Published in final edited form as:
Ann Stat. 2011 April 1; 39(2): 1180–1210. doi:10.1214/10-AOS864.

PERFORMANCE GUARANTEES FOR INDIVIDUALIZED TREATMENT RULES

Min Qian[*] and **Susan A. Murphy**[*]
Department of Statistics, University of Michigan, Ann Arbor, MI, 48109

Min Qian: minqian@umich.edu; Susan A. Murphy: samurphy@umich.edu

Abstract

Because many illnesses show heterogeneous response to treatment, there is increasing interest in individualizing treatment to patients [11]. An *individualized treatment rule* is a decision rule that recommends treatment according to patient characteristics. We consider the use of clinical trial data in the construction of an individualized treatment rule leading to highest mean response. This is a difficult computational problem because the objective function is the expectation of a weighted indicator function that is non-concave in the parameters. Furthermore there are frequently many pretreatment variables that may or may not be useful in constructing an optimal individualized treatment rule yet cost and interpretability considerations imply that only a few variables should be used by the individualized treatment rule. To address these challenges we consider estimation based on l_1 penalized least squares. This approach is justified via a finite sample upper bound on the difference between the mean response due to the estimated individualized treatment rule and the mean response due to the optimal individualized treatment rule.

Keywords and phrases

decision making; l_1 penalized least squares; Value

1. Introduction

Many illnesses show heterogeneous response to treatment. For example, a study on schizophrenia [12] found that patients who take the same antipsychotic (olanzapine) may have very different responses. Some may have to discontinue the treatment due to serious adverse events and/or acutely worsened symptoms, while others may experience few if any adverse events and have improved clinical outcomes. Results of this type have motivated researchers to advocate the individualization of treatment to each patient [16,24,11]. One step in this direction is to estimate each patient's risk level and then match treatment to risk category [5,6]. However, this approach is best used to decide whether to treat; otherwise it assumes the knowledge of the best treatment for each risk category. Alternately, there is an abundance of literature focusing on predicting each patient's prognosis under a particular treatment [10,28]. Thus an obvious way to individualize treatment is to recommend the treatment achieving the best predicted prognosis for that patient. In general the goal is to use data to construct individualized treatment rules that, if implemented in future, will optimize the mean response.

[*]Supported by NIH grants R01 MH080015 and P50 DA10075.

Consider data from a single stage randomized trial involving several active treatments. A first natural procedure to construct the optimal individualized treatment rule is to maximize an empirical version of the mean response over a class of treatment rules (assuming larger responses are preferred). As will be seen, this maximization is computationally difficult because the mean response of a treatment rule is the expectation of a weighted indicator that is non-continuous and non-concave in the parameters. To address this challenge we make a substitution. That is, instead of directly maximizing the empirical mean response to estimate the treatment rule, we use a two-step procedure that first estimates a conditional mean and then from this estimated conditional mean derives the estimated treatment rule. As will be seen in Section 3, even if the optimal treatment rule is contained in the space of treatment rules considered by the substitute two-step procedure, the estimator derived from the two-step procedure may not be consistent. However if the conditional mean is modeled correctly, then the two-step procedure consistently estimates the optimal individualized treatment rule. This motivates consideration of rich conditional mean models with many unknown parameters. Furthermore there are frequently many pretreatment variables that may or may not be useful in constructing an optimal individualized treatment rule, yet cost and interpretability considerations imply that fewer rather than more variables should be used by the treatment rule. This consideration motivates the use of l_1 penalized least squares (l_1-PLS).

We propose to estimate an optimal individualized treatment rule using a two step procedure that first estimates the conditional mean response using l_1-PLS with a rich linear model and second, derives the estimated treatment rule from estimated conditional mean. For brevity, throughout, we call the two step procedure the l_1-PLS method. We derive several finite sample upper bounds on the difference between the mean response to the optimal treatment rule and the mean response to the estimated treatment rule. All of the upper bounds hold even if our linear model for the conditional mean response is incorrect and to our knowledge are, up to constants, the best available. We use the upper bounds in Section 3 to illuminate the potential mismatch between using least squares in the two-step procedure and the goal of maximizing mean response. The upper bounds in Section 4.1 involve a minimized sum of the approximation error and estimation error; both errors result from the estimation of the conditional mean response. We shall see that l_1-PLS estimates a linear model that minimizes this approximation plus estimation error sum among a set of suitably sparse linear models.

If the part of the model for the conditional mean involving the treatment effect is correct, then the upper bounds imply that, although a surrogate two-step procedure is used, the estimated treatment rule is consistent. The upper bounds provide a convergence rate as well. Furthermore in this setting the upper bounds can be used to inform how to choose the tuning parameter involved in the l_1-penalty to achieve the best rate of convergence. As a by-product, this paper also contributes to existing literature on l_1-PLS by providing a finite sample prediction error bound for the l_1-PLS estimator in the random design setting without assuming the model class contains or is close to the true model.

The paper is organized as follows. In Section 2, we formulate the decision making problem. In Section 3, for any given decision, e.g. individualized treatment rule, we relate the reduction in mean response to the excess prediction error. In Section 4, we estimate an optimal individualized treatment rule via l_1-PLS and provide a finite sample upper bound on the maximal reduction in optimal mean response achieved by the estimated rule. In Section 5, we consider a data dependent tuning parameter selection criterion. This method is evaluated using simulation studies and illustrated with data from the Nefazodone-CBASP trial [13]. Discussions and future work are presented in Section 6.

2. Individualized treatment rules

We use upper case letters to denote random variables and lower case letters to denote values of the random variables. Consider data from a randomized trial. On each subject we have the pretreatment variables $X \in \mathcal{X}$, treatment A taking values in a finite, discrete treatment space $\mathcal{A}$, and a real-valued response R (assuming large values are desirable). An *individualized treatment rule* (ITR) d is a deterministic decision rule from $\mathcal{X}$ into the treatment space $\mathcal{A}$.

Denote the distribution of (X, A, R) by P. This is the distribution of the clinical trial data; in particular, denote the known randomization distribution of A given X by $p(\cdot|X)$. The likelihood of (X, A, R) under P is then $f_0(x)p(a|x)f_1(r|x, a)$, where f_0 is the unknown density of X and f_1 is the unknown density of R conditional on (X, A). Denote the expectations with respect to the distribution P by an E. For any ITR $d : \mathcal{X} \to \mathcal{A}$, let P^d denote the distribution of (X, A, R) in which d is used to assign treatments. Then the likelihood of (X, A, R) under P^d is $f_0(x)1_{a=d(x)}f_1(r|x, a)$. Denote expectations with respect to the distribution P^d by an E^d. The *Value* of d is defined as $V(d) = E^d(R)$. An *optimal ITR*, d_0, is a rule that has the maximal Value, i.e.

$$d_0 \in \arg\max_d V(d),$$

where the argmax is over all possible decision rules. The Value of d_0, $V(d_0)$, is the *optimal Value*.

Assume $P[p(a|X) > 0] = 1$ for all $a \in \mathcal{A}$ (i.e. all treatments in $\mathcal{A}$ are possible for all values of X a.s.). Then P^d is absolutely continuous with respect to P and a version of the Radon-Nikodym derivative is $dP^d/dP = 1_{a=d(x)}/p(a|x)$. Thus the Value of d satisfies

$$V(d)=E^d(R)=\int R dP^d=\int R\frac{dP^d}{dP}dP=E\left[\frac{1_{A=d(X)}}{p(A|X)}R\right]. \tag{2.1}$$

Our goal is to estimate d_0, i.e. the ITR that maximizes (2.1), using data from distribution P. When X is low dimensional and the best rule within a simple class of ITRs is desired, empirical versions of the Value can be used to construct estimators [21,27]. However if the best rule within a larger class of ITRs is of interest, these approaches are no longer feasible.

Define $Q_0(X,A) \triangleq E(R|X,A)$ ($Q_0(X,A)$ is sometimes called the "Quality" of treatment a at observation x). It follows from (2.1) that for any ITR d,

$$V(d)=E\left[\frac{1_{A=d(X)}}{p(A|X)}Q_0(X,A)\right]=E\left[\sum_{a\in\mathcal{A}}1_{d(X)=a}Q_0(X,a)\right]=E[\,Q_0(X,d(X))].$$

Thus $V(d_0) = E[Q_0(X, d_0(X))] \leq E[\max_{a\in\mathcal{A}} Q_0(X, a)]$. On the other hand, by the definition of d_0,

$$V(d_0) \geq V(d)\Big|_{d(X)\in\arg\max_{a\in\mathcal{A}}Q_0(X,a)}=E\left[\max_{a\in\mathcal{A}}Q_0(X,a)\right].$$

Hence an optimal ITR satisfies $d_0(X) \in \arg\max_{a\in\mathcal{A}} Q_0(X, a)$ a.s.

3. Relating the reduction in Value to excess prediction error

The above argument indicates that the estimated ITR will be of high quality (i.e. have high Value) if we can estimate Q_0 accurately. In this section, we justify this by providing a quantitative relationship between the Value and the prediction error.

Because $\mathcal{A}$ is a finite, discrete treatment space, given any ITR, d, there exists a square integrable function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ for which $d(X) \in \arg\max_a Q(X, a)$ a.s. Let $L(Q) \triangleq E[R - Q(X,A)]^2$ denote the prediction error of Q (also called the mean quadratic loss). Suppose that Q_0 is square integrable and that the randomization probability satisfies $p(a|x) \geq S^{-1}$ for an $S > 0$ and all (x, a) pairs. Murphy [23] showed that

$$V(d_0) - V(d) \leq 2S^{1/2}[L(Q) - L(Q_0)]^{1/2}. \tag{3.1}$$

Intuitively, this upper bound means that if the excess prediction error of Q (i.e. $E(R - Q)^2 - E(R - Q_0)^2$) is small, then the reduction in Value of the associated ITR d (i.e. $V(d_0) - V(d)$) is small. Furthermore the upper bound provides a rate of convergence for an estimated ITR. For example, suppose Q_0 is linear, that is $Q_0 = \Phi(X, A)\boldsymbol{\theta}_0$ for a given vector-valued basis function Φ on $\mathcal{X} \times \mathcal{A}$ and an unknown parameter θ_0. And Suppose we use a correct linear model for Q_0 (here "linear" means linear in parameters), say the model $\mathcal{Q} = \{\Phi(X, A)\boldsymbol{\theta} : \boldsymbol{\theta} \to \mathbb{R}^{dim(\Phi)}\}$ or a linear model containing $\mathcal{Q}$ with dimension of parameters fixed in n. If we estimate $\boldsymbol{\theta}$ by least squares and denote the estimator by $\hat{\boldsymbol{\theta}}$, then the prediction error of $\hat{Q} = \Phi\hat{\boldsymbol{\theta}}$ converges to $L(Q_0)$ at rate $1/n$ under mild regularity conditions. This together with inequality (3.1) implies that the Value obtained by the estimated ITR, $\hat{d}(X) \in \arg\max_a \hat{Q}(X, a)$, will converge to the optimal Value at rate at least $1/\sqrt{n}$.

In the following theorem, we improve this upper bound in two aspects. First, we show that an upper bound with exponent larger than 1/2 can be obtained under a margin condition, which implicitly implies a faster rate of convergence. Second, it turns out that the upper bound need only depend on one term in the function Q; we call this the treatment effect term, T. For any square integrable Q, the associated treatment effect term is defined as $T(X,A) \triangleq Q(X,A) - E[Q(X,A)|X]$. Note that $d(X) \in \arg\max_a T(X, a) = \arg\max_a Q(X, a)$ a.s. Similarly, the true treatment effect term is given by

$$T_0(X, A) \triangleq Q_0(X, A) - E[Q_0(X, A)|X]. \tag{3.2}$$

$T_0(x, a)$ is the centered effect of treatment $A = a$ at observation $X = x$; $d_0(X) \in \arg\max_a T_0(X, a)$.

Theorem 3.1—Suppose p(a|x) ≥ S^{-1} for a positive constant S for all (x, a) pairs. Assume there exists some constants C > 0 and α ≥ 0 such that

$$\mathbf{P}\left(\max_{a\in\mathcal{A}} T_0(X, a) - \max_{a\in\mathcal{A}\setminus \arg\max_{a\in\mathcal{A}} T_0(X,a)} T_0(X, a) \leq \varepsilon\right) \leq C\varepsilon^{\alpha} \tag{3.3}$$

for all positive ε. Then for any ITR d : $\mathcal{X} \to \mathcal{A}$ and square integrable function Q : $\mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that d(X) ∈ arg $\max_{a \in \mathcal{A}}$ Q(X, a) a.s., we have

$$V(d_0) - V(d) \le C'[L(Q) - L(Q_0)]^{(1+\alpha)/(2+\alpha)}, \tag{3.4}$$

and

$$V(d_0) - V(d) \le C'\left[E(T(X,A) - T_0(X,A))^2\right]^{(1+\alpha)/(2+\alpha)}. \tag{3.5}$$

where $C' = (2^{2+3\alpha}S^{1+\alpha}C)^{1/(2+\alpha)}$.

