Statistica Sinica **25** (2015), 901-920
doi:http://dx.doi.org/10.5705/ss.2012.364

PENALIZED Q-LEARNING FOR DYNAMIC TREATMENT REGIMENS

Rui Song[1], Weiwei Wang[2], Donglin Zeng[3] and Michael R. Kosorok[3]

[1]*North Carolina State University,* [2]*Two Sigma Investment and* [3]*University of North Carolina, Chapel Hill*

Abstract: A dynamic treatment regimen incorporates both accrued information and long-term effects of treatment from specially designed clinical trials. As these trials become more and more popular in conjunction with longitudinal data from clinical studies, the development of statistical inference for optimal dynamic treatment regimens is a high priority. In this paper, we propose a new machine learning framework called penalized Q-learning, under which valid statistical inference is established. We also propose a new statistical procedure: individual selection and corresponding methods for incorporating individual selection within penalized Q-learning. Extensive numerical studies are presented which compare the proposed methods with existing methods, under a variety of scenarios, and demonstrate that the proposed approach is both inferentially and computationally superior. It is illustrated with a depression clinical trial study.

Key words and phrases: Dynamic treatment regimen, individual selection, multi-stage, penalized Q-learning, Q-learning, shrinkage, two-stage procedure.

1. Introduction

Developing effective therapeutic regimens for diseases is one of the essential goals of medical research. Two major design and analysis challenges in this effort are taking accrued information into account in clinical trial designs and effectively incorporating long-term benefits and risks of treatment due to delayed effects. One of the most promising approaches to dealing with these challenges has been recently referred to as dynamic treatment regimens or adaptive treatment strategies (Murphy (2003)), and has been used in a number of settings, such as drug and alcohol dependency studies.

Reinforcement learning, one of the primary tools used in developing dynamic treatment regimens, is a sub-area of machine learning, where the learning behavior is through trial-and-error interactions with a dynamic environment (Kaelbling, M. and Moore (1996)). Because reinforcement learning techniques have been shown to be effective in developing optimal dynamic treatment regimens, the area is attracting increased attention among statistical researchers. As a recent example, a new approach to cancer clinical trials, based on the specific area of reinforcement learning called Q-learning, has been proposed by Zhao, Kosorok and Zeng (2009) and Zhao et al. (2011). Extensive statistical estimating methods have also been proposed for optimal dynamic treatment regimens, including, for example, Chakraborty, Murphy and Strecher (2010), who developed a Q-learning framework based on linear models. Other related literature includes likelihood-based methods (Thall, Millikan and Sung (2000); Thall, Sung and Estey (2002); Thall et al. (2007)) and semiparametric methods (Murphy (2003); Robins (2004); Lunceford, Davidian and Tsiatis (2002); Wahed and Tsiatis (2004, 2006); Moodie, Platt and Kramer (2009)).

In contrast to the substantial body of estimating methods, the development of statistical inference for optimal dynamic treatment regimens is very limited. This sequential, multi-stage decision making problem is at the intersection of machine learning, optimization and statistical inference and is thus quite challenging. As discussed in Robins (2004), and recognized by many other researchers, the challenge arises when the optimal last stage treatment is non-unique for at least some subjects in the population, causing estimation bias and failure of traditional inferential approaches. There have been a number of proposals to correct this. For example, Moodie and Richardson (2010) proposed a method called Zeroing Instead of Plugging In. This is referred to as the hard-threshold estimator by Chakraborty, Murphy and Strecher (2010), who also proposed a soft-threshold estimator and implemented several bootstrap methods. There is, however, a lack of theoretical support for these methods. Moreover, simulations indicate that neither hard-thresholding nor soft-thresholding, in conjunction with their bootstrap implementation, works uniformly well. We are therefore motivated to develop improved, asymptotically valid inference for optimal dynamic treatment regimens.

In this paper, we develop a new reinforcement learning framework for discovering optimal dynamic treatment regimens: penalized Q-learning. The major distinction of penalized Q-learning from traditional Q-learning is in the form of the objective Q-function at each stage. While the new method shares many of the properties of traditional Q-learning, it has some significant advantages. Based on penalized Q-learning, we propose effective inferential procedures for optimal dynamic treatment regimens. In contrast to existing bootstrap approaches, our variance calculations are based on explicit formulae and hence are much less time-consuming. Theoretical studies and extensive empirical evidence support the validity of the proposed methods. Since penalized Q-learning puts a penalty on each individual, it automatically initiates another procedure, individual selection, which selects those individuals without treatment effects from the population. Successful individual selection, i.e., correctly identifying individuals without treatment effects, is the key to improved statistical inference.

While the proposed individual selection procedure shares some similarities with certain commonly used variable selection methods, the approaches differ fundamentally in other ways. These issues will be addressed in greater detail below.

2. Statistical Inference with Q-learning

2.1. Personalized dynamic treatment regimens

Consider data from a sequential multiple assignment randomized trial (SMART), where treatments are randomized at multiple stages (Lavori and Dawson (2000); Murphy (2005)). The longitudinal data on each patient take the form $H = (H_1^T, H_2^T)^T$, where $H_1 = (O_1^T, A_1, R_1)^T$, $H_2 = (O_2^T, A_2, R_2)^T$ are sequences of random variables collected at two stages, $t = 1, 2$. As components of H_t, A_t is the randomly assigned treatment to patients, O_t is the observed patient covariates prior to the treatment assignment and R_t is the clinical outcome, each at stage t. The observed data are treated as n independent and identically distributed copies of H, with the goal of estimating the best treatment decision for different patients using the observed data at each stage. This is equivalent to identifying a sequence of ordered rules, which we call personalized dynamic treatment regimens, $d = (d_1, d_2)^T$, one rule for each stage, mapping from the domain of the patient history, $\mathcal{S}_t$, to the domain of treatment, $\mathcal{A}_t$, where $S_1 = O_1$ and $S_2 = (O_1^T, A_1, R_1, O_2^T)^T$.

Denote the distribution of H by P and the expectations with respect to this distribution by E. Let P^d denote the distribution of H and the expectations with respect to this distribution by E^d, where the dynamic treatment regimen $d(\cdot)$ is used to assign treatments. Take the value function to be $V(d) = E^d(R_1 + R_2)$. Thus, an optimal dynamic treatment regimen, d_0, is a rule that maximizes V. We use upper case letters to denote random variables and lower case letters to denote values of the random variables. In this two-stage setting, if we take $Q_2(s_2, a_2) = E(R_2|S_2 = s_2, A_2 = a_2)$ and $Q_1(s_1, a_1) = E(R_1 + \max_{a_2 \in \mathcal{A}_2} Q_2(S_2, a_2)|S_1 = s_1, A_1 = a_1)$, then the optimal decision rule at time t is $d_t(s_t) = \operatorname{argmax}_{a_t \in \mathcal{A}_t} Q_t(s_t, a_t)$, where Q_t are the Q-functions at time t.

