Statistical Theory and Related Fields

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

Treatment recommendation and parameter estimation under single-index contrast function

Cui Xiong, Menggang Yu & Jun Shao

To cite this article: Cui Xiong, Menggang Yu & Jun Shao (2017) Treatment recommendation and parameter estimation under single-index contrast function, Statistical Theory and Related Fields, 1:2, 171-181, DOI: 10.1080/24754269.2017.1341012

To link to this article: https://doi.org/10.1080/24754269.2017.1341012

STATISTICAL THEORY AND RELATED FIELDS, 2017
VOL. 1, NO. 2, 171–181
https://doi.org/10.1080/24754269.2017.1341012

Taylor & Francis
Taylor & Francis Group

Treatment recommendation and parameter estimation under single-index contrast function

Cui Xiong[a], Menggang Yu[b] and Jun Shao[a,c]

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

ABSTRACT

In this article, we consider a semiparametric model for contrast function which is defined as the conditional expected outcome difference under comparative treatments. The contrast function can be used to recommend treatment for better average outcomes. Existing approaches model the contrast function either parametrically or nonparametrically. We believe our approach improves interpretability over the non-parametric approach while enhancing robustness over the parametric approach. Without explicit estimation of the nonparametric part of our model, we show that a kernel-based method can identify the parametric part up to a multiplying constant. Such identification suffices for treatment recommendation. Our method is also extended to high-dimensional settings. We study the asymptotics of the resulting estimation procedure in both low- and high-dimensional cases. We also evaluate our method in simulation studies and real data analyses.

ARTICLE HISTORY
Received 12 March 2017
Accepted 12 May 2017

KEYWORDS
Kernel weighting; LASSO penalty; personalised medicine; semiparametric model

1. Introduction

For most disease conditions, the benefits of some comparative treatments can differ substantially across different patient subpopulations. Such heterogeneity of treatment effects necessitates individualised treatment assignment as an important approach to improve patient outcomes. Indeed, there have been a recent growing literature on this particular topic of individualised treatment selection.

Roughly, existing literature adopt two types of approaches. The outcome modelling approach (Lu, Zhang, & Zeng, 2013; Taylor, Cheng, & Foster, 2015; Zhang, Tsiatis, Davidian, Zhang, & Laber, 2012) assumes an underlying outcome model and then derives treatment assignment rule from exploring the fitted outcome model; even though correctly fit, the outcome model is an overachievement because optimal treatment assignment only depends on the covariate–treatment interaction part of the outcome model. Therefore, this approach has been criticised due to the need for modelling the main effect of covariates on the outcome which is not related to optimal treatment assignment.

An alternative approach, also known as A-learning (Chen, Tian, Cai, & Yu, in press; Murphy, 2003; Robins, 2004; Zhao, Zeng, Rush, & Kosorok, 2012), directly models a contrast function, the conditional expected outcome difference under comparative treatments. Because the contrast function is directly linked to the goal of optimal treatment assignment, it can be more robust and efficient compared with the outcome modelling approach because it requires less modelling. The focus has mostly been on identifying the sign of the contrast function for subgroup identification. Nevertheless, modelling of the contrast function is necessary and focuses mainly on parametric (Lu et al., 2013; Murphy, 2003; Robins, 2004; Schulte, Tsiatis, Laber, & Davidian, 2014; Xu et al., 2015) and nonparametric (Zhang et al., 2012; Zhao et al., 2012; Zhou, Mayer-Hamblett, Khan, & Kosorok, 2016). In this article, we consider a semiparametric model, the so-called single-index model, for the contrast function. We believe our approach improves interpretability over the non-parametric approach. It is also more robust than the parametric approach due to its more flexible form.

Section 2 introduces notation, our semiparametric model, and a preliminary result that motivates our methodology. Without explicit estimation of the non-parametric part of our model, we show in Section 3 that a kernel-based method can identify the parametric part of our model up to a multiplying constant. Such identification suffices for treatment recommendation. Our method is also considered when the covariate has a high dimension but the useful part of the covariate is a subvector with a much lower dimension. In Section 4, we study the asymptotics of the resulting estimation procedure in both low- and high-dimensional cases. We also evaluate our method in simulation studies and compare it with two other recently developed methods. Finally, in Section 5, we apply our method to

CONTACT Jun Shao shao@stat.wisc.edu

two data sets from the national supported work study and the mammography screening study.

2. Model and preliminaries

Consider data collected from a randomised trial with a binary treatment $A \in \{0, 1\}$ being assigned according to $P(A=1)=\pi$. The clinical outcome is Y and, associated with Y, there is a p-dimensional covariate vector $Z=(1, Z_1, \cdots, Z_p)^T$ including the constant 1 as the first component, where a^T is the transpose of a vector a. Instead of specifying a model for Y given Z, we consider the following single-index model for the contrast function:

$$\triangle(Z) = E(Y|A=1, Z) - E(Y|A=0, Z) = g(\beta^T Z), \tag{1}$$

where g is increasing and differentiable, $g(0) = 0$, but otherwise is completely unknown. We make the following important remarks regarding our model specification:

(1) Under model (1), even if g is unknown, $\beta^T Z$ is still interpretable in the sense of treatment assignment. In particular, the ranking of $\triangle(Z)$ is fully captured by $\beta^T Z$. Thus, we can rank patients' benefit, in terms of $\triangle(Z)$, by $\beta^T Z$. We can also recommend a subgroup of patients with $\beta^T Z > \delta$ to treatment $A=1$ for some constant $\delta \geq 0$. In particular, a large $\delta > 0$ may be used if treatment $A=1$ is relatively more expensive, toxic, or hard to follow.

(2) Without any further assumption, actually we are able to estimate $c\beta$, instead of β itself, for an unknown constant $c=g'(0)$. In other words, for $\beta=(\beta_0, \beta_1, \ldots, \beta_p)^T$, what we can identify is $(\beta_1/\beta_j, \beta_2/\beta_j, \ldots, \beta_p/\beta_j)$, where β_j is a non-zero component of β. As long as $c>0$, we can still use $c\beta$ for our purpose.

(3) The requirement that g is increasing is not essential in our model specification. The case of decreasing g can be similarly treated.

(4) Because an intercept is included in Z, the condition $g(0) = 0$ is not a restrictive condition. If $g(0) \neq 0$ but $g(\alpha) = 0$, then model (1) can be rewritten as $\triangle(Z) = \tilde{g}(\tilde{\beta}^T Z)$ with $\tilde{g}(\cdot) \equiv g(\cdot + \alpha)$ and a suitably defined $\tilde{\beta}$.

In many cases, we may want to assign each individual to an appropriate treatment based on Z to optimise the average clinical outcome. Let $\mathcal{D}(Z)$ be an assignment rule based on Z. The expected outcome $E^{\mathcal{D}}(Y)$ under the rule $\mathcal{D}$ is given by Qian and Murphy (2011)

$$E^{\mathcal{D}}(Y) = E\left[\frac{I(A=\mathcal{D}(Z))}{A\pi + (1-A)(1-\pi)} Y\right],$$

where $I(\cdot)$ is the indicator function. We need to find the optimal rule $\mathcal{D}^*$ that maximises $E^{\mathcal{D}}(Y)$. Under model (1), the optimal $\mathcal{D}^*(Z)$ is $\mathrm{sign}(\beta^T Z)$, where sign is the sign function. Because $\mathrm{sign}(\beta^T Z) = \mathrm{sign}(c\beta^T Z)$ for any positive constant c, the solution of

$$\arg\max_{b\in\mathbb{R}^p} E\left[\frac{I\left\{A=\mathrm{sign}(b^T Z)\right\}}{A\pi + (1-A)(1-\pi)} Y\right]$$

is the set $\{c\beta\colon c \text{ is a positive constant}\}$. Note that the previous maximisation problem is equivalent to minimising

$$R(b) = E\left[\frac{I(A \neq \mathrm{sign}(bZ))}{\pi A + (1-A)(1-\pi)} Y\right], \tag{2}$$

which is hard to solve directly due to the indicator function.

