Proper Inference for Value Function in High-Dimensional Q-Learning for Dynamic Treatment Regimes

Wensheng Zhu[a], Donglin Zeng[b], and Rui Song[c]

[a]Key Laboratory for Applied Statistics of MOE, School of Mathematics and Statistics, Northeast Normal University, Changchun, China; [b]Departments of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC; [c]Department of Statistics, North Carolina State University, Raleigh, NC

ABSTRACT

Dynamic treatment regimes are a set of decision rules and each treatment decision is tailored over time according to patients' responses to previous treatments as well as covariate history. There is a growing interest in development of correct statistical inference for optimal dynamic treatment regimes to handle the challenges of nonregularity problems in the presence of nonrespondents who have zero-treatment effects, especially when the dimension of the tailoring variables is high. In this article, we propose a high-dimensional Q-learning (HQ-learning) to facilitate the inference of optimal values and parameters. The proposed method allows us to simultaneously estimate the optimal dynamic treatment regimes and select the important variables that truly contribute to the individual reward. At the same time, hard thresholding is introduced in the method to eliminate the effects of the nonrespondents. The asymptotic properties for the parameter estimators as well as the estimated optimal value function are then established by adjusting the bias due to thresholding. Both simulation studies and real data analysis demonstrate satisfactory performance for obtaining the proper inference for the value function for the optimal dynamic treatment regimes. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received April 2017
Accepted June 2018

KEYWORDS
Hard threshold; Q-learning; Value function inference; Variable selection

1. Introduction

Heterogeneous responses to treatment, both across patients and over time, are commonly observed for many diseases including cancer, cardiovascular disease, and diabetes. Consequently, treatment decisions need to be tailored not only to patients' characteristics, but also over time according to patients' responses to previous treatments. Such adaptive treatment strategies, also referred as dynamic treatment regimes (DTRs), attract more and more attentions in concurrent clinical and statistical researches (Chakraborty and Moodie 2013). In particular, one of the most important goals is to infer the optimal dynamic treatment regimes and evaluate the resulting value function, that is, the expected reward outcome when this optimal rule is applied, using empirical data evidence.

One of the most commonly used and effective approaches to estimate the optimal DTRs in a multi-stage study is called Q-learning, which is implemented through a backward recursive fitting procedure (Watkins 1989). Specifically, a regression model is fitted to estimate the conditional expectation of the so-called Q-function at each stage, then the optimal treatment at the current stage is estimated as the one maximizing this conditional expectation. The Q-function for each state of the Q-learning procedure is defined as the reward outcome increment in future stages when the optimal treatments are implemented, given the current treatment and covariates. In practice, the regression model is often based on a parametric linear model with treatment by covariate interactions included.

There are some key challenges with existing Q-learning methods. First, with recent advances in technology, many prognostic factors and intermediate outcomes for each individual are collected but not all of them are useful for the treatment decision-making process. The ability to identify and integrate information that is relevant becomes critical when performing regression in each stage of Q-learning. Qian and Murphy (2011) proposed to estimate the conditional mean in Q-learning using an L_1-penalized regression and studied the error bound of the value function for the estimated treatment regime. However, the associated variable selection properties, such as selection consistency, convergence rate, and weak convergence are not studied. Second, it is well known that obtaining proper inference for the parameters indexing the optimal rules and the optimal value function is difficult. The challenges in the above inference problems arise when the optimal treatments in the latter stage are not unique for at least some subjects in the population, that is, nonrespondents, that fails traditional inferential approaches (Robins 2004). To remedy the inference, various approaches have been suggested, including hard threshold method (Moodie and Richardson (2010), also called Zeroing Instead of Plugging In), soft threshold method (Chakraborty, Murphy, and Strecher 2010), and resampling method (Chakraborty, Laber, and Zhao 2013; Laber et al. 2014). However, the resampling method tends to yield very conservative confidence intervals. More recently, Luedtke and Van Der Laan (2016) proposed an online one-step estimator to obtain the inference for the optimal value function

CONTACT Rui Song rsong@ncsu.edu Department of Statistics, North Carolina State University, Raleigh, NC 27695.

with split sampling. Although asymptotically valid, the resulting confidence interval for the value function can be wide due to using partial sample for inference. Furthermore, it remains unstudied and unclear how all these methods perform in high-dimensional Q-learning setting.

In this article, we propose a high-dimensional Q-learning (HQ-learning) to facilitate the inference of optimal values and parameters. The proposed method makes use of both sparsity regularization and hard-thresholding in Q-learning. Specifically, we adopt a folded-concave penalty (Fan and Lv 2011) to estimate parameters at all stages to enforce sparsity. Moreover, we remove from the regression the treatment effects for the subjects whose treatment responses are smaller than a given threshold. Through regularization, HQ-learning allows us to simultaneously estimate the optimal dynamic treatment regimes and select the important variables that truly contribute to the individual reward at each stage. With hard thresholding, we decrease the degree of nonregularity due to the nonuniqueness of the optimal treatments for those nonrespondents. The latter pays a price of introducing bias in the parameter and value estimation, however, after estimating this bias, we are able to remove this bias so still obtain proper inference for the value function estimation.

The rest of the article is organized as follows. In Section 2, we introduce the high-dimensional Q-learning and illustrate with a two-stage DTR. Asymptotic distribution of the proposed estimators is provided in Section 3 followed by the optimal value estimation result and implementation issues in Section 4. Simulation results and the application to STAR*D study are in Sections 5 and 6. We wrap up the discussions in Section 7.

2. High-Dimensional Q-learning for Dynamic Treatment Regimes

We first introduce high-dimensional Q-learning regression models for estimating optimal dynamic treatment regimes in multiple stage randomized studies. For simplicity, we only focus on a two-stage randomized trial throughout this method but generalization to more than two stages will be similar.

For tth stage ($t = 1, 2$), we let O_t denote all covariates measured prior to the tth stage and $A_t \in \{-1, 1\}$ be a dichotomized treatment assignment at the tth stage. Moreover, R_t denotes the clinical reward right after the treatment at the tth stage, for which a larger value of R_t is more desirable. Then in a two-stage trial, $(O_1, A_1, R_1, O_2, A_2, R_2)$ consists of all observed data for a single patient and the observed data consist of n trajectories of the form (O_{ti}, A_{ti}, R_{ti}) for patients $i = 1, \ldots, n$ and $t = 1, 2$. A dynamic treatment regime, denoted by $d = (d_1, d_2)$, is a sequence of treatment rules at each stage, that is, d_t is a deterministic function mapping each value of O_t to the domain of A_t, that is, $\{-1, 1\}$. Corresponding to each d, we define its value as $V(d) = E^d(R_1 + R_2)$, which is the average total reward if treatment assignments follow these rules at the two stages, that is, $A_t = d_t(O_t)$ for $t = 1, 2$. Thus, an optimal dynamic treatment regime, denoted by $d_0 = (d_{10}, d_{20})$, is a rule that maximizes $V(d)$.

One commonly used method to estimate the optimal dynamic treatment regime is Q-learning, which is a sequence of regression models to be estimated backward for each stage recursively. Specifically, in the Q-learning method (see Chakraborty, Murphy, and Strecher 2010; Song et al. 2015), we model the so-called Q-function at each stage as

$$Q_t(H_t, A_t; \boldsymbol{\theta}_t) = H_t^T \boldsymbol{\beta}_t + A_t(H_t^T \boldsymbol{\psi}_t), \tag{1}$$

where $H_t = (1, O_t)$ takes values in $\mathbb{R}^{p_t}$. Among the parameters of the above Q-function $\boldsymbol{\theta}_t = (\boldsymbol{\beta}_t^T, \boldsymbol{\psi}_t^T)^T$, $\boldsymbol{\beta}_t = (\beta_{t1}, \ldots, \beta_{tp_t})^T$ reflects the main effect of the current state on the outcome, while $\boldsymbol{\psi}_t = (\psi_{t1}, \ldots, \psi_{tp_t})^T$ reflects the interaction between the current state and the treatment choice. It is worth mentioning that a more general model can be $Q_t(S_t, H_t, A_t; \boldsymbol{\theta}_t) = S_t^T \boldsymbol{\beta}_t + A_t(H_t^T \boldsymbol{\psi}_t)$, for $t = 1, 2$, in which S_t and H_t can be different subsets of $(1, O_t)$. When H_t's dimension is not high, a typical Q-learning method is to perform a sequence least-square regression backward from the last stage. At the tth stage's regression, the mean structure in the regression is the same as (1) but the outcome variable is R_t if $t = 2$ and the summation of R_t and predicted $Q_{t+1}(H_{t+1}, A_{t+1}^*; \widehat{\boldsymbol{\theta}}_{t+1})$ if $t = 1$, where $\widehat{\boldsymbol{\theta}}_{t+1}$ is the estimated coefficient for $\boldsymbol{\theta}_{t+1}$ and $A_{t+1}^* = \text{sign}(H_{t+1}^T \widehat{\boldsymbol{\psi}}_{t+1})$, that is, the estimated optimal treatment that maximizes the Q-function at the $(t + 1)$-stage.

In most of personalized medicine, H_t includes individual genomics, baseline biomarkers, and intermediate outcomes, so its dimension can be very high. Therefore, the standard Q-learning procedure may not be feasible due to high-dimensional regression models. Furthermore, since most of H_t are not predictive of treatment effects, identifying those predictive covariates and removing other noisy covariates from the Q-learning regression will improve the estimation of the optimal dynamic treatment regime. Therefore, we propose the following Q-learning procedure with high-dimensional predictors by incorporating proper sparse penalization at each stage of Q-learning. We allow that p_t is much larger than the number of observations n, and further assume that majority of the true parameters $\boldsymbol{\theta}_{t0} = (\boldsymbol{\beta}_{t0}^T, \boldsymbol{\psi}_{t0}^T)^T$ are exactly zero for $t = 1, 2$. The proposed Q-learning procedure with high-dimensional predictors consists of the following three steps when $p_t \gg n$ for $t = 1, 2$.

Step 1. We estimate the second-stage parameters by minimizing the penalized objective function:

$$\begin{aligned}\Phi_2(\boldsymbol{\theta}_2) &= 2^{-1} \sum_{i=1}^{n} (Y_{2i} - Q_2(H_{2i}, A_{2i}; \boldsymbol{\theta}_2))^2 \\ &\quad + n \sum_{j=1}^{p_2} \left\{ p_{\lambda_{2n}}(|\beta_{2j}|) + p_{\lambda_{2n}}(|\psi_{2k}|) \right\} \\ &= 2^{-1} \|\mathbf{y}_2 - \mathbf{Z}_2 \boldsymbol{\theta}_2\|_2^2 \\ &\quad + n \sum_{j=1}^{p_2} \left\{ p_{\lambda_{2n}}(|\beta_{2j}|) + p_{\lambda_{2n}}(|\psi_{2k}|) \right\}. \end{aligned} \tag{2}$$

In the above least-square estimation, $\mathbf{Z}_2 = (Z_{21}^T, \ldots, Z_{2n}^T)^T$ is the $n \times 2p_2$ design matrix and can be also denoted by $\mathbf{Z}_2 = (\mathbf{H}_2, \mathbf{H}_2 A_2)$ with $\mathbf{H}_2 = (H_{21}^T, \ldots, H_{2n}^T)^T$ being an $n \times p_2$ matrix, and $\mathbf{y}_2 = (Y_{21}, \ldots, Y_{2n})^T$ is the optimal potential reward in the second stage. Additionally, we include two penalization terms to achieve sparsity in the Q-function estimation. Particularly, $p_{\lambda_{2n}}(\cdot)$ belongs to the class of

folded-concave penalty functions (Lv and Fan 2009), and $\lambda_{2n} \geqslant 0$ is regularization parameters indexed by sample size n. This class of penalties tends to give estimators with three desired properties advocated by Fan and Li (2001): unbiasedness, sparsity, and continuity. The SCAD (Fan and Li 2001) and MCP (Zhang 2010) with $a \geqslant 1$ belong to this class of penalties, whose derivatives of penalties are, respectively, given by

$$p'_{\lambda_{2n}}(\beta) = \lambda_{2n}\left\{I(\beta \leqslant \lambda_{2n}) + \frac{(a\lambda_{2n} - \beta)_+}{(a-1)\lambda_{2n}} I(\beta > \lambda_{2n})\right\}, \quad t \geqslant 0 \text{ for some } a \geqslant 2,$$

where often $a = 3.7$ is used, and $p'_{\lambda_{2n}}(\beta) = (a\lambda_{2n} - \beta)_+/a$ for $t \geqslant 0$. We use this class of penalties to establish the oracle property of the estimators and to make inference further in the first stage of Q-learning. We denote $\hat{\boldsymbol{\theta}}_2 = (\hat{\boldsymbol{\beta}}_2^T, \hat{\boldsymbol{\psi}}_2^T)^T$ as the minimizer of the penalized objective function (2), and $\hat{\boldsymbol{\beta}}_2^{(1)}, \hat{\boldsymbol{\psi}}_2^{(1)}$ are subvectors of $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\psi}}_2$ formed by all the nonzero components, respectively.

