Taylor & Francis
Taylor & Francis Group

Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Personalised treatment assignment maximising expected benefit with smooth hinge loss

Shixue Liu, Jun Shao & Menggang Yu

STATISTICAL THEORY AND RELATED FIELDS, 2017
VOL. 1, NO. 1, 37–47
https://doi.org/10.1080/24754269.2017.1326080

Taylor & Francis
Taylor & Francis Group

Personalised treatment assignment maximising expected benefit with smooth hinge loss

Shixue Liu[a], Jun Shao[a,b] and Menggang Yu[c]

[a]Department of Statistics, University of Wisconsin, Madison, WI, USA; [b]School of Statistics, East China Normal University, Shanghai, China; [c]Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, WI, USA

ABSTRACT

In personalised medicine, the goal is to make a treatment recommendation for each patient with a given set of covariates to maximise the treatment benefit measured by patient's response to the treatment. In application, such a treatment assignment rule is constructed using a sample training data consisting of patients' responses and covariates. Instead of modelling responses using treatments and covariates, an alternative approach is maximising a response-weighted target function whose value directly reflects the effectiveness of treatment assignments. Since the target function involves a loss function, efforts have been made recently on the choice of the loss function to ensure a computationally feasible and theoretically sound solution. We propose to use a smooth hinge loss function so that the target function is convex and differentiable, which possesses good asymptotic properties and numerical advantages. To further simplify the computation and interpretability, we focus on the rules that are linear functions of covariates and discuss their asymptotic properties. We also examine the performances of our method with simulation studies and real data analysis.

ARTICLE HISTORY
Received 12 March 2017
Revised 15 May 2017
Accepted 30 April 2017

KEYWORDS
Convex loss; linear rules; oracle property; subgroup identification; weighted outcome learning

1. Introduction

Differential treatment effects are common in many diseases, because patients with different covariates, such as demographics, genomic information, treatment and outcome history, may not respond to treatments homogeneously. For example, molecularly targeted cancer drugs are mostly effective for patients with tumours expressing the targets (Lee et al., 2016; Ulloa-Montoya et al., 2013; Zhao et al., 2016). Significant heterogeneity exists in responses among patients with different baseline levels of psychiatric symptoms (Crits-Christoph et al., 1999; Kessler et al., 2016). Personalised medicine accounts for individual heterogeneity by constructing a treatment assignment rule as a function of patient covariates to maximise the treatment benefit measured by patient's response to the treatment. Thus, it has gained considerable popularity in clinical practice and medical research in recent years.

Typically, a treatment assignment rule is built based on a training data set consisting of patients' responses and covariates from a medical or clinical study. One statistical approach that has long been discussed and explored is to fit a model based on patient's response, covariates and treatment received, since the expected patient's benefit is usually related to some characteristics under this model, e.g., the conditional expectation of patient's response given the covariates and treatment. Because this approach largely depends on the model, misspecification of the model leads to unreliable treatment assignment. This issue becomes even more prominent when the number of covariates is large, which is often the case in clinical trials and medical research. In addition, different types of data, e.g., binary, continuous, time-to-event, and possibly mixed outcomes, have to be dealt with differently in this approach.

An alternative approach circumvents the need for outcome model specification by directly searching an assignment rule that maximises the expected patient's benefit (Zhao, Zeng, Rush, & Kosorok, 2012). This approach uses patients' responses (of any type) as weights in a target function that is related to patient's benefit as well as treatment-covariate interaction effects. Instead of using the zero–one loss in the target function, which is natural but hard or impossible for numerical implementation, efforts have been made by researchers in search of a surrogate loss function for the zero–one loss that can be easily implemented and leads to solutions with good properties; for example, the hinge loss considered by Zhao et al. (2012) and the logistic loss used by Xu et al. (2015). Unlike the zero–one loss, the hinge loss is convex and continuous, but not differentiable so that numerical optimisation under the hinge loss requires some additional techniques. Although the logistic loss is convex, differentiable, and simple for optimisation, unlike the hinge loss, it does not produce a solution that is exactly the same as the solution under the zero–one loss.

CONTACT Shixue Liu shao@stat.wisc.edu

The purpose of our study is to propose a sequence of convex and differentiable functions as the surrogate loss functions in applying the previously described direct searching approach. Since the limit of this sequence of loss functions is the hinge loss function and each function in this sequence has a hinge shape, we call these functions the smooth hinge loss functions. A detailed description of the target function and related loss functions are given in Section 2, where we also establish the equivalence of the optimal treatment assignment rule under the zero–one loss and hinge loss, and show that we may focus on rules linear in covariates under some conditions. An advantage of considering rules linear in covariates is that they are simple to compute and easy to interpret. In Section 3, we introduce the smooth hinge loss and our methodology. To address high covariate dimension issue, we propose to add a LASSO penalty in the optimisation under the smooth hinge loss. Some asymptotic properties of our solution are established. Numerical performance of our approach is examined by simulation in Section 4. Section 5 contains an example. All technical proofs are given in the Appendix.

2. Target function, hinge loss, and linear rules

Let $T \in \{-1, 1\}$ be a binary treatment with $P(T = 1) = \pi$, $0 < \pi < 1$, Y be a univariate patient's response, and X be a p-dimensional covariate vector including interactions and dummy variables for categorical covariates. We focus on the case where X and T are independent, e.g., a randomised trial. Let $D(X)$ be a treatment assignment rule as a function of X, P be the joint distribution of (Y, T, X), and P^D be the conditional distribution of (Y, T, X) given $T = D(X)$. Then, the expected outcome under rule D is given by (Qian & Murphy, 2011)

$$E^D(Y) = \int y\, dP^D = \int y \frac{dP^D}{dP} dP = E\left[\frac{YI\{T = D(X)\}}{T\pi + (1-T)/2}\right],$$

where $I\{A\}$ is the indicator function of a set A. Assume that a large Y is preferable and Y is bounded. Then, $E^D(Y)$ is the expected patient's benefit under zero–one loss and assignment rule D. Our goal is to assign each patient a treatment based on X to maximise $E^D(Y)$. Hence, we aim to find the optimal rule D^* that

$$\begin{aligned} D^* &= \underset{D \in \mathcal{D}}{\operatorname{argmax}}\, E\left[\frac{YI\{T = D(X)\}}{T\pi + (1-T)/2}\right] \\ &= \underset{D \in \mathcal{D}}{\operatorname{argmin}}\, E\left[\frac{YI\{T \neq D(X)\}}{T\pi + (1-T)/2}\right], \end{aligned} \tag{1}$$

where $\mathcal{D}$ is the set of all Borel functions of X with range $\{-1, 1\}$. Note that $D^*(X)$ does not change if Y is replaced by $Y + c$ for any constant c. We assume $Y \geq 0$.

Due to the discontinuity and nonconvexity of the zero–one loss (the indicator function), it is difficult to empirically perform the minimisation in (1). A common practice to mitigate this problem is to use a continuous and convex surrogate loss. Adopting the hinge loss as a surrogate loss, Zhao et al. (2012) proved that if

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, E\left[\frac{Y\{1 - Tf(X)\}^+}{T\pi + (1-T)/2}\right], \tag{2}$$

where $(1 - a)^+ = (1 - a)I\{a \leq 1\}$ and $\mathcal{F}$ is the set of all Borel functions, then $D^*(X) = \text{sign}\{f^*(X)\}$ a.s. We establish a stronger result as the following proposition. Its proof is given in the Appendix.

Proposition 2.1: *Let D^* and f^* be given by (1) and (2), respectively. Then, $f^*(X) = D^*(X)$ a.s.*

In general, the optimal rule $D^*(X)$ could be a complicated function of X. The following proposition shows that under a mild condition, $D^*(X)$ only depends on the sign of a linear function of X. This is a slightly modified version of Proposition 1 in Xu et al. (2015). Throughout the paper, a' denotes the transpose of a vector a.

Proposition 2.2: *Suppose that*

$$E(Y|T, X) = g(l(X), TX'\beta^\dagger), \tag{3}$$

where $\beta^\dagger$ is a p-dimensional vector, $l(X)$ is a function of X, and $g(a, b)$ is a bivariate function satisfying $g(a, b) \geq g(a, -b)$ if $b \geq 0$ any real valued a. Then, $D^*(X) = sign(X'\beta^\dagger)$ a.s.

Since $\beta^\dagger$ and functions l and g can all be unknown, many useful models in application satisfy (3).

With the hinge loss, can we perform the minimisation in (2) by focusing on linear functions of X? That is, if

$$\beta^* = \underset{\beta}{\operatorname{argmin}}\, E\left[\frac{Y(1 - TX'\beta)^+}{T\pi + (1-T)/2}\right], \tag{4}$$

can we conclude that $f^*(X) = \text{sign}(X'\beta^*)$? In general, the answer is no, because Propositions 1 and 2 imply that $D^*(X) = f^*(X) = \text{sign}(X'\beta^\dagger)$, which does not necessarily imply $f^*(X) = \text{sign}(X'\beta^*)$. But the following result indicates that $D^*(X) = f^*(X) = \text{sign}(X'\beta^*)$ holds under some conditions. The proof is in the Appendix.

