), \tag{2}$$

where $U(X) = (u_1(X),\ldots,u_L(X))^{\mathrm{T}}$ is a L-variate function of X. Here $\mathrm{span}\{u_1,\ldots,u_L\}$ can be viewed as a finite-dimensional approximation space of functions in which the balancing is enforced. Practical considerations may suggest a choice of $\{u_1,\ldots,u_L\}$. In this case, we call it parametric covariate balancing. Without assumptions on the outcome regression model, the balancing of fixed and finitely many component functions u_j in (1) may not lead to consistent estimation (Hellerstein & Imbens, 1999). To allow consistent estimation in a larger family of outcome regression functions, another direction is to allow L to increase with N (Chan et al., 2016). This has a nonparametric flavor similar to regression splines for which the number of knots grows with sample size. However, the choices of L and $\{u_1,\ldots,u_L\}$ are not obvious. In this work, we aim to balance covariate functionals nonparametrically via reproducing-kernel Hilbert space modeling of the approximation space.

Let $m(X) = E\{Y(1) \mid X\}$ and $Y_i(1) = m(X_i) + \varepsilon_i$ for $i = 1,\ldots,N$. Further assume that the ε_i are independent with $E(\varepsilon_i \mid X_i) = 0$ and $E(\varepsilon_i^2 \mid X_i) = \sigma_i^2 < \infty$. All weighting estimator of $E\{Y(1)\}$ admits the decomposition

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} T_i w_i Y_i &= \left\{\frac{1}{N}\sum_{i=1}^{N} T_i w_i Y_i - \frac{1}{N}\sum_{i=1}^{N} m(X_i)\right\} + \left[\frac{1}{N}\sum_{i=1}^{N} m(X_i) - E\{Y(1)\}\right] + E\{Y(1)\} \\ &= \frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)m(X_i) + \frac{1}{N}\sum_{i=1}^{N} T_i w_i \varepsilon_i + \left[\frac{1}{N}\sum_{i=1}^{N} m(X_i) - E\{Y(1)\}\right] + E\{Y(1)\},\end{aligned} \tag{3}$$

which allows a transparent understanding of the terms that have to be controlled. The first term on the right-hand side of (3) poses a challenge since the unknown outcome regression function m is intrinsically related to the outcome data, and could be complex and high-dimensional in general. To connect with covariate balancing, if $m \in \mathrm{span}\{u_1,\ldots,u_L\}$ in (2), we can control the first term. For the second term, the $\varepsilon_i (i = 1,\ldots,N)$ are independent of the choice of $w_i (i : T_i = 1)$ if the outcome data are not used to obtain the weights. Some control over the magnitude of w_i will lead to convergence of the second term. Corresponding details will be given in §2·4. The convergence of the third term is ensured by the law of large numbers.

2·2. *Construction of the method*

We consider the following empirical validity measure for any suitable function u,

$$S_N(w,u) = \left\{\frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)\,u(X_i)\right\}^2,$$

where $w = (w_1,\ldots,w_N)^{\mathrm{T}}$. In parametric covariate balancing, weights $w_i (i : T_i = 1)$ can be constructed to satisfy

$$\sup_{u\in\mathcal{U}_L} S_N(w,u) = 0,$$

where $\mathcal{U}_L = \mathrm{span}\{u_1,\ldots,u_L\}$ with $u_1,\ldots,u_L$ being suitable basis functions. In this case, the weights attain exact covariate balance as in (2) when the dimension of $\mathcal{U}_L$ is small.

Here the overall validity of (1) is instead controlled directly on an approximation space $\mathcal{H}$, a reproducing-kernel Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$. Ideally, one would want to pick a large enough, possibly infinite-dimensional, space $\mathcal{H}$ to guarantee the control of $S_N(w,u)$ on a rich class of functions. Unlike sieve spaces, $\mathcal{H}$ is specified without reference to sample size. The matching of non-linear functions is also automatic if $\mathcal{H}$ is large enough to contain such functions, without the need to explicitly introduce particular non-linear basis functions in sieve spaces. For any Hilbert space $\mathcal{H}_1$ of functions of x_1 and any Hilbert space $\mathcal{H}_2$ of functions of x_2, the tensor product space $\mathcal{H}_1 \otimes \mathcal{H}_2$ is defined as the completion of the class $\{\sum_{k=1}^{\ell} f_1(x_1)f_2(x_2) : f_1 \in \mathcal{H}_1, f_2 \in \mathcal{H}_2, \ell = 1,2,\dots\}$ under the induced norm by $\mathcal{H}_1$ and $\mathcal{H}_2$. A popular choice of $\mathcal{H}$ is the tensor product reproducing-kernel Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2 \otimes \cdots \otimes \mathcal{H}_d$ with $\mathcal{H}_j$ being the reproducing-kernel Hilbert space of functions of the j-th component of X. Suppose the support of the covariate distribution is $[0,1]^d$ and $f^{(\ell)}$ is the ℓ-th derivative of a function f. Following Wahba (1990), one can pick $\mathcal{H}_j$ as the ℓ-th order Sobolev space $\mathcal{W}^{\ell,2}([0,1]) = \{f : f, f^{(1)}, \cdots, f^{(\ell-1)}$ are absolutely continuous, $f^{(\ell)} \in L^2[0,1]\}$ with norm

$$\|f\| = \left[\sum_{k=0}^{\ell-1}\left\{\int_0^1 f^{(k)}(t)dt\right\}^2 + \int_0^1 \left\{f^{(\ell)}(t)\right\}^2 dt\right]^{1/2}.$$

The second-order Sobolev space is one of the most common choices in practice and will be adopted in all of our numerical illustrations. Another common choice is the space generated by the Gaussian kernel, which will also be compared in numerical studies. If it is desirable to prioritize covariates based on prior beliefs, we can raise the components to different powers to reflect their relative importance. For Gaussian kernels, this is equivalent to using different bandwidth parameters for each covariate. In cases when there are binary or categorical covariates, one can choose the corresponding $\mathcal{H}_j$ as a reproducing-kernel Hilbert space with kernel $R(s,t) = I(s = t)$, for any levels s and t of such covariate, as suggested by Gu (2013); here I is an indicator function.

Ideally, we want to control $\sup_{u\in\mathcal{H}} S_N$. However, there are two issues. First, that $S_N(w,cu) = c^2 S_N(w,u)$ for any $c \geq 0$ suggests a scale issue of S_N with respect to u. Therefore, in order to use $S_N(w,u)$ to determine the weights w, the magnitude of u should be standardized. To cope with this, we notice

$$S_N(w,u) = \left\{\frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)u(X_i)\right\}^2 \leq \|u\|_N^2 \left\{\frac{1}{N}\sum_{i=1}^{N}(T_i w_i - 1)^2\right\} \tag{4}$$

due to the Cauchy–Schwarz inequality, where $\|u\|_N^2 = N^{-1}\sum_{i=1}^{N} u(X_i)^2$. In view of (4), we restrict our focus to $\widetilde{\mathcal{H}}_N = \{u \in \mathcal{H} : \|u\|_N = 1\}$. Second, similar to many statistical and machine learning frameworks, the optimization of an unpenalized sample objective function will result in overfitting. In our case, the weights become highly unstable. To alleviate this, we control $\|\cdot\|_{\mathcal{H}}$ to emphasize the balance on smoother functions. Additionally, we penalize on $V_N(w) = N^{-1}\sum_{i=1}^{N} T_i w_i^2$ to control both the variabilities of w and of the second term in the right-hand side of (3). Overall, we consider the constrained minimization,

$$\min_{w\geq 1}\left[\sup_{u\in\widetilde{\mathcal{H}}_N}\left\{S_N(w,u) - \lambda_1\|u\|_{\mathcal{H}}^2\right\} + \lambda_2 V_N(w)\right], \tag{5}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters and the above minimization is only taken over $w_i (i : T_i = 1)$. The weights w_i are restricted to be greater than or equal to 1, as their counter-

parts, inverse propensities, satisfy $\{\pi(X_i)\}^{-1} \geq 1$. We denote the solution of (5) by $\widehat{w}$. Further discussion on these tuning parameters will be given in §2.4 and §2.5. In particular, we show that the convergence to zero of the first term of (3) can be ensured even when $\lambda_2 = 0$. This indicates that this extra tuning parameter is mostly needed for our justification of the convergence of the second term in (3).

A small number of recent papers have also considered kernel-based methods for covariate balancing. An unpublished paper by Zhao (arXiv:1601.05890) considered a dual formulation of the method of Imai & Ratkovic (2014) for the estimation of $\pi(x)$ under a logistic regression model, and generalized the linear predictor into a non-linear one using the kernel trick. Since this method aims at estimating $\pi(x)$, it requires smoothness conditions on π and penalizes on smoothness of the resulting estimate. Our method does not require smoothness of π and penalizes the smoothness of the balancing functions. An unpublished paper by Kallus (arXiv:1606.05188) considered weights that minimize the dual norm of a balancing error. Given a reproducing-kernel Hilbert space, this method does not have the ability to adapt to a relevant subset of functions. An external parameter is required to index the function space, such as the dispersion parameter of a Gaussian kernel, which needs to be specified in an ad-hoc manner. Due to the lack of an explicit tuning parameter, this method will not work well for Sobolev space which does not have extra indexing parameters. Our method works for a given reproducing-kernel Hilbert space by using a data-adaptive tuning to promote balancing of smoother functions within the given space. An unpublished paper of Hazlett (arXiv: 1605.00155) proposed an extension of the moment-based balancing method of Hainmueller (2012) to balance the columns of the Gram matrix. Since the Gram matrix is $N \times N$, exact balancing of N moment conditions under additional constraints on the weights are often computationally infeasible. Balancing a low-rank approximation of the Gram matrix may be an ad-hoc solution but the theoretical properties have not been studied.

2·3. *Finite-dimensional representation*

Many common choices of reproducing-kernel Hilbert space, including Sobolev Hilbert space, are infinite-dimensional and therefore, the inner optimization in (5) is essentially an infinite-dimensional optimization which is seemingly impractical. Fortunately, we shall show that the solution of (5) enjoys a finite-dimensional representation. First, the inner optimization of (5) can be expressed as

$$\sup_{u \in \mathcal{H}} \left\{ \frac{S_N(w,u)}{\|u\|_N^2} - \lambda_1 \frac{\|u\|_{\mathcal{H}}^2}{\|u\|_N^2} \right\}.$$

Let K be the reproducing kernel of $\mathcal{H}$. By the representer theorem (Wahba, 1990), the solution lies in a finite-dimensional subspace $\mathrm{span}\{K(X_j,\cdot) : j = 1,\ldots,N\}$. Now this optimization is equivalent to:

$$\sup_{\alpha=(\alpha_1,\ldots,\alpha_N)^{\mathrm{T}} \in \mathbb{R}^N} \left[\frac{S_N\left\{w, \sum_{j=1}^N \alpha_j K(X_j,\cdot)\right\}}{\alpha^{\mathrm{T}} M^2 \alpha / N} - \lambda_1 \frac{\alpha^{\mathrm{T}} M \alpha}{\alpha^{\mathrm{T}} M^2 \alpha / N} \right], \tag{6}$$

where M is a $N \times N$ matrix with (i,j)-th element $K(X_i, X_j)$. This matrix is positive semi-definite and is commonly known as the Gram matrix. Let the eigen-decomposition of M be

$$M = \begin{pmatrix} P_1 & P_2 \end{pmatrix} \begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix} \begin{pmatrix} P_1^{\mathrm{T}} \\ P_2^{\mathrm{T}} \end{pmatrix},$$

where Q_1 and Q_2 are diagonal matrices. In particular, $Q_2 = 0$. Let r be the rank of Q_1. We remark that P_2 and Q_2 do not exist if $r = N$, but the following derivation still holds. Moreover,

$$S_N\left\{w, \sum_{j=1}^{N} \alpha_j K(X_j, \cdot)\right\} = \frac{1}{N^2}\alpha^{\mathrm{T}} M A(w) M \alpha, \tag{7}$$

where $A(w) = a(w)a(w)^{\mathrm{T}}$ with $a(w) = (T_1 w_1 - 1, T_2 w_2 - 1, \ldots, T_N w_N - 1)^{\mathrm{T}}$. Let $\beta = Q_1 P_1^{\mathrm{T}} \alpha / N^{1/2}$. The constrained optimization (6) is then equivalent to

$$\sup_{\beta \in \mathbb{R}^r : \|\beta\| \le 1} \beta^{\mathrm{T}} \left\{\frac{1}{N} P_1^{\mathrm{T}} A(w) P_1 - N\lambda_1 Q_1^{-1}\right\} \beta.$$

Therefore, the target optimization becomes

$$\min_{w \ge 1} \left[\sigma_{\max}\left\{\frac{1}{N} P_1^{\mathrm{T}} A(w) P_1 - N\lambda_1 Q_1^{-1}\right\} + \lambda_2 V_N(w)\right], \tag{8}$$

where $\sigma_{\max}(M)$ represents the maximum eigenvalue of a matrix M. Again, the above minimization is only taken over $w_i (i : T_i = 1)$. Since $P_1^{\mathrm{T}} a(w)$ is an affine transformation of w and V_N is a convex function, the objective function of this minimization is convex with respect to w, due to Proposition 1, whose proof is given in the Supplementary Material. Due to convexity and Slater's condition of strict feasibility, a necessary and sufficient condition for a global minimizer of (8) is the corresponding Karush–Kuhn–Tucker condition using subdifferentials.