Step 2. We obtain the first-stage individual pseudo-outcomes $\hat{\mathbf{y}}_1$ with hard-threshold as

$$\hat{\mathbf{y}}_1 = \mathbf{R}_1 + \mathbf{H}_{2,M}\hat{\boldsymbol{\beta}}_2^{(1)} + \left|\mathbf{H}_{2,I}\hat{\boldsymbol{\psi}}_2^{(1)}\right| \cdot 1_{|\mathbf{H}_{2,I}\hat{\boldsymbol{\psi}}_2^{(1)}| > \varrho_{2n}}, \quad (3)$$

where $\mathbf{H}_{2,M}$ and $\mathbf{H}_{2,I}$ denote the submatrices of $\mathbf{H}_2$ formed by columns corresponding to the indexes of the nonzero components of $\hat{\boldsymbol{\beta}}_2$ and $\hat{\boldsymbol{\psi}}_2$, respectively, and $|\mathbf{H}_{2,I}\hat{\boldsymbol{\psi}}_2^{(1)}| \cdot 1_{|\mathbf{H}_{2,I}\hat{\boldsymbol{\psi}}_2^{(1)}| > \varrho_{2n}}$ is an n-dimensional vector with the ith entry $|H_{2i,I}^T\hat{\boldsymbol{\psi}}_2^{(1)}| \cdot 1_{|H_{2i,I}^T\hat{\boldsymbol{\psi}}_2^{(1)}| > \varrho_{2n}}$. Here, ϱ_{2n} is a small constant depending on n. The purpose of using the truncated pseudo-outcomes is to overcome the difficulty of inference for the Q-learning due to the nature of the nonregularity when $P(H_{2i,I}^T\boldsymbol{\psi}_{20}^{(1)} = 0) > 0$ (Chakraborty, Murphy, and Strecher 2010; Moodie and Richardson 2010; Laber et al. 2014; Song et al. 2015). In contrast, the pseudo-outcomes without truncation in the standard Q-learning involves a nonsmooth function of $\hat{\boldsymbol{\psi}}_2^{(1)}$, which will cause that the estimators of the parameters in Step 3 is also a nonsmooth function of $\hat{\boldsymbol{\psi}}_2^{(1)}$. As a consequence, the asymptotic distribution of estimated parameters is not always normal distribution.

Step 3. We estimate the first-stage parameters by minimizing the following penalized objective function:

$$\begin{aligned}\Phi_1(\boldsymbol{\theta}_1) &= 2^{-1}\sum_{i=1}^{n}(\hat{Y}_{1i} - Q_1(Z_{1i};\boldsymbol{\theta}_1))^2 \\ &\quad + n\sum_{j=1}^{p_1}\left\{p_{\lambda_{1n}}(|\beta_{1j}|) + p_{\lambda_{1n}}(|\psi_{1k}|)\right\} \\ &= 2^{-1}\|\hat{\mathbf{y}}_1 - \mathbf{Z}_1\boldsymbol{\theta}_1\|_2^2 \\ &\quad + n\sum_{j=1}^{p_1}\left\{p_{\lambda_{1n}}(|\beta_{1j}|) + p_{\lambda_{1n}}(|\psi_{1k}|)\right\}, \quad (4)\end{aligned}$$

where $\mathbf{Z}_1 = (Z_{11}^T, \ldots, Z_{1n}^T)^T$ is the $n \times 2p_1$ design matrix and can be also denoted by $\mathbf{Z}_1 = (\mathbf{H}_1, \mathbf{H}_1A_1)$ with $\mathbf{H}_1 = (H_{11}^T, \ldots, H_{1n}^T)^T$ is an $n \times p_1$ matrix, and $\hat{\mathbf{y}}_1 = (\hat{Y}_{11}, \ldots, \hat{Y}_{1n})^T$ is the pseudo-outcomes from Step 2. For simplicity, here we use the same penalty functions $p_{\lambda_{1n}}(\cdot)$ given in (2), but we adopt the different regularization parameters $\lambda_{1n} \geqslant 0$. Similar to the second stage, we also denote $\hat{\boldsymbol{\theta}}_1 = (\hat{\boldsymbol{\beta}}_1^T, \hat{\boldsymbol{\psi}}_1^T)^T$ as the minimizer of the penalized objective function (4), and $\hat{\boldsymbol{\beta}}_1^{(1)}, \hat{\boldsymbol{\psi}}_1^{(1)}$ are subvectors of $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\psi}}_1$ formed by all the nonzero components, respectively.

There are two key features of our proposed Q-learning with high-dimensional predictors which make the proposed method different from the standard Q-learning method. First, we include sparsity penalization in both stages to encourage that the final model leads to a sparse Q-function with only important predictors in the function. The penalization also makes high-dimensional regression feasible. Second, we introduce the truncated pseudo-outcome in the first-stage regression. By truncation, we set the treatment effects to zeros only for those subjects whose treatment effects are very small ($< \varrho_{2n}$) so the induced bias is negligible. Furthermore, as will be shown in the next section, this simple truncation will tackle the challenge of inference due to nonregularity in the presence of positive zero-treatment effect probabilities, which is well known in the Q-learning inference literature (Chakraborty, Murphy, and Strecher 2010; Moodie and Richardson 2010; Laber et al. 2014; Song et al. 2015).

We call our approach high-dimensional Q-learning (HQ-learning). The HQ-learning allows us to simultaneously estimate the optimal dynamic treatment regimes and select the important variables that truly contribute to the individual reward at each stage.

3. Theoretical Properties

In this section, we provide theoretical justification of HQ-learning. In particular, we show that under some regularity conditions for the sparsity structure in the true Q-functions and the choice of the penalties, the proposed method possesses the oracle property of selecting important predictors with probability 1. Furthermore, we obtain the asymptotic distribution of the estimated dynamic treatment rules, which will be the key to establish a valid inference of the optimal value estimation to be given in the next section. In our results, we allow that the dimensionality of the covariates is ultra-high in terms of the exponential polynomial rate of the sample size. The latter is commonly seen if individual omics data are used to determine dynamic treatment regimes.

We first study the oracle properties of $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ and show that they converge to the true parameters $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{20}$ in probability. For the purpose of illustration, for $t = 1, 2$ we denote $\boldsymbol{\theta}_{t0} = (\boldsymbol{\theta}_{t0}^{(1)T}, \boldsymbol{\theta}_{t0}^{(2)T})^T$, $\boldsymbol{\theta}_{t0}^{(1)} = (\boldsymbol{\beta}_{t0}^{(1)T}, \boldsymbol{\psi}_{t0}^{(1)T})^T$, and $\boldsymbol{\theta}_{t0}^{(2)} = (\boldsymbol{\beta}_{t0}^{(2)T}, \boldsymbol{\psi}_{t0}^{(2)T})^T$, where each component of $\boldsymbol{\beta}_{t0}^{(1)}$ and $\boldsymbol{\psi}_{t0}^{(1)}$ is nonzero but $\boldsymbol{\beta}_{t0}^{(2)} = \mathbf{0}$ and $\boldsymbol{\psi}_{t0}^{(2)} = \mathbf{0}$, and we also denote

$$\begin{aligned}\text{supp}(\boldsymbol{\theta}_{t0}) &= \left\{j \in \{1, \ldots, 2p_t\} : \theta_{t0,j} \neq 0\right\}, \\ \text{supp}(\boldsymbol{\beta}_{t0}) &= \left\{j \in \{1, \ldots, p_t\} : \beta_{t0,j} \neq 0\right\},\end{aligned}$$

and

$$\operatorname{supp}(\boldsymbol{\psi}_{t0}) = \left\{ j \in \{1, \ldots, p_t\} : \psi_{t0,j} \neq 0 \right\}.$$

We refer to the numbers of components of $\boldsymbol{\theta}_{10}^{(1)}$ and $\boldsymbol{\theta}_{20}^{(1)}$ as s_n and r_n, respectively. We also refer to $d_{tn} = 2^{-1} \min\{|\theta_{t0,j}| : \theta_{t0,j} \neq 0\}$ as half of the minimum signal for $t = 1, 2$. In addition, we let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ represent the smallest and largest eigenvalues of a symmetric matrix respectively. Let $\mathbf{z}_{tj}$ denote the jth column of $\mathbf{Z}_t$ ($t = 1, 2$, and $j = 1, \ldots, 2p_t$). Without loss of generality, we assume that $\mathbf{z}_{tj}$ has been standardized so that $E\|\mathbf{z}_{tj}\|_2 = \sqrt{n}$. Hereafter, for any random vectors U and V with $E(U) = E(V) = 0$, we denote Σ_{UV^T} as the covariance matrix of U and V. We also denote $\mathbb{P}_n f(x) = 1/n \sum_{i=1}^n f(x_i)$ as the empirical measure.

Since the estimation of $\boldsymbol{\theta}_2$ is based on a penalized least-square estimation with a nonconcave penalty, both weak oracle property and oracle property of $\hat{\boldsymbol{\theta}}_2$ can be established under the following conditions.

(A1) For the covariance matrix $\Sigma_{Z_2 Z_2}$, $\lambda_{\min}(\Sigma_{Z_{2,(1)} Z_{2,(1)}^T}) \geqslant C_1$, $\lambda_{\max}(\Sigma_{Z_{2,(1)} Z_{2,(1)}^T}) \leqslant C_2$, $\operatorname{trace}(\Sigma_{Z_{2,(1)} Z_{2,(1)}^T}) = O(r_n)$, and $\|\Sigma_{Z_{2,(2)} Z_{2,(1)}^T}\|_\infty = O(r_n)$.

(A2) The nonsparsity size $r_n = o(n)$ and the dimensionality satisfies $\log p_2 = O(n^{\varsigma_2})$ for some $\varsigma_2 \in (0, 1/2)$. Moreover,

$$\begin{aligned} &\lambda_{2n} = o(d_{2n}), \\ &\max\left\{\sqrt{r_n/n}, n^{(\varsigma_2 - 1)/2}\sqrt{\log n}\right\} = o(\lambda_{2n}), \\ &p'_{\lambda_{2n}}(d_{2n}) = O\left(n^{-1/2}\right), \end{aligned}$$

and $\lambda_{2n}\kappa_{20} = o(1)$, where $\kappa_{20} = \max_{\boldsymbol{\delta} \in \mathcal{N}_{20}} \kappa(\rho; \boldsymbol{\delta})$. Here, $\mathcal{N}_{20} = \{\boldsymbol{\delta} \in \mathbb{R}^{r_n} : \|\boldsymbol{\delta} - \boldsymbol{\theta}_{20}\|_\infty \leqslant d_{2n}\}$, and the function $\kappa(\rho; \mathbf{v})$ is defined as

$$\kappa(\rho; \mathbf{v}) = \lim_{\epsilon \to 0+} \max_{1 \leqslant j \leqslant q} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} -\frac{\rho'(t_2) - \rho'(t_1)}{t_2 - t_1},$$

with $\|\mathbf{v}\|_0 = q$, and $\rho(\cdot) = \lambda^{-1} p_\lambda(\cdot)$ for the regularization parameter $\lambda \geqslant 0$. We assume that $p_\lambda(\cdot)$ satisfies Condition 1 in Lv and Fan (2009).

(A3) Assume $\tilde{\epsilon}_i = Y_{2i} - H_{2i}^T \boldsymbol{\beta}_{20} - A_{2i}(H_{2i}^T \boldsymbol{\psi}_{20})$. There exist some constants M and υ_0 such that

$$\max_{i=1,\ldots,n} E\left\{\exp\left(\frac{|\tilde{\epsilon}_i|}{M}\right) - 1 - \frac{|\tilde{\epsilon}_i|}{M}\right\} M^2 \leqslant \frac{\upsilon_0}{2}. \tag{5}$$

Moreover, $\max_{i=1}^n E|\tilde{\epsilon}_i|^3 = O(1)$.

(A4) As n goes to infinity, $\sum_{i=1}^n (\frac{1}{n} E(Z_{2i,(1)}^T) \Sigma_{Z_{2,(1)} Z_{2,(1)}^T}^{-1} E(Z_{2i,(1)}))^{3/2} \to 0$.

Remark 1. For more remarks about Conditions A1 and A2 please refer to Fan and Lv (2011). Condition A3 is used to bound the deviation of the p-dimensional random vector $\mathbf{Z}_2^T \mathbf{Y}_2$ from its mean $\mathbf{Z}_2^T \boldsymbol{\theta}_{20}$, which also holds if $\tilde{\epsilon}_i$ is sub-Gaussian. Condition A4 is related to the Lyapunov condition.

The following theorem gives the asymptotic property of $\hat{\boldsymbol{\theta}}_2$, whose proof is similar to Fan and Lv (2011) so is skipped.

