A Sparse Random Projection-based Test for Overall Qualitative Treatment Effects

Chengchun Shi, Wenbin Lu and Rui Song*

Abstract

In contrast to the classical "one size fits all" approach, precision medicine proposes the customization of individualized treatment regimes to account for patients' heterogeneity in response to treatments. Most of existing works in the literature focused on estimating optimal individualized treatment regimes. However, there has been less attention devoted to hypothesis testing regarding the existence of overall qualitative treatment effects, especially when there is a large number of prognostic covariates. When covariates don't have qualitative treatment effects, the optimal treatment regime will assign the same treatment to all patients regardless of their covariate values. In this paper, we consider testing the overall qualitative treatment effects of patients' prognostic covariates in a high dimensional setting. We propose a sample splitting method to construct the test statistic, based on a nonparametric estimator of the contrast function. When the dimension of covariates is large, we construct the test based on sparse random projections of covariates into a low-dimensional space. We prove the consistency of our test statistic. In the regular cases, we show the asymptotic power function of our test statistic is asymptotically the same as the "oracle" test statistic which is constructed based on the "optimal" projection matrix. Simulation studies and real data applications validate our theoretical findings.

Keywords: High-dimensional testing; Optimal treatment regime; Precision medicine; Qualitative treatment effects; Sparse random projection.

*Chengchun Shi is graduate student (E-mail: cshi4@ncsu.edu), Wenbin Lu is Professor (E-mail: lu@stat.ncsu.edu), and Rui Song is Associate Professor (rsong@ncsu.edu), Department of Statistics, North Carolina State University, Raleigh, NC 27695.

1 Introduction

In many medical studies, patients may differ significantly in the way they respond to the treatment. In contrast to the classical "one size fits all" approach, precision medicine proposes the customization of individualized treatment regimes to account for patients' heterogeneity in response to treatments. Formally speaking, a treatment regime is a function from patients' prognostic covariates to available treatment options. The optimal individualized treatment regime (OITR) is the one that maximizes patients' expected responses among all treatment regimes.

There have been increasing interests in estimating the OITR. Some common methods include Q-learning (Watkins and Dayan, 1992; Chakraborty et al., 2010), A-learning (Robins et al., 2000; Murphy, 2003) and outcome weighted learning (OWL, Zhao et al., 2012). Qian and Murphy (2011) considered a two-step procedure to construct the OITR. Their method first estimates the conditional mean of the response with l_1 penalty function and then derives the OITR from the estimated conditional mean. Zhang et al. (2012) proposed a robust method for estimating the OITR by maximizing the estimated average response of patients (i.e, the value function). Zhang et al. (2015) proposed to use decision lists to construct interpretable and parsimonious treatment regimes. Despite the popularity of estimating the OITR, there is scarce work in the literature for hypothesis testing regarding OITR. All these estimation methods implicitly assume that patients' covariates have qualitative interactions with treatment, which means that there exists a subset of patients whose "best" treatments assigned according to the OITR are different from others.

We consider testing the existence of OITR due to the following reasons. First, the OITR may not always exist in practice, see the data from the Nefazodone-CBASP clinical trial study in Section 5 for an example. In this case, one treatment is better than the other for all patients and there is no need of estimating the OITR. Second, we note that implementing the OITR requires future patients' covariates which can be expensive to collect in some cases (Baker et al., 2009; Gail, 2009; Huang et al., 2015). In these cases, we recommend to adopt the "one-size-fits-all" paradigm when the null hypothesis of no OITR is not rejected. Third, our test is constructed based on estimated value functions' difference comparing the OITR

and a fixed regime (i.e. assign all to the best treatment). The test is not significant implies that the value functions' difference is not significant. Under such a situation, although we can still estimate the OITR, the gain of the obtained OITR over the fixed regime in terms of the improvement of value is not significant. Thus, the obtained OITR under such a situation may not be of practical interest. Therefore, it is essential to test the overall qualitative treatment effects of the prognostic covariates to determine whether we need to implement the OITR for future patients. Gunter et al. (2011) developed an S-score to quantify the magnitude of the marginal qualitative treatment effects of a single covariate. However, the S-score doesn't characterize the overall qualitative treatment effects of all covariates. Besides, no theoretical guarantees were provided for the S-score.

For binary treatments, testing qualitative treatment effects is equivalent to testing whether the interaction between treatment and covariates (i.e, the contrast function) is almost surely positive or negative. To test such hypothesis, Chang et al. (2015) proposed a test based on a L_1-type functional of kernel smoothing estimators of conditional treatment effects. Hsu (2017) proposed a Kolmogorov-Smirnov type test statistic based on nonparametric estimation of conditional treatment effects with a hypercube kernel. It is well known that kernel smoothing estimators are undesirable in practice due to the curse of dimensionality. As a result, these test statistics are not reliable when the dimension of the covariates is relatively large. However, in modern biomedical applications, it is likely to obtain a large number of prognostic factors for each individual patient. To the best of our knowledge, there are lack of methods for testing the overall qualitative treatment effects in high-dimensional settings.

In this paper, we aim to test the overall qualitative treatment effects in a high dimensional setting. This is a very challenging task due to the curse of dimensionality. To better illustrate this point, consider a simple situation where patients' covariates, x, consist of p independent Rademacher variables. Then, it is equivalent to test whether the contrast as a function of the covariates is always positive or negative for any $x \in \{-1, 1\}^p$. Therefore, we need to test 2^p moment inequalities even in this very simplified situation. However, for each $x \in \{-1, 1\}^p$, we have on average $N/2^P$ observations with covariates equal to x, where

N is the total number of observations. When $N = O(2^p)$, this seems impossible without additional assumptions. We show in Lemma 3.1 that covariates have the overall qualitative treatment effects if and only if the value function under the OITR is strictly larger than those under fixed treatment regimes. This motivates us to construct test statistics based on the difference between the optimal value function and the value function under fixed treatment regimes. However, inference for such value difference is extremely difficult in the nonregular cases, that is, there is a positive probability that the contrast function is equal to zero. We use a sample-splitting method to construct the test statistic, based on a nonparametric estimator of the contrast function. As long as the estimated contrast function satisfies certain convergence rates, we show our test statistic is consistent.

When the dimension of covariates is large, we construct the test based on sparse random projections of covariates into a low-dimensional space. Random projections have been a powerful method for dimension reduction in the computer science literature. The key idea behind is given in the Johnson-Lindenstrauss Lemma (Johnson and Lindenstrauss, 1984), which states that a set of high dimensional vectors can be projected into a suitable lower-dimensional space while approximately preserve their pairwise distances. In the statistics literature, Lopes et al. (2011) proposed a high-dimensional two-sample test which integrates a random projection with the Hotelling T^2 statistic. Recently, Cannings and Samworth (2015) proposed a random projection-based method for the high-dimensional classification.

In this paper, we propose the use of random projections with sparse matrix. In contrast to the dense sketching matrix used in Lopes et al. (2011) and Cannings and Samworth (2015), only a small proportion of elements in the sparse sketching matrix are nonzero. References on sparse random projections include Omidiran and Wainwright (2010); Li et al. (2006); Nelson and Nguyên (2013). In our simulation studies, we show that our sparse random projection-based test statistics are more powerful compared to those based on dense random projection matrix, when the OITR is "sparse". Besides, we advocate using data-dependent algorithms to generate sparse sketching matrix, since most random projections will be weakly correlated with the contrast function. In theory, we prove the consistency of our sparse random projection-based test. Moreover, in the regular cases, we

show that the power function of our test statistic is asymptotically the same as the "oracle" test statistic which is constructed based on the "optimal" projection matrix.

The rest of the paper is organized as follows. In Section 2, we present the definition of the overall qualitative treatment effects. In Section 3, we introduce our test statistic and study its asymptotic properties under the null and local alternative. Simulation studies and real data applications are conducted in Section 4 and Section 5 respectively, to examine the empirical performance of the proposed testing procedure. Section 6 concludes with a summary and discussions of possible extensions.

2 Overall qualitative treatment effects

We consider a single stage study with two treatment options. Let Y be a patient's outcome of interest and $A \in \{0, 1\}$ be the treatment indicator, with 0 for the standard treatment and 1 for the new treatment. By convention, a larger value of Y indicates a better clinical outcome. Denoted by $X \in \mathbb{R}^p$ the patient's baseline covariates. We consider a high dimensional setting where p is allowed to diverge with the sample size N. Let $Y^*(0)$ and $Y^*(1)$ denote the potential outcomes of a patient that would be observed assuming s/he received treatment 0 and 1, respectively. A treatment regime $d : \mathbb{R}^p \to \{0, 1\}$ is a deterministic function from patient's covariate space to all possible treatment options. For any d, we define the expected potential outcome

$$V(d) = \mathrm{E}[d(X)Y^*(1) + \{1 - d(X)\}Y^*(0)],$$

known as the value function associated with d. The optimal treatment regime d^{opt} is defined as the maximizer of $V(d)$. Let $\tau(x)$ be the contrast function, i.e,

$$\tau(x) = \mathrm{E}(Y|A = 1, X = x) - \mathrm{E}(Y|A = 0, X = x).$$

Under the following three conditions:

(A1.) Stable Unit Treatment Value Assumption (SUTVA): $Y = AY^*(0) + (1 - A)Y^*(1)$,

(A2.) No unmeasured confounders: $Y^*(0), Y^*(1) \perp\!\!\!\perp A \mid X$,

(A3.) Positivity: there exists some constants $0 < c_1 < c_2 < 1$ such that $0 < c_1 \leq c_2 < 1$ such that $c_1 \leq P(A = 1|X = x) \leq c_2$ for any x,

we can show that $\tau(x) = \mathrm{E}\{Y^*(1) - Y^*(0) \mid X = x\}$. Since

$$V(d) = \mathrm{E}[d(X)\{Y^*(1) - Y^*(0)\} + Y^*(0)] = \mathrm{E}\{\tau(X)d(X)\} + \mathrm{E}\{Y^*(0)\},$$

it is immediate to see that $d^{opt}(x) = I\{\tau(x) > 0\}$, where $I(\cdot)$ stands for the indicator function.

Condition (A2) is satisfied in a randomized study, where the propensity score function $\pi(x) = \Pr(A = 1 \mid X = x)$ is usually a known constant by design. We assume $\pi(x)$ is known throughout this Section. In Section 3.3, we allow the propensity score to be estimated from data as in observational studies.

Covariates X are said to have the overall qualitative treatment effects (OQTE) if

$$\Pr\{\tau(X) > 0\} > 0 \quad \text{and} \quad \Pr\{\tau(X) < 0\} > 0.$$

In this paper, we consider testing the following hypothesis:

$$H_0 : \ X \text{ dosen't have OQTE} \quad \text{versus} \quad H_1 : \ X \text{ has OQTE.} \tag{1}$$

Assume (A1)-(A3) hold. Under H_0, the optimal treatment regime assigns the same treatment to all patients. Therefore, testing OQTE is equivalent to testing the existence of OITR.

3 Proposed tests

3.1 A simple value-based test statistic in fixed p case

Assume the observed data are summarized as $\{O_i = (X_i, A_i, Y_i), i = 1, \ldots, N\}$, where O_i's are i.i.d. copies of $O = (X, A, Y)$. The distribution of O is allowed to vary with N. To

illustrate the idea, we first assume p is small and fixed, and present here a value-based test statistic for the null hypothesis (1). Later in this section, we will consider the more challenging high dimensional setting. Let $V(0) = \mathrm{E}\{Y^*(0)\}$ and $V(1) = \mathrm{E}\{Y^*(1)\}$. The following lemma relates OQTE to the difference between the optimal value function and the value functions under fixed treatment regimes.