We now derive the following fundamental result that facilitates our estimation in Section 3. For the function g defined in (1), define the risk

$$R_g(b) = E\left[\frac{\{Y-(A-1/2)g(b^T Z)\}^2}{A\pi + (1-A)(1-\pi)}\right]. \tag{3}$$

By conditioning, we know that $R_g(b)$ is the expectation of

$$\begin{aligned} W_Z(b) &= E\left[\left\{Y - 2^{-1}g(b^T Z)\right\}^2 |A=1, Z\right] \\ &\quad + E\left[\left\{Y + 2^{-1}g(b^T Z)\right\}^2 |A=0, Z\right]. \end{aligned}$$

Note that

$$\begin{aligned} \frac{\partial W_Z(b)}{\partial b} &= E\left[\left\{g(b^T Z) - 2Y\right\} g'(b^T Z) Z | A=1, Z\right] \\ &\quad + E\left[\left\{2Y + g(b^T Z)\right\} g'(b^T Z) Z | A=0, Z\right] \\ &= 2g'(b^T Z)\left\{-E(Y|A=1, Z)\right. \\ &\quad \left. + \ E(Y|A=0, Z) + g(b^T Z)\right\} Z \\ &= 2g'(b^T Z)\left\{-g(\beta^T Z) + g(b^T Z)\right\} Z, \end{aligned}$$

where g' is the derivative of g. Therefore,

$$\left.\frac{\partial W_Z(b)}{\partial b}\right|_{b=\beta} = 0.$$

Assume that g is second-order differentiable and let g'' be the second-order derivative of g. Then,

$$\begin{aligned} \frac{\partial^2 W_Z(b)}{\partial b^T \partial b} &= 2\frac{\partial}{\partial b^T}\left[g'(b^T Z)\left\{-g(\beta^T Z) + g(b^T Z)\right\} Z\right] \\ &= 2g''(b^T Z) Z Z^T \left\{-g(\beta^T Z) + g(b^T Z)\right\} \\ &\quad + 2\{g'(b^T Z)\}^2 Z Z^T \end{aligned}$$

and

$$\left.\frac{\partial^2 W_Z(b)}{\partial b^T \partial b}\right|_{b=\beta} = 2\{g'(\beta^T Z)\}^2 Z Z^T.$$

If g' is always positive, then from these results, we conclude that the minimiser of $R_g(b)$ is unique and equal to β in (1). Thus, the risk function $R_g(b)$ can be viewed as an approximation to $R(b)$ in (2) in terms of their minimisers.

3. Methodology and theory

Let $\{(Y_i, X_i, A_i), i = 1, \dots, n\}$ be a random sample of size n from the distribution of (Y, Z, A). The empirical version of $R_g(b)$ in (3) is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i - (A_i - 1/2)g(b^TZ_i)\}^2}{A_i\pi + (1-A_i)(1-\pi)}$$

From the derivation in the Section 2, if g were known, then we could estimate β by finding the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i - (A_i - 1/2)g(b^TZ_i)\}}{A_i\pi + (1-A_i)(1-\pi)}(1-2A_i)g'(b^TZ_i)Z_i = 0, \tag{4}$$

where the left-hand side of (4) is the derivative of the empirical version of $R_g(b)$. However, g is unknown and we cannot solve (4) directly. Consider the Taylor expansion of $g(b^TZ)$ at 0,

$$g(b^TZ) \approx g(0) + g'(0)(b^TZ) = g'(0)(b^TZ). \tag{5}$$

If $g'(0)(b^TZ)$ is a good approximation to $g(b^TZ)$, then we can estimate $g'(0)\beta$, which, from Remark 2 in Section 2, is enough for our purpose of recommending treatments for patients and identifying subgroups of enhanced treatment effect. But (5) is accurate only when b^TZ is near to 0. To overcome this, we use a kernel-based method. Let K be a symmetric probability density function (called a kernel) with support $[-1, 1]$ and

$$B_K = \int_{-1}^{1} u^2K(u)du < \infty \quad \text{and}$$
$$V_K = \int_{-1}^{1} K^2(u)du < \infty,$$

and let $h > 0$ be a bandwidth and $K_h(t) = K(t/h)/h$. Then, we replace (4) by the following kernel weighted version:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i - (A_i - 1/2)(b^TZ_i)\}}{\pi^{A_i}(1-\pi)^{1-A_i}}(1-2A_i)Z_iK_h(b^TZ_i) = 0. \tag{6}$$

The idea is that, we apply kernel weighting that has the effect of focusing on small values of $|b^TZ_i|$ when h is chosen to satisfy $h \to 0$, the kernel forces Equation (6) involves $|b^TZ_i|$ close to 0 so that approximation (5) is good, but h should not be too small, e.g., $nh \to \infty$ as $n \to \infty$, so that there are enough observations used in solving (6). Note that the solution to (6) estimates $g'(0)\beta$, $g'(0) > 0$. Although $g'(0)$ is unknown, it follows from the previous discussion that estimating $g'(0)\beta$ is enough for treatment recommendation and subgroup identification.

Theorem 3.1: *Let $\tilde{b}$ be a solution to (6). Assume that the kernel K satisfies the previously stated conditions; $h \to 0$ and $nh \to \infty$ as $n \to \infty$; Z has a density f; $\beta_j \neq 0$ for at least one $j \geq 1$ and without loss of generality $\beta_p \neq 0$. Then, as $n \to \infty$, we have the following conclusions:*

(i) $\tilde{b}$ converges in probability to $g'(0)\beta$.

(ii) If $nh^5 \to 0$, then $(nh)^{1/2}\{\tilde{b} - g'(0)\beta\}$ converges in distribution to the p-dimensional normal distribution with mean 0 and covariance matrix $\Sigma = Q^{-1}DQ^{-1}$, where

$$D = \frac{V_K}{|\beta_p|g'(0)}\int_{-1\le u\le 1}\left\{\frac{E(Y^2|A=1)}{\pi} + \frac{E(Y^2|A=0)}{1-\pi}\right\}\omega\omega^T f(\omega)dz_{-p},$$
$$Q = \frac{1}{2|\beta_p|g'(0)}\int \omega\omega^T f(\omega)\, dz_{-p},$$

$\omega = (z_1, \dots, z_{p-1}, -(\beta_0 + \beta_1 z_1 + \cdots + \beta_{p-1}z_{p-1})/\beta_p)^T$ and $dz_{-p} = dz_1 \cdots dz_{p-1}$.

(iii) The optimal choice of h is $h \asymp n^{-1/5}$, where $a \asymp b$ means $a = O(b)$ and $b = O(a)$.

We prove Theorem 3.1 in the Appendix.

In applications, we need to choose a bandwidth h for a given sample size n. There is a rich literature on bandwidth selection in applying a kernel method. A popular method is the cross-validation, which works by leaving out q populations at a time, and choosing the value of h that minimises

$$\mathrm{CV}(h) = \frac{1}{\lceil n/q\rceil}\sum_{i=1}^{\lceil n/q\rceil}\frac{1}{q}\sum_{j=(i-1)q+1}^{iq} \times \frac{I\{(2A_j-1)\tilde{b}_{-q,i}^TZ_j < 0\}}{\pi A_j + (1-A_j)(1-\pi)}Y_j, \tag{7}$$

where $\tilde{b}_{-q,i}$ is a solution to (6) with the data from units $j = (i-1)q - 1, \dots, iq$ deleted, and $\lceil n/q\rceil$ is the integer part of n/q. Note that each term in (7) is the loss when we classify unit j by using our constructed rule based on the data set without those from units with $k = (i-1)q - 1, \dots, iq$. Thus, $\mathrm{CV}(h)$ quantifies the classification accuracy of our method based on h.

In some modern applications, the dimension of Z in (1), p, is very high, although the number of non-zero components of β is much smaller than p, i.e., β is sparse. Hence, we propose to add a LASSO penalty and solve

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i - (A_i - 1/2)(b^TZ_i)\}}{A_i\pi + (1-A_i)(1-\pi)}(1-2A_i)Z_iK_h(b^TZ_i) + \lambda s(b) = 0, \tag{8}$$

where $\lambda \geq 0$ is a tuning parameter, $s(b)$ is the subgradient of $p(b) = \sum_{j=1}^{p}|b_j|$ whose jth component is $\mathrm{sign}(b_j)$ if $b_j \neq 0$ and c if $b_j = 0$, $0 < c < 1$, and b_j is the jth component of b, $j = 1, \dots, p$.

Let $\hat{b}$ be a solution to (8). We now show that $\hat{b}$ possesses a weak oracle property, namely with probability tending to 1, and $\hat{b}$ identifies all zero components of the true β and gives consistent estimators to non-zero components of β multiplied by a positive constant.

