Nonlinear predictive directions in clinical trials

Youngjoo Cho[a,*,1], Xiang Zhan[b,**,2] and Debashis Ghosh[c]

[a]Department of Applied Statistics, Konkuk University, Seoul 05029, Republic of Korea
[b]Department of Biostatistics, School of Public Health and Beijing International Center for Mathematical Research, Peking University, Beijing 100191, China
[c]Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO 80045, USA

ARTICLE INFO

Keywords:
Causal effect
Heterogeneity of treatment effect
Machine learning
Kernel methods
Personalized medicine

ABSTRACT

In many clinical trials, individuals in different subgroups may experience differential treatment effects. This leads to the need to consider individualized differences in treatment benefit. The general concept of predictive directions, which are risk scores motivated by potential outcomes considerations, is introduced. These techniques borrow heavily from the literature from sufficient dimension reduction (SDR) and causal inference. Initially directions assuming an idealized complete data structure are formulated. Then a new connection between SDR and kernel machine methodology for detection of treatment-covariate interactions is developed. Simulation studies and a real data analysis from AIDS Clinical Trials Group (ACTG) 175 data show the utility of the proposed approach.

1. Introduction

In many clinical trials, the average treatment effect is the primary interest. After finding the effect, one of the researchers' interests would be in understanding how covariates affect the treatment effect. Developing methods for identification of appropriate patient subgroups for which the treatment might be of major benefit has become a topic of intense interest in the statistical literature. Gail and Simon (1985) introduced methods for identification of qualitative treatment covariate interactions. The Subpopulation Treatment Effect Pattern Plot (STEPP) was developed by Cai, Tian, Wong and Wei (2011a) as a graphical summary for subgroup identification with attendant permutation testing procedures. Using a working model and training/test set paradigm, Cai et al. (2011a) developed a modelling strategy to identify subgroups of patients who would benefit from the treatment; we comment on their approach in §3.2. Tree-based and related machine learning approaches (e.g., Kehl and Ulm (2006); Su, Zhou, Yan, Fan and Yang (2008); Su, Tsai, Wang, Nickerson and Li (2009); Foster, Taylor and Ruberg (2011); Imai, Ratkovic et al. (2013); Wager and Athey (2018)) for finding treatment subgroups have also been proposed. Much of these methodologies have been focused on the issue of identification of subgroups at a subpopulation level, where the subgroups are defined based on covariates that have interactions with treatment. VanderWeele, Luedtke, van der Laan and Kessler (2019) take this notion to a person-specific level and described four problems in personalized medicine. They show that for each question, the optimal rule has a form that takes the difference in individual-specific responses conditional on covariates. They use the potential outcomes framework (Rubin, 1974; Holland, 1986) to derive these results. An important takeaway from their work is the necessity of moving away from testing individual treatment-covariate interactions towards holistic testing of multiple interactions simultaneously.

In this work, inspired by ideas from causal inference and its links with sufficient dimension reduction (SDR) methods (Ghosh, 2011; Luo, Zhu and Ghosh, 2017), we develop a concept termed the predictive direction. The idea is to posit potential outcomes for the subject under each of the possible treatments and to then model their difference. In the hypothetical case where the complete potential outcomes are available, we can then exploit sufficient dimension reduction methods in order to estimate the predictive direction.

While we describe the predictive directions concept within the potential outcomes framework in Section 2, for most situations, there are two problems. First, the counterfactuals are never simultaneously observed. Second, the classical

*Corresponding author
**Co-Corresponding author
yvc5154@konkuk.ac.kr (Y. Cho); zhanx@bjmu.edu.cn (X. Zhan)
ORCID(s): 0000-0001-5667-5654 (Y. Cho); 0000-0001-9650-143X (X. Zhan)
[1]120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Republic of Korea, Tel : +82-2-450-3558, Fax : +82-2-450-4084
[2]38 Xueyuan Rd, Haidian District, Beijing 100191, China, Tel: +86 10-82802690

sufficient dimension reduction framework requires a linearity condition that might not be feasible in most applications. To deal with the former issue, we impute the outcomes using random forests (Breiman, 2001), a step that was also applied in the 'virtual twins' method of Foster et al. (2011). For the second issue, we develop a new link between sufficient dimension reduction methodology and kernel machine methods (Liu, Lin and Ghosh, 2007). This kernel machine method has an advantage over the virtual twins method by not depending on this linearity and has flexibility of accommodating nonlinear interaction. In addition, we are able to prove some technical results about our procedure in Theorem 1 in the paper.

The structure of the paper is as follows. In Section 2, we outline the background material on the potential outcomes framework as well as computation of the predictive direction using SDR methodology. Section 3 describes a general methodology to address the latter issue from Section 2. Section 4 describes a new nonlinear extension of the approach to relax the linearity assumption and yields approximations using kernel machine methods (Liu et al., 2007). Section 5 describes simulation studies to evaluate the finite-sample properties of our methodology. We apply our methodology in Section 6 to data from AIDS Clinical Trial Group (ACTG) 175. Some discussion concludes Section 7.

2. Potential outcomes framework and applications to risk modelling

We work within the potential outcomes framework of Rubin (1974) and Holland (1986). Assume that for $i = 1, \ldots, n$,

$$\{Y_i(0), Y_i(1), T_i, \mathbf{Z}_i\},$$

is a random sample from the triple $(Y(0), Y(1), T, \mathbf{Z})$, where $(Y(0), Y(1))$ represents the counterfactuals, T denotes the treatment group, and Z, a p-dimensional vector of covariates, is observed for all subjects. Let T take the values $\{0, 1\}$ so that the treatment is binary. Note that we are merely using the setup to be able to define the predictive directions. Also, we will be working within the context of a clinical trial where T will be randomized so that it can be assumed to be independent of $\mathbf{Z}$. As described in Rosenbaum and Rubin (1983), the standard assumption needed for causal inference is that

$$T \perp\!\!\!\perp \{Y(0), Y(1)\}|\mathbf{Z}, \tag{1}$$

i.e. treatment assignment is conditionally independent of the set of potential outcomes given covariates. Rosenbaum and Rubin (1983) refer to (1) as the strongly ignorable treatment assumption; it allows for the estimation of causal effects. Since we have randomized clinical trials, this strongly ignorable assumption holds.

We now exploit the work of Ghosh (2011) and impose further conditional independence assumptions from the sufficient dimension reduction literature Cook (2009). Assume that there exists a $p \times q$ matrix $\mathbf{A}$, $q \leq p$, such that treatment is conditionally independent of $\mathbf{Z}$, given $\mathbf{A}'\mathbf{Z}$. This can be expressed as

$$T \perp\!\!\!\perp \mathbf{Z}|\mathbf{A}'\mathbf{Z}. \tag{2}$$

Assumption (2) is a crucial one for defining the estimand targetted by most sufficient dimension reduction methods. In particular, if $S(\mathbf{A})$ represents the subspace generated by the columns of $\mathbf{A}$, then the smallest subspace containing all possible spaces is known as the central subspace (Cook, 2009). It most problems, the central subspace typically exists under some mild assumptions. Combining assumptions (2) and (1), we have

$$T \perp\!\!\!\perp \{Y(0), Y(1)\}|\mathbf{A}'\mathbf{Z}, \tag{3}$$

so that the columns of $\mathbf{A}$ capture the essential information about the potential outcomes. These columns are what we term the *directions* in the outcome data. Note that (3) implies that

$$T \perp\!\!\!\perp g(\{Y(0), Y(1)\})|\mathbf{A}'\mathbf{Z}$$

for any function $g(y, z)$ whose domain is R^2 and whose range is R. Next, we define the function

$$g(y, z) = y - z.$$

Of course, many other functions are possible, but in the current article, we focus on this choice of g. We then define the columns of $\mathbf{A}$ corresponding to g as the *predictive directions*.

3. Ideal algorithm and limitations

As noted by Ghosh (2011), with the sequence of conditional assumptions being invoked in §2., one can then employ sufficient dimension reduction procedures in order to compute the predictive directions. The proposed algorithm is as follows:

A. Compute $Y_i^* \equiv g\{Y_i(1), Y_i(0)\}$ for subject i, $i = 1, \ldots, n$.

B. Perform sufficient dimension reduction of Y_i^* on $\mathbf{Z}_i$ $(i = 1, \ldots, n)$ in order to estimate the directions (i.e., the columns of $\mathbf{A}$).