Lemma 3.1. *Assume $E|\tau(X)| < \infty$, and conditions (A1)-(A3) hold. Then the followings are equivalent: (i) X doesn't have OQTE; (ii) $V(d^{opt}) = \max\{V(0), V(1)\}$.*

By definition, we have $V(d^{opt}) \geq \max\{V(0), V(1)\}$. Under H_1, Lemma 3.1 implies $V(d^{opt}) > \max(V(0), V(1))$. Therefore, it suffices to test

$$H_0 : V(d^{opt}) = \max\{V(0), V(1)\} \quad \text{versus} \quad H_1 : V(d^{opt}) > \max\{V(0), V(1)\}.$$

For simplicity, we assume $V(1) \geq V(0)$. This implies that the new treatment is at least as good as the standard one on average. The hypothesis $V(1) \geq V(0)$ can be tested using historical data or data from a pilot study. When $V(0) \geq V(1)$, the test statistic can be similarly constructed.

Lemma 3.1 motivates us to consider test statistics based on some estimators for the value difference $\mathrm{VD}(d^{opt}) = V(d^{opt}) - V(1)$. For any treatment regime d, Zhang et al. (2012) proposed an inverse propensity score weighted estimator (IPSWE) for $V(d)$:

$$\widehat{V}(d) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{A_i d(X_i)}{\pi_i} Y_i + \frac{(1 - A_i)\{1 - d(X_i)\}}{1 - \pi_i} Y_i \right], \tag{2}$$

where π_i is a shorthand for $\pi(X_i)$. Plugging $d \equiv 1$, we obtain $\widehat{V}(1) = N^{-1} \sum_i A_i Y_i / \pi_i$. For any fixed d, $\sqrt{N}\widehat{\mathrm{VD}}(d) = \sqrt{N}\{\widehat{V}(d) - \widehat{V}(1)\}$ corresponds to a sum of i.i.d random variables. Therefore, its asymptotic variance can be consistently estimated by the sample variance estimator,

$$\hat{\sigma}^2(d) = \frac{1}{N-1} \sum_{i=1}^{N} \left[\left(\frac{1 - A_i}{1 - \pi_i} - \frac{A_i}{\pi_i} \right) Y_i \{1 - d(X_i)\} - \widehat{\mathrm{VD}}(d) \right]^2. \tag{3}$$

Suppose $\hat{\tau}(\cdot)$ is an estimate of $\tau(\cdot)$. Based on (2) and (3), it is natural to use $\widehat{T} = \sqrt{N}\widehat{\mathrm{VD}}(\hat{d})/\hat{\sigma}(\hat{d})$ as the test statistic, where $\hat{d}(x) = I\{\hat{\tau}(x) > 0\}$, and reject H_0 when $\widehat{T} > z_\alpha$ at a given significance level α, where z_α stands for the upper α-th quantile of the standard normal distribution.

Consistency of such a naive test requires $\mathrm{E}|\hat{d}(X) - d^{opt}(X)|^2 \to 0$. However, as commented by Luedtke and van der Laan (2016), this assumption is typically violated in the non-regular cases where $\Pr\{\tau(X) = 0\} > 0$, even when $\hat{\tau}$ is consistent to τ. To solve this problem, we consider a modified version of $\widehat{T}$ based on sample splitting and cross-validation. Let $\mathcal{I}_1$ and $\mathcal{I}_2$ be a random partition of $\{1, \ldots, N\}$ into 2 disjoint subsets of equal sizes $n = N/2$. For any $\mathcal{I} \subseteq \{1, \ldots, N\}$ and treatment regime d, define

$$\begin{aligned}
\widehat{\mathrm{VD}}_{\mathcal{I}}(d) &= \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left[\left(\frac{1-A_i}{1-\pi_i} - \frac{A_i}{\pi_i}\right) Y_i\{1 - d(X_i)\}\right], \\
\hat{\sigma}^2_{\mathcal{I}}(d) &= \frac{1}{|\mathcal{I}| - 1}\sum_{i\in\mathcal{I}}\left[\left(\frac{1-A_i}{1-\pi_i} - \frac{A_i}{\pi_i}\right) Y_i\{1 - d(X_i)\} - \widehat{\mathrm{VD}}_{\mathcal{I}}(d)\right]^2,
\end{aligned}$$

where $|\mathcal{I}|$ stands for the number of elements in $\mathcal{I}$. Let $\hat{\tau}_{\mathcal{I}}$ be the corresponding estimator of τ based on observations in $\mathcal{I}$ and $\hat{d}_{\mathcal{I}}(x) = I\{\hat{\tau}_{\mathcal{I}}(x) > 0\}$. We define our test statistic by

$$\widehat{T}_{CV} = \max\left(\frac{\sqrt{n}\widehat{\mathrm{VD}}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2})}{\max\{\hat{\sigma}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2}), \delta_n\}}, \frac{\sqrt{n}\widehat{\mathrm{VD}}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1})}{\max\{\hat{\sigma}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1}), \delta_n\}}\right), \tag{4}$$

for some positive sequence $\delta_n \to 0$, and reject H_0 when $\widehat{T}_{CV} > z_{\alpha/2}$. The sequence δ_n guarantees that the denominators in $\widehat{T}_{CV}$ are strictly greater than 0.

Alternative to the sample splitting method, one can consider a Wald-type test statistic based on the online one-step estimator proposed by Luedtke and van der Laan (2016). However, calculating such test statistic is more computationally expensive than ours. Besides, the asymptotic normality of such test statistic requires the class of functions $\left\{[(1-A)/\{1-\pi(X)\} - A/\pi(X)]Y\{1-d(X)\} : d\right\}$ to be Glivenko-Cantelli, where d varies over the range of estimators $\hat{d}$ (see Section 7.3 in Luedtke and van der Laan, 2016). In contrast, our testing procedure is valid under H_0 for any $\hat{d}$.

Theorem 3.1. *Assume conditions (A1)-(A3) hold,* $E|Y|^3 = O(1)$ *and* $\delta_n \gg n^{-1/6}$. *Then under* H_0, *for any* $0 < \alpha < 1$, *we have*

$$\limsup_n Pr(\widehat{T}_{CV} > z_{\alpha/2}) \leq \alpha.$$

Moreover, assume that

$$Var\left\{\left(\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right) Y\{1 - \hat{d}_{\mathcal{I}_j}(X)\} \mid \{O_i\}_{i\in\mathcal{I}_j}\right\} = o_p(\delta_n), \tag{5}$$

for $j = 1, 2$, *where* $Var(V_1 \mid V_2)$ *denotes the variance of* V_1 *conditional on* V_2. *Then, we have* $Pr(\widehat{T}_{CV} > z_{\alpha/2}) \to 0$.

The following theorem states the consistency of our proposed test statistic. It relies on Conditions (C1) and (C2). We provide these conditions in Section B of the Appendix to save space.

Theorem 3.2. *Assume conditions (A1)-(A3), (C1), (C2) hold,* $E|Y|^3 = O(1)$ *and* $\delta_n \to 0$. *Under* $H_1 : V(d^{opt}) = V(1) + h_n$, *if* $h_n \gg n^{-1/2}$, *then we have* $Pr(\widehat{T}_{CV} > z_{\alpha/2}) \to 1$. *Moreover, assume* $Pr\{\tau(X) = 0\} = 0$ *and* $\liminf_n \sigma_0^2 > 0$ *where*

$$\sigma_0^2 = Var\left\{\left(\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right) Y\{1 - d^{opt}(X)\}\right\}.$$

If $\sqrt{n}h_n = O(1)$, *then we have*

$$Pr\left(\widehat{T}_{CV} > z_{\alpha/2}\right) = 2\bar{\Phi}\left(z_{\frac{\alpha}{2}} - \frac{\sqrt{n}h_n}{\sigma_0}\right) - \bar{\Phi}^2\left(z_{\frac{\alpha}{2}} - \frac{\sqrt{n}h_n}{\sigma_0}\right) + o(1),$$

where $\bar{\Phi}(z) = Pr(Z \geq z)$ *for a standard normal random variable* Z.

Theorem 3.1 and 3.2 show the consistency of our testing procedure. Note that Conditions (C1) and (C2) are not required to ensure Theorem 3.1. This suggests the type-I error is well controlled regardless of any estimating procedure. On the other hand, conditions on δ_n in Theorem 3.1 are stronger than those in Theorem 3.2. In the regular cases when $\Pr\{\tau(X) = 0\} = 0$, Theorem 3.2 provides the asymptotic power function of our test.

Notice that h_n is equal to $\mathrm{E}[-\tau(X)\{\tau(X) < 0\}]$ which relies on the dependence structure of the covariates. As a result, the power of our test depends crucially on the underlying data-generating process.

In this paper, $\hat{d}$ is obtained by a plug-in estimator based on some nonparametric estimation of the contrast function. Alternatively, one can directly estimate d^{opt} using OWL. Theorem 3.2 holds as long as the estimated decision function $\hat{d}$ satisfies $V(\hat{d}_{\mathcal{I}}) = V(d^{opt}) + o_p(|\mathcal{I}|^{-1/2})$.

Since we assume $V(1) \geq V(0)$, under H_0, we have $\Pr\{\tau(X) \geq 0\} = 1$. In the regular cases where $\Pr\{\tau(X) = 0\} = 0$, we have $\Pr\{d^{opt}(X) = 1\} = 1$ and hence

$$\mathrm{Var}\left\{\left(\frac{A}{\pi(X)} - \frac{1-A}{1-\pi(X)}\right) Y\{1 - d^{opt}(X)\}\right\} = 0.$$

Besides, in the regular cases, d^{opt} can be consistently estimated by $\hat{d}_{\mathcal{I}_j}$ (see Equation (S.18) in the supplementary article). Assume conditions (C1) and (C2) hold with $\gamma \geq 1$. Then we can show (5) holds. Hence, the type-I error of our test will go to 0.

3.2 A sparse random projection-based test statistic

When p is large, it is far more challenging to estimate the contrast function $\tau(x)$ due to the curse of dimensionality. To handle high-dimensional covariates, we project the covariate space into a low dimensional vector space to construct our test statistic. Throughout this paper, we assume the dimension of the projected space, q is fixed. For a given matrix $S \in \mathbb{R}^{q\times p}$ and any $\omega \in \mathbb{R}^q$, define

$$\tau^S(\omega) = \mathrm{E}\{\tau(X)|SX = \omega\}.$$

Under (A1)-(A3), the treatment regime $d_S^{opt}(x) = I\{\tau^S(Sx) > 0\}$ is optimal in the sense that it maximizes the value function among the class of treatment regimes based only on the projected covariates SX.

Since q is small, τ^S can be consistently estimated. We can construct a value-based test statistic as discussed in Section 3.1 based on the projected data $\{O_i^S\}_{i\in\{1,\ldots,N\}}$ where

$O_i^S = (SX_i, A_i, Y_i)$. The power of such test statistic depends crucially on the sketching matrix S. To better understand this, consider the following example:

$$\tau(X) = \left\{ \left(\frac{X^{(1)} + X^{(2)}}{\sqrt{2}} \right)^2 - \delta \right\} \left(\frac{X^{(3)} + X^{(4)} + X^{(5)} + X^{(6)} + X^{(7)}}{\sqrt{5}} \right)^2, \tag{6}$$

for some $\delta > 0$, where $X^{(j)}$ denotes the j-th element of X.

Apparently, we have $\tau(X) > 0$ if $|X^{(1)} + X^{(2)}| > \sqrt{2\delta}$ and $\tau(X) < 0$ if $|X^{(1)} + X^{(2)}| < \sqrt{2\delta}$. Assume $X \sim N(0, I_p)$. Then X have the OQTE. Let $q = 1$, the "optimal" sketching matrix S^* is equal to

$$S^* = c_0(1, 1, 0, 0, \ldots, 0),$$

for any $c_0 \neq 0$. For any $S \in \mathbb{R}^{1\times p}$ such that $S^* S^T = 0$, SX is independent of $X^{(1)} + X^{(2)}$. Then, we have

$$\begin{aligned}
& \tau^S(\omega) = \mathrm{E}\{\tau(X)|SX = \omega\} \\
= \; & \mathrm{E}\left[\left| \left\{ \left(\frac{X^{(1)} + X^{(2)}}{\sqrt{2}} \right)^2 - \delta \right\} \left(\frac{X^{(3)} + X^{(4)} + X^{(5)} + X^{(6)} + X^{(7)}}{\sqrt{5}} \right)^2 \right| SX = \omega \right] \\
= \; & (1 - \delta)\mathrm{E}\left\{ \left(\frac{X^{(3)} + X^{(4)} + X^{(5)} + X^{(6)} + X^{(7)}}{\sqrt{5}} \right)^2 \middle| SX = \omega \right\}.
\end{aligned}$$

Hence, $\tau^S(\omega)$ is always nonnegative or nonpositive as a function of ω. As a result, the test statistic based on $\{O_i^S\}_i$ doesn't have any power to detect the OQTE. The challenge here lies in finding a projection matrix S that is highly correlated with S^*.

