Implementing Personalized Medicine: Estimation of Optimal Dynamic Treatment Regimes

Anastasios (Butch) Tsiatis

Department of Statistics
North Carolina State University

Optimal Regime

Assume: *Large* outcomes are *good*

An optimal regime:

- A *regime* that, if followed by all patients in the population, yields the *largest outcome on average*
- That is, yields the largest *value*

Goal: Given *data* (*evidence*) from a clinical trial or observational study, *estimate* an *optimal regime* satisfying this definition

- *For now*: Consider regimes involving a *single decision*/*rule*

Statistical Framework

Simplest setting: A *single decision* with *two* treatment options

- $\mathcal{A} = \{0, 1\}$

Observed data: (Y_i, X_i, A_i), $i = 1, \ldots, n$, iid

- Y_i outcome, X_i baseline covariates, $A_i = 0, 1$ treatment received

Treatment regime: A single *rule*

- A function $d : \mathcal{X} \to \{0,\ 1\}$

Statistical Framework

Breast cancer example: Which treatment to give patients who present with *primary operable breast cancer*?

- Two treatment options (0 or 1), x =(age, PR)
- Possible rules

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

$$d(\text{age,PR}) = I\{\text{age} + 8.7\log(\text{PR}) - 60 > 0\}$$

Goal, restated:

- Let $\mathcal{D}$ be the class of *all* possible regimes d
- Estimate $d^{opt} \in \mathcal{D}$ such that, if d^{opt} were followed by *all patients* in the population, it would lead to *largest average outcome* (*value*) among all regimes in $\mathcal{D}$

Potential Outcomes

Reminder: We can hypothesize *potential outcomes*

- $Y^*(1)$ = outcome that would be achieved if patient were to receive 1; $Y^*(0)$ defined similarly
- $E\{Y^*(1)\}$ is the *average outcome* if *all patients* in the population were to receive 1; and similarly for $E\{Y^*(0)\}$
- We *observe*

$$Y = Y^*(1)A + Y^*(0)(1 - A)$$

Potential Outcomes

No unmeasured confounders: Assume that

$$Y^*(0), Y^*(1) \perp\!\!\!\perp A | X$$

- X contains all information used to assign treatments
- Automatically satisfied for data from a *randomized trial*
- Standard but *unverifiable* assumption for *observational studies*
- Implies that

$$\begin{aligned} E\{Y^*(1)\} &= E[E\{Y^*(1)|X\}] \\ &= E[E\{Y^*(1)|X, A=1\}] \\ &= E\{\, E(Y|X, A=1)\,\} \end{aligned}$$

and similarly for $E\{Y^*(0)\}$

Potential Outcomes

$$E\{Y^*(1)\} = E\{\, E(Y|X, A = 1)\,\}$$

Implication for estimating $E\{Y^*(1)\}$: Similarly for $E\{Y^*(0)\}$

- $E(Y|X, A) = Q(X, A)$ is the *regression* of Y on X and A
- $E(Y|X, A)$ is *unknown*
- Posit a *model* $Q(X, A; \beta)$ for $Q(X, A)$
- Estimate β based on observed data $\Longrightarrow \widehat{\beta}$ (e.g., least squares)
- *Estimator* for $E\{Y^*(1)\}$

$$n^{-1} \sum_{i=1}^{n} Q(X_i, 1; \widehat{\beta})$$

Potential Outcomes

Potential outcome for a regime:

- For any $d \in \mathcal{D}$, define $Y^*(d)$ to be the *potential outcome* for a patient if s/he were given treatment according to regime d

$$Y^*(d) = Y^*(1)d(X) + Y^*(0)\{1 - d(X)\}$$

- $E\{Y^*(d)\}$ is the *average outcome for the population* if all patients were treated according to regime d
- That is, $E\{(Y^*(d)\} = V(d)$ is the *value* of regime d