For any vector $\zeta = (\zeta_1, \ldots, \zeta_p)^T$, let $\mathcal{M}_\zeta = \{j : \zeta_j \neq 0\}$, $\zeta_{(1)}$ and $\zeta_{(0)}$ be the subvectors of ζ with indices in and not in $\mathcal{M}_\zeta$, respectively, $Z^{(1)}$ and $Z^{(0)}$ be the subvectors of Z with indices in and not in $\mathcal{M}_\beta$, respectively, and let s_p be the number of elements in $\mathcal{M}_\beta$. The proof of the following theorem is given in the Appendix.

Theorem 3.2: *Assume the conditions in Theorem 3.1 and the following conditions:*

- *(C1)* $\log p \asymp n^{1-2\alpha_p}$ *and* $s_p \asymp n^{\alpha_s}$, *where* $0 < \alpha_s < \alpha_p < 1/2$.
- *(C2)* $h = o(\lambda^{1/2})$ *and* $\lambda \asymp n^{-\alpha_\lambda}$, *where* $0 < \alpha_\lambda < \min\{2\gamma - \alpha_s, \alpha_p\}$ *and* $\lambda b_n = o(n^{-\gamma})$.
- *(C3)* $\max_{1\le j\le p} E e^{tZ_j} \le e^{ct^2/2}$ *for any real number t, where c is a constant.*
- *(C4)* $|Y| \le M_1$ *and* $\sup K_h \le M_2$, *where* M_1 *and* M_2 *are constants.*
- *(C5)* $\max_{1\le j\le p} \lambda_{\max}\{E|ZZ^T Z_j|\} = O(1)$, *where* $\lambda_{\max}(A)$ *is the maximal eigenvalue of A.*
- *(C6)* $b_n = \|E\{g'(\beta_{(1)}^T Z^{(1)}) Z^{(1)} Z^{(1)T} K_h(\beta_{(1)}^T Z^{(1)})\}^{-1}\|_\infty = o(\min\{n^{1/2-\gamma}/\sqrt{\log n}, n^{\gamma-\alpha_s}\})$, *where* $\alpha_s < \gamma < \alpha_p$.
- *(C7)* $\|E\{g'(\beta_{(1)}^T Z^{(1)}) Z^{(0)} Z^{(1)T} K_h(\beta_{(1)}^T Z^{(1)})\} E\{g'(\beta_{(1)}^T Z^{(1)}) Z^{(1)} Z^{(1)T} K_h(\beta_{(1)}^T Z^{(1)})\}^{-1}\|_\infty < 1$.

Then, with probability tending to 1,

- *(a) (sparsity)* $\mathcal{M}_\beta = \mathcal{M}_{\hat b}$.
- *(b) (L_∞ consistency)* $\|g'(0)\beta_{(1)} - \hat b_{(1)}\|_\infty \le n^{-\gamma}$.

4. Simulations

In this section, we perform some simulation studies to compare our proposed method with two other recently developed subgrouping methods, the Modified Covariate Method (MCM) by Tian et al. (2014) and the FindIt by Imai and Ratkovic (2013). We consider, respectively, the low-dimensional case and the high-dimensional case under the following model:

$$Y = (\beta^T Z/2)^2 + (A - 1/2)g(\beta^T Z) + \epsilon,$$

where $\epsilon \sim N(0, 0.3^2)$, ϵ, Z and A are independent, and g has the following three forms:

1. linear model: $g(\beta^T Z) = 7\beta^T Z$;
2. logistic model: $g(\beta^T Z) = 7\{\exp(\beta^T Z)/\{1 + \exp(\beta^T Z)\} - 1/2\}$;
3. probit model: $g(\beta^T Z) = 7\{\Phi(\beta^T Z) - 1/2\}$, where Φ is the standard normal distribution.

The treatment A takes 0 and 1 with equal probability. It can be seen that $\triangle(Z) = g(\beta^T Z)$.

We first consider a low-dimensional case, where $p = 3$, $\beta = (1, 1, 1, 1)^T$, Z_1, Z_2, and Z_3 are independently distributed as the standard normal. For $n = 200$, 500, and 1000, we calculate the simulation mean and root mean squared errors (rmse) of the ratio estimators $\tilde b_j/\tilde b_0$, $j = 1, 2, 3$, and the cover probabilities (cp) of the confidence intervals based on the bootstrap variance estimators with bootstrap size 1000. Since all methods produce negligible biases, we report the simulation rmse and cp in Table 1 based on 1000 simulation runs.

It can be seen from Table 1 that, in terms of rmse, our proposed method (ours) is much better than MCM and FindIt. The cp from our method is close to 95% and is better than that from MCM or FindIt, although in many cases the cp values are comparable.

Next, we consider a high-dimensional Z with $\beta = (\beta_0, \ldots, \beta_p)^T$, where $p = 23$, $\beta_j = 1$, $j = 0, 1, 2, 3$, and $\beta_j = 0$ for $j \ge 4$. $Z_1, \ldots, Z_p$ are still independently distributed as the standard normal. Other setting are

Table 1. Simulation results for ratio estimation in low-dimensional case.

			Linear			Probit			Logistic		
n	Quantity	Estimate	Ours	MCM	FindIt	Ours	MCM	FindIt	Ours	MCM	FindIt
200	rmse	$\tilde b_1/\tilde b_0$	0.013	0.041	0.071	0.109	0.350	0.132	0.126	0.418	0.134
		$\tilde b_2/\tilde b_0$	0.014	0.041	0.072	0.109	0.340	0.132	0.131	0.462	0.143
		$\tilde b_3/\tilde b_0$	0.014	0.042	0.070	0.104	0.325	0.133	0.130	0.430	0.135
	cp	$\tilde b_1/\tilde b_0$	0.947	0.949	0.939	0.943	0.942	0.959	0.950	0.947	0.951
		$\tilde b_2/\tilde b_0$	0.948	0.944	0.942	0.945	0.951	0.951	0.960	0.940	0.961
		$\tilde b_3/\tilde b_0$	0.953	0.951	0.939	0.941	0.956	0.957	0.943	0.951	0.956
500	rmse	$\tilde b_1/\tilde b_0$	0.008	0.027	0.044	0.062	0.179	0.081	0.071	0.195	0.082
		$\tilde b_2/\tilde b_0$	0.008	0.027	0.046	0.059	0.164	0.078	0.075	0.206	0.084
		$\tilde b_3/\tilde b_0$	0.008	0.026	0.046	0.059	0.165	0.080	0.074	0.192	0.081
	cp	$\tilde b_1/\tilde b_0$	0.953	0.923	0.960	0.950	0.947	0.952	0.955	0.943	0.938
		$\tilde b_2/\tilde b_0$	0.948	0.923	0.938	0.955	0.948	0.953	0.945	0.955	0.947
		$\tilde b_3/\tilde b_0$	0.947	0.955	0.955	0.952	0.945	0.938	0.945	0.945	0.957
1000	rmse	$\tilde b_1/\tilde b_0$	0.006	0.018	0.031	0.040	0.108	0.054	0.051	0.126	0.061
		$\tilde b_2/\tilde b_0$	0.006	0.019	0.033	0.041	0.112	0.054	0.050	0.130	0.060
		$\tilde b_4/\tilde b_0$	0.006	0.019	0.033	0.041	0.112	0.053	0.051	0.131	0.061
	cp	$\tilde b_1/\tilde b_0$	0.947	0.952	0.939	0.956	0.937	0.944	0.948	0.931	0.948
		$\tilde b_2/\tilde b_0$	0.951	0.950	0.943	0.950	0.939	0.962	0.952	0.937	0.937
		$\tilde b_3/\tilde b_0$	0.949	0.960	0.931	0.957	0.964	0.956	0.956	0.937	0.950

Table 2. Simulation results for ratio estimation and $P(\mathcal{M}_\beta = \mathcal{M}_{\hat{b}})$ in high-dimensional case.