Below, we propose a data-dependent algorithm to generate S and introduce our test statistic. Our theory shows that our test statistic works as if the optimal sketching matrix S^* were known. Statistical properties of our testing procedure are formally studied in Section 3.2.2.

3.2.1 Test statistic

Assume for now, we have an estimator $\hat{\tau}^S_{\mathcal{I}}$ for τ^S based on any subset of the projected data $\{O_i^S\}_{i\in\mathcal{I}}$ and an algorithm to sample sparse sketching matrices whose distribution $G(S, \{O_i\}_{i\in\mathcal{I}})$ is allowed to depend on $\{O_i\}_{i\in\mathcal{I}}$. We describe the whole testing procedure in Algorithm 1.

Algorithm 1. Calculate the random projection-based test statistic.

1. Input observations $\{O_i\}_{i=1,\ldots,N}$, δ_n, α and a sampling distribution G.
2. Randomly partition the data into two subsets $\{O_i\}_{i\in\mathcal{I}_1}$ and $\{O_i\}_{i\in\mathcal{I}_2}$.
3. For $j = 1, 2$,
   (i) Independently sample a sparse sketching matrix $S_{\mathcal{I}_j} \sim G(S, \{O_i\}_{i\in\mathcal{I}_j})$;
   (ii) Obtain estimators $\hat{\tau}_{\mathcal{I}_j}^{S_{\mathcal{I}_j}}$ and $\hat{d}_{\mathcal{I}_j}^{S_{\mathcal{I}_j}}(x) = I\{\hat{\tau}_{\mathcal{I}_j}^{S_{\mathcal{I}_j}}(S_{\mathcal{I}_j}x) > 0\}$;
   (iii) Calculate $\widehat{T}^{S_{\mathcal{I}_j}} = \sqrt{n}\widehat{\mathrm{VD}}_{\mathcal{I}_j^c}(\hat{d}_{\mathcal{I}_j}^{S_{\mathcal{I}_j}})/\max\{\hat{\sigma}_{\mathcal{I}_j^c}(\hat{d}_{\mathcal{I}_j}^{S_{\mathcal{I}_j}}), \delta_n\}$.
4. Reject H_0 if $\widehat{T}_{SRP} = \max(\widehat{T}^{S_{\mathcal{I}_1}}, \widehat{T}^{S_{\mathcal{I}_2}}) > z_{\alpha/2}$.

Now we present our algorithm for generating sparse sketching matrix. We first introduce some notations. For any matrix Ψ with J rows, let $\Psi^{(i)}$ be the ith row of Ψ. For any vector $\psi \in \mathbb{R}^J$ and any set $\mathcal{M} \subseteq \{1, \ldots, J\}$, denote $\psi_{\mathcal{M}}$ as the subvector of ψ formed by elements in $\mathcal{M}$. Let $\mathcal{M}^c$ be the complement of $\mathcal{M}$. Let $\|\psi\|_0$ be the number of nonzero elements in ψ and $\|\psi\|_2$ be the Euclidean norm of ψ. Let $\mathcal{S}$ denote the space of sparse sketching matrices:

$$\mathcal{S} = \{S \in \mathbb{R}^{q\times p} : \|S^{(i)}\|_0 \le s, \|S^{(i)}\|_2 = 1, \forall i = 1, \ldots, q\},$$

for some fixed integer s that satisfies $2 \le s \le p$. Denoted by $N(0, I_J)$ a J-dimensional Gaussian random vector with mean zero and identity covariance matrix.

It remains to generate $S_{\mathcal{I}_j}$ based on the sub-dataset $\{O_i\}_{i\in\mathcal{I}_j}$. We first sample many sparse sketching matrices from $\mathcal{S}$. Each row of the sketching matrix is independently and uniformly distributed on the space $\{S \in \mathbb{R}^p : \|S\|_0 = s, \|S\|_2 = 1\}$. This corresponds to Step 2 in our proposed algorithm below. Then we output the sparse sketching matrix that maximizes the estimated value difference function. Specifically, we propose using data-

splitting strategy for evaluation of the value difference function. That is, for each sketching matrix, we randomly divide $\{O_i\}_{i\in\mathcal{I}_j}$ into $\mathbb{K}$ folds, use any of the $\mathbb{K}-1$ subsamples to estimate the OITR based on projected covariates, use the remaining subsamples to evaluate the corresponding value difference function, and aggregate these value difference functions over different subsamples. This corresponds to Step 3-5 in our proposed algorithm below. We summarize our procedure in Algorithm 2.

Algorithm 2. Generate data-dependent sparse random sketching matrix.

1. Input observations $\{O_i\}_{i\in\mathcal{I}}$, integers B, s, q and $\mathbb{K}\geq 2$.
2. Generate i.i.d matrices $S_1, S_2, \ldots, S_B$ according as S_0 whose distribution is described as follows. For $j=1,\ldots,q$,
   (i) Independently select a simple random sample $\mathcal{M}_j$ of size s from $\{1,\ldots,p\}$;
   (ii) Independently generate a Gaussian random vector $g_j\sim N(0,I_s)$;
   (iii) Set $S^{(j)}_{0,\mathcal{M}_j^c}=0$ and $S^{(j)}_{0,\mathcal{M}_j}=g_j/\|g_j\|_2$.
3. Randomly divide $\mathcal{I}$ into $\mathbb{K}$ subsets $\{\mathcal{I}^{(k)}\}_{k=1}^{\mathbb{K}}$ of equal sizes. Let $\mathcal{I}^{(k)-}=\mathcal{I}\cap(\mathcal{I}^{(k)})^c$.
4. For $b=1,\ldots,B$,
   (i) For $k=1,\ldots,\mathbb{K}$,
      (i.1) Obtain the estimator $\hat{\tau}^{S_b}_{\mathcal{I}^{(k)-}}$ and $\hat{d}^{S_b}_{\mathcal{I}^{(k)-}}(x)=I\{\hat{\tau}^{S_b}_{\mathcal{I}^{(k)-}}(S_bx)>0\}$;
      (i.2) Evaluate the value difference $\widehat{\mathrm{VD}}_{\mathcal{I}^{(k)}}(\hat{d}^{S_b}_{\mathcal{I}^{(k)-}})$.
   (ii) Obtain the cross-validated estimator $\widehat{\mathrm{VD}}^{S_b}_{CV}=\sum_k\widehat{\mathrm{VD}}_{\mathcal{I}^{(k)}}(\hat{d}^{S_b}_{\mathcal{I}^{(k)-}})/\mathbb{K}$.
5. Output $S_{\hat{b}}$, where $\hat{b}=\arg\max_{b=1}^{B}\widehat{\mathrm{VD}}^{S_b}_{CV}$.

3.2.2 Asymptotic properties under the null and local alternative

We first show the validity of the proposed test, which applies to any estimator $\hat{\tau}^S_{\mathcal{I}}$. For any positive sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n\gg b_n$ if and only if $\limsup_n b_n/a_n=0$.

Theorem 3.3. *Assume (A1)-(A3) hold, $E|Y|^3=O(1)$ and $\delta_n\gg n^{-1/6}$. Then under H_0, we have*

$$\limsup_n Pr\left(\widehat{T}_{SRP}>z_{\alpha/2}\right)\leq\alpha.$$

Moreover, assume that

$$Var\left\{\left(\frac{A}{\pi(X)}-\frac{1-A}{1-\pi(X)}\right)Y\{1-\hat{d}_{\mathcal{I}_j}^{S_{\mathcal{I}_j}}(X)\}\mid\{O_i\}_{i\in\mathcal{I}_j},S_{\mathcal{I}_j},\mathcal{I}_j\right\}=o_p(\delta_n),$$

for $j=1,2$. *Then we have* $Pr(\widehat{T}_{SRP}>z_{\alpha/2})\to 0$.

Let $S^*=\arg\max_{S\in\mathcal{S}}V(d_S^{opt})$ be the optimal sketching matrix. The optimal sketching matrix S^* may not be unique. To see this, for any sketching matrix $S^*\in\mathcal{S}$ that maximizes $V(d_S^{opt})$, $-S^*$ also maximizes $V(d_S^{opt})$ and we have $-S^*\in\mathcal{S}$. Moreover, when $q\geq 2$, there may exist infinitely many maximizers.

Our theoretical studies are mostly concerned with the "oracle" test statistic. The oracle knew the set $\mathcal{S}^*$ ahead of time. In Algorithm 1: Step 3(i), instead of using Algorithm 2 to sample $S^{\mathcal{I}_1}$ and $S^{\mathcal{I}_2}$, we use the oracle set $S^{\mathcal{I}_1}=S^{\mathcal{I}_2}=S^*$ for an arbitrary $S^*\in\mathcal{S}^*$. Denoted by $\widehat{T}_{oracle}$ the resulting oracle test statistic. Let $h_n^*=\arg\max_{S^*\in\mathcal{S}^*}V(d_{S^*}^{opt})-V(1)$. Similar to Theorem 3.2, under H_1, if $h_n^*\gg n^{-1/2}$, then we can show

$$\Pr\left(\widehat{T}_{oracle}>z_{\alpha/2}\right)\to 1.$$

Assume

$$V(d_{S^*}^{opt})=V(d^{opt}),\quad\forall S^*\in\mathcal{S}^*. \tag{7}$$

This condition means the optimal decision rule depends on the set of projected covariates S^*X only. It holds when $\tau(x)=\phi(S^*x)\mathrm{g}(x)$ for some function $\phi(\cdot)$ and some nonnegative function $\mathrm{g}(\cdot)$. In the regular cases where $\Pr(\tau(X)=0)=0$, (7) implies that $\Pr(d^{opt}(X)=d_{S^*}^{opt}(X))=1,\forall S^*\in\mathcal{S}^*$. Thus, the class of optimal treatment regimes $\{d_{S^*}^{opt}:S^*\in\mathcal{S}^*\}$ will almost surely recommend the same treatment to any given patient. Assume (7) holds and $\Pr(\tau(X)=0)=0$. Similar to Theorem 3.2, the asymptotic power of $\widehat{T}_{oracle}$ can be derived as

$$\Pr\left(\widehat{T}_{oracle}>z_{\alpha/2}\right)=2\bar{\Phi}\left(z_{\frac{\alpha}{2}}-\frac{\sqrt{n}h_n}{\sigma_0}\right)-\bar{\Phi}^2\left(z_{\frac{\alpha}{2}}-\frac{\sqrt{n}h_n}{\sigma_0}\right)+o(1), \tag{8}$$

where h_n and σ_0 are defined in Theorem 3.2.

In the following, we prove the consistency of our proposed testing procedure when using Algorithm 2 to generate the sparse sketching matrix. Moreover, we show our test statistic possesses the oracle property. This means the power function of $\widehat{T}_{SRP}$ is asymptotically the same as the oracle test statistic $\widehat{T}_{oracle}$.

Define the semimetric

$$d^{\tau}(S_1, S_2) = \sqrt{\mathrm{E}|\tau^{S_1}(S_1X) - \tau^{S_2}(S_2X)|^2}, \quad \forall S_1, S_2 \in \mathcal{S}.$$

We make the following assumptions.

