C-learning: Estimating Optimal Dynamic Treatment Regimes from a Classification Perspective

Min Zhang

Department of Biostatistics

University of Michigan, Ann Arbor

Joint work with Baqun Zhang

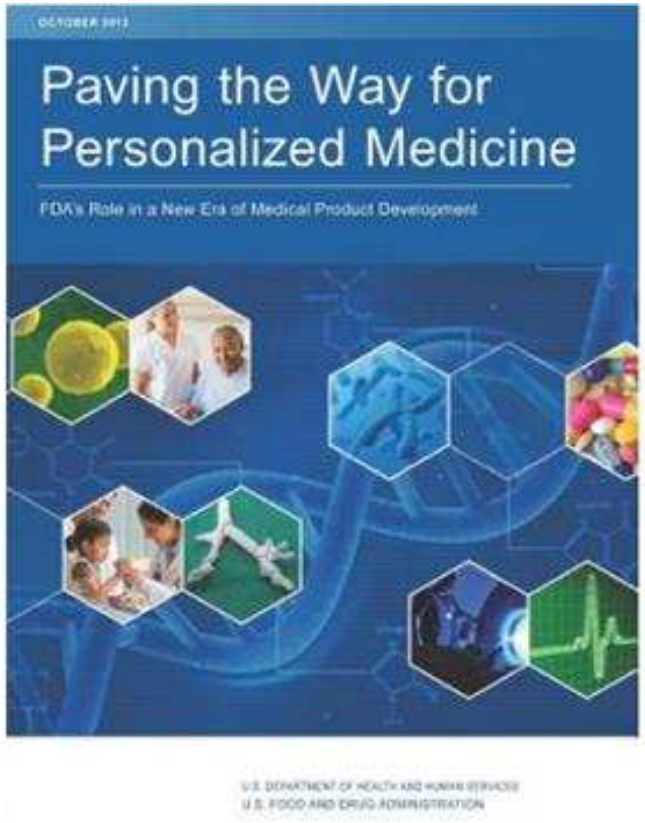
Shanghai University of Finance and Economics, China

Why Optimal Dynamic Treatment Regimes?

Goal: the right treatment for the right patient at the right time

(Personalized medicine / Precision medicine)

In contrast to: “one-size-fits all” and “once-and-for-all” approach



Why Optimal Dynamic Treatment Regimes?

Patient hetergeneity:

- Demographic characteristics
- Physiological characteristics
- Medical history, concomitant conditions
- Genetic/genomic characteristics

Why Optimal Dynamic Treatment Regimes?

Patient hetergeneity:

- Demographic characteristics
- Physiological characteristics
- Medical history, concomitant conditions
- Genetic/genomic characteristics

Clinical practice: Treatment of chronic diseases/disorders is an ongoing process and clinicians manage a patient's illness over the course of a patient's disease

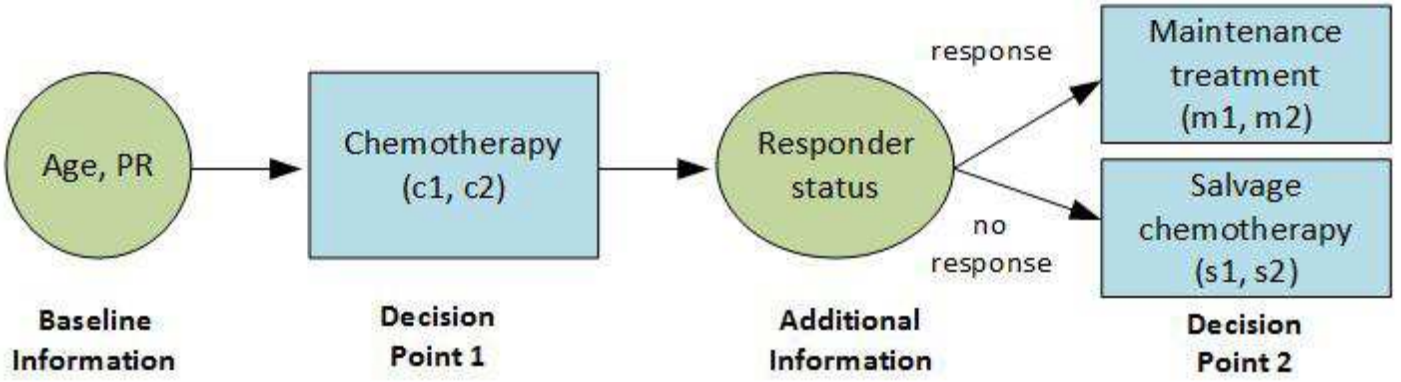
- Clinicians make (a series of) treatment decision(s) over the course
- At key decision points
- Multiple treatment options at each
- Accrued information on the patient