Value of a Regime

$$Y^*(d) = Y^*(1)d(X) + Y^*(0)\{1 - d(X)\}$$

Value of regime d: Using *no unmeasured confounders*

$$\begin{aligned} E\{Y^*(d)\} &= E[E\{Y^*(d)|X\}] \\ &= E\Big[E\{Y^*(1)|X\}d(X) + E\{Y^*(0)|X\}\{1 - d(X)\}\Big] \\ &= E\Big[\, E(Y|X, A = 1)d(X) + E(Y|X, A = 0)\{1 - d(X)\}\,\Big] \\ &= E[Q(X, 1)d(X) + Q(X, 0)\{1 - d(X)\}], \end{aligned}$$

where $E(Y|X, A) = Q(X, A)$

Estimating the Value of a Regime

$$E\{Y^*(d)\} = E[Q(X,1)d(X) + Q(X,0)\{1 - d(X)\}]$$

Again: $E(Y|X, A)$ is *not known*

- *Posit a model* $Q(X, A; \beta)$ for $E(Y|X, A)$
- *Estimate* β based on observed data $\Longrightarrow \widehat{\beta}$ (e.g., least squares)
- *Estimate* $V(d) = E\{Y^*(d)\}$ by

$$\widehat{V}(d) = n^{-1}\sum_{i=1}^{n}[Q(X_i, 1, \widehat{\beta})d(X_i) + Q(X_i, 0, \widehat{\beta})\{1 - d(X_i)\}]$$

Optimal Regime

Reminder: d^{opt} is a regime in $\mathcal{D}$ such that

- $E\{Y^*(d)\} \leq E\{Y^*(d^{opt})\}$ for all $d \in \mathcal{D}$
- $E\{Y^*(d)|X=x\} \leq E\{Y^*(d^{opt})|X=x\}$ for all $d \in \mathcal{D}$ and $x \in \mathcal{X}$

Optimal regime:

$$d^{opt}(x) = \text{arg max}_{a=\{0,1\}} E\{Y^*(a)|X=x\}$$

- *Thus*

$$\begin{aligned} d^{opt}(x) &= I[E\{Y^*(1)|X=x\} > E\{Y^*(0)|X=x\}] \\ &= I\{\, Q(x,1) > Q(x,0)\,\} \end{aligned}$$

Estimating the Optimal Regime

"Regression estimator":

- *Estimate* d^{opt} by

$$\widehat{d}^{opt}_{REG}(x) = I\{\, Q(x, 1; \widehat{\beta}) > Q(x, 0; \widehat{\beta})\,\}$$

- Estimator for $V(d^{opt}) = E\{Y^*(d^{opt})\}$

$$\widehat{V}_{REG}(\widehat{d}^{opt}_{REG}) = n^{-1} \sum_{i=1}^{n} \Big[Q(X, 1_i, \widehat{\beta}) \widehat{d}^{opt}_{REG}(X_i) + Q(X, 0_i, \widehat{\beta}) \{1 - \widehat{d}^{opt}_{REG}(X_i)\} \Big]$$

Concern: $Q(X, A; \beta)$ may be *misspecified*, so $\widehat{d}^{opt}_{REG}$ could be far from d^{opt}

Optimal Restricted Regime

Alternative perspective: $Q(X, A; \beta)$ defines a *class* of regimes

$$d(x, \beta) = I\{Q(x, 1; \beta) > Q(x, 0; \beta)\},$$

indexed by β, that *may or may not* contain d^{opt}

- E.g., suppose *in truth*

$$E(Y|X, A) = \exp\{1 + X_1 + 2X_2 + 3X_1X_2 + A(1 - 2X_1 + X_2)\}$$

$$\Longrightarrow d^{opt}(x) = I(x_2 \geq 2x_1 - 1) \;\; (\textit{hyperplane})$$

Optimal Restricted Regime

Posited model:

$$Q(X, A; \beta) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + A(\beta_3 + \beta_4 X_1 + \beta_5 X_2)$$

- Regimes $I\{Q(x, 1; \beta) > Q(x, 0; \beta)\}$ define a *class of regimes* $\mathcal{D}_\eta$ with elements

$$I(x_2 \geq \eta_1 x_1 + \eta_0) \text{ or } I(x_2 \leq \eta_1 x_1 + \eta_0), \quad \eta_0 = -\beta_3/\beta_5,\ \eta_1 = -\beta_4/\beta_5$$

  depending on the sign of β_5

- Parameter η is defined as a *function of* β
- The optimal regime *in this case* is contained in $\mathcal{D}_\eta$
- However, the estimated regime $I\{Q(x, 1; \widehat{\beta}) > Q(x, 0; \widehat{\beta}\}$ *may not* estimate the optimal regime within $\mathcal{D}_\eta$ if the posited model is *incorrect*