PROPOSITION 1. *Let $B \in \mathbb{R}^{r \times r}$ be a symmetric matrix. The function $\sigma_{\max}(vv^{\mathrm{T}} + B)$ is convex with respect to $v \in \mathbb{R}^r$.*

As for the computation, we note that the maximum eigenvalue is evaluated at a rank-one modification of a diagonal matrix, which can be computed efficiently by solving the secular equation (O'leary & Stewart, 1990) in a common linear algebra package such as LAPACK. The objective function is second-order differentiable with respect to the w_i when the maximum eigenvalue of $P_1^{\mathrm{T}} A(w) P_1 / N - N\lambda_1 Q_1^{-1}$ has multiplicity 1. Moreover, the corresponding gradient has a closed-form expression. In this case, a common and fast nonlinear optimization method such as the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with bound constraints can be applied. Non-differentiability exists when the largest two eigenvalues of $P_1^{\mathrm{T}} A(w) P_1 / N - N\lambda_1 Q_1^{-1}$ coincide. To ensure validity, one could employ the following two-part computational strategy. First, one applies the Broyden–Fletcher–Goldfarb–Shanno algorithm and checks numerically whether the maximum eigenvalue evaluated at the resulting solution is repeated. If not, the objective function is differentiable at this solution and the Karush–Kuhn–Tucker condition is satisfied. Thus, the minimizer is obtained. Otherwise, the nonlinear eigenvalue optimization method of Overton (1992, Section 5), which is applicable to the scenario of repeated eigenvalues, is initialized by the former estimate and then applied. In our practical experience, the second step is seldom needed and has negligible effect to the final solution. Therefore, for fast computation, we only apply the first part in our numerical illustrations.

2·4. *Theoretical properties*

For notational simplicity, we shall study the theoretical properties of the proposed estimator for $\mathcal{H}$ being the tensor product of ℓ-th order Sobolev spaces, as studied extensively in smoothing splines (Wahba, 1990; Gu, 2013). Our results can be extended to other choices of $\mathcal{H}$ if an entropy result and a uniform boundedness of the unit ball $\{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \le 1\}$ are supplied; see the Supplementary Material. For instance, the respective entropy result of Gaussian reproducing-kernel

Hilbert space can be obtained from Zhou (2002). As mentioned, we concentrate on $E\{Y(1)\}$. Similar conditions are required for $E\{Y(0)\}$ to obtain results on the average treatment effect $\tau = E\{Y(1) - Y(0)\}$.

Assumption 1. The propensity $\pi(\cdot)$ is uniformly bounded away from 0. That is, there exists a constant C such that $1/\pi(x) \leq C < \infty$ for all $x \in \mathcal{X}$.

Assumption 2. The ratio d/ℓ is less than 2.

Assumption 3. The regression function $m(\cdot)$ belongs to $\mathcal{H}$.

Assumption 4. The errors $\{\varepsilon_i\}$ are uncorrelated where $E(\varepsilon_i) = 0$ and $\mathrm{var}(\varepsilon_i) = \sigma_i^2 \leq \sigma^2$ for all $i = 1, \ldots, N$. Further $\{\varepsilon_i\}$ are independent of $\{T_i\}$ and $\{X_i\}$.

The above assumptions are very mild. Assumption 1 is the usual overlap condition required for identification. There are no additional smoothness assumptions on $\pi(\cdot)$ which would typically be required in propensity score or covariate balancing methods (Hirano et al., 2003; Chan et al., 2016). Assumption 2 corresponds to the weakest smoothness assumption on $m(\cdot)$ in smoothing spline regression. We use the notation $A_N \asymp B_N$ to represent $A_n = O(B_N)$ and $B_N = O(A_N)$ for some sequences A_N and B_N.

THEOREM 1. *Suppose Assumptions* 1-3 *hold. If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 = O(N^{-1})$, *then* $S_N(\widehat{w}, m) = O_p(N^{-1})\|m\|_N^2$. *If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 \asymp N^{-1}$, *then* $V_N(\widehat{w}) = O_p(1)$ *and there exist constants* $W > 0$ *and* $S^2 > 0$ *such that* $E\{V_N(\widehat{w})\} \leq W$ *and* $E\{NS_N(\widehat{w}, m)\} \leq S^2$.

Theorem 1 supplies the rate of convergence of the first term in (3), and boundedness of the expectation of the second term in (3). Convergence of $S_N(\widehat{w}, m)$ is guaranteed even if λ_2 is chosen as 0. However, to ensure the boundedness of $E\{V_N(\widehat{w})\}$, additional regularization is needed and hence $\lambda_2 > 0$ is proposed. The following theorem establishes the $N^{1/2}$-consistency of the weighting estimator. Moreover, we show that the asymptotic distribution has a finite variance.

THEOREM 2. *Suppose Assumptions* 1–4 *hold and* $m \in \mathcal{H}$. *If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 \asymp N^{-1}$,

$$\frac{1}{N}\sum_{i=1}^{N} T_i \widehat{w}_i Y_i - E\{Y(1)\} = O_p(N^{-1/2}).$$

Moreover, $N^{1/2}[\sum_{i=1}^{N} T_i \widehat{w}_i Y_i/N - E\{Y(1)\}]$ *has finite asymptotic variance.*

Although Theorem 2 only gives the rate of convergence of the estimator, it is stronger than recent results for other kernel-based methods for the estimation of average treatment effects. Zhao (arXiv:1601.05890) and Hazlett (arXiv:1605.00155) do not provide the rate of convergences of their estimators. To our knowledge, the only paper that contains a rate of convergence for kernel-based methods is Kallus (arXiv:1612.08321), who showed a root-N convergence rate under a strong assumption that $m(X)$ is linear in X and did not develop the asymptotic distribution. In fact, when linear assumptions hold, parametric covariate balancing is sufficient for estimating the average treatment effects (Qin & Zhang, 2007). When $m(\cdot)$ is a general function, the difficulty in theoretical development lies in the first term of (3), which is shown to attain the same rate of convergence as the other two terms of (3), but its asymptotic distribution is not available. For the sieve-based method (Chan et al., 2016), the growth rate of the sieve approximation space can be carefully chosen in a range such that terms analogous to the first term of (3) have a faster convergence rate than the dominating terms. In our case, similar to nonparametric regression, there is only a particular growth rate of λ_1 such that the bias and variance of the first term of (3)

are balanced. In fact, it is possible that the term has an asymptotic bias of order $N^{-1/2}$. In §2·6, a modified estimator is studied by debiasing the first term of (3), so that its rate of convergence is faster than $N^{-1/2}$ and is dominated by the other terms. In that case, the asymptotic distribution can be derived. Further discussion of the relationship between Theorem 2 and the literature is given in Remark 3.

2·5. *Tuning parameter selection*

In Theorems 1 and 2, λ_1 and λ_2 are required to decrease at the same order N^{-1}, so as to achieve the desired asymptotic results. To reduce the amount of tuning, we choose $\lambda_2 = \zeta\lambda_1$ where $\zeta > 0$ is fixed. As explained above, λ_2 is chosen to be positive mostly to ensure the boundedness of $E\{V_N(\widehat{w})\}$. From our practical experience, the term $V_N(\widehat{w})$ is usually stable and does not take large values even if λ_2 is small. Therefore, we are inclined to choose a small ζ. In all of our numerical illustrations, ζ is fixed at 0.01. Now we focus on the choice of λ_1. Note that the tuning of λ_1 is similar to choosing the dimension of the sieve space in Chan et al. (2016), which is a difficult and mostly unsolved problem. In this paper, we do not attempt to solve this problem rigorously, but to provide a reasonable solution.

By Lagrange multipliers, the optimization $\sup_{u\in\widetilde{\mathcal{H}}_N}\{S_N(w,u) - \lambda_1\|u\|^2_{\mathcal{H}}\}$ is equivalent to $\sup_{\{u\in\widetilde{\mathcal{H}}_N:\|u\|_{\mathcal{H}}\le\gamma\}} S_N(w,u)$ for some γ, where there exists a correspondence between γ and λ_1. Since a larger regularization parameter corresponds to a stricter constraint, γ decreases with λ_1. We use

$$B_N(w) = \sup_{\{u\in\widetilde{\mathcal{H}}_N:\|u\|_{\mathcal{H}}\le\gamma\}} S_N(w,u), \tag{9}$$

as a measure of the balancing error over $\{u \in \widetilde{\mathcal{H}}_N : \|u\|_{\mathcal{H}} \le \gamma\}$ with respect to the weights w. Due to the large subset of functions to balance, $B_N(\widehat{w})$ is large when γ is large, or equivalently, when λ_1 is small. When γ decreases, or equivalently, λ_1 increases, $B_N(\widehat{w})$ typically decreases to approximately zero, as the resulting weight $\widehat{w}$ approximately balances the whole subset $\{u \in \widetilde{\mathcal{H}}_N : \|u\|_{\mathcal{H}} \le \gamma\}$. An example is given in Fig. 1 which will be discussed in §3.2. When this happens, a further decrease of γ would not lead to any significant decrease in $B_N(\widehat{w})$. The key idea is to choose the smallest λ_1 that achieves such approximate balancing, to ensure the largest subset of functions being well-balanced. In practice, we compute our estimator with respect to a grid of λ_1: $\lambda_1^{(1)} < \cdots < \lambda_1^{(J)}$. Write $\widehat{w}^{(j)}$ as the estimator with respect to $\lambda_1^{(j)}$. We select $\lambda_1^{(j^*)}$ as our choice of λ_1 if j^* is the smallest j such that

$$\frac{B_N(\widehat{w}^{(j+1)}) - B_N(\widehat{w}^{(j)})}{\lambda_1^{(j+1)} - \lambda_1^{(j)}} \ge e,$$

where e is chosen as a negative constant of small magnitude. In the numerical illustrations, we set $e = -10^{-6}$.

2·6. *An efficient modified estimator*

Since the outcome regression function $m(\cdot)$ is assumed to be in a reproducing-kernel Hilbert space $\mathcal{H}$, a kernel-based estimator $\widehat{m}(\cdot)$, such as smoothing splines (Gu, 2013), can be employed, and $N^{-1}\sum_{i=1}^N \widehat{m}(X_i)$ is a natural estimator of $E\{Y(1)\} = E[E\{Y(1) \mid X\}] = E\{m(X)\}$. However, since randomization is administered before collecting any outcome data, Rubin (2007) advocated the estimation of treatment effects without using outcome data to avoid data snooping. On the other hand, Chernozhukov et al. (arXiv:1608.00060) and Athey et al. (arXiv:1604.07125) advocate the use of an estimated outcome regression function to improve the theoretical results

in high-dimensional settings. Inspired by these results, we modify the weighting estimator by subtracting $N^{-1}\sum_{i=1}^{N}(T_i w_i - 1)\widehat{m}(X_i)$ from both sides of (3), so that the first term in the decomposition becomes $N^{-1}\sum_{i=1}^{N}(T_i w_i - 1)\{m(X_i) - \widehat{m}(X_i)\}$, while the remaining two terms are unchanged. It can then be shown that the first term has a rate of convergence faster than $N^{-1/2}$ under mild assumptions, and the asymptotic distribution of the resulting estimator will be derived.

The estimator takes the form

$$\frac{1}{N}\sum_{i=1}^{N}\{T_i w_i Y_i - (T_i w_i - 1)\widehat{m}(X_i)\} = \frac{1}{N}\sum_{i=1}^{N}[T_i w_i\{Y_i - \widehat{m}(X_i)\} + \widehat{m}(X_i)]$$

which has the same form as the residual balancing estimator proposed in Athey et al. (arXiv:1604.07125). They consider a different setting of high-dimensional linear regression model with sparsity assumptions, and showed that their estimator attains the semiparametric efficiency bound.