Dynamic Treatment Regimes

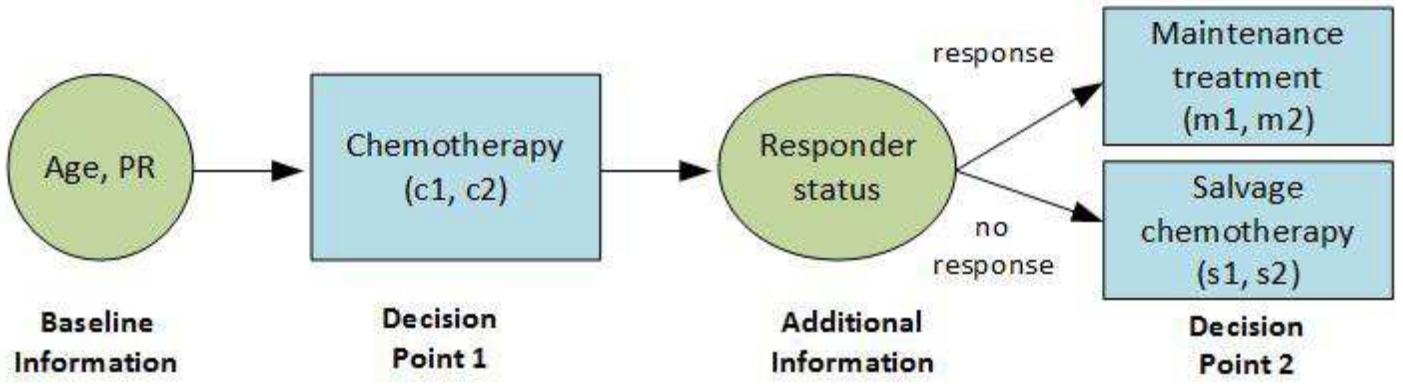
Formalizing precision medicine:

- A set of decision rules for a sequence of decision points at which decisions on treatment are made
  - Decision point $k = 1, \ldots, K$
  - Treatment options at kth stage: $a_k \in \mathcal{A}_k = \{0, 1\}$
  - Treatment history up to k: $\bar{a}_k = (a_1, \ldots, a_k)$
- At each point, the next step of treatment is determined by the decision rule according to information (variables) on the patient up to that point
  - Covariate information between decision $k-1$ and k: x_k
  - Covariate history up to k: $\bar{x}_k = (x_1, \ldots, x_k)$
  - Covariate and treatment history available before decision k: $l_k = (\bar{x}_k, \bar{a}_{k-1})$
  - Dynamic treatment regime: $g = (g_1, \ldots, g_K)$, where $g_k(l_k) \in \mathcal{A}_k$

Dynamic Treatment Regimes



Dynamic Treatment Regimes



Stage 1:

g_1(age, PR)$\in$ Two treatment options$\{$ c_1,$c_2\}$, coded as $\{0, 1\}$

Decision rule of linear form:

$$g_1(\text{age, PR}) = I\{\text{age} > 60 - 8.7\log(\text{PR})\}$$

Decision rule of a tree form:

$$g_1(\text{age, PR}) = I(\text{age} < 50 \text{ and PR} < 10)$$

Stage 2:

g_2(age, PR, treatment at decision 1, responder status...)

$\in$ Four options $\{m_1, m_2, s_1, s_2\}$

Optimal dynamic treatment regime

Potential outcomes framework:

- Potential outcome associated with any regime $g = (g_1, \ldots, g_K) \in \mathcal{G}$:

$$Y^*(g)$$

  - The outcome that would result if the subject followed g

Optimal dynamic treatment regime:

- $g^{opt} = (g_1^{opt}, \ldots, g_K^{opt}) \in \mathcal{G}$

$$E\{Y^*(g^{opt})\} \geq E\{Y^*(g)\} \text{ for all } g = (g_1, \ldots, g_K) \in \mathcal{G}$$

  - The one that would yield maximum expected outcome if were followed by all patients in the population

Notation

Observed data:

$$(\bar{A}_{Ki}, \bar{X}_{Ki}, Y_i), i = 1, \ldots, n$$

- Observed treatment at kth stage, A_k; Observed treatment history up to decision k, $\bar{A}_k = (A_1, \ldots, A_k)$
- X_k is the covariate information observed between decision $k-1$ and k; the observed covariate history up to k, $\bar{X}_k = (X_1, \ldots, X_k)$
- Covariate and treatment history available before decision k, $L_k = (\bar{X}_k, \bar{A}_{k-1})$

Assumptions

- Consistency: $Y = Y^*(\bar{A}_K)$ and $X_k = X_k^*(\bar{A}_{k-1})$
- A patient's covariates and outcome are not affected by treatments received by other patients
- No unmeasured confounders assumption

Notation

Single decision point setting

- Observed data: $(X_i, A_i, Y_i), i = 1, \ldots n$
  - X covariates
  - A treatment
  - Y outcome
- Treatment options: $a = 0, 1$
- Decision rule: $g(X) = 0, 1$
- Potential outcomes: $Y^*(g) = Y^*(1)g(X) + Y^*(0)\{1 - g(X)\}$
- Optimal treatment regime: $g^{opt}(X) = \arg\max_{g \in \mathcal{G}} E\{Y^*(g)\}$
  - the one yielding the largest expected potential outcomes among all regimes

