Precision Medicine: Lecture 06
Single Decision DTR Extensions

Michael R. Kosorok,
Nikki L. B. Freeman and Owen E. Leete

Department of Biostatistics
Gillings School of Global Public Health
University of North Carolina at Chapel Hill

Fall, 2021

Outline

Motivation

- Dose finding plays an important role in clinical trials, which aim to assess drug toxicity, identify maximum tolerated doses for safety, and determine drug efficacy
  - Overdosing can increase the risk of side effects
  - Underdosing can diminish the therapeutic effects of the drug
- Using individualized dose levels can help improve clinical practice and increase a patient's compliance
- Example: Warfarin, a common drug for the prevention of thrombosis and thromboembolism, is administered in varying doses, ranging from 10mg to 100mg per week depending on the patient's individual clinical and genetic factors

Notation

- Dose assignment $A \in \mathcal{A}$
  - We allow A to be continuous
  - $\mathcal{A}$ is a bounded interval (a safe dose range)
- Patient-level covariates $X = (X_1, X_2, \ldots, X_d)^T \in \mathcal{X}$
- The reward $R(a)$ is the outcome that would be observed if dose level a were given
- We assume the stable unit treatment value assumption (SUTVA), $R = \sum_a I(A = a)R(a)$
- We assume that the data come from a randomized dose trial design
  - Each patient receives a dose level randomly chosen from a continuous distribution within the range of safe doses

Value function

- Under a randomized dose design, since A is independent of $R(a)$ given X, we obtain the value function

$$\mathcal{V}(f) = E[R(f(X))] = E_X[E\{R|A = f(X), X\}]$$

  thus the optimal rule $f_{\text{opt}} = \text{argmax}_f \, V(f)$
- We can assume that R is positive
  - f_{opt} is invariant to a shift $(R + g(X))$ or scaling (cR) of the reward
- For any IDR f, V(f) can be estimated consistently using the mean of the reward among subjects whose dose levels are the same as $f(X)$

Complications

- Recall that OWL was designed to minimize the quantity

$$\mathcal{R}_n(f) = n^{-1} \sum_{i=1}^{n} R_i I(A_i \neq f(X_i))/\pi(A_i, X_i)$$

- Direct adaption of OWL to the dose-finding setting is not feasible when A is continuous
  - The probability $\pi(A|X)$ is always zero since A is continuous
  - Only a few subjects satisfy $A_i = f(X_i)$, so this approximation is very unstable
- What if we modify $\mathcal{R}_n(f)$ to resolve these two issues by replacing $\pi(A_i|X_i)$ with the density of $A_i|X_i$ and replace $I(A_i \neq f(X_i))$ with a smooth loss?
- The estimated dose rule resulting from the modified $\mathcal{R}_n(f)$ is not consistent for the optimal IDR

Extending the value function

- First, assuming $E[R|A = a, X]$ to be continuous in a, we note that

$$\lim_{\phi \to 0^+} \frac{E[RI[A \in (a - \phi, a + \phi)]/p(A|X)|X]}{2\phi} = E[R|A = a, X]$$

  where $p(a|X)$ is the conditional density of $A = a$ given X

- Thus

$$\lim_{\phi \to 0^+} E\left\{ \frac{RI[A \in (f(X) - \phi, f(X) + \phi)]}{2\phi p(A|X)} \right\}$$
$$= E_X[E\{R|A = f(X), X\}] = \mathcal{V}(f)$$

Extending the value function, cont.