Optimal Restricted Regime

Suggests: Consider *directly* a *restricted class of regimes* $\mathcal{D}_\eta$ with elements of form

$$d(x;\eta) = d_\eta(x) \quad \text{indexed by } \eta$$

- Such regimes may be motivated by a regression model or based on *cost*, *feasibility* in practice, *interpretability*; e.g.,

$$d(x;\eta) = I(x_1 < \eta_0, x_2 < \eta_1)$$

- $\mathcal{D}_\eta$ *may or may not* contain d^{opt} but is still of interest
- *Optimal restricted regime* $d^{opt}_\eta(x) = d(x;\eta^{opt})$,

$$\eta^{opt} = \arg\max_\eta E\{Y^*(d_\eta)\}$$

Estimating the Optimal Restricted Regime

Optimal restricted regime: $d_\eta^{opt}(x) = d(x; \eta^{opt})$,

$$\eta^{opt} = \arg\max_\eta E\{Y^*(d_\eta)\} = \arg\max_\eta V(d_\eta)$$

Approach:

- Directly estimate the *value* $V(d_\eta) = E\{Y^*(d_\eta)\}$ for any fixed $\eta \Longrightarrow \widehat{V}(d_\eta)$
- Estimate the *optimal restricted regime* by finding

$$\widehat{\eta}^{opt} = \arg\max_\eta \widehat{V}(d_\eta) \Longrightarrow \widehat{d}_\eta^{opt}(x) = d(x; \widehat{\eta}^{opt})$$

- We refer to this as a *value search estimator* for d_η^{opt}

Value Search Estimators

Required: A "*good*" estimator for $V(d_\eta)$

- *Missing data* analogy
- Let C_η denote η*-regime consistency indicator*

$$C_\eta = Ad(X;\eta) + (1 - A)\{1 - d(X;\eta)\}$$

- "*Full data*" are $\{X, Y^*(d_\eta)\}$; "*observed data*" are $(X, C_\eta, C_\eta Y)$
- $\Longrightarrow$ Only a subset of subjects have observed outcomes under d_η; the rest are *missing*

Value Search Estimators

$$C_\eta = Ad(X;\eta) + (1-A)\{1 - d(X;\eta)\}$$

Propensity scores:

- $\pi(X) = \text{pr}(A = 1|X)$ is the *propensity score* for treatment
- *Randomized trial:* $\pi(X)$ is *known*
- *Observational study:* Posit a model $\pi(X;\gamma)$ (e.g., logistic regression) and fit using (A_i, X_i), $i = 1, \dots, n \Longrightarrow \widehat{\gamma}$.
- *Propensity* of receiving treatment *consistent with* d_η

$$\begin{aligned}
\pi_c(X;\eta) &= \text{pr}(C_\eta = 1|X) = E(C_\eta|X) \\
&= E[Ad(X;\eta) + (1-A)\{1 - d(X;\eta)\}|X] \\
&= \pi(X)d(X;\eta) + \{1 - \pi(X)\}\{1 - d(X;\eta)\}
\end{aligned}$$

- Write $\pi_c(X;\eta,\gamma)$ with $\pi(X;\gamma)$

Value Search Estimators

Estimators for $V(d_\eta) = E\{Y^*(d_\eta)\}$: For fixed η

- *Inverse probability weighted* estimator

$$\widehat{V}_{IPWE}(d_\eta) = n^{-1} \sum_{i=1}^{n} \frac{C_{\eta,i} Y_i}{\pi_c(X_i; \eta, \widehat{\gamma})}.$$

- *Consistent* for $V(d_\eta)$ if $\pi(X;\gamma)$ (hence $\pi_c(X;\eta,\gamma)$) is *correct*

Value Search Estimators

Consistency:

$$
\begin{aligned}
E\left\{\frac{C_\eta Y}{\pi_c(X;\eta)}\right\} &= E\left\{\frac{C_\eta Y^*(d_\eta)}{\pi_c(X;\eta)}\right\} \\
&= E\left[E\left\{\left.\frac{C_\eta Y^*(d_\eta)}{\pi_c(X;\eta)}\right| Y^*(d_\eta), X\right\}\right] \\
&= E\left[\frac{E\{C_\eta | Y^*(d_\eta), X\} Y^*(d_\eta)}{\pi_c(X;\eta)}\right] \\
&= E\left[\frac{E\{C_\eta | X\} Y^*(d_\eta)}{\pi_c(X;\eta)}\right] \\
&= E\left\{\frac{\pi_c(X;\eta) Y^*(d_\eta)}{\pi_c(X;\eta)}\right\} = E\{Y^*(d_\eta)\}
\end{aligned}
$$