Existing Methods

$$g^{opt}(X) = I\{\mu(1,X) > \mu(0,X)\} = I\{C(X) > 0\}$$

where $\mu(a,X) = E(Y|A=a,X), C(X) = \mu(1,X) - \mu(0,X)$

Outcome regression-based methods

- Parametric regression model for $\mu(A,X) = E(Y|A,X)$, say $\mu(A,X;\beta)$
  - $\widehat{g}^{opt}(X) = I\{\mu(1,X,\widehat{\beta}) > \mu(0,X,\widehat{\beta})\}$
  - Extension to multiple decision point setting leads to Q-learning **(backward induction)**
- Semiparametric regression model for $\mu(A,X) = h_1(X) + AC(X;\psi)$
  - $h_1(X)$ is unspecified
  - $\widehat{g}^{opt}(X) = I\{C(X;\widehat{\psi}) > 0\}$
  - Extension to multiple decision point setting leads to A-learning **(backward induction)**

Existing methods

Backward induction

- Q-function: $Q_K(\bar{x}_K, \bar{a}_{K-1}, a_K) = E(Y|\bar{X}_K = \bar{x}_K, \bar{A}_K = \bar{a}_K)$
- $g_K^{opt}(\bar{x}_K, \bar{a}_{K-1}) = \arg\max_{a_K \in \Phi_K(\bar{x}_K, \bar{a}_{K-1})} Q_K(\bar{x}_K, \bar{a}_{K-1}, a_K)$
- Recursively define
  - Value function as $V_k(\bar{x}_k, \bar{a}_{k-1}) = \max_{a_k \in \mathcal{A}_k} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k)$ for $k = K, \ldots, 2$, with $\bar{a}_0$ being null,
  - Q-functions as $Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k) = E\{V_{k+1}(\bar{x}_k, X_{k+1}, \bar{a}_k)|\bar{X}_k = \bar{x}_k, \bar{A}_k = \bar{a}_k\}$ for $k = K-1, \ldots, 1$
- The optimal decision rule at the k-th point satisfies $g_k^{opt}(\bar{x}_k, \bar{a}_{k-1}) = \arg\max_{a_k \in \mathcal{A}_k} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k)$

Existing Methods

$$g^{opt}(X) = \arg\max_{g \in \mathcal{G}} E\{Y^*(g)\}$$

Direct optimization methods

- Robust method of Zhang, et al(2012 Biometrics)
  - For each regime $g \in \mathcal{G}_\eta$, estimate $E\{Y^*(g)\}$ using augmented inverse probability weighted estimator (AIPWE)

$$AIPWE(\eta) = n^{-1} \sum_{i=1}^{n} \left\{ \frac{\mathcal{C}_{\eta,i} Y_i}{\pi_c(X_i; \widehat{\gamma})} - \frac{\mathcal{C}_{\eta,i} - \pi_c(X_i; \eta, \widehat{\gamma})}{\pi_c(X_i; \widehat{\gamma})} m(X_i; \eta, \widehat{\beta}) \right\},$$

  - $\mathcal{C}_{\eta,i} = I\{A_i = g(X_i, \eta)\}$;
    $\pi_c(X_i, \gamma) = A_i \pi(X_i, \gamma_i) + (1 - A_i)\{1 - \pi(X_i, \gamma_i)\}$
  - Double robust property of AIPWE
  - Requires to prespecify $\mathcal{G}_\eta$
  - Extension to multiple decision point setting: Zhang, et al (2013, Biometrika)
    (Monotone coasening/missing)

Existing Methods

- Outcome weighted learning (OWL; Zhao et al, 2012, JASA)
  - Estimate $E\{Y^*(g)\}$ using inverse probability weighted estimator (IPWE)

$$IPWE = n^{-1} \sum_{i=1}^{n} \left\{ \frac{I\{A_i = g(X_i)\} Y_i}{A_i \pi(X_i, \gamma_i) + (1 - A_i)\{1 - \pi(X_i, \gamma_i)\}} \right\},$$

  - Equivalent to minimizing $\sum_{i=1}^{n} \left\{ \frac{I\{A_i \neq g(X_i)\} Y_i}{A_i \pi(X_i, \gamma_i) + (1 - A_i)\{1 - \pi(X_i, \gamma_i)\}} \right\}$
  - $I\{A_i \neq g(X_i)\}$ is viewed as a zero-one loss in classification, blue part as weight when Y is positive
  - No use of outcome regression model (not taking advantage of patient characteristics)
  - Extension to multiple decision point setting: BOWL, SOWL (Zhao, et al, 2015, JASA) **(Monotone coasening/missing)**