			Linear			Probit			Logistic		
n	Quantity	Estimate	Ours	MCM	FindIt	Ours	MCM	FindIt	Ours	MCM	FindIt
200	cp	$\hat{b}_1/\hat{b}_0$	0.958	0.948	0.944	0.958	0.868	0.962	0.946	0.920	0.948
		$\hat{b}_2/\hat{b}_0$	0.948	0.939	0.949	0.962	0.895	0.962	0.967	0.930	0.949
		$\hat{b}_3/\hat{b}_0$	0.953	0.932	0.948	0.954	0.923	0.890	0.960	0.925	0.793
	$P(\mathcal{M}_\beta = \mathcal{M}_{\hat{b}})$		0.988	0.984	0.691	0.913	0.804	0.295	0.828	0.706	0.378
500	cp	$\hat{b}_1/\hat{b}_0$	0.944	0.977	0.954	0.949	0.925	0.937	0.947	0.948	0.966
		$\hat{b}_2/\hat{b}_0$	0.944	0.943	0.971	0.941	0.948	0.966	0.954	0.977	0.937
		$\hat{b}_3/\hat{b}_0$	0.947	0.937	0.948	0.960	0.954	0.977	0.962	0.954	0.971
	$P(\mathcal{M}_\beta = \mathcal{M}_{\hat{b}})$		1.000	1.000	0.897	1.000	0.962	0.814	0.997	0.972	0.894
1000	cp	$\hat{b}_1/\hat{b}_0$	0.948	0.938	0.968	0.950	0.938	0.942	0.953	0.940	0.953
		$\hat{b}_2/\hat{b}_0$	0.945	0.953	0.947	0.955	0.943	0.948	0.955	0.935	0.953
		$\hat{b}_3/\hat{b}_0$	0.953	0.962	0.972	0.950	0.937	0.952	0.948	0.965	0.950
	$P(\mathcal{M}_\beta = \mathcal{M}_{\hat{b}})$		1.000	1.000	0.970	1.000	1.000	0.984	1.000	1.000	0.978

the same as that for the low-dimensional case. Table 2 lists the simulated cp and $P(\mathcal{M}_\beta = \mathcal{M}_{\hat{b}})$. The simulated rmse is omitted.

It can be seen from Table 2 that, in terms of variable selection, our proposed method is better than MCM and FindIt. When variable selection is not accurate, it affects the performance of the cp.

5. Data analysis

In this section, we apply our proposed method to two real studies. The first is the national supported work (NSW) study (LaLonde, 1986) that appeared in FindIt" package based on Imai and Ratkovic (2013). The second is the mammography screening study (Champion et al., 2007). The NSW study corresponds to the low-dimensional case, whereas the mammography screening study involves a high-dimensional covariate.

5.1. National supported work study

In the NSW study, a training programme was administered to a heterogeneous group of workers. The treatment is randomly assigned to each subject. It is of interest to investigate whether the treatment effect varies as a function of individual characteristics. The treatment and control groups consist of 297 and 425 individuals, respectively. The original data set has nine covariates. To compare the methods without variable selection, we picked five covariates in the analysis: logarithm of annual earnings (log.re75), race (white or hispanic), marriage status (married or not), and high school degree status (nodegr). The other covariates were not included in the model fitting because they were not significant in all comparison methods. The response is whether there is an increase on earnings from the years 1975 to 1978. Based on the bootstrap method, we calculate the means and stand errors of estimates for these parameters, and, at the same time, we obtain the (0.025, 0.975) quantiles of the estimates. All results are shown in Table 3.

Our method indicated that being married had positive effects from the programme; being Hispanics and having no high school degree had negative effects. The MCM method found that being married and having higher annual earnings had positive effects from the programme; but being Hispanics had negative effects. The FindIt method found that being white had positive effects from the programme.

Table 3. Data analysis of NSW.

			intercept	hisp	white	married	nodegr	log.re75
Mean	Ours		0.030	− 0.052	− 0.002	0.037	− 0.065	0.017
	MCM		0.253	− 0.375	− 0.088	0.271	− 0.088	0.148
	FindIt		− 0.006	0.024	0.132	− 0.034	− 0.039	0.000
SD	Ours		0.030	− 0.052	− 0.002	0.037	− 0.065	0.017
	MCM		0.253	− 0.375	− 0.088	0.271	− 0.088	0.148
	FindIt		− 0.006	0.024	0.132	− 0.034	− 0.039	0.000
Quantile	Ours	Lower	0.013	− 0.077	− 0.031	0.011	− 0.090	− 0.009
		Upper	0.046	− 0.028	0.025	0.063	− 0.041	0.043
	MCM	Lower	0.197	− 0.522	− 0.225	0.128	− 0.232	0.022
		Upper	0.327	− 0.254	0.016	0.418	0.032	0.266
	FindIt	Lower	− 0.058	− 0.041	0.048	− 0.109	− 0.102	0.000
		Upper	0.036	0.117	0.223	0.017	0.012	0.000

Table 4. Data analysis of mammography screening study.

Ours			intercept	edu	yearmamsum	fatal1tot6	know1tot4
	Mean		0.036	0.023	0.034	0.047	− 0.025
	SD		0.010	0.014	0.012	0.014	0.015
	Quantile	Lower	0.016	0.005	0.010	0.020	− 0.054
		Upper	0.054	0.050	0.058	0.074	− 0.005
MCM			intercept	yearmamsum	se1tot40	sus1tot6	know1tot4
	Mean		− 0.827	0.376	− 0.267	− 0.199	0.302
	SD		0.050	0.095	0.095	0.093	0.099
	Quantile	Lower	− 0.926	0.197	− 0.446	− 0.381	0.108
		Upper	− 0.734	0.560	− 0.075	− 0.008	0.508
FindIt			intercept	age65	stage	sus1tot6	fear1tot20
	Mean		0.006	0.042	0.208	− 0.057	0.024
	SD		0.057	0.063	0.061	0.047	0.053
	Quantile	Lower	− 0.109	− 0.082	0.078	− 0.150	− 0.086
		Upper	0.127	0.175	0.322	0.035	0.134

To compare the performance of various methods, we randomly used 4/5 of samples (rounded to integers) as training sets to tune penalty parameters and the rest as test sets to evaluate the statistics:

$$\hat{R}(\tilde{b}) = \hat{E}\left[\frac{I\left\{A \neq \text{sign}(\tilde{b}^T Z)\right\}}{\pi A + (1-A)(1-\pi)} Y\right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \frac{I\{(2A_i - 1)\tilde{b}^T Z_i < 0\}}{\pi A_i + (1-A_i)(1-\pi)} Y_i.$$

We repeat this process 1000 times. The corresponding $\hat{R}(\tilde{b})$ were 0.4561 for our method, 0.5878 for MCM, and 0.4931 for FindIt, indicating the empirical superiority of our method.

5.2. *Mammography screening study*

This is a randomised study that included female subjects who were non-adherent to mammography screening guidelines at baseline (i.e., no mammogram in the year prior to baseline) (Champion et al., 2007). The outcome is whether the subject took mammography screening during this time period. There are 530 subjects with 259 in the phone intervention group and 271 in the usual care group. There are 16 binary variables, including socio-demographics, health belief variables, and stage of readiness to undertake mammography screening, and one categorical variable, number of years had a mammogram in the past 2–5 years. Our method indicated that fdrhistory and fatal1tot6 had positive effects from the programme, and settot40 and know1tot4 are negatively affected by the phone intervention. The MCM method found that stage, yearmamsum, docnursespoken2years and bar1tot people tended to get benefits, but workpay are negative. The FindIt found the stage and docursespoken2years were positively affected by the programme, but bar1tot30 and ben1tot30 are negative. All estimation results are given in Table 4.

Similar to the cross-validation procedure we used for the NSW study, we report the risk function $\hat{R}(\hat{b})$ under these three methods, 0.2773 for our method, 0.3217 for MCM, and 0.2988 for FindIt. The results again indicate the empirical superiority of our method.

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

This research was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) award [ME-1409-21219]. The first and third authors' research was partially supported by the Chinese Ministry of Education 111 Project [B14019] and the US National Science Foundation [grant number DMS-1305474], [grant number DMS-1612873].

Notes on contributors

Dr Cui Xiong holds a PhD in statistics from East China Normal University. She is now a statistician at GlaxoSmithKline in Shanghai, China. Her interests include survival analysis, multiple testing, and statistical methodology related to clinical trials.

Dr Menggang Yu holds a PhD in biostatistics from the University of Michigan. He is now a professor of biostatistics at the University of Wisconsin-Madison. Besides developing statistical methodology related to cancer research and clinical trials, Dr Yu is also very interested in health services research.

Dr Jun Shao holds a PhD in statistics from the University of Wisconsin-Madison. He is a professor of statistics at the University of Wisconsin-Madison. His research interests include

variable selection and inference with high dimensional data, sample surveys, and missing data problems.