Theorem 1. Under Conditions A1–A3, there exists a minimizer $\hat{\boldsymbol{\theta}}_2 = (\hat{\boldsymbol{\theta}}_2^{(1)T}, \hat{\boldsymbol{\theta}}_2^{(2)T})^T$ of $\Phi_2(\boldsymbol{\theta}_2)$, where $\hat{\boldsymbol{\theta}}_2^{(1)}$ is the subvector of $\hat{\boldsymbol{\theta}}_2$ indexed by $\operatorname{supp}(\boldsymbol{\theta}_{20})$, such that

(i) $\hat{\boldsymbol{\theta}}_2^{(2)} = \mathbf{0}$ with probability tending to 1 as $n \to \infty$;

(ii) $\|\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_{20}\|_2 = O_P(\sqrt{r_n/n})$;

(iii) If Condition A4 also holds and $p'_{\lambda_{2n}}(d_{2n}) = o(r_n^{-1/2} n^{-1/2})$, $r_n = o(n^{1/3})$, then as $n \to \infty$ we have,

$$\sqrt{n} \mathbf{D}_n \Sigma_{Z_{2,(1)} Z_{2,(1)}^T}^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_2^{(1)} - \boldsymbol{\theta}_{20}^{(1)}\right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{J}_D),$$

where $\mathbf{D}_n$ is a $g \times r_n$ matrix such that $\mathbf{D}_n \mathbf{D}_n^T \to \mathbf{J}_D$, $\mathbf{J}_D$ is a $g \times g$ symmetric positive definite matrix, and $\Sigma_{Z_{2,(1)} Z_{2,(1)}^T}$ is the positive definite covariance matrix of $Z_{2,(1)}$.

Next, we study asymptotic properties of $\hat{\boldsymbol{\theta}}_1$. Assume that the true model for the Q-function at the first stage is

$$\mathbf{y}_1 = \mathbf{Z}_1 \boldsymbol{\theta}_1 + \boldsymbol{\epsilon}, \tag{6}$$

where $\mathbf{y}_1 = (Y_{11}, \ldots, Y_{1n})^T = \mathbf{R}_1 + \mathbf{H}_{2,M} \boldsymbol{\beta}_{20}^{(1)} + |\mathbf{H}_{2,I} \boldsymbol{\psi}_{20}^{(1)}|$ denote the n-dimensional potential rewards for the first stage, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is an n-dimensional vector of noises. For each subject i, we assume that Z_{1i} is independent to ϵ_i. To study the oracle property of $\hat{\boldsymbol{\theta}}_1$, we need the following assumptions.

(B1) We assume that there exist two positive constants C_3 and C_4 such that

$$\lambda_{\min}\left(\Sigma_{Z_{1,(1)} Z_{1,(1)}^T}\right) \geqslant C_3; \tag{7}$$

$$\operatorname{trace}\left(\Sigma_{Z_{1,(1)} Z_{1,(1)}^T}\right) = O(s_n); \tag{8}$$

$$\lambda_{\max}\left(\Sigma_{Z_{2,(1)} Z_{1,(1)}^T}^T \Sigma_{Z_{2,(1)} Z_{1,(1)}^T}\right) \leqslant C_4; \tag{9}$$

$$\left\|\Sigma_{Z_{1,(2)} Z_{2,(1)}^T}\right\|_\infty = O(r_n); \tag{10}$$

$$\left\|\Sigma_{Z_{1,(2)} Z_{1,(1)}^T}\right\|_\infty = O(s_n). \tag{11}$$

(B2) Assume $\varrho_{2n} = n^{-\vartheta}$ for some constant ϑ satisfying that $r_n n^{-1/2} = o(\varrho_{2n})$, and $\varpi_{2n} = O(\frac{r_n}{\sqrt{n}})$.

(B3) The nonsparsity size $s_n = o(n)$ and the dimensionality satisfies $\log p_1 = O(n^{\varsigma_1})$ for some $\varsigma_1 \in (0, 1/2)$. Moreover $\lambda_{1n} = o(d_{1n})$, $\max\{r_n^{3/2}/\sqrt{n}, \sqrt{s_n}/n^{\vartheta}, n^{(\varsigma_1 - 1)/2}\sqrt{\log n}\} = o(\lambda_{1n})$, $p'_{\lambda_{1n}}(d_{1n}) = O(n^{-1/2})$, and $\lambda_{1n}\kappa_{10} = o(1)$, where $\kappa_{10} = \max_{\boldsymbol{\delta} \in \mathcal{N}_{10}} \kappa(\rho; \boldsymbol{\delta})$, $\mathcal{N}_{10} = \{\boldsymbol{\delta} \in \mathbb{R}^{s_n} : \|\boldsymbol{\delta} - \boldsymbol{\theta}_{10}\|_\infty \leqslant d_{1n}\}$ and $\kappa(\rho; \mathbf{v})$ is defined in Condition A2.

(B4) Assume that $\epsilon_i = Y_{1i} - H_{1i}^T \boldsymbol{\beta}_{10} - A_{1i}(H_{1i}^T \boldsymbol{\psi}_{10})$ satisfies Equation (5) in Condition A3 and $\max_{i=1}^n E|\epsilon_i|^3 = O(1)$.

(B5) As n goes to infinity, $\sum_{i=1}^n (\frac{1}{n} E(Z_{1i,(1)}^T) \Sigma_{Z_{1,(1)} Z_{1,(1)}^T}^{-1} E(Z_{1i,(1)}))^{3/2} \to 0$.

Remark 2. Condition B1 gives the usual constraints on the corresponding covariance matrices. Condition B2 enables us to select reasonable thresholds such that we still obtain correct asymptotic properties of $\hat{\boldsymbol{\theta}}_1$ despite that the thresholding introduces us some biases. Condition B3 imposes constraints on the penalty function in terms of the minimum signal d_{1n} in order for the estimators to entail the oracle properties. Conditions B4 and B5 are very similar to Conditions A3 and A4, respectively.

Theorem 2. Under Conditions B1–B4, there exists a minimizer $\hat{\boldsymbol{\theta}}_1 = (\hat{\boldsymbol{\theta}}_1^{(1)T}, \hat{\boldsymbol{\theta}}_1^{(2)T})^T$ of $\Phi_1(\boldsymbol{\theta}_1)$ such that $\hat{\boldsymbol{\theta}}_1^{(2)} = \mathbf{0}$ with probability tending to 1 as $n \to \infty$ and $\|\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{10}\|_2 = O_P(\sqrt{s_n}/n^{\vartheta})$, where $\hat{\boldsymbol{\theta}}_1^{(1)}$ is the subvector of $\hat{\boldsymbol{\theta}}_1$ indexed by $\operatorname{supp}(\boldsymbol{\theta}_{10})$.

We are going to establish the asymptotic distribution of $\hat{\boldsymbol{\theta}}_1$ in next theorem. One difficulty is that the optimal pseudo-outcome Y_1 as well as its estimation $\hat{Y}_1$, obtained by existing methods such as the standard two-stage Q-learning procedure, involve nonsmooth function of $\boldsymbol{\psi}_{20}^{(1)}$ and $\hat{\boldsymbol{\psi}}_2^{(1)}$. Since $\hat{\boldsymbol{\theta}}_1$ is a function of $\hat{Y}_1$, it is also a nonsmooth function of $\hat{\boldsymbol{\psi}}_2^{(1)}$. Consequently, the asymptotic distribution of $\sqrt{n}(\hat{\boldsymbol{\psi}}_1^{(1)} - \boldsymbol{\psi}_{10}^{(1)})$ is not normal any more if $P(H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)} = 0) > 0$, that is, there are nonignorable nonrespondents in Stage 2. However, through the hard-threshold, our estimation automatically removes the effects of these nonrespondents in the inference with a price of inducing bias due to the truncation. Therefore, the asymptotic distribution of the estimators entails a careful control of this bias in the proof. The following theorem gives the asymptotic normality.

Theorem 3. Under the conditions of Theorem 2 and Condition B5, if $p'_{\lambda_{1n}}(d_{1n}) = o(s_n^{-1/2} n^{-1/2})$, $s_n = o(n^{1/3})$, and $r_n = o(n^{1/4})$, for a $g \times s_n$ matrix $\mathbf{A}_n$ satisfying $\mathbf{A}_n \mathbf{A}_n^T \to \mathbf{J}_A$, and $\mathbf{J}_A$ is a $g \times g$ symmetric positive definite matrix, it holds

$$\sqrt{n}\mathbf{A}_n \Sigma_{Z_{1,(1)} Z_{1,(1)}^T}^{\frac{1}{2}} \left(\hat{\boldsymbol{\theta}}_1^{(1)} - \boldsymbol{\theta}_{10}^{(1)} + \Sigma_{Z_{1,(1)} Z_{1,(1)}^T}^{-1} \mathbf{b}\right) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \Sigma_{\text{cov}}),$$

where $\mathbf{b} = E(Z_{1,(1)} | H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| \cdot 1_{|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}| \leqslant \varrho_{2n}} 1_{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \varpi_{2n}})$ is the bias, and

$$\bar{Z}_{2,(1)} = \left(H_{2,M}^T, H_{2,I}^T \operatorname{sgn}\left(H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}\right) \cdot 1_{|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}| > \varrho_{2n}} 1_{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \varpi_{2n}}\right)^T.$$

The covariance matrix

$$\Sigma_{\text{cov}} = \operatorname{cov}\left\{ \lim_n \mathbf{A}_n \Sigma_{Z_{1,(1)} Z_{1,(1)}^T}^{-\frac{1}{2}} \left[F_1(\boldsymbol{\theta}_{10}) + E\left[Z_{1,(1)} \bar{Z}_{2,(1)}^T\right] F_2(\boldsymbol{\theta}_{20})\right]\right\},$$

where $F_1(\boldsymbol{\theta}_{10}) = 2^{-1} \nabla_{\boldsymbol{\theta}_1} Q_1(Z_1; \boldsymbol{\theta}_{10})(Y_1 - Q_1(Z_1; \boldsymbol{\theta}_{10}))$, $F_2(\boldsymbol{\theta}_{20}) = \Sigma_{Z_{2,(1)} Z_{2,(1)}^T}^{-1} 2^{-1} \nabla_{\boldsymbol{\theta}_2} Q_2(Z_2; \boldsymbol{\theta}_{20})(Y_2 - Q_2(Z_2; \boldsymbol{\theta}_{20}))$.

Remark 3. Although $\mathbf{b}$ will go to 0 as n goes to infinity, it is nonnegligible after multiplied by a factor of $\sqrt{n}$. This is because

$$\sqrt{n} E\left(Z_{1,(1)} \left| H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)} \right| \cdot 1_{|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}| \leqslant \varrho_{2n}} 1_{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \varpi_{2n}}\right) = O_P\left(n^{1/2 - \vartheta}\right)$$

and $\vartheta < 1/2$. We refer to $\mathbf{b}$ here as the bias induced by the truncation in the pseudo-outcomes of the first stage.

Theorem 3 shows that the linear combination $\mathbf{A}_n \hat{\boldsymbol{\theta}}_1^{(1)}$ is asymptotically normal for any finite g as the length of $\hat{\boldsymbol{\theta}}_1^{(1)}$ depends on n. This is very essential to construct the confidence interval for the optimal value function of the first stage.

4. Optimal Value Estimation

In the dynamic treatment regimes, one important goal is to evaluate the value function of the obtained treatment regimes, which is characterized as the average reward if the estimated treatment rule would be applied to the same population in future. Simple algebra gives that under the proposed model, the value function associated with the true optimal treatment regimes is $E[Q_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_{10})]$, where

$$Q_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_{10}) = \max_{a_1} Q_1(\mathcal{Z}_1; \boldsymbol{\theta}_{10}) = \mathcal{H}_{1,M}^T \boldsymbol{\beta}_{10}^{(1)} + \left|\mathcal{H}_{1,I}^T \boldsymbol{\psi}_{10}^{(1)}\right|.$$

Therefore, one natural estimator of the value function for the estimated optimal dynamic treatment regime in our approach is

$$\mathbb{P}_n \widehat{Q}_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_1) = \mathbb{P}_n \mathcal{H}_{1,M}^T \hat{\boldsymbol{\beta}}_1^{(1)} + \mathbb{P}_n \left|\mathcal{H}_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}\right| 1_{\{|\mathcal{H}_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}| > \varrho_{1n}\}}. \tag{12}$$

The following theorem gives the asymptotic distribution of this estimator.