C-learning

Theorem 1: *Let* $g^* = (g_1^*, \ldots, g_K^*)$*, be a treatment regime that satisfies*

$$g_k^*(L_k) = \arg \min_{g_k \in \mathcal{G}_k} E[|C_k(L_k)| I\{Z_k \neq g_k(L_k)\}]$$

where $Z_k = I\{C_k(L_k) > 0\}$*,* $k = K, \ldots, 1$*, then* g^* *is the optimal dynamic treatment regime.*

- $Q_K(\bar{x}_K, \bar{a}_K) = E(Y | \bar{X}_K = \bar{x}_K, \bar{A}_K = \bar{a}_K)$
- $V_k(\bar{x}_k, \bar{a}_{k-1}) = \max_{a_k \in \mathcal{A}_k} Q_k(\bar{x}_k, \bar{a}_{k-1}, a_k)$
- $Q_k(L_k, a_k) = E\{V_{k+1}(L_{k+1}) | L_k, a_k\}$
- $C_k(L_k) = Q_k(L_k, 1) - Q_k(L_k, 0)$

C-learning

$$g_k^{opt}(L_k) = \arg\min_{g_k \in \mathcal{G}_k} E[|C_k(L_k)| I\{Z_k \neq g_k(L_k)\}], Z_k = I\{C_k(L_k) > 0\}$$

Intuition: Consider the last stage,

- Directly optimizing estimate of $E\{Y^*(\bar{A}_{K-1}, g_K)\}$ over a class of regimes
- Retain only part of $E\{Y^*(\bar{A}_{K-1}, g_K)\}$ relevant for decision making
- $g_K^{opt}(X) = \arg\max_{g_K \in \mathcal{G}_K} E\{Y^*(\bar{A}_{K-1}, g_K)\}$
  $= \arg\max_{g_K \in \mathcal{G}_K} [E\{g_K(L_K) C_K(L_K)\}] + Q_K(L_K, 0)$
- Decompose $C_K(L_K)$ into magnitude $|C_K(L_K)|$ and sign $I\{C_K(L_K) > 0\}$
- $g(L_K) C_K(L_K) = Z_K |C_K(L_K)| - |C_K(L_K)| I\{Z_K \neq g(L_K)\}$
- $g_K^{opt}(X) = \arg\min_{g_K \in \mathcal{G}_K} E[|C_K(L_K)| I\{Z_K \neq g(L_K)\}]$
- Classification perspective leads to powerful and flexible learning algorithms

C-learning

Proposition 1: *The value functions satisfy the following condition:*

$$E[V_{k+1}(L_{k+1}) + \{Q_k(L_k, 1) - Q_k(L_k, 0)\}\{g_k^{opt}(L_k) - A_k\}|L_k] = V_k(L_k),$$

$k = K, \ldots, 1, V_{K+1} \equiv Y$, *where* g_k^{opt} *is the optimal decision rule at stage* k.

C-learning

At stage K:

- $g_K^{opt}(L_K) = \arg\min_{g_K \in \mathcal{G}_K} E[|C_K(L_K)|\{Z_K \neq g_K(L_K)\}]$
- Considering data (Y_i, L_{Ki}, A_{Ki}), estimate $C_K(L_{Ki})$ by the AIPWE estimate

$$\begin{aligned}\widehat{C}_K(L_{Ki}) &= \frac{A_{Ki}}{\widehat{\pi}_K(L_{Ki})}Y_i - \frac{A_{Ki}-\widehat{\pi}_K(L_{Ki})}{\widehat{\pi}_K(L_{Ki})}\widehat{Q}_K(L_{Ki},1) \\ &- \{\frac{1-A_{Ki}}{1-\widehat{\pi}_K(L_{Ki})}Y_i - \frac{A_{Ki}-\widehat{\pi}_K(L_{Ki})}{1-\widehat{\pi}_K(L_{Ki})}\widehat{Q}_K(L_{Ki},0)\},\end{aligned}$$

- Weighted classification:

$$\widehat{g}_{C,K}^{opt} = \arg\min_{g_K \in \mathcal{G}_K} \sum_{i=1}^n [\widehat{W}_{Ki}\{\widehat{Z}_{Ki} \neq g_K(L_{Ki})\}],$$

  - Class label: $\widehat{Z}_{Ki} = I\{\widehat{C}_K(L_{Ki}) > 0\}$
  - Weight: $\widehat{W}_{Ki} = |\widehat{C}_K(L_{Ki})|$

using by existing optimization/classification techniques

C-learning

After obtaining $\widehat{g}_{C,K}^{opt}$, the C-learning moves backward till the first stage