(A4.) For any sketching matrices $S_1, S_2, \ldots, S_B \in \mathcal{S}$ and any $\mathcal{I} \subseteq \{1, 2, \ldots, N\}$ with $|\mathcal{I}| \geq n/2$, assume the following event holds with probability tending to 1,

$$\max_{b=1}^{B} \mathrm{E}^X |\hat{\tau}_{\mathcal{I}}^{S_j}(S_jX) - \tau^{S_j}(S_jX)|^2 = O\left(n^{-r_0} \log n\right),$$

where the expectation E^X is taken with respect to X, and the little-o term is uniform in $\mathcal{I}$ and $S_1, \ldots, S_B$.

(A5.) Assume $B \gg (p\sqrt{n})^{(s-1)q}$. In addition, assume there exist some constant $\bar{C} > 0$ and some sketching matrix $S^* \in \mathcal{S}^*$ such that

$$d^{\tau}(S, S^*) \leq \bar{C} \left(\sum_{j=1}^{q} \|S^{(j)} - S^{*(j)}\|_2^2 \right)^{1/2}, \quad \forall S \in \mathcal{S}. \tag{9}$$

(A6.) Assume there exist some constants γ, ε_0, $\delta_0 > 0$ such that for any sketching matrix S satisfying $V(d_S^{opt}) \geq V(d_{S^*}^{opt}) - \varepsilon_0$, we have $\Pr\{0 < |\tau^S(SX)| \leq t\} = O(t^{\gamma})$, where the big-$O$ term is uniform in $0 < t < \delta_0$ and S.

Condition (A4) assumes the uniform convergence rate of $\hat{\tau}_{\mathcal{I}}^{S_b}$ for $b = 1, \ldots, B$. Since the uniform convergence rate increases as B increases, Condition (A4) gives the upper bound for B. On the contrary, Condition (A5) gives the lower bound for B. It requires B to diverge at a proper rate, to give us a good chance for finding a random projection with a

large value function. More specifically, under (A5), we can show that

$$\Pr\left\{\max_{b=1}^{B} V(d_{S_b}^{opt}) = V(d_{S^*}^{opt}) + o(n^{-1/2})\right\} \to 1.$$

In Section C.3 of the Appendix, we show (A5) holds when $\tau(x) = \phi(S^*X)$ for some sketching matrix $S^* \in \mathcal{S}^*$ and some Lipschitz continuous function $\phi(\cdot)$.

Condition (A6) holds with $\gamma = 1$ when $\tau^S(SX)$ has a uniformly bounded density function near 0 for any sketching matrix S that nearly maximizes the value function (see Section C.4 in the Appendix for detailed discussion). Assume $\tau(X) \geq \delta_0$ almost surely or $\tau(X) \leq -\delta_0$ almost surely. Then for any sketching matrix S, we have $\tau^S(SX) \geq \delta_0$ almost surely or $\tau^S(SX) \leq -\delta_0$. As a result, (A6) automatically holds for any $\gamma > 0$.

In Section C.2 of the Appendix, we consider a simple model and show (A4)-(A6) holds.

Theorem 3.4. *Assume Conditions (A1)-(A5) hold,* $E|Y|^3 = O(1)$, $\log B = o(n^{1/3})$ *and* $\delta_n \to 0$. *If* $h_n^* \gg \max(\sqrt{\log B}/\sqrt{n}, n^{-r_0/2}\sqrt{\log n})$, *then we have*

$$Pr\left(\widehat{T}_{SRP} > z_{\alpha/2}\right) \to 1.$$

Moreover, assume (7) *and (A6) hold,* $Pr\{\tau(X) = 0\} = 0$, $\sqrt{n}h_n = O(1)$, $B = O(n^{\kappa_B})$ *for some* $\kappa_B > 0$, $r_0 > \frac{\gamma+2}{2\gamma+2}$ *and* $\liminf_n \sigma_0 > 0$. *Then we have*

$$Pr\left(\widehat{T}_{SRP} > z_{\alpha/2}\right) = 2\bar{\Phi}\left(z_{\frac{\alpha}{2}} - \frac{\sqrt{n}h_n}{\sigma_0}\right) - \bar{\Phi}^2\left(z_{\frac{\alpha}{2}} - \frac{\sqrt{n}h_n}{\sigma_0}\right) + o(1).$$

Assume $p = O(n)$ and we set $B = c_* n^{\{3q(s-1)+\epsilon\}/2}$ for any $c_*, \epsilon > 0$. Then the conditions $B \gg (p\sqrt{n})^{(s-1)q}$ in (A5) and $B = O(n^{\kappa_B})$ in Theorem 3.4 automatically hold. It is worth mentioning that when h_n and σ_0 don't depend on p, Theorem 3.4 implies that the asymptotic power of our test is independent of p.

3.3 Some implementation issues

3.3.1 Doubly-robust test statistics

So far we have assumed that the propensity scores are known for all patients. In the following, we introduce a doubly-robust test statistic to deal with data from an observational study. We begin by introducing a doubly-robust value difference estimator, which requires the estimation of the propensity score and the conditional mean functions $h_0(x) = \mathrm{E}(Y|A=0, X=x)$ and $h_1(x) = \mathrm{E}(Y|A=1, X=x)$. Denoted by $\hat{\pi}(\cdot)$, $\hat{h}_0(\cdot)$ and $\hat{h}_1(\cdot)$ the corresponding estimators. Zhang et al. (2012) proposed a doubly-robust estimator for the value function under a given treatment regime d,

$$\begin{aligned}\widehat{V}^{dr}(d) &= \frac{1}{N}\sum_{i=1}^{N}\left\{\left(\frac{A_i}{\hat{\pi}(X_i)}d_i + \frac{1-A_i}{1-\hat{\pi}(X_i)}(1-d_i)\right)Y_i\right. \\ &- \left.\left(\frac{A_i}{\hat{\pi}(X_i)}d_i + \frac{1-A_i}{1-\hat{\pi}(X_i)}(1-d_i) - 1\right)\{\hat{h}_0(X_i)(1-d_i) + \hat{h}_1(X_i)d_i\}\right\},\end{aligned}$$

where d_i is a shorthand for $d(X_i)$. When either the propensity score or the conditional mean models are correctly specified, $\widehat{V}^{dr}(d)$ is consistent to $V(d)$ (Zhang et al., 2012). Based on $\widehat{V}^{dr}$, for any $\mathcal{I} \subset [1, \ldots, N]$ and a given treatment regime d, we define our doubly-robust value difference estimator as

$$\widehat{\mathrm{VD}}^{dr}_{\mathcal{I}}(d) = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left\{\left(\frac{1-A_i}{1-\hat{\pi}_i^{\mathcal{I}}} - \frac{A_i}{\hat{\pi}_i^{\mathcal{I}}}\right)Y_i - \left(\frac{1-A_i}{1-\hat{\pi}_i^{\mathcal{I}}} - 1\right)\hat{h}^{\mathcal{I}}_{0,i} + \left(\frac{A_i}{\hat{\pi}_i^{\mathcal{I}}} - 1\right)\hat{h}^{\mathcal{I}}_{1,i}\right\}(1-d_i),$$

where $\hat{\pi}_i^{\mathcal{I}} = \hat{\pi}^{\mathcal{I}}(X_i)$, $\hat{h}^{\mathcal{I}}_{0,i} = \hat{h}^{\mathcal{I}}_0(X_i)$, $\hat{h}^{\mathcal{I}}_{1,i} = \hat{h}^{\mathcal{I}}_1(X_i)$, and $\hat{\pi}^{\mathcal{I}}$, $\hat{h}^{\mathcal{I}}_0$, $\hat{h}^{\mathcal{I}}_1$ are obtained based on $\{O_i\}_{\mathcal{I}}$. When p is large, we recommend to estimate π, h_0 and h_1 via penalized regression. The asymptotic variance of $\sqrt{|\mathcal{I}|}\widehat{\mathrm{VD}}^{dr}(d)$ can be consistently estimated by $\hat{\sigma}^{dr}_{\mathcal{I}}(d)$ whose exact form is given in Section A of the Appendix.

We briefly summarize our test procedures. Similar to Algorithm 1, we first randomly partition the data into two halves $\{O_i\}_{\mathcal{I}_1}$ and $\{O_i\}_{\mathcal{I}_2}$, and obtain the estimators $\hat{\pi}^{\mathcal{I}_j}$, $\hat{h}_0^{\mathcal{I}_j}$, $\hat{h}_1^{\mathcal{I}_j}$ based on $\{O_i\}_{i\in\mathcal{I}_j}$ for $j = 1, 2$. Then we independently sample the sparse sketching matrices $S_{\mathcal{I}_1}$ and $S_{\mathcal{I}_2}$. The sampling algorithm is similar to Algorithm 2. Specifically,

for $j = 1, 2$, we randomly divide $\mathcal{I}_j$ into $\{\mathcal{I}_j^{(k)}\}_{k=1}^{\mathbb{K}}$ and independently sample $S_1, \ldots, S_B$ as Steps 2 and 3 of Algorithm 2. Then we calculate the doubly-robust value difference estimator,

$$\widehat{\mathrm{VD}}_{CV}^{dr,S_b} = \mathbb{K}^{-1} \sum_k \widehat{\mathrm{VD}}_{\mathcal{I}_j^{(k)}}^{dr}(\hat{d}_{\mathcal{I}_j^{(k)-}}^{S_b}), \tag{10}$$

for each S_b where $\mathcal{I}_j^{(k)-} = \mathcal{I}_j \cap (\mathcal{I}_j^{(k)})^c$, and set $S_{\mathcal{I}_j} = S_{\hat{b}}$ where $\hat{b} = \arg\max_{b=1}^{B} \widehat{\mathrm{VD}}_{CV}^{dr,S_b}$. Finally, we define our test statistic by

$$\widehat{T}_{SRP}^{dr} = \max\left(\frac{\sqrt{n}\widehat{\mathrm{VD}}_{\mathcal{I}_2}^{dr}(\hat{d}_{\mathcal{I}_1}^{S_{\mathcal{I}_1}})}{\max\{\hat{\sigma}_{\mathcal{I}_2}^{dr}(\hat{d}_{\mathcal{I}_1}^{S_{\mathcal{I}_1}}), \delta_n\}}, \frac{\sqrt{n}\widehat{\mathrm{VD}}_{\mathcal{I}_1}^{dr}(\hat{d}_{\mathcal{I}_2}^{S_{\mathcal{I}_2}})}{\max\{\hat{\sigma}_{\mathcal{I}_1}^{dr}(\hat{d}_{\mathcal{I}_2}^{S_{\mathcal{I}_2}}), \delta_n\}} \right), \tag{11}$$

and reject the test if $\widehat{T}_{SRP}^{dr} > z_{\alpha/2}$ for a given significance level $\alpha > 0$. Statistical properties of $\widehat{T}_{SRP}^{dr}$ can be similarly established.