As pointed out before, in practice, we cannot implement the high-level algorithm in the previous paragraph due to the inability to observe both potential outcomes. Instead of $\{Y_i(0), Y_i(1)\}$, we observe $Y_i = T_i Y_i(1) + (1 - T_i)Y_i(0)$. We thus modify the algorithm by including an imputation step before step A.

We make several remarks about this algorithm. First, since the data come from a randomized clinical trial, separate prediction within treatment arms is a valid approach for imputing potential outcomes. Second, the approach is agnostic to the choice of imputation algorithm; one could use other alternatives (e.g., Raghunathan, Lepkowski, Van Hoewyk, Solenberger et al. (2001); Van Buuren (2018)). Third, the imputation step corresponds to that needed in algorithms such as the 'virtual twins' algorithm of Foster et al. (2011); however, their subsequent steps are different from ours.

For the choice of sufficient dimension reduction procedure, one can consider sliced inverse regression (SIR) (Li, 1991). Alternative methods could also be used, such as SAVE (Cook and Weisberg, 1991) and MAVE (Xia, Tong, Li and Zhu, 2002). However, SIR requires the linearity condition for its validity. For a $p-$dimensional random vector $\mathbf{x}$, the linearity condition assumes that

$$E(\mathbf{x}|\boldsymbol{\beta}'\mathbf{x}) = \mathbf{P}\mathbf{x},$$

where $\mathbf{P} \equiv \boldsymbol{\Sigma}\boldsymbol{\beta}(\boldsymbol{\beta}'\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1}\boldsymbol{\beta}'$ is a $p \times p$ matrix and $\boldsymbol{\Sigma}$ is the covariance matrix of $\mathbf{x}$. The linearity condition is viewed as restrictive, as it is mainly satisfied by elliptically symmetric distributions. The $p-$dimensional random vector $\mathbf{x}$ is elliptically symmetric distributed if and only if there exists a p-dimensional vector $\boldsymbol{\mu}$, a $p \times r$ matrix $\mathbf{B}$ with maximal rank r, and a nonnegative random variable V, such that

$$\mathbf{x} = \boldsymbol{\mu} + V\mathbf{B}\mathbf{u},$$

in distribution, where the $r-$dimensional random vector $\mathbf{u}$ is independent of V and is uniformly distributed on the $r-$dimensional unit sphere (Paindaveine, 2014).

4. Methodology

4.1. SDR, metric spaces and a new link

To overcome this restriction of linear SDR, we consider nonlinear SDR. An alternative for SIR has been kernel-based SDR approaches. Such approaches can be found in Ferré and Yao (2003), Fukumizu, Bach and Jordan (2004); Fukumizu, Bach, Jordan et al. (2009), Wu (2008), Wu, Liang and Mukherjee (2013), Li, Artemiou and Li (2011) and Lee, Li, Chiaromonte et al. (2013). The approaches of Li et al. (2011) and Lee et al. (2013) completely avoid the linearity condition by formulating a more general notion of sufficiency and develop estimation procedures based on constructional of kernel matrices and eigenvalue/eigenvector decomposition to estimate fitted functions.

While the starting points of SDR and kernel methods appear different at first glance, we show how the two are connected in this section. To prove the main result in this section, we will require the definitions of positive definite and completely monotone functions.

Definition 1. A real-valued function f is said to be positive definite if for any set of real numbers $x_1, \ldots, x_n$, the $n \times n$ matrix A with (i, j)th entry $a_{ij} = f(x_i - x_j)$ $(i = 1, \ldots, n; j = 1, \ldots, n)$ is positive definite.

Definition 2. A real-valued function f is said to be completely monotone if for all $r \in \{0, 1, 2, \ldots\}$,

$$(-1)^r f^{(r)}(x) \geq 0,$$

where $f^{(r)}$ denotes the $r-$th derivative of f.

A function $f(t)$ ($t \in R$) is positive definite if and only if $f(t) = g(t^2)$, where g is completely monotone. The other key fact is that any positive definite function will define a kernel (Aronszajn, 1950). Thus, for any positive definite function f, we have that $K(Z, \tilde{Z}) = f(\|Z - \tilde{Z}\|)$ is a proper kernel. As in Schoenberg (1938), we will study spaces of positive definite functions that are defined on proper metric spaces. The space R^p with the Euclidean distance can also be viewed as a metric space. Let $\mathcal{B}(E)$ denote the space of positive definite functions for a metric space E. One result of Schoenberg (1938) was that if E_1 and E_2 are metric spaces with $E_1 \subset E_2$, then $\mathcal{B}(E_1) \supset \mathcal{B}(E_2)$. If we take E_1 to be the restriction of R^p to random vectors Z that satisfy the linearity condition and E_2 to be random vectors which are elliptically symmetric, then we have $\mathcal{B}(E_1) \supset \mathcal{B}(E_2)$. For $\mathcal{B}(E_2)$, we have the following characterization from Schoenberg (1938):

Lemma 1. *A $p-$dimensional random vector $\mathbf{W}$ is elliptically symmetric if and only if its characteristic function can be written as $\psi(\|\mathbf{w}\|^2)$, where $\mathbf{w} \in R^p$ and $\psi(t)$ has the form*

$$\psi(t) = \int_0^\infty \omega_p(r^2 t) dF(r), \tag{4}$$

where ω_p is the characteristic function for a $p-$dimensional random vector that is distributed uniformly on the unit sphere in R^p, and $F(r)$ is a distribution function on $[0, \infty)$. We note that the form of $\omega_p(t)$ is given by

$$\omega_p(t) = \Gamma\left(\frac{p}{2}\right)\left(\frac{2}{t}\right)^{(p-2)/2} J_{(p-2)/2}(t),$$

where $\Gamma(a) \equiv \int_0^\infty u^{a-1}\exp(-u)du$ denotes the Gamma function and

$$J_\alpha(x) \equiv \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\Gamma(m+\alpha+1)}\left(\frac{x}{2}\right)^{2m+\alpha},$$

represents the Bessel function.

Given the definitions of E_1 and E_2 above, we define a sequence of metric spaces in the following way: let E_{2+i} be a metric space consisting of elliptically symmetric random vectors in R^{p+i} for $i = 1, 2, \ldots,$. We have that elliptical symmetry in higher dimensions implies elliptical symmetry in lower dimensions. This yields the following chain of inclusion relations:

$$\mathcal{B}(E_1) \supset \mathcal{B}(E_2) \supset \mathcal{B}(E_3) \supset \cdots \supset \mathcal{B}(E_\infty). \tag{5}$$

In addition, Schoenberg (1938) provides a characterization of $\mathcal{B}(E_\infty)$ in (5), which is given in the following result:

Lemma 2. *A random element $\mathcal{W}$ exists in E_∞ if and only if $\mathcal{W}$'s characteristic function can be written as $\psi(\|w\|^2)$, where $\psi(t)$ has the form*

$$\psi(t) = \int_0^\infty \exp(-r^2 t) dF(r), \quad t > 0, \tag{6}$$

and $F(r)$ is a distribution function on $[0, \infty)$.

Note that by the nested structure of the space of positive definite functions in (5), it is also the case that

$$\mathcal{B}(E_\infty) = \cap_{i=1}^\infty \mathcal{B}(E_i).$$

Thus, $\mathcal{B}(E_\infty)$ is the smallest space containing $\mathcal{B}(E_i)$ for all i. In this sense, the object $\mathcal{B}(E_\infty)$ can be interpreted as an infinite-dimensional analog to the central subspace that was described in §2. Combining all the results above leads us to the following result.

Proposition. A random element exists in $\mathcal{B}(E_\infty)$ if and only if its associated kernel is of the form

$$K(X, \tilde{X}) = \psi(\|X - \tilde{X}\|), \tag{7}$$

Table 1
Examples of kernels that are members of $\mathcal{B}(E_\infty)$. Here K_ν denotes the modified Bessel function of the second kind of order ν.