Estimators for $V(d_\eta) = E\{Y^*(d_\eta)\}$: For fixed η

- *Doubly robust augmented inverse probability weighted* estimator

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

$$m(X;\eta,\beta) = E\{Y^*(d_\eta)|X\} = Q(X,1;\beta)d(X;\eta)+Q(0,X;\beta)\{1-d(X;\eta)\}$$

and $Q(X,A;\beta)$ is a model for $E(Y|X,A)$

- *Consistent* if *either* $\pi(X,\gamma)$ or $Q(X,A;\beta)$ is *correct*

Augmented Estimator

Under MAR: $Y^*(d_\eta) \perp\!\!\!\perp C_\eta | X$

- If $\widehat{\gamma} \xrightarrow{p} \gamma^*$ and $\widehat{\beta} \xrightarrow{p} \beta^*$, this estimator $\xrightarrow{p}$

$$
\begin{aligned}
& E\left\{ \frac{C_\eta Y}{\pi_c(X;\eta,\gamma^*)} - \frac{C_\eta - \pi_c(X;\eta,\gamma^*)}{\pi_c(X;\eta,\gamma^*)} m(X;\eta,\beta^*) \right\} \\
&= E\left[Y^*(d_\eta) + \left\{ \frac{C_\eta - \pi_c(X;\eta,\gamma^*)}{\pi_c(X;\eta,\gamma^*)} \right\} \{Y^*(d_\eta) - m(X;\eta,\beta^*)\} \right] \\
&= E\{Y^*(d_\eta)\} + E\left[\left\{ \frac{C_\eta - \pi_c(X;\eta,\gamma^*)}{\pi_c(X;\eta,\gamma^*)} \right\} \{Y^*(d_\eta) - m(X;\eta,\beta^*)\} \right]
\end{aligned}
$$

- Hence the estimator is *consistent* if *either*
  - $\pi(X;\gamma^*) = \pi(X) \Rightarrow \pi_c(X;\eta,\gamma^*) = \pi_c(X;\eta)$ (*propensity correct*)
  - $Q(X,A;\beta^*) = Q(X,A) \Rightarrow m(X;\eta,\beta^*) = m(X;\eta)$ (*regression correct*)
  - *Double robustness*

Value Search Estimators

Result: Estimators $\widehat{\eta}^{opt}$ for η^{opt} obtained by *maximizing* $\widehat{V}_{IPWE}(d_\eta)$ or $\widehat{V}_{AIPWE}(d_\eta)$ in η

- Estimated optimal restricted regime $\widehat{d}_\eta^{opt}(x) = d(x; \widehat{\eta}^{opt})$
- *Non-smooth* functions of η; must use suitable *optimization techniques*
- Estimators for $V(d_\eta^{opt}) = E\{Y^*(d_\eta^{opt})\}$

$$\widehat{V}_{IPWE}(\widehat{d}^{opt}_{\eta,IPWE}) \quad \text{or} \quad \widehat{V}_{AIPWE}(\widehat{d}^{opt}_{\eta,AIPWE})$$

  Can calculate *standard errors*

- *Semiparametric theory*: *AIPWE* is *more efficient* than *IPWE* for estimating $V(d_\eta) = E\{Y^*(d_\eta)\}$
- $\Longrightarrow$ Estimating regimes based on *AIPWE* should be "*better*"

Empirical Studies

Extensive simulations: Qualitative conclusions

- Estimated optimal regime based on *regression* can achieve the true $E\{Y^*(d^{opt})\}$ *if* $Q(X, A; \beta)$ is *correctly specified*
- But performs *poorly* when $Q(X, A; \beta)$ is *misspecified*
- Estimated regimes based on *IPWE*(η) are *so-so* even if propensity model is *correct*
- Estimated regimes based on *AIPWE*(η) achieves the true $E\{Y^*(d^{opt})\}$ if $Q(X, A; \beta)$ is *correctly specified* even if the propensity model is *misspecified*
- And are *much better* than the regression estimator when $Q(X, A; \beta)$ is *misspecified*

Discussion

- Two approaches to estimation of optimal regimes for a *single decision point*
- *Regression methods* – estimate an optimal regime based on a *posited regression model*
- *Value search methods* – estimate an optimal treatment regime within a specified class by *maximizing the value*
- Robustness to *misspecification* (*AIPWE*)
- Both methods may be extended to *multiple decision points* (later)
- *Next:* Alternative *classification* perspective for single decision

References

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2012). A robust method for estimating optimal treatment regimes. *Biometrics* **68**, 1010–1018.