3.3.2 Estimation of τ^S

The projected contrast function τ^S can be estimated by any machine learning or statistical nonparametric methods. In our implementation, we estimate τ^S using cubic B-splines. Let $\mathcal{I}$ be an arbitrary subset of $\{1, \ldots, N\}$. Based on the dataset $\{O_i\}_{i \in \mathcal{I}}$, we first estimate π using the penalized logistic regression, and estimate h_0, h_1 using the penalized linear regression, with SCAD penalty functions (Fan and Li, 2001). These penalized regressions are implemented by the R package `ncvreg` and the tuning parameters are selected via 10-folded cross-validation. Let $\hat{\pi}_i^{\mathcal{I}}$, $\hat{h}_{0,i}^{\mathcal{I}}$ and $\hat{h}_{1,i}^{\mathcal{I}}$ be the corresponding estimators for $\pi(X_i)$, $h_0(X_i)$ and $h_1(X_i)$, respectively. Recall that $S^{(j)} \in \mathbb{R}^{1 \times p}$ is the jth row of sketching matrix S. We define the pseudo outcome

$$\hat{\tau}_i^{\mathcal{I}} = \left(\frac{A_i}{\hat{\pi}_i^{\mathcal{I}}} - \frac{1 - A_i}{1 - \hat{\pi}_i^{\mathcal{I}}} \right) Y_i - \left(\frac{A_i}{\hat{\pi}_i^{\mathcal{I}}} - 1 \right) \hat{h}_{1,i}^{\mathcal{I}} + \left(\frac{1 - A_i}{1 - \hat{\pi}_i^{\mathcal{I}}} - 1 \right) \hat{h}_{0,i}^{\mathcal{I}}, \tag{12}$$

and minimize

$$(\hat{\xi}_1^{\mathcal{I}}, \cdots, \hat{\xi}_q^{\mathcal{I}}) = \underset{\xi_1,\cdots,\xi_q \in \mathbb{R}^k}{\arg\min} \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(\hat{\tau}_i^{\mathcal{I}} - \sum_{j=1}^{q} \sum_{k=1}^{K+4} N_k^{S^{(j)}}(S^{(j)} X_i) \xi_{j,k} \right)^2, \tag{13}$$

where $N_1^{S^{(j)}}(\cdot), \ldots, N_{K+4}^{S^{(j)}}(\cdot)$ are cubic B-spline bases of $S^{(j)}X_i$ and K is the number of interior knots. Given K, we place the interior knots at equally-spaced sample quantiles of the projected covariates $\{SX_i\}_{i \in \mathcal{I}}$. After solving (13), we set $\hat{\tau}_{\mathcal{I}}^S(Sx) = \sum_{j=1}^{q} \sum_{k=1}^{K+4} N_k^{S^{(j)}}(S^{(j)}x) \hat{\xi}_{j,k}^{\mathcal{I}}$.

Based on the B-spline methods, we show in Section C.2.1 of the Appendix (A4) holds with $r_0 = 4/5$ when $q = 1$ and $B = O(n^{\kappa_B})$ for any $\kappa_B > 0$. Assume (A6) holds with $\gamma > 2/3$. The condition $r_0 > (\gamma + 2)/(2\gamma + 2)$ in Theorem 3.4 is thus satisfied. More generally, we may use series estimator (Belloni et al., 2015) to estimate τ^S. Then the rate r_0 in (A4) will decrease as the number of projected dimension q increases.

3.3.3 Choice of s

Our testing procedure requires specification of s, which determines the number of nonzero elements in each row of the sketching matrix. Ideally, one could treat s as a tuning parameter and choose s to maximize the estimated value difference defined in (10). However, this approach would be time-consuming. In our implementation, we set s as a discrete random variable when sampling $S_1, \ldots, S_B$. More specifically, for $b = 1, \ldots, B$, we first independently sample s according as some random variable s_0, and then sample S_b according to Step 3 of Algorithm 2.

We recommend to set $s_0 = 2 + \text{Binom}(p - 2, p_0)$, where $\text{Binom}(m, p_0)$ is a binomial random variable with the total number of trials equal to m and the probability of success equal to p_0. In our simulation study, we set $p_0 = 2/(p - 2)$.

3.3.4 Choice of q

The choice of the projection dimension q involves a trade-off. If q is too large, then the curse of dimensionality will affect the uniform convergence rates of $\hat{\tau}_{\mathcal{I}}^{S_j}$ in (A8), resulting in decreased power of the corresponding test. If q is too small, then the OITR is not well

approximated. In our numerical experiments, we set $q = 1$. In the supplementary article, we examine the performance of the proposed test with difference choices of q. Results show that the optimal choice of q depends on the number of covariates involved in the OITR and varies across different simulation settings. We further propose a method that adaptively determines q. Detailed algorithm is given in Section E.2 of the supplementary article. In our simulations, we find such adaptive method is no worse than any fixed choice of q and has nearly optimal performance in some cases.

3.3.5 Choices of other hyperparameters

We recommend to set the number of folds $\mathbb{K}$ in Algorithm 2 to be 5 or 10. The number of sketching matrices B shall diverge as $N, p \to \infty$. In practice, we recommend to set $B = N^{\kappa_N} p^{\kappa_p}$ for some $\kappa_N, \kappa_p \geq 1$.

4 Simulations

4.1 Settings

We examine the finite sample performance of the proposed tests via Monte Carlo simulations. Simulated data with sample size N were generated from

$$Y = 1 + (X^{(1)} - X^{(2)})/2 + A\tau(X) + e,$$

where $X \sim N(0, I_p)$, $A \sim \text{Binom}(1, 0.5)$ and $e \sim N(0, 0.5^2)$. Here, we set $p = 50$ or 100.

We consider four scenarios. In the first three scenarios, we set

$$\tau(X) = \phi_\delta\{(X^{(1)} + X^{(2)})/\sqrt{2}\}(X^{(3)} + X^{(4)} + X^{(5)} + X^{(6)} + X^{(7)})^2/5,$$

for some function ϕ_δ parameterized by some $\delta > 0$. More specifically, we set $\phi_\delta(x) = x^2 - \delta$ in Scenario 1, $\phi_\delta(x) = \delta\cos(\pi x)$ in Scenario 2, and $\phi_\delta(x) = \delta\sqrt{2\pi}x$ in Scenario 3.

In Scenario 4, we set

$$\tau(X) = \delta \left\{ \left(\sum_{j=1}^{2} \frac{X^{(j)}}{\sqrt{2}} \right)^2 - \left(\sum_{j=3}^{20} \frac{X^{(j)}}{\sqrt{18}} \right)^2 \right\} (X^{(21)} + X^{(22)} + X^{(23)} + X^{(24)} + X^{(25)})^2/5.$$

It is immediate to see that the OITR is sparse and is a function of $X^{(1)}$ and $X^{(2)}$ in the first three scenarios. In Scenario 4, however, a total of 20 variables are involved in the OITR. In addition, the true OITR is linear in X under Scenario 3, but non-linear under Scenarios 1, 2 and 4. We set $N = 500$ in Scenarios 1, 2 and 3, and $N = 1000$ in Scenario 4.

For all scenarios, the parameter δ controls the degree of overall qualitative treatment effects. Specifically, H_0 holds if $\delta = 0$ and H_1 holds if $\delta > 0$. For each scenario, we further consider four cases by setting $\text{VD}(d^{opt}) = V(d^{opt}) - V(1) = 0, 0.2, 0.35$ and 0.5. Note that in Scenarios 2, 3 and 4, the settings for $\text{VD}(d^{opt}) = 0$ are the same. Hence, in Scenarios 3 and 4, we only report the simulation results for $\text{VD}(d^{opt}) = 0.2, 0.35$ and 0.5.

We set $q = 1$ and calculate $\widehat{T}^{dr}_{SRP}$ as described in Section 3.3. The number of interior knots K in the cubic B-spline bases is specified in the following fashion. When generating $S_{\mathcal{I}_1}$ or $S_{\mathcal{I}_2}$, we fix $K = 3$ when estimating τ^{S_b} for $b = 1, \ldots, B$. After obtaining $S_{\mathcal{I}_1}$ and $S_{\mathcal{I}_2}$, K is tuned with cross-validation when estimating $\tau^{S_{\mathcal{I}_1}}$ and $\tau^{S_{\mathcal{I}_2}}$. We set $B = 10^5$ for $p = 50$ and $B = 4 \times 10^5$ for $p = 100$.

The whole simulation program is implemented in R. Some subroutines, including sampling data-dependent sketching matrices $S_{\mathcal{I}_1}$ and $S_{\mathcal{I}_2}$ and estimating $\tau^{S_{\mathcal{I}_1}}$ and $\tau^{S_{\mathcal{I}_2}}$, are written in C with the GNU Scientific Library (GSL, Galassi et al., 2015).

4.2 Competing methods

Comparison is made among the following five test statistics:

(i) The proposed sparse random projection-based test statistic $\widehat{T}^{dr}_{SRP}$.

(ii) The dense random projection-based test statistic, denoted by $\widehat{T}^{dr}_{RP}$.

(iii) The cross-validated test statistic with the OITR estimated by the penalized least square method developed in Shi et al. (2016), denoted by $\widehat{T}_{PLS}$.

(iv) The cross-validated test statistic based on step-wise variable selection, denoted by $\widehat{T}_{VS}$.

(v) The supremum-type test statistic $\widehat{T}_{DL}$ based on the desparsified Lasso estimator (Zhang and Zhang, 2014; van de Geer et al., 2014).

$\widehat{T}_{RP}^{dr}$ is computed in a similar fashion as $\widehat{T}_{SRP}^{dr}$. We randomly partition $\{1,\ldots,N\}$ into $\mathcal{I}_1 \cup \mathcal{I}_2$ of equal size, generate some data dependent sketching matrices $S_{\mathcal{I}_1}$ and $S_{\mathcal{I}_2}$, and construct the test statistic as in (11). When generating $S_{\mathcal{I}_1}$ or $S_{\mathcal{I}_2}$, instead of sampling B sparse sketching matrices as described in Step 3 of Algorithm 2, we generate B dense sketching matrices $S_1,\ldots,S_B$ according to $Z_0/\|Z_0\|_2$, where $Z_0 \in \mathbb{R}^p$ is a Gaussian random vector with mean zero and identity covariance matrix, and set $S_{\mathcal{I}_1}$ or $S_{\mathcal{I}_2}$ to be the one that gives the largest cross-validated value difference as in (10). Similar to $\widehat{T}_{SRP}^{dr}$, we set $B = 10^5$ for $p = 50$ and set $B = 4 \times 10^5$ for $p = 100$, and use cubic B-splines to estimate τ^S for any sketching matrix S.

To calculate $\widehat{T}_{PLS}$, we first partition the data into two halves $\{O_i\}_{i\in\mathcal{I}_1}$ and $\{O_i\}_{i\in\mathcal{I}_2}$. Then for $j = 1, 2$, we set $\hat{d}_{\mathcal{I}_j}(x) = I(\bar{x}^T\hat{\beta}^{\mathcal{I}_j} > 0)$ where $\bar{x} = (1, x^T)^T$, $\hat{\beta}^{\mathcal{I}_j}$ is computed by

$$\hat{\beta}^{\mathcal{I}_j} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i\in\mathcal{I}_j} \frac{1}{|\mathcal{I}_j|} \left(Y_i - \bar{X}_i^T\hat{\theta}^{\mathcal{I}_j} - (A_i - \hat{\pi}_i^{\mathcal{I}_j})\bar{X}_i^T\beta\right)^2 + \sum_{j=2}^{p+1} p_{\lambda_{n,1}}(|\beta_j|), \tag{14}$$

for some penalty functions p_λ, where $\bar{X}_i = (1, X_i^T)^T$, $\hat{\pi}_i^{\mathcal{I}_j}$ is the estimated propensity score for the ith patient based on a penalized logistic regression with SCAD penalty function, and $\hat{\theta}^{\mathcal{I}_j}$ is calculated by

$$\hat{\theta}^{\mathcal{I}_j} = \underset{\theta\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i\in\mathcal{I}_j} \frac{1}{|\mathcal{I}_j|}(Y_i - \bar{X}_i^T\theta)^2 + \sum_{j=2}^{p+1} p_{\lambda_{n,2}}(|\theta_j|). \tag{15}$$

We use the SCAD penalty in both (14) and (15). The tuning parameters $\lambda_{n,1}$ and $\lambda_{n,2}$ were selected via 10-folded cross-validation. Finally, define $\widehat{T}_{PLS}$ by

$$\widehat{T}_{PLS} = \max\left(\frac{\sqrt{n}\widehat{\mathrm{VD}}_{\mathcal{I}_2}^{dr}(\hat{d}_{\mathcal{I}_1})}{\max\{\hat{\sigma}_{\mathcal{I}_2}^{dr}(\hat{d}_{\mathcal{I}_1}), \delta_n\}}, \frac{\sqrt{n}\widehat{\mathrm{VD}}_{\mathcal{I}_1}^{dr}(\hat{d}_{\mathcal{I}_2})}{\max\{\hat{\sigma}_{\mathcal{I}_1}^{dr}(\hat{d}_{\mathcal{I}_2}), \delta_n\}}\right). \tag{16}$$