Theorem 4. In addition to the conditions of Theorem 3, we further assume that $s_n = o(n^{(\upsilon - 1/4)})$, $\varrho_{1n} = O(\frac{1}{n^{\gamma}})$ for some constant γ satisfying that $s_n n^{-\vartheta} = o(\varrho_{1n})$, and $\varpi_{1n} = O(\frac{s_n}{n^{\vartheta}})$. Then

$$\sqrt{n}\Big(\mathbb{P}_n \widehat{Q}_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_1) - E(Q_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_{10})) + b + E\left(\bar{\mathcal{Z}}_{1,(1)}^T\right) \Sigma_{Z_{1,(1)} Z_{1,(1)}^T}^{-1} \mathbf{b}\Big) \xrightarrow{\mathcal{D}} N\left(0, \sigma_{\mathcal{U}}^2\right),$$

where

$$\bar{\mathcal{Z}}_{1,(1)} = \left(\mathcal{H}_{1,M}^T, \mathcal{H}_{1,I}^T \operatorname{sgn}\left(\mathcal{H}_{1,I}^T \boldsymbol{\psi}_{10}^{(1)}\right) \cdot 1_{\{|\mathcal{H}_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}| > \varrho_{1n}\}} 1_{\{|\mathcal{H}_{1,I}^T \boldsymbol{\psi}_{10}^{(1)}| > \varpi_{1n}\}}\right)^T,$$

$$b = E\left(\left|\mathcal{H}_{1,I}^T \boldsymbol{\psi}_{10}^{(1)}\right| \cdot 1_{\{|\mathcal{H}_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}| \leqslant \varrho_{1n}\}} \cdot 1_{\{|\mathcal{H}_{1,I}^T \boldsymbol{\psi}_{10}^{(1)}| > \varpi_{1n}\}}\right),$$

and

$$\sigma_{\mathcal{U}}^2 = \operatorname{cov}\Big\{Q_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_1) + E\left(\bar{\mathcal{Z}}_{1,(1)}^T\right) \Sigma_{Z_{1,(1)} Z_{1,(1)}^T}^{-1} \times \left[F_1(\boldsymbol{\theta}_{10}) + E\left[Z_{1,(1)} \bar{Z}_{2,(1)}^T\right] F_2(\boldsymbol{\theta}_{20})\right]\Big\}.$$

In addition, $\mathbf{b}$, $F_1(\boldsymbol{\theta}_{10})$, $F_2(\boldsymbol{\theta}_{20})$, and $\bar{Z}_{2,(1)}$ are given in Theorem 3.

Remark 4. As we shown in Remark 3, b is also not ignorable after $\sqrt{n}$-scale in Theorem 4, which is another part of bias induced by the truncation in the estimation of value function (12). Consequently, a better estimator of the value function for the estimated optimal dynamic treatment regime is the estimator of (12) adjusted by the bias given in Theorem 4.

Theorem 4 shows that the average of the optimal value function estimation $\mathbb{P}_n \widehat{Q}_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_1)$ is asymptotically normal under some mild conditions. Following Theorem 4, we can construct asymptotically valid confidence intervals by correcting the bias for the optimal value function of the first stage.

To construct the confidence interval of $E(Q_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_{10}))$, we need first to estimate the variance $\sigma_{\mathcal{U}}^2$ and the bias $\text{Bias} = b + E(\bar{\mathcal{Z}}_{1,(1)}^T) \Sigma_{Z_{1,(1)} Z_{1,(1)}^T}^{-1} \mathbf{b}$. The estimated bias is

$$\widehat{\text{Bias}} = \hat{b} + \left(\mathbb{P}_n[\hat{\bar{\mathcal{Z}}}_{1,(1)}]\right)^T \left(\mathbb{P}_n\left[Z_{1,(1)} Z_{1,(1)}^T\right]\right)^{-1} \widehat{\mathbf{b}},$$

where

$$\hat{b} = \mathbb{P}_n\left(\left|H_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}\right| \cdot 1_{\{|H_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}| \leqslant \varrho_{1n}\}} \cdot 1_{\{|H_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}| > \varpi_{1n}\}}\right),$$

$$\widehat{\mathbf{b}} = \mathbb{P}_n\left(Z_{1,(1)} \left|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}\right| \cdot 1_{|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}| \leqslant \varrho_{2n}} 1_{|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}| > \varpi_{2n}}\right),$$

and

$$\widehat{\tilde{Z}}_{1,(1)} = \mathbb{P}_n\Big(H_{1,M}^T, H_{1,I}^T \mathrm{sgn}\big(H_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}\big) \cdot 1_{\{|H_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}| > \varrho_{1n}\}} 1_{\{|H_{1,I}^T \hat{\boldsymbol{\psi}}_1^{(1)}| > \varpi_{1n}\}}\Big)^T.$$

The estimated variance is

$$\hat{\sigma}_{\mathcal{U}}^2 = \widehat{\mathrm{cov}}\big\{\widehat{Q}_1^*(Z_1; \hat{\boldsymbol{\theta}}_1) + \big(\mathbb{P}_n\big[\widehat{\tilde{Z}}_{1,(1)}\big]\big)^T \big(\mathbb{P}_n\big[Z_{1,(1)} Z_{1,(1)}^T\big]\big)^{-1} \times \big[F_1(\hat{\boldsymbol{\theta}}_1) + \mathbb{P}_n[Z_{1,(1)} \widehat{\tilde{Z}}_{2,(1)}^T] \hat{F}_2(\hat{\boldsymbol{\theta}}_2)\big]\big\},$$

where

$$\hat{F}_2(\hat{\boldsymbol{\theta}}_2) = \big(\mathbb{P}_n\big[Z_{2,(1)} Z_{2,(1)}^T\big]\big)^{-1} \nabla_{\boldsymbol{\theta}_2} Q_2(Z_2; \hat{\boldsymbol{\theta}}_2)(Y_2 - Q_2(Z_2; \hat{\boldsymbol{\theta}}_2)),$$

and

$$\widehat{\tilde{Z}}_{2,(1)} = \mathbb{P}_n\Big(H_{2,M}^T, H_{2,I}^T \mathrm{sgn}\big(H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}\big) \cdot 1_{|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}| > \varrho_{2n}} 1_{|H_{2,I}^T \hat{\boldsymbol{\psi}}_2^{(1)}| > \varpi_{2n}}\Big)^T.$$

Given the estimated variance $\hat{\sigma}_{\mathcal{U}}^2$ and the estimated bias $\widehat{\mathrm{Bias}}$, the $100 \times (1-\alpha)\%$ confidence interval is

$$\Big(\mathbb{P}_n \widehat{Q}_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_1) + \widehat{\mathrm{Bias}} - \frac{1}{\sqrt{n}} \hat{\sigma}_{\mathcal{U}} z_{\alpha/2}, \mathbb{P}_n \widehat{Q}_1^*(\mathcal{Z}_1; \boldsymbol{\theta}_1) + \widehat{\mathrm{Bias}} + \frac{1}{\sqrt{n}} \hat{\sigma}_{\mathcal{U}} z_{\alpha/2}\Big),$$

where $z_{\alpha/2}$ satisfies $P(T \geqslant z_{\alpha/2}) = \alpha/2$ with $T \sim N(0,1)$.

Remark 5. According to Condition B2 and the assumptions in Theorem 3, $\varpi_{2n} = O(\frac{r_n}{\sqrt{n}})$, $\varrho_{2n} = n^{-\vartheta}$ with $\varpi_{2n} = o(\varrho_{2n})$, and $r_n = o(n^{1/4})$. In addition, according to the assumptions of Theorem 4, it requires $\varpi_{1n} = O(\frac{s_n}{n^{\vartheta}})$, $\varrho_{1n} = O(n^{-\gamma})$ with $\varpi_{1n} = o(\varrho_{1n})$, and $s_n = o(n^{(\nu - 1/4)})$. These asymptotic rates of tuning parameters can serve as a general guide of how to select $(\varrho_{2n}, \varpi_{2n})$ and $(\varrho_{1n}, \varpi_{1n})$. Thus, we can choose $\varrho_{2n} = C_{21} n^{-\tau_{21}}$, $\varpi_{2n} = C_{22} n^{-\tau_{22}}$, and $\varrho_{1n} = C_{11} n^{-\tau_{11}}$, $\varpi_{1n} = C_{12} n^{-\tau_{12}}$, where C_{ij} and τ_{ij} for $i, j = 1, 2$ are positive constants with the constraint of $\tau_{11} < \tau_{12} \leqslant \tau_{21} < \tau_{22} \leqslant 1/2$. In our simulation studies and real data analysis, for simplicity, we fix $\tau_{11} = 1/3$, $\tau_{12} = 2/5$, $\tau_{21} = 2/5$, and $\tau_{22} = 1/2$. It remains to choose the tuning constants C_{ij}, $i, j = 1, 2$. To this end, we use bootstrapping method by constructing bootstrapped confidence intervals for each choice of tuning parameter value, then select the best coverage rates (Altman and Leger 1994). Because the bootstrap is time-consuming especially for the large-scale simulation studies, we recommend to choose (C_{21}, C_{22}) and (C_{11}, C_{12}) separately in different bootstrap procedures instead of choosing them simultaneously. That is, we first choose the tuning parameters C_{21} and C_{22} that give the best average coverage rates of coefficients related to the selected variables in Stage 1 through bootstrapping among a candidate set. By fixing C_{22} and C_{21}, we next choose C_{11} and C_{12} that give the best coverage rates for estimating value function through another bootstrap procedure among a given candidate set. Finally, to choose a penalty ($p_{\lambda_{2n}}(\cdot)$ or $p_{\lambda_{1n}}(\cdot)$) in our HQ-learning procedure, we use the SCAD penalty (Fan and Li 2001) in the following simulation and real data analysis, in which the regularization parameter (λ_{2n} or λ_{1n}) are selected by Bayesian information criterion (BIC).

5. Simulation Studies

We conduct simulation studies to assess the performance of penalization and truncation of the HQ-learning in this section, which are exactly the two key features of our proposed method.

5.1. Simulation Settings

We first simulate m_1 baseline covariates $\mathbf{O}_1 = (O_{11}, \ldots, O_{1m_1})^T$ and the treatment A_1 for the first stage. We further simulate m_2 baseline covariates $\mathbf{O}_2 = (O_{21}, \ldots, O_{2m_2})^T$ and the treatment A_2 in the second stage. The binary treatments A_t's are generated according to $P(A_t = 1) = 1 - P(A_t = -1) = \pi$ for $t = 1, 2$. Without loss of generality, we assume that O_{21} is related to O_{11} and is generated as follows,

$$O_{21} = \delta_1 O_{11} + \delta_2 A_1 + e$$

with $e \sim N(0, \sigma_2^2)$, and $O_{11}, \ldots, O_{1m_1}, O_{22}, \ldots, O_{2m_2}$ are generated independently from $N(0,1)$.

We set $R_1 = 0$. To consider the model sparsity, we treat $O_{12}, \ldots, O_{1m_1}, O_{21}, \ldots, O_{2m_2}$ as noise covariates, and $O_{22}A_2, \ldots, O_{2m_2}A_2$ as noise interactions with A_2 when simulating R_2. Moreover, to evaluate the performance of the variable selections of both stages and statistical inference of the optimal value function for the first stage in the presence of nonregularity, we simulate R_2 from the following model:

$$R_2 = \gamma_1 + \gamma_2 O_{11} + \gamma_3 A_1 + \gamma_4 O_{11} A_1 + (\gamma_5 O_{11} + \gamma_6 O_{21} + \gamma_7 A_1) A_2 \cdot 1_{\{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \alpha_0\}} + \epsilon, \quad (13)$$

where $\epsilon \sim N(0, \sigma_1^2)$, $H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)} = \gamma_5 O_{11} + \gamma_6 O_{21} + \gamma_7 A_1$, and α_0 is a small number used to control the degree of nonregularity, which is defined as the proportion of those subjects without treatment effects among all the subjects.

Based on the sparsity model (13), the Q-function and the optimal Q-function of Stage 2 are, respectively,

$$\begin{aligned} Q_2(\mathbf{O}_1, \mathbf{O}_2, A_1, A_2) &= \gamma_1 + \gamma_2 O_{11} + \gamma_3 A_1 + \gamma_4 O_{11} A_1 \\ &\quad + (\gamma_5 O_{11} + \gamma_6 O_{21} + \gamma_7 A_1) A_2 \\ &\quad \cdot 1_{\{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \alpha_0\}}, \\ Q_2^*(\mathbf{O}_1, \mathbf{O}_2, A_1, A_2) &= \gamma_1 + \gamma_2 O_{11} + \gamma_3 A_1 + \gamma_4 O_{11} A_1 \\ &\quad + |\gamma_5 O_{11} + \gamma_6 O_{21} + \gamma_7 A_1| \\ &\quad \cdot 1_{\{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \alpha_0\}}. \end{aligned}$$

Note that a linear model with covariate $\mathbf{O}_1$, $\mathbf{O}_2$, A_1, and A_2 is fitted for the Q function in the second stage, which is misspecified unless $\alpha_0 = 0$ or the degree of nonregularity is 100% in this simulation. Following this optimal Q-function at Stage 2, the Q-function at Stage 1 is

$$\begin{aligned} &Q_1(\mathbf{O}_1, A_1) \\ &= \gamma_1 + \gamma_2 O_{11} + \gamma_3 A_1 + \gamma_4 O_{11} A_1 \\ &\quad + E\big(|\gamma_5 O_{11} + \gamma_6 O_{21} + \gamma_7 A_1| \cdot 1_{\{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \alpha_0\}} | \mathbf{O}_1, A_1\big) \\ &= \gamma_1 + \gamma_2 O_{11} + \gamma_3 A_1 + \gamma_4 O_{11} A_1 \\ &\quad + |\gamma_5 O_{11} + \gamma_7 A_1| \cdot 1_{\{|H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)}| > \alpha_0\}} \cdot 1_{\{\gamma_6 = 0\}} \\ &\quad + |\gamma_6| \left\{ \mu \left[1 - \Phi\left(\frac{\mu - \gamma_6'}{\sigma_2} \right) - \Phi\left(\frac{-\mu - \gamma_6'}{\sigma_2} \right) \right] \right. \end{aligned}$$

Table 1. Simulation results on variable selection of the second stage for $n = 200$.