Our analysis requires the additional technical assumption such that $\widehat{w}$ is $o_p(N^{1/2})$. To achieve this, we adopt an assumption as in Athey et al. (arXiv:1604.07125) that $\widehat{w} \le BN^{1/3}$ for a prespecified large positive constant B. This can be enforced in the optimization (8) easily together with the constraint $\widehat{w} \ge 1$. For clarity, we call this estimator $\widetilde{w} = (\widetilde{w}_1, \ldots, \widetilde{w}_N)^{\mathrm{T}}$.

THEOREM 3. *Suppose Assumptions* 1, 2 *and* 4 *hold with* $\sigma_i^2 = \sigma^2$ *for all* i. *Also, assume* $\max_i E|\varepsilon_i|^3 < \infty$. *Let* $h = m - \widehat{m} \in \mathcal{H}$ *such that* $\|h\|_N = o_p(1)$ *and* $\|h\|_{\mathcal{H}} = O_p(1)$. *Further, assume* $\lambda_1 = o(N^{-1})$, $\lambda_2\|h\|_N^2 = o_p(N^{-1})$, *and* $\lambda_1^{-1} = o(\lambda_2^{(2\ell-d)/d}N^{2\ell/d})$. *Write*

$$J_N = N^{1/2}\left(\left[\frac{1}{N}\sum_{i=1}^{N}T_i\widetilde{w}_i\{Y_i - \widehat{m}(X_i)\} + \frac{1}{N}\sum_{i=1}^{N}\widehat{m}(X_i)\right] - E\{Y(1)\}\right),$$

$$J_N^* = [\mathrm{var}\{m(X_1)\}]^{1/2}F + \sigma N^{-1/2}\sum_{j=1}^{N}T_j\widetilde{w}_j G_j,$$

where $F, G_1, \ldots, G_N$ *are independent and identically distributed standard normal random variables independent of* $X_1, \ldots, X_N$, $T_1, \ldots, T_N$ *and* $\varepsilon_1, \ldots, \varepsilon_N$. *Let* ψ_N *and* ψ_N^* *be the corresponding characteristic function of* J_N *and* J_N^* *respectively. Then*

$$|\psi_N(t) - \psi_N^*(t)| \to 0, \quad t \in \mathbb{R},$$

where ψ_N^* *is twice differentiable, and*

$$\limsup_N \mathrm{var}(J_N) \le \mathrm{var}\{m(X_1)\} + \sigma^2 V, \tag{10}$$

where $V = E\{1/\pi(X_1)\}$.

COROLLARY 1. *Under the assumptions of Theorem* 3,

$$N^{1/2}\{\sigma^2 V_N(\widetilde{w})\}^{-1/2}\left[\frac{1}{N}\sum_{i=1}^{N}T_i\widetilde{w}_i\{Y_i - \widehat{m}(X_i)\} + \frac{1}{N}\sum_{i=1}^{N}\{\widehat{m}(X_i) - m(X_i)\}\right]$$

converges in distribution to a standard normal distribution as $N \to \infty$.

Remark 1. In Theorem 3, the estimand is $E\{Y(1)\}$, whereas in Corollary 1, the estimand is a finite-sample conditional average, $N^{-1}\sum_{i=1}^{N}E(Y_i(1) \mid X_i) = N^{-1}\sum_{i=1}^{N}m_1(X_i)$. Athey et al.

(arXiv: 1604.07125) considered a finite-sample conditional average treatment effect and obtained a result similar to Corollary 1. Normalization by $V_N(\widetilde{w})$ is possible in Corollary 1 following a conditional central limit theorem, since $\widetilde{w}$ depends only on (T_i, X_i) $(i = 1,\ldots,N)$ and can be treated as constants upon conditioning. To derive the limiting distribution of J_N in Theorem 3, one cannot use a similar normalization because the handling of extra term $N^{-1}\sum_{i=1}^{N}\{\widehat{m}(X_i) - m(X_i)\}$ requires averaging across the X distribution. If $V_N(\widetilde{w})$ converges to a constant in probability, one could use Slutsky's theorem to claim asymptotic normality of J_N. Theorem 3 requires a partially conditional central limit theorem which is proven in the Supplementary Material and the distribution of J_N can be approximated by a weighted sum of independent standard normal random variables. The asymptotic variance is bounded above by the right-hand side of (10), which is the semiparametric efficiency bound (Robins et al., 1994; Hahn, 1998).

Remark 2. Compared to Theorem 2, Theorem 3 requires different conditions on the orders of λ_1 and λ_2. These order specifications, together with a diminishing $\|h\|_N$, allow a direct asymptotic comparison between $V_N(\widetilde{w})$ and V, which leads to $V_N(\widetilde{w}) \le V\{1 + o_p(1)\}$. This is essential for achieving (10) in our proof. To make sense of the theorem, the conditions $\lambda_1 = o(N^{-1})$ and $\lambda_1^{-1} = o\{\lambda_2^{(2\ell-d)/d} N^{2\ell/d}\}$ should not lead to a null set of λ_1. As an illustration, suppose $\widehat{m}$ achieves the optimal rate $\|h\|_N \asymp N^{-\ell/(2\ell+d)}$, then one can take $\lambda_2 = o\{N^{-d/(2\ell+d)}\}$, which suggests $\lambda_2^{(2\ell-d)/d} N^{2\ell/d} = o\{N^{(d^2+4\ell^2)/(d^2+2\ell d)}\}$. Due to Assumption 2, $(d^2+4\ell^2)(d^2+2\ell d)^{-1} > 1$. Therefore, there exist choices of λ_1 and λ_2 that fulfill the assumption of Theorem 3. We found in simulations that the practical performance of the modified estimator is not sensitive to λ_1 and λ_2, and we thus use the method described in §2.5 to obtain these tuning parameters.

Remark 3. Most existing efficient methods require explicit or implicit estimation of both $\pi(\cdot)$ and $m(\cdot)$. Chernozhukov et al. (arXiv:1608.00060) gave a general result on the convergence rate required on both $\pi(\cdot)$ and $m(\cdot)$ for efficient estimation. Even though weighting methods do not explicitly estimate $m(\cdot)$, estimating-equation-based methods would give rise to implicit estimators of $m(\cdot)$ that attain good rates of convergence (Hirano et al., 2003; Chan et al., 2016). However, weights constructed based on complex optimization problems may not even converge to the true inverse propensities, see Athey et al. (arXiv: 1604.07125) who, under a sparse linear model assumption, proposed an efficient estimator by controlling the balancing error of linear functions and the estimation error for $m(\cdot)$. Although our modified estimator is not a direct kernel-based extension of their method, we have arrived at a similar conclusion. Our method only requires $\|\widehat{m} - m\|_N = o_p(1)$ and does not require the smoothness of $\pi(\cdot)$ or linearity of $m(\cdot)$, and is therefore less vulnerable to the curse of dimensionality. Note that the weighting estimator as described in Theorem 2 corresponds to $\widehat{m} = 0$, and therefore $\|\widehat{m}-m\|_N$ is not $o_p(1)$ when $m(\cdot)$ is not the zero function. In view of the current literature, Theorem 2 is an interesting result because convergence to neither $\pi(\cdot)$ nor $m(\cdot)$ is established.

3. NUMERICAL EXAMPLES

3·1. *Simulation study*

Simulation studies were conducted to evaluate the finite sample performance of the proposed estimator. We considered simulation settings where the propensity score and outcome regression models are non-linear functions of the observed covariates, with possibly non-smooth propensity score functions. For each observation, we generated a ten-dimensional multivariate standard Gaussian random vector $Z = (Z_1,\ldots,Z_{10})^{\mathrm{T}}$. The observed covariates are

$X = (X_1, \ldots, X_{10})^{\mathrm{T}}$ where $X_1 = \exp(Z_1/2), X_2 = Z_2/\{1 + \exp(Z_1)\}, X_3 = (Z_1 Z_3/25 + 0.6)^3$, $X_4 = (Z_2 + Z_4 + 20)^2$ and $X_j = Z_j$ $(j = 5, \ldots, 10)$. Three propensity score models are studied; model 1 is $\mathrm{pr}(T = 1 \mid Z) = \exp(-Z_1 - 0.1Z_4)/(1 + \exp(-Z_1 - 0.1Z_4)$, model 2 is $\mathrm{pr}(T = 1 \mid Z) = \exp\{-Z_1 - 0.1Z_4 + \eta_2(\widetilde{Z})\}/[1 + \exp\{-Z_1 - 0.1Z_4 + \eta_2(\widetilde{Z})\}]$, and model 3 is $\mathrm{pr}(T = 1 \mid Z) = \exp\{-Z_1 - 0.1Z_4 + \eta_3(\widetilde{Z})\}/[1 + \exp\{-Z_1 - 0.1Z_4 + \eta_3(\widetilde{Z})\}]$, where $\widetilde{Z} = (Z_2 + Z_4 + Z_6 + Z_8 + Z_{10})/5$, η_2 is the scaling function of the Daubechies 4-tap wavelet (Daubechies, 1992), and η_3 is the Weierstrass function with parameters $a = 2$ and $b = 13$. The functions η_2 and η_3 are chosen such that the propensity functions in models 2 and 3 are non-smooth. Two outcome regression models are studied: model A is $Y = 210 + (1.5T - 0.5)(27.4Z_1 + 13.7Z_2 + 13.7Z_3 + 13.7Z_4) + \epsilon$, and model B is $Y = Z_1 Z_2^3 Z_3^2 Z_4 + Z_4 |Z_1|^{0.5} + \epsilon$, where ϵ has standard normal distribution. For each scenario, we compared the proposed weighting and modified estimators using two commonly employed kernels: the second-order Sobolev kernel and the Gaussian kernel with bandwidth parameter chosen via the median heuristics (Gretton et al., 2005). We also compared the Horvitz–Thompson estimator where the weights are the inverse of propensity scores estimated by maximum likelihood under a working logistic regression model with X being the predictors, the Hájek estimator which is a normalized version of the Horvitz–Thompson estimator with weights summing up to N, the inverse probability weighting estimator using covariate balancing propensity score of Imai & Ratkovic (2014), the stable balancing weights of Zubizarreta (2015), and the nonparametric covariate balancing estimator of Chan et al. (2016) with exponential weights. The first moment of X was balanced explicitly for these methods. We compared the bias, root mean squared error and covariate balance of the methods, where covariate balance is evaluated at the true conditional mean function. In particular, we calculate $S_N(\widehat{w}, m)$ to evaluate the covariate balance of the treatment and the combined groups, also its counterpart for the covariate balance of the controls and the combined groups, and report the sum of these two measures. The reason for comparing the covariate balance at the true conditional mean function is that it is the optimal function to balance but is unknown in practice. For each scenario, 1000 independent data sets are generated, and the results for outcome models A and B with sample size $N = 200$ are given in Tables 1 and 2 respectively.

The results show that the empirical performance of the estimators are related to the degree of covariate balancing. Without any explicit covariate balancing, the Horvitz–Thompson estimator can be highly unstable. The Hájek estimator balances the constant function, the Imai–Ratkovic estimator balances X, the estimators of Zubizarreta and Chan et al. balance both the constant function and X. For outcome model A, the balance of both constant and X is important and the omission of either constraints can lead to a poor performance. For outcome model B, the balance of X often implies approximate balance of the constant and therefore the estimators of Imai and Ratkovic, as was Zubizarreta and Chan et al. had similar performance. However, in both cases, the proposed method outperformed the other estimators because it can also control the balance of nonlinear and higher-order moments. We attempted to compute a Horvitz–Thompson estimator using a smoothing spline logistic regression model with the same kernel as the proposed method using the R package `gss`, but the program did not converge in reasonable time. We also tried to exactly balance the second moments in addition to the first moments of ten baseline covariates in the existing methods, but the algorithms did not converge in a substantial fraction of simulations. This shortcoming of the existing methods can be circumvented by the proposed methods.

3·2. *Data analysis*

We compare the proposed methods with others using a study of the impact of child abduction by a militant group on the future income of abductees who escape later (Blattman & Annan, 2010). The data contain 741 males in Uganda collected during 2005–2006, of which 462 had

Table 1. *Biases, root mean squared errors and overall covariate balancing measures of various weighting estimators for outcome model A; the reported numbers are averages obtained from 1000 simulated datasets*

	PS1			PS2			PS3		
	Bias	RMSE	Bal	Bias	RMSE	Bal	Bias	RMSE	Bal
Proposed weighting (S)	−192	470	28	−114	422	20	−86	420	21
Proposed weighting (G)	−362	599	86	−232	505	58	−201	504	67
Proposed modified (S)	150	512	28	243	486	20	255	480	21
Proposed modified (G)	−91	378	86	45	368	58	−28	370	67
Horvitz–Thompson	>9999	>9999	>9999	>9999	>9999	>9999	7298	>9999	>9999
Hájek	837	1792	275	693	1429	165	662	1480	175
Imai–Ratkovic	−527	1720	331	−187	1523	274	224	1516	253
Zubizarreta	444	715	28	389	658	23	381	630	21
Chan et al.	262	594	21	226	549	17	244	533	16

The sample sizes were $N = 200$. The values of Bias and RMSE were multiplied by 100, RMSE represents the root mean squared error, and Bal represents an overall covariate balancing measure. S: Sobolev kernel; G: Gaussian kernel.