- If we let

$$\widetilde{\mathcal{V}}_\phi(f) = E\left\{\frac{RI[A \in (f(X)-\phi, f(X)+\phi)]}{2\phi p(A|X)}\right\}$$

  then $\widetilde{\mathcal{V}}_\phi(f)$ approximates $\mathcal{V}(f)$ when ϕ is sufficiently small
- Hence, an IDR maximizing $\widetilde{\mathcal{V}}_\phi(f)$, or equivalently, minimizing

$$E\left[\frac{R}{2\phi p(A|X)}\right] - \widetilde{\mathcal{V}}_\phi(f) = E\left[\frac{RI(|A-f(X)| > \phi)}{2\phi p(A|X)}\right]$$

  will be close to the optimal IDR

Surrogate loss function

- The zero-one loss $I(|A - f(X)| > \phi)$ often causes difficulty when optimizing using empirical data
- We can replace the zero-one loss with a continuous surrogate loss, specifically

$$\ell_\phi(A - f(X)) = \min\left(\frac{|A - f(X)|}{\phi}, 1\right)$$

- Note that $\ell_\phi(x)$ is the difference of two convex functions $|x|/\phi$ and $(|x|/\phi - 1)_+$
  - This allows us to use a difference of convex (DC) optimization algorithm

Surrogate loss function, cont.

- With the surrogate loss, we can define the ϕ-risk

$$\mathcal{R}_{\phi_n}(f) = E\left[\frac{R\ell_{\phi_n}[A - f(X)]}{\phi_n p(A|X)}\right]$$

- With data from a randomized dose trial we can minimize the empirical version of $\mathcal{R}_{\phi_n}(f)$

$$\widehat{\mathcal{R}}_{\phi_n}(f) = n^{-1}\sum_{i=1}^{n}\frac{R_i\ell_{\phi_n}[A_i - f(X_i)]}{\phi_n p(A_i|X_i)}$$

- To prevent overfitting, we penalize the complexity of $f(X)$

$$\min_f\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{R_i\ell_{\phi_n}[A_i - f(X_i)]}{\phi_n p(A_i|X_i)} + \lambda_n\|f\|^2\right\} \qquad (1)$$

Tuning parameter selection

- There are two tuning parameters in equation (1)
- Ideally, we would evaluate the prediction performance associated with each tuning parameter with an unbiased estimate for the true value function
- The true value function, $\mathcal{V}(f)$, cannot be well estimated with continuous dose levels, instead we use an approximate estimate $\mathcal{V}_\epsilon(f)$ where $\epsilon = 0.01$
- Divide the data into training and tuning sets
  - λ_n is selected using cross-validation with the training data only
  - The predicted value function is evaluated using the empirical estimate of $V_\epsilon(f_{\phi_n})$ using the tuning data
  - The optimal choice of ϕ_n is the one maximizing the average of the predicted values

Learning a linear IDR

- Consider $f(X) = X^T\mathbf{w} + b$, let $\mathbf{\Theta} = (\mathbf{w}^T, b)^T$. We can formulate the objective function as

$$S(\mathbf{\Theta}) = \frac{\lambda_n}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n\phi_n}\sum_{i=1}^{n} R_i \min\left(\frac{|A_i - X_i^T\mathbf{w} + b|}{\phi_n p(A_i|X_i)}, 1\right)$$

- Express the objective function as the difference of two convex functions, $S(\mathbf{\Theta}) = S_1(\mathbf{\Theta}) - S_2(\mathbf{\Theta})$

$$S_1(\mathbf{\Theta}) = \frac{\lambda_n}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n\phi_n}\sum_{i=1}^{n} R_i \frac{|A_i - X_i^T\mathbf{w} + b|}{\phi_n p(A_i|X_i)}$$

$$S_2(\mathbf{\Theta}) = \frac{1}{n\phi_n}\sum_{i=1}^{n} R_i \left(\frac{|A_i - X_i^T\mathbf{w} + b|}{\phi_n p(A_i|X_i)} - 1\right)_+$$

Learning a linear IDR, cont.

