Paper Review: New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes

Ying-Qi Zhao   Donglin Zeng
Eric B. Lader   Michael R. Kosorok

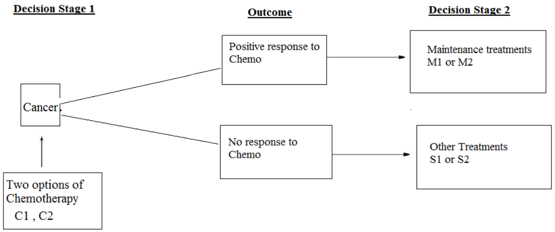
April 13, 2016

Presentation by Hilda Ibriga

Overview

- A Dynamic Treatment Regime (DTR) is a sequence of decision rules, which at each step indicates what treatment to give a patient based on historical data collected from the patient up to that time point.
- At each decision point, **prognostic** variables and **treatment** histories of a patient are used as input for the decision rule, which outputs an individualized treatment recommendation for the patient.

Cancer Treatment



Two-stage Decision Process

- Decision point 1 : Introduce Chemotherapy (C_1 or C_2)
- Decision point 2
  - Maintenance treatment (M_1 or M_2) for patients who respond.
  - Alternative treatment (S_1 or S_2) for patients who do not respond.
- Objective : Maximize survival time

- More realistic and flexible treatment plan than a **once-and-for-all** strategy.
- Accounts for heterogeneity across patients and heterogeneity over time within each patient.
- Allows different aspect of treatment to vary over time according to patient specific information. Example: Treatment types, dosage levels, timing of delivery etc.
- *" The right treatment for the right patient at the right time".*

- **Identify an optimal DTR**: the sequence of decision rules which will maximize expected **long-term outcome** in the patients.
- Outcome can be defined in a variety of ways:
  - For cancer patients, outcome could be time before relapse.
  - For smoking patients, outcome could be quitting.
  - Outcome could also be the reduced side effects of a treatment.

Data Collection Procedure (SMART)

The algorithms considered assume the data is collected from Sequential Multiple Assignment Clinical Trials (SMART).

Features of SMART data:

- Randomizations are done at multiple decision time points.
- Choice of treatments to randomize at each decision point can depend on success or failure of previously randomized treatments.
- Treatments are independent of potential future outcomes.

Mathematical framework

Consider a **multistage** decision problem where decisions are made in T stages.

Let

- A_j be the treatment assigned at the *jth* stage with domain $\mathcal{A}_j = \{-1, 1\}$.
- X_j be the observation after treatment A_{j-1} but prior to the *jth* stage.
- R_j the observed outcome/reward in the patient following *jth* treatment. Each R_j is assumed to be bounded and positive.

Notice that $\sum_{j=1}^{T} R_j$ is the cumulative reward up to stage T.

A **DTR** is a sequence of decision rules $d = (d_1, d_2, ..., d_T)$, with

$$d_j : \mathcal{O}_j \mapsto \mathcal{A}_j,$$

where $\mathcal{O}_j$ is the space of history information $H_j = (X_1, A_1, R_1 ..., A_{j-1}, X_j)$.

Mathematical framework

The **Value function** V^d is defined to be:

$$V^d = E_d \left[\sum_{j=1}^{T} R_j \right].$$

- The expectation is with respect to the measure P_d generated by the random variables $(X_1, A_1, R_1..., A_T, X_T + 1, R_T)$ under $A_j = d_j(H_j)$.
- V^d can be interpreted as the expected long-term benefit if the population were to follow the regime d.

Let P be the measure generated by $(X_1, A_1, R_1..., A_T, X_{T+1}, R_T)$ and E the expectation with respect to that measure. Under some assumptions which are met when the data is collected in a SMART, P dominates P_d and,

$$\frac{dP_d}{dP} = \frac{\prod_{j=1}^{T} I(A_j = d_j(H_j))}{\prod_{j=1}^{T} \pi_j(A_j, H_j)},$$

where $I(.)$ is the indicator function.
The Value function can therefore be written as

$$V^d = E\left[\frac{\left(\sum_{j=1}^{T} R_j\right) \prod_{j=1}^{T} I(A_j = d_j(H_j))}{\prod_{j=1}^{T} \pi_j(A_j, H_j)}\right] \tag{1}$$

Interpretation: for $T = 1$, the value function of assigning treatment $A_1 = 1$ to all patients is a *weighted average* of all outcomes R_j among those patients that received $A_1 = 1$ with weights $\pi_1(A_1, H_1)^{-1}$.