Proposition 2.3: *Suppose that (3) holds and that*

$$E(X'\beta|X'\beta^\dagger) = c_\beta X'\beta^\dagger \quad \textit{for any } \beta, \tag{5}$$

where c_β is a constant that depends on β. Let ϕ be a continuous and convex function satisfying

$$[g(w, z) - g(w, -z)][\phi(cz) - \phi(-cz)] \leq 0 \tag{6}$$

for any real valued w and z and any $c > 0$. Then, there exist

$$\beta_\phi = \underset{\beta}{argmin}\, E\left[\frac{Y\phi(TX'\beta)}{T\pi + (1-T)/2}\right]$$

and $\tilde{c} \geq 0$ such that $\beta_\phi = \tilde{c}\beta^\dagger$.

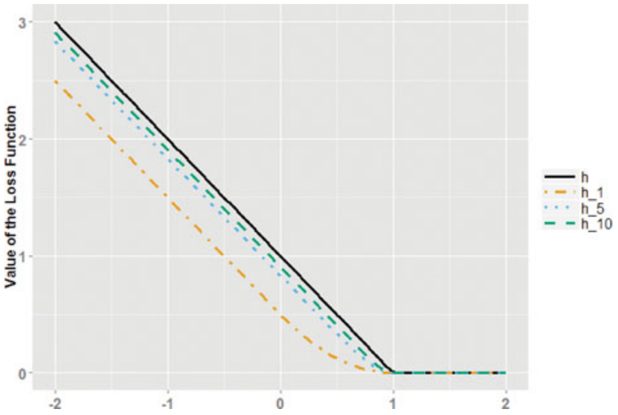

Figure 1. Hinge loss h and smooth hinge loss h_k.

For the hinge loss $\phi(z) = (1 - z)^+$, condition (6) is satisfied. Hence, $f^*(X) = \text{sign}(X'\beta^*)$ if condition (5) holds. Li (1991) proved that condition (5) is satisfied when the distribution of X is elliptically symmetric, e.g., X is multivariate normal.

The conclusion is that the hinge loss is a good surrogate for the zero–one loss and, under condition (5), we can focus on linear functions of X in the minimisation in (2).

3. Smooth hinge loss and optimal linear rule

Unfortunately, if we focus on linear rules, solving the minimisation problem (4) empirically is still difficult, because hinge loss is not always differentiable. We therefore propose to use a differentiable loss function to establish desirable asymptotic properties, as well as enhancing computational performance. We consider the following smooth hinge loss introduced by Rennie (2005):

$$h_k(z) = \begin{cases} \frac{k}{k+1} - z & z \leq 0 \\ \frac{k}{k+1} - z + \frac{1}{k+1} z^{k+1} & 0 < z < 1 \\ 0 & z \geq 1 \end{cases} \tag{7}$$

where $k \geq 1$ is an integer. The smooth hinge loss is twice differentiable for every k. For a fixed z, the smooth hinge loss function increases with k, and converges to the hinge loss function as k goes to infinity. Figure 1 shows the hinge loss and the smooth hinge loss with $k = 1, 5$, and 10.

With the smooth hinge loss being a surrogate loss, we focus on the optimisation problem

$$\beta_k^* = \underset{\beta}{\text{argmin}}\, E\left[\frac{Y h_k(TX'\beta)}{T\pi + (1 - T)/2}\right]. \tag{8}$$

The following proposition shows that, as $k \to \infty$, the minimum of the risk function with smooth hinge loss (achieved at β_k^*) converges to the minimum of the risk function with hinge loss (achieved at β^*). Furthermore, with the hinge loss, the risk function at β_k^* converges to the minimum of the risk function. The proof is in the Appendix.

Proposition 3.1: *Define* $R(\beta) = E[\frac{Y(1-TX'\beta)^+}{T\pi+(1-T)/2}]$ *and* $R_k(\beta) = E[\frac{Yh_k(TX'\beta)}{T\pi+(1-T)/2}]$ *to be the risk functions for the hinge loss and smooth hinge loss h_k, respectively. Then,*

$\lim_{k\to\infty} R_k(\beta_k^*) = R(\beta^*)$ *and* $\lim_{k\to\infty} R(\beta_k^*) = R(\beta^*)$.

In application, we can solve (8) with the expectation replaced by the sample average based on an independent sample (Y_i, T_i, X_i), $i = 1, \ldots, n$, identically distributed as (Y, T, X). To perform variable selection, we add the LASSO penalty and solve

$$\hat{\beta}_k = \underset{\beta}{\text{argmin}} \left\{\frac{1}{n}\sum_i \frac{Y_i h_k(T_i X_i'\beta)}{T_i\hat{\pi} + (1 - T_i)/2} + \lambda_n \|\beta\|_1\right\}, \tag{9}$$

where $\|\beta\|_1$ is the L_1 norm of β, λ_n is a LASSO tuning parameter, and $\hat{\pi}$ is the proportion of treatment group subjects in the data set. We define the final empirical treatment assignment rule as $\hat{D}(X) = \text{sign}(X'\hat{\beta}_k)$. Because a smooth hinge loss is used as the surrogate loss, we name this method as the sHinge method.

In the rest of this section, we show that $\hat{\beta}_k$ possesses a weak oracle property. We need some notations. For any β, let $\mathcal{M}_\beta = \{j : \beta_{(j)} \neq 0\}$ be the index set of non-zero components of β, where $\beta_{(j)}$ is the jth component of β. Let β_k^{*I} be the subvector of β_k^* with indices in $\mathcal{M}_{\beta_k^*}$, X_I be the subvector of X with indices in $\mathcal{M}_{\beta_k^*}$, and X_0 be the subvector of X with components not in X_I. Let s_p be the cardinality of $\mathcal{M}_{\beta_k^*}$ and $d_p = \min_{j\in\mathcal{M}_{\beta_k^*}} |\beta_{k,(j)}^*|$ be the minimal signal. For any $a = (a_{(1)}, \ldots, a_{(q)})$, let $\|a\|_\infty = \max_{j\leq q}|a_{(j)}|$. For any two sequence a_n and b_n, $a_n \asymp b_n$ denotes $a_n = O(b_n)$ and $b_n = O(a_n)$. Without loss of generality, we assume $E(X) = 0$.

The proof of the following result is in the Appendix.

Theorem 3.1 (Weak oracle property): *Assume that the following conditions hold.*

(C1) $\max_{1\leq j\leq p} Ee^{tX_{(j)}} \leq e^{ct^2/2}$ *for any t, where c is a constant and $X_{(j)}$ is the jth component of X.*

(C2) $0 \leq Y \leq M$ with a constant M.

(C3) $\log p \asymp n^{1-2\alpha_p}$, $s_p \asymp n^{\alpha_s}$, *and* $d_p \asymp n^{-\alpha_d}$, *where α_s, α_p, and α_d are constants satisfying $0 < \alpha_s < \gamma < \alpha_p < 1/2$ and $0 < \alpha_d < \gamma$ for a constant $\gamma > 0$.*

(C4) $\max_{\delta_1,\delta_2\in\mathcal{N}_0} \|[E\{W\ddot{h}_k(TX_I'\delta_1)X_0X_I'\}][E\{W\ddot{h}_k(TX_I'\delta_2)X_IX_I'\}]^{-1}\|_\infty < 1$ *and* $\max_{\delta\in\mathcal{N}_0} \|[E\{W\ddot{h}_k(TX_I'\delta)X_IX_I'\}]^{-1}\|_\infty = O(b_n)$, *where* $b_n = o(n^{1/2-\gamma}/(\log n)^{1/2})$, $\mathcal{N}_0 = \left\{\delta : \|\delta - \beta_k^{*I}\|_\infty \leq n^{-\gamma}\right\}$, *$\ddot{h}_k(x)$ is the second-order derivative of $h_k(x)$, and* $W = Y/\{T\pi + (1 - T)/2\}$.

(C5) $k \asymp n^\xi$ *with a constant* $\xi < \gamma - \alpha_s$.

If we choose λ_n such that

$$\lambda_n = o(n^{-\alpha_p}(\log n)^{1/2}) \quad \text{and} \quad \lambda_n b_n = o(n^{-\gamma}), \tag{10}$$

then, when n is sufficiently large, with probability greater than $1 - 4\{s_p/n + (p - s_p)e^{-n^{1-2\alpha_p}\log n}\}$, *there exists a* $\hat{\beta}_k$ *satisfying* (9) *and*

(a) (sparsity) $\mathcal{M}_{\hat{\beta}_k} = \mathcal{M}_{\beta_k^*}$;

(b) (L_∞ *consistency)* $\|\hat{\beta}_k^I - \beta_k^{*I}\|_\infty \le n^{-\gamma}$.