		Setting 1		Setting 2		Setting 3	
(p_1, p_2)	NR	FN	FP	FN	FP	FN	FP
(51, 127)	0	0.0000	0.0218	0.0011	0.0206	0.1049	0.0168
	0.25	0.0000	0.0219	0.0018	0.0203	0.1052	0.0168
	0.5	0.0000	0.0221	0.0015	0.0202	0.1674	0.0169
	0.75	0.0000	0.0217	0.0060	0.0200	0.3146	0.0162
	1	0.2428	0.0289	0.3900	0.0251	0.4993	0.0271
(161, 402)	0	0.0000	0.0143	0.0031	0.0130	0.1249	0.0082
	0.25	0.0000	0.0142	0.0032	0.0132	0.1330	0.0081
	0.5	0.0000	0.0143	0.0034	0.0129	0.2138	0.0081
	0.75	0.0000	0.0144	0.0082	0.0127	0.3531	0.0080
	1	0.2467	0.0146	0.3942	0.0141	0.5023	0.0136
(401, 1002)	0.0000	0.0000	0.0090	0.0030	0.0100	0.1320	0.0040
	0.25	0.0000	0.0090	0.0030	0.0100	0.1380	0.0040
	0.5	0.0000	0.0090	0.0030	0.0100	0.2440	0.0050
	0.75	0.0000	0.0090	0.0080	0.0090	0.3740	0.0050
	1	0.2480	0.0120	0.3950	0.0120	0.5080	0.0090

NOTES: FN: percentage of important variables that are missed. FP: percentage of unimportant variables that are selected. NR: the degree of nonregularity.

$$+\frac{\sigma_2}{\sqrt{2\pi}}\left[\exp\left(-\frac{(\mu-\gamma_6')^2}{2\sigma_2^2}\right)+\exp\left(-\frac{(\mu+\gamma_6')^2}{2\sigma_2^2}\right)\right]\Bigg\}\cdot 1_{\{\gamma_6\neq 0\}},$$

where $\mu = (\gamma_5 O_{11} + \gamma_7 A_1)/\gamma_6 + (\delta_1 O_{11} + \delta_2 A_1)$, $\gamma_6' = \alpha_0/|\gamma_6|$, and $\Phi(\cdot)$ is the cumulative distribution function of standard normal. Note that a linear model is also used to fit the Q-function $Q_1(\mathbf{O}_1, A_1)$ in the first stage, which is misspecified unless $\mu = 0$ or the degree of nonregularity is 100% in this simulation.

We choose $\pi = 0.5, \sigma_1 = \sigma_2 = 1$, and consider the following three settings with different levels of model misspecification:

- Setting 1: $\gamma_5 = \gamma_7 = \delta_1 = \delta_2 = 0$, and all other γ's equal 1;
- Setting 2: $\gamma_6 = 0, \delta_1 = \delta_2 = 0.5$, and all other γ's equal 1;
- Setting 3: $\gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = 0.5$, $\gamma_5 = \gamma_6 = \gamma_7 = 0.8$, and $\delta_1 = \delta_2 = 0.8$.

The model is correctly specified under Setting 1, while the level of model misspecification under Setting 3 is higher than that under Setting 2 since our HQ-learning uses a linear model to fit the Q-function $Q_1(\mathbf{O}_1, A_1)$ in the first stage. To choose the tuning parameters, we set $C_{11} \in \{1, \ldots, 10\}$, $C_{12} \in \{1, \ldots, 6\}$, $C_{21} \in \{5, \ldots, 15\}$, and $C_{22} \in \{1, \ldots, 10\}$, then we try different combinations of C_{21} and C_{22} in the first bootstrap procedure, and different combinations of C_{11} and C_{12} in the next bootstrap procedure, respectively. All results are based on 5000 replications in the simulation studies.

5.2. Comparison With HQ-learning Without Truncation

As mentioned previously, our HQ-learning used a truncated pseudo-outcome in Step 2 of the first-stage regression to ease the inference in the presence of nonregularity. For the purpose of comparison, we consider an alternative procedure, which used the pseudo-outcome without truncation in Step 2 of the first-stage regression. We call the alternative procedure HQ-learning without truncation, and compare it with our HQ-learning in the simulation studies. The simulation study compares the competing estimators on a variety of settings under different degrees of nonregularity. To control the degrees of nonregularity, we choose $\alpha_0 = 0, 0.32, 0.68, 1.15$, and 5 in Setting 1, $\alpha_0 = 0, 0.52, 1.05, 1.7$, and 5 in Setting 2, and $\alpha_0 = 0, 0.7625, 1.58, 2.588$, and 10 in Setting 3, which correspond to the proportions of nonregularity 0%, 25%, 50%, 75%, and 100%, respectively. We consider sample size $n = 200$, and set $m_1 = m_2 = 25$, 80, and 200, then the dimensions of the covariates are 51, 161, and 401 for the first state, and 127, 402, and 1002 for the second stage, respectively. Additional results for $n = 400$ and 600 are provided in the supplementary material. We not only evaluate the performance of the variable selections of both stages, but also consider estimations and inference for ψ_{11}, ψ_{12}, and the optimal value function for the first stage in the presence of nonregularity. The population parameter ψ_{11}, ψ_{12} and the true value function are estimated based on Monte Carlo procedure by choosing linear model as the working model.

Tables 1 and 2 summarize variable selection results for estimated optimal Q-functions of both second and first stages, respectively. Specifically, we report the false negative (FN) rate (the percentage of important variables that are missed) and false positive (FP) rate (the percentage of unimportant variables that are selected) of the 5000 replications. Based on the Q-function $Q_1(\mathbf{O}_1, A_1)$ given in Section 5.1, we clearly know which variables are important and which are not although the models are misspecified under settings 2 and 3. The results of variable selection of HQ-learning are identical with that of HQ-learning without truncation in Stage 2 because both methods are exactly the same for Step 1. We can easily see from Table 1 that FN and FP rates are very small in most cases, which shows that both methods can identify the important predictors and remove noise covariates properly for Stage 2 in the estimation of the optimal dynamic treatment regime when the dimensional covariates is very high. However, the FN rates become higher when the degree of nonregularity is 100% for all three settings as well as when the degree of nonregularity is higher (75%, even for 50%) for Setting 3. This is not surprising as $H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)} = 0$ for all subjects if the degree of nonregularity is 100%, which is equivalent that $\gamma_5 = \gamma_6 = \gamma_7 = 0$ in Stage 2. At the same time, $H_{2,I}^T \boldsymbol{\psi}_{20}^{(1)} = 0$ for most of subjects if the degree of nonregularity is 75%. As a result, the variables attached to these coefficients (especially for γ_6) will be missed during selecting variables. Table 2 shows that the results of variable selection for HQ-learning and HQ-learning without truncation are very similar for all settings, and both methods have very good performance on variable selection in Stage 1 whatever the degree of nonregularity is in

Table 2. Simulation results on variable selection via HQ-learning and HQ-learning without truncation of the first stage for $n = 200$.

		Setting 1				Setting 2				Setting 3			
		FN		FP		FN		FP		FN		FP	
(p_1, p_2)	NR	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL
(51, 127)	0	0.0000	0.0000	0.0422	0.0397	0.0002	0.0003	0.0505	0.0458	0.0735	0.0807	0.0341	0.0314
	0.25	0.0000	0.0000	0.0421	0.0401	0.0008	0.0009	0.0501	0.0468	0.0713	0.0851	0.0348	0.0310
	0.5	0.0000	0.0000	0.0430	0.0406	0.0003	0.0003	0.0503	0.0487	0.0946	0.0986	0.0366	0.0354
	0.75	0.0000	0.0000	0.0439	0.0433	0.0002	0.0002	0.0499	0.0494	0.0887	0.0981	0.0371	0.0357
	1	0.0000	0.0000	0.0466	0.0480	0.0000	0.0000	0.0460	0.0484	0.0002	0.0002	0.0448	0.0467
(161, 402)	0	0.0000	0.0000	0.0311	0.0294	0.0015	0.0015	0.0359	0.0317	0.086	0.0917	0.0244	0.0237
	0.25	0.0000	0.0000	0.0309	0.0291	0.0013	0.0016	0.0361	0.0331	0.0884	0.1016	0.0245	0.0228
	0.5	0.0000	0.0000	0.0318	0.0306	0.0006	0.0005	0.0357	0.0323	0.1161	0.1287	0.0249	0.0239
	0.75	0.0000	0.0000	0.0335	0.0342	0.0004	0.0003	0.0359	0.0349	0.1054	0.1194	0.0267	0.0255
	1	0.0000	0.0000	0.0354	0.0356	0.0000	0.0000	0.0353	0.0352	0.0003	0.0003	0.0338	0.0344
(401, 1002)	0	0.0000	0.0000	0.0190	0.0180	0.0010	0.0010	0.0200	0.0190	0.0900	0.1060	0.0150	0.0140
	0.25	0.0000	0.0000	0.0190	0.0180	0.0010	0.0010	0.0200	0.0190	0.0940	0.1160	0.0150	0.0140
	0.5	0.0000	0.0000	0.0190	0.0180	0.0010	0.0010	0.0210	0.0190	0.1270	0.1530	0.0150	0.0140
	0.75	0.0000	0.0000	0.0200	0.0190	0.0000	0.0000	0.0210	0.0200	0.1200	0.1360	0.0160	0.0160
	1	0.0000	0.0000	0.0210	0.0210	0.0000	0.0000	0.0210	0.0210	0.0000	0.0000	0.0220	0.0220

NOTES: FN: percentage of important variables that are missed. FP: percentage of unimportant variables that are selected. NR: the degree of nonregularity. HQLN: HQ-learning without truncation. HQL: HQ-learning.

Stage 2. Combining Table 1 and Table 2, we conclude that the truncated pseudo-outcome used in the first-stage regression did not deteriorate variable selection performance for the estimation of the optimal dynamic treatment regime, which is consistent with the theoretical observations in terms of variable selection of our HQ-learning.

Tables 3–5 list the biases of the estimates of ψ_{11}, ψ_{12}, the optimal value function for the first stage as well as the empirical coverage probability of 95% nominal percentile confidence interval for ψ_{11}, ψ_{12}, and the optimal value function of Stage 1. From the results we can see that although our HQ-learning induced some biases due to the truncation, after carefully adjusting for the biases, the biases of HQ-learning are comparable with that of HQ-learning without truncation. It should be pointed out that the bias of HQ-learning without truncation becomes very large when the degree of nonregularity is 100% for most cases, but the HQ-learning still has reasonable bias under these cases. Moreover, the standard errors of the estimates of both HQ-learning and HQ-learning without truncation are also reasonable and comparable.

In settings 1 and 2, the coverage rates of ψ_{11} and ψ_{12} of HQ-learning are a bit closer to the nominal percentile 95% than that of HQ-learning without truncation, but the coverage rates of both methods are basically comparable in most cases. In Setting 3, the coverages of ψ_{11} and ψ_{12} for both methods are a little low except the case of 100% nonregularity, but HQ-learning still has better coverages than HQ-learning without truncation for some cases. Most likely, the problem of model misspecification becomes more severe for Setting 3, so the performance of variable selection is not perfect (see Tables 1 and 2), however, the Q-function $Q_1(\mathbf{O}_1, A_1)$ in the first stage is exactly linear when the degree of nonregularity is 100%, which was correctly specified for the estimation of the optimal dynamic treatment regime in the first stage.

For the coverage of the optimal value function, the coverage rates of HQ-learning are higher and closer to 95% than that of HQ-learning without truncation for all three settings when the degree of nonregularity is 100%. For the other degrees of nonregularity, the coverages of the optimal value function for HQ-learning are slightly closer to the nominal level than that of

Table 3. Summary statistics and coverage rates of 95% nominal percentile for ψ_{11}, ψ_{12}, and the optimal value function of the first stage for $n = 200$ (Setting 1).