To compute $\widehat{T}_{VS}$, we similarly split the observations into two sub-datasets $\{O_i\}_{i\in\mathcal{I}_1}$ and

$\{O_i\}_{i\in\mathcal{I}_2}$. For each sub-dataset, we apply the sequential advantage selection (SAS, Fan et al., 2016) to select variables with a qualitative interaction with the treatment. SAS is a greedy stepwise selection procedure and uses a BIC-type criterion to choose the best candidate subset of variables. Denoted by $\widehat{\mathcal{M}}_{\mathcal{I}_1}, \widehat{\mathcal{M}}_{\mathcal{I}_2} \subseteq \{1,\ldots,p\}$ the corresponding sets of selected variables. Then for each $j = 1, 2$, we calculate the pseudo responses $\hat{\tau}_i^{\mathcal{I}_j}, \forall i \in \mathcal{I}_j$ (see the definition in (12)) and compute

$$\hat{\tau}_{\mathcal{I}_j} = \arg\min_{f\in\mathbb{H}_j} \frac{1}{n}\sum_{i\in\mathcal{I}_j}\{\hat{\tau}_i^{\mathcal{I}_j} - f(X_{i,\widehat{\mathcal{M}}_{\mathcal{I}_j}})\}^2 + \lambda_j\|f\|^2_{\mathbb{H}_j},$$

where $\lambda_j > 0$ is a tuning parameter, $\mathbb{H}_j$ is the reproducing kernel Hilbert space with the reproducing kernel $K_j(X_{i,\widehat{\mathcal{M}}_{\mathcal{I}_j}}, X_{k,\widehat{\mathcal{M}}_{\mathcal{I}_j}}) = \exp\{-\sum_{l\in\widehat{\mathcal{M}}_{\mathcal{I}_j}} \eta_{j,l}(X_i^{(l)} - X_k^{(l)})^2\}$ where $X_i^{(l)}, X_k^{(l)}$ denote the l-th element in X_i, X_k and $\eta_{j,l} > 0, \forall l \in \widehat{\mathcal{M}}_{\mathcal{I}_j}$ are tuning parameters. The estimating procedure is implemented by the R package `listdtr` and the tuning parameters are selected via leave-one-out cross validation. Then we define $\hat{d}_{\mathcal{I}_j}(x) = I\{\hat{\tau}_{\mathcal{I}_j}(x_{\widehat{\mathcal{M}}_{\mathcal{I}_j}}) > 0\}$ and set

$$\widehat{T}_{VS} = \max\left(\frac{\sqrt{n}\widehat{\mathrm{VD}}^{dr}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1})}{\max\{\hat{\sigma}^{dr}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1}), \delta_n\}}, \frac{\sqrt{n}\widehat{\mathrm{VD}}^{dr}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2})}{\max\{\hat{\sigma}^{dr}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2}), \delta_n\}}\right). \tag{17}$$

We set $\delta_n = \log(\log_{10}(2n))/(2n)^{1/6}$ in (11), (16) and (17), where $\log_{10}$ denotes the logarithm with base 10.

$\widehat{T}_{DL}$ tests the overall treatment effects by fitting the following linear regression model for the response:

$$\mathrm{E}(Y|A, X) \approx \beta_0 + X^T\beta_x + A\beta_a + AX^T\beta_{ax}.$$

Based on this model, testing the overall treatment effects is equivalent to test $H_0^* : \beta_{ax} = 0$. Denoted by $\beta = (\beta_0, \beta_x^T, \beta_a, \beta_{ax}^T)^T$. To deal with high dimensionality, we estimate β by the desparsified Lasso estimator $\widehat{\beta}^{DL}$ and test H_0^* based on the following supremum-type test statistic, $\max_{j\in\mathcal{M}_{ax}} \sqrt{n}|\widehat{\beta}_j^{DL}|$, where $\mathcal{M}_{ax} = \{p+3,\ldots,2p+2\}$ and $\widehat{\beta}_j^{DL}$ is the j-th element of $\widehat{\beta}^{DL}$. The critical value of $\widehat{T}^{DL}$ is approximated via bootstrap. Detailed implementation

of the test can be found in Zhang and Cheng (2017).

4.3 Results

We conduct 500 simulations for each setting and report the proportions of rejecting the null hypothesis (%) in Table 1 and Table 2, with standard errors in parenthesis (%). Under H_0, the type-I errors of our test statistic is well controlled. Specifically, in Scenario 1 when VD = 0, the rejection probability of $\widehat{T}^{dr}_{SRP}$ is exactly zero. This is in line with our theory which suggests that the type-I error of our test statistics will converge to 0 in the regular cases where $\Pr\{\tau(X) = 0\} = 0$. In Scenario 2 when VD = 0, the rejection probability of $\widehat{T}^{dr}_{SRP}$ is close to the nominal level.

Table 1: Rejection probabilities (%) of the sparse random projection-based test, dense random projection-based test, penalized least square-based test, step-wise selection-based test and the supremum-type test based on the desparsified Lasso estimator, with standard errors in parenthesis (%), under Scenarios 1 and 2 where $X \sim N(0, I_p)$.

Scenario 1		VD = 0		VD = 20%		VD = 35%		VD = 50%	
		α level		α level		α level		α level	
	p	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
$\widehat{T}^{dr}_{SRP}$	50	0(0)	0(0)	24(1.9)	39.6(2.2)	71(2.0)	81(1.8)	90.8(1.3)	95.2(1.0)
	100	0(0)	0(0)	17.4(1.7)	29.6(2.0)	60.8(2.2)	73.8(2.0)	86.6(1.5)	92.4(1.2)
$\widehat{T}^{dr}_{RP}$	50	0(0)	0(0)	0.2(0.2)	0.6(0.4)	0.8(0.4)	3.2(0.8)	7.2(1.2)	18.6(1.7)
	100	0(0)	0(0)	0.4(0.3)	0.4(0.3)	0.4(0.3)	4(0.9)	6.8(1.1)	19(1.8)
$\widehat{T}_{PLS}$	50	0(0)	0(0)	0(0)	0(0)	0.4(0.3)	0.8(0.4)	6(1.1)	17.6(1.7)
	100	0(0)	0(0)	0(0)	0(0)	0.8(0.4)	2.4(0.7)	8.6(1.3)	19.8(1.8)
$\widehat{T}_{VS}$	50	0(0)	0(0)	1.2(0.5)	3.8(0.9)	16 (1.6)	29.4 (2.0)	36.6(2.2)	50.8(2.2)
	100	0(0)	0(0)	0(0)	0.6(0.3)	8.4 (1.2)	17.4 (1.7)	23.8(1.9)	36.4(2.2)
$\widehat{T}_{DL}$	50	10.2(1.4)	22.4(1.9)	11.2(1.4)	22.8(1.9)	10.8 (1.4)	21.8 (1.9)	9.8(1.3)	22.4(1.9)
	100	7.6(1.2)	20.0(1.8)	7.8(1.2)	21.4(1.8)	7.6 (1.2)	22.0 (1.9)	6.8(1.1)	21.6(1.8)
Scenario 2		VD = 0		VD = 20%		VD = 35%		VD = 50%	
		α level		α level		α level		α level	
	p	0.01	0.05	0.01	0.05	0.01	0.05	0.01	0.05
$\widehat{T}^{dr}_{SRP}$	50	1.2(0.5)	5.4(1)	24(1.9)	35.8(2.1)	76.4(1.9)	84.6(1.6)	90.2(1.3)	94(1.1)
	100	0.6(0.3)	5.2(1)	15.2(1.6)	28.2(2)	67(2.1)	78.8(1.8)	84.2(1.6)	90.4(1.3)
$\widehat{T}^{dr}_{RP}$	50	1.8(0.6)	4.6(0.9)	2(0.6)	4.8(1)	1.6(0.6)	5.4(1)	1(0.4)	6(1.1)
	100	1.2(0.5)	4.2(0.9)	1.2(0.5)	5.4(1)	0.6(0.3)	4.8(1)	0.8(0.4)	4.4(0.9)
$\widehat{T}_{PLS}$	50	1.8(0.6)	6(1.1)	1.2(0.5)	4.4(0.9)	1(0.4)	4.2(0.9)	0.8(0.4)	3.8(0.9)
	100	1.2(0.5)	4.2(0.9)	0.8(0.4)	4.6(0.9)	0.6(0.3)	5.6(1)	0.6(0.3)	5(1)
$\widehat{T}_{VS}$	50	1.2(0.5)	6.4(1.1)	0.6(0.3)	4(0.9)	1(0.4)	6.6(1.1)	1(0.4)	5(1)
	100	1.4(0.5)	5(1.0)	1.0(0.4)	5(1.0)	1.4(0.5)	6.4(1.1)	0.6(0.3)	4.6(0.9)
$\widehat{T}_{DL}$	50	1.6(0.6)	6.4(1.1)	2.8(0.7)	11.8(1.4)	4.4 (0.9)	15.4 (1.6)	5.4 (1.0)	17(1.7)
	100	1.2(0.5)	3.6(0.8)	2.8(0.7)	11.8(1.4)	5.2(1.0)	17.6(1.7)	7.2(1.2)	19.8(1.8)

Table 2: Rejection probabilities (%) of the sparse random projection-based test, dense random projection-based test, penalized least square-based test, step-wise selection-based test and the supremum-type test based on the desparsified Lasso estimator, with standard errors in parenthesis (%), under Scenarios 3 and 4 where $X \sim N(0, I_p)$.

Scenario 3		VD = 20%		VD = 35%		VD = 50%	
		α level		α level		α level	
	p	0.01	0.05	0.01	0.05	0.01	0.05
$\widehat{T}^{dr}_{SRP}$	50	47.2(2.2)	71.8(2)	92.4(1.2)	97.8(0.7)	99(0.4)	100(0)
	100	42.4(2.2)	61.2(2.2)	89.8(1.4)	96.2(0.9)	97.2(0.7)	99.4(0.3)
$\widehat{T}^{dr}_{RP}$	50	4.4(0.9)	16.2(1.6)	13.4(1.5)	35.8(2.1)	22(1.9)	49.4(2.2)
	100	3(0.8)	8.4(1.2)	4(0.9)	14.2(1.6)	5.4(1)	19.6(1.8)
$\widehat{T}_{PLS}$	50	76.4(1.9)	92(1.2)	97.8(0.7)	99.4(0.3)	99.4(0.3)	100(0)
	100	64.8(2.1)	87(1.5)	97(0.8)	99.4(0.3)	98.6(0.5)	99.8(0.2)
$\widehat{T}_{VS}$	50	55.6(2.2)	81.8(1.7)	93(1.1)	99(0.4)	97.8(0.7)	100(0)
	100	49.8(2.2)	74.2(2.0)	90(1.3)	98.6(0.5)	99(0.4)	100(0)
$\widehat{T}_{DL}$	50	99.8(0.7)	100(0)	100(0)	100(0)	100(0)	100(0)
	100	99.2(0.4)	100(0)	100(0)	100(0)	100(0)	100(0)
Scenario 4		VD = 20%		VD = 35%		VD = 50%	
		α level		α level		α level	
	p	0.01	0.05	0.01	0.05	0.01	0.05
$\widehat{T}^{dr}_{SRP}$	50	22.4(1.9)	41.8(2.2)	60.4(2.2)	76.6(1.9)	72.4(2)	87.2(1.5)
	100	15.2(1.6)	28(2)	49.6(2.2)	70.2(2)	70(2)	84(1.6)
$\widehat{T}^{dr}_{RP}$	50	0.4(0.3)	6.2(1.1)	0.6(0.3)	5.4(1)	0.2(0.2)	5.4(1)
	100	1.2(0.5)	6(1.1)	0.8(0.4)	3.8(0.9)	1.2(0.5)	5.2(1)
$\widehat{T}^{dr}_{PLS}$	50	1.2(0.5)	5.4(1)	1.2(0.5)	6(1.1)	1.4(0.5)	4.8(1)
	100	1.6(0.6)	5.8(1)	1.8(0.6)	6(1.1)	1.4(0.5)	5.2(1)
$\widehat{T}^{dr}_{VS}$	50	10.4(1.4)	24.2(1.9)	13.6(1.5)	30.6(2.1)	13.2(1.5)	29.4(2)
	100	5(1)	15.6(1.6)	4.6(0.9)	20(1.8)	8.2(1.2)	18.4(1.7)
$\widehat{T}^{dr}_{DL}$	50	4.2(0.9)	11.4(1.4)	5.4(1)	14.2(1.6)	6.4(1.1)	15.8(1.6)
	100	6.2(1.1)	16(1.6)	6.4(1.1)	18.2(1.7)	6.8(1.1)	19.6(1.8)