References

Champion, V., Skinner, C. S., Hui, S., Monahan, P., Juliar, B., Daggy, J., & Menon, U. (2007). The effect of telephone vs. print tailoring for mammography adherence. *Patient Education and Counseling, 65*(3), 416.

Chen, S., Tian, L., Cai, T., & Yu, M. (in press). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics*. doi:10.1111/biom.12676

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

LaLonde, R. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review, 76*, 604–20.

Lu, W., Zhang, H. H., & Zeng, D. (2013). Variable selection for optimal treatment decision. *Statistical Methods in Medical Research, 22*(5), 493–504.

Murphy, S. A. (2003). Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65*(2), 331–355.

Qian, M., & Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of statistics, 39*(2), 1180.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics* (pp. 189–326). New York, NY: Springer.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., & Davidian, M. (2014). **Q**- and **A**-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science, 29*(4), 640–661.

Taylor, J. M. G., Cheng, W., & Foster, J. C. (2015). Reader reaction to a robust method for estimating optimal treatment regimes by Zhang et al. (2012). *Biometrics, 71*(1), 267–273.

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*, 645–53.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., & Laber E,. (2012). Estimating optimal treatment regimes from a classification perspective. *Statistics, 1*(1), 103–114.

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*(499), 1106–1118.

Zhou, X., Mayer-Hamblett, N., Khan, U., & Kosorok, M. R. (2016). Residual weighted learning for estimating individualized treatment rules. *Journal of the American Statistical Association, 39*(2), 1180–1210.

Appendix

Proof of Theorem 3.1: For the result in (i), it suffices to show that

$$G\left(g'(0)\beta\right) = \frac{1}{n}\sum_{i=1}^{n} \frac{\{Y_i - (A_i - 1/2)g'(0)\beta^T Z_i\}}{\pi^{A_i}(1-\pi)^{1-A_i}} \times (1-2A_i)Z_i K_h\left\{g'(0)\beta^T Z_i\right\} = o_p(1). \tag{A1}$$

Let $U = g'(0)\{\beta^T Z\}/h$. Then,

$$\begin{aligned}
&E\left[\frac{\{Y-(A-1/2)g'(0)\beta^T Z\}}{\pi^A(1-\pi)^{1-A}} \times (1-2A)ZK_h\left\{g'(0)\beta^T Z\right\}\right] \\
&= E\left[\{Y - g'(0)\beta^T Z/2\}(-Z)K_h\left\{g'(0)\beta^T Z\right\}\right] + E\left[\{Y + g'(0)\beta^T Z/2\}ZK_h\left\{g'(0)\beta^T Z\right\}\right] \\
&= E\left[\left\{-g(\beta^T Z) + g'(0)\beta^T Z\right\} ZK_h\left\{g'(0)\beta^T Z\right\}\right] \\
&= E\left[\left\{-g\left(\frac{Uh}{g'(0)}\right) + Uh\right\} Z\frac{K(U)}{h}\right] \\
&= E\left[\left\{-g(0) - g'(0)\frac{Uh}{g'(0)} - g''(0)\frac{U^2h^2}{2g'^2(0)} - g''(\xi)\frac{U^3h^3}{6g'^3(0)} + Uh\right\} Z\frac{K(U)}{h}\right] \\
&= h^2 E\left[\left\{-\frac{g''(0)}{2g'^2(0)} - \frac{g'''(\xi)Uh}{6g'^3(0)}\right\} U^2 Z\frac{K(U)}{h}\right] \\
&= -\frac{h^2 g''(0)}{2g'^2(0)} E\left[U^2 Z\frac{K(U)}{h}\right] + \frac{h^3 c_0}{6g'^3(0)} E\left|U^3 Z\frac{K(U)}{h}\right|,
\end{aligned} \tag{A2}$$

where ξ is between 0 and $hU/g'(0)$ and $c_0 = \max_t |g'''(t)|$. Consider the transformation

$$u = g'(0)\frac{\beta_1 + \beta_2 z_2 + \cdots + \beta_p z_p}{h}, \quad \text{and} \quad z_j = z_j, \; j = 2, \ldots, p-1.$$

Let $dz_{-p} = dz_2 \cdots dz_{p-1}$. For $j = 2, \ldots, p-1$, the jth component of $E\left[U^2 Z\frac{K(U)}{h}\right]$ is the integral

$$\begin{aligned}
&\frac{1}{h}\int_{-1\le u\le 1} u^2 z_j K(u) f(z) dz \\
&= \frac{1}{h}\int_{-1\le u\le 1} u^2 z_j K(u) f\left(z_2, \ldots, z_{p-1}, \frac{2uh}{\beta_p g'(0)} - \frac{(\beta^T z)_{-p}}{\beta_p}\right) \frac{2h}{|\beta_p| g'(0)} du dz_{-p} \\
&\xrightarrow{h\to 0} \int_{-1\le u\le 1} u^2 z_j K(u) f\left(z_2, \ldots, z_{p-1}, -\frac{(\beta^T z)_{-p}}{\beta_p}\right) \frac{2}{|\beta_p| g'(0)} du dz_{-p} \\
&= \frac{B_k}{|\beta_p| g'(0)} \int z_j f\left(z_2, \ldots, z_{p-1}, -\frac{(\beta^T z)_{-p}}{\beta_p}\right) dz_{-p} \\
&= \frac{B_k}{|\beta_p| g'(0)} \int z_j f(\omega)\, dz_{-p},
\end{aligned}$$

where $\omega = (z_2, \ldots, z_{p-1}, -(\beta^T z)_{-p}/\beta_p)^T$. Similarly, the first component of $E[U^2 Z\frac{K(U)}{h}]$ is $\frac{B_k}{|\beta_p| g'(0)} \int f(\omega)\, dz_{-p}$, and the pth component of $E\left[U^2 Z\frac{K(U)}{h}\right]$ is the integral

$$\frac{1}{h}\int_{-1\le u\le 1} u^2 z_p K(u) f(z) dz$$
$$= \frac{1}{|\beta_p| g'(0)} \int_{-1\le u\le 1} u^2 \left(\frac{uh}{\beta_p g'(0)} - \frac{(\beta^T z)_{-p}}{\beta_p}\right) K(u)$$
$$f\left(z_2, \ldots, z_{p-1}, \frac{uh}{\beta_p g'(0)} - \frac{(\beta^T z)_{-p}}{\beta_p}\right) dudz_{-p}$$
$$\xrightarrow{h\to 0} \frac{1}{|\beta_p| g'(0)} \int_{-1\le u\le 1} u^2 \left(-\frac{(\beta^T z)_{-p}}{\beta_p}\right) K(u)$$
$$f\left(z_2, \ldots, z_{p-1}, -\frac{(\beta^T z)_{-p}}{\beta_p}\right) dudz_{-p}$$
$$= \frac{1}{|\beta_p| g'(0)} B_k \int \left(-\frac{(\beta^T z)_{-p}}{\beta_p}\right) f(\omega)\, dz_{-p}.$$

Combining these results, we obtain that

$$E\left[U^2 Z\frac{K(U)}{h}\right] \to \frac{B_k}{|\beta_p| g'(0)} \int \omega f(\omega)\, dz_{-p}. \qquad \text{(A3)}$$

Combining these results, we obtain that the first term on the right-hand side of (A2)$\asymp h^2 \mathbf{1}_p$. Similarly, we can show that the second term on the right-hand side of (A2)$\asymp h^2 \mathbf{1}_p$. Hence,

$$E\Bigg[\frac{\{Y - (A - 1/2)(g'(0)\beta^T Z)\}}{\pi^A (1-\pi)^{1-A}} \times (1-2A) Z K_h\left\{g'(0)\beta^T Z\right\}\Bigg] \asymp h^2 \mathbf{1}_p. \qquad \text{(A4)}$$

Then, (A1) follows from the law of large numbers and (i) is proved. To prove the result in (ii), we can calculate that

$$\left.\frac{\partial G(b)}{\partial b}\right|_{g'(0)\beta} = \frac{1}{2n}\sum_{i=1}^n \left.\frac{(1-2A_i)^2 Z_i Z_i^T}{\pi^{A_i}(1-\pi)^{1-A_i}} K_h(b^T Z_i)\right|_{g'(0)\beta}$$
$$+ \frac{1}{n}\sum_{i=1}^n \frac{\{Y_i - (A_i - 1/2)(b^T Z_i)\}}{\pi^{A_i}(1-\pi)^{1-A_i}}$$
$$\times (1-2A_i) Z_i Z_i^T K_h'(b^T Z_i)\Big|_{g'(0)\beta}.$$