Kernel	K$(z, \tilde{z})$	Parameter ranges
Gaussian	$\exp\{-\|\|z-\tilde{z}\|\|^2/\rho\}$	$\rho > 0$
Matérn	$\frac{2^{\nu-1}}{\Gamma(\nu)}\left(\frac{\|\|z-\tilde{z}\|\|}{c}\right)^{\nu} K_\nu\left(\frac{\|\|z-\tilde{z}\|\|}{c}\right)$	$c, \nu > 0$
Generalized Cauchy	$\left[1+\left(\frac{\|\|z-\tilde{z}\|\|}{c}\right)^{\alpha}\right]^{-\tau/\alpha}$	$c, \tau > 0, 0 < \alpha \leq 2$
Dagum		
Powered Exponential	$\exp\{-\left(\frac{\|\|z-\tilde{z}\|\|}{c}\right)^{\alpha}\}$	$c > 0, 0 < \alpha \leq 2$

where ψ is generated via (6). The proposition shows that the kernels in $\mathcal{B}(E_\infty)$ only depend on the interpoint distances between points.

We note that we arrive at kernels as in Lee et al. (2013) but with a very different starting point and a different set of assumptions. We do so through an alternative construction that did not rely on the generalized notions of sufficiency that are considered by Lee et al. (2013). The nesting function space argument in this paper allows one to transition from distributional assumptions (e.g., elliptical distribution for X) to functional definitions that can be characterized using kernels. We also note that because we are modelling the difference in mean potential outcomes using kernel machines, we are less reliant on the central subspace object in sufficient dimension reduction and could instead have started with the central mean subspace (Cook and Li, 2002) instead. This is what was used in Luo et al. (2017).

We recall an earlier example from the SDR literature that violates the linearity condition. The example is the regression model

$$Y = (\boldsymbol{\beta}'\mathbf{X})^2 + \epsilon,$$

where $\mathbf{X}$ and ϵ have normal distributions. A method such as SIR will estimate the direction to be zero. Using the theoretical framework that is presented here, we would see that this regression relationship would not exist in $\mathcal{B}(E_\infty)$. Formally, the regression model would correspond to a kernel of the form $K(X, \tilde{X}) = (< X, \tilde{X} > +1)$ which has been referred to as the polynomial kernel of order one in the machine learning literature. By the proposition, such a kernel does not have the form (7) so that it would not be in $\mathcal{B}(E_\infty)$. Thus, the theorem provides new insights as to situations in which SDR methodologies will fail to capture the correct directions.

Each element of $\mathcal{B}(E_\infty)$ will have a unique kernel associated with it and vice versa. One example of a kernel that would exist in $\mathcal{B}(E_\infty)$ is the Gaussian Kernel, whose kernel is given by

$$K(z, \tilde{z}) = \exp\{-\|z - \tilde{z}\|^2/\rho\},$$

where $\|z - \tilde{z}\|^2 = \sum_{k=1}^{p}(z_k - \tilde{z}_k)^2$ and $\rho > 0$ represents a scale parameter. The Gaussian kernel generates the function space spanned by radial basis functions, a complete overview for which can be found in Bühmann (2003). Other examples of kernels that reside in $\mathcal{B}(E_\infty)$ can be found in Table 1.

Remark 1. While we have defined the predictive direction in a linear way in Section 2, the development here allows us for one to define a nonlinear predictive direction. In particular, it will be an element $\tilde{b} \in \mathcal{B}(E_\infty)$ such that T is independent of $\{Y(0), Y(1)\}$ given the $\sigma-$algebra generated by $\tilde{b}$. This is in spirit to a definition of nonlinear sufficient dimension reduction given in Lee et al. (2013).

4.2. Proposed Algorithm and some theoretical guarantees

The results in the previous section lead to a modification of the algorithm in Section 3. It now proceeds as follows:

1. Fit random forests for Y_i as a function of $T_i, \mathbf{Z}_i$, and $T_i\mathbf{Z}_i$, $i = 1, \ldots, n$. Such an algorithm will allow for computation of $(\hat{Y}_i(1), \hat{Y}_i(0))$ based on the observed covariates $\mathbf{Z}_i$, $i = 1, \ldots, n$.
2. Compute the variable $\tilde{Y}_i = g\{\hat{Y}_i(1), \hat{Y}_i(0)\}$ for subject i, $i = 1, \ldots, n$.
3. Divide the dataset into training data $\{(\tilde{Y}_i, \mathbf{Z}_i)\}_{i=1}^{n_1}$ and test data $\{(\tilde{Y}_i, \mathbf{Z}_i)\}_{i=n_1+1}^{n_1+n_2}$ where n_1 and n_2 are number of observations in the training and test datasets, respectively. Let $n = n_1 + n_2$. Fit a kernel machine regression model of $\tilde{Y}_i$ on $\mathbf{Z}_i$, $i = 1, \ldots, n_1$ using the training data.

4. Predict the outcome based on the fitted kernel machine regression model with the test data $\mathbf{Z}_i$, $i = n_1+1, \ldots, n_1+n_2$ and obtain predictions for $\tilde{Y}_{n_1+1}, \ldots \tilde{Y}_{n_1+n_2}$.

One then gets fitted values from the kernel machine model applied to the input covariate vectors, and these can be treated as functionals of nonlinear extensions of the predictive directions defined in §2.1. Note that the third step amounts to fitting a support vector regression model (Cristianini, Shawe-Taylor et al., 2000). More details of kernel machine regression are provided in Appendix A of in this paper.

Without loss of generality, we picked a Gaussian kernel provided in Table 1 to build our algorithm. The Gaussian kernel is one of the most widely used kernels in the literature mainly because the corresponding RKHS $\mathcal{H}_K$ is universal. This means that the function space $\mathcal{H}_K$ spanned by the Gaussian kernel is dense in $C(R^p)$, the collection of all continuous functions defined on R^p (Steinwart, 2001). Equivalently, for any arbitrary function $g \in C(R^p)$ and every $\epsilon > 0$, there exists always a function $f \in \mathcal{H}_K$ such that $sup_{\mathbf{z}\in R^p}|f(\mathbf{z}) - g(\mathbf{z})| < \epsilon$. The universality property states that there always exists a function in the Gaussian RKHS that is arbitrarily close to the true functional relation between potential outcomes and covariates as long as such a functional relationship is continuous. This is a much weaker assumption than that used for most parametric outcome regression models in the causal inference literature. This universal property of Gaussian RKHS is very appealing in our framework as we want our regression model to be correctly specified and a broad function space would gain some robustness against model misspecification, which is a crucial factor in causal inference.

To study the theoretical guarantees of estimator $\hat{h}(\cdot)$ of the kernel machine regression $y = \beta_0 + h(z) + \epsilon$, we first calculate the L^2-distance between the estimated function $\widehat{h}(\cdot)$ (explicit form given in Appendix A) and the true underlying $h(\cdot)$ using

$$||\widehat{h} - h||_{L^2(P_{\mathbf{Z}})} := \left[\int_{R^p} |\widehat{h}(\mathbf{z}) - h(\mathbf{z})|^2 d\, P_{\mathbf{Z}}(\mathbf{z})\right]^{1/2},$$

where $P_{\mathbf{Z}}$ is a probability distribution for covariate vector $\mathbf{Z}$. This metric has been widely used in the literature of causal inference (Hill, 2011; Alaa and Schaar, 2018). The major result in this paper on the kernel machine regression estimator is presented in the following theorem:

Theorem 1: Let $(Y_i, \mathbf{Z}_i)$, $i = 1, \ldots, n$ be i.i.d. samples randomly drawn from the joint distribution $P_{Y\mathbf{Z}}$ satisfying $|Y| \leq M$ almost surely and $P_{\mathbf{Z}}$ be the marginal distribution of $\mathbf{Z}$, whose domain is compact. Assume the true regression surface of regression model $h(\cdot) = E[Y|\mathbf{Z} = \cdot] \in \mathcal{H}_K$, the RKHS spanned by the Gaussian kernel. Then, for any $0 < \delta < 1$, the following error bound holds with probability $1 - \delta$:

$$||\widehat{h} - h||_{L^2(P_{\mathbf{Z}})} \leq C n^{-\frac{1}{4}}, \tag{8}$$

where $C = \log(4/\delta)(8M + \sqrt{8}||h||_{\mathcal{H}_K})$ is a function that does not depend on n.