The proof of Theorem 3.1 is in Appendix A.1.

Remarks

1. We set the second maximum in (3.3) to $-\infty$ if for an x, $T_0(x, a)$ is constant in a and thus the set $\mathcal{A} \backslash \arg\max_{a \in \mathcal{A}} T_0(x, a) = \varnothing$.

2. Condition (3.3) is similar to the margin condition in classification [25,18,32]; in classification this assumption is often used to obtain sharp upper bounds on the excess 0 − 1 risk in terms of other surrogate risks [1]. Here
$$\max_{a \in \mathcal{A}} T_0(x,a) - \max_{a \in \mathcal{A} \backslash \arg\max_{a \in \mathcal{A}} T_0(x,a)} T_0(x,a)$$
can be viewed as the "margin" of T_0 at observation $X = x$. It measures the difference in mean responses between the optimal treatment(s) and the best suboptimal treatment(s) at x. For example, suppose $X \sim U[-1, 1]$, $P(A = 1|X) = P(A = -1|X) = 1/2$ and $T_0(X, A) = XA$. Then the margin condition holds with $C = 1/2$ and $\alpha = 1$. Note the margin condition does not exclude multiple optimal treatments for any observation x. However, when α > 0, it does exclude suboptimal treatments that yield a conditional mean response very close to the largest conditional mean response for a set of x with nonzero probability.

3. For $C = 1$, $\alpha = 0$, Condition (3.3) always holds for all $\varepsilon > 0$; in this case (3.4) reduces to (3.1).

4. The larger the α, the larger the exponent $(1 + \alpha)/(2 + \alpha)$ and thus the stronger the upper bounds in (3.4) and (3.5). However the margin condition is unlikely to hold for all ε if α is very large. An alternate margin condition and upper bound are as follows.

   Suppose $p(a|x) \ge S^{-1}$ *for all* (x, a) *pairs. Assume there is an* $\varepsilon > 0$, *such that*

$$\mathbf{P}\left(\max_{a \in \mathcal{A}} T_0(X,a) - \max_{a \in \mathcal{A} \backslash \arg\max_{a \in \mathcal{A}} T_0(X,a)} T_0(X,a) < \varepsilon\right) = 0. \tag{3.6}$$

   Then $V(d_0) - V(d) \le 4S[L(Q) - L(Q_0)]/\varepsilon$ *and* $V(d_0) - V(d) \le 4SE(T - T_0)^2/\varepsilon$.

   The proof is essentially the same as that of Theorem 3.1 and is omitted. Condition (3.6) means that T_0 evaluated at the optimal treatment(s) minus T_0 evaluated at the best suboptimal treatment(s) is bounded below by a positive constant for almost all X observations. If X assumes only a finite number of values, then this condition always holds, because we can take ε to be the smallest difference in T_0 when

evaluated at the optimal treatment(s) and the suboptimal treatment(s) (note that if $T_0(x, a)$ is constant for all $a \in \mathcal{A}$ for some observation $X = x$, then all treatments are optimal for that observation).

5. Inequality (3.5) cannot be improved in the sense that choosing $T = T_0$ yields zero on both sides of the inequality. Moreover an inequality in the opposite direction is not possible, since each ITR is associated with many non-trivial T-functions. For example, suppose $X \sim U[-1, 1]$, $P(A = 1|X) = P(A = -1|X) = 1/2$ and $T_0(X, A) = (X - 1/3)^2A$. The optimal ITR is $d_0(X) = 1$ a.s. Consider $T(X, A) = \theta A$. Then maximizing $T(X, A)$ yields the optimal ITR as long as $\theta > 0$. This means that the left hand side (LHS) of (3.5) is zero, while the right hand side (RHS) is always positive no matter what value θ takes.

Theorem 3.1 supports the approach of minimizing the estimated prediction error to estimate Q_0 or T_0 and then maximizing this estimator over $a \in \mathcal{A}$ to obtain an ITR. It is natural to expect that even when the approximation space used in estimating Q_0 or T_0 does not contain the truth, this approach will provide the best (highest Value) of the considered ITRs. Unfortunately this does not occur due to the mismatch between the loss functions (weighted 0–1 loss and the quadratic loss). This mismatch is indicated by remark 5 above. More precisely, note that the approximation space, say $\mathcal{Q}$ for Q_0, places implicit restrictions on the class of ITRs that will be considered. In effect the class of ITRs is $\mathcal{D}_{\mathcal{Q}} = \{d(X) \in \arg\max_a Q(X, a) : Q \in \mathcal{Q}\}$. It turns out that minimizing the prediction error may not result in the ITR in $\mathcal{D}_{\mathcal{Q}}$ that maximizes the Value. This occurs when the approximation space $\mathcal{Q}$ does not provide a treatment effect term close to the treatment effect term in Q_0. In the following toy example, the optimal ITR d_0 belongs to $\mathcal{D}_{\mathcal{Q}}$, yet the prediction error minimizer over $\mathcal{Q}$ does not yield d_0.

A toy example

Suppose X is uniformly distributed in $[-1, 1]$, A is binary $\{-1, 1\}$ with probability 1/2 each and is independent of X, and R is normally distributed with mean $Q_0(X, A) = (X - 1/3)^2A$ and variance 1. It is easy to see that the optimal ITR satisfies $d_0(X) = 1$ a.s. and $V(d_0) = 4/9$. Consider approximation space $\mathcal{Q} = \{Q(X, A; \boldsymbol{\theta}) = (1, X, A, XA)\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^4\}$ for Q_0. Thus the space of ITRs under consideration is $\mathcal{D}_{\mathcal{Q}} = \{d(X) = sign(\theta_3+\theta_4X) : \theta_3, \theta_4 \in \mathbb{R}\}$. Note that $d_0 \in \mathcal{D}_{\mathcal{Q}}$ since $d_0(X)$ can be written as $sign(\theta_3 + \theta_4X)$ for any $\theta_3 > 0$ and $\theta_4 = 0$. d_0 is the best treatment rule in $\mathcal{D}_{\mathcal{Q}}$. However, minimizing the prediction error $L(Q)$ over $\mathcal{Q}$ yields $Q^*(X, A) = (4/9-2/3X)A$. The ITR associated with Q^* is $d^*(X) = \arg\max_{a\in\{-1,1\}} Q^*(X, a) = sign(2/3 - X)$, which has lower Value than d_0 ($V(d^*) = E\left[\frac{1_{A(2/3-X)>0}R}{1/2}\right] = 29/81 < V(d_0)$).

4. Estimation via l_1-penalized least squares

To deal with the mismatch between minimizing the prediction error and maximizing the Value discussed in the prior section, we consider a large linear approximation space $\mathcal{Q}$ for Q_0. Since overfitting is likely (due to the potentially large number of pretreatment variables and/or large approximation space for Q_0) we use penalized least squares (see Section S.1 of the supplementary material for further discussion of the overfitting problem). Furthermore we use l_1 penalized least squares (l_1-PLS, [31]) as the l_1 penalty does some variable selection and as a result will lead to ITRs that are cheaper to implement (fewer variables to collect per patient) and easier to interpret. See Section 6 for the discussion of other potential penalization methods.

Let $\{(X_i, A_i, R_i)\}_{i=1}^n$ represent i.i.d. observations on n subjects in a randomized trial. For convenience, we use E_n to denote the associated empirical expectation (i.e.

$E_n f = \sum_{i=1}^{n} f(X_i, A_i, R_i)/n$ for any real-valued function f on $\mathcal{X} \times \mathcal{A} \times \mathbb{R}$). Let $\mathcal{Q} := \{Q(X, A; \boldsymbol{\theta}) = \Phi(X, A)\,\boldsymbol{\theta}, \boldsymbol{\theta} \in \mathbb{R}^J\}$ be the approximation space for Q_0, where $\varphi(X, A) = (\varphi_1(X, A), \ldots, \varphi_J(X, A))$ is a 1 by J vector composed of basis functions on $\mathcal{X} \times \mathcal{A}$, $\boldsymbol{\theta}$ is a J by 1 parameter vector, and J is the number of basis functions (for clarity here J will be fixed in n, see Appendix A.2 for results with J increasing as n increases). The l_1-PLS estimator of $\boldsymbol{\theta}$ is

$$\widehat{\theta}_n = \arg\min_{\theta \in \mathbb{R}^J} \left\{ E_n[R - \Phi(X, A)\theta]^2 + \lambda_n \sum_{j=1}^{J} \widehat{\sigma}_j |\theta_j| \right\}, \tag{4.1}$$

where $\hat{\sigma}_j = [E_n \varphi_j(X, A)^2]^{1/2}$, θ_j is the j^{th} component of $\boldsymbol{\theta}$ and λ_n is a tuning parameter that controls the amount of penalization. The weights $\hat{\sigma}_j$'s are used to balance the scale of different basis functions; these weights were used in Bunea et al. [4] and van de Geer [33]. In some situations, it is natural to penalize only a subset of coefficients and/or use different weights in the penalty; see Section S.2 of the supplementary material for required modifications. The resulting estimated ITR satisfies

$$\widehat{d}_n(X) \in \arg\max_{a \in \mathcal{A}} \Phi(X, a)\widehat{\theta}_n. \tag{4.2}$$

4.1. Performance guarantee for the l_1-PLS

In this section we provide finite sample upper bounds on the difference between the optimal Value and the Value obtained by the l_1-PLS estimator in terms of the prediction errors resulting from the estimation of Q_0 and T_0. These upper bounds guarantee that if Q_0 (or T_0) is consistently estimated, the estimator of d_0 will be consistent and will inherit a rate of convergence from the rate of convergence of the estimator of Q_0 (or T_0). Perhaps more importantly, the finite sample upper bounds provided below do *not* require the assumption that either Q_0 or T_0 is consistently estimated. Thus each upper bound includes approximation error as well as estimation error. The estimation error decreases with decreasing model sparsity and increasing sample size. An "oracle" model for Q_0 (or T_0) minimizes the sum of these two errors among suitably sparse linear models (see remark 2 after Theorem 4.3 for a precise definition of the oracle model). In finite samples, the upper bounds imply that $\hat{d}_n$, the ITR produced by the l_1-PLS method, will have Value roughly as if the l_1-PLS method detects the sparsity of the oracle model and then estimates from the oracle model using ordinary least squares (see remark 3 below).

Define the prediction error minimizer $\boldsymbol{\theta}^* \in \mathbb{R}^J$ by

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^J} L(\Phi\theta) = \arg\min_{\theta \in \mathbb{R}^J} E(R - \Phi\theta)^2. \tag{4.3}$$

For expositional simplicity assume that $\boldsymbol{\theta}^*$ is unique, and define the sparsity of $\boldsymbol{\theta} \in \mathbb{R}^J$ by its l_0 norm, $\|\boldsymbol{\theta}\|_0$ (see Appendix A.2 for a more general setting, where $\boldsymbol{\theta}^*$ is not unique and a laxer definition of sparsity is used). As discussed above, for finite n, instead of estimating $\boldsymbol{\theta}^*$, the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$ estimates a parameter, θ_n^{**}, possessing small prediction error but with controlled sparsity. For any bounded function f on $\mathcal{X} \times \mathcal{A}$, let $\|f\|_\infty \triangleq \sup_{x \in \mathcal{X}, a \in \mathcal{A}} |f(x,a)|$. θ_n^{**} lies in the set of parameters Θ_n defined by

$$\Theta_n \triangleq \left\{\theta \in \mathbb{R}^J : \left\|\Phi(\theta-\theta^*)\right\|_\infty \le \eta,\ \max_{j=1,\ldots,J}\left|\frac{E[\varphi_j\Phi(\theta-\theta^*)]}{\sigma_j}\right| \le 15\eta\sqrt{\frac{\log(Jn)}{n}} \text{ and } \left\|\theta\right\|_0 \le \frac{\beta}{640U}\sqrt{\frac{n}{\log(nJ)}}\right\}, \tag{4.4}$$

where $\sigma_j=(E\varphi_j^2)^{1/2}$, and η, β and U are positive constants that will be defined in Theorem 4.1.

The first two conditions in (4.4) restrict Θ_n to $\boldsymbol{\theta}$'s with controlled distance in sup norm and with controlled distance in prediction error via first order derivatives (note that $|E[\varphi_j\Phi(\theta-\theta^*)]/\sigma_j| = |\partial L(\Phi\theta)/\partial\theta_j - \partial L(\Phi\theta^*)/\partial\theta_j^*|/2\sigma_j$). The third condition restricts Θ_n to sparse $\boldsymbol{\theta}$'s. Note that as n increases this sparsity requirement becomes laxer, ensuring that $\boldsymbol{\theta}^* \in \Theta_n$ for sufficiently large n.