From Theorem 3.1, we know that under certain conditions, as n, p and k diverge to infinity in a certain way, with probability converging to one, $\hat{\beta}_k$ correctly identifies all the zero components of β_k^* and estimates the non-zero components of β_k^* consistently in the rate $n^{-\gamma}$. In other words, the penalised empirical solution $\hat{\beta}_k$ is a good estimate of the theoretical optimiser β_k^*, in the sense of weak oracle property. Furthermore, with the risk $R(\beta) = E[\frac{Y(1-TX'\beta)^+}{T\pi+(1-T)/2}]$ for the hinge loss, it follows from Theorem 3.1 and Proposition 3.1 that $\hat{\beta}_k$ is risk-consistent.

Corollary 3.1: *Under the conditions of Theorem 3.1,* $R(\hat{\beta}_k)$ *converges to* $R(\beta^*)$ *in probability as* $n \to \infty$.

4. Simulation results

In this section, we compare by simulation the proposed sHinge method with the ROWSi method in Xu et al. (2015), which solves (9) with h_k replaced by the logistic loss $\phi(t) = \log(1 + e^{-t})$. The reason why we compare our sHinge method with the ROWSi is that Xu et al. (2015) showed by simulation that the ROWSi was superior over the method solving (9) with h_k replaced by the hinge loss, which was proposed in Zhao et al. (2012) except that LASSO penalty instead of L_2 penalty was used for variable selection. Xu et al. (2015) also showed by simulation that ROWSi was superior over other four recently proposed methods, the interaction tree by Su, Tsai, Wang, Nickerson, and Li (2009), the virtual twins by Foster, Taylor, and Ruberg (2011), the logistic regression with LASSO penalty by Qian and Murphy (2011), and the FindIt by Imai and Ratkovic (2013).

For our method, $\hat{\beta}_k$ was computed by optimising the target function in (9) with coordinate descent algorithm (Friedman, Hastie, Höfling, & Tibshirani, 2007). The parameter k in h_k was set to be 2 in all the settings. To implement ROWSi, we used the R codes provided by the authors, in which a group LASSO procedure was used to pre-screen the interaction terms. For both methods, we chose the tuning parameter λ by a 10-fold cross-validation. The criterion used in validation was the prediction accuracy, which is defined as the proportion of empirical treatment assignments $\hat{D}(X) = \text{sign}(X'\hat{\beta}_k)$ that are consistent with the oracle assignments $\text{sign}(x'\beta^\dagger)$ (Xu et al., 2015).

We considered three types of covariates: binary covariates ($X_A, X_B, X_C, \ldots$), discrete covariates ($X_a, X_b, X_c, \ldots$) with four categories, and continuous covariates ($X_{Ca}, X_{Cb}, X_{Cc}, \ldots$). Deviation coding was used for all discrete covariates: a binary X_A was coded as ±1 and a X_a with four categories was coded as

	Coded variables		
X_a	X_{a1}	X_{a2}	X_{a3}
1	1	0	0
2	0	1	0
3	0	0	1
4	−1	−1	−1

All discrete covariates were simulated from the uniform distribution over their categories and all continuous covariates were simulated from the standard normal distribution. All covariates were generated independently. The treatment T takes values ±1 with equal probability and is independent of the covariates. We considered the following eight models between a binary response Y and (covariates, treatment) with up to three-way interactions among treatment and covariates. In the following, ϵ denotes a standard normal random variable independent of treatment and all covariates, and $I\{A\}$ denotes the indicator function of set A.

I. $X = (X_a, X_b, X_A, X_B, X_{Ca}, X_{Cb})$ and

$$Y = I\{(-0.5X_A + 0.5X_B)^2 + T(X_{Ca} + X_{Cb} + 0.9X_{a2}X_{Ca} + 0.9X_{b3}X_{Cb}) + \sqrt{2}\epsilon \ge 0\}$$

II. $X = (X_a, X_b, X_A, X_B, X_{Ca}, X_{Cb})$ and

$$\log\frac{P(Y=1|X)}{P(Y=0|X)} = -0.5X_A + 0.5X_B + T(X_{Ca} + X_{Cb} + 0.9X_{a2}X_{Ca} + 0.9X_{b3}X_{Cb})$$

III. $X = (X_a, X_b, X_c, X_A, X_B, X_C, X_{Ca}, X_{Cb}, X_{Cc})$ but Y is the same as that in (I), i.e., covariates X_c, X_C, and X_{Cc} are actually not related with Y.

IV. $X = (X_a, X_b, X_c, X_A, X_B, X_C, X_{Ca}, X_{Cb}, X_{Cc})$ but Y is the same as that in (II), i.e., covariates X_c, X_C, and X_{Cc} are actually not related with Y.

V. $X = (X_a, X_b, X_c, X_d, X_e, X_A, X_B, X_C, X_D, X_E, X_{Ca})$ and

$$Y = I\{(-0.5X_A + 0.5X_B)^2 + T(X_{a2} + X_{a3} + X_{b2}X_{Ca} + X_{b3}X_{Ca}) + \sqrt{2}\epsilon \ge 0\}$$

VI. $X = (X_a, X_b, X_c, X_d, X_e, X_A, X_B, X_C, X_D, X_E, X_{Ca})$ and

$$\log\frac{P(Y=1|X)}{P(Y=0|X)} = -0.5X_A + 0.5X_B + T(X_{a2} + X_{a3} + X_{b2}X_{Ca} + X_{b3}X_{Ca})$$

VII. $X = (X_{Ca}, X_{Cb}, X_{Cc}, X_{Cd}, X_{Ce}, X_{Cf}, X_{Cg}, X_{Ch})$ and

$$Y = I\{(-0.5X_{Ca} + 0.5X_{Cb})^2 + T(X_{Cd} + 0.9X_{Cd}X_{Ce}) + \sqrt{2}\epsilon \ge 0\}$$

VIII. $X = (X_{Ca}, X_{Cb}, X_{Cc}, X_{Cd}, X_{Ce}, X_{Cf}, X_{Cg}, X_{Ch})$ and

$$\log\frac{P(Y=1|X)}{P(Y=0|X)} = -0.5X_{Ca} + 0.5X_{Cb} + T(X_{Cd} + 0.9X_{Cd}X_{Ce})$$