The crucial factor in the error bound of Theorem 1 is $||h||_{\mathcal{H}_K}$, which is finite by the assumption that true regression surface h lies in the RKHS $\mathcal{H}_K$. This assumption is relatively weak given the fact that the Gaussian RKHS is dense in the collection of all continuous functions defined on R^p. The results in Theorem 1 is consistent with findings reported in the literature of causal inference that the learning rate of estimating causal effects is mainly determined by the more complex of the surfaces of outcome regression models (Alaa and Schaar, 2018). The proof of this theorem is provided in Appendix B in this paper. In our algorithm, the working outcomes $\tilde{Y}_i = \hat{Y}_i(1) - \hat{Y}_i(0)$ of the kernel machine regression model are calculated based on imputed values from random forests models. The following remark accounts for this fact.

Remark 2. Let $\widehat{h}_{\tilde{Y}}$ denote the kernel machine regression estimator calculated from the working outcomes $\tilde{Y}$. Then, under the assumptions of Theorem 1, we have

$$Pr\left[||\widehat{h}_{\tilde{Y}} - h||_{L^2(P_{\mathbf{Z}})} \leq C n^{-\frac{1}{4}} | \tilde{Y}\right] \geq 1 - \delta.$$

Finally, we note that a related approach to using kernel machines was taken in Shen and Cai (2016). While their approach shares similarities with the algorithm developed here, we note that the motivation and starting points are completely different. Furthermore, they were focused more on the issue of testing, while our goal here is that of computing and estimating of the nonlinear directions.

4.3. Treatment selection rules and evaluation of predictive directions

Based on our approach to predictive direction estimation, we can use the estimated directions to guide optimal treatment strategies inspired by VanderWeele et al. (2019). Our proxy rule is to use

$$D^{10} > k, \tag{9}$$

where D_{10} is the predictive direction-derived score.

To evaluate the predictive direction as a scoring rule, we need a training and testing set in which both studies are randomized and consist of the same treatments. In addition, outcome variables need to be measured in both studies. The proposal is related to one discussed in Vickers, Kattan and Sargent (2007). To simplify the discussion, we will deal with the case of two treatment groups. The procedure works as follows:

(a). Estimate the predictive direction using the training dataset.

(b). Using the estimated direction, compute scores for all subjects in the test set.

(c). Based on the scores, determine which treatment each subject should receive in the test set using treatment rules of the form (9).

(d). For the subjects whose predicted treatment match their randomized treatment in the test set, compare the outcomes between the two treatment groups.

We mention some points at this stage. First, we note that for step (b), the outcome information in the test set is not used at all. Only the covariate information is used to compute the scores. The outcome information is needed in step (d) in order to compute the measure of treatment effect between the two groups. Note also that the fact that the test set also comes from a clinical trial is a necessary feature here. In step (d), we will be excluding two types of subjects in the test set: those who were predicted to have greatest benefit from one treatment group but were observed to receive the other one. Thus, we are performing a subgroup analysis in step (d) based on subjects in the test set whose predicted and actual treatment assignments are concordant. The randomization of treatment is necessary in order to ensure that the subgroup analysis will also be the same as the overall treatment effect.

5. Simulation studies

5.1. Simulation I

We first conducted numerical studies to evaluate the theoretical guarantees stated in Theorem 1 on the kernel-based estimation algorithm of predictive directions presented in §4.2. There are many aspects of the design that may be important to consider including the sample size, forms of function $h(\cdot)$ and the correlations between the $\mathbf{Z}$'s. Specifically, we considered $p = 10$ covariates with varied sample sizes $n = 100$, 200, 400 or 800. The p covariates $(Z_1, \ldots, Z_p)$ were simulated from two different distributions: 1) independently form uniform distribution between 0 and 1, that is, $Z_j \sim unif(0, 1)$, i.i.d.; 2) multivariate normal distribution $\mathbf{Z}_0 = (Z_{01}, \ldots, Z_{0p}) \sim N_p(\mathbf{0}, \mathbf{\Sigma}_r)$, where $\mathbf{\Sigma}_r$ was the compound symmetry covariance structure with all diagonal elements being 1 and off-diagonal elements being r. Here we used $r = 0$, 0.4 and 0.8 to represent low (or independence), moderate and high correlation level among covariates. Notice that Theorem 1 requires the domain of $\mathbf{Z}$ to be compact. Hence, we made further transformation from $\mathbf{Z}_0$ to $\mathbf{Z}$ by setting $Z_j = 1$ if $Z_{0j} > 1$, $Z_j = -1$ if $Z_{0j} < -1$, and $Z_j = Z_{0j}$ otherwise, $j = 1, \ldots, p$. For ease of presentation, we term this distribution of covariates as the truncated normal (TN) hereafter. After covariates were simulated, we next generated the difference in potential outcomes as

$$\tilde{Y}_i = 1 + h(Z_{i1}, \ldots, Z_{ip}) + \epsilon_i, \quad i = 1, \ldots, n,$$

where error ϵ_i was generated from the standard normal distribution $N(0, 1)$, and $h(\cdot) = h_1(\cdot)$ or $h_2(\cdot)$ given by:

$$h_1(Z_1, \ldots, Z_p) = 2cos(Z_1) - 3Z_2^2 + 2Z_4 e^{-Z_3} - 1.6sin(Z_5)cos(Z_3) + 4Z_1 Z_5,$$

and

$$h_2(Z_1, \ldots, Z_p) = 2Z_1 - 3Z_2 - Z_3 + 2Z_4 + 4Z_5.$$

Table 2
Mean squared error of kernel machine regression (KMR) and linear regression (LR). The results are averaged over 200 replicates.

		$h = h_1$				$h = h_2$			
P_Z	Method	$n = 100$	200	400	800	$n = 100$	200	400	800
Unif	KMR	0.241	0.167	0.116	0.072	0.127	0.065	0.037	0.022
	LR	1.048	1.119	1.155	1.167	0.898	0.949	0.972	0.981
TN0	KMR	0.428	0.302	0.190	0.115	0.263	0.172	0.110	0.063
	LR	6.278	6.583	6.931	6.998	0.901	0.944	0.979	0.986
TN4	KMR	0.393	0.311	0.217	0.133	0.272	0.177	0.114	0.072
	LR	7.205	8.563	8.847	9.100	0.910	0.942	0.965	0.981
TN8	KMR	0.290	0.245	0.184	0.121	0.216	0.151	0.097	0.062
	LR	5.222	6.311	6.943	7.112	0.901	0.947	0.974	0.989

It is of note that function $h(\cdot)$ is sometime referred to as the interaction term between treatment and covariates on the outcome in literature (Foster et al., 2011). Our simulation design allows for nonlinear interaction terms, which is a feature typically ignored or infeasible in many other existing methods. Finally, we used the simulated data $(\tilde{Y}_i, \mathbf{Z}_i), i = 1, \ldots, n$ to fit a kernel machine regression (KMR) model described in Appendix A. As a comparison, we included a linear regression (LR) model fitted to the same data $(\tilde{Y}_i, \mathbf{Z}_i), i = 1, \ldots, n$. To evaluate the criterion in Theorem 1, we calculated the mean squared errors (MSE) $\sum_{i=1}^{n}[\tilde{Y}_i - \hat{\tilde{Y}}_i]^2/n$ of two regression models and report their values across different simulation configurations in Table 2.

As can be seen in Table 2, under the distributions of the covariate (P_Z) and forms of the interactions (h_1/h_2), the mean squared error of kernel machine regression decays towards zero when the sample size increases, as guaranteed by Theorem 1. As an example, taking $\mathbf{P}_Z$ to correspond to the Unif(0,1) distribution and $h = h_1$ as an example, it can be easily calculated that the MSEs of KMR under different sample sizes satisfy the following relationship:

$$\frac{0.241}{0.167} \approx \frac{0.167}{0.116} \approx \frac{0.116}{0.072} \approx \sqrt{2}.$$

In other words, the MSE of KMR decays nearly at a rate that is around the square root of ratio of sample sizes, which is the rate guaranteed by Theorem 1. On the other hand, traditional linear regression does not enjoy such a property. Finally, when the linear model is true and the sample size goes to infinity, the MSE of linear regression will converge to the variance of error, which seems to be the case in Table 2.