Under H_1, we can see that our test statistic is much more powerful compared to other competing test statistics in Scenarios 1, 2 and 4. For example, when VD = 0.35 and $\alpha = 0.05$, the rejection probabilities of our test are around 75% in Scenario 1. On the other hand, $\widehat{T}^{dr}_{RP}$, $\widehat{T}_{PLS}$ and $\widehat{T}_{VS}$ fail in Scenario 2. Specifically, the rejection probabilities of these three tests are no more than 6% in all settings. The rejection probabilities of $\widehat{T}_{DL}$ are around 10%-20% in Scenario 2 under H_1. However, $\widehat{T}_{DL}$ doesn't have valid type-I error rates under H_0. Here, the test statistics $\widehat{T}_{PLS}$ and $\widehat{T}_{DL}$ fail mainly due to the fact that the true OITR is not linear, while $\widehat{T}^{dr}_{RP}$ and $\widehat{T}_{VS}$ fail partly because the dense projection and greedy stepwise variable selection cannot correctly identify the variables with qualitative interactions.

In Scenario 3, $\widehat{T}_{DL}$ and $\widehat{T}_{PLS}$ achieve the greatest power in all settings as expected since the true OITR is linear in this scenario. Notice that $X^{(1)}, X^{(2)}, \cdots, X^{(7)}$ are independent.

Although the contrast function is not linear, the estimated contrast functions via the penalized least squares (see (14) and (15)) will converge to $\mathrm{E}\{\tau(X)|X^{(1)}, X^{(2)}\}$. As a result, the estimated OITR is consistent. When VD = 0.35 and 0.5, the rejection probabilities of $\widehat{T}^{dr}_{SRP}$ are slightly smaller when compared to $\widehat{T}_{PLS}$, $\widehat{T}_{DL}$ and $\widehat{T}_{VS}$, but are much larger than those of $\widehat{T}^{dr}_{RP}$.

In Section E of the supplementary article, we report the rejection probabilities of $\widehat{T}^{dr}_{SRP}$, $\widehat{T}^{dr}_{RP}$, $\widehat{T}_{PLS}$, $\widehat{T}_{VS}$ and $\widehat{T}_{DL}$ under the scenario where $X \sim N(0, \{0.5^{|i-j|}\}_{i,j=1,\ldots,p})$. Results are similar to those presented in Table 1 and 2.

4.4 Computation time

Our tests are computed on a 32 core 2.2GHz machine with 512GB RAM. Fixing $B = 10^5$, it took approximately 3 minutes to implement the test in Scenarios 1-3 where $N = 500$ and 5 minutes in Scenario 4 where $N = 1000$. The computation time can be largely reduced if we use a much smaller B. For example, if we set $B = 10^4$ in some simulation settings, the computation is 10 times faster and the test performance is still satisfactory. Moreover, since our testing procedure independently generates many sketching matrices and retains the one that maximizes the estimated value function, it can be naturally implemented in parallel. This scalability can further effectively reduce the computational cost.

5 Real data

We apply our proposed test to the data from the Nefazodone-CBASP clinical trial study (Keller et al., 2000), which enrolled 681 patients with nonpsychotic chronic major depressive disorder (MDD). Patients were randomized to three treatments, including Nafazodone (coded as 0), Cognitive Behavioral-Analysis System of Psychotherapy (CBASP, coded as 1), and the combination of Nefazodone and CBASP (2). The outcome of interests were patients' scores on the 24-item Hamilton Rating Scale for Depression (HRSD). The maximum value of HRSD was 43 and we set $Y = 43 - \mathrm{HRSD}$ as our response. Larger value of Y indicates better clinical outcome. Similarly as in Zhao et al. (2012), we use a subset

of 647 patients that have complete records of 50 baseline covariates for analysis. Among them, 216 were treated with Nafazodone, 220 with CBASP and 211 with the combination.

Our objective was to test whether the baseline covariates X have overall qualitative treatment effects. This is equivalent to test $H_0 : V(d^{opt}) = \max\{V(0), V(1), V(2)\}$, where $V(d^{opt})$ is the optimal value function, and $V(j)$ denotes the value function under the fixed treatment regimes by assigning all patients to treatment j, for $j = 0, 1, 2$. Patients' average responses under treatment $0, 1, 2$ are 27.14, 27.27 and 32.13, respectively. Besides, pairwise t tests show that $V(2)$ is significantly larger than $V(0)$ and $V(1)$. Therefore, it suffices to test $H_0 : V(d^{opt}) = V(2)$. This is equivalent to test the intersection of the following two hypotheses:

$$H_0^{(j)} : V(d^{opt,(j)}) = \max_{k\in\{0,1,2\},k\neq j} V(k),$$

for $j = 0, 1$, where $d^{opt,(j)}$ is the optimal treatment regime comparing Treatment 2 with Treatment j. For testing $H_0^{(j)}$, we computed the test statistic $\widehat{T}_{SRP}^{dr,j}$ as described in Section 3.3 and 4.1. We set $B = 100000$ and $\delta_n = \log(\log_{10}(2n))/(2n)^{1/6}$. For a given $0 < \alpha < 1$, we reject H_0 if

$$\max_{j=0,1} \widehat{T}_{SRP}^{dr,j} > z_{\alpha/4}.$$

By Bonferroni's inequality, the type-I error is well-controlled.

The two test statistics are equal to -0.67 and 0.31, respectively. We fail to reject H_0 at a significance level of 0.1. Therefore, we suspect that the prognostic covariates in this study might not have qualitative treatment effects. Zhao et al. (2012) performed pairwise comparisons between the combination treatment and any single treatment, and estimate the OITR by the outcome weighted learning. Their estimated optimal treatment regime recommended the combination treatment to all the patients. Our tests formally verify their findings.

6 Discussion

In this paper, we develop tests for overall qualitative treatment effects. The test statistics are constructed by a sample-splitting method. In the high-dimensional setting, we use sparse random projections of the covariate space to construct the test statistic and introduce a data-dependent way to sample sparse projection matrices. In theory, we show the consistency of the proposed test statistic and prove its "oracle" property in the regular cases.

6.1 Nonnegative average treatment effects

In this paper, we assume $V(1) \geq V(0)$ (the new treatment is on average better than the standard control) and consider the test statistic based on estimators for the value difference $V(d^{opt}) - V(1)$. When such prior information is not available, let $\hat{a}_{\mathcal{I}_j} = \arg\max_{a \in \{0,1\}} \widehat{V}_{\mathcal{I}_j}(a)$ for $j = 0, 1$ where $\mathcal{I}_1$ and $\mathcal{I}_2$ stands for a random partition of the dataset, $\widehat{V}_{\mathcal{I}_j}(a)$ denotes the estimated value function based on observations in $\mathcal{I}_j$ under the decision rule $d(x) = a, \forall x$. We can consider the following test statistic,

$$\widehat{T}_{CV} = \max \left\{ \frac{\sqrt{|\mathcal{I}_2|}\{\widehat{V}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1}) - \widehat{V}_{\mathcal{I}_2}(\hat{a}_{\mathcal{I}_1})\}}{\hat{\sigma}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1}, \hat{a}_{\mathcal{I}_1})}, \frac{\sqrt{|\mathcal{I}_1|}\{\widehat{V}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2}) - \widehat{V}_{\mathcal{I}_1}(\hat{a}_{\mathcal{I}_2})\}}{\hat{\sigma}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2}, \hat{a}_{\mathcal{I}_2})} \right\},$$

where $\hat{\sigma}^2_{\mathcal{I}}(d, a)$ denotes some consistent estimator for the asymptotic variance of $\sqrt{|\mathcal{I}|}\{\widehat{V}_{\mathcal{I}}(d) - \widehat{V}_{\mathcal{I}}(a)\}$ for a given regime d and $a \in \{0, 1\}$. The null is rejected if $\widehat{T}_{CV} > z_{\alpha/2}$ for a given significance level α. Using similar arguments in Theorem 3.1 and Theorem 3.2, we can show that such a testing procedure is consistent.

6.2 Multi-stage studies

Currently, we only consider a single stage study. For multiple-stage studies, it suffices to test whether the value function under the optimal dynamic treatment regime is strictly larger than those under nondynamic treatment regimes. Zhang et al. (2013) proposed an inverse propensity-score weighted estimator for the value function under an arbitrary

dynamic treatment regime. Denoted by $\widehat{\mathrm{VD}}_{\mathcal{I}}(d_1, d_2)$ the corresponding estimator for the value difference between two dynamic treatment regimes d_1 and d_2, and $\hat{d}_{\mathcal{I}}$ the estimated optimal dynamic treatment regime, based on the sub-dataset $\mathcal{I}$. Consider the following test statistic:

$$\widehat{T}_{CV} = \max\left\{ \min_{d\in\mathcal{D}_{nd}} \frac{\sqrt{|\mathcal{I}_2|}\widehat{\mathrm{VD}}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1}, d)}{\hat{\sigma}_{\mathcal{I}_2}(\hat{d}_{\mathcal{I}_1}, d)}, \min_{d\in\mathcal{D}_{nd}} \frac{\sqrt{|\mathcal{I}_1|}\widehat{\mathrm{VD}}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2}, d)}{\hat{\sigma}_{\mathcal{I}_1}(\hat{d}_{\mathcal{I}_2}, d)} \right\},$$

where $\mathcal{I}_1$ and $\mathcal{I}_2$ stand for a random partition of the dataset, $\hat{\sigma}^2_{\mathcal{I}}(d_1, d_2)$ some consistent estimator for the asymptotic variance of $\sqrt{|\mathcal{I}|}\widehat{\mathrm{VD}}_{\mathcal{I}}(d_1, d_2)$ and $\mathcal{D}_{nd}$ denotes the set of non-dynamic treatment regimes.

Note that for $j = 1, 2$, we have that under the null,

$$\min_{d\in\mathcal{D}_{nd}} \frac{\sqrt{|\mathcal{I}_j|}\widehat{\mathrm{VD}}_{\mathcal{I}_j}(\hat{d}_{\mathcal{I}_j^c}, d)}{\hat{\sigma}_{\mathcal{I}_j}(\hat{d}_{\mathcal{I}_j^c}, d)} \leq \min_{d\in\mathcal{D}_{nd}} \frac{\sqrt{|\mathcal{I}_j|}\{\widehat{\mathrm{VD}}_{\mathcal{I}_j}(\hat{d}_{\mathcal{I}_j^c}, d) - \mathrm{VD}(\hat{d}_{\mathcal{I}_j^c}, d)\}}{\hat{\sigma}_{\mathcal{I}_j}(\hat{d}_{\mathcal{I}_j^c}, d)} \xrightarrow{L} \min_{d\in\mathcal{D}_{nd}} Z_d, \qquad (18)$$

where $\mathrm{VD}(d_1, d_2) = \mathrm{E}\widehat{\mathrm{VD}}_{\mathcal{I}}(d_1, d_2)$ and $\{Z_d\}_{d\in\mathcal{D}_{nd}}$ is a set of mean zero Gaussian random variables whose covariance matrix can be consistently estimated from data. For a given significance level α, we reject the null if $\widehat{T}_{CV} > \hat{c}_{\alpha/2}$ where $\hat{c}_\alpha$ corresponds to some consistent estimator for $\Pr(\min_{d\in\mathcal{D}_{nd}} Z_d > z_\alpha)$. It follows from the Bonferroni's inequality and (18) that the type-I error of $\widehat{T}_{CV}$ is well-controlled. In the high-dimensional setting, we can calculate $\widehat{T}_{CV}$ based on sparse random projections of the covariate space. Details are omitted for brevity.