2.2. Q-learning for personalized dynamic treatment regimens

Q-learning is a backward recursive approach commonly used for estimating the optimal personalized dynamic treatment regimens. Following Chakraborty, Murphy and Strecher (2010), let the Q-function for time $t = 1, 2$ be modeled as

$$Q_t(S_t, A_t; \beta_t, \psi_t) = \beta_t^T S_{t(1)} + (\psi_t^T S_{t(2)}) A_t, \tag{2.1}$$

where S_t is the full state information at time t introduced in the previous section and $S_{t(1)}$ and $S_{t(2)}$ are given features as functions of S_t. For example, they can be subsets of S_t selected for the model, and can be identical or different. Moreover, the constant 1 is included in $S_{t(1)}$ and $S_{t(2)}$. The action A_t takes value 1 or -1. The parameters of the Q-function are $\theta_t = (\beta_t^T, \psi_t^T)^T$, where β_t reflects the main effect of current state on outcome, and ψ_t reflects the interaction effect between current state and treatment choice. The true values of these parameters are denoted θ_{t0}, β_{t0}, and ψ_{t0} respectively. We note that the additive formulation of rewards is not restrictive. In fact, we can always define the intermediate rewards to be zeros with the final stage reward to as the final outcome of interest. This does not change the value function we aim to maximize. The linear models studied here are also general if we let state variables in the regression be basis functions of historical variables (for instance, using kernel machine). One can always perform model diagnostics to check the linearity assumption.

Suppose that the observed data consist of (S_{ti}, A_{ti}, R_{ti}) for patients $i = 1, \ldots, n$ and $t = 1, 2$, from a sample of n independent patients. The two-stage empirical version of the Q-learning procedure is summarized as follows.

Step 1. Start with a regular non-shrinkage estimator, based on least squares, for the second stage:

$$\begin{aligned}\tilde{\theta}_2 = (\tilde{\beta}_2^T, \tilde{\psi}_2^T)^T &= \underset{\beta_2,\psi_2}{\operatorname{argmin}} \sum_{i=1}^{n} \{R_{2i} - Q_2(S_{2i}, A_{2i}; \beta_2, \psi_2)\}^2 \\ &= \left(Z_2^T Z_2\right)^{-1} Z_2^T R_2,\end{aligned}$$

where $\tilde{\theta}_2$ is the least squares estimator, Z_2 is the stage-2 design matrix with each row of $(S_{2i(1)}^T, A_{2i} S_{2i(2)}^T)$ and $R_2 = (R_{21}, \ldots, R_{2n})^T$. We use $S_{ti(k)}$ to denote the kth component of S_t for subject i, where $k = 1, 2, t = 1, 2$ and $i = 1, \ldots, n$.

Step 2. Estimate the first-stage individual pseudo-outcome by $\widehat{Y}_1^{\mathrm{HM}} = (\widehat{Y}_{11}^{\mathrm{HM}}, \ldots, \widehat{Y}_{1n}^{\mathrm{HM}})^T$, where

$$\widehat{Y}_{1i}^{\mathrm{HM}} = R_{1i} + \max_{a \in \{-1,1\}} Q_2(S_{2i}, a; \tilde{\theta}_2) = R_{1i} + \tilde{\beta}_2^T S_{2i(1)} + |\tilde{\psi}_2^T S_{2i(2)}|, \qquad (2.2)$$

with ^{HM} as the index for the hard-max estimator.

Step 3. Estimate the first-stage parameters by least squares estimation:

$$\widehat{\theta}_1^{\mathrm{HM}} = \underset{\beta_1,\psi_1}{\operatorname{argmin}} \sum_{i=1}^{n} \{\widehat{Y}_{1i}^{\mathrm{HM}} - Q_1(S_{1i}, A_{1i}; \beta_1, \psi_1)\}^2 = \left(Z_1^T Z_1\right)^{-1} Z_1^T \widehat{Y}_1^{\mathrm{HM}},$$

where Z_1 is the stage-1 design matrix whose ith row is $(S_{1i(1)}^T, A_{1i} S_{1i(2)}^T)$. The corresponding estimator of ψ_1, denoted by $\widehat{\psi}_1^{\mathrm{HM}}$, is referred to as the hard max estimator in Chakraborty, Murphy and Strecher (2010).

2.3. Challenges in statistical inference

When the Q-function takes the form (2.1), the optimal dynamic treatment regimen for patient i is

$$d_i(s_{ti}) = \underset{a_i\in\{-1,1\}}{\operatorname{argmax}}\,(\psi_t^T s_{ti(2)})a_i = \operatorname{sgn}(\psi_t^T s_{ti(2)}),\ t=1,2,\ i=1,\ldots,n,$$

where $\operatorname{sgn}(x)=1$ if $x>0$ and -1 otherwise. We use s_{ti} to denote the observed value of S_t for patient i and $s_{ti(k)}$ denotes the observed value of $S_{t(k)}$ for stage $t=1,2$, component $k=1,2$ and patient i. The parameters ψ_2 are of particular interest for inference on the optimal dynamic treatment regimen, as ψ_2 is the interaction effect of the treatment and covariates.

During the Q-learning procedure, when there is a positive probability that $\psi_{20}^T S_{2(2)}=0$, the first-stage hard max pseudo-outcome $\widehat{Y}_1^{\rm HM}$ is a non-smooth function of $\tilde{\psi}_2$. As a linear function of $\widehat{Y}_1^{\rm HM}$, the hard max estimator $\widehat{\psi}_1^{\rm HM}$ is also a non-smooth function of $\tilde{\psi}_2$. Consequently, the asymptotic distribution of $n^{1/2}(\widehat{\psi}_1^{\rm HM}-\psi_{10})$ is neither normal nor any well-tabulated distributions if $\Pr(\psi_{20}^T S_{2(2)}=0)>0$. In this non-standard case, such tools as Wald-type confidence intervals are no longer valid.

2.4. Review of existing approaches

To overcome the difficulty of inference for ψ_1 in Q-learning, several methods have been proposed, that we briefly review in the two-stage set-up. Since all the methods are also nested in the Q-learning procedure, we update the two-stage version of Q-learning as follows.

Step 1. Estimate the first-stage individual pseudo-outcome by shrinking the second-stage regular estimators via hard-thresholding or soft-thresholding. The hard-threshold pseudo-outcome, $\widehat{Y}_1^{\rm HT}=(\widehat{Y}_{11}^{\rm HT},\ldots,\widehat{Y}_{1n}^{\rm HT})^T$, is

$$\widehat{Y}_{1i}^{\rm HT} = R_{1i} + \tilde{\beta}_2^T S_{2i(1)} + |\tilde{\psi}_2^T S_{2i(2)}| 1\Big\{\frac{n^{1/2}|\tilde{\psi}_2^T S_{2i(2)}|}{(S_{2i(2)}^T\widehat{\Sigma}_2 S_{2i(2)})^{1/2}} > z_{\alpha/2}\Big\}, \tag{2.3}$$

where $\widehat{\Sigma}_2/n$ is the estimated covariance matrix of $\tilde{\psi}_2$, α is a pre-specified significance level and $z_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution. The soft-threshold pseudo-outcome, $\widehat{Y}_1^{\rm ST}=(\widehat{Y}_{11}^{\rm ST},\ldots,\widehat{Y}_{1n}^{\rm ST})^T$, is

$$\widehat{Y}_{1i}^{\rm ST} = R_{1i} + \tilde{\beta}_2^T S_{2i(1)} + |\tilde{\psi}_2^T S_{2i(2)}|\Big(1-\frac{\lambda_i}{|\tilde{\psi}_2^T S_{2i(2)}|}\Big)_+,\ i=1,\ldots,n, \tag{2.4}$$

where $x_+ = xI\{x>0\}$ and λ_i is a tuning parameter.

Step 2. Estimate the first-stage parameters by least squares estimation:

$$\widehat{\theta}_1^{\circ} = (\widehat{\beta}_1^{\circ T}, \widehat{\psi}_1^{\circ T})^T = \underset{\beta_1,\psi_1}{\operatorname{argmin}} \sum_{i=1}^{n} \{\widehat{Y}_{1i}^{\circ} - Q_1(S_{1i}, A_{1i}; \beta_1, \psi_1)\}^2 = \left(Z_1^T Z_1\right)^{-1} Z_1^T \widehat{Y}_1^{\circ},$$

where $\widehat{Y}_1^{\circ}$ is either a hard-threshold or soft-threshold pseudo-outcome, as (2.3) or (2.4). The corresponding estimator of ψ_1, $\widehat{\psi}_1^{\circ}$, can be the hard-threshold estimator $\widehat{\psi}_1^{\mathrm{HT}}$ or the soft-threshold estimator $\widehat{\psi}_1^{\mathrm{ST}}$.