5.2. Simulation II

In the second set of simulations, we conducted numerical experiments to evaluate the potential of using predictive directions as a guideline for optimal treatment selection or subgroup analysis as described in §4.3. We mainly followed the design of the original virtual twins-based subgroup identification paper (Foster et al., 2011) and generated randomized trials of $(Y_i, T_i, \mathbf{Z}_i)$ with $n = 400$, 800 or 1600 patients, where T was generated from Bernoulli (0.5) and $\mathbf{Z} = (Z_1, \ldots, Z_{10})$ were generated as independent $Z_j \sim Unif(0, 1)$. We considered the following model for outcome data generation

$$Y = -1 + 0.1T + 0.5Z_1 + 0.5Z_2 - 0.5Z_3 + 0.5Z_2Z_4 + Th(\mathbf{Z}) + \epsilon,$$

where function $h(\mathbf{Z})$ is given in Simulation I. Then, we split n data points into two parts: a training set $(Y_i^t, T_i^t, \mathbf{Z}_i^t), i = 1, \ldots, n_1$ and a test set $(Y_i, T_i, \mathbf{Z}_i), i = 1, \ldots, n_2$, where $n_1 = n_2 = n/2$. We applied the predictive directions estimation algorithm in §4.2 the training set $(Y_i^t, T_i^t, \mathbf{Z}_i^t), i = 1, \ldots, n_1$ to obtain $\hat{h}(\cdot)$, and then applied this estimated function to test set $\mathbf{Z}_i^t, i = 1, \ldots, n_2$ to obtain the predicted direction scores $D_i^{10} = \hat{\beta}_0 + \hat{h}(\mathbf{Z}_i), i = 1, \ldots, n_2$, which were used to perform the subgroup analysis. Our proxy rule is to define subgroup using $\{D^{10} > k\}$. Without specific context-dependent information, we selected the threshold k in a data-adaptive way by using quantiles of the predicted predicted direction scores $D_i^{10}, i = 1, \ldots, n_2$. Recall that predictive scores are estimations of differences in potential outcomes and a larger

Table 3
Estimation of enhanced treatment effects by KMR and VTR. The results are averaged over 200 replicates. A paired two-sample t-test was used to calculate the p-values. Larger values correspond to better performance.

	$h = h_1$			$h = h_2$		
n	KMR	VTR	p-value	KMR	VTR	p-value
400	0.799	0.755	7.3e-02	1.797	1.316	5.8e-44
800	0.816	0.782	4.9e-02	1.796	1.320	8.8e-60
1600	0.849	0.810	3.0e-04	1.811	1.289	5.0e-93

score indicates that the subject is more likely to benefit from treatment $T = 1$ than $T = 0$. Therefore, we assigned $\hat{T}_i = 1$ for subject i of the test set if its predicted direction score D_i^{10} is greater than the 66.6%-quantile of all predicted direction scores $D_i^{10}, i = 1, \ldots, n_2$, and $\hat{T}_i = 0$ if its score is lower than the 33.3%-quantile. Here, quantiles (33.3% and 66.6%) are arbitrarily picked to evaluate the potential usefulness of predictive directions in subgroup analysis. For the remaining one third of subjects, more information is needed to determine the appropriate treatment they should receive. Finally, following step (d) of the procedure described in §4.3, we selected subjects whose predicted treatment match their randomized treatment and compare the outcomes between two groups. That is,

$$Q = \left[E(Y|T = 1 \cap \hat{T} = 1) - E(Y|T = 0 \cap \hat{T} = 0)\right] - \left[E(Y|T = 1) - E(Y|T = 0)\right].$$

The difference Q is also referred to as the enhanced treatment effect (Foster et al., 2011). As pointed out by a reviewer, it is possible that other forms of Q statistics are appropriate for evaluation, such as replacing the second term with $E(Y|T = 0 \cap \hat{T} = 1)$. We stick with the current form to match with the procedure described in §4.3. Clearly, larger values of Q are more desirable if our predictive directions-guided treatment selection rule is to be useful.

A natural competitor of our predictive directions-based subgroup analysis rules is the original virtual twins method Foster et al. (2011), which uses the virtual twins in a different manner to estimate a region $\hat{A}$ of covariates such that a subject should receive a treatment if his/her covariates measurement belong to that region. Correspondingly, their enhanced treatment effect associated with the subgroup (or region) is defined as

$$Q = \left[E(Y|T = 1 \cap \mathbf{Z} \in \hat{A}) - E(Y|T = 0 \cap \mathbf{Z} \in \hat{A})\right] - \left[E(Y|T = 1) - E(Y|T = 0)\right].$$

In this Simulation II, we compared the enhanced treatment effect (i.e., Q value) of our KMR-based subgroup analysis and that of the original virtual twins method (Foster et al., 2011). We adapted the original virtual twins regression (VTR) method a little bit to handle continuous outcomes, and also used the threshold $c = E[Y^t|T = 1] - E[Y^t|T = 0]+0.05$ calculated from training set to determine $\hat{A}$ in VTR. When sample size is relatively small, $\{T = 1 \cap \mathbf{Z} \in \hat{A}\} = \emptyset$ or $\{T = 0 \cap \mathbf{Z} \in \hat{A}\} = \emptyset$ occasionally happened in VTR. If either case happened, we set the corresponding VTR Q value as zero.

The enhanced treatment effects of KMR and VTR are reported in Table 3. On the basis of the table, the enhanced treatment effect of KMR is consistently better than that of VTR across all scenarios being evaluated. Such an improvement is significant under the nominal level $\alpha = 0.05$ (except for the scenario with $n = 400$ and $h = h_1$). It seems that the predictive directions is a better subgroup analysis tools when the interaction function $h = h_2$ compared to h_1. One possible reason is because the estimation error is controlled by the complexity of the interaction function (i.e., term $||h||_{\mathcal{H}_K}$ in the constant C in Theorem 1). The function h_2 is simpler than h_1 and hence the estimation of predictive directions of h_2 is more accurate leading to more powerful subgroup identification analysis in the sense of larger enhanced treatment effects.

6. Data analysis

To illustrate the methods, we consider data from the AIDS Clinical Trial Group (ACTG) 175 study (Hammer, Katzenstein, Hughes, Gundacker, Schooley, Haubrich, Henry, Lederman, Phair, Niu et al., 1996). This dataset was analyzed in Wang, Zhou, Song and Sherwood (2018) and Cho and Ghosh (2021). The response variable is CD4 cell count from 200-500 per cubic millimeter from patients with human immunodeficiency virus type 1 (HIV-1). By

removing ineligible patients, we obtain 2467 patients. We use the following 13 covariates in the analysis: age, existence of hemophilia, Karnofsky score, sex, days of antiretroviral therapy before ACTG 175, dichotomized race (white vs non-white), homosexual activity, history of intravenous drug use, symptomatic/asymptomatic status, antiretroviral therapy history, weight, zidovudine (ZDV) use prior to ACTG 175 (yes/no), Non-ZDV antiretroviral therapy prior to ACTG 175 (yes/no). the original treatment has four levels : Zidovudine, Zidovudine & Didanosine, Zidovudine & Zalcitabine, Didanosine, but we dichotomize our treatment variable into Zidovudine only versus the others (Cho and Ghosh, 2021). We also standardized these 13 variables. Note that due to the presence of discrete variables, the linearity condition in Section 3 is violated, necessitating the new methods developed in the paper.

We use random forests to impute potential outcomes. As in the simulation section, we used the mean squared error to evaluate performance of our proposed estimator and the virtual twins method. The mean squared error of our RKHS method is 0.178 and that using linear regression is 0.266.