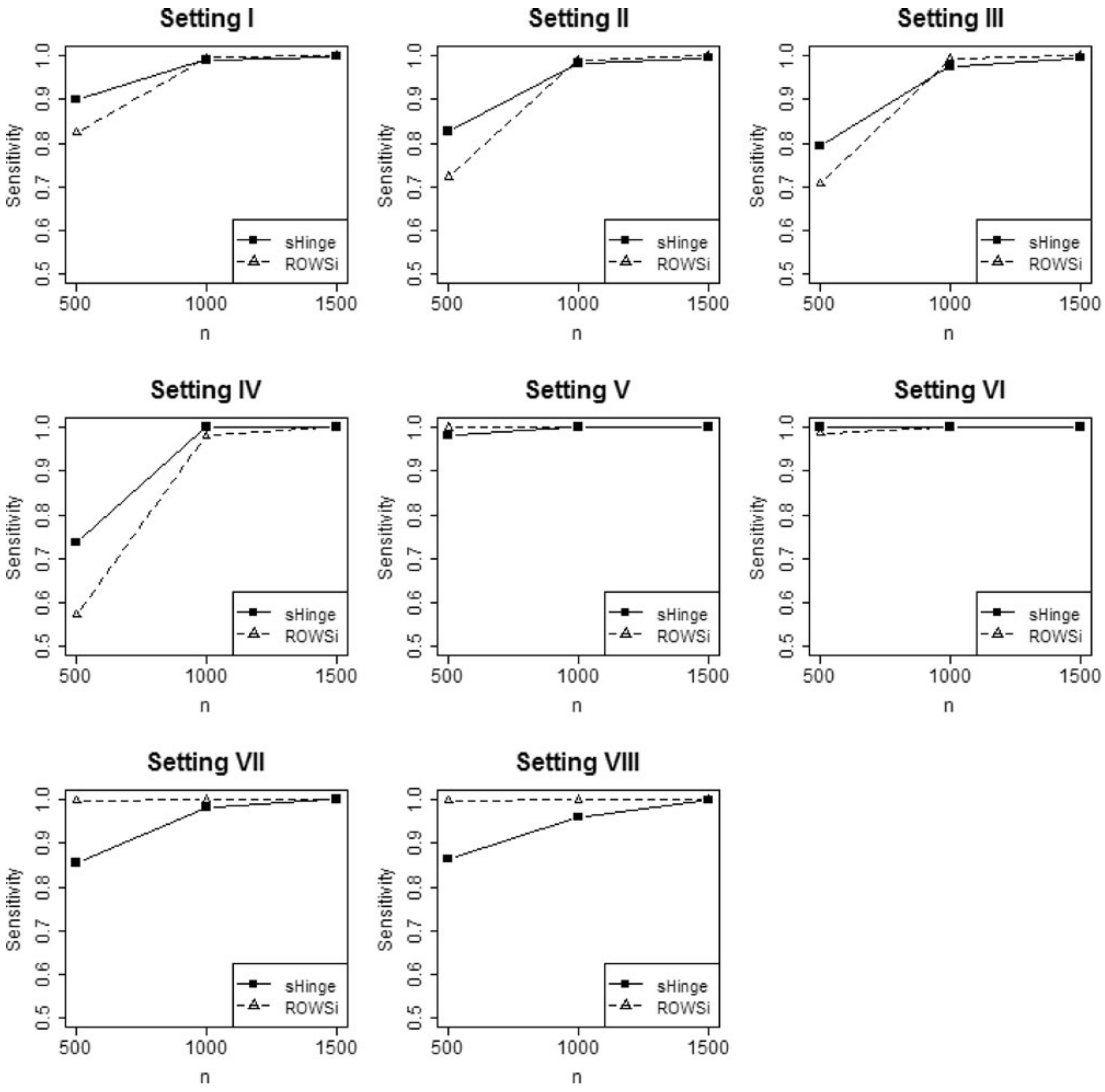

Figure 2. Comparison of sensitivity (sHinge vs. ROWSi).

In settings II, IV, VI and VIII, Y follows a logistic model. In settings I, III, V and VII, Y is an indicator of a regression; in these models, the conditions in Proposition 2.2 are violated. The purpose of including them was to test the robustness of the methods. In settings I and II, all covariates show up in the model, whereas in settings III and IV, one of three types of covariates is not related with Y. Models in V and VI contain only one continuous covariate and discrete covariates are dominant in the interaction terms. The last two models contain only continuous variables.

For each setting, three sample sizes, $n = 500$, 1000 and 1500, were used and the number of simulation runs was 200.

We evaluated the performances of both methods by three criteria: sensitivity, specificity and prediction accuracy. Sensitivity is defined as the proportion of true non-zero interactions being estimated as non-zero. Specificity is defined as the proportion of true zero interactions being estimated as zero. Prediction accuracy follows the definition in the discussion of the cross-validation procedure.

To compare the ability of variable selection, Figures 2 and 3 show sensitivity and specificity, respectively, based on 200 simulations. Both methods demonstrated exceptional ability to identify non-zero interaction terms, as shown in Figure 2. When the sample size is large, the sensitivity can even reach 1, meaning that all non-zero components are successfully identified. Figure 3 indicates that our method greatly surpasses ROWSi in the ability to identify unimportant interaction terms. The performance of our method is shown

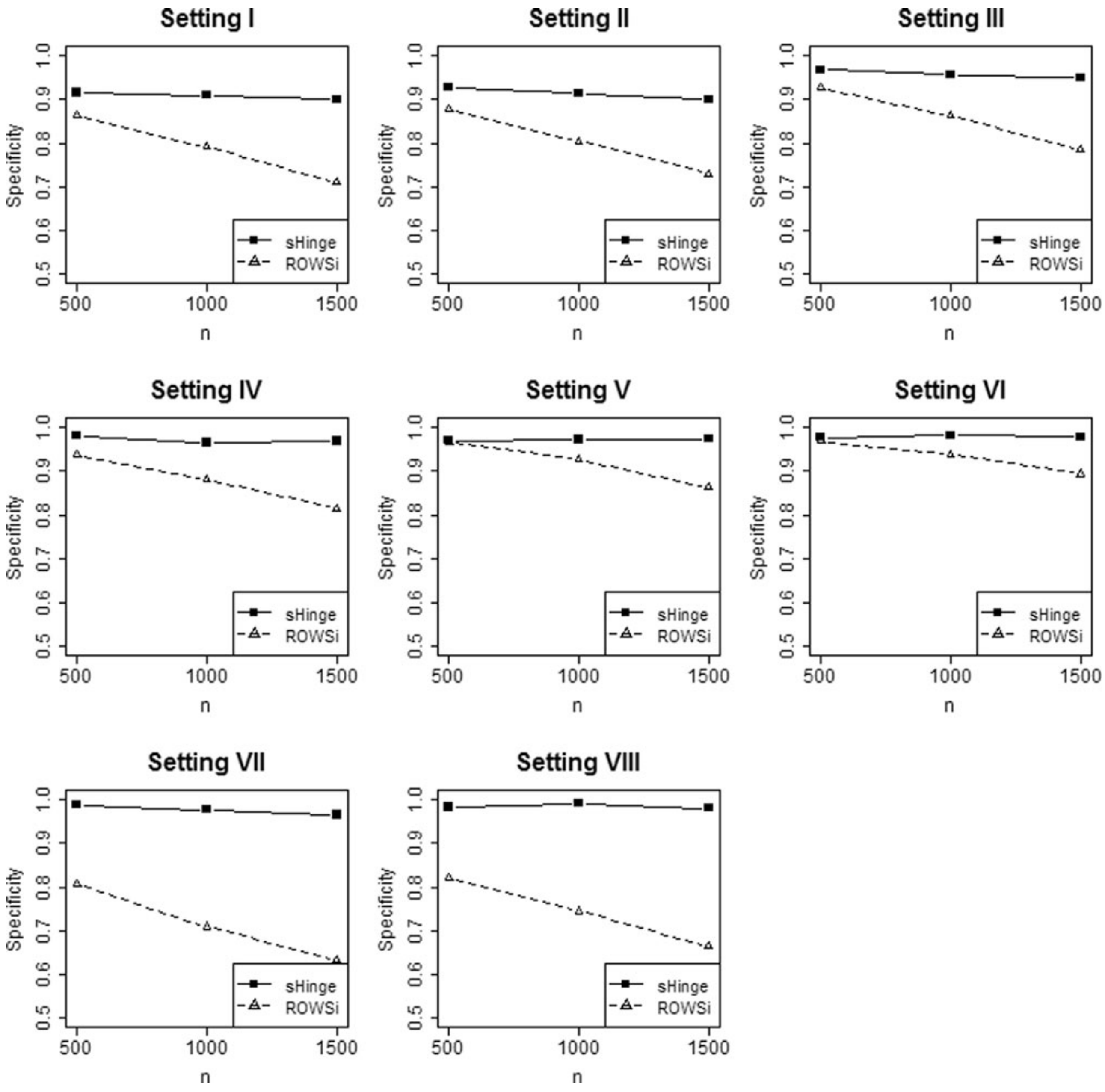

Figure 3. Comparison of specificity (sHinge vs. ROWSi).

to be stable across different sample sizes, while ROWSi has exhibited a decreasing trend in specificity as the sample size increases, which indicates that ROWSi may not have the consistency property described in our Theorem 3.1.

As for prediction accuracy, as shown in Figure 4, our method again excels ROWSi in almost all settings. The only exceptions are settings V and VI with $n = 500$, where the prediction accuracy of our method is lower by a small amount no larger than 0.0083.

5. Example

In this section, we apply our proposed method to a real data set from a large randomised trial that evaluates the efficacy of the telephone intervention at promoting mammography screening for women who were 51–75 years of age but non-adherent to breast cancer screening guidelines at baseline (Champion et al., 2016). The original study has three arms. We consider two arms that consist of 574 subjects in the phone intervention group and 544 in the usual care (control) group. The primary outcome of interest Y is a binary variable for mammography screening ($1 = \text{yes}$, $0 = \text{no}$) at 21 months post-baseline. The covariate vector X consists of 21 covariates, of which 16 are binary, one is categorical, and four are numerical. The covariates names and descriptions are listed in Table 1.