Generic classification situation:

- Z = *outcome*, *class*, *label*; here, $Z = \{0, 1\}$ (*binary*)
- X = vector of covariates, *features* taking values in $\mathcal{X}$, the *feature space*
- d is a *classifier:* $d : \mathcal{X} \rightarrow \{0, 1\}$
- $\mathcal{D}$ is a *family of classifiers*, e.g.,
  - *Hyperplanes* of the form

$$I(\eta_0 + \eta_1 X_1 + \eta_2 X_2 > 0)$$

  - *Rectangular regions* of the form

$$I(X_1 < a_1) + I(X_1 \geq a_1, X_2 < a_2)$$

Classification Methods

Generic classification problem:

- *Training set:* $(X_i, Z_i),\ i = 1, \ldots, n$
- *Find* classifier $d \in \mathcal{D}$ that minimizes
  - *Classification error*

$$\sum_{i=1}^{n} \{Z_i - d(X_i)\}^2$$

  - *Weighted classification error*

$$\sum_{i=1}^{n} w_i \{Z_i - d(X_i)\}^2$$

Classification Methods

Approaches:

- This problem has been studied extensively by *statisticians* and *computer scientists*
- *Machine learning* (*supervised* learning)
- Many methods and software are available
- *Recursive partitioning* (*CART*): Rectangular regions
- *Support vector machines:* Hyperplanes, etc.

Value Search Estimators, Revisited

Recall: Estimation of $d_\eta \in$ *restricted class* $\mathcal{D}_\eta$

$$\eta^{opt} = \arg\max_\eta V(d_\eta) = \arg\max_\eta E\{Y^*(d_\eta)\}$$

- Doubly robust *AIPWE*

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

$$\begin{aligned}
C_{\eta,i} &= A_i d(X_i;\eta) + (1-A_i)\{1-d(X_i;\eta)\} \\
\pi_c(X_i;\eta,\widehat{\gamma}) &= \pi(X_i;\widehat{\gamma})d(X_i;\eta) + \{1-\pi(X_i;\widehat{\gamma})\}\{1-d(X_i;\eta)\} \\
m(X_i;\eta,\widehat{\beta}) &= Q(X_i,1;\widehat{\beta})d(X_i;\eta) + Q(X_i,0,\widehat{\beta})\{1-d(X_i;\eta)\}
\end{aligned}$$

Value Search Estimators, Revisited

Algebra: $\widehat{V}_{AIPWE}(d_\eta)$ may be *rewritten* as

$$n^{-1}\sum_{i=1}^{n} d(X_i;\eta)\widehat{\mathcal{C}}(X_i) + \text{terms not involving } d$$

$$\begin{aligned}\widehat{\mathcal{C}}(X_i) &= \left\{\frac{A_i Y_i}{\pi(X_i,\widehat{\gamma})} - \frac{A_i - \pi(X_i,\widehat{\gamma})}{\pi(X_i;\widehat{\gamma})} Q(X_i,1;\widehat{\beta})\right\} \\ &- \left\{\frac{(1-A_i)Y_i}{1-\pi(X_i;\widehat{\gamma})} + \frac{A_i - \pi(X_i;\widehat{\gamma})}{1-\pi(X_i;\widehat{\gamma})} Q(X_i,0;\widehat{\beta})\right\},\end{aligned}$$

- The *contrast function* is

$$E\{\widehat{\mathcal{C}}(X_i)|X_i\} \approx \mathcal{C}(X_i) = Q(X_i,1) - Q(X_i,0)$$

Contrast Function

$$E\{\widehat{\mathcal{C}}(X_i)|X_i\} \approx \mathcal{C}(X_i) = Q(X_i, 1) - Q(X_i, 0)$$

Result: $\widehat{\mathcal{C}}(X_i)$ can be viewed as an *estimator* for the *contrast function* for subject i

- If we *knew* the functions $Q(X_i, 1)$ and $Q(X_i, 0)$, we should assign treatment

  $$I\{\mathcal{C}(X_i) > 0\} = I\{Q(X_i, 1) - Q(X_i, 0) > 0\}$$

  to patient i.