Next, we divided the data into training (1233 observations) and test datasets (1234 observations). We fit both RKHS and linear regression on the training data and computed mean square error with the imputed test dataset. For mimicking the simulation study, we repeat this process 100 times. Figure 1 shows the boxplot of RKHS and linear

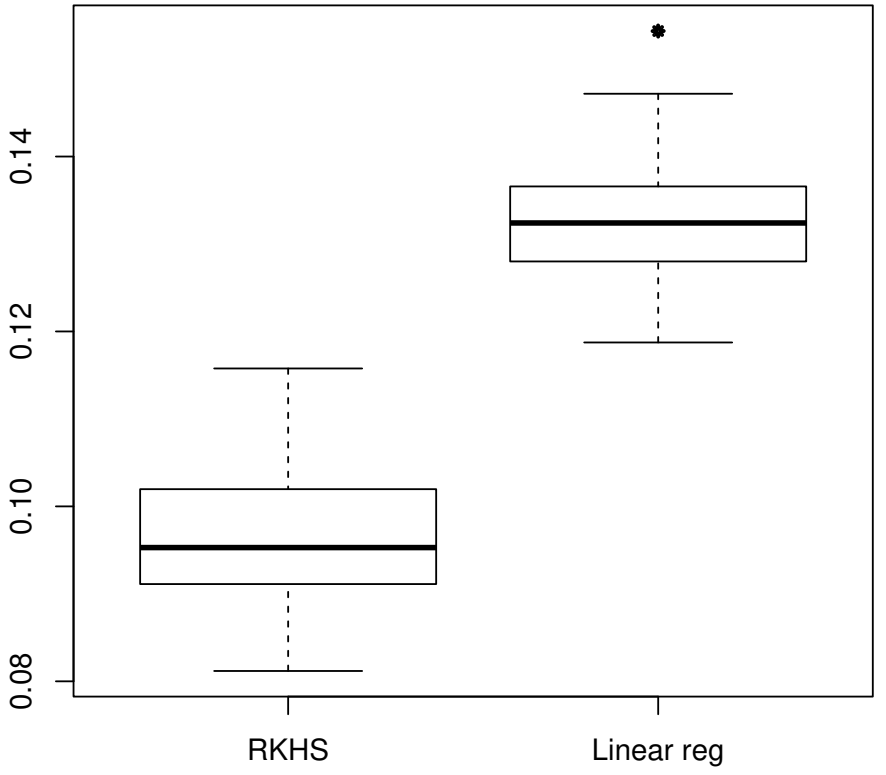

Figure 1: MSE comparison with 100 imputation of ACTG 175 dataset.

regression. Clearly, our approach is better than linear regression.

We also calculated Q (defined in Section 5) to compare the performance between our proposed approach and virtual twins for several cutoff values. We use the same threshold (33.3% and 66.6%) as simulation section for assignment of treatment. We repeated this process 100 times. Note that as in the simulation studies, $\{T = 1 \cap \mathbf{Z} \in \hat{A}\} = \emptyset$ or $\{T = 0 \cap \mathbf{Z} \in \hat{A}\} = \emptyset$ occasionally happened in VTR. If either case happened, we set the corresponding VTR Q value as zero. We examine various c values to see change of number of these empty sets. Figure 2 shows the boxplot of our proposed method and VTR for various values of c.

Table 4shows that our proposed method yields a higher Q-value than VTR at any cutoff level. The enhanced treatment effect of KMR is better than that of VTR as well. The p-value associated with Figure 2 for comparing the difference in metrics is less than 0.005, which shows statistically significant evidence of a benefit of our procedure at a 5% level of significance.

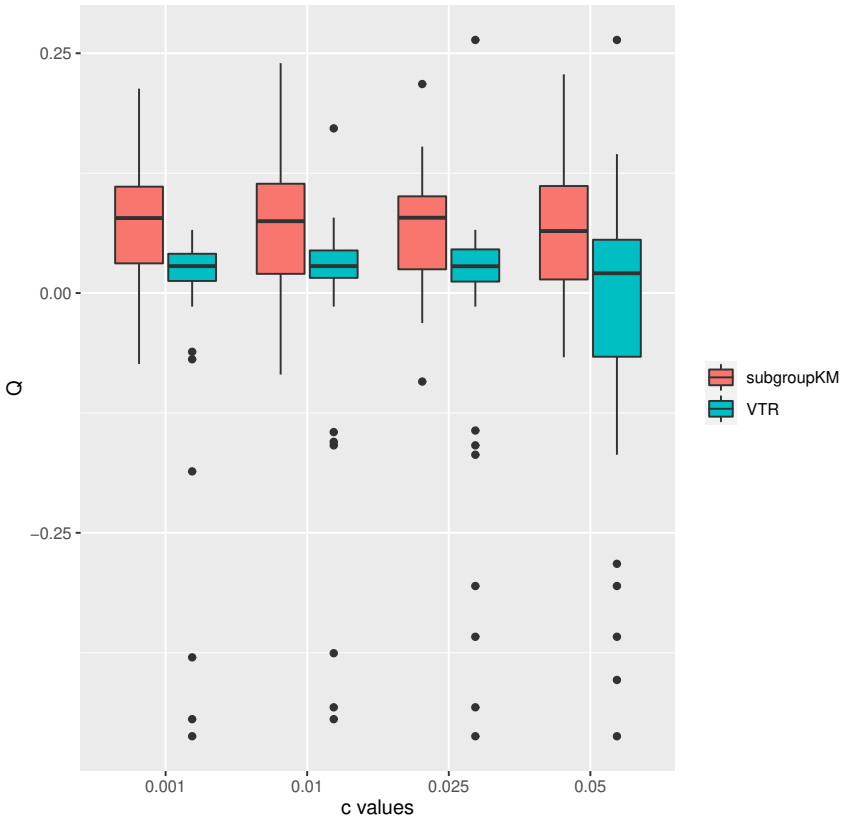

Figure 2: Enhanced treatment comparison with 100 imputation of ACTG 175 dataset. Higher values denote better performance.

c value	test statistic	p-value	number of VTR Q values being 0
$\delta + 0.05$	0.097	0.001	69
$\delta + 0.025$	0.074	0.0001	46
$\delta + 0.01$	0.064	2.363e-05	29
$\delta + 0.001$	0.064	5.714e-07	23

Table 4
Test statistic for paired t-test, p-values and number of VTR Q values being 0 with respect to various c values

7. Discussion

In this article, we have developed the concept of predictive directions for identification of person-specific effects in clinical trials. In a linear, idealized case, we are able to obtain linear combinations of the covariates that have a risk score interpretation. However, the validity of predictive directions requires strong distributional assumptions, so we proposed a novel nonlinear extension based on the connection between nonlinear SDR and kernel machine. We also showed that our approach is advantageous over traditional subgroup approach.

There are several potential extensions of this work that are currently under investigation. First, the issue of dimension estimation and subsequent post-model selection inference has not been addressed. In the current paper, we have bypassed the issue by fixing the dimension to be one. In the situation where there are multiple directions (i.e., multiple columns of $\mathbf{A}$ in (3)), a natural question arises as to how to use them to inform selection of optimal treatment as discussed in §4. Second, the aforementioned advantages of nonlinear predictive direction extension come at a price. The computational complexity of our method is $O(n^3)$, which mainly depends on calculating the inversion of a $n \times n$ kernel matrix. Taking the real data application in Section 6 as an example, the average computing

time handling around 2500 samples in Figure 1 is about 5 hours and the average computation handling around 1250 samples in Figure 2 is about 1.5 hours. Moreover, comparing to a linear counterpart, it is much more difficult to evaluate the importance of each individual covariate in the nonlinear kernel machine regression. While the KNIFE approach (Allen, 2013) can be potentially useful, more investigation is needed in future research. Third, as pointed out by a referee, it is promising to use more than one imputation for the counterfactual difference and to account for the impact of uncertainties of imputation in Theorem 1. However, there is substantial work to do in terms of evaluation of its properties and performance, as it represents a counterfactual extension of stability selection (Meinshausen and Bühlmann, 2010). We leave these to future investigations.

Moreover, it may be interesting to extend our method in survival data. Cai, Tonini and Lin (2011b) propose kernel machine methodology by using Cox model. We can adopt potential outcome into survival context. However, that use of that framework would then require rephrasing the potential outcomes model and attendant assumptions in §2.1.

Acknowledgement

The authors would like to thank the Editor, the Associate Editor and two referees for their constructive comments, which greatly improved the quality of the paper. This research is supported by National Cancer Institute (NCI) R01-CA129102 and NSF DMS 1914937.