		ψ_{11}						ψ_{12}						Q1					
		Bias		SE		CI		Bias		SE		CI		Bias		SE		CI	
(p_1, p_2)	NR	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL
(51, 127)	0	−0.0080	−0.0079	0.0801	0.0859	93.2	94.7	−0.0076	−0.0076	0.0799	0.0858	91.6	93.8	0.0234	−0.0036	0.1695	0.1689	93.8	93.8
	0.25	−0.0084	−0.0083	0.0802	0.0842	93.4	94.7	−0.0079	−0.0078	0.0800	0.0841	92.1	93.7	0.0219	0.0053	0.1698	0.1723	93.7	94.2
	0.5	−0.0088	−0.0087	0.0807	0.0856	93.0	94.5	−0.0080	−0.0080	0.0805	0.0855	91.9	93.6	0.0213	0.0143	0.1715	0.1723	94.3	94.3
	0.75	−0.0093	−0.0094	0.0810	0.0839	92.5	93.4	−0.0064	−0.0065	0.0809	0.0838	92.1	93.1	0.0273	0.0120	0.1751	0.1755	93.7	93.5
	1	−0.0094	−0.0092	0.0681	0.0675	92.7	92.4	−0.0054	−0.0052	0.0679	0.0673	92.1	91.9	0.1438	0.0159	0.1575	0.1569	81.2	94.5
(161, 402)	0	−0.0073	−0.0071	0.0774	0.0829	90.5	92.7	−0.0052	−0.0053	0.0774	0.0829	89.9	92.1	0.0482	0.0381	0.1673	0.1693	93.1	93.8
	0.25	−0.0071	−0.0071	0.0776	0.0831	89.7	92.3	−0.0058	−0.0059	0.0775	0.0830	90.0	92.3	0.0470	0.0120	0.1676	0.1718	93.0	93.6
	0.5	−0.0074	−0.0077	0.0782	0.0819	90.0	91.7	−0.0058	−0.0059	0.0781	0.0818	89.7	91.1	0.0461	0.0052	0.1691	0.1724	92.6	93.5
	0.75	−0.0089	−0.0089	0.0785	0.0766	90.0	88.8	−0.0052	−0.0051	0.0784	0.0765	89.3	88.4	0.0582	0.0391	0.1721	0.1637	92.2	91.6
	1	−0.0086	−0.0087	0.0659	0.0655	90.2	89.8	−0.0033	−0.0032	0.0659	0.0655	89.3	88.8	0.2198	0.1638	0.1574	0.1554	69.7	78.6
(401, 1002)	0	0.0017	0.0022	0.0936	0.0995	87.7	90.1	−0.0017	−0.0017	0.0937	0.0986	87.4	89.3	0.0723	0.0452	0.1788	0.1937	91.0	92.1
	0.25	0.0020	0.0022	0.0935	0.0994	88.0	90.0	−0.0012	−0.0009	0.0941	0.0993	87.6	89.1	0.0720	0.0273	0.1801	0.2003	90.3	92.5
	0.5	0.0022	0.0024	0.0943	0.1002	87.7	90.0	−0.0024	−0.0030	0.0950	0.1010	87.1	89.4	0.0711	0.0043	0.1819	0.2095	90.7	92.3
	0.75	0.0052	0.0050	0.0956	0.0994	87.4	89.0	−0.0027	−0.0025	0.0953	0.0989	87.3	88.9	0.0876	0.0515	0.1871	0.2141	89.7	91.2
	1	−0.0097	−0.0097	0.0812	0.0801	86.4	85.5	−0.0055	−0.0053	0.0811	0.0803	86.8	86.1	0.2668	0.1550	0.1878	0.1694	56.4	75.4

NOTES: NR: the degree of nonregularity. HQLN: HQ-learning without truncation. HQL: HQ-learning.

Table 4. Summary statistics and coverage rates of 95% nominal percentile for ψ_{11}, ψ_{12}, and the optimal value function of the first stage for $n = 200$ (Setting 2).

		ψ_{11}						ψ_{12}						Q1					
		Bias		SE		CI		Bias		SE		CI		Bias		SE		CI	
(p_1, p_2)	NR	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL
(51, 127)	0	−0.0111	−0.0111	0.0745	0.0834	92.7	95.6	−0.0174	−0.0160	0.0951	0.1025	91.2	93.0	0.0037	−0.0393	0.1890	0.2029	93.7	93.4
	0.25	−0.0117	−0.0117	0.0754	0.0834	92.9	95.0	−0.0171	0.0018	0.0962	0.1038	91.3	92.8	0.0024	−0.0016	0.1899	0.2043	93.5	94.8
	0.5	−0.0105	−0.0106	0.0801	0.0859	92.8	94.5	−0.0123	0.0162	0.1013	0.1098	91.7	92.5	0.0046	−0.0073	0.1958	0.2053	93.2	93.9
	0.75	−0.0123	−0.0124	0.0867	0.0898	92.7	93.5	−0.0241	−0.0129	0.1128	0.1199	90.4	91.8	0.0000	−0.0158	0.2133	0.2113	92.6	91.2
	1	−0.0089	−0.0089	0.0681	0.0679	92.6	92.6	−0.0046	−0.0047	0.0682	0.0680	92.2	91.8	0.1550	0.0251	0.1584	0.1567	80.8	92.9
(161, 402)	0	−0.0120	−0.0121	0.0725	0.0816	89.6	92.8	−0.0274	−0.0172	0.0914	0.0989	88.0	91.2	0.0141	0.0023	0.1870	0.2028	92.7	94.1
	0.25	−0.0118	−0.0116	0.0734	0.0805	89.7	92.2	−0.0256	−0.0201	0.0922	0.0994	88.7	91.0	0.0137	0.0026	0.1877	0.1928	93.0	93.4
	0.5	−0.0111	−0.0111	0.0781	0.0848	89.9	92.5	−0.0187	−0.0175	0.0973	0.1049	88.7	91.3	0.0162	−0.0053	0.1932	0.1990	92.6	92.9
	0.75	−0.0120	−0.0120	0.0843	0.0874	90.3	91.5	−0.0309	−0.0261	0.1082	0.1137	86.3	87.4	0.0150	0.0112	0.2082	0.2050	91.0	90.0
	1	−0.0077	−0.0077	0.0659	0.0662	90.3	90.0	−0.0040	−0.0039	0.0661	0.0663	89.4	89.2	0.2147	0.1747	0.1576	0.1575	70.6	77.9
(401, 1002)	0	−0.0030	−0.0033	0.0963	0.1037	87.7	90.8	−0.0518	−0.0263	0.1424	0.1563	84.5	88.6	0.0179	−0.0087	0.2116	0.2651	92.0	94.1
	0.25	−0.0022	−0.0021	0.0962	0.1018	87.1	90.0	−0.0449	−0.0279	0.1395	0.1490	85.8	88.9	0.0203	0.0084	0.2085	0.2298	92.2	93.1
	0.5	−0.0004	0.0001	0.0939	0.1005	87.7	90.7	−0.0233	−0.0218	0.1462	0.1639	86.1	88.7	0.0321	0.0002	0.2168	0.2789	91.6	92.8
	0.75	0.0015	0.0014	0.1027	0.1067	87.9	89.1	−0.0374	−0.0362	0.1670	0.1838	82.0	83.8	0.0286	0.0322	0.2453	0.2993	89.6	89.5
	1	−0.0074	−0.0076	0.0815	0.0817	86.1	86.3	−0.0045	−0.0046	0.0825	0.0827	85.8	86.0	0.2823	0.1344	0.1798	0.1889	53.3	80.6

NOTES: NR: the degree of nonregularity. HQLN: HQ-learning without truncation. HQL: HQ-learning.

HQ-learning without truncation for most of cases in Setting 1, and the coverages of the optimal value function for both methods are comparable in settings 2 and 3. Again, due to model misspecification, the coverages of the optimal value function for both HQ-learning and HQ-learning without truncation become low in Setting 3, especially for HQ-learning without truncation with 100% degree of nonregularity. Besides, for our HQ-learning in Setting 3, the results under 100% degree of nonregularity are better than that of 75% degree of nonregularity (even others) based on the coverages of ψ_{11}, ψ_{12}, and value function (Table 5) as well as the values of FN in the first stage (Table 2), which is also because the model was correctly specified under this case.

In conclusion, by truncation, our HQ-learning can improve the coverage rates via bias adjustment in the presence of nonregularity in most of cases, especially when the underling model is correctly specified and/or the degree of nonregularity is very high. Furthermore, the running time of the HQ-learning is acceptable. For example, for the case of $n = 200$, $(p_1, p_2) = (401, 1002)$ at the degree of nonregularity 50%, the simulation with 5000 replications took 5.68 hr to run on a computer with a 2.3GHz dual core processor.

5.3. Comparison With the Nonpenalized Q-learning

In this subsection, we compare our HQ-learning with the standard Q-learning (without penalty and without truncation) and the soft-threshold Q-learning (Chakraborty, Murphy, and Strecher 2010). The simulation setting is the same as in Section 5.2, except that $m_1 = m_2 = 15$ and the dimensions of the covariates are 31 and 77 for each stage, respectively, since the nonpenalized methods are not applicable for a higher dimension. Moreover, for the standard Q-learning and the soft-threshold Q-learning, as the confidence interval constructions are not based on explicit formulas, we consider percentile and hybrid bootstrap confidence intervals similar to Chakraborty, Murphy, and Strecher (2010).

The simulation results are presented in Figures 1 and 2, where we report the mean square errors (MSEs) and the coverage rates

Table 5. Summary statistics and coverage rates of 95% nominal percentile for ψ_{11}, ψ_{12}, and the optimal value function of the first stage for $n = 200$ (Setting 3).

		ψ_{11}						ψ_{12}						Q1					
		Bias		SE		CI		Bias		SE		CI		Bias		SE		CI	
(p_1, p_2)	NR	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL	HQLN	HQL
(51, 127)	0	−0.0764	−0.0784	0.0828	0.0895	79.5	81.5	−0.1122	−0.1108	0.1169	0.1260	78.5	80.4	−0.1014	−0.1106	0.1791	0.1829	82.2	82.7
	0.25	−0.0745	−0.0787	0.0841	0.0943	79.4	82.3	−0.1089	−0.0867	0.1186	0.1326	78.8	81.2	−0.1012	−0.0994	0.1812	0.1886	82.3	83.5
	0.5	−0.0949	−0.0960	0.0865	0.0895	74.6	75.1	−0.1563	−0.1214	0.1269	0.1315	71.1	72.8	−0.1515	−0.1916	0.1979	0.1986	78.3	75.4
	0.75	−0.0905	−0.0920	0.0964	0.1014	74.7	76.3	−0.2137	−0.2084	0.1368	0.1489	66.7	70.3	−0.1895	−0.2062	0.2315	0.2297	81.4	79.5
	1	−0.0171	−0.0172	0.0676	0.0671	89.7	89.6	−0.0061	−0.0063	0.0682	0.0676	91.7	91.4	0.1390	0.0189	0.1162	0.1073	69.8	90.6
(161, 402)	0	−0.0909	−0.0934	0.0801	0.0837	75.3	76.5	−0.1367	−0.1008	0.1130	0.1182	74.0	74.7	−0.1079	−0.1251	0.1780	0.1696	79.5	76.9
	0.25	−0.0956	−0.1007	0.0808	0.0880	74.7	77.1	−0.1402	−0.1027	0.1148	0.1257	73.6	75.4	−0.1175	−0.1202	0.1803	0.1791	79.0	78.7
	0.5	−0.1205	−0.1247	0.0835	0.0877	70.1	71.9	−0.2104	−0.2043	0.1235	0.1322	62.7	65.2	−0.1893	−0.2013	0.1968	0.2002	73.7	73.4
	0.75	−0.1154	−0.1194	0.0912	0.0956	68.1	70.5	−0.2395	−0.2042	0.1308	0.1469	62.2	68.7	−0.1895	−0.3002	0.2276	0.2200	80.1	67.7
	1	−0.0204	−0.0206	0.0658	0.0654	86.2	85.9	−0.0119	−0.0121	0.0663	0.0659	88.1	87.8	0.2019	0.0301	0.1153	0.1037	54.1	88.7
(401, 1002)	0	−0.0897	−0.0968	0.1962	0.2062	72.1	75.2	−0.1523	−0.1162	0.2820	0.3065	70.6	73.6	−0.0939	−0.1190	0.2986	0.3498	79.3	79.6
	0.25	−0.0909	−0.0999	0.1971	0.2095	71.9	76.0	−0.1562	−0.1341	0.2831	0.3144	68.3	73.3	−0.1007	−0.1106	0.2993	0.3710	79.2	80.1
	0.5	−0.1216	−0.1317	0.2147	0.2257	66.2	70.3	−0.2453	−0.2420	0.2915	0.3186	55.6	59.7	−0.1842	−0.1909	0.3082	0.3672	74.3	74.7
	0.75	−0.1257	−0.1303	0.2132	0.2188	65.6	66.5	−0.2537	−0.2307	0.2585	0.2641	57.1	62.0	−0.1652	−0.2167	0.2978	0.3363	80.4	73.6
	1	−0.0232	−0.0234	0.0934	0.0927	80.5	79.5	−0.0212	−0.0212	0.0879	0.0863	82.5	81.9	0.2491	−0.0050	0.1643	0.1328	40.2	83.1

NOTES: NR: the degree of nonregularity. HQLN: HQ-learning without truncation. HQL: HQ-learning.