Mathematical framework

The **optimal value function** is defined as

$$V^* = sup_{d \in \mathcal{D}} V^d,$$

where $\mathcal{D}$ consists of all possible treatment regimes.
The goal is to estimate the **optimal** d^* from the data.

$$d^* = argmax_{d \in \mathcal{D}} V_d$$

Classical Approach to DTR

Q-learning

Q-learning is a machine learning method for estimating the optimal DTR. It involves two steps:

- **Step 1.** Estimate the Q-function which in the case of T=1 is defined as :

$$Q(x, a) = E[R|X = x, A = x]$$

  by regressing R on (X, A)
- **Step 2.** Maximize the estimated Q-function in order to infer the optimal DTR.

Limitations:

- The estimated Q-function often poorly fits the data in the case of high-dimensional covariate space.
- If the regression model is misspecified, Q-learning could generate an inconsistent estimator of the optimal DTR

Reformulation as weighted classification problem

Advantages:

- Takes advantage of existing machine learning algorithm in order to estimate DTR.
- Directly estimates the optimal regime unlike other regression based methods such as Q-learning.
- Implementation involves easy-to-use algorithms similar to Support Vector Machine.

Reformulation as a Weighted Classification Problem

Let $T = 1$.

Identifying the optimal treatment regime d^* is equivalent to finding d^* which minimizes the following weighted classification error

$$E\left[\frac{RI(A \neq d(X))}{\pi(A, X)}\right] \tag{2}$$

This is similar to minimizing a non convex and discontinuous 0-1 loss function.

One approach is to use a convex surrogate loss function instead of the 0-1 loss function to obtain an empirical analog to (2).

Mathematical framework

Based on data on n subjects, the empirical analog to (2) using the hinge loss function as surrogate function is

$$n^{-1}\sum_{i=1}^{n}\frac{R_i}{\pi(A_i,X_i)}\phi(A_i f(X_i))+\lambda_n\|f\|^2,$$

where

- $f(x)$ is the decision function
- $d(x)=sign(f(x))$
- The hinge loss function is $\phi(v)=\max(1-v,0)$
- λ_n is the tuning parameter controlling the severity of the penalty
- $\|f\|$ is the norm in a reproducing kernel Hilbert space (RKHS).

Example: $\|f\|$ can be the Euclidean norm of β if f(x) is linear $f(x)=\langle\beta,x\rangle+\beta_0$

New Approach 1

Backward Outcome Weighted Learning (BOWL)

- **Backward Outcome Weighted Learning** known as **BOWL** is a statistical learning method for finding the optimal DTR.
- It uses a backward recursive procedure which estimates the optimal decision rule at future stage first, and then optimal decision rule at the current stage.
- At each stage, analysis is restricted to the subjects who have followed the estimated optimal decision rules thereafter.

Let $(X_{i1}, A_{i1}, R_{i1} ..., A_{iT}, X_{i,T+1}, R_{iT})$, $i = 1, ..., n$ be n *iid* patients trajectories from a SMART. Suppose we already know the optimal regimes at stages $t+1, ..., T$. The optimal decision rule at stage t, d^* is a map from $\mathcal{O}_t$ to $-1, 1$ which minimizes

$$E\left[\frac{\left(\sum_{j=t}^{T} R_j\right)\prod_{j=t+1}^{T} I(A_j = d_j^*(H_j))}{\prod_{j=t+1}^{T} \pi_j(A_j, H_j)} I(A_t \neq d_t(H_t))|H_t = h_t\right] \qquad (3)$$

Hinge loss function as a convex surrogate for the 0-1 loss is used to get the following counterpart of (3).