Table 2. *Biases, root mean squared errors and overall covariate balancing measures of various weighting estimators for outcome model B; the reported numbers are averages obtained from 1000 simulated datasets*

	PS1			PS2			PS3		
	Bias	RMSE	Bal	Bias	RMSE	Bal	Bias	RMSE	Bal
Proposed weighting (S)	−10	85	1	−7	81	0	−8	85	0
Proposed weighting (G)	−7	92	1	−3	88	0	−5	94	0
Proposed modified (S)	3	82	1	−8	82	0	−9	85	0
Proposed modified (G)	−6	89	1	−2	85	0	−4	92	0
Horvitz–Thompson	131	3629	1151	−1	881	66	−35	2092	451
Hájek	4	392	14	−3	221	4	−2	367	12
Imai–Ratkovic	−7	110	1	−6	97	1	−6	108	1
Zubizarreta	−8	115	1	−9	98	1	−10	108	1
Chan et al.	− 8	123	1	−9	99	1	−8	111	1

The sample sizes were $N = 200$. The values of Bias and RMSE were multiplied by 100, RMSE represents the root mean squared error, and Bal represents an overall covariate balancing measure. S: Sobolev kernel; G: Gaussian kernel.

been abducted by militant groups before 2005 but had escaped by the time of the study. Covariates include geographical region, age at 1996, father's education, mother's education, whether the parents had died during or before 1996, whether the father is a farmer and household size in 1996. The investigators chose to collect covariate values in 1996 because it predates most abductions and is also easily recalled as the year of the first election since 1980. The authors discuss the plausibility of the unconfounded treatment assignment since abduction is mostly due to random night raids on rural homes. The outcome of interest here is the daily wage of the study participants in Ugandan shillings in 2005. We compared the estimators as in the simulation studies in §3.1. Table 3 shows the point estimates and 95% confidence intervals based on 1000 bootstrap samples. All methods comparing the abducted to the non-abducted group give a small but non-significant decrease in income. However, a small difference is noted between the proposed method and other methods, indicating that a mild non-linear effect is possibly present, especially in the non-abducted group. To further illustrate this point, we compared the maximal balancing error $B_N(w)$ as a function of λ_1, which, as defined by (9), measures the balancing error as a func-

Table 3. *The effect of child abduction to income in Uganda.*

	$E\{Y(1)\}$	$E\{Y(0)\}$	τ
Proposed weighting (S)	1530 (1219, 1886)	1851 (1247, 2354)	−321 (−893, 431)
Proposed weighting (G)	1516 (1212, 1822)	1671 (1231, 2204)	−156 (−809, 382)
Proposed modified (S)	1532 (1239, 1945)	1867 (1256, 2489)	−355 (−993, 444)
Proposed modified (G)	1536 (1238, 1845)	1689 (1269, 2249)	−153 (−819, 380)
Horvitz–Thompson	1573 (1234, 2033)	2135 (1478, 3075)	−562 (−1667, 242)
Hájek	1573 (1234, 2027)	2131 (1471, 3064)	−558 (−1614, 241)
Imai and Ratkovic	1599 (1256, 2062)	1998 (1381, 2857)	−399 (−1312, 365)
Zubizarreta	1591 (1253, 2073)	2165 (1492, 3046)	−574 (−1613, 229)
Chan et al.	1580 (1246, 2060)	2144 (1485, 3034)	−564 (−1576, 238)

Numbers in parentheses are 95% confidence intervals. S: Sobolev kernel; G: Gaussian kernel.

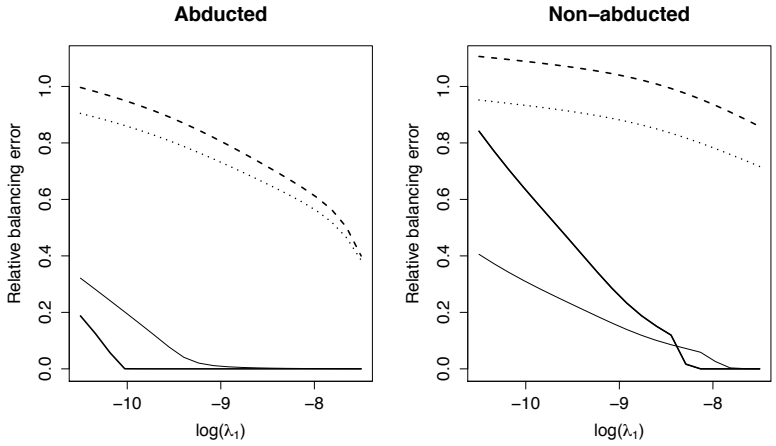

Fig. 1. Supremum balancing error for the child abduction data. Solid bold curves correspond to the proposed estimator using a Sobolev kernel, solid unbold curves correspond to the proposed estimator using a Gaussian kernel, dashed curves correspond to the Horvitz–Thompson estimator, dotted curves correspond to the Imai–Ratkovic estimator.

tion of the size of nested subsets of $\widetilde{\mathcal{H}}_N$, which is chosen as the Sobolev space. The subspace is smaller with an increasing λ_1, containing smoother functions. We standardize the comparisons by dividing the balancing error of constant weights that are used in unweighted comparisons. As seen in Fig. 1, the proposed estimator had approximately no balancing error after reaching a data-dependent threshold, so that any smoother functions can be approximately balanced. This is not the case for other estimators, since there will be residual imbalance for non-linear functions of a given smoothness. We note that the Imai–Ratkovic estimator has less balancing error than the Horvitz–Thompson estimator with maximum likelihood weights, because the former explicitly balances more moments than the latter, including linear and some non-linear covariate functionals. The Imai–Ratkovic estimator gives the closest result to the proposed estimators, but their confidence intervals are consistently narrower than other methods.

ACKNOWLEDGEMENT

The authors thank the editor, an associate editor and a reviewer for their helpful comments and suggestions. The work of the first author was partially supported by the U.S. National Science

Foundation. In addition, most of this work was conducted while the first author was affiliated with Iowa State University. The work of the second author was partially supported by the National Heart, Lung, and Blood Institute of the U.S. National Institutes of Health and the U.S. National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes the proofs of Proposition 1 and Theorems 1–3.

REFERENCES

ARONSZAJN, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.* **68**, 337–404.

BLATTMAN, C. & ANNAN, J. (2010). The consequences of child soldiering. *Rev. Econ. Stat.* **92**, 882–898.

CHAN, K. C. G., YAM, S. C. P. & ZHANG, Z. (2016). Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting. *J. R. Statist. Soc. B* **78**, 673–700.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets*. Philadelphia: SIAM.

GRAHAM, B. S., PINTO, C. C. D. X. & EGEL, D. (2012). Inverse probability tilting for moment condition models with missing data. *Rev. Econ. Stud.* **79**, 1053–1079.

GRETTON, A., HERBRICH, R., SMOLA, A., BOUSQUET, O. & SCHÖLKOPF, B. (2005). Kernel methods for measuring independence. *J. Mach. Learn. Res.* **6**, 2075–2129.

GU, C. (2013). *Smoothing Spline ANOVA Models*. New York: Springer.

HAHN, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* **66**, 315–332.

HAINMUELLER, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Polit. Anal.* **20**, 25–46.

HAN, P. & WANG, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* **100**, 417–430.

HELLERSTEIN, J. K. & IMBENS, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Rev. Econ. Stat.* **81**, 1–14.

HIRANO, K., IMBENS, G. & RIDDER, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161–1189.

IACUS, S. M., KING, G. & PORRO, G. (2011). Multivariate matching methods that are monotonic imbalance bounding. *J. Am. Statist. Ass.* **106**, 345–361.

IMAI, K. & RATKOVIC, M. (2014). Covariate balancing propensity score. *J. R. Statist. Soc. B* **76**, 243–263.

KANG, J. D. & SCHAFER, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion). *Statist. Sci.* **22**, 523–539.

O'LEARY, D. & STEWART, G. (1990). Computing the eigenvalues and eigenvectors of symmetric arrowhead matrices. *J. Comput. Phys.* **90**, 497–505.

OVERTON, M. L. (1992). Large-scale optimization of eigenvalues. *SIAM J. Optim.* **2**, 88–120.

QIN, J. & ZHANG, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies. *J. R. Statist. Soc. B* **69**, 101–122.

ROBINS, J., ROTNITZKY, A. & ZHAO, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Ass.* **89**, 846–866.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

RUBIN, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.* **26**, 20–36.

TAN, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–682.

WAHBA, G. (1990). *Spline Models for Observational Data*. Philadelphia: SIAM.

ZHOU, D.-X. (2002). The covering number in learning theory. *J. Complexity* **18**, 739–767.

ZUBIZARRETA, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *J. Am. Statist. Ass.* **110**, 910–922.

[*Received x x. Revised x x*]

Biometrika (2016), *pp*. 1–11

Printed in Great Britain

Advance Access publication on xx xx xxxx

Supplementary material for 'Kernel-based covariate functional balancing for observational studies'

BY RAYMOND K. W. WONG

Department of Statistics, Texas A&M University, College Station, Texas 77843, U.S.A.
raywong@stat.tamu.edu

AND KWUN CHUEN GARY CHAN

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
kcgchan@u.washington.edu

S1. PROOFS AND TECHNICAL RESULTS

S1·1. *Proof of Proposition* 1

Proof of Proposition 1. Consider any $v_1, v_2 \in \mathbb{R}^r$, and $t \in [0,1]$. For $\beta \in \mathbb{R}^r$,

$$\begin{aligned}\beta^{\mathrm{T}}[\{tv_1+(1-t)v_2\}\{tv_1+(1-t)v_2\}^{\mathrm{T}}+B]\beta &= [\{tv_1+(1-t)v_2\}^{\mathrm{T}}\beta]^2+\beta^{\mathrm{T}}B\beta\\ &= \{tv_1^{\mathrm{T}}\beta+(1-t)v_2^{\mathrm{T}}\beta\}^2+\beta^{\mathrm{T}}B\beta\\ &\le t(v_1^{\mathrm{T}}\beta)^2+(1-t)(v_2^{\mathrm{T}}\beta)^2+\beta^{\mathrm{T}}B\beta\\ &= t\beta^{\mathrm{T}}(v_1v_1^{\mathrm{T}}+B)\beta+(1-t)\beta^{\mathrm{T}}(v_2v_2^{\mathrm{T}}+B)\beta\end{aligned}$$

Therefore, $\sigma_{\max}[\{tv_1+(1-t)v_2\}\{tv_1+(1-t)v_2\}^{\mathrm{T}}+B] \le t\sigma_{\max}(v_1v_1^{\mathrm{T}}+B)+(1-t)\sigma_{\max}(v_2v_2^{\mathrm{T}}+B)$.□

S1·2. *Proof of Theorems* 1 *and* 2

We begin with several definitions that will be used throughout the theoretical development. Write $w^* = (w_1^*,\ldots,w_N^*)^{\mathrm{T}} = [\{\pi(X_1)\}^{-1},\ldots,\{\pi(X_N)\}^{-1}]^{\mathrm{T}}$ and

$$F_{N,\lambda_1,\lambda_2}(w) = \sup_{u\in\widetilde{\mathcal{H}}_N}\left\{S_N(w,u)-\lambda_1\|u\|_{\mathcal{H}}^2\right\}+\lambda_2 V_N(w).$$

Obviously $w^* \ge 1$. Due to the definition of the proposed estimator, we have $F_{N,\lambda_1,\lambda_2}(\widehat{w}) \le F_{N,\lambda_1,\lambda_2}(w^*)$. This implies that for any $f \in \widetilde{\mathcal{H}}_N$,

$$S_N(\widehat{w},f)-\lambda_1\|f\|_{\mathcal{H}}^2+\lambda_2V_N(\widehat{w}) \le S_N(w^*,u^*)-\lambda_1\|u^*\|_{\mathcal{H}}^2+\lambda_2V_N(w^*), \tag{S1}$$

where $u^* = \operatorname{argmin}_{u\in\widetilde{\mathcal{H}}_N}\{S_N(w^*,u)-\lambda_1\|u\|_{\mathcal{H}}^2\}$ and its existence is shown in §2·3. Since $S_N(\widehat{w},u)=0$ for any $u\in\mathcal{H}$ such that $\|u\|_N=0$, (S1) also implies that, for any $u\in\mathcal{H}$,

$$S_N(\widehat{w},u)-\lambda_1\|u\|_{\mathcal{H}}^2+\lambda_2V_N(\widehat{w})\|u\|_N^2 \le \left\{S_N(w^*,u^*)-\lambda_1\|u^*\|_{\mathcal{H}}^2+\lambda_2V_N(w^*)\right\}\|u\|_N^2. \tag{S2}$$

In below, we adopt several choices of $f\in\widetilde{\mathcal{H}}_N$ in (S1) and $u\in\mathcal{H}$ in (S2) to obtain various results. For instance, one obvious candidate of $f\in\widetilde{\mathcal{H}}$ is the constant function z where $z(x)\equiv 1$. On the other hand, the control of $S_N(w^*,u^*)$ is given in the following Lemma S1, whose proof is given in §S1·4, so as to control the right-hand side of (S1) and (S2).