- The DC algorithm is an iterative sequence of convex minimization problems for solving the original nonconvex minimization problem
- Initialize $\mathbf{\Theta}^0$, then repeatedly update $\mathbf{\Theta}$ via

$$\mathbf{\Theta}^{t+1} = \underset{\mathbf{\Theta}}{\operatorname{argmin}}(S_1(\mathbf{\Theta}) - [\nabla S_2(\mathbf{\Theta}^t)]^T(\mathbf{\Theta} - \mathbf{\Theta}^t))$$

- Finding the update involves SVM like optimization
- We can also learn a non-linear IDR using the "kernel trick"

Extension to observational studies

- We assumed that the training data was from a randomized dose trial
- What if the data come from an observational study?
- Similar to causal inference using observational data, we make the no unobserved confounders assumption (i.e. A is independent of the potential outcomes $R(a)$ given X)
- Under this assumption, the proposed approach remains valid except that the density $p(a|X)$ needs to be estimated by the observed data

Theoretical results

- For any measurable function f,

$$\mathcal{V}(f_{\phi_n}) - \mathcal{V}(f) \leq C\phi_n$$

  where C controls how fast $E[R|A = a]$ changes with respect to a
- Under ideal circumstances (f_{opt} is smooth)

$$\mathcal{V}(f_{\text{opt}}) - \mathcal{V}(\widehat{f}) = O_p\left(\left(\frac{1}{n}\right)^{\frac{1}{4}}\right)$$

- The optimal convergence rate of OWL for binary outcomes is closer to n^{-1}, but the efficiency of this method is much lower due to the continuous nature of the dose levels.

Conclusions

- The proposed method appears to be more effective than alternative approaches especially when the training sample size is relatively small
- Compared to regression-based methods, the proposed method is more robust to model specification of the reward
- When the training data comes from an observational study, this method can yield a less efficient rule
  - A doubly robust version of the estimator could help improve efficiency

Outline

Personalized dose finding

Outcome weighted learning with censored data

Feature construction for dynamic treatment regimes

Balanced policy evaluation and learning

Estimating optimal treatment regimes using lists

Motivation

- In clinical trials, right censored survival data are frequently observed as primary outcomes
- Outcome weighted learning has been adapted to right censored data using inverse censoring weighted (ICW) OWL and doubly robust (DR) OWL
- Both the ICW and DR versions require semiparametric estimation of the conditional censoring probability given the patient characteristics and treatment
- The DR estimator additionally involves semiparametric estimation of the conditional failure time expectation
- If either or both models are misspecified, these methods can produce results that are unstable numerically

Data and notation

- Observed patient-level covariate vector, $X \in \mathcal{X}^d$
- Binary treatment indicator, $A \in \{-1, +1\}$
- True survival time $\widetilde{T}$
- In most studies the true survival time is not observable for all subjects beause there is a maximum follow-up time τ
- We consider a truncated version of the survival time, $T = \min(\widetilde{T}, \tau)$
- we wish to identify a treatment rule $\mathcal{D}$, which maximizes the expected reward.
- In the survival outcome setting, we use $R = T$ or $\log(T)$

Value function under right censoring

- Consider a censoring time C that is independent of T given (X, A)
- Rather than the failure time T we observe $Y = \min(T, C)$, and the censoring indicator $\delta = I(T \leq C)$
- Recall the value function for an ITR $\mathcal{D}$

$$\mathcal{V}(\mathcal{D}) = E^{\mathcal{D}}(R) = E\left[\frac{RI[A = \mathcal{D}(X)]}{\pi(A, X)}\right]$$

- In the right censored data setting we cannot directly estimate $\mathcal{V}(\mathcal{D})$ because $R = T$ is not observed for all subjects

Inverse censoring weights

- Current extensions of OWL to right censored data use weights to account for the censored observations
- Let $S(Y|A,X)$ be the survival function of the censoring distribution, this quantity describes the probability of observing a failure at time Y conditional on A and X
- The ICW approach uses only the observed failure times, but gives increased weight to subjects that are more likely to be censored

$$\mathcal{V}_{\text{ICW}}(\mathcal{D}) = E\left[\frac{I(\delta = 1)YI[A = \mathcal{D}(X)]}{\pi(A,X)S(Y|A,X)}\right]$$