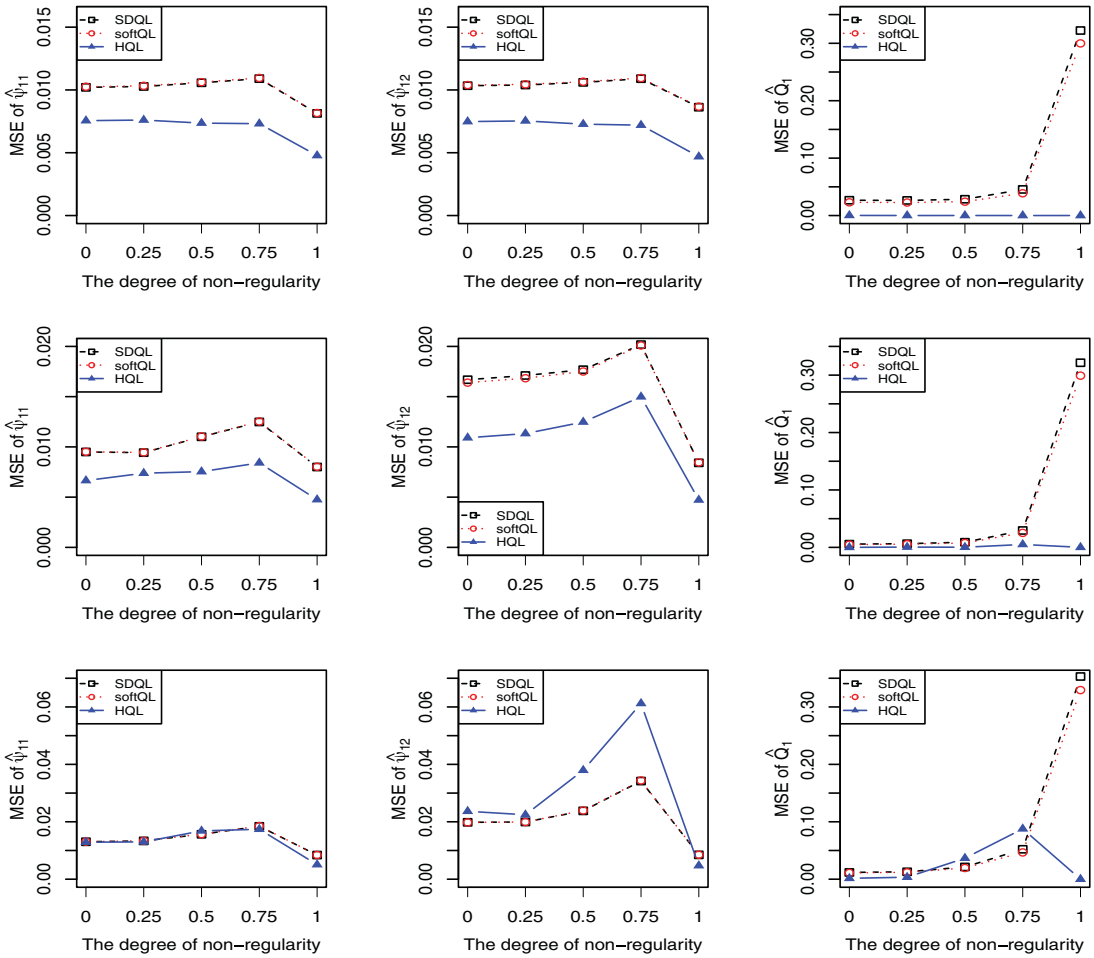

Figure 1. MSE of the estimates of ψ_{11}, ψ_{12}, and the optimal value function for three settings: Setting 1 (first row), Setting 2 (second row), and Setting 3 (third row). HQL: HQ-learning. SDQL: the standard Q-learning. softQL: the soft-threshold Q-learning.

of ψ_{11}, ψ_{12}, and the optimal value function for the first stage. In settings 1 and 2, we can see that our HQ-learning has better performance, especially for the inference of ψ_{11} and ψ_{12}, than the other two nonpenalized methods, from the points of producing smaller MSEs and better coverage rates of 95% confidence intervals. For the standard Q-learning and the soft-threshold Q-learning, although the coverage rates of the hybrid bootstrap confidence intervals for the optimal value function are reasonable, the results for MSEs and the coverage rates of the percentile bootstrap confidence intervals become worse when the degree of nonregularity increases. The HQ-learning does not perform very well in Setting 3, because the performance of variable selection is heavily affected by model misspecification. In contrast, the nonpenalized methods are nearly not affected by model misspecification, but they generally fail without variable selection when the dimension is high.

6. Application to STAR*D Study

6.1. Background

Sequenced Treatment Alternatives to Relieve Depression (STAR*D) is a multisite, sequentially randomized study that aims to explore dynamically which is the most effective treatment plan for each patient with major depressive disorder (MDD) (Fava et al. 2003; Rush et al. 2004). This study initially recruited a total of 4041 patients with nonpsychotic MDD and those who have adequate clinical responses after each treatment level will exempt from future randomization for the next treatment level and enter the 12-month naturalistic follow-up phase. For each treatment level, protocol medication clinic visits of the participants without a satisfactory clinical outcome are required at weeks 0, 2, 4, 6, 9, and 12.

At level 1, all 4041 participants were treated with citalopram (CIT) and those who without a satisfactory clinical outcome to CIT are eligible to enter treatment level 2. At level 2, patients were randomly assigned to one of seven treatment options including four switch options (venlafaxine (VEN), sertraline (SER), bupropion (BUP), and cognitive therapy (CT)) and three augment options (augmenting CIT with CT, BUP, or buspirone (BUS)). Patients who were assigned to CT or CT+CIT in level 2 are eligible for a supplementary level treatment (level 2A) if they did not have a satisfactory response in level 2 and will further be treated with VEN or BUP in level 2A. At level 3, patients without a satisfactory response to level 2 or level 2A would continue to be randomly assigned to one of four treatments: mirtazapine (MIRT), nortriptyline (NTP), and augmentation their previous treatment with either lithium (Li) or thyroid hormone (THY). Patients who did not respond satisfactorily at level 3 would move to level 4 treatment, which included

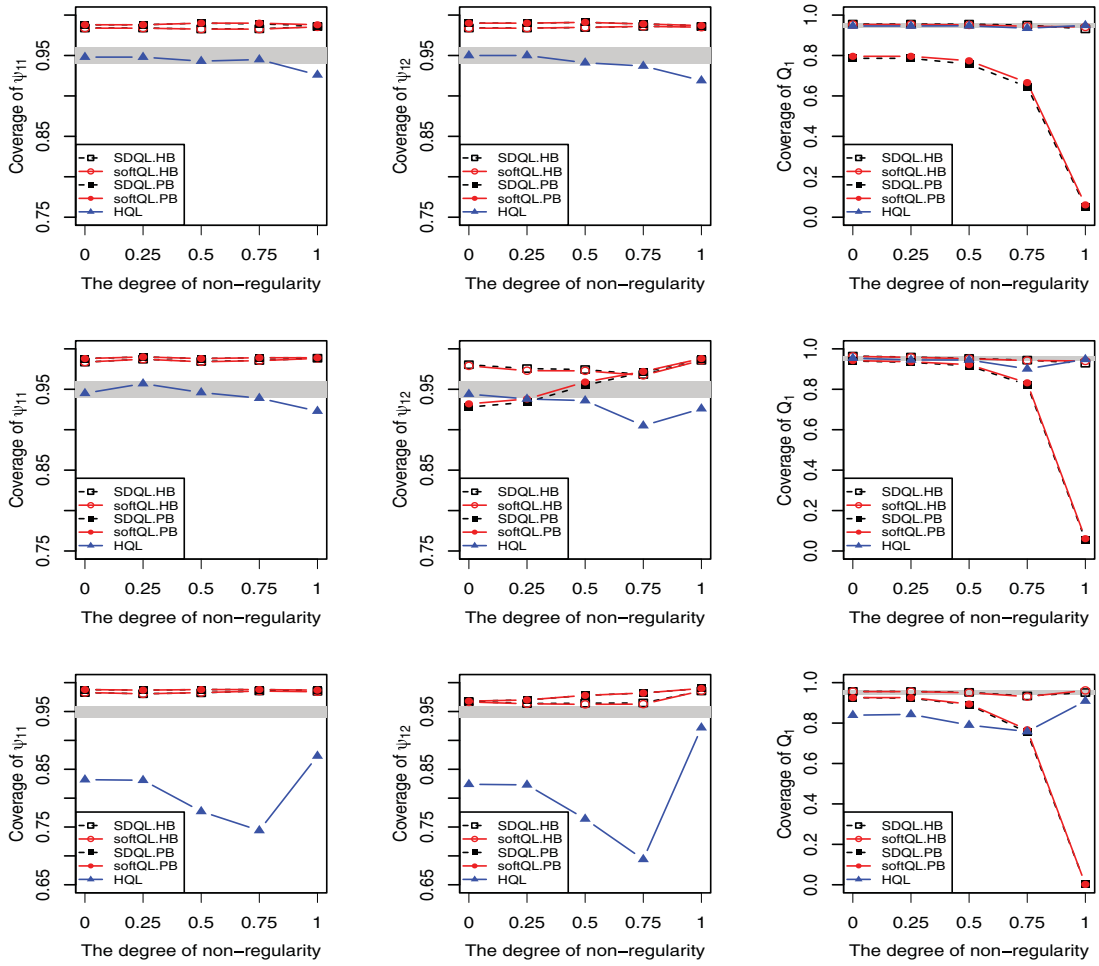

Figure 2. Coverage rates of 95% confidence intervals for ψ_{11}, ψ_{12}, and the optimal value function for three settings: Setting 1 (first row), Setting 2 (second row), and Setting 3 (third row). HQL: HQ-learning. SDQL.PB: the standard Q-learning with percentile bootstrap. SDQL.HB: the standard Q-learning with hybrid bootstrap. softQL.PB: the soft-threshold Q-learning with percentile bootstrap. softQL.HB: the soft-threshold Q-learning with hybrid bootstrap.

two options: tranylcypromine (TCP) or MIRT+VEN. For more details of the STAR*D study, see Fava et al. (2003) and Rush et al. (2004).

6.2. Data Analysis

In our data analysis, we use a subset of patients from level 2 and level 3 of the STAR*D data to illustrate how HQ-learning can be used to simultaneously estimate an optimal dynamic treatment regime and select the important variables in the presence of nonregularity as well as high-dimensional covariates. We refer to level 2 and level 3 of the trial as stages 1 and 2, respectively. We will not consider treatments CT and CT+CIT in our study because patients received treatment CT or CT+CIT in level 2 moved to level 2A. Since our HQ-learning assumes a dichotomized treatment design at each stage, we combine the rest of five treatment options into two treatments in level 2 and combine four treatments into two treatments in level 3. Specifically, let the first-stage treatment $A_1 = 1$ if level 2 treatment belongs to switch options (VEN, SER, and BUP) and $A_1 = -1$ if it belongs to augment options (CIT+BUP and CIT+BUS). For level 3, let the second-stage $A_2 = 1$ if the treatment is MIRT or NTP and $A_2 = -1$ if it is the previous treatment plus Li or THY. To access the performance of variable selection and value estimation of the proposed procedure in the high-dimensional case, we collected 212 covariates in total which come from the baseline and intermediate levels in this study. For treatment regime at level 2, 206 of 212 covariates that were collected before giving treatment at level 2, the treatment A_1, as well as all the interactions between 206 covariates and A_1 were considered to fit the penalized regression model, in which the dimensionality is 413. In the penalized regression model of level 3, all 212 covariates, the treatments A_1 and A_2, as well as all the 206 interactions with A_1 in level 2 were treated as the main predictors, and their interactions with A_2 were also included in the model. The maximum dimensionality involved in the model of level 3 is 837, which is much greater than the sample size involved in the following analysis. Since the severity of depression was measured by Quick Inventory of Depressive Symptomatology-Clinician (QIDS-C16), where higher values of QIDS-C16 score correspond to more severe negative symptoms, we use the negative QIDS-C16 score as the rewards right after the treatment at each level such that higher values correspond to better clinical rewards. By removing the patients whose information relative to any one of 212 covariates is missing, only a subset containing 261 patients with complete information remained in the final data analysis. Of the total 261 patients, 164 were assigned switch treatments ($A_1 = 1$)

Table 6. Coefficient estimates of the selected variables as well as their standard errors and 95% confidence intervals for HQ-learning.