$$n^{-1}\sum_{i=1}^{n}\frac{\left(\sum_{j=t}^{T}R_{ij}\right)\prod_{j=t+1}^{T}I(A_{ij}=d_j^*(H_{ij}))}{\prod_{j=t+1}^{T}\pi_j(A_{ij},H_{ij})}\phi(A_{it}f_t(H_{it}))+\lambda_{t,n}\|f_t\|^2 \quad (4)$$

and $d(h_t)=sign(f_t(h_t))$
This is an empirical weighted average of the loss function ϕ with weights $(\sum_{j=t}^{T}R_j)\prod_{j=t+1}^{T}I(A_j=d_j^*(H_j))/\prod_{j=t+1}^{T}\pi_j(A_j,H_j)$ for each individual.

Step 1. Minimize

$$n^{-1} \sum_{i=1}^{n} \frac{R_{iT} \phi(A_{iT} quad f_t(H_{iT}))}{\pi_j(A_{iT}, H_{iT})} + \lambda_{T,n} \|f_T\|^2$$

with respect to f_T. Then let $\hat{d}_T(h_T) = sign(\hat{f}_T(h_T))$
Step 2. For $t = T-1, T-2, ..., 1$ minimize

$$n^{-1} \sum_{i=1}^{n} \frac{\left(\sum_{j=t}^{T} R_{ij}\right) \prod_{j=t+1}^{T} I(A_{ij} = \hat{d}_j(H_{ij}))}{\prod_{j=t+1}^{T} \pi_j(A_{ij}, H_{ij})} \phi(A_{it} f_t(H_{it})) + \lambda_{t,n} \|f_t\|^2$$

- The minimization at each step has a similar dual objective function to the usual **SVM**, and can be implemented via quadratic programming

Backward Outcome Weighted Learning

- The number of subjects used to learn the optimal decision rules decreases geometrically as t decreases.
- IOWL is an iterative version of BOWL which eventually uses the entire sample of patients to learn the optimal decision rule.

Step 1. Estimate the optimal DTR $\tilde{d} = (\tilde{d}_1)(\tilde{d}_2)$ using BOWL. The corresponding decision functions are $(\tilde{f}_1, \tilde{f}_2)$. Set $\tilde{d}_1^{new} = \tilde{d}_1$.

Step 2. Given $\tilde{d}_1^{new}$, find an updated optimal stage 2 treatment decision by minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{R_{i2}I(A_{i1}=\tilde{d}_1^{new}(H_{i1}))}{\pi_2(A_{i2},H_{i2})}\phi(A_{i2}f_2(H_{i2}))+\lambda_{2,n}\|f_2\|^2$$

to obtain $\tilde{f}_2$. Set $\tilde{d}_2^{new} = \mathrm{sign}(\tilde{f}_2)$.

Step 3. Given $\tilde{d}_2^{new}$, find an updated optimal stage 1 treatment decision by minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{(R_{i1}+R_{i2})I(A_{i2}=\tilde{d}_2^{new}(H_{i2}))}{\prod_{j=1}^{2}\pi_j(A_{ij},H_{ij})}\phi(A_{i1}f_1(H_{i1}))+\lambda_{1,n}\|f_2\|^2$$

to obtain $\tilde{f}_1$. Set $\tilde{d}_1^{new} = \mathrm{sign}(\tilde{f}_1)$.

Step 4. Iterate between Steps 2 and 3 until the value function V^{d^*} does not increase significantly.

New Approach 2

Simultanous Outcome Weighted Learning (SOLW)

- The SOWL algorithm determines the optimal regimes at all stages simultaneously instead of sequentially using a classification method.
- SOWL aims at directly optimizing the empirical counterpart of (1) in one step.
- A concave surrogate function is used instead of the product of indicators.