LEMMA S1. *Suppose Assumptions* 1 *and* 2 *hold. Let* $w^* = (w_1^*, \ldots, w_N^*)^{\mathrm{T}} = [\{\pi(X_1)\}^{-1}, \ldots, \{\pi(X_N)\}^{-1}]^{\mathrm{T}}$. *There exists a constant* $c \geq 0$ *such that for all* $T \geq c$,

$$\mathrm{pr}\left\{\sup_{u \in \widetilde{\mathcal{H}}_N} \frac{N S_N(w^*, u)}{\|u\|_{\mathcal{H}}^{d/\ell}} \geq T^2\right\} \leq c \exp\left(-\frac{T^2}{c^2}\right).$$

Moreover, by central limit theorem, $V_N(w^*) = V + O_p(N^{-1/2})$ where $V = E\{\pi(X_1)^{-1}\}$. To prove Theorem 1, it suffices to establish the following two lemmas (Lemmas S2 and S3). The proof is given in §S1·4.

LEMMA S2. *Suppose Assumptions* 1 *and* 2 *hold. If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 \asymp N^{-1}$, *we have* $S_N(\widehat{w}, z) = O_p(N^{-1})$ *and* $V_N(\widehat{w}) = O_p(1)$. *Moreover, there exists a constant* $W > 0$ *such that* $E\{V_N(\widehat{w})\} \leq W$.

*Proof of Lemma S*2. Taking f as z (constant function of value 1) in (S1), we obtain a basic inequality:

$$S_N(\widehat{w}, z) + \lambda_1 \|u^*\|_{\mathcal{H}}^2 + \lambda_2 V_N(\widehat{w}) \leq S_N(w^*, u^*) + \lambda_1 \|z\|_{\mathcal{H}}^2 + \lambda_2 V_N(w^*). \tag{S3}$$

By Lemma S1, there exists a constant c such that for all $T \geq c$, $\mathrm{pr}\{S_N(w^*, u^*) \leq T^2 N^{-1} \|u^*\|_{\mathcal{H}}^{d/\ell}\} \geq 1 - c\exp(-T^2/c^2)$.

Let $\widetilde{E}_{N,1}$, $\widetilde{E}_{N,2}$ and $\widetilde{E}_{N,3}$ be the events that $S_N(w^*, u^*)$ is the largest in right-hand side of (S3), that $\lambda_1 \|z\|_{\mathcal{H}}^2$ is the largest in right-hand side of (S3), and that $\lambda_2 V_N(w^*)$ is the largest in right-hand side of (S3), respectively. Note that they are not necessarily disjoint. We write $E_{N,1} = \widetilde{E}_{N,1}$, $E_{N,2} = \widetilde{E}_{N,2} \setminus \widetilde{E}_{N,1}$ and $E_{N,3} = \widetilde{E}_{N,3} \setminus (\widetilde{E}_{N,1} \cup \widetilde{E}_{N,2})$. Therefore $\{E_1, E_2, E_3\}$ forms a partition of the sample space. We can further divide the event $E_{N,1}$ into two disjoint events, $E_{N,1,T} = E_{N,1} \cap \{S_N(w^*, u^*) \leq T^2 N^{-1} \|u^*\|_{\mathcal{H}}^{d/\ell}\}$ and $\check{E}_{N,1,T} = E_{N,1} \cap \{S_N(w^*, u^*) > T^2 N^{-1} \|u^*\|_{\mathcal{H}}^{d/\ell}\}$. Note that $\{E_{N,1,T} \cup \check{E}_{N,1,T} \cup E_{N,2} \cup E_{N,3}\}$ forms a partition of the sample space. We analyze (S3) on these events.

Case (i): On $E_{N,1,T}$, (S3) leads to $S_N(\widehat{w}, z) \leq T^{4d/(2\ell-d)} \lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)}$, $\|u^*\|_{\mathcal{H}} \leq T^{2\ell/(2\ell-d)} \lambda_1^{-\ell/(2\ell-d)} N^{-\ell/(2\ell-d)}$ and $\lambda_2 V_N(\widehat{w}) \leq T^{4d/(2\ell-d)} \lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)}$.

Case (ii): On $E_{N,2}$, (S3) leads to $S_N(\widehat{w}, z) \leq 3\lambda_1 \|z\|_{\mathcal{H}}^2$, $\|u^*\|_{\mathcal{H}} \leq 3\|z\|_{\mathcal{H}}$ and $\lambda_2 V_N(\widehat{w}) \leq 3\lambda_1 \|z\|_{\mathcal{H}}^2$.

Case (iii): On $E_{N,3}$, (S3) leads to $S_N(\widehat{w}, z) \leq 3\lambda_2 V_N(w^*)$, $\lambda_1 \|u^*\|_{\mathcal{H}}^2 \leq 3\lambda_2 V_N(w^*)$ and $\lambda_2 V_N(\widehat{w}) \leq 3\lambda_2 V_N(w^*)$.

Now, we focus on $S_N(\widehat{w}, z)$:

$$\begin{aligned}
&\mathrm{pr}\left[S_N(\widehat{w}, z) \leq \max\left\{T^{4d/(2\ell-d)} \lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)}, 3\lambda_1 \|z\|_{\mathcal{H}}^2, 3\lambda_2 V_N(w^*)\right\}\right] \\
&= \sum_{i=1}^{3} \mathrm{pr}\left[S_N(\widehat{w}, z) \leq \max\left\{T^{4d/(2\ell-d)} \lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)}, 3\lambda_1 \|z\|_{\mathcal{H}}^2, 3\lambda_2 V_N(w^*)\right\} \cap E_{N,i}\right] \\
&\geq \mathrm{pr}(E_{N,1,T}) + \mathrm{pr}\left[\left\{S_N(\widehat{w}, z) \leq T^{4d/(2\ell-d)} \lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)}\right\} \cap \check{E}_{N,1,T}\right] + \mathrm{pr}(E_{N,2}) + \mathrm{pr}(E_{N,3}) \qquad \text{(S4)} \\
&= 1 - \mathrm{pr}\left[\left\{S_N(\widehat{w}, z) > T^{4d/(2\ell-d)} \lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)}\right\} \cap \check{E}_{N,1,T}\right]
\end{aligned}$$

$$\geq 1 - \mathrm{pr}(\check{E}_{N,1,T}) \geq 1 - \mathrm{pr}\left\{S_N(w^*, u^*) > T^2 N^{-1}\|u^*\|_{\mathcal{H}}^{d/\ell}\right\} = \mathrm{pr}\left\{S_N(w^*, u^*) \leq T^2 N^{-1}\|u^*\|_{\mathcal{H}}^{d/\ell}\right\} \quad (60)$$
$$\geq 1 - c\exp(-T^2/c^2),$$

for all $T \geq c$, where (S4) follows from the above analyses of Cases (i), (ii) and (iii). We can show that $N^{1/2}\{V_N(w^*) - V\}$ converges to $N(0, \sigma_V^2)$ in distribution by central limit theorem, where $V = E[\{\pi(X_1)\}^{-1}] < \infty$ and $\sigma_V^2 = \{1 - \pi(X_1)\}/\pi(X_1)^3 < \infty$. Therefore $S_N(\widehat{w}, z) = O_p(N^{-1})$ under the condition that $\lambda_1 \asymp N^{-1}$ and $\lambda_2 = O(N^{-1/2})$. If $\lambda_1 \asymp N^{-1}$ and $\lambda_2 \asymp N^{-1}$, similar (65) arguments lead to $V_N(\widehat{w}) = O_p(1)$.

Now, we focus on $E\{V_N(\widehat{w})\}$ under $\lambda_2 > 0$. The requirements that $\lambda_1 \asymp N^{-1}$ and $\lambda_2 \asymp N^{-1}$ imply $\widetilde{B}_1 N^{-1} \leq \lambda_1 \leq B_1 N^{-1}$ and $\widetilde{B}_2 N^{-1} \leq \lambda_2 \leq B_2 N^{-1}$ for some positive constants $\widetilde{B}_1$, B_1, $\widetilde{B}_2$ and B_2. We derive bounds for each term in the following decomposition of $E\{V_N(\widehat{w})\}$:

$$E\{V_N(\widehat{w}) \mid E_{N,1}\}\mathrm{pr}(E_{N,1}) + E\{V_N(\widehat{w}) \mid E_{N,2}\}\mathrm{pr}(E_{N,2}) + E\{V_N(\widehat{w}) \mid E_{N,3}\}\mathrm{pr}(E_{N,3}).$$

The first term: Fix $\widetilde{c} = \max\{c, 2\widetilde{B}_1^{-d/(2\ell-d)}\widetilde{B}_2^{-1}, 2\}$ and $a > 0$ such that $\widetilde{c} > (70)$ $\widetilde{B}_1^{-d/(2\ell-d)}\widetilde{B}_2^{-1}\widetilde{c}^{4da/(2\ell-d)}$ and $4da/(2\ell-d) < 1$. That means, a is a constant fulfilling $\min[(2\ell - d)\log\{\widetilde{c}\widetilde{B}_1^{d/(2\ell-d)}\widetilde{B}_2\}/4d\log\widetilde{c}, (2\ell-d)/4d] > a > 0$.

$$\begin{aligned} E\{V_N(\widehat{w}) \mid E_{N,1}\}\mathrm{pr}(E_{N,1}) &= \int_0^\infty \mathrm{pr}\{V_N(\widehat{w}) > t \mid E_{N,1}\}\mathrm{pr}(E_{N,1})dt \\ &= \int_0^\infty \mathrm{pr}[\{V_N(\widehat{w}) > t\} \cap E_{N,1}]dt \\ &\leq \widetilde{c} + \int_{\widetilde{c}}^\infty \mathrm{pr}[\{V_N(\widehat{w}) > t\} \cap E_{N,1,t^a}]dt + \int_{\widetilde{c}}^\infty \mathrm{pr}[\{V_N(\widehat{w}) > t\} \cap \check{E}_{N,1,t^a})]dt \end{aligned}$$

Due to the construction of a,

$$\int_{\widetilde{c}}^\infty \mathrm{pr}[\{V_N(\widehat{w}) > t\} \cap E_{N,1,t^a}]dt \leq \int_{\widetilde{c}}^\infty \mathrm{pr}\{t^{4da/(2\ell-d)}\widetilde{B}_1^{-d/(2\ell-d)}\widetilde{B}_2^{-1} \geq V_N(\widehat{w}) > t\}dt = 0.$$

Now, we look at the last term

$$\int_{\widetilde{c}}^\infty \mathrm{pr}[\{V_N(\widehat{w}) > t\} \cap \check{E}_{N,1,t^a}]dt \leq \int_{\widetilde{c}}^\infty \mathrm{pr}(\check{E}_{N,1,t^a})dt \leq \int_{\widetilde{c}}^\infty c\exp\left(\frac{-t^{2a}}{c^2}\right)dt = -c\frac{c^{1/a}\Gamma(1/2a, t^{2a}/c^2)}{2a}\Bigg|_{\widetilde{c}}^\infty,$$

which is bounded due to the fact that $\Gamma(s,x)/(x^{s-1}e^{-x}) \to 1$ as $x \to \infty$. Therefore $E\{V_N(\widehat{w}) \mid (80)$ $E_{N,1}\}\mathrm{pr}(E_{N,1}) < \infty$.