The hard-thresholding and soft-thresholding methods can be viewed as upgraded versions of the hard max methods in terms of reducing the degree of non-differentiability of the absolute value function at zero. The first-stage pseudo-outcome for these three existing methods can be viewed as shrinkage functionals of certain standard estimators. Even if these estimators form shrinkage estimators under certain conditions, they are not optimizers of reasonable objective functions in general. Consequently, even if these estimators can successfully achieve shrinkage, two drawbacks remain that negate their ability to be used for statistical inference for optimal dynamic treatment regimens. First, their bias can be large in finite samples, leading to further bias in the first stage estimator of ψ_1 in regular settings; see point has been demonstrated in the empirical studies of Chakraborty, Murphy and Strecher (2010). More importantly, these shrinkage functional estimators may not possess the oracle property that, with probability tending to one, the set $\mathcal{M}_\star = \{i : |\psi_{20}^T S_{2i(2)}| > 0\}$ can be correctly identified and the resulting estimator performs as well as the estimator that knows the true set $\mathcal{M}_\star$ in advance.

3. Inference Based on Penalized Q-Learning

3.1. Estimation procedure

We focus on the two-stage setting as given in Section 2.2 and use the same notation. As a backward recursive reinforcement learning procedure, our method follows the three steps of the usual Q-learning method, except that it replaces Step 1 of the standard Q-learning procedure with

Step 1_{p}. We minimize the penalized objective function

$$W_2(\theta_2) = \sum_{i=1}^{n} \{R_{2i} - Q_2(S_{2i}, A_{2i}; \beta_2, \psi_2)\}^2 + \sum_{i=1}^{n} p_{\lambda_n}(|\psi_2^T S_{2i(2)}|), \tag{3.1}$$

where $p_{\lambda_n}(\cdot)$ is a pre-specified penalty function and λ_n is a tuning parameter.

Because of this penalized estimation, we call our approach penalized Q-learning. Since the penalty is put on each individual, we also call Step 1_{p} individual selection.

Using individual selection enjoys similar shrinkage advantages as do penalized methods described by Frank and Friedman (1993), Tibshirani (1996), Fan and Li (2001), Candes and Tao (2007), Zou (2006) and Zou and Li (2008). In the first step of the proposed penalized Q-learning approach, penalized estimation allows us simultaneously to estimate the second-stage parameters θ_2 and select individuals whose value functions are not affected by treatments, individuals whose true values of $\psi_2^T S_{2(2)}$ are zero.

Statistical inference in the usual Q-learning is mainly challenged by difficulties in obtaining the correct asymptotic distribution of $n^{1/2}(|\widehat{\psi}_2^T S_{2(2)}|-|\psi_{20}^T S_{2(2)}|)$, where $\widehat{\psi}_2$ is an estimator for ψ_{20}. Via our penalized Q-learning method, we can identify individuals whose $\widehat{\psi}_2^T S_{2(2)} = \psi_{20}^T S_{2(2)}$ takes value zero; moreover, we know that for all individuals, $\widehat{\psi}_2^T S_{2(2)}$ has the same sign as $\psi_{20}^T S_{2(2)}$ asymptotically. Then, $n^{1/2}(|\widehat{\psi}_2^T S_{2(2)}| - |\psi_{20}^T S_{2(2)}|)$ is equivalent to

$$n^{1/2}(\widehat{\psi}_2 - \psi_{20})^T S_{2(2)} \mathrm{sgn}(\psi_{20}^T S_{2(2)}).$$

Hence, correct inference can be obtained following standard arguments, see Section 3.3.

The choice of the penalty function $p_{\lambda_n}(\cdot)$ can be that of popular variable selection methods. Specifically, we require $p_{\lambda_n}(\cdot)$ to possess the following properties.

A1. For non-zero fixed θ, $\lim_{n\to\infty} n^{1/2} p_{\lambda_n}(|\theta|) = 0$, $\lim_{n\to\infty} n^{1/2} p'_{\lambda_n}(|\theta|) = 0$, and $\lim_{n\to\infty} p''{}_{\lambda_n}(|\theta|) = 0$.

A2. For any $M > 0$, $\inf_{|\theta|\leq Mn^{-1/2}} p_{\lambda_n}(|\theta|) \to \infty$, as $n \to \infty$.

Among penalty functions satisfying A1 and A2 are the smoothly clipped absolute deviation penalty (Fan and Li (2001)) and the adaptive lasso penalty (Zou (2006)), where $p_{\lambda_n}(\theta) = \lambda_n \theta / |\theta^{(0)}|^{\phi}$ with $\phi > 0$ and $\theta^{(0)}$ a root-n consistent estimator of θ. To achieve both sparsity and oracle properties, the tuning parameter λ_n in these examples should be taken correspondingly. The adaptive lasso method will be implemented in this paper, where λ_n can be taken as scalars satisfying $n^{1/2}\lambda_n \to 0$ and $n\lambda_n \to \infty$, as $n \to \infty$.

3.2. Implementation

The minimization in Step 1_{p} of the penalized Q-learning procedure has some unique features which distinguish it from the optimization done in the variable selection literature. The component to be shrunk, $\psi_2^T S_{2i(2)}$, is subject-specific, and this component is a linear combination of the parameters.

To deal with these issues, in this section, we propose an algorithm for the minimizing problem of (3.1) based on local quadratic approximation. Following

Fan and Li (2001), we first calculate an initial estimator $\widehat{\psi}_{2(0)}$ via the standard least squares estimation. We then obtain the following local quadratic approximation to the penalty terms in (3.1):

$$p_{\lambda_n}(|\psi_2^T S_{2i(2)}|) \approx p_{\lambda_n}(|\widehat{\psi}_{2(0)}^T S_{2i(2)}|) + \frac{1}{2}\frac{p'_{\lambda_n}(|\widehat{\psi}_{2(0)}^T S_{2i(2)}|)}{|\widehat{\psi}_{2(0)}^T S_{2i(2)}|}\{(\psi_2^T S_{2i(2)})^2 - (\widehat{\psi}_{2(0)}^T S_{2i(2)})^2\}$$

for ψ_2 close to $\widehat{\psi}_{2(0)}$. Thus, (3.1) can be locally approximated up to a constant by

$$\sum_{i=1}^n \{Y_{2i} - Q_2(S_{2i}, A_{2i}; \beta_2, \psi_2)\}^2 + \frac{1}{2}\sum_{i=1}^n \frac{p'_{\lambda_n}(|\widehat{\psi}_{2(0)}^T S_{2i(2)}|)}{|\widehat{\psi}_{2(0)}^T S_{2i(2)}|}(\psi_2^T S_{2i(2)})^2. \qquad (3.2)$$

The updated estimators for ψ_2 and β_2 can be obtained by minimizing the above approximation. When $Q(\cdot)$ is (3.2), this minimization problem has closed form solution

$$\widehat{\psi}_2 = \left[X_{22}^T\{I - X_{21}(X_{21}^T X_{21})^{-1}X_{21}^T + D\}X_{22}\right]^{-1} X_{22}^T\{I - X_{21}(X_{21}^T X_{21})^{-1}X_{21}^T\}Y_2,$$

and

$$\widehat{\beta}_2 = (X_{21}^T X_{21})^{-1}X_{21}^T(Y_2 - X_{22}\widehat{\psi}_2),$$

where X_{22} is a matrix with i-th row equal to $S_{2i(2)}^T A_{2i}$, X_{21} is a matrix with i-th row equal to $S_{2i(1)}^T$, I is the $n \times n$ identity matrix, and D is an $n \times n$ diagonal matrix with $D_{ii} = (1/2)p'_{\lambda_n}(|\widehat{\psi}_{2(0)}^T S_{2i(2)}|)/|\widehat{\psi}_{2(0)}^T S_{2i(2)}|$.