  SOWL optimal regime estimator maximizes

$$n^{-1} \sum_{i=1}^{n} \left[\frac{\left(\sum_{j=1}^{2} R_{ij} \right) \psi(A_{i1} f_1(H_{i1}), A_{i2} f_2(H_{i2}))}{\prod_{j=1}^{2} \pi_j(A_{ij}, H_{ij})} \right] - \lambda_n(\|f_1\|^2 + \|f_2\|^2) \tag{5}$$

where, $\psi(Z_1, Z_2) = min(Z_1 - 1, Z_2 - 1, 0) + 1$

Simultanous Outcome Weighted Learning

Concave surrogate function

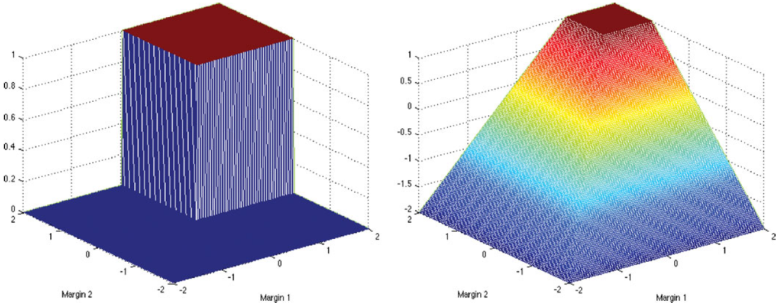


Figure: **Left**: Non smooth indicator function $I(Z_1 > 0, Z_2 > 0)$; **Right**: Smooth concave surrogate $min(Z_1 - 1, Z_2 - 1, 0) + 1$

If the decision functions f_j, $j = 1, ..., T$ are restricted to linear functions of the form $f_j(H_j) = \langle \beta_j, H_j \rangle + \beta_{0j}$ for $j = 1, 2$.
Then the norm of f_1 and f_2 in (5) are the Euclidean norms of β_1 and β_2 respectively.
The optimization problem can be rewritten as:

$$max \quad \gamma \sum_{i=1}^{n} W_i \xi_i - \|\beta_1\|^2 - \|\beta_2\|^2$$

subject to, $\quad \xi_i \leq 0$, $\xi_i \leq A_{i1}(\langle \beta_1, H_{i1} \rangle + \beta_{01}) - 1$,
$\xi_i \leq A_{i2}(\langle \beta_2, H_{i2} \rangle + \beta_{02}) - 1$,
where γ is a constant that depends on λ_n.
This is also a quadratic programming problem with quadratic objective function and linear constraints.

Simultanous Outcome Weighted Learning
SOWL

The dual problem is given by

$$\max_{\alpha_1,\alpha_2} \quad \sum_{i=1}^{n}(\alpha_{i1}+\alpha_{i2}) - \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{n}\sum_{j=1}^{2}\alpha_{ij}\alpha_{lj}A_{ij}A_{lj}\langle H_{ij}, H_{lj}\rangle$$

subject to $quad \alpha_{i1},\, \alpha_{i2} \geq 0, \quad \sum_{i=1}^{n}\alpha_{i1}A_{i1} = 0, \quad \sum_{i=1}^{n}\alpha_{i2}A_{i2} = 0,$
$\alpha_{i1}+\alpha_{i2} \leq \gamma W_i \quad$ for $i = 1, ..., n$.
Remark: Non-linear decision functions can be used by selecting a nonlinear kernel function and the associated RKHS. In this case, $\langle H_{ij}, H_{lj}\rangle$ is replaced by the inner product in the RKHS.

To generalize SOWL to a T-stage decision problem, with $T > 2$, use the surrogate reward function

$$\phi(Z_1, ..., Z_T) = \min(Z_1 - 1, ..., Z_T - 1, 0) + 1$$

and an objective function analogous to the one in (5) when T=2.

Properties of BOWL and SOWL

- Fisher Consistency
- Asymptotic Consistency
- Risk bound (bound on both estimation error and approximation error)

Fisher Consistency

Fisher consistency states that the population optimizer in BOWL and SOWL is the optimal DTR.