- The denominator, $\pi(A,X)S(Y|A,X)$, can be small, which leads to numerical instability

Expected survival time

- Rather than discard the censored observations, we can replace the missing data
- Is there a sensible replacement which maintains as close as possible the same value function as if T were fully observed?
- The first approach is to obtain a nonparametric estimated conditional expectation $\widehat{E}(T|X, A)$
- Letting $R_1 = E(T|X, A)$ and bringing the expectation of T inside, we have

$$E\left[\frac{T\mathbb{I}[A = \mathcal{D}(X)]}{\pi(A, X)}\right] = E\left[\frac{R_1\mathbb{I}[A = \mathcal{D}(X)]}{\pi(A, X)}\right] \quad (2)$$

- This approach replaces both the failure and censoring times with their expected values

Conditional expected survival time

- Another approach is to replace only the censored observations conditioning on the observed data
- The conditional expectation of T, given Y and δ, can be written as

$$\begin{aligned} R_2 &= E(T|X, A, Y, \delta) \\ &= \mathbb{I}(\delta = 1)Y + \mathbb{I}(\delta = 0)E(T|X, A, Y, T > Y) \end{aligned}$$

- With the information of $Y = y$ given, and knowing that $\delta = 0$, the conditional distribution of T is defined on the interval $(y, \tau]$
- As with R_1, we can write the value function in terms of R_2

$$E\left[\frac{T\mathbb{I}[A = \mathcal{D}(X)]}{\pi(A, X)}\right] = E\left[\frac{R_2\mathbb{I}[A = \mathcal{D}(X)]}{\pi(A, X)}\right] \qquad (3)$$

Estimating the survival time distribution

- By replacing all observations with an estimate of the mean, using R_1 appears to be very similar to regression based methods
- Because of this connection, it is reasonable to assume that using R_1 would lead to estimates that are sensitive to misspecification of the mean model
- To avoid misspecification, the authors recommend using a nonparametric random forest based model to estimate R_1 and R_2
- Nonparametric models are often not very efficient, but combining the nonparametric estimates with the OWL framework can lead to good performance

Estimating the survival time distribution, cont.

- The authors recommend estimating R_1 and R_2 with recursively imputed survival trees (RIST) (Zhu & Kosorok 2012)
- It is worth noting that the conditional expectation of T defined in R_2 shares the same logical underpinnings as the imputation step in RIST
- The imputation step in RIST replaces censored observations with a random draw from the conditional survival distribution, while R_2 is the mean of the conditional survival distribution
- Note that the above arguments remain unchanged if we replace T, C and Y with $\log(T)$, $\log(C)$, and $\log(Y)$

Tree based outcome weighted learning

- For rewards R_1 and R_2, respectively, we solve for the optimal decision $\mathcal{D}^*$ by minimizing

$$n^{-1}\sum_{i=1}^{n}\frac{\widehat{E}(T_i|A_i,X_i)\mathbb{I}(A_i\neq\mathcal{D}(X_i))}{\pi(A_i,X_i)} \quad \text{or}$$

$$n^{-1}\sum_{i=1}^{n}\frac{[\delta Y+(1-\delta)\widehat{E}(T|X,A,Y,T>Y)]\mathbb{I}(A_i\neq\mathcal{D}(X_i))}{\pi(A_i,X_i)}$$

  respectively

- As with OWL, we replace the 0-1 loss $\mathbb{I}(A_i\neq\mathcal{D}(X_i))$ with the hinge loss $\phi\{A_i f(X_i)\}$, regularize the complexity of f with $\lambda_n\|f\|^2$, and use SVM methods for optimization

Tree based OWL algorithm

Step 1: Use $\{(X_i^T, A_i, A_i X_i^T)^T, Y_i, \delta_i\}_{i=1}^n$ to fit recursively imputed survival trees. Obtain the estimation $\widehat{E}(T_i|A_i, X_i)$ for reward R_1 or the estimation $\widehat{E}(T_i|X_i, A_i, T_i > Y_i, Y_i)$ for reward R_2.