This minimization procedure can be continued until convergence. However, as discussed in Fan and Li (2001), either the one-step or k-step estimator will be as efficient as the fully iterative method as long as the initial estimators are consistent. Since in practice, the local quadratic approximation algorithm shrinks $|\widehat{\psi}_2^T S_{2i(2)}|$ to a very small value instead of exactly zero even if the true value is zero, we set $|\widehat{\psi}_2^T S_{2i(2)}| = 0$ once the value is below a pre-specified tolerance threshold.

The choice of local quadratic approximation is mainly for convenience in solving the penalized least squares estimation in (3.1). If least absolute deviation estimation or some other quantile regression approach is used in place of least squares, then the local linear approximation of the penalty function described in Zou and Li (2008) can be used instead of local quadratic approximation, and the resulting minimization problem can be solved by linear programming.

We use five-fold cross-validation to choose the tuning parameter, where we partition data into five folds, perform the estimation on four folds, and validate the least squares fitting on the other fold. We set $\phi = 2$ as the parameter used in adaptive lasso. We acknowledge the insufficient theory support for using this method. The general guideline for choosing tuning parameters and ϕ is of great research interest but it is beyond the scope of the current paper.

3.3. Asymptotic results

In this section, we establish the asymptotic properties for the parameter estimators in our penalized Q-learning method. We assume that the penalty function $p_{\lambda_n}(x)$ satisfies A1 and A2, and that the following conditions hold.

B1. The support of $S_{2(2)}$ contains a finite number of vectors, say, $v_1, \ldots, v_K$. Moreover, $\psi_{20}^T v_k \neq 0$ for $k \leq K_1$ and $\psi_{20}^T v_k = 0$ for $k > K_1$. Let $n_k = \#|\{i: \ S_{2i(2)} = v_k, \ i = 1, \ldots, n\}|$, where for a set A, $\#|A|$ is defined as its cardinality.

B2. The true value for θ_2, $\theta_{20} = (\psi_{20}^T, \beta_{20}^T)^T$, minimizes

$$\lim_n \sum_{i=1}^{n} n^{-1} \left\{ R_{2i} - Q_2(S_{2i}, A_{2i}; \beta_2, \psi_2) \right\}^2 ,$$

while the true value for θ_1, $\theta_{10} = (\psi_{10}^T, \beta_{10}^T)^T$ minimizes

$$\lim_n n^{-1} \sum_{i=1}^{n} \left\{ R_{1i} + \max_{a \in \{-1,1\}} Q_2(S_{2i}, a; \beta_{20}, \psi_{20}) - Q_1(S_{1i}, A_{1i}; \beta_1, \psi_1) \right\}^2 ,$$

limits existing.

B3. For $t = 1, 2$, with probability one, $Q_t(S_t, A_t; \theta_t)$ is twice-continuously differentiable with respect to θ_t in a neighborhood of θ_{t0}; the Hessian matrices of the limiting functions in B2 are continuous and their values at $\theta_t = \theta_{t0}$, denoted I_{t0}, are nonsingular.

B4. With probability one, $n_k/n = p_k + O_p(n^{-1/2})$ for some constant p_k in $[0, 1]$.

Condition B2 says that θ_{10} and θ_{20} are the target values in the dynamic treatment regimens. Condition B3 can be verified via the design matrix in the two-stage setting: if Q_t takes the form of (3.2), this condition is equivalent to linear independence of $[S_{t(1)}, S_{t(2)} A_t]$ with positive probability. The numerical performance for data from population with a small probability of linear independence is likely to be unstable with small sample sizes.

Theorem 1. *Under conditions* A1−A2 *and* B1−B4*, there exists a local minimizer* $\widehat{\theta}_2$ *of* $W_2(\theta_2)$ *such that* $\|\widehat{\theta}_2 - \theta_{20}\| = O_P(n^{-1/2} + a_n)$*, where* $a_n = \max_{k=1}^{K_1} \left\{ p'_{\lambda_n}(|\psi_{20}^T v_k|) \right\}$.

According to the properties of $p_{\lambda_n}(\cdot)$, $\widehat{\theta}_2$ is $n^{1/2}$-consistent.

Theorem 2. *If* $\mathcal{M}_\star^c = \left\{i : i = 1, \ldots, n, \psi_{20}^T S_{2i(2)} = 0\right\}$, *then under conditions* A1−A2 *and* B1−B4, $\lim_{n\to\infty} \Pr(\widehat{\psi}_2^T S_{2i(2)} = 0,$ *for any* $i \in \mathcal{M}_\star^c) = 1$.

The set $\mathcal{M}_\star^c$ consists of those individuals whose true value functions at the second stage have no effect from treatment. Thus Theorem 2 states that with probability tending to one, we can identify these individuals in $\mathcal{M}_\star^c$. We need the asymptotic distribution of $\widehat{\theta}_2$ in order to make inference.

Theorem 3. *Under conditions* A1−A2 *and* B1−B4, $n^{1/2}(I_{20} + \Sigma)\{\widehat{\theta}_2 - \theta_{20} + (I_{20} + \Sigma)^{-1} b\}$ *converges in distribution to* $N(0, I_{20})$, *where*

$$b = \left(0_p^T, \sum_{k=1}^{K_1} p_k p'_{\lambda_n}(|\psi_{20}^T v_k|) sgn(\psi_{20}^T v_k) v_k\right)^T,$$

and

$$\Sigma = \text{diag}\{0_{p\times p}, \sum_{k=1}^{K_1} p_k p''_{\lambda_n}(|\psi_{20}^T v_k|) v_k v_k^T\}.$$

Theorem 4. *Under conditions* A1−A2 *and* B1−B4, *if* $\bar{S}_{1i} = (S_{1i(1)}^T, S_{1i(2)}^T A_{1i})^T$ *and* $\bar{S}_{2i} = (S_{2i(1)}^T, S_{2i(2)}^T sgn(\psi_{20}^T S_{2i(2)}))^T$, *then* $n^{1/2}(\widehat{\theta}_1 - \theta_{10})$ *converges in distribution to* $I_{10}^{-1}\mathcal{G}$, *where*

$$\mathcal{G} \sim N\Big[0, \text{Cov}\,\Big\{F_1(\theta_{10}) + \lim_n \frac{1}{n}\sum_{i=1}^n \bar{S}_{1i}\bar{S}_{2i}^T F_2(\theta_{20})\Big\}\Big],$$

with

$$F_1(\theta_{10}) = \nabla_{\theta_1} Q_1(S_1, A_1; \theta_{10})\Big\{Y_1 - Q_1(S_1, A_1; \theta_{10})\Big\},$$
$$F_2(\theta_{20}) = (I_{20} + \Sigma)^{-1}\nabla_{\theta_2} Q_2(S_2, A_2; \theta_{20})(R_2 - Q_2(S_2, A_2; \theta_{20})).$$