Theorem (BOWL)

Assume $(\tilde{f}_1, .., \tilde{f}_T)$ *is a sequence of decision functions obtained by taking the supremum over* $\mathcal{F}_1 \times \mathcal{F}_2 \times ... \times \mathcal{F}_T$ *of*

$$E\left[\frac{\left(\sum_{j=1}^{T} R_j\right)\prod_{j=t+1}^{T} \mathrm{I}(A_j = sign(\tilde{f}_j(H_j)))}{\prod_{j=t}^{T} \pi_j(A_j, H_j)}\phi(A_t f_t(H_t))\right]$$

backward through time for $t = T, T-1, ..., 1$,
then

$$d_j^*(h_j) = sign(\tilde{f}_j(h_j))$$

for all $j = 1, ..., T$

Fisher Consistency

Theorem (SOWL)

If $(\tilde{f}_1, .., \tilde{f}_T) \in \mathcal{F}_1 \times \mathcal{F}_2 \times ... \times \mathcal{F}_T$ *maximizes,*

$$V_\psi(f_1, ..., f_T) = E\left[\frac{\left(\sum_{j=1}^T R_j\right)\psi(A_1 f_1(H_1), ..., A_T f_T(H_T))}{\prod_{j=1}^T \pi_j(A_j, H_j)}\right]$$

then for $h_j \in \mathcal{O}_j$,

$$d_j^*(h_j) = sign(\tilde{f}_j(h_j))$$

for $j = 1, ..., T$.

Relationship between Excess Values

The theorem below shows that the difference between the value function for any decision rules $(f_1, ..., f_T)$ and the optimal value function $(f_1^*, ..., f_T^*)$ with 0-1 reward function is no larger than under the surrogate reward function ψ times a constant.

Theorem (SOWL Excess Values)

$$V(f_1^*, ..., f_T^*) - V(f_1, ..., f_T) \leq (1 + (T-1)c_0^{-1})[V_\psi(f_1^*, ..., f_T^*) - V_\psi(f_1, ..., f_T)]$$

where $(f_1^*, ..., f_T^*)$ *is the optima over* $\mathcal{F}_1 \times \mathcal{F}_2 \times ... \times \mathcal{F}_T$

- This guarantees that if the V_ψ value of a given decision rule is fairly close to V_ψ^*, then the decision rule is also close to the optimal value under the 0-1 loss function.

Theorem

Assume that at stage t , $t = 1, ..., T$, the sequence $\lambda_{j,n}$ satisfies $\lambda_{j,n} \to 0$ and $n\lambda_{j,n} \to \infty$ for $j = 1, ..., T$. Moreover, assume $\hat{f}_j$ is obtained within an RKHS $\mathcal{H}_{k_j}$ associated with a kernel function k_j and that f_j^* belongs to the closure of $\limsup_n \mathcal{H}_{k_j}$ where $d_j^* = sign(f_j^*)$ and $\mathcal{H}_{k_j}$ may depend on n. Then for all distributions P,

$$\lim_{n\to\infty} V_t(\hat{f}_t, ..., \hat{f}_T) = V_t^* \text{ in probability.}$$

Theorem

Assume that the sequence λ_n satisfies $\lambda_n \to 0$ and $n\lambda_n \to \infty$. Moreover, assume $(\hat{f}_1, ..., \hat{f}_T)$ is obtained my maximizing (5) within $\mathcal{H}_{k_j} \times ... \times \mathcal{H}_{k_T}$ and that $(f_1^*, ..., f_T^*)$ belongs to the closure of $\limsup_n \mathcal{H}_{k_j} \times ... \times \mathcal{H}_{k_T}$ where $\mathcal{H}_{k_j}, ..., \mathcal{H}_{k_T}$ are associated with kernel functions $k_1, ..., k_T$, respectively and may depend on n. Then for all distributions P,

$$\lim_{n\to\infty} V_t(\tilde{f}_t, ..., \tilde{f}_T) = V_t^* \text{ in probability.}$$

Risk Bound

Theorem (BOWL)