We intended to apply our method to obtain $\hat{D}(X) = \text{sign}(X'\hat{\beta}_k)$ with $\hat{\beta}_k$ given by (9), which determines the subgroup of subjects that would benefit from the phone intervention. However, the Y values were missing for 122 subjects. Assuming that missingness depends on

Table 1. Variables in the mammography screening data set.

resp21	Mammography screening 21 months post-baseline (yes/no)
treatment	1 = phone, −1 = control
age	Age
educyrs	Years of education
collegeormore	Four-year college or more (yes/no)
raceaa	African American (yes/no)
married	Married or in long-term relationship (yes/no)
income3	Household income ($1 \le 30$ K, $2 = 30$–75 K, $3 \ge 75$ K)
workpay	Currently working for pay (yes/no)
stgpca1	Baseline stage of behaviour change, 0 = precontemplation (not planning), 1 = contemplation (planning)
mediapaper	Exposure to paper media (yes/no)
mediatv	Exposure to TV media (yes/no)
mediainternet	Exposure to Internet media (yes/no)
hadmamm1	Ever had a mammogram (yes/no)
yearmamsum	Number of years had a mammogram in the past 2–5 years
doceversug	Doctor ever suggest you have a mammogram (yes/no)
docspoke	Doctor/nurse spoke to you in the last 2 years about mammogram (yes/no)
famhist	Family history of breast cancer (yes/no)
hcreminder	Received reminders of mammogram from a health care facility (yes/no)

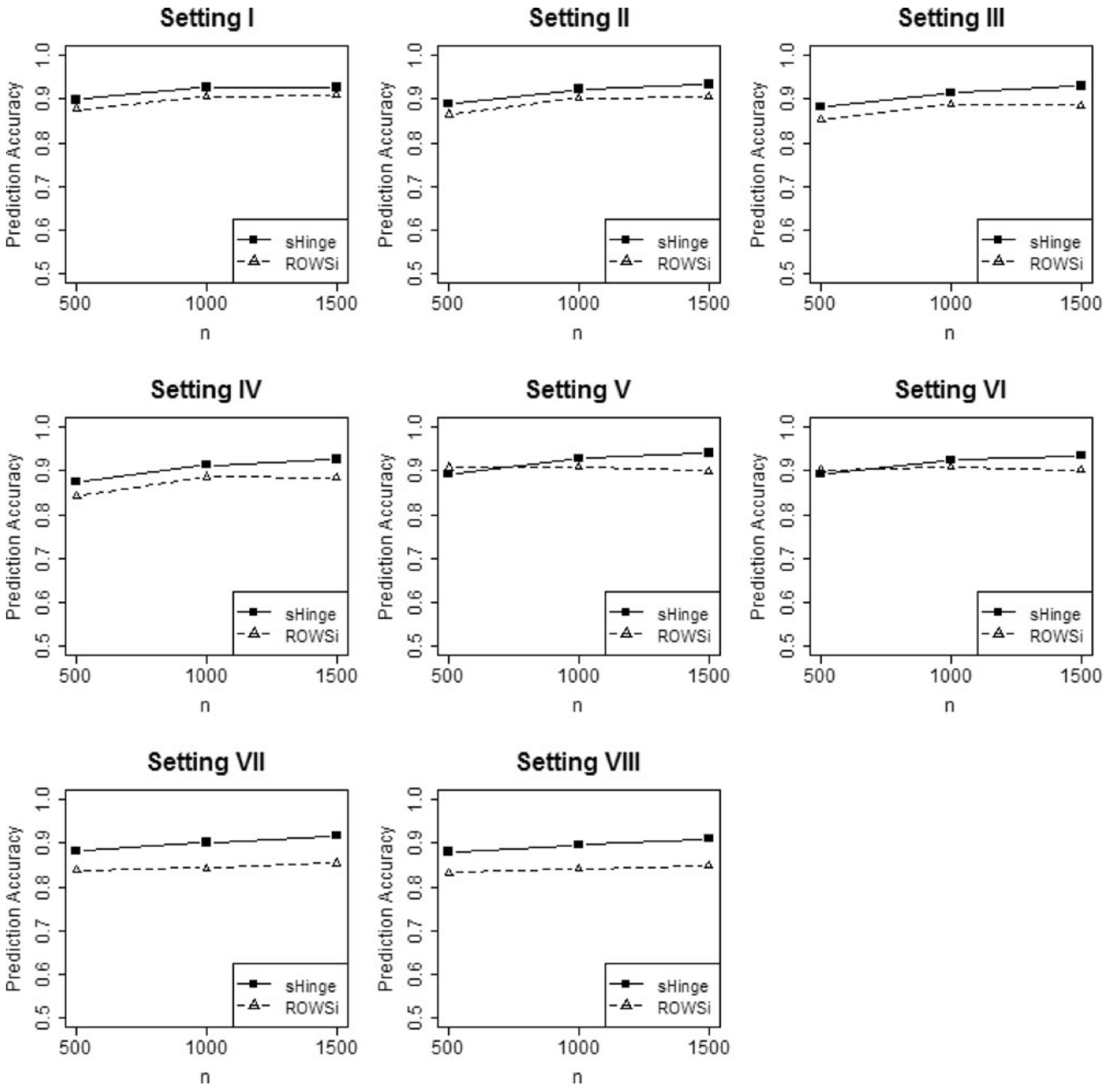

Figure 4. Comparison of prediction accuracy (sHinge vs. ROWSi).

covariates, we imposed a weight to each observed Y value and solved a modified version of (9):

$$\hat{\beta} \in \operatorname*{argmin}_{\beta} \left\{ \frac{1}{n} \sum_{i \in \mathcal{R}} \frac{Y_i h_1(T_i X_i' \beta)}{[T_i \hat{\pi} + (1 - T_i)/2] \hat{p}(X_i)} + \lambda_n \|\beta\|_1 \right\}, \tag{11}$$

where $\mathcal{R} = \{i : Y_i \text{ is observed}\}$, h_1 is the smooth hinge loss with parameter $k = 1$, $\hat{\pi}$ is the proportion of treatment group subjects in the data set, $\hat{p}(X_i)$ is an estimate of the propensity $P(Y_\text{i} \text{ is observed}|X_i)$ given by $\hat{p}(X_i) = \exp(X_i' \hat{\gamma})/[1 + \exp(X_i' \hat{\gamma})]$ with

$$\hat{\gamma} = \operatorname*{argmin}_{\gamma} \left\{ \frac{1}{n} \sum_{i \in \mathcal{R}}^{n} \left[\log(1 + \exp(X_i' \gamma)) - X_i' \gamma \right] + \upsilon_n \|\gamma\|_1 \right\}.$$

The tuning parameters λ_n and υ_n were selected by five-fold cross-validation with BIC being the criterion.

The resulting assignment rule is

$$\hat{D}_{\text{sHinge}} = \text{sign}(0.0065 \times \text{age} - 0.0241 \times \text{educyrs}).$$

This indicates that, for mammography screening, women who are non-adherent to breast cancer screening guidelines will benefit more from the phone intervention if they are

- in an older age group,
- had fewer years of education.

We also applied the ROWSi method in Xu et al. (2015), which simply replaces the loss function $h_1(\cdot)$ in (11) by the logistic loss $\log(1 + \exp(\cdot))$. However, it failed to select any covariates, i.e., it assigned all the subjects to the control group.

Acknowledgments

The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the Patient-Centered Outcomes Research Institute (PCORI), its Board of Governors or Methodology Committee.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Research reported in this article was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award [ME-1409-21219]. The second author's research was also partially supported by the Chinese 111 Project [B14019] and the US National Science Foundation [grant number DMS-1612873].

References

Champion, V. L., Rawl, S. M., Bourff, S. A., Champion, K. M., Smith, L. G., Buchanan, A. H., … Skinner, C. S. (2016). Randomized trial of dvd, telephone, and usual care for increasing mammography adherence. *Journal of Health Psychology, 21*, 916–926.

Crits-Christoph, P., Siqueland, L., Blaine, J., Frank, A., Luborsky, L. S., Onken, L. S., … Beck, A. T. (1999). Psychosocial treatments for cocaine dependence. *Archives of General Psychiatry, 56*, 493–502.

Foster, J., Taylor, J., & Ruberg, S. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine, 30*, 2867–2880.

Friedman, J., Hastie, T., Höfling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics, 1*, 302–332.

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

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Ebert, D. D., … Zaslavsky, A. M. (2016). Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. *Epidemiology and Psychiatric Sciences, 26*, 22–36.

Lee, C. K., Davies, L., Gebski, V. J., Lord, S. J., Di Leo, A., Johnston, S., … de Souza, P. (2016). Serum human epidermal growth factor 2 extracellular domain as a predictive biomarker for lapatinib treatment efficacy in patients with advanced breast cancer. *Journal of Clinical Oncology, 34*, 936–944.

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

Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *The Annals of Statistics, 39*, 1180–1210.

Rennie, J. D. M. (2005). Smooth hinge classification. Retrieved from http://people.csail.mit.edu/jrennie/writing

Su, X., Tsai, C.-L., Wang, H., Nickerson, D., & Li, B. (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research, 10*, 141–158.

Ulloa-Montoya, F., Louahed, J., Dizier, B., Gruselle, O., Spiessens, B., Lehmann, F. F., … Brichard, V. G. (2013). Predictive gene signature in mage-a3 antigen-specific cancer immunotherapy. *Journal of Clinical Oncology, 31*, 2388–2395.