3.4. Variance estimation

The standard errors can be obtained directly since we are estimating parameters and selecting individuals simultaneously. A sandwich type plug-in estimator can be used as the variance estimator for $\widehat{\theta}_2$:

$$\widehat{\text{Cov}}\,(\widehat{\theta}_2) = (\widehat{I}_{20} + \widehat{\Sigma})^{-1}\widehat{I}_{20}(\widehat{I}_{20} + \widehat{\Sigma})^{-1},$$

where $\widehat{I}_{20} = n^{-1}\sum_{i=1}^n [\nabla^2_{\theta_2\theta_2}\{R_2 - Q_2(S_{2i}, A_{2i}; \theta_2)\}^2]$ is the empirical Hessian matrix and $\widehat{\Sigma} = \text{diag}\{0_{p\times p}, n^{-1}\sum_{i=1}^n p''_{\lambda_n}(|\widehat{\psi}_2^T S_{2i(2)}|) S_{2i(2)} S_{2i(2)}^T\}$. As $\hat{\Sigma}$ converges to zero as n goes to infinity, hence often negligible, we use

$$\widehat{\text{Cov}}\,(\widehat{\theta}_2) = \widehat{I}_{20}^{-1} \tag{3.3}$$

instead, and this performs well in practice. The estimated variance for $\widehat{\theta}_1$ is then

$$\widehat{\text{Cov}}\,(\widehat{\theta}_1) = \widehat{I}_{10}^{-1}\widehat{\text{Cov}}\left\{F_1(\widehat{\theta}_1) + n^{-1}\sum_{i=1}^{n}\bar{S}_{1i}\bar{S}_{2i}^T\widehat{F}_2(\widehat{\theta}_2)\right\}, \tag{3.4}$$

where $\widehat{I}_{10}$ is the empirical estimator for I_{10} and $\widehat{F}_2(\widehat{\theta}_2) = (\widehat{I}_{20}+\widehat{\Sigma})^{-1}\nabla_{\theta_2}Q_2(S_2, A_2; \widehat{\theta}_2)\{R_2-Q_2(S_2, A_2; \widehat{\theta}_2)\}$. These variance estimators have good accuracy for moderate sample sizes; see section 4. This success of direct inference for the estimated parameters makes inference for optimal dynamic treatment regimens possible in the multi-stage setting.

4. Numerical Studies

We apply the proposed method to the simulation study conditions of Chakraborty, Murphy and Strecher (2010). A total of 500 subjects were generated for each dataset. We set $R_1 = 0$ and $(O_1, A_1, O_2, A_2, R_2)$ was collected on each subject, where (O_t, A_t) denotes the covariates and treatment status at stage t ($t = 1, 2$). The binary covariates O_t's and the binary treatments A_t's were generated as follows:

$$\begin{aligned}\Pr(O_1 = 1) = P(O_1 = -1) &= \frac{1}{2},\\ \Pr(A_t = 1) = P(A_t = -1) &= \frac{1}{2},\ t = 1, 2,\\ \Pr(O_2 = 1|O_1, A_1) = 1 - \Pr(O_2 = -1|O_1, A_1) &= \text{expit}(\delta_1 O_1 + \delta_2 A_1),\end{aligned}$$

where $\text{expit}(x) = \exp(x)/\{1 + \exp(x)\}$.

$$R_2 = \gamma_1 + \gamma_2 O_1 + \gamma_3 A_1 + \gamma_4 O_1 A_1 + \gamma_5 A_2 + \gamma_6 O_2 A_2 + \gamma_7 A_1 A_2 + \varepsilon,$$

where $\varepsilon \sim N(0, 1)$. Under this setting, the true Q-functions for time $t = 1, 2$ are

$$\begin{aligned}Q_2(S_2, A_2; \beta_2, \psi_2) =& \beta_{21} + \beta_{22}O_1 + \beta_{23}A_1 + \beta_{24}O_1A_1\\ &+ \psi_{21}A_2 + \psi_{22}O_2A_2 + \psi_{23}A_1A_2, \qquad (4.1)\\ Q_1(S_1, A_1; \beta_1, \psi_1) =& \beta_{11} + \beta_{12}O_1 + \psi_{11}A_1 + \psi_{12}O_1A_1. \qquad (4.2)\end{aligned}$$

The true values β_{11}^0, β_{12}^0, ψ_{11}^0 and ψ_{12}^0 are

$$\begin{aligned}\beta_{11}^0 &= \gamma_1 + q_1|f_1| + q_2|f_2| + (0.5 - q_1)\,|f_3| + (0.5 - q_2)\,|f_4|,\\ \beta_{12}^0 &= \gamma_2 + q_1'|f_1| + q_2'|f_2| - q_1'|f_3| - q_2'|f_4|,\\ \psi_{11}^0 &= \gamma_3 + q_1|f_1| - q_2|f_2| + (0.5 - q_1)|f_3| - (0.5 - q_2)|f_4|,\\ \psi_{12}^0 &= \gamma_4 + q_1'|f_1| - q_2'|f_2| - q_1'|f_3| + q_2'|f_4|,\end{aligned} \tag{4.3}$$

Table 1. Values of $\psi_{20}^T S_{2(2)}$ in the six simulation settings. In Setting 1, $\gamma = (0,0,0,0,0,0,0)^T, \delta_1 = \delta_2 = 0.5$. In Setting 2, $\gamma = (0,0,0,0,0.01,0,0)^T, \delta_1 = \delta_2 = 0.5$. In Setting 3, $\gamma = (0,0,-0.5,0,0.5,0,0.5)^T, \delta_1 = \delta_2 = 0.5$. In Setting 4, $\gamma = (0,0,-0.5,0,0.5,0,0.49)^T, \delta_1 = \delta_2 = 0.5$. In Setting 5, $\gamma = (0,0,-0.5,0,1,0.5,0.5)^T, \delta_1 = 1, \delta_2 = 0$. In Setting 6, $\gamma = (0,0,-0.5,0,0.25,0.5,0.5)^T, \delta_1 = \delta_2 = 0.1$.

	$S_{2(2)} = (1, O_2, A_1)^T$			
Setting	(1,1,1)	(1,1,-1)	(1,-1,1)	(1,-1,-1)
1	0	0	0	0
2	0.01	0.01	0.01	0.01
3	1	0	1	0
4	0.99	0.01	0.99	0.01
5	2	1	1	0
6	1.25	0.25	0.25	-0.75

where $q_1 = 0.25(\text{expit}(\delta_1 + \delta_2) + \text{expit}(-\delta_1 + \delta_2))$, $q_2 = 0.25(\text{expit}(\delta_1 - \delta_2) + \text{expit}(-\delta_1 - \delta_2))$, $q_1' = 0.25(\text{expit}(\delta_1 + \delta_2) - \text{expit}(-\delta_1 + \delta_2))$, $q_2' = 0.25(\text{expit}(\delta_1 - \delta_2) - \text{expit}(-\delta_1 - \delta_2))$, $f_1 = \gamma_5 + \gamma_6 + \gamma_7$, $f_2 = \gamma_5 + \gamma_6 - \gamma_7$, $f_3 = \gamma_5 - \gamma_6 + \gamma_7$, $f_4 = \gamma_5 - \gamma_6 - \gamma_7$. Let $\gamma = (\gamma_1, \ldots, \gamma_7)^T$. We consider six settings, with values of $\psi_{20}^T S_{2(2)}$ and γ for each setting listed in Table 1.