Let the distribution of (H_j, A_j, R_j), $j = 1, \ldots, T$ satisfy some regularity conditions, with noise exponent $q_j > 0$. Then for any $\delta > 0$, $0 < \nu \leq 2$, there exists a constant K_j depending on ν, δ, p_j and π_j, such that for all $\tau \geq 1$, $\pi_j(a_j, h_j) > c_0$ and $\sigma_{j,n} = \lambda_{j,n}^{-1/(q_j+1)p_j}$, $j \geq t$,

$$P\left(V_t(\hat{f}_t, \ldots, \hat{f}_T) \geq V_t^* - \sum_{j=t}^{T}(3^{-1}c_0)^{t-j}\epsilon_j\right) \geq 1 - \sum_{j=t}^{T} 2^{j-t}e^{-\tau} \qquad (6)$$

$$\epsilon_j = K_j\left[\lambda_{j,n_j}^{-\frac{2}{2+\nu}+\frac{(2-\nu)(1+\delta)}{(2+\nu)(1+q_j)}} n_j^{-\frac{2}{2+\nu}} + \frac{\tau}{n_j\lambda_{j,n_j}} + \lambda_{j,n_j}^{\frac{q_j}{q_j+1}}\right] \qquad (7)$$

where

- *n_j is the available sample size at stage j.*
- *δ is a free parameter.*
- *q_j is the geometric noise condition which describes the behavior of the data near the true decision boundary at each stage.*
- *ν measures the order of complexity for the associated RKHS.*

Risk Bound

Theorem (SOWL)

Let the distribution of (H_j, A_j, R_j), $j = 1, ..., T$ *satisfy some regularity conditions, with noise exponent* $q_j > 0$. *Then for any* $\delta > 0$, $0 < \nu \leq 2$, *there exists a constant* K *depending on* ν, δ, p_j *and* π_j, *such that for all* $\tau \geq 1$, $\pi_j(a_j, h_j) > c_0$ *and* $\sigma_{j,n} = \lambda_{j,n}^{-1/(q_j+1)p_j}$,

$$P\left(V(\hat{f}_t, ..., \hat{f}_T) \geq V^* - \epsilon\right) \geq 1 - e^{-\tau} \tag{8}$$

$$\epsilon = K\left[\lambda_n^{-\frac{2}{2+\nu}}\left(\sum_{j=1}^{T}\lambda_n^{\frac{(2-\nu)(1+\delta)}{2+2q_j}}\right)^{frac22+\nu} n^{-\frac{2}{2+\nu}} + \frac{\tau}{n\lambda_n} + \sum_{j=1}^{T}\lambda_n^{\frac{q_j}{q_j+1}}\right] \tag{9}$$

Convergence Rate

Under the following assumptions in (7) and (9)

- $q_j = q$ for $j = 1, \ldots, T$,
- $\lambda_{j,n} = n_j^{-\frac{2(1+q)}{(4+\nu)q+2+(2-\nu)(1=\delta)}}$

The optimal rate for the value of the estimated DTRs using both BOWL and SOWL is,

$$O_p(n_1^{-\frac{2q}{(4+\nu)q+2+(2-\nu)(1+\delta)}}) \tag{10}$$

Example: if there is no data near the true decision boundary across all stages, then $q = \infty$ and the rate is approximately $n_1^{2+\nu}$

Consider the two-stage process with

- Treatments: A_1 and $A_2 \sim unif\{1, -1\}$
- Covariates: $X_1 = (X_{1,1}, ..., X_{1,50})$ with $X_{1,j} \sim N(0, 1)$.
- Outcomes: $R_1 \sim N(0.5X_1, 3A_1, 1)$ and $R_2 \sim N(((X_{1,1}^2 + X_{1,2}^2 - 0.2)(0.5 - X_{1,1}^2 - X_{1,2}^2) + R_1)A_2, 1)$

The covariates are the same across all stages and there is a nonlinear relationship between the covariates and stage 2 treatment A_2.