Step 2: Let the weights W_i be either $\widehat{E}(T_i|A_i, X_i)$ or $\delta_i Y_i + (1-\delta_i)\widehat{E}(T_i|A_i, X_i, T_i > Y_i, Y_i)$, depending on which of the two proposed approaches is used. Minimize the following weighted misclassification error:

$$\widehat{f}(x) = \underset{f}{\text{argmax}} \sum_{i=1}^{n} W_i \frac{\phi\{A_i f(X_i)\}}{\pi(A_i, X_i)} + \lambda_n \|f\|^2$$

Step 3: Output the estimated optimal treatment rule $\widehat{\mathcal{D}}(x) = \text{sign}\{\widehat{f}(X)\}$

Theoretical results

- Because of equations (2) and (3) showing the consistency of tree based outcome weighted learning for R_1 and R_2 only requires the consistency of $\widehat{E}(T_i|A_i, X_i)$ and $\widehat{E}(T_i|X_i, A_i, T_i > Y_i, Y_i)$ respectively
- The authors showed that tree-based survival models are consistent under general settings with restrictions only on the splitting rules

Discussion

- Tree based OWL gives better performance than existing approaches
- Numerical experiments demonstrate that the difference in performance between R_1 and R_2 is small
- The tree based OWL method was designed to maximize the expected failure time, but this choice could be more flexible (i.e. maximize the median survival time or a different quantile)
  - Under the presented framework, this is achievable by replacing the censored observations with a suitable estimate of the quantile

Outline

Motivation

- Most existing methods for estimating optimal individualized treatment regimes (ITRs) assume that each patient is measured at the same, fixed time points and that the patient measurements are made without error.
- In practice, however, patient measurements are often sparsely observed and irregularly spaced (e.g. at clinic visits).
- A standard approach to dealing with longitudinal measurements is to summarise them as a scalar (e.g. last observed measurement)
- Laber and Staicu (2018) propose a framework for treating subject-specific longitudinal measures as a realization of a stochastic process observed with error and using functionals of this latent process with outcome models to estimate ITRs.

Data

- Observed data $\mathcal{D}_n = \{(\mathbf{X}_i, W_i(\mathbf{T}_i), A_i, Y_i)\}_{i=1}^n$ are n iid copies of a trajectory $\{\mathbf{X}, W(\mathbf{T}), A, Y\}$.
  - $\mathbf{X} \in \mathbb{R}^p$ denotes pretreatment subject covariate information.
  - $W(\mathbf{T}) = \{W(T_1), \ldots, W(T_M)\}$ denotes M pretreatment proxy measurements taken at times $\mathbf{T} = (T_1, \ldots, T_M)$.
  - $A \in \{-1, 1\}$ denotes treatment assigned.
  - $Y \in \mathbb{R}$ denotes an outcome coded so that higher values are better.
- Both M and $\mathbf{T}$ are treated as random variables as the number and timing of observations varies across subjects.
- Observation times $\mathbf{T}$ are allowed to vary in number, be sparse, and irregularly spaced.

Notation

- An ITR π is a map, $\pi : \text{dom}\mathbf{X} \times \text{dom}W(\mathbf{T}) \mapsto \text{dom}A$.
- The potential outcome under ITR π is $Y^*(\pi) = \sum_{a \in \{-1,1\}} Y^*(a) I(\pi\{\mathbf{X}, W(\mathbf{T})\} = a)$.
- The optimal ITR, π^{opt} satisfies $\mathbb{E}Y^*(\pi^{\text{opt}}) \geq \mathbb{E}Y^*(\pi)$ for all π.