We applied penalized Q-learning with adaptive lasso to these settings. The one-step local quadratic approximation algorithm was used with least squares estimation for the initial values. The tuning parameter λ in the adaptive lasso penalty was chosen by five-fold cross-validation. We took $\phi = 2$ as the parameter in the adaptive least absolute shrinkage and selection operator penalty. When the estimated value $|\widehat{\psi}_2^T S_{2(2)}| < 0.001$, it was set as zero in the stage-1 estimation. The simulation results shown in Tables 4.2 and 4.3 were summarized over 2,000 replications. We included the oracle estimator that knows the true $\mathcal{M}_\star = \{i : |\psi_{20}^T S_{2i(2)}| > 0\}$, the hard max estimator, and the soft-threshold estimator for comparison. Theoretical standard errors for the hard max estimator and the soft-threshold estimator are not available. Results on average length of the confidence intervals are presented in the Web Appendix.

The true values β_{11}^0, β_{12}^0, ψ_{11}^0, and ψ_{12}^0 of stage-1 parameters are linear combinations of four absolute value functions $|f_1|$, $|f_2|$, $|f_3|$, and $|f_4|$ from (4.3). It can be shown that, with a bias of order $o(n^{-1/2})$, the hard max estimators of stage-1 parameters are linear combinations of four corresponding absolute value functions, with stage-2 estimators rather than true values to plug into $|f_1|$, $|f_2|$, $|f_3|$, and $|f_4|$. The performances are greatly affected by the estimation of each of the four absolute value functions, especially as to bias.

Setting 1 is a setting where there is no second-stage treatment effect, as $\psi_{20}^T S_{2(2)} = 0$ for all values of $S_{2(2)}$. The hard-max estimator incurs asymptotic biases for all the four terms $|f_1|$, $|f_2|$, $|f_3|$ and $|f_4|$, all four at about the same order of $\sqrt{2/(n\pi)} = 0.036$, as in this case $\{|f_i|\}_{i=1}^4 = 0$. As shown in (13), the biases in the estimation of $|f_1|$, $|f_2|$, $|f_3|$ and $|f_4|$ are almost completely canceled out in the estimation of β_{12}^0 and ψ_{12}^0, due to the fact that the sum of the coefficients is zero. These biases are largely canceled out in the estimation of ψ_{11}^0, as the sum of the coefficients are close to zero. The hard-max estimator of β_{11}^0 has a significant bias because the coefficients of the four absolute value terms, q_1, q_2, $0.5 - q_1$ and $0.5 - q_2$, are all positive and sum to 1.

The simulation results of Setting 1 are consistent with the theoretical observations in terms of the hard-max estimation. The oracle estimator automatically sets the estimator of ψ_2 to be zero. It has no significant bias, with standard errors accurately predicted by the theory and 95% confidence interval coverage close to the nominal value. The penalized Q-learning based estimator's performance is actually identical to the oracle estimator's. The hard-max estimator has a significant bias and inferior mean square error in $\hat{\beta}_{11}$ while remaining consistent for estimation of the other three stage-1 parameters.

Setting 2 is close to Setting 1, with $\psi_{20}^T S_{2(2)}$ equal to 0.01 for all values of $S_{2(2)}$. The hard-max estimator's 95% confidence interval shows poor coverage for β_{11}^0 and ψ_{11}^0. As the value of $\psi_{20}^T S_{2(2)}$ is nonzero, the oracle estimator reduces to the hard-max in this setting. Although the penalized Q-learning based estimator demonstrates a small bias (-0.009) in the estimation of β_{11}^0, the bias is less than one fifth of that of the oracle estimator and the mean square error is less than half of the oracle estimator. Its standard error estimate remains close to the empirical values.

There is no second-stage treatment effect for a positive proportion of subjects in the population in Setting 3. The value of $\psi_{20}^T S_{2(2)}$ is equal to 0 when $A_1 = -1$ with probability one half. The hard-max estimator incurs bias on the order of $O(n^{-1/2})$ in the estimation of $|f_2|$ and $|f_4|$, but not $|f_1|$ and $|f_3|$, as $f_2 = f_4 = 0$ and $f_1 = f_3 = 1$. The hard-max estimation of β_{12}^0 and ψ_{12}^0 is still approximately unbiased, due to the canceling-out of the coefficients of the absolute value terms. The estimation of β_{11}^0 is biased from the true value at approximately half of the bias of Setting 1, due to the values of the $|f_i|$'s and their coefficients. The estimation of ψ_{11}^0 is also biased, with similar magnitude of bias as in $\hat{\beta}_{11}$ but with reversed sign. The penalized Q-learning based estimator has a bias in $\hat{\psi}_{11}$ but the bias is three times smaller than that of the hard-max estimator. Otherwise, the penalized Q-learning based estimator has almost exactly the same performance as the oracle estimator.

Setting 4 is close to Setting 3. The hard-max estimator's performance is similar to Setting 3, the oracle estimator reduces to the hard-max estimator, and the penalized Q-learning based estimator outperforms the oracle estimator, with both a smaller bias (5 times smaller), and a correctly predicted standard error.

In Setting 5, the term $\psi_{20}^T S_{2(2)}$ is zero when $(O_2, A_1) = (-1, -1)$ with probability one fourth. The hard-max estimator incurs bias in the estimation of $|f_4|$, since $f_4 = 0$. Consequently, all four stage-1 parameter estimators are biased. The bias in $\hat{\beta}_{11}$ is approximately a quarter of that in Setting 1. The bias in $\hat{\beta}_{12}$ is about half of that of $\hat{\beta}_{11}$, with reversed sign. The bias in $\hat{\psi}_{11}$ is about the same magnitude as that of $\hat{\beta}_{11}$, with reversed sign. The bias in $\hat{\psi}_{12}$ is about half of that of $\hat{\beta}_{11}$. In this setting, the oracle estimator has the best performance, with no significant bias and well predicted standard errors. The penalized Q-learning based estimator has a bias in $\hat{\psi}_{11}$ but the bias is much smaller than that of the hard-max estimator. The penalized Q-learning based estimator has no noticeable bias in the other three parameter estimations and the standard error calculations are accurate when compared to Monte-Carlo errors.

Setting 6 is a completely regular setting with values of $\psi_{20}^T S_{2(2)}$ well above zero. The penalized Q-learning based estimator has almost identical performance as the oracle estimator, which is the same as the hard-max estimator. Both estimators are unbiased with accurately calculated standard errors.

Chakraborty, Murphy and Strecher (2010) proposed several bootstrapped confidence intervals for the hard max estimator as well as hard-threshold estimators, with α in Step 2 set to be 0.08 or 0.20, and the soft-threshold estimator. To compare the confidence intervals from the proposed penalized Q-learning based estimator with these bootstrapped methods, we re-ran the simulations with the penalized Q-learning based estimator at sample size $n = 300$ and 1,000 replications. The coverage probabilities from different inferential methods in the six settings are compared in Figure 1, where the results from the hard max, hard threshold and soft threshold methods based on hybrid bootstrapping for variance estimation are shown. Overall, the competing methods cannot provide consistent coverage rates across all six settings while the penalized Q-learning based method always gives coverage probabilities that are not significantly different from the nominal level.

We also applied percentile bootstrapping or double bootstrapping variance estimation in the other competing methods and found that the hard max estimator with the double bootstrapped confidence interval and the soft-threshold estimator with the percentile bootstrapped can give reasonable coverage probabilities. Nonetheless, the penalized Q-learning based estimator has a significant computational advantage. In a comparison run analyzing one dataset with sample size 300, the hard max with double bootstrap confidence interval, at 500 first-stage and 100 second-stage bootstrap iterations needed 316.35 seconds. The soft thresholding with percentile confidence interval at 1,000 bootstrap iterations took 10.98 seconds, and the penalized Q-learning based estimator took only 0.14 second.