At stage k: $k = K-1, \ldots, 1,$

- $g_k^{opt}(L_k) = \arg\min_{g_k \in \mathcal{G}_k} E[|C_k(L_k)|\{Z_k \neq g_k(L_k)\}]$
- $C_k(L_k) = Q_k(L_k, 1) - Q_k(L_k, 0)$, and
  $Q_k(L_k, a_k) = E\{V_{k+1}(L_{k+1})|L_k, a_k\}$
- Denoting $\widetilde{V}_{(K+1)i} = Y_i$, estimate $V_k(L_{ki})$ recursively by
  $\widetilde{V}_{ki} \equiv \widetilde{V}_k(L_{ki}) = \widetilde{V}_{(k+1)i} + \{\widehat{Q}_k(L_{ki}, 1) - \widehat{Q}_k(L_{ki}, 0)\}\{\widehat{g}_{C,k}^{opt}(L_{ki}) - A_{ki}\},$
- Treating $(\widetilde{V}_{k+1,i}, L_{ki}, A_{ki})$ as "data", estimate $C_k(L_{ki})$ by AIPWE estimate $\widehat{C}_k\{L_{ki}\}$
- $\widehat{g}_{C,k}^{opt} = \arg\min_{g_k \in \mathcal{G}_k} \sum_{i=1}^n [\widehat{W}_{ki}\{\widehat{Z}_{ki} \neq g_k(L_{ki})\}]$, where
  - Class label: $\widehat{Z}_{ki} = I\{\widehat{C}_k(L_{ki}) > 0\}$
  - Weight: $\widehat{W}_{ki} = |\widehat{C}_k(L_{ki})|$

C-learning

Algorithm:

1. At stage K, based on data $(Y_i, L_{Ki}, A_{Ki}), i = 1, \ldots, n$,

   1.1 Build model for $P(A_K = 1|L_K)$ to estimate propensity score, $\hat{\pi}_K(L_{Ki})$

   1.2 Build model for $Q_K(L_K, A_K) = E(Y|L_K, A_K)$ and obtain estimate of $Q_K(L_K, a_K), a_K = 0, 1$, for each subject, $\widehat{Q}_K(L_{Ki}, a_K)$

   1.3 Estimate the contrast function $C_K(L_{Ki})$ for each subject by the AIPWE $\widehat{C}_K(L_{Ki})$

   1.4 Estimate g_K^{opt} by minimizing a weighted misclassification error

   * classification data set $(\widehat{Z}_i, \widehat{W}_i, L_i)$
   * $\widehat{Z}_i$ is class label: $I(\widehat{C}_K(L_{Ki}) > 0)$
   * $\widehat{W}_i$ is weight: $|\widehat{C}_K(L_{Ki})|$
   * L_i covariates and treatment history

   1.5 Estimate $\widetilde{V}_{K,i}$

2. Repeat 1.1-1.4 for stage $k = K - 1, \ldots, 1$ sequentially, based on "data" $(\widetilde{V}_{k+1,i}, L_{ki}, A_{ki}), i = 1, \ldots, n$, to obtain estimate of g_k^{opt}

C-learning and other methods

C-learning and Zhang et al. (2013)

- Similarity:
  - Direct optimization
  - AIPWE
- Difference:
  - Weighted classification perspective
  - Sequential optimization vs. simultaneous optimization across stages

C-learning and other methods

C-learning and BOWL:

- AIPWE vs. IPWE
- Classification perspective
  - BOWL: misclassify if $A_i \neq g(X_i)$
    - Undesirable feature: estimated regime tries to keep observed treatment assignments (Zhou et al., 2015, JASA)
    - Dependent on the IPWE of $E\{Y^*(g)\}$
  - C-learning: misclassify if $I\{C(X_i) > 0)\} \neq g(X_i)$
    - Theorem 1 is a general result and does not depend on estimator of $E\{Y^*(g)\}$
    - Meaningful and consistent with the goal of optimizing treatment decisions
- Backward sequential optimization
  - BOWL loses sample size geometrically with stages

C-learning: high dimensionality

Regimes of linear form: Variable selection

- Target selecting prescriptive variables as opposed to predictive variables
- weighted misclassification error rate corresponding to $\{X_{j^1}, ..., X_{j^m}\}$ as

$$err(X_{j^1}, ..., X_{j^m}) = \min_{\beta} \sum_{i=1}^{n} \widehat{W}_i I\{\widehat{Z}_i \neq I(\beta_0 + \beta_1 X_{j^1} + ... + \beta_m X_{j^m} > 0)\},$$

- Forward Minimal Misclassification Error Rate (ForMMER) Selection (Zhang and Zhang, 2018)
  - First selected variable: $X_{j^1} = \arg\min_{X_j \in (X_1, ..., X_p)} err(X_j)$
  - m-th selected variable: $X_{j^m} = \arg\min_{X_j \in \mathcal{F}^0 \setminus \mathcal{S}^{(m-1)}} err(\mathcal{S}^{(m-1)}, X_j)$
  - $\mathcal{S}^{(m)} = \{X_{j^1}, ..., X_{j^m}\}$: set of selected variables up to steps m-th step

Regimes of the form of a decision tree

- Existing classification algorithms capable of handling high dimensional set of covariates
- e.g., CART

C-learning: Simulations (scenario I)

Simulation setting I:

- Adopted from Zhao et al. (2015; JASA)
- Treatments A_1,A_2 and A_3 are generated from $\{1, 0\}$ with equal probability 0.5.
- $X_{1,1}, X_{1,2}, X_{1,3}$ are generated from $N(45, 15^2)$. X_2 is generated according to $X_2 \sim N(1.5X_{1,1}, 10^2)$ and X_3 is generated according to $X_3 \sim N(0.5X_2, 10^2)$
- Outcome: $Y = \mu(\bar{A}_3, \bar{X}_3) + \epsilon$ for ϵ standard normal and $\mu(\bar{A}_3, \bar{X}_3) = 20 - |0.6X_{1,1} - 40|(A_1 - g_1^{opt})^2 - |0.8X_2 - 60|(A_2 - g_2^{opt})^2 - |1.4X_3 - 40|(A_3 - g_3^{opt})^2$, where
  - $g_1^{opt} = I(X_{1,1} - 30 > 0)$
  - $g_2^{opt} = I(X_2 - 40 > 0)$
  - $g_3^{opt} = I(X_3 - 40 > 0)$
- The optimal treatment decision rule at each stage depends only on a single covariate

C-learning: Simulations (scenario I)

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	10.84(1.85)	12.13(1.54)	13.02(1.36)
Q-learning	12.49(1.83)	12.76(1.46)	13.05(1.14)
Zhang et al.(2013)	13.25(2.12)	15.08(1.46)	16.28(1.01)
C-learning	17.27(0.97)	18.52(0.74)	19.37(0.41)

- $E\{Y^*(g^{opt})\} = 20$
- Specification of Q-functions the same as in Zhao, et al (2015, JASA)
- AIPWE in Zhang et al. (2013) and C-learning use the same Q-functions for augmentation terms
- In method of Zhang, et al (2013), consider all available covariates at each stage in parameterizing regimes
- In C-learning, the optimization step is carried out by a genetic algorithm (R package Rgenoud)

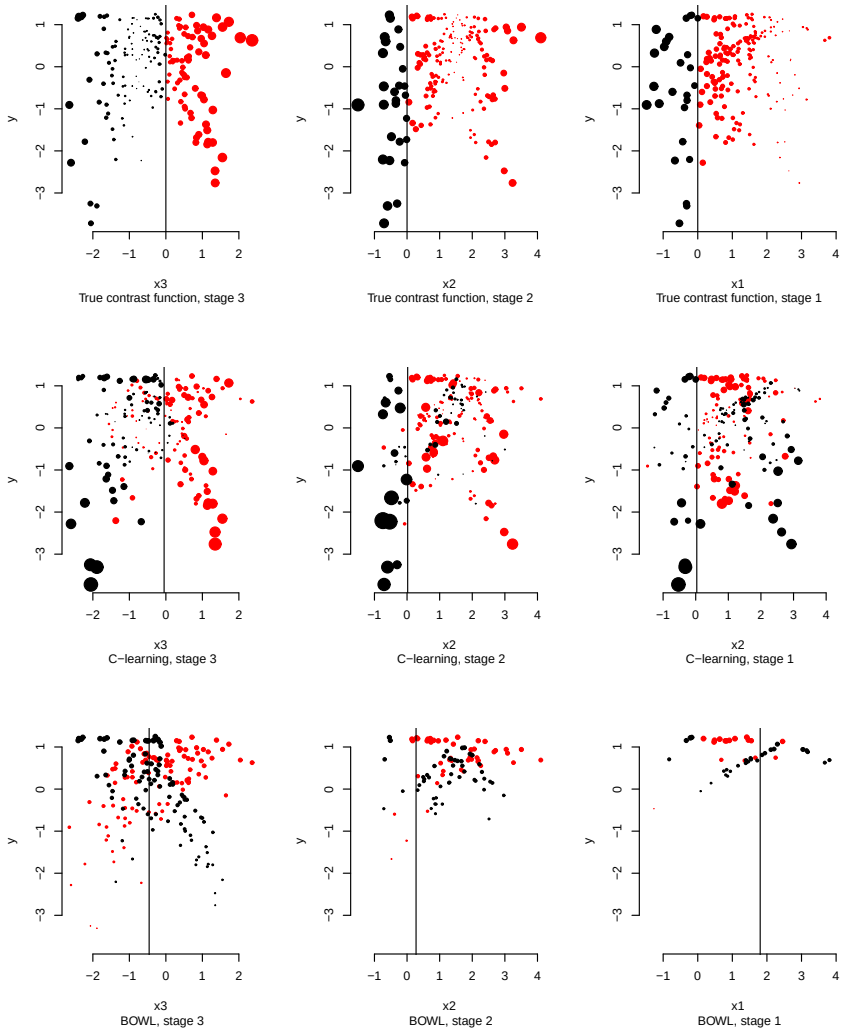
C-learning: Simulations (scenario I)

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	10.84(1.85)	12.13(1.54)	13.02(1.36)
Q-learning	12.49(1.83)	12.76(1.46)	13.05(1.14)
Zhang et al.(2013)	13.25(2.12)	15.08(1.46)	16.28(1.01)
C-learning	17.27(0.97)	18.52(0.74)	19.37(0.41)