The second term: As shown in Case (ii):

$$E\{V_N(\widehat{w}) \mid E_{N,2}\}\mathrm{pr}(E_{N,2}) \leq 3B_1\widetilde{B}_2^{-1}\|z\|_{\mathcal{H}}^2 < \infty$$

The third term: As shown in Case (iii):

$$E\{V_N(\widehat{w}) \mid E_{N,3}\}\mathrm{pr}(E_{N,3}) \leq 3B_2E\{V_N(w^*) \mid E_{N,3}\}\mathrm{pr}(E_{N,3}) \leq 3B_2E\{V_N(w^*)\} = B_2V < \infty. \quad (85)$$

Combining the above results, there exists a constant $W > 0$ such that $E\{V_N(\widehat{w})\} \leq W$. □

LEMMA S3. *Suppose Assumptions* 1-3 *hold. If* $\lambda_1 \asymp N^{-1}$ *and* $\lambda_2 = O(N^{-1})$, *then* $S_N(\widehat{w}, m) = O_p(N^{-1})\|m\|_N^2$. *Further, if* $\lambda_2 \asymp N^{-1}$, *there exists a constant* $S^2 > 0$ *such that* $E\{NS_N(\widehat{w}, m)\} \leq S^2$.

Proof of Lemma S3. Rearranging the terms in (S2), we obtain the basic inequality:

$$S_N(\widehat{w}, m) + \lambda_1 \|u^*\|_{\mathcal{H}}^2 \|m\|_N^2 + \lambda_2 V_N(\widehat{w}) \|m\|_N^2 \leq S_N(w^*, u^*) \|m\|_N^2 + \lambda_1 \|m\|_{\mathcal{H}}^2 + \lambda_2 V_N(w^*) \|m\|_N^2. \tag{S5}$$

By Lemma S1, we have $S_N(w^*, u^*) = O_p(N^{-1}) \|u^*\|_{\mathcal{H}}^{d/\ell}$. Now we compare different scenarios of (S5):

Case (i): Suppose that $S_N(w^*, u^*)\|m\|_N^2$ is the largest in right-hand side of (S5). If $\|m\|_N \neq 0$, we have $\|u^*\|_{\mathcal{H}} \leq \lambda_1^{-\ell/(2\ell-d)} O_p\{N^{-\ell/(2\ell-d)}\}$ and therefore $S_N(\widehat{w}, m) \leq \lambda_1^{-d/(2\ell-d)} O_p\{N^{-2\ell/(2\ell-d)}\}\|m\|_N^2$. As if $\|m\|_N = 0$, we have $S_N(\widehat{w}, m) = 0 \leq \lambda_1^{-d/(2\ell-d)} O_p\{N^{-2\ell/(2\ell-d)}\}\|m\|_N^2$.

Case (ii): Suppose that $\lambda_1\|m\|_{\mathcal{H}}^2$ is the largest in right-hand side of (S5). We obtain $S_N(\widehat{w}, m) \leq 3\lambda_1\|m\|_{\mathcal{H}}^2$.

Case (iii): Suppose that $\lambda_2 V_N(w^*)\|m\|_N^2$ is the largest in right-hand side of (S5). We obtain $S_N(\widehat{w}, m) \leq 3\lambda_2\{V + O_p(N^{-1/2})\}\|m\|_N^2$.

Due to Lemma S7 in §S1·4, $\|m\|_N \leq R\|m\|_{\mathcal{H}} < \infty$. Overall, we have

$$S_N(\widehat{w}, m) = O_p\left[\max\left\{\lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)} \|m\|_N^2,\, \lambda_1\|m\|_{\mathcal{H}}^2,\, \lambda_2\|m\|_N^2\right\}\right].$$

Since $S_N(\widehat{w}, m) = 0$ if $\|m\|_N^2 = 0$, we have $S_N(\widehat{w}, m) = O_p(N^{-1})\|m\|_N^2$ due to the conditions of λ_1 and λ_2.

Next, suppose $\lambda_2 \asymp N^{-1}$. Based on the exponential inequality in Lemma S1, one could apply a similar argument of Lemma S2, and show that there exists a constant $\widetilde{S}^2 > 0$ such $E\{N^2 \widetilde{S}_N^2(\widehat{w}, m)\} \leq \widetilde{S}^2$ where

$$\widetilde{S}_N(\widehat{w}, m) = \begin{cases} S_N(\widehat{w}, m/\|m\|_N), & \text{if } \|m\|_N \neq 0, \\ 0, & \text{if } \|m\|_N = 0. \end{cases}$$

Moreover,

$$\begin{aligned} E\{N S_N(\widehat{w}, m)\} = E\left\{N \widetilde{S}_N(\widehat{w}, m) \|m\|_N^2\right\} &\leq \frac{1}{2}\left[E\{N^2 \widetilde{S}_N^2(\widehat{w}, \widetilde{m})\} + E(\|m\|_N^4)\right] \\ &\leq \frac{1}{2}\left\{\widetilde{S}^2 + \frac{\int m^4 dP}{N} + \frac{N-1}{N}\left(\int m^2 dP\right)^2\right\} \\ &\leq \frac{1}{2}\left\{\widetilde{S}^2 + \int m^4 dP + \left(\int m^2 dP\right)^2\right\}. \end{aligned}$$

Due to Lemma S7 in §S1·4, $\int m^2 dP < \infty$ and $\int m^4 dP < \infty$. □

Proof of Theorem 2. Recall the decomposition:

$$\frac{1}{N}\sum_{i=1}^N T_i \widehat{w}_i Y_i = \frac{1}{N}\sum_{i=1}^N (T_i \widehat{w}_i - 1) m(X_i) + \frac{1}{N}\sum_{i=1}^N T_i \widehat{w}_i \varepsilon_i + \left[\frac{1}{N}\sum_{i=1}^N m(X_i) - E\{Y(1)\}\right] + E\{Y(1)\}.$$

Due to Lemma S7 in §S1·4, $\|m\|_2^2 = \int m^2 dP < \infty$. Since $X_1, \ldots, X_N$ are i.i.d., we can show that $\|m\|_N = \int m^2 dP + o_p(1)$. Therefore, the first term can be controlled:

$$\left| \frac{1}{N} \sum_{i=1}^{N} (T_i \widehat{w}_i - 1) m(X_i) \right| = S_N(\widehat{w}, m)^{1/2} = O_p(N^{-1/2}) \|m\|_2 + o_p(N^{-1/2}),$$

due to Theorem 1. Moreover, $E\{N S_N(\widehat{w}, m)\} < \infty$ due to Theorem 1. As for the second term, we write $\widehat{\delta}_i = T_i \widehat{w}_i$. Under Assumption 4, we have $E(\varepsilon_i \mid \widehat{\delta}_1, \ldots, \widehat{\delta}_N) = 0$. Therefore,

$$\operatorname{var}\left(\frac{1}{N} \sum_{i=1}^{N} T_i \widehat{w}_i \varepsilon_i \right) = E\left\{ \operatorname{var}\left(\frac{1}{N} \sum_{i=1}^{N} \widehat{\delta}_i \varepsilon_i \,\middle|\, \widehat{\delta}_1, \ldots, \widehat{\delta}_N \right) \right\} \leq \frac{\sigma^2}{N} E\{V_N(\widehat{w})\} \leq \frac{\sigma^2 W}{N},$$

due to Theorem 1. Therefore, $N^{-1} \sum_{i=1}^{N} T_i \widehat{w}_i \varepsilon_i = O_p(N^{-1/2})$. The above derivation also implies that

$$E\left\{ N^{-1/2} \sum_{i=1}^{N} (T_i \widehat{w}_i - 1) \varepsilon_i \right\}^2 < \infty.$$

Finally, by central limit theorem, we have

$$\left[N^{-1} \sum_{i=1}^{N} m(X_i) - E\{Y(1)\} \right] = O_p(N^{-1/2}),$$

due to Assumption 4. Also,

$$E\left[N^{-1/2} \sum_{i=1}^{N} m(X_i) - E\{Y(1)\} \right]^2 < \infty.$$

Therefore, $\sum_{i=1}^{N} T_i \widehat{w}_i Y_i / N - E\{Y(1)\} = O_p(N^{-1/2})$ and $N^{1/2}[\sum_{i=1}^{N} T_i \widehat{w}_i Y_i / N - E\{Y(1)\}]$ has bounded variance. □

S1·3. *Proof of Theorem* 3

LEMMA S4. *Suppose Assumptions* 1 *and* 2 *hold. Assume* $\lambda_1 = O(N^{-1})$ *and* $\lambda_1^{-1} = o\{\lambda_2^{(2\ell-d)/d} N^{2\ell/d}\}$. *We have* $V_N(\widetilde{w}) \leq V\{1 + o_p(1)\}$ *where* $V = E[\{\pi(X_1)\}^{-1}]$. *Moreover, there exists a constant* $W' > 0$ *such that* $E\{V_N(\widetilde{w})\} \leq W'$.

*Proof of Lemma S*4. Taking f as z (constant function of value 1) in (S1), we obtain the basic inequality:

$$S_N(\widetilde{w}, z) + \lambda_1 \|u^*\|_{\mathcal{H}}^2 + \lambda_2 V_N(\widetilde{w}) \leq S_N(w^*, u^*) + \lambda_1 \|z\|_{\mathcal{H}}^2 + \lambda_2 V_N(w^*), \tag{S6}$$

for all large N such that $1 \leq BN^{1/3}$. By Lemma S1, $S_N(w^*, u^*) = O_p(N^{-1}) \|u^*\|_{\mathcal{H}}^{d/\ell}$. Moreover, it is easy to show that $V_N(w^*) = V + O_p(N^{-1/2})$. Due to the condition of λ_1 and λ_2, we have $\lambda_1^{-d/(2\ell-d)} N^{-2\ell/(2\ell-d)} = o(\lambda_2)$ which implies $(\lambda_1 N)^{-d/(2\ell-d)} = o(\lambda_1^{-1} \lambda_2)$. As $\lambda_1 N = O(1)$, therefore $\lambda_1^{-1} \lambda_2 \to \infty$.

$$\lambda_1 \|z\|_{\mathcal{H}}^2 + \lambda_2 V_N(\widetilde{w}) = \lambda_2 \{o(1) + V + O_p(N^{-1/2})\} = \lambda_2 V\{1 + o_p(1)\}.$$

Now, we come back to (S6). Let $\mathcal{A}$ be the event that $S_N(w^*, u^*) \leq \lambda_1 \|z\|_{\mathcal{H}}^2 + \lambda_2 V_N(\widetilde{w})$. On the event $\mathcal{A}^c$, from (S6), we obtain $\|u^*\|_{\mathcal{H}} \leq \lambda_1^{-\ell/(2\ell-d)} O_p\{N^{-\ell/(2\ell-d)}\}$ which implies

$S_N(\widetilde{w}, u^*) \le \lambda_1^{-d/(2\ell-d)} O_p\{N^{-2\ell/(2\ell-d)}\} = o_p(\lambda_2)$ due to the conditions of λ_1 and λ_2. Notice that, on $\mathcal{A}^c$, we also have $S_N(w^*, u^*) > \lambda_1\|z\|_{\mathcal{H}}^2 + \lambda_2 V_N(\widetilde{w}) = \lambda_2 V\{1 + o_p(1)\}$. This implies that $\mathrm{pr}(\mathcal{A}^c) \to 0$ as $N \to \infty$. Therefore we only have to focus on $\mathcal{A}$. From (S6), we obtain $\lambda_1\|u^*\|_{\mathcal{H}}^2 \le 2\lambda_2\{V + o_p(1)\}$. In this case, $\|u^*\|_{\mathcal{H}}^2 \le 2\lambda_1^{-1}\lambda_2 V\{1 + o_p(1)\} = O_p(\lambda_1^{-1}\lambda_2)$. Therefore $S_N(w^*, u^*) = O_p(N^{-1})\|u^*\|_{\mathcal{H}}^{d/\ell} = O_p\{N^{-1}(\lambda_1^{-1}\lambda_2)^{d/(2\ell)}\} = o_p(\lambda_2)$. This implies that the right-hand side of (S6) is $\lambda_2 V\{1 + o_p(1)\}$. Hence $\lambda_2 V_N(\widetilde{w}) \le \lambda_2 V\{1 + o_p(1)\}$. Finally, using a similar but simpler argument in Lemma S2, one can show that there exists a constant $W' > 0$ such that $E\{V_N(\widetilde{w})\} \le W'$. □

LEMMA S5. *Suppose Assumptions* 1 *and* 2 *hold. Let* $h = m - \widehat{m} \in \mathcal{H}$ *such that* $\|h\|_N = o_p(1)$ *and* $\|h\|_{\mathcal{H}} = O_p(1)$. *Further, assume* $\lambda_1 = o(N^{-1})$, $\lambda_1^{-1}\|h\|_N^{2(2\ell-d)/d} = o_p(N)$ *and* $\lambda_2\|h\|_N^2 = o_p(N^{-1})$. *Then* $S_N(\widetilde{w}, h) = o_p(N^{-1})$. *Moreover, there exists a constant* $S' > 0$ *such that* $E\{NS_N(\widetilde{w}, h)\} \le S'$.