Xu, Y., Yu, M., Zhao, Y.-Q., Li, Q., Wang, S., & Shao, J. (2015). Regularized outcome weighted subgroup identification for differential treatment effects. *Biometrics, 71*, 645–653.

Zhao, S. G., Chang, S. L., Spratt, D. E., Erho, N., Yu, M., Ashab, H. A.-D., … Feng, F. Y. (2016). Development and validation of a 24-gene predictor of response to postoperative radiotherapy in prostate cancer: A matched, retrospective analysis. *The Lancet Oncology, 17*, 1612–1620.

Zhao, Y., Zeng, D., Rush, A. J., & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association, 107*, 1106–1118.

Appendix

Proof of Proposition 2.1

For a.s. any fixed x,

$$E\left(\frac{Y(1 - Tf(x))^+}{T\pi + (1 - T)/2} \middle| X = x \right)$$
$$= E(Y|X = x, T = 1)(1 - f(x))^+$$

$$+\, E(Y|X = x, T = -1)(1 + f(x))^{+}.$$

Then, for f such that $f(x) \geq 1$,

$$E\left(\frac{Y(1 - Tf(x))^{+}}{T\pi + (1 - T)/2}\middle| X = x\right) = E(Y|X = x, T = -1)(1 + f(x)). \quad \text{(A1)}$$

For f such that $f(x) \leq -1$,

$$E\left(\frac{Y(1 - Tf(x))^{+}}{T\pi + (1 - T)/2}\middle| X = x\right) = E(Y|X = x, T = 1)(1 - f(x)). \quad \text{(A2)}$$

For f such that $-1 < f(x) < 1$,

$$\begin{aligned} &E\left(\frac{Y(1 - Tf(x))^{+}}{T\pi + (1 - T)/2}\middle| X = x\right) \\ &= (E(Y|X = x, T = -1) - E(Y|X = x, T = 1))f(x) \\ &\quad + E(Y|X = x, T = 1) + E(Y|X = x, T = -1) \\ &= -\mathcal{D}^{*}(x)f(x) \\ &\quad + E(Y|X = x, T = 1) + E(Y|X = x, T = -1). \end{aligned} \quad \text{(A3)}$$

The minimum of (A1) is $= 2E(Y|X = x, T = -1)$, obtained at f such that $f(x) = 1$. The minimum of (A2) is $= 2E(Y|X = x, T = 1)$, obtained at f such that $f(x) = -1$. When $\mathcal{D}^{*}(x) = 1$, the minimum of (A3) is $2E(Y|X = x, T = -1)$ obtained at f such that $f(x) = 1$. When $\mathcal{D}^{*}(x) = -1$, the minimum of (A3) is $2E(Y|X = x, T = 1)$ obtained at f such that $f(x) = -1$. Also, notice that when $\mathcal{D}^{*}(x) = 1$, the minimum of (A1) is smaller than the minimum of (A2), and when $\mathcal{D}^{*}(x) = -1$, the minimum of (A2) is smaller than the minimum of (A1). Therefore,

$$\min_{f} E\left(\frac{Y(1 - Tf(x))^{+}}{T\pi + (1 - T)/2}\middle| X = x\right) = \begin{cases} 2E(Y|X = x, T = -1) \text{ at } f^{*} \text{ such that } f^{*}(x) = 1 \\ \qquad\qquad \text{when } D^{*}(x) = 1 \\ 2E(Y|X = x, T = 1) \quad \text{at } f^{*} \text{ such that } f^{*}(x) = -1 \\ \qquad\qquad \text{when } D^{*}(x) = -1 \end{cases}.$$

Hence, $f^{*}(x) = \mathcal{D}^{*}(x)$ for a.s. any fixed x.

Proof of Proposition 2.3

For any β,

$$\begin{aligned} &E\left[\frac{Y\phi(TX'\beta)}{T\pi + (1 - T)/2}\right] \\ &= E\left\{E\left[\frac{Y\phi(TX'\beta)}{T\pi + (1 - T)/2}\middle| T, X\right]\right\} \\ &= E\left\{\frac{\phi(TX'\beta)}{T\pi + (1 - T)/2} g(l(X), TX'\beta^{\dagger})\right\} \\ &= E\left\{E\left[\frac{\phi(TX'\beta)}{T\pi + (1 - T)/2} g(l(X), TX'\beta^{\dagger})\middle| X'\beta^{\dagger}\right]\right\} \\ &= E\left\{g(l(X), X'\beta^{\dagger})E\left[\phi(X'\beta|X'\beta^{\dagger}, T = 1)\right]\right. \\ &\quad \left.+ g(l(X), -X'\beta^{\dagger})E\left[\phi(-X'\beta|X'\beta^{\dagger}, T = -1)\right]\right\} \\ &= E\left\{g(l(X), X'\beta^{\dagger})E\left[\phi(X'\beta|X'\beta^{\dagger})\right]\right. \\ &\quad \left.+\ g(l(X), -X'\beta^{\dagger})E\left[\phi(-X'\beta|X'\beta^{\dagger})\right]\right\} \\ &\geq E\left\{g(l(X), X'\beta^{\dagger})\phi\left(E(X'\beta|X'\beta^{\dagger})\right)\right. \\ &\quad \left.+\ g(l(X), -X'\beta^{\dagger})\phi\left(E(-X'\beta|X'\beta^{\dagger})\right)\right\} \\ &= E\left\{g(l(X), X'\beta^{\dagger})\phi(c_{\beta}X'\beta^{\dagger})\right. \\ &\quad \left.+\ g(l(X), -X'\beta^{\dagger})\phi(-c_{\beta}X'\beta^{\dagger})\right\}, \end{aligned}$$

where the fifth equality follows from the independence of T and X, the inequality is the result of an application of Jensen's inequality with conditional expectation, and the last equality follows from condition (5). Similarly, we can show that

$$\begin{aligned} &E\left[\frac{Y\phi(c_{\beta}TX'\beta^{\dagger})}{T\pi + (1 - T)/2}\right] \\ &\quad = E\left\{g(l(X), X'\beta^{\dagger})\phi(c_{\beta}X'\beta^{\dagger})\right. \\ &\qquad \left.+\, g(l(X), -X'\beta^{\dagger})\phi(-c_{\beta}X'\beta^{\dagger})\right\}. \end{aligned}$$

Thus, for any β,

$$\begin{aligned} E\left[\frac{Y\phi(TX'\beta)}{T\pi + (1 - T)/2}\right] &\geq E\left[\frac{Y\phi(c_{\beta}TX'\beta^{\dagger})}{T\pi + (1 - T)/2}\right] \\ &\geq \min_{c} E\left[\frac{Y\phi(cTX'\beta^{\dagger})}{T\pi + (1 - T)/2}\right]. \end{aligned}$$

Suppose that the minimum on the far right side of the previous expression is achieved at $\tilde{c}$. By definition, $\beta_{\phi} = \tilde{c}\beta^{\dagger}$. It remains to prove that $\tilde{c} \geq 0$. This follows because for any $c > 0$,

$$\begin{aligned} &E\left[\frac{Y\phi(cTX'\beta^{\dagger})}{T\pi + (1 - T)/2}\right] - E\left[\frac{Y\phi(-cTX'\beta^{\dagger})}{T\pi + (1 - T)/2}\right] \\ &\quad = \left[g(l(X), X'\beta^{\dagger}) - g(l(X), -X'\beta^{\dagger})\right] \\ &\qquad \times \left[\phi(cX'\beta^{\dagger}) - \phi(-cX'\beta^{\dagger})\right] \leq 0 \end{aligned}$$

by condition (6).

Proof of Proposition 3.1

By the definition, $\beta_{k}^{*} = \operatorname{argmin}_{\beta} R_{k}(\beta)$ and $\beta^{*} = \operatorname{argmin}_{\beta} R(\beta)$. From

$$h_{k}(z) - (1 - z)^{+} = \begin{cases} -\dfrac{1}{k + 1} & z \leq 0 \\ \dfrac{1}{k + 1} z^{k+1} - \dfrac{1}{k + 1} & 0 < z < 1 \\ 0 & z \geq 1 \end{cases},$$

we obtain that $\sup_{z} |h_{k}(z) - h(z)| = (k + 1)^{-1}$ and for any β,

$$\begin{aligned} |R_{k}(\beta) - R(\beta)| &\leq E\left[\frac{|Y||h_{k}(TX'\beta) - (1 - TX'\beta)^{+}|}{\min\{\pi, 1 - \pi\}}\right] \\ &\leq \frac{E|Y|}{(k + 1)\min\{\pi, 1 - \pi\}}. \end{aligned}$$

Hence, $\sup_{\beta} |R_{k}(\beta) - R(\beta)| \to 0$ as $k \to \infty$, i.e., R_{k} converges to R uniformly in β. Then,

$$\liminf_{k\to\infty} R_{k}(\beta_{k}^{*}) = \liminf_{k\to\infty} R(\beta_{k}^{*}) \geq R(\beta^{*})$$

by the definition of β^*. On the other hand, by the definition of β_k^*,

$$\limsup_{k\to\infty} R_k(\beta_k^*) \le \limsup_{k\to\infty} R_k(\beta^*) = R(\beta^*).$$

This proves $\lim_{k\to\infty} R_k(\beta_k^*) = R(\beta^*)$, which is the first conclusion. From the uniform convergence, $\lim_{k\to\infty}[R_k(\beta_k^*) - R(\beta_k^*)] = 0$. This together with the first conclusion proves the second conclusion $\lim_{k\to\infty} R(\beta_k^*) = R(\beta^*)$.