- Zhang et al. (2013) and C-learning outperform Q-learning, even though the same Q-functions are used: direct optimization vs. outcome regression based method
- C-learning outperforms Zhang et al. (2013), even though both are based on the same AIPWE: sequential vs. simultaneous optimization
- C-learning outperforms BOWL: classification perspective; AIPWE vs. IPWE; sequential optimization

C-learning: Simulations (scenario I)

Figure 1: Classification data set (n=200)



- C-learning vs. BOWL: classification perspective; AIPWE vs. IPWE; sequential optimization

C-learning: Simulations (scenario II)

Simulation setting II:

- Increase the dimension of covariates at each stage so that the total number is 50
- 40 baseline covariates $X_{1,1}, ..., X_{1,40}$ are generated from $N(45, 15^2)$. At stage 2, $X_{2,j} \sim N(1.5X_{1,j}, 10^2), j = 1, ..., 5$. At stage 3, $X_{3,j} \sim N(0.5X_{2,j}, 10^2)$, $j = 1, ..., 5$.
- The outcome was generated as $Y = \mu(\bar{A}_3, \bar{X}_3) + \epsilon$ for ϵ standard normal and $\mu(\bar{A}_3, \bar{X}_3) = 20 - |0.6X_{1,1} - 40|(A_1 - g_1^{opt})^2 - |0.8X_{2,1} - 60|(A_2 - g_2^{opt})^2 - |1.4X_{3,1} - 40|(A_3 - g_3^{opt})^2$, where
  - $g_1^{opt} = I(X_{1,1} - X_{1,2} > 0)$
  - $g_2^{opt} = I(X_{2,1} - X_{2,2} > 0)$
  - $g_3^{opt} = I(X_{3,1} - X_{3,2} > 0)$
- Otherwise similar to scenario 1, except that the optimal decision rule at each stage a little bit more complicated in that it depends on linear combination of two covariates

C-learning: Simulations (scenario II)

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	3.38(1.62)	5.93(1.37)	7.79(1.10)
BOWL†	14.76(1.74)	15.43(1.38)	15.74(1.12)
Q-learning†	14.01(1.05)	13.94(0.78)	13.78(0.56)
Zhang et al.(2013)†	17.98(1.42)	18.83(0.87)	19.35(0.45)
C-learning-Q	17.70(1.75)	19.45(0.61)	19.78(0.22)
C-learning-RF	16.59(2.14)	19.21(0.80)	19.75(0.14)

- In specification of Q-functions, use only "important covariates"
- Methods† consider only regimes constructed by "important covariates" (not feasible in practice)
- In C-learning-RF, the augmentation term is fit by random forest
- In C-learning, use data-driven way to choose important covariates from the high dimension of covariates and search the optimal regime among all regimes of the linear form

C-learning: Simulations (scenario II)

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	3.38(1.62)	5.93(1.37)	7.79(1.10)
BOWL†	14.76(1.74)	15.43(1.38)	15.74(1.12)
Q-learning†	14.01(1.05)	13.94(0.78)	13.78(0.56)
Zhang et al.(2013)†	17.98(1.42)	18.83(0.87)	19.35(0.45)
C-learning-Q	17.70(1.75)	19.45(0.61)	19.78(0.22)
C-learning-RF	16.59(2.14)	19.21(0.80)	19.75(0.14)

- BOWL has considerable worse performance when the dimension of covariates is high
- C-learning-Q outperforms Q-learning, even though they both use use the same outcome-regression model: direct optimization vs. outcome regression based method
- C-learning is comparable to Zhang et al. (2013)†, even though Zhang et al. (2013)† considers a much smaller class of regimes: sequential vs. simultaneous optimization

C-learning: Simulations (scenario III)

Simulation setting III:

- Data generating scenario is the same as scenario II except that
  - $g_1^{opt} = I(X_{1,1} > 40)I(X_{1,2} < 60)$
  - $g_2^{opt} = I(X_{2,1} > 60)I(X_{2,2} < 90)$
  - $g_3^{opt} = I(X_{3,1} > 30)I(X_{3,2} < 50)$ in $\mu(\bar{A}_3, \bar{X}_3)$
- The optimal decision rule at each stage is of the form of a decision tree

C-learning: Simulations (scenario III)

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	3.01(1.63)	5.02(1.42)	6.73(1.15)
BOWL†	12.55(1.28)	12.91(0.95)	13.12(0.72)
Q-learning†	13.12(0.45)	13.08(0.35)	13.07(0.23)
Zhang et al.(2013)†	17.02(1.25)	18.02(0.90)	18.71(0.63)
C-learning-Q	17.44(1.29)	18.91(0.73)	19.52(0.32)
C-learning-RF	16.94(1.48)	18.92(0.63)	19.61(0.24)

- Methods† limit the search among regimes constructed using only "important covariates" (not feasible in practice)
- In C-learning-RF, the augmentation term is fit by random forest
- In C-learning-Q, the augmentation term uses the same model as in Q-learning
- C-learning uses CART to choose important covariates from the high dimensional set of covariates and optimizes regimes across all regimes of the form of a decision tree