In order to determine the optimal DTRs for the simulated data,

- BOWL, IOWL and SOWL were applied using a linear kernel $f_j = \langle \beta_j, H_j \rangle + \beta_{0j}$ for $j = 1, 2$
- The weighted SVM procedure was implemented using LIBSVM.
- A five fold cross-validation was used in order to choose the tuning parameters $\lambda_{t,n}$ in each stage.
- The Q-learning algorithm was carried using the following linear model $Q_j(H_j, A_j; \alpha_j, \gamma_j) = \alpha_j H_j + \gamma_j H_j A_j, \quad j = 1, \ldots T.$

Simulation Study 1 results

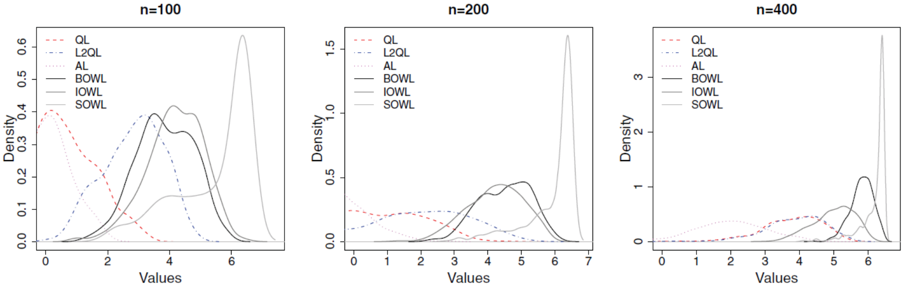


Figure: Smooth histograms of the optimal values of estimated DTRs for Study 1. The optimal value $V^* = 6.695$. DTRs were constructed using the different methods based on training sets of size n replicated 500 times. A validation dataset of size $n = 10,000$ was used. For each training set the values were computed by averaging the 10,000 subjects' outcomes. The histograms above represent the empirical distribution of the 500 values.

Simulation Study 2

Time varying covariates with non-linear stage 2 model

Consider a two stage process with

- Treatments: A_1 and $A_2 \sim unif\{1, -1\}$
- Covariates: $X_1 = (X_{1,1}, ..., X_{1,50})$ with $X_{1,j} \sim N(0, 1)$.
- Outcomes: $R_1 \sim N((1 + X_{1,3})A_1, 1)$ and
  $R_2 \sim N((0.5 + R_1 + 0.5A_1 + 0.5X_{2,1} - 0.5X_{2,2})A_2, 1)$

with $X_{2,1} \sim I\{N(-1.25X_{1,1}A_1, 1) > 0\}$ and
$X_{2,2} \sim I\{N(-1.75X_{1,2}A_1, 1) > 0\}$

Simulation Study 2 results

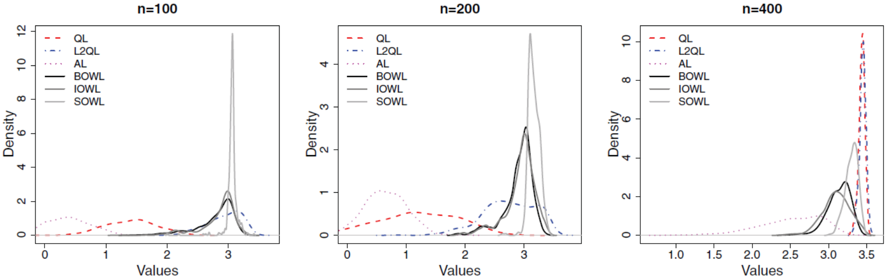


Figure: Smoothed histograms of the values of estimated DTRs for study 2. The optimal value $V^* = 3.667$

Simulation

Mean and Standard Errors

Table 1. Mean values of the estimated DTR for Scenarios 1-2

	n	Q-learning	$L_2 Q$-learning	A-learning	BOWL	IOWL	SOWL
Scenario 1	100	0.692(0.972)	2.831(0.972)	−0.298(0.984)	3.849(0.918)	4.166(0.921)	5.428(1.234)
	200	0.583(1.476)	1.928(1.533)	−0.650(0.991)	4.502(0.768)	4.210(0.840)	5.933(0.755)
	400	3.766(0.896)	3.859(0.897)	1.973(1.072)	5.811(0.331)	4.996(0.602)	6.189(0.388)
Scenario 2	100	1.462(0.361)	2.857(0.248)	0.369(0.318)	2.709(0.340)	2.777(0.314)	3.026(0.141)
	200	1.122(0.679)	2.650(0.547)	0.631(0.322)	2.847(0.269)	2.871(0.278)	3.119(0.084)
	400	3.435(0.041)	3.449(0.043)	2.549(0.394)	3.105(0.131)	3.049(0.197)	3.212(0.089)

The data is from a two-stage randomized trial of the effectiveness of a web-based smoking intervention.