Proof of Lemma S5. Rearranging the terms in (S2), we obtain the basic inequality:

$$S_N(\widetilde{w}, h) + \lambda_1\|u^*\|_{\mathcal{H}}^2\|h\|_N^2 + \lambda_2 V_N(\widetilde{w})\|h\|_N^2 \le S_N(w^*, u^*)\|h\|_N^2 + \lambda_1\|h\|_{\mathcal{H}}^2 + \lambda_2 V_N(w^*)\|h\|_N^2, \quad \text{(S7)}$$

for all large N such that $C \le BN^{1/3}$. The rest of the proof is similar to the proof of Lemma S3 but with different conditions of λ_1 and λ_2. □

Proof of Theorem 3. Recall the decomposition:

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N T_i\widetilde{w}_i\{Y_i - \widehat{m}(X_i)\} + \frac{1}{N}\sum_{i=1}^N \widehat{m}(X_i) \\
&= \frac{1}{N}\sum_{i=1}^N (T_i\widetilde{w}_i - 1)h(X_i) + \frac{1}{N}\sum_{i=1}^N T_i\widetilde{w}_i\varepsilon_i + \left[\frac{1}{N}\sum_{i=1}^N m(X_i) - E\{Y(1)\}\right] + E\{Y(1)\}\,.
\end{aligned}$$

Note that the assumed conditions imply the conditions of Lemmas S4 and S5. By Lemma S5, the first term of the decomposition is $o_p(N^{-1/2})$. By dominated convergence theorem, with Skorohod Representation Theorem to extend its result to weakly convergent sequence of random variables, we have $\mathrm{var}\{N^{-1/2}\sum_{i=1}^N (T_i\widetilde{w}_i - 1)h(X_i)\} \le E\{NS_N(\widetilde{w}, h)\} \to 0$ using Lemma S5. Write

$$Z_N = N^{1/2}\left(\frac{1}{N}\sum_{i=1}^N T_i\widetilde{w}_i\varepsilon_i + \left[\frac{1}{N}\sum_{i=1}^N m(X_i) - E\{Y(1)\}\right]\right).$$

It is obvious that $\mathrm{var}(Z_N) = \mathrm{var}\{m(X_1)\} + \sigma^2 E V_N(\widetilde{w})$. By Lemma S4, we have $\limsup_N E V_N(\widetilde{w}) \le E\{V + o_p(1)\} = V$ using dominated convergence theorem.

Now, since the first term of the decomposition is $o_p(N^{-1/2})$, we focus on Z_N. We will utilize Theorem S1, by setting $\tau^2 = \mathrm{var}\{m(X_1)\}$, $g^2(\mathcal{D}_N) = \sigma^2 V_N(\widetilde{w})$, and

$$A_j = \frac{m(X_j) - E\{Y(1)\}}{[N\mathrm{var}\{m(X_1)\}]^{1/2}}, \quad \mathcal{B}_j = \{X_j, T_j\}, \quad C_j = \frac{T_j\widetilde{w}_j\varepsilon_j}{(\sigma^2\sum_{i=1}^N T_i\widetilde{w}_i^2)^{1/2}}, \quad (j = 1, \ldots, N).$$

Write $\mathcal{D}_N = \{A_1, \ldots, A_N, \mathcal{B}_1, \ldots, \mathcal{B}_N\}$. By the definition of $\widetilde{w}_i$ $(i : T_i = 1)$, $1 \le \widetilde{w}_i \le BN^{1/3}$ for all i. Therefore, $(\sum_{i=1}^N T_i\widetilde{w}_i^2)^{-1} = O_p(N)$ and $\max_i|\widetilde{w}_i| = o_p(N^{1/2})$. Moreover, $\max_i E|\varepsilon_i|^3 < \infty$ by

assumption. Hence

$$0 \le E\left(\sum_{i=1}^{N}|C_i|^3 \mid \mathcal{D}_N\right) = \frac{(\max_i E|\varepsilon_i|^3)\sum_{j=1}^{N} T_j\widetilde{w}_j^3}{\sigma^3(\sum_{i=1}^{N} T_i\widetilde{w}_i^2)^{3/2}} \le \frac{(\max_i E|\varepsilon_i|^3)\max_i|\widetilde{w}_i|}{\sigma^3(\sum_{i=1}^{N} T_i\widetilde{w}_i^2)^{1/2}} = o_p(1).$$

By Lemma S4, we have $E\{g^2(\mathcal{D}_N)\} \le M$ and $g^2(\mathcal{D}_N) \le M + o_p(1)$, by taking $M = \sigma^2\max\{W', V\}$. Write

$$Z_n = \tau\sum_{j=1}^{n} A_j + g(\mathcal{D}_N)\sum_{j=1}^{n} C_j = N^{1/2}\left(\frac{1}{N}\sum_{i=1}^{N} T_i\widetilde{w}_i\varepsilon_i + \left[\frac{1}{N}\sum_{i=1}^{N} m(X_i) - E\{Y(1)\}\right]\right)$$

$$Z_n^* = \tau F + g(\mathcal{D}_N)\sum_{j=1}^{n}\mathrm{var}(C_j \mid \mathcal{D}_N)^{1/2}G_j = [\mathrm{var}\{m(X_1)\}]^{1/2}F + \sigma N^{-1/2}\sum_{j=1}^{N} T_j\widetilde{w}_j G_j$$

where $F, G_1, \ldots, G_N$ are i.i.d. standard normal random variables independent of $C_1, \ldots, C_N$ and $\mathcal{D}_N$. Let ϕ_N and ϕ_N^* be the corresponding characteristic function of Z_N and Z_N^* respectively. Applying Theorem S1, we have $|\phi_N(t) - \phi_N^*(t)| \to 0$ for every $t \in \mathbb{R}$ and ϕ_N^* is twice differentiable.

S1·4. *Proof of Lemma S*1

LEMMA S6. *For $d/\ell < 2$, there exists a constant A such that the uniform entropy $H_\infty(\xi, \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \le 1\}) \le A\xi^{-d/\ell}$ for $\xi > 0$, where the uniform entropy is defined in Definition 2.3 of van de Geer (2000).*

*Proof of Lemma S*6. This is shown by Birman & Solomyak (1967) and the fact that $\mathcal{H}$ is a subspace of the Sobolev space $\mathcal{W}^{\ell,2}([0,1]^d)$. □

LEMMA S7. *There exists a constant R such that* $\sup_{\{u\in\mathcal{H}:\|u\|_{\mathcal{H}}\le 1\}}\|u\|_\infty \le R$.

*Proof of Lemma S*7. This is due to Lemma 2.1 of Lin (2000) and norm equivalence. □

*Proof of Lemma S*1. Let $\delta_i = T_i w_i^* - 1$. Note that the conditional expectation $E(\delta_i \mid X_i) = 0$. We will focus on the empirical process $\{N^{-1/2}\sum_{i=1}^{N}\delta_i u(X_i) : u \in \widetilde{\mathcal{H}}\}$. Due to Assumption 1, $0 < w_i^* \le C$ for all $i = 1, \ldots, N$. Therefore, $\delta_i (i = 1, \ldots, N)$ are uniformly sub-Gaussian: there exist constants K and σ_0^2, independent of X_i $(i = 1, \ldots, N)$, such that

$$\max_{i=1,\ldots,N} K^2\left[E\left(e^{|\delta_i|^2/K^2} \mid \{X_i\}_{i=1}^N\right) - 1\right] \le \sigma_0^2.$$

For instance, take $K = \max\{|C-1|, 1\}$ and $\sigma_0^2 = K^2(e-1)$, we have

$$K^2\left[E\left(e^{|\delta_i|^2/K^2} \mid \{X_i\}_{i=1}^N\right) - 1\right] \le K^2(e-1) = \sigma_0^2.$$

To derive the modulus of continuity of the aforementioned empirical process, we need upper bound on the entropy results related to $\mathcal{H}$ supplied by Lemma S6. Namely, under Assumption 2, there exists a constant A such that $H_\infty(\xi, \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \le 1\}) \le A\xi^{-d/\ell}$ for $\xi > 0$. Due to Lemma S7, there exists a constant R, independent of X_i $(i = 1, \ldots, N)$, such that $\sup_{\{u\in\mathcal{H}:\|u\|_{\mathcal{H}}\le 1\}}\|u\|_\infty \le R$. Now, we apply Lemma 8.4 of van de Geer (2000). For some constant c depending on A, d, ℓ,

R, K and σ_0, we have for all $T \geq c$,

$$\mathrm{pr}\left\{ \sup_{u \in \widetilde{\mathcal{H}}_N} \frac{|N^{-1/2} \sum_{i=1}^N \delta_i u(X_i)|}{\|u\|_{\mathcal{H}}^{d/2\ell}} \geq T \;\middle|\; X_1, \ldots, X_N \right\} \leq c \exp\left(-\frac{T^2}{c^2}\right).$$

Note that the constant c is independent of $\{X_i\}$ which leads to the unconditional probability inequality that, for all $T \geq c$,

$$\mathrm{pr}\left\{ \sup_{u \in \widetilde{\mathcal{H}}_N} \frac{|N^{-1/2} \sum_{i=1}^N \delta_i u(X_i)|}{\|u\|_{\mathcal{H}}^{d/2\ell}} \geq T \right\} \leq c \exp\left(-\frac{T^2}{c^2}\right).$$

This implies the desired result. □

S1·5. *Partially Conditional Central Limit Theorem*

THEOREM S1. *Let $(A_1, \mathcal{B}_1), \ldots, (A_n, \mathcal{B}_n)$ be independent and identically distributed where $A_1, \ldots, A_n$ are random variables and $\mathcal{B}_1, \ldots, \mathcal{B}_n$ are sets of random variables. Let $\{C_1, \ldots, C_n\}$ be another set of random variables. Write $\mathcal{D}_n = \{A_1, \ldots, A_n, \mathcal{B}_1, \ldots, \mathcal{B}_n\}$. Assume these variables satisfy*

$$E(A_j) = 0, \quad E(C_j \mid \mathcal{D}_n) = 0, \qquad (j = 1, \ldots, n),$$
$$\sum_{j=1}^n \mathrm{var}(A_j) = 1, \quad \sum_{j=1}^n \mathrm{var}(C_j \mid \mathcal{D}_n) = 1,$$

and there exists $\delta > 0$ such that $\sum_{j=1}^n E\left(|C_j|^{2+\delta} \mid \mathcal{D}_n\right) \to 0$ in probability. Moreover, $C_1, \ldots, C_n$ are conditionally independent given $\mathcal{D}_n$. Let g be a (non-random) function mapping from the support of $\mathcal{D}_n$ to $\mathbb{R}^+$ such that there exists a constant $M > 0$ such that $Eg^2(\mathcal{D}_n) \leq M$ and $g^2(\mathcal{D}_n) \leq M + o_p(1)$. For any positive real number τ, consider two random variables:

$$Z_n = \tau \sum_{j=1}^n A_j + g(\mathcal{D}_n) \sum_{j=1}^n C_j \quad \textit{and} \quad Z_n^* = \tau F + g(\mathcal{D}_n) \sum_{j=1}^n \{\mathrm{var}(C_j \mid \mathcal{D}_n)\}^{1/2} G_j,$$

where $F, G_1, \ldots, G_n$ are i.i.d. standard normal random variables independent of $C_1, \ldots, C_n$ and $\mathcal{D}_n$. Let ϕ_n and ϕ_n^* be the corresponding characteristic function of Z_n and Z_n^* respectively. Then $|\phi_n(t) - \phi_n^*(n)| \to 0$ for every $t \in \mathbb{R}$. Moreover, $E(Z_n^{*2}) = \tau^2 + E\{g^2(\mathcal{D}_n)\} \leq \tau^2 + M$ and therefore ϕ_n^* is twice differentiable.