Acknowledgment

The authors are grateful for helpful feedback from the Associate Editor and anonymous referees, which lead to significant improvement of this work.

A Variance estimator in Section 3.3.1

Define $\hat{\alpha}_{\mathcal{I}}$ to be the penalized logistic regression estimator based on $\{(X_i, A_i)\}_{i\in\mathcal{I}}$, $\hat{\theta}_{0,\mathcal{I}}$ and $\hat{\theta}_{1,\mathcal{I}}$ to be the penalized linear regression estimators based on $\{(X_i, Y_i)\}_{i\in\mathcal{I},A_i=0}$ and $\{(X_i, Y_i)\}_{i\in\mathcal{I},A_i=1}$ respectively. Denoted by $M_{\alpha,\mathcal{I}}$ the support of $\hat{\alpha}_{\mathcal{I}}$, i.e, $M_{\alpha,\mathcal{I}} = \{j = 1,\ldots,p : \hat{\alpha}_{\mathcal{I},j} \neq 0\}$. Similarly define $M_{\theta_0,\mathcal{I}}$ and $M_{\theta_1,\mathcal{I}}$ to be the supports of $\hat{\theta}_{0,\mathcal{I}}$ and $\hat{\theta}_{1,\mathcal{I}}$ respectively. Let

$$\hat{\pi}_i = \frac{\exp(X_i^T\hat{\alpha}_{\mathcal{I}})}{1+\exp(X_i^T\hat{\alpha}_{\mathcal{I}})},$$

For any treatment regime d, we define

$$\hat{\sigma}^2_{DR,\mathcal{I}}(d) = \frac{1}{|\mathcal{I}|-1}\sum_{i\in\mathcal{I}}\kappa_i^2 - \frac{1}{|\mathcal{I}|(|\mathcal{I}|-1)}\left(\sum_{i\in\mathcal{I}}\kappa_i\right)^2,$$

where

$$\begin{aligned}
\kappa_i &= \left\{\left(\frac{1-A_i}{1-\hat{\pi}_i} - \frac{A_i}{\hat{\pi}_i}\right)Y_i - \left(\frac{1-A_i}{1-\hat{\pi}_i} - 1\right)X_i^T\hat{\theta}_{0,\mathcal{I}} + \left(\frac{A_i}{\hat{\pi}_i} - 1\right)X_i^T\hat{\theta}_{1,\mathcal{I}}\right\}\{1-d(X_i)\} \\
&+ \bar{I}_1^T\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}X^T_{iM_{\alpha,\mathcal{I}}}\hat{\pi}_i(1-\hat{\pi}_i)X_{iM_{\alpha,\mathcal{I}}}\right)^{-1}X_{iM_{\alpha,\mathcal{I}}}(A_i-\hat{\pi}_i) \\
&- \bar{I}_2^T\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}(1-A_i)X^T_{iM_{\theta_0,\mathcal{I}}}X_{iM_{\theta_0,\mathcal{I}}}\right)^{-1}X_{iM_{\theta_0,\mathcal{I}}}(1-A_i)(Y_i - X_i^T\hat{\theta}_0) \\
&+ \bar{I}_3^T\left(\frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}A_iX^T_{iM_{\theta_0,\mathcal{I}}}X_{iM_{\theta_0,\mathcal{I}}}\right)^{-1}X_{iM_{\theta_0,\mathcal{I}}}A_i(Y_i - X_i^T\hat{\theta}_1),
\end{aligned}$$

and $\bar{I}_j = \sum_{i\in\mathcal{I}} I_{i,j}/n$ where

$$\begin{aligned}
I_{i,1} &= \left\{\frac{\hat{\pi}_i(1-A_i)}{1-\hat{\pi}_i}\{Y_i - X_i^T\hat{\theta}_{0,\mathcal{I}}\} + \frac{A_i(1-\hat{\pi}_i)}{\hat{\pi}_i}\{Y_i - \hat{\theta}_{1,\mathcal{I}}\}\right\}X_{iM_{\alpha,\mathcal{I}}}\{1-d(X_i)\}, \\
I_{i,2} &= \left(\frac{1-A_i}{1-\hat{\pi}_i} - 1\right)X_{iM_{\theta_0,\mathcal{I}}}\{1-d(X_i)\}, \qquad I_{i,3} = \left(\frac{A_i}{\hat{\pi}_i} - 1\right)X_{iM_{\theta_1,\mathcal{I}}}\{1-d(X_i)\}.
\end{aligned}$$

B Technical conditions

(C1.) Assume there exist some positive constants γ and δ_0 such that

$$\Pr\{0 < |\tau(X)| \leq t\} = O(t^{\gamma}),$$

where the big-O term is uniform in $0 < t < \delta_0$.

(C2.) Assume $\hat{\tau}$ satisfies

$$\mathrm{E}|\hat{\tau}_{\mathcal{I}}(X) - \tau(X)|^2 = o\left(|\mathcal{I}|^{-(2+\gamma)/(2+2\gamma)}\right) \qquad \text{as } |\mathcal{I}| \to \infty,$$

where the little-o term is uniform in the training samples $\mathcal{I}$.

Condition (C1) is closely related to the margin assumption (Tsybakov, 2004; Audibert and Tsybakov, 2007) in the classification literature. It is often used to obtain sharp upper bounds on the difference between the value function under d^{opt} and that under an estimated OITR (Qian and Murphy, 2011; Luedtke and van der Laan, 2016). The larger the structure parameter γ in (C1), the sharper the upper bounds. When $\tau(X)$ has a bounded density function near 0, (C1) holds with $\gamma = 1$. If there exists some $\delta_0 > 0$ such that $|\tau(X)| \geq \delta_0$ almost surely, then (C1) holds with $\gamma = +\infty$.

Condition (C2) depends on the "structural" parameter γ in (C1) and the convergence rates of the estimated contrast function. The larger the γ, the more likely (C2) holds. When $\gamma = 1$, (C2) requires $\mathrm{E}|\hat{\tau}_{\mathcal{I}}(X) - \tau(X)|^2 = o(|\mathcal{I}|^{-3/4})$. The rates of convergence of the estimated contrast function are available for most often used machine learning or statistical methods, such as spline methods (Zhou et al., 1998), kernel ridge regression (Steinwart and Christmann, 2008; Zhang et al., 2013) and random forests (Biau, 2012). In Section C.1 of the Appendix, we show (C2) holds when $\hat{\tau}$ is computed by some of the aforementioned methods. Combining (C1) together with (C2) gives $V(\hat{d}_{\mathcal{I}}) = V(d^{opt}) + o_p(|\mathcal{I}|^{-1/2})$.

References

Audibert, J.-Y. and A. B. Tsybakov (2007). Fast learning rates for plug-in classifiers. *Ann. Statist. 35*(2), 608–633.

Baker, S. G., N. R. Cook, A. Vickers, and B. S. Kramer (2009). Using relative utility curves to evaluate risk prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 172*(4), 729748.

Belloni, A., V. Chernozhukov, D. Chetverikov, and K. Kato (2015). Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics 186*(2), 345–366.

Biau, G. (2012). J. mach. learn. res. *Journal of Machine Learning Research 13*, 1063–1095.

Cannings, T. I. and R. J. Samworth (2015). Random projection ensemble classification. *arXiv preprint arXiv:1504.04595*.

Chakraborty, B., S. Murphy, and V. Strecher (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Stat. Methods Med. Res. 19*(3), 317–343.

Chang, M., S. Lee, and Y.-J. Whang (2015). Nonparametric tests of conditional treatment effects with an application to single-sex schooling on academic achievements. *Econom. J. 18*(3), 307–346.

Fan, A., W. Lu, and R. Song (2016). Sequential advantage selection for optimal treatment regime. *Ann. Appl. Stat. 10*(1), 32–53.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc. 96*(456), 1348–1360.

Gail, M. H. (2009). Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *Journal of the National Cancer Institute 101*(13), 959–963.

Galassi, M., J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, F. Rossi, and R. Ulerich (2015). *GNU Scientific Library Reference Manual (Version 2.1)*.

Gunter, L., J. Zhu, and S. A. Murphy (2011). Variable selection for qualitative interactions. *Stat. Methodol. 8*(1), 42–55.

Hsu, Y.-C. (2017). Consistent tests for conditional treatment effects. *The Econometrics Journal 20*(1), 1–22.

Huang, Y., E. B. Laber, and H. Janes (2015). Characterizing expected benefits of biomarkers in treatment selection. *Biostatistics 16*(2), 383–399.

Johnson, W. B. and J. Lindenstrauss (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics 26*(189-206), 1.

Keller, M. B., J. P. McCullough, D. N. Klein, B. Arnow, D. L. Dunner, A. J. Gelenberg, J. C. Markowitz, C. B. Nemeroff, J. M. Russell, M. E. Thase, et al. (2000). A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. *New England Journal of Medicine 342*(20), 1462–1470.

Li, P., T. J. Hastie, and K. W. Church (2006). Very sparse random projections. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 287–296. ACM.

Lopes, M., L. Jacob, and M. J. Wainwright (2011). A more powerful two-sample test in high dimensions using random projection. In *Advances in Neural Information Processing Systems*, pp. 1206–1214.

Luedtke, A. R. and M. J. van der Laan (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist. 44*(2), 713–742.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *J. R. Stat. Soc. Ser. B Stat. Methodol. 65*(2), 331–366.

Nelson, J. and H. L. Nguyên (2013). Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*, pp. 117–126. IEEE.

Omidiran, D. and M. J. Wainwright (2010). High-dimensional variable selection with sparse random projections: measurement sparsity and statistical efficiency. *Journal of Machine Learning Research 11*(Aug), 2361–2386.

Qian, M. and S. A. Murphy (2011). Performance guarantees for individualized treatment rules. *Ann. Statist. 39*(2), 1180–1210.

Robins, J., M. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology. *Epidemiol. 11*, 550–560.

Shi, C., A. Fan, R. Song, and W. Lu (2016). High-dimensional a-learning for optimal dynamic treatment regimes. *Annals of Statistics accepted*.

Steinwart, I. and A. Christmann (2008). *Support vector machines*. Information Science and Statistics. Springer, New York.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist. 32*(1), 135–166.

van de Geer, S., P. Bühlmann, Y. Ritov, and R. Dezeure (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist. 42*(3), 1166–1202.

Watkins, C. and P. Dayan (1992). Q-learning. *Mach. Learn. 8*, 279–292.

Zhang, B., A. A. Tsiatis, E. B. Laber, and M. Davidian (2012). A robust method for estimating optimal treatment regimes. *Biometrics 68*(4), 1010–1018.

Zhang, C.-H. and S. S. Zhang (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 76*(1), 217–242.

Zhang, X. and G. Cheng (2017). Simultaneous inference for high-dimensional linear models. *J. Amer. Statist. Assoc. 112*(518), 757–768.

Zhang, Y., J. Duchi, and M. Wainwright (2013). Divide and conquer kernel ridge regression. In *Conference on Learning Theory*, pp. 592–617.

Zhang, Y., E. B. Laber, A. Tsiatis, and M. Davidian (2015). Using decision lists to construct interpretable and parsimonious treatment regimes. *Biometrics 71*(4), 895–904.

Zhao, Y., D. Zeng, A. J. Rush, and M. R. Kosorok (2012). Estimating individualized treatment rules using outcome weighted learning. *J. Amer. Statist. Assoc. 107*(499), 1106–1118.

Zhou, S., X. Shen, and D. A. Wolfe (1998). Local asymptotics for regression splines and confidence regions. *Ann. Statist. 26*(5), 1760–1782.