- **Stage 1**-(Project Quit): Find the best intervention out of two $A_1 \in \{1, -1\}$ which helps adult smokers quit smoking.
- **Stage 2**-(Forever Free) : After stage 1 is completed, find the best intervention out of two $A_2 \in \{1, -1\}$ to help those who quit in stage 1 stay quit and those who failed in stage 1 to quit.

- Sample size: 479 patients went through both stages.
- Covariates:
  - Stage 1: 8 covariates $(X_{1,1}, ..., X_{1,8})$ Including Age, Gender, Education, Race, initial motivation to quit etc.
  - Stage 2: all covariates from stage 1 plus two additional covariates measured after stage 1 was completed.
- **Outcome 1**: $R_{Q1}(1 = quit, 0 = no \quad quit)$ and $R_{Q2}(1 = quit, 0 = no \quad quit)$
- **Outcome 2**: R_{S1}(1 = satisfied, 0=otherwise) and R_{S2}(1 = satisfied, 0=otherwise)
- Model: $H_1 = (1, X_1)$ and $H_2 = (H_1, H_1A_1, X_{2,1}, X_{2,2}, R_.1)$
  $\hat{\pi}_j(a_j, H_j) = \sum_j I(A_j = a_j)/n_j$

- BOWL, IOWL and SOWL were applied using a linear kernel $f_j = \langle \beta_j, H_j \rangle + \beta_{0j}$ for $j = 1, 2$.
- Cross validation was implemented using training and validation sets of equal size with 100 replications.
- For Q-learning $Q_j(H_j, A_j; \alpha_j, \gamma_j) = \alpha_j H_j + \gamma_j H_j A_j, \quad j = 1, \ldots T$.

Data Analysis

Mean and Standard error

Table 2. Mean (s.e.) cross-validated values using different methods

	Mean (s.e.) cross-validated values					
Outcome	BOWL	IOWL	SOWL	Q-learning	$L_2 Q$-learning	A-learning
R_Q	0.747 (0.099)	0.768 (0.101)	0.751 (0.073)	0.692 (0.089)	0.696 (0.093)	0.709 (0.090)
R_S	1.262 (0.093)	1.288 (0.114)	1.254 (0.091)	1.216 (0.087)	1.231 (0.094)	1.183 (0.084)

Possible Extension

- Developing tools for statistical inference for DTRs
- Developing methods for estimating DTRs in the case of high dimensional predictor spaces.
- Estimating required sample size for multidecision problems.
- Determining DTR using purely observational data.
- Methods for dealing with missing data (non-compliance)
- Analysis on right censored data

Further Reading I

- Murphy SA, Lynch KG, Oslin D, McKay JR, Ten Have T. Developing adaptive treatment strategies in substance abuse research. Drug Alcohol Depend. 2007a; 88S:S24S30.
- Nahum-Shani I, Qian M, Almirall D, Pelham WE, Gnagy B, Fabiano G, Waxmonsky J, Yu J, Murphy SA. Q-Learning: A data analysis method for constructing adaptive interventions.
- Zhao YQ, Zeng D, Laber EB, Kosorok MR. Journal of the American Statistical Association. 2015/01/01 00:00; 110(510)583-598
- Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Amer. Statist. Assoc. 2012; 107:11061118.
- Schulte PJ, Tsiatis AA, Laber EB, Davidian M. Q- and A-learning Methods for Estimating Optimal Dynamic Treatment Regimes. Statistical science: a review journal of the Institute of Mathematical Statistics. 2014;29(4):640-661. doi:10.1214/13-STS450.