When Θ_n is non-empty, θ_n^{**} is given by

$$\theta_n^{**} = \arg\min_{\theta\in\Theta_n}[L(\Phi\theta)+3\left\|\theta\right\|_0\lambda_n^2/\beta]. \tag{4.5}$$

Note that θ_n^{**} is at least as sparse as $\boldsymbol{\theta}^*$ since by (4.3), $L(\Phi\theta)+3\left\|\theta\right\|_0\lambda_n^2/\beta > L(\Phi\theta^*)+3\left\|\theta^*\right\|_0\lambda_n^2/\beta$ for any $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta}\|_0 > \|\boldsymbol{\theta}^*\|_0$.

The following theorem provides a finite sample performance guarantee for the ITR produced by l_1-PLS method. Intuitively, this result implies that if Q_0 can be well approximated by the sparse linear representation θ_n^{**} (so that both $L(\Phi\theta_n^{**}) - L(Q_0)$ and $\left\|\theta_n^{**}\right\|_0$ are small), then $\hat{d}_n$ will have Value close to the optimal Value in finite samples.

Theorem 4.1—Suppose $p(a|x) \ge S^{-1}$ for a positive constant S for all (x, a) pairs and the margin condition (3.3) holds for some $C > 0$, $\alpha \ge 0$ and all positive ε. Assume

1. the error terms $\varepsilon_i = R_i - Q_0(X_i, A_i)$, i = 1, …, n, are independent of (X_i, A_i), i …, n and are i.i.d. with $E(\varepsilon_i) = 0$ and $E[|\varepsilon_i|^l] \le \frac{l!}{2}c^{l-2}\sigma^2$ for some c, $\sigma^2 > 0$ for all $l \ge 2$;
2. there exist finite, positive constants U and η such that $\max_{j=1,\ldots,J} \|\varphi_j\|_\infty/\sigma_j \le U$ and $\|Q_0 - \Phi\boldsymbol{\theta}^*\|_\infty \le \eta$; and
3. $E[(\varphi_1/\sigma_1, \ldots, \varphi_J/\sigma_J)^T (\varphi_1/\sigma_1,\ldots, \varphi_J/\sigma_J)]$ is positive definite, and the smallest eigenvalue is denoted by β.

Consider the estimated ITR $\hat{d}_n$ defined by (4.2) with tuning parameter

$$\lambda_n \ge k\sqrt{\frac{\log(Jn)}{n}}, \tag{4.6}$$

where $k = 82\max\{c, \sigma, \eta\}$. Let Θ_n be the set defined in (4.4). Then for any $n \ge 24U^2\log(Jn)$ and for which Θ_n is non-empty, we have, with probability at least $1 - 1/n$, that

$$V(d_0) - V(\widehat{d_n}) \le C' \left[\min_{\theta \in \Theta_n} \left(L(\Phi\theta) - L(Q_0) + 3\left\|\theta\right\|_0 \lambda_n^2/\beta \right) \right]^{\frac{1+\alpha}{2+\alpha}}, \tag{4.7}$$

where $C' = (2^{2+3\alpha} S^{1+\alpha} C)^{1/(2+\alpha)}$.

The result follows from inequality (3.4) in Theorem 3.1 and inequality (4.10) in Theorem 4.3. Similar results in a more general setting can be obtained by combining (3.4) with inequality (A.7) in Appendix A.2.

Remarks

1. Note that θ_n^{**} is the minimizer of the upper bound on the RHS of (4.7) and that θ_n^{**} is contained in the set $\{\theta_n^{*,(m)}: m \subset \{1, \ldots, J\}\}$. Each $\theta_n^{*,(m)}$ satisfies $\theta_n^{*,(m)} = \arg\min_{\{\theta \in \Theta_n : \theta_j = 0 \text{ for all } j \notin m\}} L(\Phi\theta)$; that is, $\theta_n^{*,(m)}$ minimizes the prediction error of the model indexed by the set m (i.e. model $\{\Sigma_{j \in m} \varphi_j \theta_j : \theta_j \in \mathbb{R}\}$) (within Θ_n). For each $\theta_n^{*,(m)}$, the first term in the upper bound in (4.7) (i.e. $L(\Phi\theta_n^{*,(m)}) - L(Q_0)$) is the approximation error of the model indexed by m within Θ_n. As in van de Geer [33], we call the second term $3\left\|\theta_n^{*,(m)}\right\|_0 \lambda_n^2/\beta$ the estimation error of the model indexed by m. To see why, first put $\lambda_n = k\sqrt{\log(Jn)/n}$. Then, ignoring the $\log(n)$ factor, the second term is a function of the sparsity of model m relative to the sample size, n. Up to constants, the second term is a "tight" upper bound for the estimation error of the OLS estimator from model m, where "tight" means that the convergence rate in the bound is the best known rate. Note that θ_n^{**} is the parameter that minimizes the sum of the two errors over all models. Such a model (the model corresponding to θ_n^{**}) is called an oracle model. The $\log(n)$ factor in the estimation error is the price paid for not knowing the sparsity of the oracle model. By using the l_1-PLS method, we pay by a factor of $\log(n)$ in the estimation error and as an exchange, the l_1-PLS estimator behaves roughly as if it knew the sparsity of the oracle model and as if is was estimated from the oracle model using OLS. Thus the $\log(n)$ factor can be viewed as the price paid for not knowing the sparsity of the oracle model and thus having to conduct model selection. See remark 2 after Theorem A.1 for the precise definition of the oracle model and its relationship to θ_n^{**}.

2. Suppose $\lambda_n = o(1)$. Then in large samples the estimation error term $3\left\|\theta\right\|_0 \lambda_n^2/\beta$ is negligible. In this case, θ_n^{**} is close to $\boldsymbol{\theta}^*$. When the model $\Phi\boldsymbol{\theta}^*$ approximates Q_0 sufficiently well, we see that setting λ_n equal to its lower bound in (4.6) provides the fastest rate of convergence of the upper bound to zero. More precisely, suppose $Q_0 = \Phi\boldsymbol{\theta}^*$ (i.e. $L(\Phi\boldsymbol{\theta}^*) - L((Q_0) = 0$). Then inequality (4.7) implies that $V(d_0) - V(\hat{d}_n) \le O_p((\log n/n)^{(1+\alpha)/(2+\alpha)})$. A convergence in mean result is presented in Corollary 4.1.

3. In finite samples, the estimation error $3\left\|\theta\right\|_0 \lambda_n^2/\beta$ is nonnegligible. The argument of the minimum in the upper bound (4.7), θ_n^{**}, minimizes prediction error among parameters with controlled sparsity. In remark 2 after Theorem 4.3, we discuss how this upper bound is a tight upper bound for the OLS estimator from an oracle model in the step-wise model selection setting. In this sense, inequality (4.7) implies that decision rule produced by the l_1-PLS method will have a reduction in Value

roughly as if it knew the sparsity of the oracle model and were estimated from the oracle model using OLS.

4. Assumptions 1–3 in Theorem 4.1 are employed to derive the finite sample prediction error bound for the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$ defined in (4.1). Below we briefly discuss these assumptions.

Assumption 1 implicitly implies that the error terms do not have heavy tails. This condition is often assumed to show that the sample mean of a variable is concentrated around its true mean with a high probability. It is easy to verify that this assumption holds if each ε_i is bounded. Moreover, it also holds for some commonly used error distributions that have unbounded support, such as the normal or double exponential.

Assumption 2 is also used to show the concentration of the sample mean around the true mean. It is possible to replace the boundedness condition by a moment condition similar to Assumption 1. This assumption requires that all basis functions and the difference between Q_0 and its best linear approximation are bounded. Note that we do not assume $\mathcal{Q}$ to be a good approximation space for Q_0. However, if $\Phi\boldsymbol{\theta}^*$ approximates Q_0 well, η will be small, which will result in a smaller upper bound in (4.7). In fact, in the generalized result (Theorem A.1) we allow U and η to increase in n.

Assumption 3 is employed to avoid collinearity. In fact, we only need

$$E[\Phi(\theta' - \theta)]^2 \|\theta\|_0 \geq \beta \left(\sum_{j \in M_0(\theta)} \sigma_j |\theta'_j - \theta_j| \right)^2, \tag{4.8}$$

for $\boldsymbol{\theta}, \boldsymbol{\theta}'$ belonging to a subset of $\mathbb{R}^J$ (see Assumption A.3), where $M_0(\boldsymbol{\theta}) \triangleq \{j = 1, \ldots, J: \boldsymbol{\theta}_j \neq 0\}$. Condition (4.8) has been used in van de Geer [33]. This condition is also similar to the restricted eigenvalue assumption in Bickel et al. [3] in which E is replaced by E_n, and a fixed design matrix is considered. Clearly, Assumption 3 is a sufficient condition for (4.8). In addition, condition (4.8) is satisfied if the correlation $|E\varphi_j\varphi_k|/(\sigma_j\sigma_k)$ is small for all $k \in M_0(\boldsymbol{\theta})$, $j \neq k$ and a subset of $\boldsymbol{\theta}$'s (similar results in a fixed design setting have been proved in Bickel et al. [3]. The condition on correlation is also known as "mutual coherence" condition in Bunea at al. [4]). See Bickel et al. [3] for other sufficient conditions for (4.8).

The above upper bound for $V(d_0) - V(\hat{d}_n)$ involves $L(\Phi\boldsymbol{\theta}) - L(Q_0)$, which measures how well the conditional mean function Q_0 is approximated by $\mathcal{Q}$. As we have seen in Section 3, the quality of the estimated ITR only depends on the estimator of the treatment effect term T_0. Below we provide a strengthened result in the sense that the upper bound depends only on how well we approximate the treatment effect term.

First we identify terms in the linear model $\mathcal{Q}$ that approximate T_0 (recall that $T_0(X,A) \triangleq Q_0(X,A) - E[Q_0(X,A)|X]$). Without loss of generality, we rewrite the vector of basis functions as $\Phi(X, A) = (\Phi^{(1)}(X), \Phi^{(2)}(X, A))$, where $\Phi^{(1)} = (\varphi_1(X), \ldots, \varphi_{J^{(1)}}(X))$ is composed of all components in Φ that do not contain A and $\Phi^{(2)} = (\varphi_{J^{(1)}+1}(X, A), \ldots, \varphi_J(X, A))$ is composed of all components in Φ that contain A. Since A takes only finite values and the randomization distribution $p(a|x)$ is known, we can code A so that $E[\Phi^{(2)}(X, A)^T|X] = \mathbf{0}$ a.s. (see Section 5.2 and Appendix A.3 for examples). For any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)^T \in \mathbb{R}^J$, denote $\boldsymbol{\theta}^{(1)} = (\theta_1, \ldots, \theta_{J^{(1)}})^T$ and $\boldsymbol{\theta}^{(2)} = (\theta_{J^{(1)}+1}, \ldots, \theta_J)^T$. Then $\Phi^{(1)}\boldsymbol{\theta}^{(1)}$ approximates $E(Q_0(X, A)|X)$ and $\Phi^{(2)}\boldsymbol{\theta}^{(2)}$ approximates T_0.

The following theorem implies that if the treatment effect term T_0 can be well approximated by a sparse representation, then $\hat{d}_n$ will have Value close to the optimal Value.

Theorem 4.2—Suppose $p(a|x) \geq S^{-1}$ for a positive constant S for all (x, a) pairs and the margin condition (3.3) holds for some $C > 0$, $\alpha \geq 0$ and all positive ε. Assume $E[\Phi^{(2)}(X, A)^T|X] = \mathbf{0}$ a.s. Suppose Assumptions 1 – 3 in Theorem 4.1 hold. Let $\hat{d}_n$ be the estimated ITR with λ_n satisfying condition (4.6). Let Θ_n be the set defined in (4.4). Then for any $n \geq 24U^2 \log(Jn)$ and for which Θ_n is non-empty, we have, with probability at least $1 - 1/n$, that

$$V(d_0) - V(\widehat{d_n}) \leq C' \left[\min_{\theta \in \Theta_n} \left(E(\Phi^{(2)}\theta^{(2)} - T_0)^2 + 5\left\|\theta^{(2)}\right\|_0 \lambda_n^2/\beta \right) \right]^{\frac{1+\alpha}{2+\alpha}}, \tag{4.9}$$

where $C' = (2^{2+3\alpha} S^{1+\alpha} C)^{1/(2+\alpha)}$.

The result follows from inequality (3.5) in Theorem 3.1 and inequality (4.11) in Theorem 4.3.

Remarks

1. Inequality (4.9) improves inequality (4.7) in the sense that it guarantees a small reduction in Value of $\hat{d}_n$ as long as the treatment effect term T_0 is well approximated by a sparse linear representation; it does not require that the entire conditional mean function Q_0 be well approximated. In many situations Q_0 may be very complex, but T_0 could be very simple. This means that T_0 is much more likely to be well approximated as compared to Q_0 (indeed, if there is no difference between treatments, then $T_0 \equiv 0$).
2. Inequality (4.9) cannot be improved in the sense that if there is no treatment effect (i.e. $T_0 \equiv 0$), then both sides of the inequality are zero. This result implies that minimizing the penalized empirical prediction error indeed yields high Value (at least asymptotically) if T_0 can be well approximated.