Table 2. Summary statistics and empirical coverage probability of 95% nominal percentile confidence intervals for β_{11}^0 and β_{12}^0 using the oracle estimator, the proposed penalized Q-learning based estimator, the hard max estimator and the soft-threshold estimator. "PQ" refers to the penalized Q-learning based estimator, "HM" refers to the hard max estimator, "MSE" refers to the mean squares error, "Std" refers to the average of the 2,000 standard error estimates and "CP" refers to the empirical coverage probability of 95% nominal percentile confidence interval. A "*" indicates a significantly different coverage rate than the nominal.

	β_{11}				β_{12}			
	bias($\times$1000)	MSE($\times$1000)	std($\times$100)	CP	bias($\times$1000)	MSE($\times$1000)	std($\times$100)	CP
Setting 1								
Oracle	1	2.0	4.5	94.7	1	2.0	4.5	94.9
PQ	1	2.0	4.5	94.7	1	2.0	4.5	94.9
HM	61	6.6	−	88.7*	1	2.1	−	95.2
ST	7	2.3	−	96.2*	−1	2.0	−	94.9
Setting 2								
Oracle	52	5.5	6.2	90.0*	1	2.1	4.6	95.3
PQ	−9	2.1	4.5	94.7	1	2.0	4.5	94.8
HM	52	5.5	−	90.8*	1	2.1	−	95.1
ST	−3	2.2	−	94.8	−1	2.0	−	95.1
Setting 3								
Oracle	0	3.0	5.5	94.6	1	2.0	4.5	95.1
PQ	0	3.1	5.5	94.2	2	2.0	4.5	95.2
HM	30	4.3	−	93.0*	2	2.1	−	95.2
ST	−5	3.3	−	93.5*	−1	2.1	−	94.9
Setting 4								
Oracle	26	4.0	6.2	93.8*	2	2.1	4.6	95.2
PQ	−5	3.1	5.5	94.3	2	2.0	4.5	95.1
HM	26	4.0	−	93.4*	2	2.1	−	95.1
ST	−10	3.4	−	93.5*	−1	2.1	−	95.0
Setting 5								
Oracle	0	3.8	6.1	94.7	2	2.7	5.2	95.1
PQ	−2	3.8	6.1	94.4	0	2.7	5.2	95.0
HM	15	4.1	−	94.1	−5	2.7	−	94.3
ST	−8	4.0	−	94.9	−3	2.8	−	94.9
Setting 6								
Oracle	2	3.8	6.2	94.8	−1	2.3	4.8	94.7
PQ	1	3.8	6.2	94.7	−1	2.3	4.8	94.7
HM	2	3.8	−	94.3	−1	2.3	−	95.2
ST	45	6.3	−	87.2*	−1	2.3	−	95.2

We analyzed data from the mental health study described in Fava et al. (2003) using the proposed method. The details are given in the online Supplemental Material.

Table 3. Summary statistics and empirical coverage probability of 95% nominal percentile confidence intervals for ψ_{11}^0 and ψ_{12}^0, using the oracle estimator, the proposed penalized Q-learning based estimator, the hard max estimator and the soft-threshold estimator. The notations are the same as in Table 2.

	ψ_{11}				ψ_{12}			
	bias×1000	MSE×1000	std×100	CP	bias×1000	MSE×1000	std×100	CP
Setting 1								
Oracle	−1	1.9	4.5	95.3	0	2.0	4.5	94.7
PQ	−1	1.9	4.5	95.3	0	2.0	4.5	94.7
HM	0	2.4	–	93.7*	−1	2.1	–	94.5
ST	1	2.3	–	94.8	−1	2.1	–	94.6
Setting 2								
Oracle	0	2.5	5.8	97.3*	−1	2.1	4.6	95.0
PQ	−1	1.9	4.5	95.3	0	2.0	4.5	94.8
HM	0	2.5	–	94.8	−1	2.1	–	94.4
ST	1	2.3	–	94.8	0	2.1	–	94.4
Setting 3								
Oracle	−1	2.9	5.5	95.0	0	2.0	4.5	94.8
PQ	−10	3.1	5.5	94.0	−1	2.0	4.5	94.6
HM	−31	4.2	–	93.8*	−1	2.1	–	94.0
ST	−11	4.0	–	95.0	0	2.1	–	94.0
Setting 4								
Oracle	−26	4.0	6.2	94.9	−1	2.1	4.5	95.0
PQ	−6	3.1	5.5	94.6	0	2.0	4.5	94.6
HM	−26	4.0	–	94.5	−1	2.1	–	94.0
ST	−7	3.4	–	95.1	0	2.1	–	93.9
Setting 5								
Oracle	−1	3.5	6.1	95.8	−1	2.5	4.9	94.3
PQ	−5	3.6	6.1	95.2	0	2.5	4.9	94.3
HM	−16	4.0	–	94.6	6	2.6	–	94.0
ST	−3	4.0	–	95.2	−3	2.6	–	94.2
Setting 6								
Oracle	2	3.9	6.2	95.0	0	2.4	4.8	94.2
PQ	2	4.0	6.2	94.6	0	2.4	4.8	94.2
HM	2	3.9	–	93.7*	0	2.4	–	94.1
ST	1	4.6	–	91.4*	2	2.5	–	94.1

5. Discussion

The proposed penalized Q-learning provides valid inference based on an approximate normal distribution for the estimators of the regression coefficients. Recently while this paper was under review, Chakraborty, Laber and Zhao (2013) proposed m-out-of-n bootstrap as a remedy to the non-regular inference in Q-learning. This modified bootstrap is consistent, and can be used in conjunction with the simple hard-max estimator.

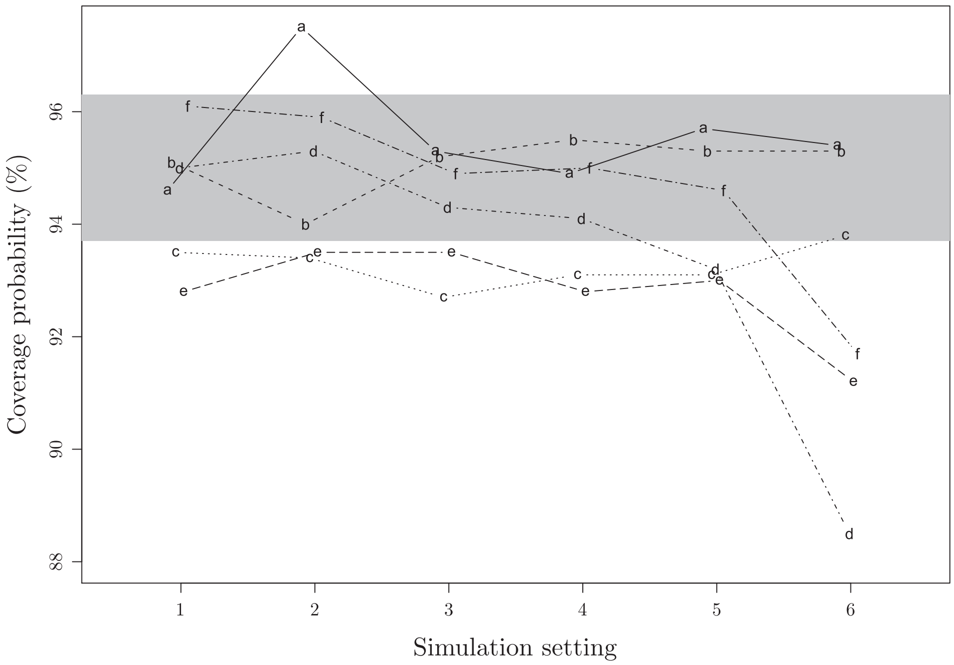

Figure 1. Plot of 95% confidence interval's coverage rates with several inference methods in six simulation settings. The shaded area indicates coverage rates considered to be nonsignificantly different from nominal rate 0.95. Curve "a": the coverage probabilities from the oracle estimator; curve "b": the coverage probabilities from the penalized Q-learning method; curve "c": the coverage probabilities from the hard-max estimator; curve "d": the coverage probabilities from the hard-threshold estimator with $\alpha = 0.08$; curve "e": the coverage probabilities from the hard-threshold estimator with $\alpha = 0.2$; curve "f": the coverage probabilities from the soft-threshold estimator. Curves "c"–"f" all use hybrid bootstrapping.