	Variable	Estimate	SE	95% CI
Level 3	*In protocol eligibility form at baseline*			
	Fatigue or loss energy (DSMLE)	− 3.8852	3.4419	(−10.6312, 2.8607)
	In clinic visit clinical record form at baseline			
	QIDS-C beginning score (QCBEG-0)	− 0.5573	0.3090	(−1.1630, 0.0484)
	QIDS-C current score (QCCUR-0)	− 0.2057	0.2377	(−0.6716, 0.2601)
	In clinic visit clinical record form at Level 1			
	QIDS-C percent improvement (QCIMP-1)	0.0722	0.0476	(−0.0211, 0.1655)
	In clinic visit clinical record form at Level 2			
	QIDS-C percent improvement (QCIMP-2)	0.0855	0.0589	(−0.0300, 0.2010)
	CGII score (CGI-I-2)	− 0.5938	0.8227	(−2.2063, 1.0186)
	In psychiatric diagnostic screening questionnaire form at baseline			
	TE Avoid activities remind of trauma (TERMD)	− 1.0049	1.1197	(−3.1994, 1.1896)
	The interaction between "EM Worry about embarrassing self" and A_2 (EMWRY*A_2)	− 0.8596	1.4378	(−3.6777, 1.9584)
	The interaction between "Worry daily" and A_2 (WYDLY*A_2)	0.2508	1.3261	(−2.3483, 2.8498)
	In cumulative illness rating scale form at baseline			
	The interaction between "Upper GI" and A_2 (UGI*A_2)	− 0.8812	0.9226	(−2.6894, 0.9270)
Level 2	*In cumulative illness rating scale form at baseline*			
	Upper GI (UGI)	0.8109	0.2787	(0.4241, 1.5165)
	In protocol eligibility form at baseline			
	Highest degree received (DEGREE)	0.2058	0.0891	(0.0392, 0.3886)
	Fatigue or loss energy (DSMLE)	− 4.5944	1.4319	(−7.6173, −2.0044)
	In clinic visit clinical record form at baseline			
	QIDS-C beginning score (QCBEG-0)	− 0.4278	0.1354	(−0.6991, −0.1683)
	QIDS-C current score (QCCUR-0)	− 0.2886	0.0982	(−0.4867, −0.1016)
	In clinic visit clinical record form at Level 1			
	QIDS-C percent improvement (QCIMP-1)	0.0333	0.0177	(−0.0014, 0.0680)
	In psychiatric diagnostic screening questionnaire form at baseline			
	Did you get less joy or pleasure from almost all of the things you normally enjoy during the past two weeks? (JOY2W)	− 0.8967	0.3873	(−1.6804, −0.1622)
	The interaction between "TE Avoid activities remind of trauma" and A_1 (TERMD*A_1)	− 1.1912	0.4684	(−2.0987, −0.2626)
	Value function	− 10.1136	0.35771	(−10.8147, −9.4125)

and 97 were assigned augment treatments ($A_1 = -1$) at level 2, and 158 were assigned switch treatments MIRT or NTP ($A_2 = 1$) and 103 were assigned the previous treatment plus Li or THY ($A_2 = -1$) at level 3. In this data analysis, we set $C_{11} \in \{1, \ldots, 15\}$, $C_{12} \in \{1, \ldots, 10\}$, $C_{21} \in \{1, \ldots, 25\}$, and $C_{22} \in \{1, \ldots, 10\}$ to choose tuning parameters ϱ_{1n}, ϖ_{1n}, ϱ_{2n}, and ϖ_{2n} by using bootstrap procedure mentioned above.

Our HQ-learning selected ten and eight variables for levels 3 and 2, respectively. Table 6 lists all the selected variables, their coefficient estimates, standard errors of these estimates, and 95% confidence intervals for each stage. Based on these 95% confidence intervals, none of the selected variables at level 3 have significant effect on the clinical outcome, while all eight selected variables at level 2 have strong effect on the pseudo-outcome. This shows that no significant level 3 treatment effect was found at Stage 2, which further indicates possible existence of nonregularity. In particular, we found significant level 2 treatment effect (TERMD*A_1) on the pseudo-outcome. We thus summarize the estimated optimal rule based on HQ-learning as follows. At level 3, there is no difference in switching MIRT and NTP or continuing the previous treatment plus Li or THY. If a patient tried to avoid activities, places, or people that reminded him/her of a traumatic event (i.e., TERMD = 1) during the past 2 weeks before the patient was enrolled in the study, and remained unsatisfactory in level 1, augment treatment (CIT+BUP or CIT+BUS) is a better option for level 2 treatment when compared to switch option (VEN, SER, or BUP). We also list the value estimation, its standard error, and the corresponding 95% confidence interval in Table 6.

To further examine the merits of our HQ-learning, we compared it with HQ-learning without truncation, the standard Q-learning, and the soft-threshold Q-learning (Chakraborty, Murphy, and Strecher 2010). With the same dataset, HQ-learning without truncation selected the same variables for each stage as HQ-learning. By truncation, our HQ-learning got a smaller standard error for the value estimation, as a result, the length of the confidence interval becomes shorter. It is very likely that this simple truncation reduced the difficulty of inference for the Q-learning due to nonregularity in the data analysis by simultaneously adjusting for bias induced during truncation. Because both the standard Q-learning and the soft-threshold Q-learning failed in handling the high dimensionality of the dataset, we first applied principal components analysis to each stage and chose age, gender, and the top 15 principal component scores as the covariates of interest. Based on these covariates of interest as well as their interactions with treatment A_1 and/or A_2 (the dimensionalities of Stage 1 and Stage 2 are 35 and 71, respectively), we compared HQ-learning, HQ-learning without truncation, the standard Q-learning, and the soft-threshold Q-learning. Figure 3 shows that the widths of the 95% confidence intervals of HQ-learning are obviously shorter than that of the nonpenalized Q-learning for all the components selected by HQ-learning. Besides, both the standard Q-learning and the soft-threshold Q-learning missed the ninth component and the significant level 2 treatment effect (Component 1*A_1) on the pseudo-outcome. All these results indicate that our HQ-learning outperforms the nonpenalized Q-learning in general.

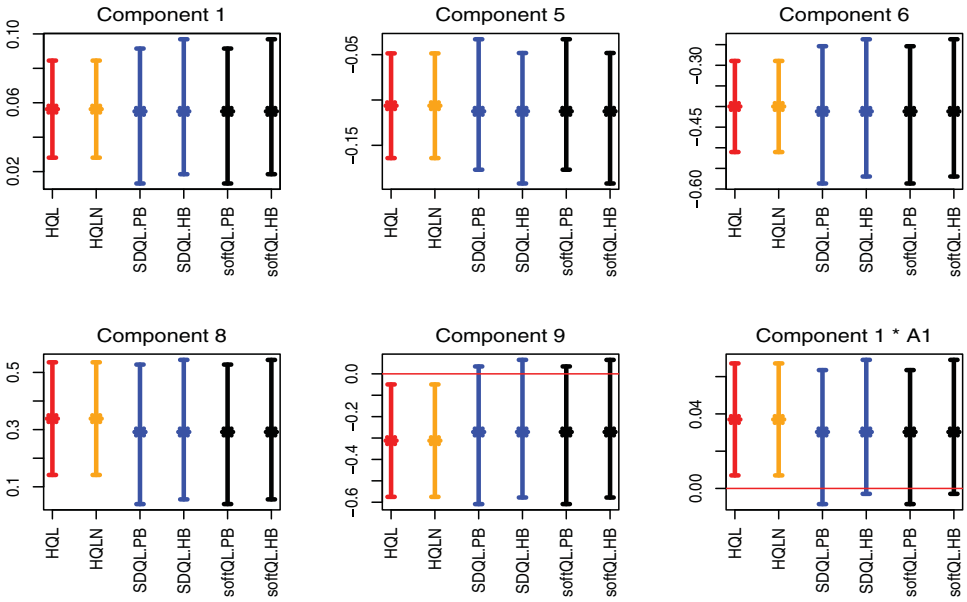

Figure 3. Width of 95% confidence intervals of HQ-learning (HQL), HQ-learning without truncation (HQLN), the standard Q-learning with percentile bootstrap (SDQL.PB) and with hybrid bootstrap (SDQL.HB), and the soft-threshold Q-learning with percentile bootstrap (softQL.PB) and with hybrid bootstrap (softQL.HB) for the components selected by HQ-learning at level 2.

7. Discussion

In this article, we proposed a high-dimensional Q-learning procedure to estimate optimal dynamic treatment regimes when the number of covariates involved in each stage is huge. To reduce the effect of nonregularity on the statistical inference of Q-learning, we removed the treatment effects for those subjects with very small treatment effects by using truncated pseudo-outcome in the first-stage penalized regression. Although the truncation induced some bias during the inference, the asymptotic properties for the parameter estimators as well as the estimated optimal value function were established by adjusting for the bias carefully. The simulation studies and real data analysis showed that our HQ-learning simultaneously estimated the optimal dynamic treatment regimes and selected the important variables very well in the presence of high-dimensional case and nonregularity. While the inference of the HQ-learning without truncation was affected by the nonregularity, in particular, the bias of the estimates became large and the coverage rates of the confidence intervals for the parameters and the optimal value function were low when the degree of nonregularity is very high (e.g., 100%).

We note that the penalized Q-learning of Song et al. (2015) also minimized a penalized objective function in Step 1 of the Q-learning procedure, which aimed to handle the problem of nonregularity by using penalization to shrink some of individual-level treatment effects to zeros. While our HQ-learning used penalization at each stage to shrink some of regression coefficients to zeros to select the important variables that truly contribute to the individual reward in the face of a large number of covariates. It is worthwhile yet beyond the scope of the current article to combine these two kinds of penalizations to simultaneously perform individual selection and variable selection in high-dimensional Q-learning for dynamic treatment regimes.

We use the linear model as the working model for each stage of our HQ-learning. Although the linearity may not hold for Q-function, linear models are also most commonly used working models to help understand the treatment heterogeneity among patients. Such models are particularly useful for high-dimensional settings when variable selection is necessary. Actually, for the dynamic treatment regimes, even when the dimension is low, linear models are widely used in each of stage for many methods including the standard Q-learning, the soft-threshold method proposed in Chakraborty, Murphy, and Strecher (2010). Furthermore, according to our simulation studies, the performance of the proposed method remains good even if the linearity assumption does not hold for the true Q-function. Nevertheless a more flexible model such as semiparametric or nonparametric models will be a better alternative in real data analysis. It is of great interest to investigate Q-learning procedures for dynamic treatment regimes with high-dimensional variables under more flexible models. We shall leave this promising question for future work.

This article only focuses on the two-stage randomized trial, but the HQ-learning can be generalized to the case of multiple stages. Meanwhile, generalizations of Q-functions to deal with data with multiple treatments as well as other type of outcome (ordinal, censored outcomes) are interesting research topics as well.

Supplementary Material

The online supplement contains additional simulation results, and the proofs for the theorems discussed in the article.

Acknowledgment

The authors thank the associate editor and two referees for their constructive comments that led to a significantly improved article.

Funding

This work is supported in part by the National Natural Science Foundation of China (11771072 and 11371083), Foundation from China Scholarship Council (No. 201406625026), National Institute of Health and National Science Foundation (2R01NS073671-05A1, 1R01GM124104-01A1, NCI-P01-CA142538, NSF-DMS-1555244).

References

Altman, T., and Leger, C. (1994), "Cross-validation, the Bootstrap, and Related Methods for Tuning Parameter Selection," Technical Report, The Cornell University Library, 1–23. [1409]

Chakraborty, B., Laber, E., and Zhao, Y. (2013), "Inference for Optimal Dynamic Treatment Regimes using an Adaptive m-out-of-n Bootstrap Scheme," *Biometrics*, 69, 714–723. [1404]

Chakraborty, B., and Moodie, E. E. M. (2013), *Statistical Methods for Dynamic Treatment Regimes*, New York: Springer. [1404]

Chakraborty, B., Murphy, S., and Strecher, V. (2010), "Inference for Non-Regular Parameters in Optimal Dynamic Treatment Regimes," *Statistical Methods in Medical Research*, 19, 317–343. [1404,1405,1412,1415,1416]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [1409]

Fan, J., and Lv, J. (2011), "Non-Concave Penalized Likelihood with NP-Dimensionality," *IEEE Transactions on Information Theory*, 57, 5467–5484. [1405,1407]

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sackeim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F., and Kupfer, D. J. (2003), "Background and Rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) Study," *Psychiatric Clinics of North America*, 26, 457–494. [1413]

Laber, E., Lizotte, D., Qian, M., Pelham, W., and Murphy, S. (2014), "Dynamic Treatment Regimes: Technical Challenges and Applications," *Electronic Journal of Statistics*, 8, 1225–1272. [1404]

Luedtke, A. R., and Van Der Laan, M. J. (2016), "Statistical Inference for the Mean Out- Come under a Possibly Non-Unique Optimal Treatment Strategy," *The Annals of Statistics*, 44, 713–742. [1404]

Lv, J., and Fan, Y. (2009), "A Unified Approach to Model Selection and Sparse Recovery using Regularized Least Squares," *The Annals of Statistics*, 37, 3498–3528. [1406,1407]

Moodie, E., and Richardson, T. (2010), "Estimating Optimal Dynamic Regimes: Correcting Bias under the Null," *Scandinavian Journal of Statistics*, 37, 126–146. [1404]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *Annals of Statistics*, 39, 1180–1210. [1404]

Robins, J. M. (2004), "Optimal Structural Nested Models for Optimal Sequential Decisions," in *Proceedings of the Second Seattle Symposium in Biostatistics*, Springer, pp. 189–326. [1404]

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sackeim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M., Kupfer, D. J., Rosenbaum, J. F., Alpert, J., Stewart, J. W., McGrath, P. J., Biggs, M. M., Shores-Wilson, K., Lebowitz, B. D., Ritz, L., and Niederehe, G. (2004), "Sequenced Treatment Alternatives to Relieve Depression (STAR*D): Rationale and Design," *Controlled Clinical Trials*, 25, 119–142. [1414]

Song, R., Wang, W., Zeng, D., and Kosorok, M. R. (2015), "Penalized Q-Learning for Dynamic Treatment Regimens," *Statistica Sinica*, 25, 901–920. [1405,1416]

Watkins, C. J. (1989), "Learning from Delayed Rewards," Ph.D. dissertation, University of Cambridge, England. [1404]

Zhang, C. (2010), "Nearly Unbiased Variable Selection under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [1406]