The following asymptotic result follows from Theorem 4.2. Note that when $E[\Phi^{(2)}(X, A)^T|X] = \mathbf{0}$ a.s. (see Section 5 for examples), $E(\Phi\boldsymbol{\theta} - Q_0)^2 = E(\Phi^{(1)}\boldsymbol{\theta}^{(1)} - [Q_0 - E(Q_0|X)])^2 + E(\Phi^{(2)}\boldsymbol{\theta}^{(2)} - T_0)^2$. Thus the estimation of the treatment effect term T_0 is asymptotically separated from the estimation of the main effect term $Q_0 - E(Q_0|X)$. In this case, $\Phi^{(2)}\boldsymbol{\theta}^{(2),*}$ is the best linear approximation of the treatment effect term T_0, where $\boldsymbol{\theta}^{(2),*}$ is the vector of components in $\boldsymbol{\theta}^*$ corresponding to $\Phi^{(2)}$.

Corollary 4.1—Suppose $p(a|x) \geq S^{-1}$ for a positive constant S for all (x, a) pairs and the margin condition (3.3) holds for some $C > 0$, $\alpha \geq 0$ and all positive ε. Assume $E[\Phi^{(2)}(X, A)^T|X] = \mathbf{0}$ a.s. In addition, suppose Assumptions 1 – 3 in Theorem 4.1 hold. Let $\hat{d}_n$ be the estimated ITR with tuning parameter $\lambda_n = k_1 \sqrt{\log(Jn)/n}$ for a constant $k_1 \geq 82 \max\{c, \sigma, \eta\}$. If $T_0(X, A) = \Phi^{(2)}\boldsymbol{\theta}^{(2),*}$, then

$$V(d_0) - \mathbf{E}[V(\widehat{d_n})] = O((\log n/n)^{(1+\alpha)/(2+\alpha)}).$$

This result provides a guarantee on the convergence rate of $V(\hat{d}_n)$ to the optimal Value. More specifically, it means that if T_0 is correctly approximated, then the Value of $\hat{d}_n$ will

converge to the optimal Value in mean at rate at least as fast as $(\log n/n)^{(1+\alpha)/(2+\alpha)}$ with appropriate choice of λ_n.

4.2. Prediction error bound for the l_1-PLS estimator

In this section we provide a finite sample upper bound for the prediction error of the l_1-PLS estimator $\hat{\boldsymbol{\theta}}_n$. This result is needed to prove Theorem 4.1. Furthermore this result strengthens existing literature on l_1-PLS method in prediction. Finite sample prediction error bounds for the l_1-PLS estimator in the random design setting have been provided in Bunea et al. [4] for quadratic loss, van de Geer [33] mainly for Lipschitz loss, and Koltchinskii [15] for a variety of loss functions. With regards quadratic loss, Koltchinskii [15] requires the response Y is bounded, while both Bunea et al. [4] and van de Geer [33] assumed the existence of a sparse $\boldsymbol{\theta} \in \mathbb{R}^J$ such that $E(\Phi\boldsymbol{\theta} - Q_0)^2$ is upper bounded by a quantity that decreases to 0 at a certain rate as $n \to \infty$ (by permitting J to increase with n so Φ depends on n as well). We improve the results in the sense that we do not make such assumptions (see Appendix A.2 for results when Φ, J are indexed by n and J increases with n).

As in the prior sections, the sparsity of $\boldsymbol{\theta}$ is measured by its l_0 norm, $\|\boldsymbol{\theta}\|_0$ (see the Appendix A.2 for proofs with a laxer definition of sparsity). Recall that the parameter, θ_n^{**} defined in (4.5) has small prediction error and controlled sparsity.

Theorem 4.3—Suppose Assumptions 1–3 in Theorem 4.1 hold. For any $\eta_1 \geq 0$, Let $\hat{\boldsymbol{\theta}}_n$ be the l_1-PLS estimator defined by (4.1) with tuning parameter λ_n satisfying condition (4.6). Let Θ_n be the set defined in (4.4). Then for any $n \geq 24U^2 \log(Jn)$ and for which Θ_n is non-empty, we have, with probability at least $1 - 1/n$, that

$$L(\Phi\widehat{\theta}_n) \leq \min_{\theta \in \Theta_n} \left(L(\Phi\theta) + 3\left\|\theta\right\|_0 \lambda_n^2/\beta \right) = L(\Phi\theta_n^{**}) + 3\left\|\theta_n^{**}\right\|_0 \lambda_n^2/\beta. \tag{4.10}$$

Furthermore, suppose $E[\Phi^{(2)}(X, A)^T|X] = \mathbf{0}$ a.s. Then with probability at least $1 - 1/n$,

$$E(\Phi^{(2)}\widehat{\theta}_n^{(2)} - T_0)^2 \leq \min_{\theta \in \Theta_n} \left(E(\Phi^{(2)}\theta^{(2)} - T_0)^2 + 5\left\|\theta^{(2)}\right\|_0 \lambda_n^2/\beta \right), \tag{4.11}$$

The results follow from Theorem A.1 in Appendix A.2 with $\rho = 0$, $\gamma = 1/8$, $\eta_1 = \eta_2 = \eta$, $t = \log 2n$ and some simple algebra (notice that Assumption 3 in Theorem 4.1 is a sufficient condition for Assumptions A.3 and A.4).

Remarks: Inequality (4.11) provides a finite sample upper bound on the mean square difference between T_0 and its estimator. This result is used to prove Theorem 4.2. The remarks below discuss how inequality (4.10) contributes to the l_1-penalization literature in prediction.

1. The conclusion of Theorem 4.3 holds for all choices of λ_n that satisfy (4.6). Suppose $\lambda_n = o(1)$, then $L(\Phi\theta_n^{**}) - L(\Phi\theta^*) \to 0$ as $n \to \infty$ (since $\|\boldsymbol{\theta}\|_0$ is bounded). Then (4.10) implies that $L(\Phi\hat{\boldsymbol{\theta}}_n) - L(\Phi\boldsymbol{\theta}^*) \to 0$ in probability. To achieve the best rate of convergence, equal sign should be taken in (4.6).
2. Note that θ_n^{**} minimizes $L(\Phi\theta) - L(Q_0) + 3\left\|\theta\right\|_0 \lambda_n^2/\beta$. Below we demonstrate that the minimum of $L(\Phi\theta) - L(Q_0) + 3\left\|\theta\right\|_0 \lambda_n^2/\beta$ can be viewed as the approximation error

plus a "tight" upper bound of the estimation error of an "oracle" in the stepwise model selection framework (when "=" is taken in (4.6)). Here "tight" means the convergence rate in the bound is the best known rate, and "oracle" is defined as follows. Let m denote a non-empty subset of the index set $\{1, \ldots, J\}$. Then each m represents a model which uses a non-empty subset of $\{\varphi_1, \ldots, \varphi_J\}$ as basis functions (there are $2^J - 1$ such subsets). Define $\widehat{\theta}_n^{(m)} = \arg\min_{\{\theta \in \mathbb{R}^J : \theta_j = 0 \text{ for all } j \notin m\}} E_n(R - \Phi\theta)^2$ and $\boldsymbol{\theta}^{*,(m)} = \arg\min_{\{\boldsymbol{\theta} \in \mathbb{R}^J : \theta_j = 0 \text{ for all } j \notin m\}} L(\Phi\boldsymbol{\theta})$. In this setting, an ideal model selection criterion will pick model m^* such that $L(\Phi\widehat{\theta}_n^{(m^*)}) = \inf_m L(\Phi\widehat{\theta}_n^{(m)})$. $\widehat{\theta}_n^{(m^*)}$ is referred as an "oracle" in Massart [20]. Note that the excess prediction error of each $\widehat{\theta}_n^{(m)}$ can be written as

$$L(\Phi\widehat{\theta}_n^{(m)}) - L(Q_0) = [L(\Phi\theta^{*,(m)}) - L(Q_0)] + [L(\Phi\widehat{\theta}_n^{(m)}) - L(\Phi\theta^{*,(m)})],$$

where the first term is called the approximation error of model m and the second term is the estimation error. It can be shown that [2] for each model m and $x_m > 0$, with probability at least $1 - \exp(-x_m)$,

$$L(\Phi\widehat{\theta}_n^{(m)}) - L(\Phi\theta^{*,(m)}) \le constant \times \left(\frac{x_m + |m| \log(n/|m|)}{n} \right)$$

under appropriate technical conditions, where $|m|$ is the cardinality of the index set m. To our knowledge this is the best rate known so far. Taking $x_m = \log n + |m| \log J$ and using the union bound argument, we have with probability at least $1 - O(1/n)$,

$$\begin{aligned} & L(\Phi_n \widehat{\theta}_n^{(m^*)}) - L(Q_0) \\ = & \min_m \left([L(\Phi\theta^{*,(m)}) - L(Q_0)] + L(\Phi\widehat{\theta}_n^{(m)}) - L(\Phi\theta^{*,(m)}) \right) \\ \le & \min_m \left([L(\Phi\theta^{*,(m)}) - L(Q_0)] + constant \times \frac{|m| \log(Jn)}{n} \right) \\ = & \min_\theta \left([L(\Phi\theta) - L(Q_0)] + constant \times \frac{\|\theta\|_0 \log(Jn)}{n} \right). \end{aligned} \tag{4.12}$$

On the other hand, take λ_n so that condition (4.6) holds with "=". (4.10) implies that, with probability at least $1 - 1/n$,

$$L(\Phi\widehat{\theta}_n) - L(Q_0) \le \min_{\theta \in \Theta_n} \left([L(\Phi\theta) - L(Q_0)] + constant \times \frac{\|\theta\|_0 \log(Jn)}{n} \right),$$

which is essentially (4.12) with the constraint of $\boldsymbol{\theta} \in \Theta_n$. (The "*constant*" in the above inequalities may take different values.) Since $\theta = \theta_n^{**}$ minimizes the approximation error plus a tight upper bound for the estimation error in the oracle model, within $\boldsymbol{\theta} \in \Theta_n$, we refer to θ_n^{**} as an oracle.

3. The result can be used to emphasize that l_1 penalty behaves similarly as the l_0 penalty. Note that $\hat{\boldsymbol{\theta}}_n$ minimizes the empirical prediction error, $E_n(R - \Theta\boldsymbol{\theta})^2$, plus an l_1 penalty whereas $\boldsymbol{\theta}^{**}(u_n)$ minimizes the prediction error $L(\Phi\boldsymbol{\theta})$ plus an l_0 penalty. We provide an intuitive connection between these two quantities. First

note that $E_n(R - \Phi\boldsymbol{\theta})^2$ estimates $L(\Phi\boldsymbol{\theta})$ and $\hat{\sigma_j}$ estimates σ_j. We use "≈" to denote this relationship. Thus

$$E_n(R - \Phi\theta)^2 + \lambda_n \sum_{j=1}^{J} \widehat{\sigma}_j |\theta_i| \approx L(\Phi\theta) + \lambda_n \sum_{j=1}^{J} \sigma_j |\theta_j| \leq L(\Phi\theta) + \lambda_n \sum_{j=1}^{J} \sigma_j |\widehat{\theta}_{n,j} - \theta_j| + \lambda_n \sum_{j=1}^{J} \sigma_j |\widehat{\theta}_{n,j}|, \tag{4.13}$$

where $\hat{\theta}_{n,j}$ is the j_{th} component of $\hat{\boldsymbol{\theta}}_n$. In Appendix B we show that for any $\boldsymbol{\theta} \in \Theta_n$, $\lambda_n \sum_{j=1}^{J} \sigma_j |\widehat{\theta}_{n,j} - \theta_j|$ is upper bounded by $\|\theta\|_0 \lambda_n^2 / \beta$ up to a constant with a high probability. Thus $\hat{\boldsymbol{\theta}}_n$ minimizes (4.13) and $\boldsymbol{\theta}^{**}(u_n)$ roughly minimizes an upper bound of (4.13).

4. The constants involved in the theorem can be improved; we focused on readability as opposed to providing the best constants.

5. A Practical Implementation and an Evaluation

In this section we develop a practical implementation of the l_1-PLS method, compare this method to two commonly used alternatives and lastly illustrate the method using the motivating data from the Nefazodone-CBASP trial [13].

A realistic implementation of l_1-PLS method should use a data-dependent method to select the tuning parameter, λ_n. Since the primary goal is to maximize the Value, we select λ_n to maximize a cross validated Value estimator. For any ITR d, it is easy to verify that $E[(R - V(d))1_{A=d(X)}/p(A|X)] = 0$. Thus an unbiased estimator of $V(d)$ is

$$E_n[1_{A=d(X)} R / p(A|X)] / E_n[1_{A=d(X)} / p(A|X)]$$

[21] (recall that the randomization distribution $p(a|X)$ is known). We split the data into 10 roughly equal-sized parts; then for each λ_n we apply the l_1-PLS based method on each 9 parts of the data to obtain an ITR, and estimate the Value of this ITR using the remaining part; the λ_n that maximizes the average of the 10 estimated Values is selected. Since the Value of an ITR is noncontinuous in the parameters, this usually results in a set of candidate λ_n's achieving maximal Value. In the simulations below the resulting λ_n is nonunique in around 97% of the data sets. If necessary, as a second step we reduce the set of λ_n's by including only λ_n's leading to the ITR's using the least number of variables. In the simulations below this second criterion effectively reduced the number of candidate λ_n's in around 25% of the data sets, however multiple λ_n's still remained in around 90% of the data sets. This is not surprising since the Value of an ITR only depends on the relative magnitudes of parameters in the ITR. In the third step we select the λ_n that minimizes the 10-fold cross validated prediction error estimator from the remaining candidate λ_n's; that is, minimization of the empirical prediction error is used as a final tie breaker.