Using almost the same proof as that for (A4), we obtain that

$$E\Bigg[\frac{\{Y - (A - 1/2) g'(0)\beta^T Z\}}{\pi^A (1-\pi)^{1-A}} \times (1-2A) Z Z^T K_h'\left(\beta^T Z g'(0)\right)\Bigg]$$
$$= E\left[\left\{-g(\beta^T Z) + g'(0)(\beta^T Z)\right\} Z Z^T K_h'\left(\beta^T Z g'(0)\right)\right]$$
$$= E\left[\left\{-g\left(\frac{Uh}{g'(0)}\right) + Uh\right\} Z Z^T \frac{K'(U)}{h}\right]$$
$$= h^2 E\left[\left\{-\frac{g''(0)}{2g'^2(0)} - \frac{g'''(\xi) U h}{6 g'^3(0)}\right\} U^2 Z Z^T \frac{K'(U)}{h}\right]$$
$$\to 0$$

and, similar to the proof (A3),

$$E\left[Z Z^T K_h\left\{g'(0)\beta^T Z\right\}\right]$$
$$= \frac{1}{h} E\left[Z Z^T K(U)\right]$$
$$= \frac{1}{h}\int_{-1\le u\le 1} \begin{pmatrix} 1 & z^T \\ z & z z^T \end{pmatrix} K(u) f(z) dz$$
$$\to \frac{1}{|\beta_p| g'(0)} \int_{-1\le u\le 1} \begin{pmatrix} 1 & \omega^T \\ \omega & \omega\omega^T \end{pmatrix} K(u)$$
$$\times f\left(z_2, \ldots, z_{p-1}, -\frac{(\beta^T z)_{-p}}{\beta_p}\right) dudz_{-p}$$
$$= \frac{1}{|\beta_p| g'(0)} \int \begin{pmatrix} 1 & \omega^T \\ \omega & \omega\omega^T \end{pmatrix} f(\omega)\, dz_{-p}.$$

By the law of large numbers,

$$\left.\frac{\partial G(\beta)}{\partial \beta}\right|_{g'(0)\beta} \to \frac{1}{|\beta_p| g'(0)} \int \begin{pmatrix} 1 & \omega^T \\ \omega & \omega\omega^T \end{pmatrix} f(\omega)\, dz_{-p}$$

in probability. Let Q be the matrix on the right-hand side of the previous expression. By using the Taylor expansion of $G(\beta)$ at $\beta = g'(0)\beta$ and a standard argument, we can show that the asymptotic distribution of $(nh)^{1/2}\{\tilde{b} - g'(0)\beta\}$ is the same as the asymptotic distribution of $(nh)^{1/2} Q^{-1} G(g'(0)\beta)$, provided that this asymptotic distribution is not degenerated. To find the asymptotic distribution of $G(g'(0)\beta)$, we calculate the covariance matrix of $G(g'(0)\beta)$. From the result in the proof of (i), $E\{G(g'(0)\beta)\} \asymp h^2 \mathbf{1}_{p+1}$. Let $E_{21} = E[Y^2|A=1]$, $E_{20} = E[Y^2|A=0]$, and $E_{11} = E[Y|A=1]$ and $E_{10} = E[Y|A=0]$. Then,

$$\mathrm{Cov}\{G(g'(0)\beta)\}$$
$$= \frac{1}{nh^2} E\Bigg\{\Bigg[\frac{1}{\pi}E_{21} + \frac{1}{1-\pi}E_{20} - \left(\frac{1}{\pi}E_{11} - \frac{1}{1-\pi}E_{10}\right)$$
$$\times \left\{g'(0)\beta^T Z\right\} + \frac{1}{\pi(1-\pi)}\left\{g'(0)\beta^T Z\right\}^2\Bigg]$$
$$Z Z^T K^2\left\{\frac{g'(0)}{h}\beta^T Z\right\}\Bigg\} - \frac{1}{n} E\left\{G(g'(0)\beta) G(g'(0)\beta)\right\}^T$$
$$= \frac{1}{nh^2} E\Bigg\{\Bigg[\frac{1}{\pi}E_{21} + \frac{1}{1-\pi}E_{20} - \left(\frac{1}{\pi}E_{11} - \frac{1}{1-\pi}E_{10}\right) Uh$$
$$+ \frac{1}{\pi(1-\pi)}(Uh)^2\Bigg] Z Z^T K^2(U)\Bigg\}$$
$$-\frac{1}{n} E\left\{G(g'(0)\beta) G(g'(0)\beta)\right\}^T.$$

Also,

$$\frac{1}{h}E\Bigg\{\Bigg[\frac{1}{\pi}E_{21}+\frac{1}{1-\pi}E_{20}-\left(\frac{1}{\pi}E_{11}-\frac{1}{1-\pi}E_{10}\right)Uh + \frac{1}{\pi(1-\pi)}(Uh)^2\Bigg]ZZ^TK^2(U)\Bigg\}$$
$$\to \int_{-1\le u\le 1}\left(\frac{1}{\pi}E_{21}+\frac{1}{1-\pi}E_{20}\right)\omega\omega^T K^2(u) f(\omega)\frac{1}{|\beta_p|g'(0)}dudz_{-p}$$
$$=\frac{V_K}{|\beta_p|g'(0)}\int_{-1\le u\le 1}\left(\frac{1}{\pi}E_{21}+\frac{1}{1-\pi}E_{20}\right)\omega\omega^T f(\omega)dz_{-p}. \quad \text{(A5)}$$

As a result, the covariance matrix depends on $E[Y^2|A]$. Let D be the quantity in (A5). Then, the asymptotic covariance matrix Σ for $(nh)^{1/2}\{\tilde{b}-g'(0)\beta\}$ is $(nh)^{1/2}Q^{-1}DQ^{-1}$. This shows that each component of the matrix $\mathrm{Cov}\{G(g'(0)\beta)\}$ has the order $1/(nh)$, because $E\{G(g'(0)\beta)\}\{G(g'(0)\beta)\}^T$ has the order h^4. By the central limit theorem,

$$\sqrt{nh}[G(g'(0)\beta)-E\{G(g'(0)\beta)\}]\to N_p(0,D)$$

in distribution. Since $E\{G(g'(0)\beta)\}\asymp h^2\mathbf{1}_{p+1}$,

$$\sqrt{nh}G(g'(0)\beta)\to N_p(0,D)$$

in distribution, under the assumed condition on h. Therefore,

$$\sqrt{nh}\{\tilde{b}-g'(0)\beta\}\to N_p\left(0,Q^{-1}DQ^{-1}\right)$$

in distribution. This proves the result in (ii). From the proofs of (i)–(ii), the bias of $\tilde{b}$ as an estimator of $g'(0)\beta$ is of the order h^2 and the covariance matrix of $\tilde{b}$ is of the order $(nh)^{-1}$. Hence, the asymptotic mean squared error of $\tilde{b}$ is of the order $nh^{-1}+h^4$. Therefore, the best rate of convergence to 0 in mean squared error is achieved when $h\asymp n^{-1/5}$. This proves (iii). □

Proof of Theorem 3.2 By the classical optimisation theory, any vector $\hat{b}\in\mathcal{R}^p$ satisfying the following KKT conditions is a solution to (8):

$$\frac{1}{n}\sum_{i=1}^n\frac{y_i-(A_i-1/2)\hat{b}_{(1)}^T\hat{z}_i^{(1)}}{\pi_i^A(1-\pi)^{(1-A_i)}}\times(1-2A_i)\hat{z}_i^{(1)}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)})+\lambda_{1n}\mathrm{sign}(\hat{b}_{(1)})=0, \quad \text{(A6)}$$

$$\left\|\frac{1}{n}\sum_{i=1}^n\frac{y_i-(A_i-1/2)\hat{b}_{(1)}^T\hat{z}_i^{(1)}}{\pi_i^A(1-\pi)^{(1-A_i)}}\times(1-2A_i)\hat{z}_i^{(0)}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)})\right\|_\infty<\lambda_{1n}. \quad \text{(A7)}$$