Under some special cases, the proposed method is the same as variable selection but, in general, it is for individual subject selection instead of individual variable selection. In small samples, our penalization on the linear predictor would possibly impose some constraints as demonstrated in the following example provided by a referee. Suppose that $\psi_2 = (\epsilon, -\epsilon)^T$ and that the following four feature vectors are in the support of $S_{2(2)}$: $(1,1)^T$, $(1,1+a)^T$, $(1+a,1)^T$, $(1+a,1+a)^T$ for $a > 0$. For small value of ϵ and large value of a, say $a = 100/\epsilon$, it is possible that the penalized estimator $\hat{\psi}_2$ shrinks the entire coefficient vector to zero due to the penalty put on $|\psi^T S_{2(2)}|$. Every patient is thus deemed to have no treatment effect. The true treatment effect for subject $(1,1+a)^T$ and $(1+a,1)^T$, however is non-negligible.

This dilemma is due to the finite rank of the covariate space and lack of power to distinguish groups with small effects in small samples. One possibility here is to use a penalty of the form

$$\sum_{i=1}^{n}\{p_{\lambda_1}(|S_{2i(2)}^T\psi_2|)+p_{\lambda_2}(|S_{2i(2)}^T\psi_2-S_{2i(2)}^T\psi_2^\circ|)\},$$

where ψ_2° is a consistent initial estimator of ψ_2. This penalty can shrink estimated $S_{2i(2)}^T\psi_2$ to zero if the truth is zero; otherwise, it will force $S_{2i(2)}^T\psi_2$ to be not far from $S_{2i(2)}^T\psi_2^\circ$.

Although the linear model form of the Q-functions presented here is an important first step, as well as being useful for illustrating the ideas of this paper, this form may not be sufficiently flexible for certain practical settings. Semiparametric models are an alternative in many such settings because such models involve both a parametric component, which is usually easy to interpret, and a nonparametric component that allows greater flexibility. Generalizations of Q-functions to allow diverse data such as ordinal outcome, censored outcome, and semiparametric modeling, are thus future research topics of practical importance.

The theoretical framework is based on discrete covariates, but this is so restrictive. Thus in a two-stage setting with continuous covariates, outside the rare case with ψ_{20} zero, the set $\mathcal{M}_\star^c=\{i:\psi_{20}^TS_{2i(2)}=0\}$ does not have positive probability. That said, we can always discretize continuous covariates, though with a loss of information. Future research to extend our work to continuous covariates would be useful. The framework works for two-level treatments, and the generalization to multilevel treatments is a natural and useful next step.

In many clinical studies, the state space is of high dimension. Then to develop optimal dynamic treatment regimes, it is important to develop simultaneous variable selection and individual selection. Such modern machine learning techniques as support vector regression and random forests can be nested into our penalized Q-learning framework as powerful tools to develop optimal dynamic treatment regimes.

Our method is proposed for the general setting of randomized clinical trials, but it would be useful to generalize the proposed methods to observational studies. Under certain assumptions (for example, no-unobserved confounders), a propensity scores weighted approach can be incorporated into the proposed PQ-learning. We are currently investigating this topic.

References

Candes, E. and Tao, T. (2007). The dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-2404.

Chakraborty, B., Murphy, S. and Strecher, V. (2010). Inference for non-regular parameters in optimal dynamic treatment regimes. *Statist. Meth. Medical Res.* **19**, 317-343.

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.

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

Fava, M., Rush, A., Trivedi, M., Nierenberg, A., Thase, M., Sackeim, H., Quitkin, F., Wisniewski, S., Lavori, P., Rosenbaum, J., Kupfer, D. and STAR D Invest Grp (2003). Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. *Psychiatric Clinics of North America* **26**, 457-494.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics* **35**, 109-148.

Hirano, K. and Porter, J. R. (To appear). Impossibility results for nondifferentiable functionals. *Econometrica*.

Kaelbling, P. L., M., L. and Moore, A. (1996). Reinforcement learning: A survey. *J. Artificial Intelligence Research* **4**, 237-285.

Lavori, P.W. and Dawson, A. (2000). A design for testing clinical strategies: biased adaptive withinsubject randomization. *J. Roy. Statist. Soc. A* **163**, 29-38.

Lunceford, J., Davidian, M. and Tsiatis, A. (2002). Estimation of survival distributions of treatment policies in two-stage randomization designs in clinical trials. *Biometrics* **58**, 48-57.

Moodie, E. E. M., Platt, R. W. and Kramer, M. S. (2009). Estimating response-maximized decision rules with applications to breastfeeding. *J. Amer. Statist. Assoc.* **104**, 155-165.

Moodie, E. E. M. and Richardson, T. S. (2010). Estimating optimal dynamic regimes: correcting bias under the null. *Scand. J. Statist.* **37**, 126-146.

Murphy, S. (2003). Optimal dynamic treatment regimes. *J. Roy. Statist. Soc. Ser. B* **65**, 331-355.

Murphy, S. (2005). An experimental design for the development of adaptive treatment strategies. *Statist. Medicine* **24**, 1455-1481.

Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium on Biostatistics*. Springer.

Thall, P., Millikan, R. and Sung, H. (2000). Evaluating multiple treatment courses in clinical trials. *Statist. Medicine* **19**, 1011-1028.

Thall, P., Sung, H. and Estey, E. (2002). Selecting therapeutic strategies based on efficacy and death in multicourse clinical trials. *J. Amer. Statist. Assoc.* **97**, 29-39.

Thall, P. F., Wooten, L. H., Logothetis, C. J., Millikan, R. E. and Tannir, N. M. (2007). Bayesian and frequentist two-stage treatment strategies based on sequential failure times subject to interval censoring. *Statist. Medicine* **26**, 4687-4702.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Wahed, A. and Tsiatis, A. (2006). Semiparametric efficient estimation of survival distributions in two-stage randomisation designs in clinical trials with censored data. *Biometrika* **93**, 163-177.

Wahed, A. S. and Tsiatis, A. A. (2004). Optimal estimator for the survival distribution and related quantities for treatment policies in two-stage randomised designs in clinical trials. *Biometrics* **60**, 124-133.

Zhao, Y., Kosorok, M. R. and Zeng, D. (2009). Reinforcement learning design for cancer clinical trials. *Statist. Medicine* **28**, 3294-3315.

Zhao, Y., Zeng, D., Socinski, M. and Kosorok, M. (2011). Reinforcement learning strategies for clinical trials in non-small cell lung cancer. *Biometrics* **67**, 1422-1433.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36**, 1509-1533.

Department of Statistics, North Carolina State University, Box 8203, 2311 Stinson Drive, 5120 SAS Hall, Raleigh NC 27695-8203, USA.

E-mail: rsong@ncsu.edu

Two Sigma Investment, 101 Avenue of America, 16th FL, New York, NY 10012, USA.

E-mail: Weiwei.Wang@uth.tmc.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: dzeng@bios.unc.edu

Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

E-mail: kosorok@unc.edu

(Received December 2012; accepted May 2014)