5.1. Simulations

A first alternative to l_1-PLS is to use ordinary least squares (OLS). The estimated ITR is $\hat{d}_{OLS} \in \arg\max_a \Phi(X, a)\hat{\boldsymbol{\theta}}_{OLS}$ where $\hat{\boldsymbol{\theta}}_{OLS}$ is the OLS estimator of $\boldsymbol{\theta}$. A second alternative is called "prognosis prediction" [14]. Usually this method employees multiple data sets, each of which involves one active treatment. Then the treatment that is associated with the best predicted prognosis [14] is selected. We implement this method by estimating $E(R|X, A = a)$ via least squares with l_1 penalization for each treatment group (each $a \in \mathcal{A}$) separately. The tuning parameter involved in each treatment group is selected by minimizing the 10-fold

cross-validated prediction error estimator. The resulting ITR satisfies $\hat{d}_{PP}(X) \in \arg\max_{a \in \mathcal{A}} \hat{E}(R|X, A = a)$ where the subscript "PP" denotes prognosis prediction.

For simplicity we consider binary A. All three methods use the same number of data points and the same number of basis functions but use these data points/basis functions differently. l_1-PLS and OLS use all J basis functions to conduct estimation with all n data points whereas the prognosis prediction method splits the data into the two treatment groups and uses $J/2$ basis functions to conduct estimation with the $n/2$ data points in each of the two treatment groups. To ensure the comparison is fair across the three methods, the approximation model for each treatment group is consistent with the approximation model used in both l_1-PLS and OLS (e.g. if Q_0 is approximated by $(1, X, A, XA)\boldsymbol{\theta}$ in l_1-PLS and OLS, then in prognosis prediction we approximate $E(R|X, A = a)$ by $(1, X)\boldsymbol{\theta}_{PP}$ for each treatment group). We do not penalize the intercept coefficient in either prognosis prediction or l_1-PLS.

The three methods are compared using two criteria: 1) Value maximization; and 2) simplicity of the estimated ITRs (measured by the number of variables/basis functions used in the rule).

We illustrate the comparison of the three methods using 4 examples selected to reflect three scenarios; please see Section S.3 of the supplementary material for 4 further examples.

1. There is no treatment effect (i.e. Q_0 is constructed so that $T_0 = 0$; example 1). In this case, all ITRs yield the same Value. Thus the simplest rule is preferred.
2. There is a treatment effect and the treatment effect term T_0 is correctly modeled (example 4 for large n, and example 2). In this case, minimizing the prediction error will yield the ITR that maximizes the Value.
3. There is a treatment effect and the treatment effect term T_0 is misspecified (example 4 for small n, and example 3). In this case, there might be a mismatch between prediction error minimization and Value maximization.

The examples are generated as follows. The treatment A is generated uniformly from $\{-1, 1\}$ independent of X and the response R. The response R is normally distributed with mean $Q_0(X, A)$. In examples 1–3, $X \sim U[-1, 1]^5$ and we consider three simple examples for Q_0. In example 4, $X \sim U[0, 1]$ and we use a complex Q_0, where $Q_0(X, 1)$ and $Q(X, -1)$ are similar to the blocks function used in Donoho and Johnstone [8]. Further details of the simulation design are provided in Appendix A.3.

We consider two types of approximation models for Q_0. In examples 1–3, we approximate Q_0 by $(1, X, A, XA)\boldsymbol{\theta}$. In example 4, we approximate Q_0 by Haar wavelets. The number of basis functions may increase as n increases (we index J, Φ and $\boldsymbol{\theta}^*$ by n in this case). Plots for $Q_0(X, A)$ and the associated best wavelet fits $\Phi_n(X, A)\theta_n^*$ are provided in Figure 1.

For each example, we simulate data sets of sizes $n = 2^k$ for $k = 5, \ldots, 10$. 1000 data sets are generated for each sample size. The Value of each estimated ITR is evaluated via Monte Carlo using a test set of size 10, 000. The Value of the optimal ITR is also evaluated using the test set.

Simulation results are presented in Figure 2. When the approximation model is of high quality, all methods produce ITRs with similar Value (see examples 1, 2 and example 4 for large n). However, when the approximation model is poor, the l_1-PLS method may produce highest Value (see example 3). Note that in example 3 settings in which the sample size is small, the Value of the ITR produced by l_1-PLS method has larger median absolute

deviation (MAD) than the other two methods. One possible reason is that due to the mismatch between maximizing the Value and minimizing the prediction error, the Value estimator plays a strong role in selecting λ_n. The non-smoothness of the Value estimator combined with the mismatch results in very different λ_ns and thus the estimated decision rules vary greatly from data set to data set in this example. Nonetheless, the l_1-PLS method is still preferred after taking the variation into account; indeed l_1-PLS produces ITRs with higher Value than both OLS and PP in around 46%, 55% and 67% in data sets of sizes n = 32, 64 and 128, respectively. Furthermore, in general the l_1-PLS method uses much fewer variables for treatment assignment than the other two methods. This is expected because the OLS method does not have variable selection functionality and the PP method will use all variables that are predictive of the response R whereas the use of the Value in selecting the tuning parameter in l_1-PLS discounts variables that are only useful in predicting the response (and less useful in selecting the best treatment).

5.2. Nefazodone-CBASP trial example

The Nefazodone-CBASP trial was conducted to compare the efficacy of several alternate treatments for patients with chronic depression. The study randomized 681 patients with non-psychotic chronic major depressive disorder (MDD) to either Nefazodone, cognitive behavioral-analysis system of psychotherapy (CBASP) or the combination of the two treatments. Various assessments were taken throughout the study, among which the score on the 24-item Hamilton Rating Scale for Depression (HRSD) was the primary outcome. Low HRSD scores are desirable. See Keller et al. [13] for more detail of the study design and the primary analysis.

In the data analysis, we use a subset of the Nefazodone-CBASP data consisting of 656 patients for whom the response HRSD score was observed. In this trial, pairwise comparisons show that the combination treatment resulted in significantly lower HRSD scores than either of the single treatments. There was no overall difference between the single treatments.

We use l_1-PLS to develop an ITR. In the analysis the HRSD score is reverse coded so that higher is better. We consider 50 pretreatment variables $X = (X_1,\ldots, X_{50})$. Treatments are coded using contrast coding of dummy variables $A = (A_1, A_2)$, where $A_1 = 2$ if the combination treatment is assigned and -1 otherwise and $A_2 = 1$ if CBASP is assigned, -1 if nefazodone and 0 otherwise. The vector of basis functions, $\Phi(X, A)$, is of the form $(1, X, A_1, XA_1, A_2, XA_2)$. So the number of basis functions is $J = 153$. As a contrast, we also consider the OLS method and the PP method (separate prognosis prediction for each treatment). The vector of basis functions used in PP is $(1, X)$ for each of the three treatment groups. Neither the intercept term nor the main treatment effect terms in l_1-PLS or *PP* is penalized (see Section S.2 of the supplementary material for the modification of the weights $\hat{\sigma_j}$ used in (4.1)).

The ITR given by the l_1-PLS method recommends the combination treatment to all (so none of the pretreatment variables enter the rule). On the other hand, the PP method produces an ITR that uses 29 variables. If the rule produced by PP were used to assign treatment for the 656 patients in the trial, it would recommend the combination treatment for 614 patients and nefazodone for the other 42 patients. In addition, the OLS method will use all the 50 variables. If the ITR produced by OLS were used to assign treatment for the 656 patients in the trial, it would recommend the combination treatment for 429 patients, nefazodone for the 145 patients and CBASP for the other 82 patients.

6. Discussion

Our goal is to construct a high quality ITR that will benefit future patients. We considered an l_1-PLS based method and provided a finite sample upper bound for $V(d_0) - V(\hat{d}_n)$, the excess Value of the estimated ITR.

The use of an l_1 penalty allows us to consider a large model for the conditional mean function Q_0 yet permits a sparse estimated ITR. In fact, many other penalization methods such as SCAD [9] and l_1 penalty with adaptive weights (adaptive Lasso; [37]) also have this property. We choose the non-adaptive l_1 penalty to represent these methods. Interested readers may justify other PLS methods using similar proof techniques.

The high probability finite sample upper bounds (i.e. (4.7) and (4.9)) cannot be used to construct a prediction/confidence interval for $V(d_0) - V(\hat{d}_n)$ due to the unknown quantities in the bound. How to develop a tight computable upper bound to assess the quality of $\hat{d}_n$ is an open question.

We used cross validation with Value maximization to select the tuning parameter involved in the l_1-PLS method. As compared to the OLS method and the PP method, this method may yield higher Value when T_0 is misspecified. However, since only the Value is used to select the tuning parameter, this method may produce a complex ITR for which the Value is only slightly higher than that of a much simpler ITR. In this case, a simpler rule may be preferred due to the interpretability and cost of collecting the variables. Investigation of a tuning parameter selection criterion that trades off the Value with the number of variables in an ITR is needed.

This paper studied a one stage decision problem. However, it is evident that some diseases require time-varying treatment. For example, individuals with a chronic disease often experience a waxing and waning course of illness. In these settings the goal is to construct a sequence of ITRs that tailor the type and dosage of treatment through time according to an individual's changing status. There is an abundance of statistical literature in this area [29,30,22,23,26,17,34,35]. Extension of the least squares based method to the multi-stage decision problem has been presented in Murphy [23]. The performance of l_1 penalization in this setting is unclear and worth investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Martin Keller and the investigators of the Nefazodone-CBASP trial for use of their data. The authors also thank John Rush, MD, for the technical support and Bristol-Myers Squibb for helping fund the trial. The authors thank valuable comments from Eric B. Laber and Peng Zhang.

References

1. Bartlett PL, Jordan ML, McAuliffe JD. Convexity, classification, and risk bounds. Journal of the American Statistical Association. 2006; 135(3):311–334.
2. Bartlett PL. Fast rates for estimation error and oracle inequalities for model selection. Econometric Theory. 2008; 24(2):545–552.
3. Bickel PJ, Ritov Y, Tsybakov AB. Simultaneous analysis of lasso and dantzig selector. The Annals of Statistics. 2009; 37(4):1705–1732.

4. Bunea F, Tsybakov AB, Wegkamp MH. Sparsity oracle inequalities for the Lasso. Electronic Journal of Statistics. 2007; 1:169–194.
5. Cai T, Tian L, Lloyd-Jones DM, Wei LJ. Evaluating Subject-level Incremental Values of New Markers for Risk Classification Rule. Harvard University Biostatistics Working Paper Series. Working Paper 91. 2008a
6. Cai T, Tian L, Uno H, Solomon SD, Wei LJ. Calibrating Parametric Subject-specific Risk Estimation. Harvard University Biostatistics Working Paper Series. Working Paper 92. 2008b
7. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc; 1988.
8. Donoho D, Johnstone I. Ideal spatial adaptation by wavelet shrinkage. Biometrika. 1994; 81(3):425–455.
9. Fan J, Li R. Variable selection via nonconcave penalized likelihood and it oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.
10. Feldstein ML, Savlov ED, Hilf R. A statistical model for predicting response of breast cancer patients to cytotoxic chemotherapy. Cancer Research. 1978; 38(8):2544–2548. [PubMed: 667849]
11. Insel TR. Translating scientific opportunity into public health impact: a strategic plan for research on mental illness. Archives of General Psychiatry. 2009; 66(2):128–33. [PubMed: 19188534]
12. Ishigooka J, Murasaki M, Miura S. The Olanzapine Late-Phase II Study Group. Olanzapine optimal dose: Results of an open-label multicenter study in schizophrenic patients. Psychiatry and Clinical Neurosciences. 2001; 54(4):467–478. [PubMed: 10997865]
13. Keller MB, McCullough JP, Klein DN, Arnow B, Dunner DL, Gelenberg AJ, Markowitz JC, Nemeroff CB, Russell JM, Thase ME, Trivedi MH, Zajecka J. A comparison of nefazodone, the cognitive behavioral-analysis system of psychotherapy, and their combination for the treatment of chronic depression. The New England Journal of Medicine. 2000; 342(20):1462–1470. [PubMed: 10816183]
14. Kent DM, Hayward RA, Griffith JL, Vijan S, Beshansky JR, Califf RM, Selker HP. An independently derived and validated predictive model for selecting patients with myocardial infarction who are likely to benefit from tissue plasminogen activator compared with streptokinase. The American Journal of Medicine. 2002; 113(2):104–11. [PubMed: 12133748]
15. Koltchinskii V. Sparsity in penalized empirical risk minimization. Annales de l'Institut Henri Poincaré Probabilitiés et Statistiques. 2009; 45(1):7–57.
16. Lesko LJ. Personalized Medicine: Elusive Dream or Imminent Reality? Clinical Pharmacology and Therapeutics. 2007; 81:807–816. [PubMed: 17505496]
17. Lunceford JK, Davidian M, Tsiatis AA. Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. Biometrics. 2002; 58:48–57. [PubMed: 11890326]
18. Mammen E, Tsybakov A. Smooth discrimination analysis. The Annals of Statistics. 1999; 27:1808–1829.
19. Massart, P. Ecole d'Eté de Probabilités de Saint-Flour XXXIII, Concentration inequalities and model selection. Springer; 2003.
20. Massart, P. A non asymptotic theory for model selection. Proceedings of the 4th European Congress of Mathematicians (Ed. Ari Laptev); European Mathematical Society; 2005. p. 309-323.
21. Murphy SA, van der Laan MJ, Robins JM. CPPRG. Marginal mean models for dynamic regimes. Journal of the American Statistical Association. 2001; 96:1410–1423. [PubMed: 20019887]
22. Murphy SA. Optimal Dynamic Treatment Regimes. Journal of the Royal Statistical Society, Series B (with discussion). 2003; 65(2):331–366.
23. Murphy SA. A Generalization error for Q-Learning. Journal of Machine Learning Research. 2005; 6:1073–1097. [PubMed: 16763665]
24. Piquette-Miller P, Grant DM. The Art and Science of Personalized Medicine. Clinical Pharmacology and Therapeutics. 2007; 81:311–315. [PubMed: 17339856]
25. Polonik W. Measuring mass concentrations and estimating density contour clusters - an excess mass approach. The Annals of Statistics. 1995; 23(3):855–881.