Proof of Theorem S1. We extend the arguments of Dvoretzky (1972) to our partially conditional setting. Let $F_1, \ldots, F_n, G_1, \ldots, G_n$ be i.i.d. standard normal random variables independent of $C_1, \ldots, C_n$ and $\mathcal{D}_n$. Write $\sigma_{C,j}^2(\mathcal{D}_n) = \mathrm{var}(C_j \mid \mathcal{D}_n)$ for all $j = 1, \ldots, n$. Throughout this proof, i represents the complex number such that $i^2 = -1$. Let $t \in \mathbb{R}$. First,

$$\exp\left\{ it \left(\tau \sum_{j=1}^n A_j + g(\mathcal{D}_n) \sum_{j=1}^n C_j \right) \right\} - \exp\left[it \left\{ \tau \sum_{j=1}^n n^{-1/2} F_j + g(\mathcal{D}_n) \sum_{j=1}^n \sigma_{C,j}(\mathcal{D}_n) G_j \right\} \right]$$

$$= \left(\exp\left[it \left\{ \tau \sum_{j=1}^{n} A_j + g(\mathcal{D}_n) \sum_{j=1}^{n} C_j \right\} \right] - \exp\left[it \left\{ \tau \sum_{j=1}^{n} A_j + g(\mathcal{D}_n) \sum_{j=1}^{n} \sigma_{C,j}(\mathcal{D}_n) G_j \right\} \right] \right) \quad 220$$

$$+ \left(\exp\left[it \left\{ \tau \sum_{j=1}^{n} A_j + g(\mathcal{D}_n) \sum_{j=1}^{n} \sigma_{C,j}(\mathcal{D}_n) G_j \right\} \right] - \exp\left[it \left\{ \tau \sum_{j=1}^{n} n^{-1/2} F_j + g(\mathcal{D}_n) \sum_{j=1}^{n} \sigma_{C,j}(\mathcal{D}_n) G_j \right\} \right] \right).$$

Denote the first bracket and the second bracket as Q_1 and Q_2 respectively. Write $\tilde{C}_k = g(\mathcal{D}_n) \sum_{j=1}^{k} C_j$ and $\tilde{G}_k = g(\mathcal{D}_n) \sum_{j=k+1}^{n} \sigma_{C,j}(\mathcal{D}_n) G_j$.

$$\begin{aligned} Q_1 &= \exp\left(it\tau \sum_{j=1}^{n} A_j \right) \left\{ \exp\left(it\tilde{C}_n \right) - \exp\left(it\tilde{G}_0 \right) \right\} \\ &= \exp\left(it\tau \sum_{j=1}^{n} A_j \right) \sum_{k=1}^{n} \left[\exp\left\{ it(\tilde{C}_k + \tilde{G}_k) \right\} - \exp\left\{ it(\tilde{C}_{k-1} + \tilde{G}_{k-1}) \right\} \right] \quad 225 \\ &= \exp\left(it\tau \sum_{j=1}^{n} A_j \right) \sum_{k=1}^{n} \exp\left\{ it(\tilde{C}_{k-1} + \tilde{G}_k) \right\} \left[\exp\left\{ itg(\mathcal{D}_n) C_k \right\} - \exp\left\{ itg(\mathcal{D}_n) \sigma_{C,k}(\mathcal{D}_n) G_k \right\} \right] \end{aligned}$$

Therefore,

$$\begin{aligned} &|E(Q_1)| \\ &\leq \sum_{k=1}^{n} \left| E\left(\exp\left(it\tau \sum_{j=1}^{n} A_j \right) E\left[\exp\left\{ it(\tilde{C}_{k-1} + \tilde{G}_k) \right\} \mid \mathcal{D}_n \right] \right. \right. \\ &\qquad \left. \left. \times E\left[\exp\left\{ itg(\mathcal{D}_n) C_k \right\} - \exp\left\{ itg(\mathcal{D}_n) \sigma_{C,k}(\mathcal{D}_n) G_k \right\} \mid \mathcal{D}_n \right] \right) \right| \quad 230 \\ &\leq \sum_{k=1}^{n} E\left(\left| \exp\left(it\tau \sum_{j=1}^{n} A_j \right) \right| E\left[\left| \exp\left\{ it(\tilde{C}_{k-1} + \tilde{G}_k) \right\} \right| \mid \mathcal{D}_n \right] \right. \\ &\qquad \left. \times \left| E\left[\exp\left\{ itg(\mathcal{D}_n) C_k \right\} - \exp\left\{ itg(\mathcal{D}_n) \sigma_{C,k}(\mathcal{D}_n) G_k \right\} \mid \mathcal{D}_n \right] \right| \right) \\ &\leq \sum_{k=1}^{n} E\left| E\left[\exp\left\{ itg(\mathcal{D}_n) C_k \right\} - \exp\left\{ itg(\mathcal{D}_n) \sigma_{C,k}(\mathcal{D}_n) G_k \right\} \mid \mathcal{D}_n \right] \right| \end{aligned} \quad \text{(S8)}$$

Similarly,

$$\begin{aligned} |E(Q_2)| &\leq \left| E\left[\exp\left\{ itg(\mathcal{D}_n) \sum_{j=1}^{n} \sigma_{C,k}(\mathcal{D}_n) G_j \right\} \left\{ \exp\left(it\tau \sum_{j=1}^{n} A_j \right) - \exp\left(it\tau \sum_{j=1}^{n} n^{-1/2} F_j \right) \right\} \right] \right| \quad 235 \\ &\leq \left| E\left\{ \exp\left(it\tau \sum_{j=1}^{n} A_j \right) - \exp\left(it\tau \sum_{j=1}^{n} n^{-1/2} F_j \right) \right\} \right| \end{aligned}$$

$$\leq \sum_{k=1}^{n} \left| E\left\{ \exp\left(it\tau A_k\right) - \exp\left(it\tau n^{-1/2} F_k\right) \right\} \right|, \tag{S9}$$

where the last inequality is due to a similar argument applied to Q_1.

Now, we focus on (S8). As

$$\left| \exp(it) - \sum_{k=0}^{K} \frac{(it)^k}{k!} \right| \leq 2 \min\left\{ \frac{|t|^{n+1}}{(n+1)!}, \frac{2|t|^n}{n!} \right\}.$$

For any $\varepsilon > 0$,

$$\begin{aligned}
&\sum_{k=1}^{n} \left| E\left[\exp\left\{ itg(\mathcal{D}_n) C_k \right\} \mid \mathcal{D}_n \right] - 1 + \frac{1}{2} t^2 g^2(\mathcal{D}_n) \sigma_{C,k}^2(\mathcal{D}_n) \right| \\
&\leq \frac{1}{6} |t|^3 \sum_{k=1}^{n} E\left[|g(\mathcal{D}_n) C_k|^3 I\{ |g(\mathcal{D}_n) C_k| \leq \varepsilon \} \mid \mathcal{D}_n \right] + t^2 \sum_{k=1}^{n} E\left[g^2(\mathcal{D}_n) C_k^2 I\{ |g(\mathcal{D}_n) C_k| > \varepsilon \} \mid \mathcal{D}_n \right] \\
&\leq \frac{1}{6} \varepsilon^3 |t|^3 + t^2 g^2(\mathcal{D}_n) \sum_{k=1}^{n} E\left[C_k^2 I\{ |g(\mathcal{D}_n) C_k| > \varepsilon \} \mid \mathcal{D}_n \right].
\end{aligned}$$

Since $|g(\mathcal{D}_n) C_k| > \varepsilon$ implies $|g(\mathcal{D}_n) C_k / \varepsilon|^\delta > 1$, we have

$$\begin{aligned}
0 \leq g^2(\mathcal{D}_n) \sum_{k=1}^{n} E\left[C_k^2 I\{ |g(\mathcal{D}_n) C_k| > \varepsilon \} \mid \mathcal{D}_n \right] &\leq \frac{g^{2+\delta}(\mathcal{D}_n)}{\varepsilon^\delta} \sum_{k=1}^{n} E\left[|C_k|^{2+\delta} I\{ |g(\mathcal{D}_n) C_k| > \varepsilon \} \mid \mathcal{D}_n \right] \\
&\leq \frac{g^{2+\delta}(\mathcal{D}_n)}{\varepsilon^\delta} \sum_{k=1}^{n} E\left(|C_k|^{2+\delta} \mid \mathcal{D}_n \right),
\end{aligned}$$

where the rightmost expression converges to 0 in probability, since $\sum_{k=1}^{n} E(|C_k|^{2+\delta} \mid \mathcal{D}_n) \to 0$ in probability and $g^2(\mathcal{D}_n) \leq M + o_p(1)$. Moreover,

$$g^2(\mathcal{D}_n) \sum_{k=1}^{n} E\left[C_k^2 I\{ |g(\mathcal{D}_n) C_k| > \varepsilon \} \mid \mathcal{D}_n \right] \leq g^2(\mathcal{D}_n),$$

where $E g^2(\mathcal{D}_n) \leq M$. By dominated convergence theorem, with Skorohod Representation Theorem to extend its result to weakly convergent sequence of random variables, we have $E(g^2(\mathcal{D}_n) \sum_{k=1}^{n} E[C_k^2 I\{ |g(\mathcal{D}_n) C_k| > \varepsilon \} \mid \mathcal{D}_n]) \to 0$. As $\varepsilon > 0$ is arbitrary,

$$E \sum_{k=1}^{n} \left| E\left[\exp\left\{ itg(\mathcal{D}_n) C_k \right\} \mid \mathcal{D}_n \right] - 1 + \frac{1}{2} t^2 g^2(\mathcal{D}_n) \sigma_{C,k}^2(\mathcal{D}_n) \right| \to 0. \tag{S10}$$

Similarly, we have

$$\begin{aligned}
&\sum_{k=1}^{n}\left|E\left[\exp\{itg(\mathcal{D}_n)\sigma_{C,k}(\mathcal{D}_n)G_k\}\mid\mathcal{D}_n\right]-1+\frac{1}{2}t^2g^2(\mathcal{D}_n)\sigma_{C,k}^2(\mathcal{D}_n)\right| \\
&\le\frac{1}{6}\varepsilon^3|t|^3+t^2g^2(\mathcal{D}_n)\sum_{k=1}^{n}\sigma_{C,k}^2(\mathcal{D}_n)E\Big[G_k^2I\{|g(\mathcal{D}_n)\sigma_{C,k}(\mathcal{D}_n)G_k|>\varepsilon\}\mid\mathcal{D}_n\Big] \\
&\le\frac{1}{6}\varepsilon^3|t|^3+t^2\frac{g^{2+\delta}(\mathcal{D}_n)}{\varepsilon^\delta}E\left(|G_1|^{2+\delta}\right)\sum_{k=1}^{n}\sigma_{C,k}^{2+\delta}(\mathcal{D}_n) \\
&\le\frac{1}{6}\varepsilon^3|t|^3+t^2\frac{g^{2+\delta}(\mathcal{D}_n)}{\varepsilon^\delta}E\left(|G_1|^{2+\delta}\right)\sum_{k=1}^{n}E\left(|C_k|^{2+\delta}\mid\mathcal{D}_n\right)
\end{aligned}$$

where the last equality is due to Jensen's inequality as $(2+\delta)/2>1$. As G_1 is a standard normal random variable, $E|G_1|^{2+\delta}=\Gamma\{(3+\delta)/2\}\pi^{-1/2}$ where Γ is the Gamma function. Therefore $E|G_1|^{2+\delta}<\infty$. Hence, by a similar argument using dominated convergence theorem, we conclude

$$E\sum_{k=1}^{n}\left|E\left[\exp\{itg(\mathcal{D}_n)\sigma_{C,k}(\mathcal{D}_n)G_k\}\mid\mathcal{D}_n\right]-1+\frac{1}{2}t^2g^2(\mathcal{D}_n)\sigma_{C,k}^2(\mathcal{D}_n)\right|\to 0.$$

Combining with (S10), $|E(Q_1)|\to 0$. Similar but simpler argument can be used to control (S9) and conclude that $|E(Q_2)|\to 0$. As a result,

$$\left|E\exp\left[it\left\{\tau\sum_{j=1}^{n}A_j+g(\mathcal{D}_n)\sum_{j=1}^{n}C_j\right\}\right]-E\exp\left[it\left\{\tau\sum_{j=1}^{n}n^{-1/2}F_j+g(\mathcal{D}_n)\sum_{j=1}^{n}\sigma_{C,j}(\mathcal{D}_n)G_j\right\}\right]\right|\to 0,$$

for every t. Write $F=\sum_{j=1}^{n}n^{-1/2}F_j$ which is a standard normal random variable. Note that $E\{\tau F+g(\mathcal{D}_n)\sum_{j=1}^{n}\sigma_{C,j}(\mathcal{D}_n)G_j\}^2=\tau^2+E\{g^2(\mathcal{D}_n)\}\le\tau^2+M$, and therefore the second moment exists. Hence the characteristic function of $\tau F+g(\mathcal{D}_n)\sum_{j=1}^{n}\sigma_{C,j}(\mathcal{D}_n)G_j$ is at twice differentiable. Hence we obtain the desired result. Note that F is independent of both $G_1,\ldots,G_n$ and $\mathcal{D}_n$. Hence we obtain the desired result. □

References

Birman, M. S. & Solomyak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes w_p^α. *Matematicheskii Sbornik* **115**, 331–355.

Dvoretzky, A. (1972). Asymptotic normality for sums of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Probab* **2**, eds. L. M. Le Cam, J. Neyman and E. L. Scott, pp. 513–35. Berkeley: University of California Press.

Lin, Y. (2000). Tensor product space ANOVA models. *Ann. Statist.* **28**, 734–755.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*. Cambridge: Cambridge University Press.

[*Received x x. Revised x x*]