Appendix A: Kernel machine methodology

The kernel machine regression model (Liu et al., 2007) is given by

$$Y_i = \beta_0 + h(\mathbf{Z}_i) + \epsilon_i, \tag{10}$$

where β_0 is an intercept term, $h(\mathbf{Z}_i)$ is an unknown centered smooth function, and the error term ϵ_i $(i = 1, \ldots, n)$ is assumed to be a random sample from a $N(0, \sigma^2)$ distribution. The kernel machine methodology assumes that $h(\cdot)$ lies in a reproducing Kernel Hilbert space (RKHS) and further details about RKHS can be found in literature (Aronszajn, 1950; Wahba, 1990; Berlinet and Thomas-Agnan, 2011). Let $\mathcal{H}_K$ denote the corresponding RKHS, which is a Hilbertian function space that satisfies the property that for any function in $\mathcal{H}_K$, its pointwise evaluation is a continuous linear functional. As shown in Aronszajn (1950), there exists a one-to-one correspondence between $\mathcal{H}_K$ with a so-called kernel function $K(\mathbf{z}, \mathbf{z}^*)$ is a bounded, symmetric, positive function satisfying

$$\int K(\mathbf{z}, \mathbf{z}^*)h(\mathbf{z})h(\mathbf{z}^*)d\mathbf{z}d\mathbf{z}^* \geq 0, \tag{11}$$

for any arbitrary square integrable function $h(\mathbf{z})$ and all $\mathbf{z}, \mathbf{z}^* \in R^p$. The kernel function can be viewed as a measure of similarity between two values of the covariate vector $\mathbf{z}$ and $\mathbf{z}^*$.

Any function $h(\mathbf{z})$ in the function space $\mathcal{H}_K$ defined by a kernel $K(\cdot, \cdot)$ can have a primal representation directly using the basis functions (features) of $\mathcal{H}_K$, and it can equivalently have a dual representation using the kernel function $K(\mathbf{z}, \mathbf{z}^*)$ directly. Specifically, for an arbitrary function $h(\mathbf{z}) \in \mathcal{H}_K$, its primal representation takes the form

$$h(\mathbf{z}) = \sum_{j=1}^{J} \omega_j \phi_j(\mathbf{z}) = \boldsymbol{\phi}(\mathbf{z})^T \boldsymbol{\omega}, \tag{12}$$

where $\boldsymbol{\phi}(\cdot) = \{\phi_1(\cdot), \cdots, \phi_J(\cdot)\}^T$ is a $J \times 1$ vector of the standardized orthogonal basis functions (features), i.e., standardized Mercer features of the function space $\mathcal{H}_k$, and the $\boldsymbol{\omega} = (\omega_1, \cdots, \omega_J)'$ is a vector of some constants. The square norm of $h(\cdot)$ can be written as

$$\|h\|^2_{\mathcal{H}_K} = \sum_{j=1}^{J} \omega_j^2 = \boldsymbol{\omega}^T \boldsymbol{\omega}. \tag{13}$$

Alternatively, the same $h(\mathbf{z})$ can be equivalently written in a dual representation using the kernel function $K(\cdot, \cdot)$ directly as

$$h(\mathbf{z}) = \sum_{l=1}^{L} \alpha_l K(\mathbf{z}_l^*, \mathbf{z}), \tag{14}$$

for some integer L, some constants α_l and some $\{\mathbf{z}_1^*, \cdots, \mathbf{z}_L^*\} \in R^p$. Justifications of these results and more details about the RKHS can be found in Chapter 3 of Cristianini et al. (2000).

Based on regression model (10), it is common to study the following Tikhonov regularized least squares problem (Smale and Zhou, 2007) with $\lambda > 0$ to avoid model over-fitting:

$$\widehat{\mathbf{h}} = \underset{h \in \mathcal{H}_K}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - \beta_0 - h(\mathbf{Z}_i)]^2 + \lambda \|h\|_{\mathcal{H}_K}^2 \right\}. \tag{15}$$

Exploiting a primal/dual equivalence from Karush-Kuhn-Tucker theory for (15), one can show that the estimator of the nonparametric function $h(\cdot)$ evaluated at the design points $(\mathbf{Z}_1, \cdots, \mathbf{Z}_n)^T$ is estimated as

$$\widehat{\mathbf{h}} = K(K + \lambda I)^{-1}\mathbf{y}, \tag{16}$$

where $\mathbf{y} \equiv (Y_1, \ldots, Y_n)$. For a new observation $\mathbf{Z}^*$, the predicted function value is given by:

$$\widehat{h}(\mathbf{Z}^*) = [K(\mathbf{Z}^*, \mathbf{Z}_1), \cdots, K(\mathbf{Z}^*, \mathbf{Z}_n)](K + \lambda I)^{-1}\mathbf{y}. \tag{17}$$

It has been shown that the estimates of h in (16) can be derived as arising from a random effects model of the following form (Liu et al., 2007):

$$\mathbf{y} = \beta_0 + \mathbf{h} + \mathbf{e}, \tag{18}$$

where $\mathbf{h}$ is an $n \times 1$ vector of random effects following $\mathbf{h} \sim N(0, \tau K)$, τ is a scale parameter, and $e \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. Because of this equivalence, all regression parameters in the model can be estimated by maximum likelihood, while the variance component parameters (including τ, σ^2 and other potential parameters involved in the kernel function $K(\cdot, \cdot)$ or equivalently the kernel matrix K) can be estimated by restricted maximum likelihood (Liu et al., 2007).

Appendix B: Proof of Theorem 1

The proof of Theorem 1 mostly follows the pioneering work of Smale and Zhou (2007) on the learning theories of integral operators. We first build a connection between the RKHS considered in the current paper with integral operators and then utilize the previous results on integral operators (Smale and Zhou, 2007) to prove our results on RKHS estimators as presented in Theorem 1 of the main text.

For ease of presentation, let X denote the covariate vector lies in a compact metric space $\mathcal{X}$ (e.g., a closed set in R^p) with probability measure P_X, and $L^2(P_X)$ denote the collection of all square-integrable functions, that is,

$$L^2(P_X) := \left\{ f(x) : \int_{\mathcal{X}} f^2(x) dP_X < \infty \right\}.$$

The type of integral operators considered in Smale and Zhou (2007) is $L_K : L^2(P_X) \mapsto \mathcal{H}_K$ defined by:

$$L_K(f)(x) := \int_{\mathcal{X}} K(x, x') f(x') dP_X(x'), x \in \mathcal{X},$$

where $K(x, x')$ is a reproducing (Mercer) kernel considered in §3 of the main text.

Since T_k is a linear operator, the eigenvalues and eigenfunctions of T_k are well-defined. Let $\phi_i \in L^2(P_X)$ be the normalized eigenfunctions of T_k (in fact $\{\phi_i\}$ form an orthonormal basis of $L^2(P_X)$) associated with eigenvalues $\lambda_i > 0$ sorted in non-increasing order. Then, the Mercer's theorem state that kernel K has the representation

$$K(x, x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(x'),$$

where the convergence is absolute and uniform. This representation can be used to define integral operator $L_K^{1/2}$ by

$$L_K^{1/2}(f)(x) := \int_{\mathcal{X}} \sum_{i=1}^{\infty} \sqrt{\lambda_i} \phi_i(x) \phi_i(x') f(x') dP_X(x'). \tag{19}$$

Furthermore, the Mercer's theorem implies that the corresponding RKHS $\mathcal{H}_K$ can be characterized as

$$\mathcal{H}_K = \left\{ f \in L^2(P_X) : ||f||^2_{\mathcal{H}_\mathcal{K}} = \sum_{i=1}^{\infty} \frac{< f, \phi_i >^2}{\lambda_i} < \infty \right\}, \tag{20}$$

where $< \cdot, \cdot >$ is the inner product in $L^2(P_X)$, and the specific inner product in this RKHS is given by

$$< f, g >_{\mathcal{H}_\mathcal{K}} = \sum_{i=1}^{\infty} \frac{< f, \phi_i >< g, \phi_i >}{\lambda_i}.$$

We now elaborate on (19) and (20) to show that $\mathcal{H}_K = Im(T_K^{1/2})$, where $Im(T_K^{1/2})$ denotes the range of integral operator $L_K^{1/2}$. On the one hand, according to (20), $\forall f \in \mathcal{H}_K$, we have $f(x) = \sum_{i=1}^{\infty} f_i \phi_i(x)$ with $\sum_{i=1}^{\infty} f_i^2/\lambda_i < \infty$, where $f_i =< f, \phi_i >$. On the other hand, according to (19), $\forall g \in Im(T_K^{1/2})$, there exists a function $h \in L^2(P_X)$ such that $g(x) = L_K^{1/2}(h)(x) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\phi_i(x)h_i$, where $h_i = \int_{\mathcal{X}} \phi_i(x')h(x')dP_X(x')$. The fact that $h \in L^2(P_X)$ implies that $\sum_{i=1}^{\infty} h_i^2 < \infty$. Combining both characterizations of f and g, we have:

$$\mathcal{H}_K = Im(T_k^{1/2}) := \left\{ f \in L^2(P_X) : f = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i}\phi_i, s.t. \sum_{i=1}^{\infty} a_i^2 < \infty \right\}. \tag{21}$$