26. Robins, JM. Optimal-regime structural nested models. In: Lin, DY.; Haegerty, P., editors. Lecture notes in Stastitics; Proceedings of the Second Seattle Symposium on Biostatistics; New York: Springer; 2004.
27. Robins JM, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. Statistics in Medicine. 2008; 27(23):4678–4721. [PubMed: 18646286]
28. Stoehlmacher J, Park DJ, Zhang W, Yang D, Groshen S, Zahedy S, Lenz HJ. A multivariate analysis of genomic polymorphisms: prediction of clinical outcome to 5-FU/oxaliplatin combination chemotherapy in refractory colorectal cancer. British Journal of cancer. 2004; 91(2): 344–354. [PubMed: 15213713]
29. Thall PF, Millikan RE, Sung HG. Evaluating multiple treatment courses in clinical trials. Statistics in Medicine. 2000; 19:1011–1028. [PubMed: 10790677]
30. Thall PF, Sung HG, Estey EH. Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. Journal of the American Statistical Association. 2002; 97:29–39.
31. Tibshirani R. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1996; 32:135–166.
32. Tsybakov AB. Optimal aggregation of classifiers in statistical learning. The Annals of Statistics. 2004; 32:135–166.
33. van de Geer S. High-dimensional generalized linear models and the Lasso. The Annals of Statistics. 2008; 36(2):614–645.
34. van der Laan MJ, Petersen ML, Joffe MM. History-Adjusted Marginal Structural Models and Statically-Optimal Dynamic Treatment Regimens. The International Journal of Biostatistics. 2005; 1(1) Article 4.
35. Wahed AS, Tsiatis AA. Semiparametric efficient estimation of survival distribution for treatment policies in two-stage randomization designs in clinical trials with censored data. Biometrika. 2006; 93:163–177.
36. Zhang CH, Huang J. The sparsity and bias of the lasso selection in high-dimensional linear regression. The Annals of Statistics. 2008; 36(4):1567–1594.
37. Zou H. The Adaptive Lasso and its Oracle Properties. Journal of the American Statistical Association. 2006; 101(476):1418–1429.

APPENDIX

A.1. Proof of Theorem 3.1

For any ITR d: $\mathcal{X} \to \mathcal{A}$, denote $\Delta T_d(X) \triangleq \max_{a\in\mathcal{A}} T_0(X, a) - T_0(X, d(X))$. Using similar arguments to that in Section 2, we have $V(d_0) - V(d) = E(\Delta T_d)$. If $V(d_0) - V(d) = 0$, then (3.4) and (3.5) automatically hold. Otherwise, $E(\Delta T_d)^2 \geq (E\Delta T_d)^2 > 0$. In this case, for any $\varepsilon > 0$, define the event

$$\Omega_\varepsilon = \left\{ \max_{a\in\mathcal{A}} T_0(X,a) - \max_{a\in\mathcal{A}\setminus \arg\max_{a\in\mathcal{A}} T_0(X,a)} T_0(X,a) \leq \varepsilon \right\}.$$

Then $\Delta T_d \leq (\Delta T_d)^2/\varepsilon$ on the event Ω_ε^C. This together with the fact that $\Delta T_d \leq (\Delta T_d)^2/\varepsilon + \varepsilon/4$ implies

$$\begin{aligned} V(d_0) - V(d) &= E(1_{\Omega_\varepsilon^C}\Delta T_d) + E(1_{\Omega_\varepsilon}\Delta T_d) \\ &\leq \tfrac{1}{\varepsilon}E[1_{\Omega_\varepsilon^C}(\Delta T_d)^2] + E\left[1_{\Omega_\varepsilon}\left(\tfrac{(\Delta T_d)^2}{\varepsilon} + \tfrac{\varepsilon}{4}\right)\right] \\ &= \tfrac{1}{\varepsilon}E[(\Delta T_d)^2] + \tfrac{\varepsilon}{4}P(\Omega_\varepsilon) \leq \tfrac{1}{\varepsilon}E[(\Delta T_d)^2] + \tfrac{C}{4}\varepsilon^{1+\alpha}, \end{aligned}$$

where the last inequality follows from the margin condition (3.3). Choosing $\varepsilon = (4E(\Delta T_d)^2/C)^{1/(2+\alpha)}$ to minimize the above upper bound yields

$$V(d_0) - V(d) \le 2^{\alpha/(2+\alpha)} C^{1/(2+\alpha)} [E(\Delta T_d)^2]^{(1+\alpha)/(2+\alpha)}. \tag{A.1}$$

Next, for any d and Q such that $d(X) \in \max_{a\in\mathcal{A}} Q(X, a)$ and decomposition $Q(X, A)$ into $W(X) + T(X, A)$,

$$\begin{aligned} E(\Delta T_d)^2 &= E\left[\left(\max_{a\in\mathcal{A}} T_0(X,a) - \max_{a\in\mathcal{A}} T(X,a) + T(X,d(X)) - T_0(X,d(X))\right)^2\right] \\ &\le 2E\left[\left(\max_{a\in\mathcal{A}} T_0(X,a) - \max_{a\in\mathcal{A}} T(X,a)\right)^2 + (T(X,d(X)) - T_0(X,d(X)))^2\right] \\ &\le 4E\left[\max_{a\in\mathcal{A}} (T(X,a) - T_0(X,a))^2\right], \end{aligned}$$

where the last inequality follows from the fact that neither $|\max_a T_0(X, a) - \max_a T(X, a)|$ nor $|T(X, d(X)) - T_0(X, d(X))|$ is larger than $\max_a |T(X, a) - T_0(X, a)|$. Since $p(a|x) \ge S^{-1}$ for all (x, a) pairs, we have

$$E(\Delta T_d)^2 \le 4S\,E\left[\sum_{a\in\mathcal{A}} (T(X,a) - T_0(X,a))^2 p(a|X)\right] = 4S\,E(T(X,A) - T_0(X,A))^2. \tag{A.2}$$

Inequality (3.5) follows by substituting (A.2) into (A.1) and setting $W(X, A) = E[Q(X, A)|X]$. Inequality (3.4) follows by setting $W(X) = 0$ and noticing that $\Delta T_d(X) = \max_{a\in\mathcal{A}} Q_0(X, a) - Q_0(X, d(X))$.

A.2. Generalization of Theorem 4.3

In this section, we present a generalization of Theorem 4.3 where J may depend on n and the sparsity of any $\boldsymbol{\theta} \in \mathbb{R}^J$ is measured by the number of "large" components in $\boldsymbol{\theta}$ as described in Zhang and Huang [36]. In this case, J, F and the prediction error minimizer $\boldsymbol{\theta}^*$ are denoted as J_n, Φ_n and θ_n^*, respectively. All relevant quantities and assumptions are re-stated below.

Let $|M|$ denote the cardinality of any index set $M \subseteq \{1,\ldots, J_n\}$. For any $\boldsymbol{\theta} \in \mathbb{R}^{J_n}$ and constant $\rho \ge 0$, define

$$M_{\rho\lambda_n}(\theta) \in \arg \min_{\{M\subseteq\{1,\ldots,J_n\}:\sum_{j\in\{1,\ldots,J_n\}\setminus M}\sigma_j|\theta_j|\le\rho|M|\lambda_n\}} |M|.$$

Then $M_{\rho\lambda_n}(\boldsymbol{\theta})$ is the smallest index set that contains only "large" components in $\boldsymbol{\theta}$. $|M_{\rho\lambda_n}(\boldsymbol{\theta})|$ measures the sparsity of $\boldsymbol{\theta}$. It is easy to see that when $\rho = 0$, $M_0(\boldsymbol{\theta})$ is the index set of nonzero components in $\boldsymbol{\theta}$ and $|M_0(\boldsymbol{\theta})| = \|\boldsymbol{\theta}\|_0$. Moreover, $M_{\rho\lambda_n}(\boldsymbol{\theta})$ is an empty set if and only if $\boldsymbol{\theta} = \mathbf{0}$.

Let $[\theta_n^*]$ be the set of most sparse prediction error minimizers in the linear model, i.e.

$$[\theta_n^*] = \arg \min_{\theta \in \arg\min_\theta L(\Phi_n \theta)} |M_{\rho\lambda_n}(\theta)|. \tag{A.3}$$

Note that $[\theta_n^*]$ depends on $\rho\lambda_n$.

To derive the finite sample upper bound for $L(\Phi_n \hat{\boldsymbol{\theta}}_n)$, we need the following assumptions.

Assumption A.1

The error terms ε_i, i = 1,…, n are independent of (X_i, A_i), i = 1,…, n and are i.i.d. with $E(\varepsilon_i) = 0$ and $E[|\varepsilon_i|^l] \le \frac{l!}{2} c^{l-2}\sigma^2$ for some c, $\sigma^2 > 0$ for all $l \ge 2$.

Assumption A.2

For all n ≥ 1,

a. there exists an $1 \le U_n < \infty$ such that $\max_{j=1,\ldots,J_n} \|\varphi_j\|_\infty / \sigma_j \le U_n$, where $\sigma_j \triangleq (E\varphi_j^2)^{1/2}$.

b. there exists an $0 < \eta_{1,n} < \infty$, such that $\sup_{\theta \in [\theta_n^*]} \left\| Q_0 - \Phi_n \theta \right\|_\infty \le \eta_{1,n}$.

For any $0 \le \gamma < 1/2$, $\eta_{2,n} \ge 0$ (which may depend on n) and tuning parameter λ_n, define

$$\Theta_n^o = \{\theta \in \mathbb{R}^{J_n} : \exists \theta^o \in [\theta_n^*] \text{ s.t.} \left\| \Phi_n(\theta - \theta^o) \right\|_\infty \le \eta_{2,n} \text{ and } \max_{j=1,\ldots,J_n} \left| E\left[\Phi_n(\theta - \theta^o) \frac{\varphi_j}{\sigma_j} \right] \right| \le \gamma\lambda_n \}.$$

Assumption A.3

For any n ≥ 1, there exists a $\beta_n > 0$ such that

$$E[\Phi_n(\tilde{\theta} - \theta)]^2 |M_{\rho\lambda_n}(\theta)| \ge \beta_n \left[\left(\sum_{j \in M_{\rho\lambda_n}(\theta)} \sigma_j |\tilde{\theta}_j - \theta_j| \right)^2 - \rho^2 |M_{\rho\lambda_n}(\theta)|^2 \lambda_n^2 \right]$$

for all $\theta \in \Theta_n^o \setminus \{\mathbf{0}\}$, $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{J_n}$ and $\sum_{j \in \{1,\ldots,J_n\} \setminus M_{\rho\lambda_n}(\theta)} \sigma_j |\tilde{\theta}_j| \le \frac{2\gamma+5}{1-2\gamma} \left(\sum_{j \in M_{\rho\lambda_n}(\theta)} |\tilde{\theta}_j - \theta_j| + \rho |M_{\rho\lambda_n}(\theta)| \lambda_n \right)$.