In the following, we show that within a neighbourhood of $g'(0)\beta$, such a vector exists and satisfies (a) and (b). The result follows since the original problem (8) has a unique solution. Let

$$\epsilon_0=\frac{1}{n}\sum_{i=1}^n\frac{y_i(1/2-A_i)}{\pi^{A_i}(1-\pi)^{1-A_i}}\hat{z}_i^{(0)}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)}) - E\left[\frac{Y(1/2-A)}{\pi^A(1-\pi)^{1-A}}\hat{Z}^{(0)}K_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right],$$
$$\epsilon_1=\frac{1}{n}\sum_{i=1}^n\frac{y_i(1/2-A_i)}{\pi^{A_i}(1-\pi)^{1-A_i}}\hat{z}_i^{(1)}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)}) - E\left[\frac{Y(1/2-A)}{\pi^A(1-\pi)^{1-A}}\hat{Z}^{(1)}K_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right],$$
$$\xi_0=\frac{1}{n}\sum_{i=1}^n\frac{\hat{b}_{(1)}^T\hat{z}_i^{(1)}}{\pi^{A_i}(1-\pi)^{1-A_i}}\hat{z}_i^{(0)}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)}) - E\left[\frac{\hat{b}_{(1)}^T\hat{Z}^{(1)}}{\pi^A(1-\pi)^{1-A}}\hat{Z}^{(0)}K_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right],$$
$$\xi_1=\frac{1}{n}\sum_{i=1}^n\frac{\hat{b}_{(1)}^T\hat{z}_i^{(1)}}{\pi^{A_i}(1-\pi)^{1-A_i}}\hat{z}_i^{(1)}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)}) - E\left[\frac{\hat{b}_{(1)}^T\hat{Z}^{(1)}}{\pi^A(1-\pi)^{1-A}}\hat{Z}^{(1)}K_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right].$$

$E_1=\{\|\epsilon_1\|_\infty\le C_1\sqrt{\log n/n}\}$, $E_2=\{\|\epsilon_0\|_\infty\le C_1n^{-\alpha_p}\sqrt{\log n}\}$, $E_3=\{\|\xi_1\|_\infty\le C_2\sqrt{\log n/n}\}$ and $E_4=\{\|\xi_0\|_\infty\le C_2n^{-\alpha_p}\sqrt{\log n}\}$, where C_1 and C_2 are constants depending on c, M_1 and M_2. Condition (C3) ensures that Z_j is a sub-Gaussian random variable. It then follows from (C4) that $\frac{Y(1/2-A)}{\pi^A(1-\pi)^{1-A}}Z_jK_h(\beta^TZ)$ and $\frac{\beta^TZ}{\pi^A(1-\pi)^{1-A}}Z_jK_h(\beta^TZ)$ are also sub-Gaussian, i.e., there exist constants c_1 and c_2 depending on c, M_1 and M_2 that

$$\max_{1\le j\le p}E\exp\left\{t\frac{Y(1/2-A)}{\pi^A(1-\pi)^{1-A}}Z_jK_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right\}\le e^{c_1t^2/2}$$

and

$$\max_{1\le j\le p}E\exp\left\{t\frac{\hat{b}_{(1)}^T\hat{Z}^{(1)}}{\pi^A(1-\pi)^{1-A}}Z_jK_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right\}\le e^{c_2t^2/2}$$

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

$$\max_{1\le j\le s_p}P\left[\left|\frac{1}{n}\sum_{i=1}^n\frac{y_i(1/2-A_i)}{\pi^{A_i}(1-\pi)^{1-A_i}}z_{ij}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)}) - E\left\{\frac{Y(1/2-A)}{\pi^A(1-\pi)^{1-A}}Z_jK_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right\}\right| >\sqrt{2c_1\log n/n}\right]\le 2\exp(-\log n)=2/n.$$

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

$$P\left(\|\epsilon_1\|_\infty > C_1\sqrt{\log n/n}\right) \leq s_p \max_{1\leq j\leq s_p} P\Bigg[\Bigg|\frac{1}{n}\sum_{i=1}^{n}\frac{y_i(1/2-A_i)}{\pi^{A_i}(1-\pi)^{1-A_i}}z_{ij}K_h(\hat{b}_{(1)}^T\hat{z}_i^{(1)}) - E\left\{\frac{Y(1/2-A)}{\pi^A(1-\pi)^{1-A}}Z_jK_h(\hat{b}_{(1)}^T\hat{Z}^{(1)})\right\}\Bigg| \geq 2C_1\sqrt{\log n/n}\Bigg] \leq 2s_p/n.$$

Similarly, we can show that

$$P\left(\|\epsilon_0\|_\infty \leq C_1 n^{-\alpha_p}\sqrt{\log n}\right) \leq 2(p-s_p)e^{-n^{1-2\alpha_p}\log n}.$$

Following the same technique as in the above, we can show that

$$P\left(\|\xi_1\|_\infty > C_2\sqrt{\log n/n}\right) \leq 2s_p/n$$
$$P\left(\|\xi_0\|_\infty \leq C_2 n^{-\alpha_p}\sqrt{\log n}\right) \leq 2(p-s_p)e^{-n^{1-2\alpha_p}\log n}.$$

Therefore,

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

Next, we show that within event $E_1\cap E_2\cap E_3\cap E_4$, there exists a solution to (A6) and satisfies (a) and (b).

Step 1: we will prove that, when n is sufficiently large, there exists a solution to (A6) in the hypercube

$$\mathcal{N} = \left\{\delta \in \mathcal{R}^{s_p} : \left\|\delta - g'(0)\beta_{(1)}\right\|_\infty = n^{-\gamma}\right\}.$$

Based on (A6), we know that

$$E\left[\frac{Y-(A-1/2)\delta^TZ^{(1)}}{\pi^A(1-\pi)^{(1-A)}} \times (1-2A)Z^{(1)}K_h(\delta^TZ^{(1)})\right] = -\epsilon_1 - \xi_1 - \lambda_{1n}\text{sign}(\delta),$$

the left on is equal to

$$E\left[\left\{-g(\beta_{(1)}^TZ^{(1)}) + \delta^TZ^{(1)}\right\}Z^{(1)}K_h(\delta^TZ^{(1)})\right] = E\Bigg[\Bigg\{-g\left(\frac{\delta^TZ^{(1)}}{g'(0)}\right) - g'\left(\frac{\delta^TZ^{(1)}}{g'(0)}\right)Z^{(1)T} \times \left(\beta_{(1)} - \frac{\delta}{g'(0)}\right) - \frac{1}{2}g''\left(\frac{\tilde{\delta}^TZ^{(1)}}{g'(0)}\right) \times \left(\beta_{(1)}-\frac{\delta}{g'(0)}\right)^T Z^{(1)}Z^{(1)T}\left(\beta_{(1)}-\frac{\delta}{g'(0)}\right) + \delta^TZ^{(1)}\Bigg\}Z^{(1)}K_h(\delta^TZ^{(1)})\Bigg],$$

where $\tilde{\delta}$ lies on the line segment connecting δ and $g'(0)\beta_{(1)}$. Let

$$\tau = E\left[\left\{-g\left(\frac{\delta^TZ^{(1)}}{g'(0)}\right) + \delta^TZ^{(1)}\right\}Z^{(1)}K_h(\delta^TZ^{(1)})\right]$$
$$\omega = E\Bigg[\Bigg\{-\frac{1}{2}g''\left(\frac{\tilde{\delta}^TZ^{(1)}}{g'(0)}\right)\left(\beta_{(1)}-\frac{\delta}{g'(0)}\right)^T \times Z^{(1)}Z^{(1)T}\left(\beta_{(1)}-\frac{\delta}{g'(0)}\right)\Bigg\}Z^{(1)}K_h(\delta^TZ^{(1)})\Bigg],$$

where $\tau = (\tau_1,\ldots,\tau_{s_p})^T$ and $\omega = (\omega_1,\ldots,\omega_{s_p})^T$, then we can have

$$E\left\{g'\left(\frac{\delta^TZ^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\delta^TZ^{(1)})\right\}\left(\beta_{(1)}-\frac{\delta}{g'(0)}\right) = \tau+\omega+\epsilon_1+\xi_1+\lambda_{1n}\text{sign}(\delta).$$