Proof of Theorem 3.1

Let $p_1(\beta) = \sum_{j=1}^{p} |\beta_{(j)}|$. Then, the subgradient of $p_1(\beta)$ is $\partial p_1(\beta_{(j)})$, a set-valued function such that

$$\partial p_1(\beta_{(j)}) = \begin{cases} \{1\}, & \text{if } \beta_{(j)} > 0; \\ \{-1\}, & \text{if } \beta_{(j)} < 0; \\ [-1, 1], & \text{if } \beta_{(j)} = 0. \end{cases} \tag{A4}$$

By the classical optimisation theory, the KKT condition is

$$n^{-1}\tilde{\boldsymbol{X}}'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}\beta) \circ \boldsymbol{W} - \boldsymbol{W}] + \lambda_n \boldsymbol{s} = 0, \tag{A5}$$

where $s_j \in \partial p_1(\beta_{(j)})$, $\boldsymbol{W} = (\frac{Y_1}{T_1\pi+(1-T_1)/2}, \ldots, \frac{Y_n}{T_n\pi+(1-T_n)/2})'$, $\tilde{\boldsymbol{X}} = (T_1X_1, \ldots, T_nX_n)'$, $\boldsymbol{\mu}(x)$ is a function from $\mathcal{R}^n$ to $\mathcal{R}^n$ such that its ith element is $\mu(x) = \dot{h}_k(x) + 1$, $\dot{h}_k$ is the first-order derivative of h_k, and $\circ$ denotes componentwise product. Then, any vector $\beta \in \mathcal{R}^p$ satisfying the following KKT conditions is a solution to (9):

$$n^{-1}\tilde{\boldsymbol{X}}_I'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}_I\beta^I) \circ \boldsymbol{W} - \boldsymbol{W}] + \lambda_n \text{sign}(\beta^I) = 0, \tag{A6}$$

$$\left\| n^{-1}\tilde{\boldsymbol{X}}_0'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}_I\beta^I) \circ \boldsymbol{W} - \boldsymbol{W}] \right\|_\infty < \lambda_n, \tag{A7}$$

where $\tilde{\boldsymbol{X}}_I$ is the submatrix of $\tilde{\boldsymbol{X}}$ with columns in $\mathcal{M}_\beta$ and $\tilde{\boldsymbol{X}}_0$ is the submatrix of $\tilde{\boldsymbol{X}}$ with columns not in $\mathcal{M}_\beta$, and β^I is the subvector of β with indices in $\mathcal{M}_\beta$.

In the following, we show that within a neighbourhood of β_k^*, a vector satisfying the KKT conditions exists and it also satisfies conclusions (a) and (b) in Theorem 3.1. Define

$$\boldsymbol{\epsilon}_1 = n^{-1}\tilde{\boldsymbol{X}}_I'\boldsymbol{W} - E(n^{-1}\tilde{\boldsymbol{X}}_I'\boldsymbol{W}),$$
$$\boldsymbol{\epsilon}_0 = n^{-1}\tilde{\boldsymbol{X}}_0'\boldsymbol{W} - E(n^{-1}\tilde{\boldsymbol{X}}_0'\boldsymbol{W}),$$
$$\boldsymbol{\xi}_1 = n^{-1}\tilde{\boldsymbol{X}}_I'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}\beta) \circ \boldsymbol{W}] - E\left(n^{-1}\tilde{\boldsymbol{X}}_I'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}\beta) \circ \boldsymbol{W}]\right),$$
$$\boldsymbol{\xi}_0 = n^{-1}\tilde{\boldsymbol{X}}_0'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}\beta) \circ \boldsymbol{W}] - E\left(n^{-1}\tilde{\boldsymbol{X}}_0'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}\beta) \circ \boldsymbol{W}]\right),$$

and events

$$E_1 = \{\|\boldsymbol{\epsilon}_1\|_\infty \le C_1\sqrt{\log n/n}\},$$
$$E_2 = \{\|\boldsymbol{\epsilon}_0\|_\infty \le C_1 n^{-\alpha_p}\sqrt{\log n}\},$$
$$E_3 = \{\|\boldsymbol{\xi}_1\|_\infty \le C_2\sqrt{\log n/n}\},$$
$$E_4 = \{\|\boldsymbol{\xi}_0\|_\infty \le C_2 n^{-\alpha_p}\sqrt{\log n}\},$$

where C_1 and C_2 are constants depending on c and M. Let ϵ_j be the jth component of $\boldsymbol{\epsilon}_1$. Under conditions (C1) and (C2), ϵ_j is sub-Gaussian, i.e., there exists a constant c_1 depending on c and M that

$$\max_{1\le j\le s_p} Ee^{t\epsilon_j} \le e^{c_1t^2/2}.$$

By the Hoeffding's bound for sub-Gaussian random variables, it holds that

$$\max_{1\le j\le s_p} P\left(|\epsilon_j| > \sqrt{2c_1 \log n/n}\right) \le 2\exp(-\log n) = 2/n.$$

Let $C_1 = \sqrt{2c_1}$, it follows from Bonferroni inequality that

$$P(E_1^c) = P\left(\|\boldsymbol{\epsilon}_1\|_\infty > C_1\sqrt{\log n/n}\right)$$
$$\le s_p \max_{1\le j\le s_p} P\left(|\epsilon_j| > \sqrt{2c_1 \log n/n}\right) \le 2s_p/n.$$

Similarly, we can show that

$$P(E_2^c) = P\left(\|\boldsymbol{\epsilon}_0\|_\infty > C_1 n^{-\alpha_p}\sqrt{\log n}\right)$$
$$\le 2(p - s_p)e^{-n^{1-2\alpha_p}\log n}.$$

Since $|\mu(x)| \le 1$, following the same technique, we can show that

$$P(E_3^c) = P\left(\|\boldsymbol{\xi}_1\|_\infty > C_2\sqrt{\log n/n}\right) \le 2s_p/n,$$

$$P(E_4^c) = P\left(\|\boldsymbol{\xi}_0\|_\infty > C_2 n^{-\alpha_p}\sqrt{\log n}\right)$$
$$\le 2(p - s_p)e^{-n^{1-2\alpha_p}\log n}.$$

Therefore,

$$P(E_1 \cap E_2 \cap E_3 \cap E_4)$$
$$\ge 1 - 4\{s_p/n + (p - s_p)e^{-n^{1-2\alpha_p}\log n}\}.$$

Next, we show that within event $E_1 \cap E_2 \cap E_3 \cap E_4$, there exists a solution to (A6) and (A7), and it satisfies (a) and (b). First, we prove that, when n is sufficiently large, there exists a solution to (A6) in the hypercube

$$\mathcal{N} = \{\delta \in \mathcal{R}^{s_p} : \|\delta - \beta_k^{*I}\|_\infty = n^{-\gamma}\}.$$

Since

$$\beta_k^* = \operatorname{argmin} E(Wh_k(\tilde{X}'\beta)),$$
$$W = \frac{Y}{T\pi + (1 - T)/2}, \quad \tilde{X} = TX,$$

and

$$\left|\frac{\partial}{\partial\beta} Wh_k(\tilde{X}'\beta)\right| = |\{\mu(\tilde{X}'\beta) - 1\}W\tilde{X}| \le C|X|,$$

which is integrable, it follows that

$$E\left(\{\mu(\tilde{X}'\beta_k^*) - 1\}W\tilde{X}\right) = \frac{\partial}{\partial\beta} E\left(Wh_k(\tilde{X}'\beta)\right)\Big|_{\beta=\beta_k^*} = 0. \tag{A8}$$