When $E(\Phi_n^{(2)}(X, A)^T | X) = \mathbf{0}$ a.s. ($\Phi_n^{(2)}$ is defined in Section 4.1), we need an extra assumption to derive the finite sample upper bound for the mean square error of the treatment effect estimator, $E[\Phi_n^{(2)} \widehat{\theta}_n^{(2)} - T_0(X, A)]^2$ (recall that $T_0(X,A) \triangleq Q_0(X, A) - E[Q_0(X,A)|X]$).

Assumption A.4

For any n ≥ 1, there exists a $\beta_n > 0$ such that

$$E[\Phi_n^{(2)}(\tilde{\theta}^{(2)} - \theta^{(2)})]^2 |M_{\rho\lambda_n}^{(2)}(\theta)| \geq \beta_n \left[\left(\sum_{j \in M_{\rho\lambda_n}^{(2)}(\theta)} \sigma_j |\tilde{\theta}_j - \theta_j| \right)^2 - \rho^2 |M_{\rho\lambda_n}^{(2)}(\theta)|^2 \lambda_n^2 \right]$$

for all $\theta \in \Theta_n^o \backslash \{\mathbf{0}\}$, $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{J_n}$ and $\sum_{j \in \{1,\ldots,J_n\} \backslash M_{\rho\lambda_n}(\theta)} \sigma_j |\tilde{\theta}_j| \leq \frac{2\gamma+5}{1-2\gamma} \left(\sum_{j \in M_{\rho\lambda_n}(\theta)} |\tilde{\theta}_j - \theta_j| + \rho |M_{\rho\lambda_n}(\theta)| \lambda_n \right)$, where

$$M_{\rho\lambda_n}^{(2)}(\theta) \in \arg \min_{\{M \subseteq \{J_n^{(1)}+1,\ldots,J_n\}: \sum_{j \in (J_n^{(1)}+1,\ldots,J_n) \backslash M} \sigma_j |\theta_j| \leq \rho |M| \lambda_n\}} |M|$$

is the smallest index set that contains only large components in $\boldsymbol{\theta}^{(2)}$.

Note that here for simplicity, we assume that Assumptions A.3 and A.4 hold with the same value of β_n. And with out loss of generality, we can always choose a small enough β_n so that $\rho\beta_n \leq 1$ for a given ρ.

For any $t > 0$, define

$$\Theta_n = \{\theta \in \Theta_n^o : |M_{\rho\lambda_n}(\theta)| \leq \frac{(1-2\gamma)^2 \beta_n}{120} \left[\sqrt{\frac{1}{9} + \frac{n}{2U_n^2[\log(3J_n(J_n+1)) + t]}} - \frac{1}{3} \right]\}. \tag{A.4}$$

Note that we allow U_n, $\eta_{1,n}$, $\eta_{2,n}$ and β_n^{-1} to increase as n increases. However, if those quantities are small, the upper bound in (A.7) will be tighter.

Theorem A.1

Suppose Assumptions A.1 and A.2 hold. For any given $0 \leq \gamma < 1/2$, $\eta_{2,n} > 0$, $\rho \geq 0$ and $t > 0$, let $\hat{\boldsymbol{\theta}}_n$ be the l_1-PLS estimator defined in (4.1) with tuning parameter

$$\lambda_n \geq \frac{8 \max\{3c, 2(\eta_{1,n}+\eta_{2,n})\} U_n (\log 6J_n + t)}{(1-2\gamma)n} + \frac{12 \max\{\sigma, (\eta_{1,n}+\eta_{2,n})\}}{(1-2\gamma)} \sqrt{\frac{2(\log 6J_n + t)}{n}}. \tag{A.5}$$

Suppose Assumption A.3 holds with $\rho\beta_n \leq 1$. Let Θ_n be the set defined in (A.4) and assume Θ_n is non-empty. If

$$\frac{\log 2J_n}{n} \leq \frac{2(1-2\gamma)^2}{27U_n^2 - 10\gamma - 22}, \tag{A.6}$$

then with probability at least $1 - \exp(-k_n' n) - \exp(-t)$, we have

$$L(\Phi_n\widehat{\theta}_n) \le \min_{\theta\in\Theta_n}\left[L(\Phi_n\theta)+K_n\frac{|M_{\rho\lambda_n}(\theta)|}{\beta_n}\lambda_n^2\right], \tag{A.7}$$

where $k_n^{'}=13(1-2\gamma)^2[6(27U_n^2-10\gamma-22)]$ and $K_n = [40\gamma(12\beta_{n\rho}+2\gamma+5)]/[(1-2\gamma)(2\gamma+19)] + 130(12\beta_{n\rho}+2\gamma+5)^2/[9(2\gamma+19)^2]$.

Furthermore, suppose $E(\Phi_n^{(2)}(X,A)^T|X)=\mathbf{0}$ a.s. If Assumption A.4 holds with $\rho\beta_n \le 1$, then with probability at least $1-\exp(-k_n^{'}n)-\exp(-t)$, we have

$$E(\Phi_n^{(2)}\widehat{\theta}_n^{(2)}-T_0)^2 \le \min_{\theta\in\Theta_n}\left[E(\Phi_n^{(2)}\theta^{(2)}-T_0)^2+K_n^{'}\frac{|M_{\rho\lambda_n}^2(\theta)|}{\beta_n}\lambda_n^2\right].$$

where $K_n^{'}=20(12\beta_n\rho+2\gamma+5)\{\gamma/[(1-2\gamma)(7-6\beta_n\rho)]+[3(1-2\gamma)\beta_n\rho+10(2\gamma+5)]/[9(2\gamma+19)^2]\}$.

Remark

1. Note that K_n is upper bounded by a constant under the assumption $\beta_n\rho \le 1$. In the asymptotic setting when $n\to\infty$ and $J_n\to\infty$, (A.7) implies that with probability tending to 1, $L(\Phi_n\widehat{\theta}_n)-L(\Phi_n\theta_n^*)\to 0$ if (i) $|M_{\rho\lambda_n}(\theta_n^*)|\lambda_n^2/\beta_n=o(1)$, (ii) $U_n^2\log J_n/n \le k_1$ and $|M_{\rho\lambda_n}(\theta_n^*)| \le k_2\beta_n\sqrt{n/(U_n^2\log J_n)}$ for some sufficiently small positive constants k_1 and k_2, and (iii) $\lambda_n \ge k_3\max\{1,\eta_{1,n}+\eta_{2,n}\}\sqrt{\log J_n/n}$ for a sufficiently large constant k_3, where $\theta_n^*\in[\theta_n^*]$ (take $t=\log J_n$).

2. Below we briefly discuss Assumptions A.2 – A.4.

   Assumption A.2 is very similar to Assumption 2 in Theorem 4.1 (which is used to prove the concentration of the sample mean around the true mean), except that U_n and $\eta_{1,n}$ may increase as n increases. This relaxation allows the use of basis functions for which the sup norm $\max_j ||\varphi_j||_\infty$ is increasing in n (e.g. the wavelet basis used in example 4 of the simulation studies).

   Assumption A.3 is a generalization of condition (4.8) (which has been discussed in remark 4 following Theorem Theorem 4.1)) to the case where J_n may increase in n and the sparsity of a parameter is measured by the number of "large" components as described at the beginning of this section. This condition is used to avoid the collinearity problem. It is easy to see that when $\rho=0$ and β_n is fixed in n, this assumption simplifies to condition (4.8).

   Assumption A.4 puts a strengthened constraint on the linear model of the treatment effect part, as compared to Assumption A.3. This assumption, together with Assumption A.3, is needed in deriving the upper bound for the mean square error of the treatment effect estimator. It is easy to verify that if $E[\Phi_n^T\Phi_n]$ is positive definite, then both A.3 and A.4 hold. Although the result is about the treatment effect part, which is asymptotically independent of the main effect of X (when $E[\Phi_n^{(2)}(X,A)|X]=\mathbf{0}$ a.s.), we still need Assumption A.3 to show that the cross product term $E_n[(\Phi_n^{(1)}\widehat{\theta}_n^{(1)}-\Phi_n^{(1)}\theta^{(1)})(\Phi_n^{(2)}\widehat{\theta}_n^{(2)}-\Phi_n^{(2)}\theta^{(2)})]$ is upper bounded by a quantity converging to 0 at the desired rate. We may use a really poor model for the

main effect part $E(Q_0(X, A)|X)$ (e.g. $\Phi_n^{(1)} \equiv 1$), and Assumption A.4 implies Assumption A.3 when $\rho = 0$. This poor model only effects the constants involved in the result. When the sample size is large (so that λ_n is small), the estimated ITR will be of high quality as long as T_0 is well approximated.

Proof

For any $\boldsymbol{\theta} \in \Theta_n$, define the events

$$\Omega_1 = \bigcap_{j=1}^{J_n} \left\{ \frac{2(1+\gamma)}{3}\sigma_j \le \widehat{\sigma}_j \le \frac{2(2-\gamma)}{3}\sigma_j \right\} (\text{where } \widehat{\sigma}_j \triangleq (E_n\varphi_j^2)^{1/2}),$$
$$\Omega_2(\theta) = \left\{ \max_{j,k=1,\ldots,J_n} \left| (E - E_n)\left(\frac{\varphi_j\varphi_k}{\sigma_j\sigma_k}\right) \right| \le \frac{(1-2\gamma)^2\beta_n}{120|M_{\rho\lambda_n}(\theta)|} \right\},$$
$$\Omega_3(\theta) = \left\{ \max_{j=1,\ldots,J_n} \left| E_n\left[(R - \Phi_n\theta)\frac{\varphi_j}{\sigma_j} \right] \right| \le \frac{4\gamma+1}{6}\lambda_n \right\}.$$

Then there exists a $\theta^o \in [\theta_n^*]$ such that

$$\begin{aligned} L(\Phi_n\widehat{\theta}_n) &= L(\Phi_n\theta) + 2E[(\Phi_n\theta^o - \Phi_n\theta)\Phi_n(\theta - \widehat{\theta}_n)] + E[\Phi_n(\widehat{\theta}_n - \theta)]^2 \\ &\le L(\Phi_n\theta) + 2\max_{j=1,\ldots,J_n} \left| E\left[\Phi_n(\theta^o - \theta)\frac{\varphi_j}{\sigma_j} \right] \right| \left(\sum_{j=1}^{J_n} \sigma_j|\widehat{\theta}_{n,j} - \theta_j| \right) + E[\Phi_n(\widehat{\theta}_n - \theta)]^2 \\ &\le L(\Phi_n\theta) + 2\gamma\lambda_n \left(\sum_{j=1}^{J_n} \sigma_j|\widehat{\theta}_{n,j} - \theta_j| \right) + E[\Phi_n(\widehat{\theta}_n - \theta)]^2, \end{aligned}$$

where the first equality follows from the fact that $E[(R - \Phi_n\boldsymbol{\theta}^o)\varphi_j] = 0$ for any $\theta^o \in [\theta_n^*]$ for $j = 1,\ldots, J_n$ and the last inequality follows from the definition of Θ_n^o.

Based on Lemma A.1 below, we have that on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$,

$$L(\Phi_n\widehat{\theta}_n) \le L(\Phi_n\theta) + K_n\frac{|M_{\rho\lambda_n}(\theta)|}{\beta_n}\lambda_n^2.$$

Similarly, when $E[\Phi_2^{(2)}(X, A)^T|X] = \mathbf{0}$, by Lemma A.2, we have that on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$,

$$\begin{aligned} E(\Phi_n^{(2)}\widehat{\theta}_n^{(2)} - T_0)^2 &\le E(\Phi_n^{(2)}\theta^{(2)} - T_0)^2 + 2\gamma\lambda_n \left(\sum_{j=J_n^{(1)}+1}^{J_n} \sigma_j|\widehat{\theta}_{n,j} - \theta_j| \right) + E[\Phi_n^{(2)}(\widehat{\theta}_n^{(2)} - \theta^{(2)})]^2 \\ &\le E(\Phi_n^{(2)}\theta^{(2)} - T_0)^2 + K_n'\frac{|M_{\rho\lambda_n}^{(2)}(\theta)|}{\beta_n}\lambda_n^2. \end{aligned}$$

The conclusion of the theorem follows from the union probability bounds of the events Ω_1, $\Omega_2(\boldsymbol{\theta})$ and $\Omega_3(\boldsymbol{\theta})$ provided in Lemmas A.3, A.4 and A.5.

Below we state the lemmas used in the proof of Theorem A.1. The proofs of the lemmas are given in Section S.3 of the supplementary material.