Classification Perspective

$$\widehat{\eta}^{opt} = \text{arg max}_{\eta} \sum_{i=1}^{n} d(X_i;\eta)\widehat{\mathcal{C}}(X_i)$$

Further algebra: Another *identity*

$$\begin{aligned} d(X_i;\eta)\widehat{\mathcal{C}}(X_i) &= -|\widehat{\mathcal{C}}(X_i)|[I\{\widehat{\mathcal{C}}(X_i) > 0\} - d(X_i;\eta)]^2 \\ &+ |\widehat{\mathcal{C}}(X_i)|I\{\widehat{\mathcal{C}}(X_i) > 0\} \end{aligned}$$

- Hence

$$\eta^{\hat{o}pt} = \text{arg min}_{\eta} \sum_{i=1}^{n} |\widehat{\mathcal{C}}(X_i)|\,[I\{\widehat{\mathcal{C}}(X_i) > 0\} - d(X_i;\eta)]^2,$$

Classification Perspective

$$\widehat{\eta}^{opt} = \text{arg min}_\eta \sum_{i=1}^{n} |\widehat{\mathcal{C}}(X_i)| \left[I\{\widehat{\mathcal{C}}(X_i) > 0\} - d(X_i; \eta) \right]^2$$

Alternative formulation: This can be viewed as a *weighted classification problem* with

- *Label* $I\{\widehat{\mathcal{C}}(X_i) > 0\}$
- *Classifier* $d(X_i; \eta)$
- *Weight* $|\widehat{\mathcal{C}}(X_i)|$

Discussion

- Estimation of optimal regime using "*off-the-shelf*" classification methods
- Estimated contrast functions constructed *independently* of class of regimes
- Form of estimated optimal regime *determined by classification method*
- Extension to *multiple decisions* ongoing

References

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

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107**, 1106–1118.

Recap: Multiple Decision Points

In general: K decision points

- *Baseline information* x_1, *intermediate information* x_k between decisions $k-1$ and k, $k = 2, \ldots, K$
- Set of *treatment options* at decision k $a_k \in \mathcal{A}_k$
- *Accrued information* $h_1 = x_1 \in \mathcal{H}_1$,

$$h_k = \{x_1, a_1, x_2, a_2, \ldots, x_{k-1}, a_{k-1}, x_k\} \in \mathcal{H}_k, \quad k = 2, \ldots, K$$

- *Decision rules* $d_1(h_1), d_2(h_2), \ldots, d_K(h_K)$, $d_k : \mathcal{H}_k \rightarrow \mathcal{A}_k$
- *Dynamic treatment regime* $d = (d_1, d_2, \ldots, d_K)$
- $\mathcal{D}$ is the set of *all possible* K-decision regimes

Recap: Optimal Regime for Multiple Decisions

Optimal regime: $d^{opt} \in \mathcal{D}$ such that a patient with *baseline information* $X_1 = x_1$ who receives *all K treatments* according to d^{opt} has *expected outcome as large as possible*

Potential outcomes under a regime $d \in \mathcal{D}$:

- *Baseline information* X_1, *potential outcomes*

$$X_2^*(d_1), \ldots, X_K^*(\bar{d}_{K-1}), Y^*(d)$$

d^{opt} satisfies:

- $E\{Y^*(d)\} \leq E\{Y^*(d^{opt})\}$ for all $d \in \mathcal{D}$
- $E\{Y^*(d) | X_1 = x_1\} \leq E\{Y^*(d^{opt}) | X_1 = x_1\}$ for all $d \in \mathcal{D}$ and $x_1 \in \mathcal{H}_1$

K decisions: *Data*

$$(X_{1i}, A_{1i}, X_{2i}, A_{2i}, \ldots, X_{(K-1)i}, A_{(K-1)i}, X_{Ki}, A_{Ki}, Y_i), \quad i = 1, \ldots, n$$

- X_{1i} = *Baseline information* observed on subject i
- X_{ki}, $k = 2, \ldots, K$ = *intermediate information* between decisions $k-1$ and k on subject i
- A_{ki}, $k = 1, \ldots, K$ = *observed treatment* actually received by subject i at decision k
- H_i = *accrued information* for subject i up to decision k

$$H_{1i} = X_{1i}, \quad H_{ki} = (X_{1i}, A_{1i}, \ldots, A_{(k-1)i}, X_{ki}), \ k = 2, \ldots, K$$

- Y_i = *observed outcome* for subject i; can be *ascertained after* decision K or can be a *function* of $X_{2i}, \ldots, X_{Ki}$