Then, the condition in (A6) is equivalent to

$$\begin{aligned} & n^{-1}\tilde{\boldsymbol{X}}_1'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}_I\delta) \circ \boldsymbol{W}] - n^{-1}\tilde{\boldsymbol{X}}_I'\boldsymbol{W} \\ & \quad - E\left(\{\mu(\tilde{X}_I'\beta_k^{*I}) - 1\}W\tilde{X}_I\right) = -\lambda_n \mathrm{sign}(\delta), \end{aligned}$$

or

$$\begin{aligned} & E\left(\mu(\tilde{X}_I'\delta)W\tilde{X}_I\right) - E\left(\mu(\tilde{X}_I'\beta_k^{*I})W\tilde{X}_I\right) \\ & \quad = \boldsymbol{\epsilon}_1 - \boldsymbol{\xi}_1 - \lambda_n \mathrm{sign}(\delta), \end{aligned}$$

where $\tilde{X}_I = TX_I$. By Taylor expansion,

$$\begin{aligned} & E\left(\mu(\tilde{X}_I'\delta)W\tilde{X}_I\right) - E\left(\mu(\tilde{X}_I'\beta_k^{*I})W\tilde{X}_I\right) \\ & \quad = E\left(WX_I\mu'(\tilde{X}_I'\bar{\delta})X_I'\right)(\delta - \beta_k^{*I}), \end{aligned}$$

where $\bar{\delta}$ lies on the line segment connecting δ and β_k^{*I} and the fact that $T^2 = 1$ is used. Thus, (A6) is equivalent to

$$E\left(W\mu'(\tilde{X}_I'\bar{\delta})X_IX_I'\right)(\delta - \beta_k^{*I}) - \boldsymbol{\epsilon}_1 + \boldsymbol{\xi}_1 + \lambda_n \mathrm{sign}(\delta) = 0$$

or $\boldsymbol{\Psi}(\delta) = 0$, where

$$\begin{aligned} \boldsymbol{\Psi}(\delta) = \delta - \beta_k^{*I} - & \left\{E\left(W\mu'(\tilde{X}_I'\bar{\delta})X_IX_I'\right)\right\}^{-1} \\ & (\boldsymbol{\epsilon}_1 - \boldsymbol{\xi}_1 - \lambda_n \mathrm{sign}(\delta)). \end{aligned}$$

Since $\bar{\delta} \in \mathcal{N}_0$, (C4) implies that

$$\left\|\left\{E\left(WX_1\mu'(\tilde{X}_I'\bar{\delta})X_I'\right)\right\}^{-1}\right\|_\infty = O(b_n).$$

With the choice of λ_n, it follows that

$$\begin{aligned} & \left\|\left\{E\left(WX_I\mu'(\tilde{X}_I'\bar{\delta})X_I'\right)\right\}^{-1}(\boldsymbol{\epsilon}_1 - \boldsymbol{\xi}_1 - \lambda_n \mathrm{sign}(\delta))\right\|_\infty \\ & \quad \le \left\|\left\{E\left(WX_I\mu'(\tilde{X}_I'\bar{\delta})X_I'\right)\right\}^{-1}\right\|_\infty \\ & \qquad \times (\|\boldsymbol{\epsilon}_1\|_\infty + \|\boldsymbol{\xi}_1\|_\infty + \lambda_n) = o(n^{-\gamma}). \end{aligned}$$

Then, for sufficiently large n, if $\delta_j - \beta_{k,(j)}^* = n^{-\gamma}$, $\boldsymbol{\Psi}(\delta_j) > 0$; if $\delta_j - \beta_{k,(j)}^* = -n^{-\gamma}$, $\boldsymbol{\Psi}(\delta_j) < 0$. By continuity of $\Psi(\delta)$, an application of Miranda's existence theorem shows that $\boldsymbol{\Psi}(\delta) = \boldsymbol{0}$ has a solution in $\mathcal{N}$, which is also a solution to (A6). Denote this solution by $\hat{\beta}_k^I$. Let $\hat{\beta}_k = (\hat{\beta}_k^I, 0)'$. Then, $\hat{\beta}_k$ satisfies (a) by definition.

Next, we prove that $\hat{\beta}_k$ satisfies (A7) and (b). By (A8),

$$E\left(\{\mu(\tilde{X}_I'\beta_k^{*I}) - 1\}W\tilde{X}_0\right) = 0.$$

Then,

$$\begin{aligned} & n^{-1}\tilde{\boldsymbol{X}}_0'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}_I\hat{\beta}_k^I) \circ \boldsymbol{W} - \boldsymbol{W}] \\ & \quad = n^{-1}\tilde{\boldsymbol{X}}_0'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}_I\hat{\beta}_k^I) \circ \boldsymbol{W} - \boldsymbol{W}] \\ & \qquad - E\left(\{\mu(\tilde{X}_I'\beta_k^{*I}) - 1\}W\tilde{X}_0\right) \\ & \quad = E\left(\mu(\tilde{X}_I'\hat{\beta}_k^I)W\tilde{X}_0\right) - E\left(\mu(\tilde{X}_I'\beta_k^{*I})W\tilde{X}_0\right) \\ & \qquad + \boldsymbol{\xi}_0 - \boldsymbol{\epsilon}_0. \end{aligned} \tag{A9}$$

By Taylor expansion,

$$\begin{aligned} & E\left(\mu(\tilde{X}_I'\hat{\beta}_k^I)W\tilde{X}_0\right) - E\left(\mu(\tilde{X}_I'\beta_k^{*I})W\tilde{X}_0\right) \\ & \quad = E\left(W\mu'(\tilde{X}_I'\tilde{\delta})X_0X_I'\right)(\hat{\beta}_k^I - \beta_k^{*I}), \end{aligned} \tag{A10}$$

where $\tilde{\delta}$ lies on the line segment connecting $\hat{\beta}_k^I$ and β_k^{*I}. Since $\hat{\beta}_k^I$ is the solution to $\boldsymbol{\Psi}(\delta) = 0$, it holds that

$$\begin{aligned} \hat{\beta}_k^I - \beta_k^{*I} = & \left\{E\left(W\mu'(\tilde{X}_I'\beta_k^{*I})X_IX_I'\right)\right\}^{-1} \\ & \times (\boldsymbol{\epsilon}_1 - \boldsymbol{\xi}_1 - \lambda_n \mathrm{sign}(\hat{\beta}_k^I)). \end{aligned} \tag{A11}$$

So $\hat{\beta}_k^I$ satisfies (b); that is, $\left\|\hat{\beta}_k^I - \beta_k^{*I}\right\|_\infty \le n^{-\gamma}$. By (A9), (A10) and (A11),

$$\begin{aligned} & (n\lambda_n)^{-1}\tilde{\boldsymbol{X}}_0'[\boldsymbol{\mu}(\tilde{\boldsymbol{X}}_I\hat{\beta}_k^I) \circ \boldsymbol{W} - \boldsymbol{W}] \\ & \quad = \lambda_n^{-1}E\left(W\mu'(\tilde{X}_I'\tilde{\delta})X_0X_I'\right)\left\{E\left(W\mu'(\tilde{X}_I'\bar{\delta})X_IX_I'\right)\right\}^{-1} \\ & \qquad \times (\boldsymbol{\epsilon}_1 - \boldsymbol{\xi}_1 - \lambda_n \mathrm{sign}(\hat{\beta}_k^I)) - \lambda_n^{-1}\boldsymbol{\epsilon}_0 + \lambda_n^{-1}\boldsymbol{\xi}_0. \end{aligned}$$

In the event $E_1 \cap E_2 \cap E_3 \cap E_4$, by the choice of λ_n,

$$\left\|\lambda_n^{-1}\boldsymbol{\epsilon}_0\right\|_\infty = o(1), \ \left\|\lambda_n^{-1}\boldsymbol{\xi}_0\right\|_\infty = o(1).$$

Note that $\tilde{\delta}, \bar{\delta} \in \mathcal{N}_0$. (C4) indicates that

$$\begin{aligned} & \lambda_n^{-1}\left\|E\left(W\mu'(\tilde{X}_I'\tilde{\delta})X_0X_I'\right)\right. \\ & \qquad \left. \times \left\{E\left(W\mu'(\tilde{X}_I'\bar{\delta})X_IX_I'\right)\right\}^{-1}(\boldsymbol{\epsilon}_1 - \boldsymbol{\xi}_1)\right\|_\infty \\ & \quad < \lambda_n^{-1}\left\|\boldsymbol{\epsilon}_1 - \boldsymbol{\xi}_1\right\|_\infty = o(1). \end{aligned}$$

Finally, again by (C4),

$$\begin{aligned} & \lambda_n^{-1}\left\|E\left(W\mu'(\tilde{X}_I'\tilde{\delta})X_0X_I'\right)\right. \\ & \qquad \left. \times \left\{E\left(W\mu'(\tilde{X}_I'\bar{\delta})X_IX_I'\right)\right\}^{-1}\lambda_n \mathrm{sign}(\hat{\beta}_k^I)\right\|_\infty < 1. \end{aligned}$$

Therefore, $\hat{\beta}_k$ satisfies (A7). This completes the proof.

Proof of Corollary 3.1

The result follows from Theorem 3.1, the dominated convergence theorem under (C1) and (C2), and Proposition 3.1.