Lemma A.1

Suppose Assumption A.3 holds with $\rho\beta_n \le 1$. Then for any $\boldsymbol{\theta} \in \Theta_n$, on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$, we have

$$\sum_{j=1}^{J_n} \sigma_j |\widehat{\theta}_{n,j} - \theta_j| \le \frac{20(12\rho\beta_n + 2\gamma + 5)}{(1-2\gamma)(19+2\gamma)\beta_n} |M_{\rho\lambda_n}(\theta)| \lambda_n \tag{A.8}$$

and

$$E[\Phi_n(\widehat{\theta}_n - \theta)]^2 \le \frac{130(12\rho\beta_n + 2\gamma + 5)^2}{9(19+2\gamma)^2\beta_n} |M_{\rho\lambda_n}(\theta)| \lambda_n^2 \tag{A.9}$$

Remark—This lemma implies that $\hat{\boldsymbol{\theta}}_n$ is close to each $\boldsymbol{\theta} \in \Theta_n$ on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$. The intuition is as follows. Since $\hat{\boldsymbol{\theta}}_n$ minimizes (4.1), the first order conditions imply that $\max_j |E_n(R - \Phi_n \hat{\boldsymbol{\theta}}_n)\varphi_j/\hat{\sigma}_j| \le \lambda_n/2$. Similar property holds for $\boldsymbol{\theta}$ on the event $\Omega_1 \cap \Omega_3(\boldsymbol{\theta})$. Assumption A.3 together with event $\Omega_2(\boldsymbol{\theta})$ ensures that there is no collinearity in the $n \times J_n$ design matrix $(\Phi_n(X_i, A_i))_{i=1}^n$. These two aspects guarantee the closeness of $\hat{\boldsymbol{\theta}}_n$ to $\boldsymbol{\theta}$.

Lemma A.2

Suppose $E[\Phi_n^{(2)}(X, A)^T | X] = \mathbf{0}$ a.s. and Assumption A.4 holds with $\rho\beta_n \le 1$. Then for any $\boldsymbol{\theta} \in \Theta_n$, on the event $\Omega_1 \cap \Omega_2(\boldsymbol{\theta}) \cap \Omega_3(\boldsymbol{\theta})$, we have

$$\sum_{j=J_n^{(1)}+1}^{J_n} \sigma_j |\widehat{\theta}_{n,j} - \theta_j| \le \frac{10(12\beta_n\rho + 2\gamma + 5)}{(1-2\gamma)(7-6\beta_n\rho)\beta_n} |M^{(2)}_{\rho\lambda_n}(\theta)| \lambda_n \tag{A.10}$$

and

$$E[\Phi_n^{(2)}(\widehat{\theta}_n^{(2)} - \theta^{(2)})]^2 \le \frac{20(12\rho\beta_n + 2\gamma + 5)[3(1-2\gamma)\beta_n\rho + 10(2\gamma+5)]}{9(2\gamma+19)^2\beta_n} |M^{(2)}_{\rho\lambda_n}(\theta)| \lambda_n^2 \tag{A.11}$$

Lemma A.3

Suppose Assumption A.2(a) and inequality (A.6) hold. Then $\mathbf{P}(\Omega_1^C) \le \exp(-k_n' n)$, where $k_n' = 13(1-2\gamma)^2/[6(27U_n^2 - 10\gamma - 22)]$.

Lemma A.4

Suppose Assumption A.2(a) holds. Then for any $t > 0$ and $\boldsymbol{\theta} \in \Theta_n$, $\mathbf{P}(\{\Omega_2(\boldsymbol{\theta})\}^C) \le 2\exp(-t)/3$.

Lemma A.5

Suppose Assumptions A.1 and A.2 hold. For any $t > 0$, if λ_n satisfies condition (A.5), then for any $\boldsymbol{\theta} \in \Theta_n$, we have $\mathbf{P}(\{\Omega_3(\boldsymbol{\theta})\}^C) \le 2\exp(-t)/3$.

A.3. Design of simulations in Section 5.1

In this section, we present the detailed simulation design of the examples used in Section 5.1. These examples satisfy all assumptions listed in the Theorems (it is easy to verify that for examples 1–3. Validity of the assumptions for example 4 is addressed in the remark after example 4). In addition, Θ_n defined in (4.4) is non-empty as long as n is sufficiently large (note that the constants involved in Θ_n can be improved and are not that meaningful. We focused on a presentable result instead of finding the best constants).

In examples 1 – 3, $X = (X_1,\ldots, X_5)$ is uniformly distributed on $[-1, 1]^5$. The treatment A is then generated independently of X uniformly from $\{-1, 1\}$. Given X and A, the response R is generated from a normal distribution with mean $Q_0(X, A) = 1+2X_1 +X_2 +0.5X_3 +T_0(X, A)$ and variance 1. We consider the following three examples for T_0.

1. $T_0(X, A) = 0$ (i.e. there is no treatment effect).
2. $T_0(X, A) = 0.424(1 - X_1 - X_2)A$.
3. $T_0(X, A) = 0.446sign(X_1)(1 - X_1)^2A$.

Note that in each example $T_0(X, A)$ is equal to the treatment effect term, $Q_0(X, A) - E[Q_0(X, A)|X]$. We approximate Q_0 by $\mathcal{Q} = \{(1, X, A, XA)\boldsymbol{\theta}: \boldsymbol{\theta} \in \mathbb{R}^{12}\}$. Thus in examples 1 and 2 the treatment effect term T_0 is correctly modeled, while in example 3 the treatment effect term T_0 is misspecified.

The parameters in examples 2 and 3 are chosen to reflect a medium effect size according to Cohen's d index. When there are two treatments, the Cohen's d effect size index is defined as the standardized difference in mean responses between two treatment groups, i.e.

$$es=\frac{E(R|A=1) - E(R|A=-1)}{([Var(R|A=1)+Var(R|A=-1)]/2)^{1/2}}.$$

Cohen [7] tentatively defined the effect size as "small" if the Cohen's d index is 0.2, "medium" if the index is 0.5 and "large" if the index is 0.8.

In example 4, X is uniformly distributed on [0, 1]. Treatment A is generated independently of X uniformly from $\{-1, 1\}$. The response R is generated from a normal distribution with mean $Q_0(X, A)$ and variance 1, where

$Q_0(X, 1)=\sum_{j=1}^{8}\vartheta_{(1),j}1_{X<u_{(1),j}}$, $Q_0(X, -1)=\sum_{j=1}^{8}\vartheta_{(-1),j}1_{X<u_{(-1),j}}$, and ϑ's and u's are parameters specified in (A.12). The effect size is small.

$$\begin{aligned}(\vartheta_{(1),1},\ldots,\vartheta_{(1),8})&=(-0.781, 0.730, 0.635, 0.512, -2.278, 1.347, 1.155, -0.030);\\ (\vartheta_{(-1),1},\ldots,\vartheta_{(-1),8})&=(-2.068, 1.520, -0.072, -0.637, 1.003, -0.611, -0.305, 1.016);\\ (u_{(1),1},\ldots,u_{(1),8})&=(0.028, 0.144, 0.171, 0.298, 0.421, 0.443, 0.463, 0.758);\\ (u_{(-1),1},\ldots,u_{(-1),8})&=(0.061, 0.215, 0.492, 0.544, 0.6302, 0.650, 0.785, 0.909).\end{aligned} \quad (A.12)$$

We approximate Q_0 by Haar wavelets

$$Q=\left\{\theta_{(0),0}h_0(X)+\sum_{lk}\theta_{(0),lk}h_{lk}(X)+\left(\theta_{(0),1}h_0(X)+\sum_{lk}\theta_{(1),lk}h_{lk}(X)\right)A:\theta_{\cdot,\cdot}\in\mathbb{R}\right\},$$

where $h_0(x) = 1_{x\in[0,1]}$ and $h_{lk}(x) = 2^{l/2}\,(1_{2^l x\in[k+1/2,k+1)} - 1_{2^l x\in[k,k+1/2)})$ for $l = 0,\ldots, \bar{l}_n$. We choose $\bar{l}_n = \lfloor 3\log_2 n/4\rfloor - 2$. For a given l and sample $(X_i, A_i, R_i)_{i=1}^n$, k takes integer values from $\lfloor 2^l \min_i X_i\rfloor$ to $\lceil 2^l \max_i X_i\rceil - 1$. Then $J_n = 2^{\lfloor 3\log_2 n/4\rfloor} \le n^{3/4}$.

Remark

In example 4, we allow the number of basis functions J_n to increase with n. The corresponding theoretical result can be obtained by combining Theorem 3.1 and Theorem A. 1. Below we demonstrate the validation of the assumptions used in the theorems.

Theorem 3.1 requires that the randomization probability $p(a|x) \ge S^{-1}$ for a positive constant for all (x, a) pairs and the margin condition (3.3) or (3.6) holds. According the generative model, we have that $p(a|x) = 1/2$ and condition (3.6) holds.

Theorem A.1 requires Assumptions A.1 - Assumptions A.4 hold and Θ_n defined in (A.4) is non-empty. Since we consider normal error terms, Assumption A.1 holds. Note that the basis functions used in Haar wavelet are orthogonal. It is also easy to verify that Assumptions A.3 and A.4 hold with $\beta_n = 1$ and Assumption A.2 holds with $U_n = n^{3/8}/2$ and $\eta_{1,n} \le constant + constant \times \left\|\theta_n^*\right\|_0$ (since each $|\varphi_j\theta^*_{n,j}| = |\varphi_j E(\varphi_j R)| \le constant \times |\varphi_j|E|\varphi_j| \le O(1)$). Since Q_0 is piece-wise constant, we can also verify that $\left\|\theta_n^*\right\|_0 \le O(\log n)$. Thus for sufficiently large n, Θ_n is non-empty and (A.6) holds. The RHS of (A.5) converges to zero as $n \to \infty$.

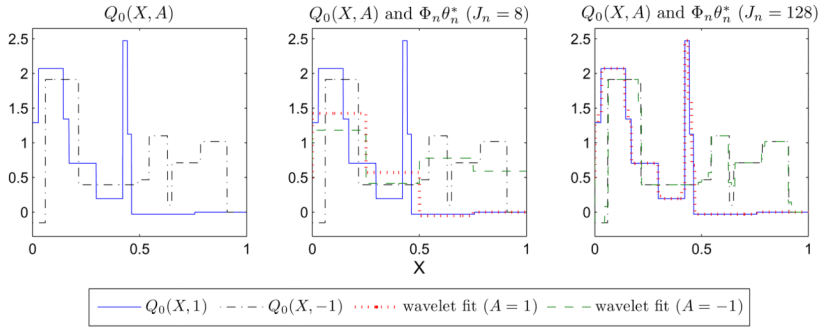

Fig 1.

Plots for: the conditional mean function $Q_0(X; A)$ (left), $Q_0(X; A)$ and the associated best wavelet fit when $J_n = 8$ (middle), and $Q_0(X; A)$ and the associated best wavelet fit when $J_n = 128$ (right) (example 4).

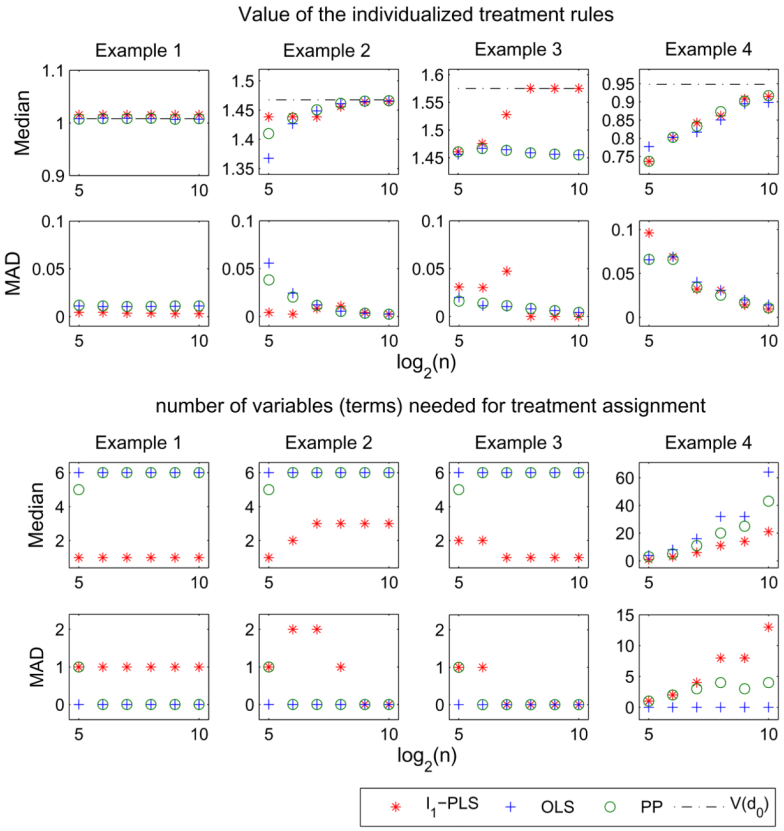

Fig 2.

Comparison of the l_1-PLS based method with the OLS method and the PP method (examples 1 – 4): Plots for medians and median absolute deviations (MAD) of the Value of the estimated decision rules (top panels) and the number of variables (terms) needed for treatment assignment (including the main treatment effect term, bottom panels) over 1000 samples versus sample size on the log scale. The black dash-dotted line in each plot on the first row denotes the Value of the optimal treatment rule, V (d_0), for each example. (n = 32; 64; 128; 256; 512; 1024. The corresponding numbers of basis functions in example 4 are J_n = 8; 16; 32; 64; 64; 128).