Estimation of Optimal Treatment Regimes

Goal, restated: Estimate d^{opt} satisfying

- $E\{Y^*(d)\} \leq E\{Y^*(d^{opt})\}$ for all $d \in \mathcal{D}$
- $E\{Y^*(d)|X_1 = x_1\} \leq E\{Y^*(d^{opt})|X_1 = x_1\}$ for all $d \in \mathcal{D}$ and $x_1 \in \mathcal{H}_1$

Sequential randomization assumption: *Data* from

- A *SMART*
- A *fabulous* longitudinal observational study

For definiteness: Take $K = 2$ and $\mathcal{A}_k = \{0, 1\}$, $k = 1, 2$

- Recall *accrued information*

$$H_{1i} = X_{1i}, \quad H_{2i} = (X_{1i}, A_{1i}, X_{2i})$$

Characterizing the Optimal Regime

Optimal regime d^{opt}: Follows from *backward induction* (*dynamic programming*)

- Formally in terms of *potential outcomes*
- *Sequential randomization* assumption allows equivalent expressions in terms of *observed data* (X_1, A_1, X_2, A_2, Y) (as for *single decision* and *no unmeasured confounders*)

Characterizing the Optimal Regime

Optimal regime d^{opt}: *Backward induction*

- *Decision 2:* $Q_2(H_2, A_2) = E(Y|H_2, A_2)$

$$d_2^{opt}(h_2) = I\{Q_2(h_2, 1) > Q_2(h_2, 0)\} = \text{arg max}_{a_2=\{0,1\}} Q_2(h_2, a_2)$$

$$\widetilde{Y}_2(h_2) = \max\{Q_2(h_2, 0), Q_2(h_2, 1)\}$$

- *Decision 1:* $Q_1(H_1, A_1) = E\{\widetilde{Y}_2(H_2)|H_1, A_1\}$

$$d_1^{opt}(h_1) = I\{Q_1(h_1, 1) > Q_1(h_1, 0)]\} = \text{arg max}_{a_1=\{0,1\}} Q_1(h_1, a_1)$$

$$\widetilde{Y}_1(h_1) = \max\{Q_1(h_1, 0), Q_1(h_1, 1)\}$$

- $d^{opt} = (d_1^{opt}, d_2^{opt})$
- The *value* of d^{opt} is $V(d^{opt}) = E\{\widetilde{Y}_1(H_1)\}$
- $\widetilde{Y}_2(h_2)$ and $\widetilde{Y}_1(h_1)$ are referred to as the *value functions*

Q-Learning

Q-learning: May be thought of as a generalization of the *regression estimator* to *sequential decisions*

- *Reinforcement learning* in computer science
- Posit models for the "*Q-functions*"
- Involves some *complications* not present in the *single decision* case

Q-Learning

Estimation of d^{opt}:

- *Decision 2: Posit and fit a model* $Q_2(H_2, A_2; \beta_2)$ by regressing Y on H_2, A_2 (e.g., least squares) and *estimate*

$$\widehat{d}^{opt}_{Q,2}(h_2) = I\{Q_2(h_2, 1; \widehat{\beta}_2) > Q_2(h_2, 0; \widehat{\beta}_2)\}$$

- For each i, form "*predicted value*"

$$\widehat{\widetilde{Y}}_{2i} = \widetilde{Y}_{2i}(H_{2i}; \widehat{\beta}_2) = \max\{Q_2(H_{2i}, 0; \widehat{\beta}_2), Q_2(H_{2i}, 1; \widehat{\beta}_2)\}$$

- *Decision 1: Posit and fit a model* $Q_1(H_1, A_1; \beta_1)$ by regressing $\widehat{\widetilde{Y}}_2$ on H_1, A_1 (e.g., least squares) and *estimate*

$$\widehat{d}^{opt}_{Q,1}(h_1) = I\{Q_1(h_1, 1; \widehat{\beta}_1) > Q_1(h_1, 0; \widehat{\beta}_1)\}$$

- *Estimated regime* $\widehat{d}^{opt}_{Q} = (\widehat{d}^{opt}_{Q,1}, \widehat{d}^{opt}_{Q,2})$

Q-Learning

Issues and challenges:

- Regardless, as in the *single decision* case, *incorrect model specification* will impact quality of estimation of d^{opt}
- Modeling at decisions $K-1,\ldots,1$ challenging due to need to model *max*
- More *flexible models* for Q-functions can be used
- Because of *nonsmooth max operator*, standard asymptotic theory is *invalid*
- Considerable *current research*

Value Search Methods

Generalization to $K > 1$:

- Consider *directly* a *restricted class of regimes* $\mathcal{D}_\eta$ with elements $d_\eta = (d_{\eta,1}, \ldots, d_{\eta,K})$; at decision k

$$d_{\eta,k}(h_k) = d_k(h_k; \eta_k)$$

- Based on *cost*, *feasibility*, *interpretability* at each decision
- *Optimal restricted regime* d_η^{opt}

$$\eta^{opt} = \text{arg max}_\eta E\{Y^*(d_\eta)\} = \text{arg max}_\eta V(d_\eta)$$

- *Estimator* $\widehat{V}(d_\eta)$ for fixed η; *maximize* in η to obtain $\widehat{\eta}^{opt}$
- *Required:* A "*good*" $\widehat{V}(d_\eta)$

Extend missing data analogy to monotone dropout: $K = 2$

- "*Full data*"

$$\{X_1, X_2^*(d_\eta), Y^*(d_\eta)\}$$

- Define *η-regime consistency indicator* C_η
- $C_\eta = \infty$: If a patient's *actual treatments* A_1, A_2 are *all consistent with* following d_η, then

$$(X_1, X_2, Y) = \{X_1, X_2^*(d_\eta), Y^*(d_\eta)\}$$

- $C_\eta = 2$: If *actual* A_1 is *consistent with* following d_η but A_2 is *not*, then

$$(X_1, X_2) = \{X_1, X_2^*(d_\eta)\}$$

  but $Y^*(d_\eta)$ is "*missing*" ("*dropout*" before decision 2)
- $C_\eta = 1$: If *neither* of A_1, A_2 is *consistent with* following d_η, *both* $X_2^*(d_\eta), Y^*(d_\eta)$ are "*missing*" ("*dropout*" before decision 1)

Value Search Methods

Propensity scores: At decision $k = 1, \dots, K$

$$\pi_k(H_k) = \text{pr}(A_k = 1 | H_k)$$

- *Randomized trial (SMART):* $\pi_k(h_k)$ is *known*
- *Observational study:* Posit and fit models $\pi_k(h_k; \gamma_k)$
- Can express *propensities* of receiving treatment *consistent with d_η through decision k* in terms of $\pi_k(h_k)$

Result: Can develop *IPWE* and doubly-robust *AIPWE* estimators for $V(d_\eta)$ in terms of C_η and $\pi_k(h_k)$

Augmented Inverse Probability Weighted Estimators

$$
\begin{aligned}
&\widehat{V}_{AIPWE}(d_\eta) \\
&\quad = \sum_{i=1}^{n} \left(\frac{I(C_{\eta,i} = \infty) Y_i}{\prod_{k=1}^{K} [\pi_k(H_{ki}) d_{\eta,k}(H_{ki}) + \{1 - \pi_k(H_{ki})\}\{1 - d_{\eta,k}(H_{ki})\}]} \right) \\
&\quad + \text{ augmentation terms}
\end{aligned}
$$

Value Search Methods

Issues and challenges:

- As for $K = 1$, is *nonstandard optimization problem*
- *IPWE* (leading term for *AIPWE*) involves *only* subjects with $C_\eta = \infty$ (*consistent* with following regime for *all K decisions*)
- May become *infeasible for* $K > 3$
- *Simulation evidence:* Performance comparable to Q-learning with *correct models*; *AIPWE* is *robust* to model model misspecification while Q-learning is *not*

Discussion

- Two classes of methods for estimation of optimal regimes for *multiple decision points*
- *Q- and A-learning* (*sequential regression* methods) – estimate an optimal regime based on *sequential* posited regression models
- Potential for *model misspecification* is high
- *Value search methods* – robustness to *misspecification*
- *Limitation* to small K due to need for "*regime consistency*"

References

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

Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in Medicine* **24**, 1455–1481.

Schulte, P. J., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2014). Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, in press.

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. (2013). Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. *Biometrika* **100**, 681–694.

Closing Remarks

- Estimation of optimal treatment regimes is a *wide open* area of research
- *SMARTs* are the "*gold standard*" data source for estimation of optimal regimes
- *Design considerations* for SMARTs?
- *High-dimensional* covariate information? Regression *model selection*?
- "*Black box*" vs. *restricted class* of regimes?
- *Inference*?
- Balancing *multiple outcomes* (e.g., *efficacy* vs. *toxicity*)?
- ...

Thought Leaders

2013 MacArthur Fellow *Susan Murphy* and *Jamie Robins*