This established connection between RKHS and integral operator T_K makes it possible to utilize previous learning theories of integral operators. Specifically, the RKHS scenario considered in Theorem 1 corresponds to the case of $r = 1/2$ in Corollary 5 of Smale and Zhou (2007). At first glance, the main result presented in Theorem 1 looks a bit different from the $r = 1/2$ result presented in Corollary 5 of Smale and Zhou (2007). Notice that the constant $\kappa := \sqrt{sup_{x\in\mathcal{X}} K(x,x)} = 1$ for the Gaussian kernel used in this paper, and the only remaining thing we need to prove is:

$$||T_K^{-1/2} f||_{L^2(P_X)} = ||f||_{\mathcal{H}_\mathcal{K}}.$$

To show this, we use our results on the connection between $\mathcal{H}_K$ and $Im(T_k^{1/2})$. For an arbitrary $f \in \mathcal{H}_K$, the equivalence result (21) indicates that $f = \sum_{i=1}^{\infty} a_i \sqrt{\lambda_i}\phi_i$ with $||f||_{\mathcal{H}_\mathcal{K}} = \sqrt{\sum_{i=1}^{\infty} a_i^2} < \infty$. On the other hand, equation (19) implies that $L_K^{1/2}(\sum_{i=1}^{\infty} a_i\phi_i)(x) = f(x)$. That is, $L_K^{-1/2}(f) = \sum_{i=1}^{\infty} a_i\phi_i$ and $||T_K^{-1/2} f||_{L^2(P_X)} = \sqrt{\sum_{i=1}^{\infty} a_i^2} = ||f||_{\mathcal{H}_\mathcal{K}}$. Combining all equivalence/equality results, the integral operator statement of Corollary 5 exactly reduces to the RKHS statement made in Theorem 1 of the current paper, which completes the proof to Theorem 1.

References

Alaa, A., Schaar, M., 2018. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design, in: International Conference on Machine Learning, PMLR. pp. 129–138.

Allen, G.I., 2013. Automatic feature selection via weighted kernels and regularization. Journal of Computational and Graphical Statistics 22, 284–299.

Aronszajn, N., 1950. Theory of reproducing kernels. Transactions of the American mathematical society 68, 337–404.

Berlinet, A., Thomas-Agnan, C., 2011. Reproducing kernel Hilbert spaces in probability and statistics. Springer Science & Business Media.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Bühmann, M.D., 2003. Radial basis functions: theory and implementations. volume 12. Cambridge university press.

Cai, T., Tian, L., Wong, P.H., Wei, L., 2011a. Analysis of randomized comparative clinical trial data for personalized treatment selections. Biostatistics 12, 270–282.

Cai, T., Tonini, G., Lin, X., 2011b. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. Biometrics 67, 975–986.

Cho, Y., Ghosh, D., 2021. Quantile-based subgroup identification for randomized clinical trials. Statistics in Biosciences 13, 90–128.

Cook, R., Weisberg, S., 1991. Discussion of 'sliced inverse regression for dimension reduction' by li. Journal of the American Statistical Association 86, 328–332.

Cook, R.D., 2009. Regression graphics: Ideas for studying regressions through graphics. volume 482. John Wiley & Sons.

Cook, R.D., Li, B., 2002. Dimension reduction for conditional mean in regression. The Annals of Statistics 30, 455–474.

Cristianini, N., Shawe-Taylor, J., et al., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Ferré, L., Yao, A.F., 2003. Functional sliced inverse regression analysis. Statistics 37, 475–488.

Foster, J.C., Taylor, J.M., Ruberg, S.J., 2011. Subgroup identification from randomized clinical trial data. Statistics in medicine 30, 2867–2880.

Fukumizu, K., Bach, F.R., Jordan, M.I., 2004. Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. Journal of Machine Learning Research 5, 73–99.

Fukumizu, K., Bach, F.R., Jordan, M.I., et al., 2009. Kernel dimension reduction in regression. The Annals of Statistics 37, 1871–1905.

Gail, M., Simon, R., 1985. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics , 361–372.

Ghosh, D., 2011. Propensity score modelling in observational studies using dimension reduction methods. Statistics & probability letters 81, 813–820.

Hammer, S.M., Katzenstein, D.A., Hughes, M.D., Gundacker, H., Schooley, R.T., Haubrich, R.H., Henry, W.K., Lederman, M.M., Phair, J.P., Niu, M., et al., 1996. A trial comparing nucleoside monotherapy with combination therapy in hiv-infected adults with cd4 cell counts from 200 to 500 per cubic millimeter. New England Journal of Medicine 335, 1081–1090.

Hill, J.L., 2011. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics 20, 217–240.

Holland, P.W., 1986. Statistics and causal inference. Journal of the American statistical Association 81, 945–960.

Imai, K., Ratkovic, M., et al., 2013. Estimating treatment effect heterogeneity in randomized program evaluation. The Annals of Applied Statistics 7, 443–470.

Kehl, V., Ulm, K., 2006. Responder identification in clinical trials with censored data. Computational statistics & data analysis 50, 1338–1355.

Lee, K.Y., Li, B., Chiaromonte, F., et al., 2013. A general theory for nonlinear sufficient dimension reduction: Formulation and estimation. Annals of Statistics 41, 221–249.

Li, B., Artemiou, A., Li, L., 2011. Principal support vector machines for linear and nonlinear sufficient dimension reduction. The Annals of Statistics 39, 3182–3210.

Li, K.C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86, 316–327.

Liu, D., Lin, X., Ghosh, D., 2007. Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics 63, 1079–1088.

Luo, W., Zhu, Y., Ghosh, D., 2017. On estimating regression-based causal effects using sufficient dimension reduction. Biometrika 104, 51–65.

Meinshausen, N., Bühlmann, P., 2010. Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72, 417–473.

Paindaveine, D., 2014. Elliptical symmetry. Wiley StatsRef: Statistics Reference Online .

Raghunathan, T.E., Lepkowski, J.M., Van Hoewyk, J., Solenberger, P., et al., 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27, 85–96.

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., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688.

Schoenberg, I.J., 1938. Metric spaces and completely monotone functions. Annals of Mathematics , 811–841.

Shen, Y., Cai, T., 2016. Identifying predictive markers for personalized treatment selection. Biometrics 72, 1017–1025.

Smale, S., Zhou, D.X., 2007. Learning theory estimates via integral operators and their approximations. Constructive approximation 26, 153–172.

Steinwart, I., 2001. On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research 2, 67–93.

Su, X., Tsai, C.L., Wang, H., Nickerson, D.M., Li, B., 2009. Subgroup analysis via recursive partitioning. Journal of Machine Learning Research 10.

Su, X., Zhou, T., Yan, X., Fan, J., Yang, S., 2008. Interaction trees with censored survival data. The International Journal of Biostatistics 4.

Van Buuren, S., 2018. Flexible imputation of missing data. CRC press.

VanderWeele, T.J., Luedtke, A.R., van der Laan, M.J., Kessler, R.C., 2019. Selecting optimal subgroups for treatment using many covariates. Epidemiology (Cambridge, Mass.) 30, 334.

Vickers, A.J., Kattan, M.W., Sargent, D.J., 2007. Method for evaluating prediction models that apply the results of randomized trials to individual patients. Trials 8, 1–11.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113, 1228–1242.

Wahba, G., 1990. Spline models for observational data. SIAM.

Wang, L., Zhou, Y., Song, R., Sherwood, B., 2018. Quantile-optimal treatment regimes. Journal of the American Statistical Association 113, 1243–1254.

Wu, H.M., 2008. Kernel sliced inverse regression with applications to classification. Journal of Computational and Graphical Statistics 17, 590–610.

Wu, Q., Liang, F., Mukherjee, S., 2013. Kernel sliced inverse regression: Regularization and consistency. Abstract and Applied Analysis 2013, 1–11.

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