C-learning: Simulations (scenario III)

	n=200	n=400	n=800
Estimator	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$	$E(\widehat{g}^{opt})$
BOWL	3.01(1.63)	5.02(1.42)	6.73(1.15)
BOWL†	12.55(1.28)	12.91(0.95)	13.12(0.72)
Q-learning†	13.12(0.45)	13.08(0.35)	13.07(0.23)
Zhang et al.(2013)†	17.02(1.25)	18.02(0.90)	18.71(0.63)
C-learning-Q	17.44(1.29)	18.91(0.73)	19.52(0.32)
C-learning-RF	16.94(1.48)	18.92(0.63)	19.61(0.24)

- BOWL has considerable worse performance when the dimension of covariates is high
- C-learning-Q outperforms Q-learning, even though they both use use the same outcome-regression model: direct optimization vs. outcome regression based method
- C-learning is comparable to Zhang et al. (2013)†, even though Zhang et al. (2013)† considers a much smaller class of regimes: sequential vs. simultaneous optimization
- C-learning-RF is data-driven

C-learning: Simulations (variable selection)

Method	ρ	Size	TP	ER	VR
		n=200			
SAS (Fan, Lu and Song, 2016)	0.2	14.49 (2.07)	0.72 (0.60)	43.9 (5.3)	71.8 (5.6)
C-learning ForMMER-AIPWE		3.99 (0.95)	1.97 (0.16)	**11.8** (5.1)	**97.3** (2.5)
C-learning ForMMER-reg		4.39 (1.04)	**1.98** (0.15)	12.5 (6.3)	96.7 (2.9)
SAS	0.8	8.55 (2.03)	0.96 (0.35)	21.3 (6.3)	90.9 (5.1)
C-learning ForMMER-AIPWE		3.57 (1.00)	1.81 (0.39)	7.7 (2.9)	98.7 (0.9)
C-learning ForMMER-reg		3.84 (1.00)	**1.99** (0.11)	**4.4** (1.8)	**99.6** (0.4)

- Single decision point setting, $K = 1$.
- A total of 500 covariates.
- Size: number of selected prescriptive variables; TP: number of true positive (important) prescriptive variables; ER: error rate of the treatment decision; VR (value ratio): ratio of the value of the estimated regime relative to that of the true optimal regime. Numbers in parenthesis are Monte Carlo standard deviations.

Discussion

- The proposed C-learning is a powerful and flexible learning method
  - Existing powerful and off-the-shelf classification/optimization techniques can be used to optimize the decision rules (eg, CART, the genetic algorithm discussed by Goldberg, 1989)
  - Existing model building techniques (parametric and nonparametric, eg, random forest) can be used to best estimate the Q-function
  - Accommodate variable selection targeting selecting prescriptive variables (variables relevant for treatment decision making) as opposed to predictive variables
  - Each step (estimate Q-function and optimization) targets the right goal
    - Q-function: best model outcomes and make predictions
    - Optimization: Optimize decision rules
    - However, as pointed out by Murphy (2005), there is a mismatch between outcome-regression based methods (eg, Q- and A-learning) and the goal of optimizing decision rules

References

- Zhang, B. & Zhang, M. (2018). C-learning: a New Classification Framework to Estimate Optimal Dynamic Treatment Regimes. *Biometrics,*74:891-899.
- Zhang, B. & Zhang, M. (2018). Variable Selection for Estimating the Optimal Treatment Regimes in the Presence of a Large Number of Covariate. *Annals of Applied Statistics*.
- Murphy, S. A. (2003). Optimal dynamic treatment regimes (with discussion). *J. Royal Statist. Soc., Ser. B* **58**, 331–366.
- Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Mach. Learn.* **8**, 279–292.
- Zhang, B., Tsiatis, A. A., Laber, E. B. & Davidian, M. (2012a). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 1010–1018.
- Zhang, B., Tsiatis, A. A., Laber, E. B. Davidian, M., Zhang, M. & Laber, E. B.(2012b). Estimating optimal treatment regimes from a classification perspective *Stat* **1**, 103–114.
- Zhang, B.,Tsiatis, A. A., Laber, E. B., & Davidian, M.(2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions *Biometrika* **100**, 681–694.
- Zhao, Y., Zeng, D., Rush, A. J. & Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.
- Zhao, Y., Zeng, D., Laber, E. B & Kosorok, M. R. (2015). New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes. *Journal of the American Statistical Association* **510**, 583–598.
- Zhou, X., Mayer-Hamblett, N., Khan, U. & Kosorok, M. R. (2015). Residual Weighted Learning for Estimating Individualized Treatment Rules. *Journal of the American Statistical Association* DOI: 10.1080/01621459.2015.1093947.
- Fan, A., Lu, W. & Song, R. (2016). Sequential Advantage Selection for Optimal Treatment Regimens. *Annals of Applied Statistics,* 10, 32-53.