Based on the proof of Theorem 3.1, we know that $\tau_j = O(h^2)$ for $j = 1,\ldots,s_p$, so $\|\tau\|_\infty = O(h^2)$. For ω, since $g''(\delta^TZ^{(1)}/g'(0)) = O(1)$ for all $\delta \in \mathcal{N}$, based on the proof of Theorem 3.1, we can have

$$\lambda_{\max}\left[E\left|\left\{-\frac{1}{2}g''\left(\frac{\tilde{\delta}^TZ^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}\right\}Z_j^{(1)}K_h(\delta^TZ^{(1)})\right|\right] = O\left[\lambda_{\max}\left\{E\left|Z^{(1)}Z^{(1)T}Z_j^{(1)}\right|\right\}\right] = O\left[\lambda_{\max}\left\{E\left|ZZ^TZ_j\right|\right\}\right].$$

Then, by (C5), $\|\omega\|_\infty = O(\|\delta - g'(0)\beta_{(1)}\|^2) = O(s_pn^{-2\gamma})$. Let

$$\Psi(\delta) = \beta_{(1)} - \frac{\delta}{g'(0)} - E\left\{g'\left(\frac{\delta^TZ^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\delta^TZ^{(1)})\right\}^{-1} \times (\tau+\omega+\epsilon_1+\xi_1+\lambda_{1n}\text{sign}(\delta)).$$

Then, if δ solves $\Psi(\delta) = 0$, it also solves (A6). It follows from (C2), (C6) and the choice of λ_{1n} that

$$\left\|E\left\{g'\left(\frac{\delta^TZ^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\delta^TZ^{(1)})\right\}^{-1} \times (\tau+\omega+\epsilon_1+\xi_1+\lambda_{1n}\text{sign}(\delta))\right\|_\infty \leq \left\|E\left\{g'\left(\frac{\delta^TZ^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\delta^TZ^{(1)})\right\}^{-1}\right\| \times (\|\tau\|_\infty+\|\omega\|_\infty+\|\epsilon_1\|_\infty+\|\xi_1\|_\infty+\lambda_{1n}) = o(n^{-\gamma}).$$

Then, for sufficiently large n, if $\beta_{(1)} - \delta/g'(0) = n^{-\gamma}$, $\Psi(\delta) > 0$; if $\beta_{(1)} - \delta/g'(0) = -n^{-\gamma}$, $\Psi(\delta) < 0$. By continuity of $\Psi(\delta)$, an application of Miranda's existence theorem shows that $\Psi(\delta) = 0$ has a solution in $\mathcal{N}$, which is also the solution to (A6).

Step 2: Let $\hat{b} = (\hat{b}_{(1)}, 0)^T$, where $\hat{b}_{(1)}$ is the solution to (A6) as shown above, then $\hat{b}$ will be the solution to

(A7).

$$\frac{1}{n}\sum_{i=1}^{n}\frac{y_i-(A_i-1/2)\hat{b}_{(1)}^T z_i^{(1)}}{\pi_i^A(1-\pi)^{(1-A_i)}}(1-2A_i)z_i^{(0)}K_h(\hat{b}_{(1)}^T z_i^{(1)})$$
$$=E\left[\frac{Y-(A-1/2)\hat{b}_{(1)}^T Z^{(1)}}{\pi^A(1-\pi)^{(1-A)}}(1-2A)Z^{(0)}K_h(\hat{b}_{(1)}^T Z^{(1)})\right]$$
$$-\epsilon_0-\xi_0.$$

By Taylor expansion,

$$E\left[\frac{Y-(A-1/2)\hat{b}_{(1)}^T Z^{(1)}}{\pi^A(1-\pi)^{(1-A)}}(1-2A)Z^{(0)}K_h(\hat{b}_{(1)}^T Z^{(1)})\right]$$
$$=E\left[\left\{-g(\beta_{(1)}^T Z^{(1)})+\hat{b}_{(1)}^T Z^{(1)}\right\}Z^{(0)}K_h(\hat{b}_{(1)}^T Z^{(1)})\right]$$
$$=-E\left\{g'\left(\hat{b}_{(1)}^T Z^{(1)}\right)Z^{(0)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}$$
$$\times\left(\beta_{(1)}-\frac{\hat{b}_{(1)}}{g'(0)}\right)+\varsigma+\varpi,$$

where

$$\varsigma=E\left[\left\{-g\left(\frac{\hat{b}_{(1)}^T Z^{(1)}}{g'(0)}\right)+\hat{b}_{(1)}^T Z^{(1)}\right\}Z^{(0)}K_h(\hat{b}_{(1)}^T Z^{(1)})\right]$$
$$\varpi=E\left[\left\{-\frac{1}{2}g''\left(\frac{\tilde{\delta}^T Z^{(1)}}{g'(0)}\right)\left(\beta_{(1)}-\frac{\hat{b}_{(1)}}{g'(0)}\right)^T\right.\right.$$
$$\left.\left.\times Z^{(1)}Z^{(1)T}\left(\beta_{(1)}-\frac{\hat{b}_{(1)}}{g'(0)}\right)\right\}Z^{(0)}K_h(\hat{b}_{(1)}^T Z^{(1)})\right]$$

and $\tilde{\delta}$ lies on the line segment connecting $\hat{b}_{(1)}$ and $g'(0)\beta_{(1)}$, $\varsigma=(\varsigma_1,\ldots,\varsigma_{s_p})^T$, $\varpi=(\varpi_1,\ldots,\varpi_{s_p})^T$. Based on the proof above, it is not difficult to prove that $\|\varsigma\|_\infty=O(h^2)$ and $\|\varpi\|_\infty=O(s_p n^{-2\gamma})$. Since $\hat{b}_{(1)}$ is the solution to $\Psi(\delta)=0$, it holds that

$$\beta_{(1)}-\frac{\hat{b}_{(1)}}{g'(0)}=E\left\{g'\left(\frac{\hat{b}_{(1)}^T Z^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}^{-1}$$
$$\times(\tau+\omega+\epsilon_1+\xi_1+\lambda_{1n}\mathrm{sign}(\hat{b}_{(1)}))$$

Then, we have

$$\frac{1}{n\lambda_{1n}}\sum_{i=1}^{n}\frac{y_i-(A_i-1/2)\hat{b}_{(1)}^T z_i^{(1)}}{\pi_i^A(1-\pi)^{(1-A_i)}}(1-2A_i)z_i^{(0)}K_h(\hat{b}_{(1)}^T z_i^{(1)})$$
$$=-\frac{1}{\lambda_{1n}}E\left\{g'\left(\hat{b}_{(1)}^T Z^{(1)}\right)Z^{(0)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}$$
$$E\left\{g'\left(\frac{\hat{b}_{(1)}^T Z^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}^{-1}$$
$$*(\tau+\omega+\epsilon_1+\xi_1+\lambda_{1n}\mathrm{sign}(\hat{b}_{(1)}))$$
$$+\frac{1}{\lambda_{1n}}(\varsigma+\varpi-\epsilon_0-\xi_0).$$

In the event $E_1\cap E_2\cap E_3\cap E_4$, by the choice of λ_{1n},

$$\|\lambda_{1n}^{-1}\epsilon_1\|_\infty=o(1),\quad \|\lambda_{1n}^{-1}\xi_1\|_\infty=o(1),$$
$$\|\lambda_{1n}^{-1}\varsigma_1\|_\infty=o(1),\quad \|\lambda_{1n}^{-1}\varpi_1\|_\infty=o(1).$$

By (C7),

$$\frac{1}{\lambda_{1n}}\left\|E\left\{g'\left(\hat{b}_{(1)}^T Z^{(1)}\right)Z^{(0)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}\right.$$
$$\times E\left\{g'\left(\frac{\hat{b}_{(1)}^T Z^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}^{-1}$$
$$\left.*(\tau+\omega+\epsilon_1+\xi_1)\right\|_\infty$$
$$<\frac{1}{\lambda_{1n}}\|\tau+\omega+\epsilon_1+\xi_1\|_\infty$$
$$=o(1).$$

Finally, by (C7),

$$\frac{1}{\lambda_{1n}}\left\|E\left\{g'\left(\hat{b}_{(1)}^T Z^{(1)}\right)Z^{(0)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}\right.$$
$$E\left\{g'\left(\frac{\hat{b}_{(1)}^T Z^{(1)}}{g'(0)}\right)Z^{(1)}Z^{(1)T}K_h(\hat{b}_{(1)}^T Z^{(1)})\right\}^{-1}$$
$$\left.*\lambda_{1n}\mathrm{sign}(\hat{b}_{(1)})\right\|_\infty<1.$$

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