Regression-based methods aren't ideal in this setting

- Define

$$Q\{\mathbf{x}, w(\mathbf{t}), a\} = \mathbb{E}\{Y|\mathbf{X} = x, W(\mathbf{T}) = w(\mathbf{t}), A = a\}. \quad (4)$$

- Under our usual causal assumptions (i.e. consistency, positivity, and ignorability), we know from previous lectures that

$$\pi^{\text{opt}}\{\mathbf{x}, w(\mathbf{t})\} = \underset{a \in \{-1,1\}}{\text{argmax}}\ Q\{\mathbf{x}, w(\mathbf{t}), a\}$$

- Postulating a working model for $Q\{\mathbf{x}, w(\mathbf{t}), a\}$ when $w(\mathbf{t})$ is sparsely observed and irregularly spaced is difficult.

Regression-based methods aren't ideal (cont.)

- A common strategy to overcome this is to use a scalar summary $s\{w(\mathbf{t})\}$ and assume

$$Q\{\mathbf{x}, w(\mathbf{t}), a\} = Q\{\mathbf{x}, s\{w(\mathbf{t})\}, a\}. \tag{5}$$

- Under our usual causal assumptions, ignorability means $\{Y^*(1), Y^*(0)\}$ is independent of A conditional on $\mathbf{X}$ and $W(\mathbf{T})$.
- Assuming (5) implicitly requires a stronger causal assumption: $\{Y^*(1), Y^*(0)\}$ is independent of A conditional on $\mathbf{X}$ and $s\{W(\mathbf{T})\}$.
- The method proposed by Laber and Staicu (2018) uses $w(\mathbf{t})$ directly in the model and is thereby more consistent with our usual assumption of ignorability.

Longitudinal measurements as a noisy, latent process

Treating patient longitudinal information as sparse functional data, we want to derive the optimal ITR that does not require an ad hoc summary of $w(\mathbf{t})$.

- Suppose that $0 \leq T_1 < \cdots < T_M \leq 1$.
- Assume that $W(t) = Z(t) + \epsilon(t)$ for all $t \in [0, 1]$, where
  - $Z(\cdot)$ is a square integrable latent process with smooth mean and covariance functions
  - $\epsilon(\cdot)$ is mean-zero white-noise process. (We will see the importance of this choice later.)

Using this representation of patient longitudinal measurements, in the following slides we will consider a general linear model for $\mathbb{E}\{Y|\mathbf{X}, Z(\cdot), A\}$ that is linear in $\mathbf{X}$, A, and $Z(\mathbf{T})$.

- Note that a hybrid model that combines linear effects for $\mathbf{X}$ and A and a linear effect of $f\{Z(\cdot); \eta\}$ for a parametric functional $f(\cdot; \eta)$ indexed by unknown parameters η is also explored in Laber and Staicu (2018), but we will focus on the linear working model for ease of exposition.

Linear working model for $\mathbb{E}\{Y|\mathbf{X}, Z(\cdot), A\}$

Assume a linear model of the form

$$\begin{aligned}&\mathbb{E}\{Y|\mathbf{X} = \mathbf{x}, Z(\cdot) = z(\cdot), A = a\} \\ &= \mathbf{x}^\intercal\alpha^* + \int_0^1 z(t)\beta^*(t)dt + a\left\{\mathbf{x}^\intercal\delta^* + \int_0^1 z(t)\gamma^*(t)dt\right\} \qquad (6)\end{aligned}$$

- If $\beta^*(\cdot)$ and $\gamma^*(\cdot)$ known and $Z(\cdot)$ observed, (6) would correspond to a linear model.
- Now we are ready to consider how to estimate this model.

Estimation of the linear working model

Estimation of the linear working model can be summarized as follows

- Use orthogonal basis function expansions for both the latent process and unknown coefficient functions.
- The basis function expansions are infinite sums; reduce those to finite sums by truncation.
- You will end up with something that looks linear—just one "missing piece".
- Use functional principle components to get the missing piece.
- Use regression to estimate the parameters in the Q function.
- The optimal decision rule is the decision rule that maximizes the predicted Q function.

The next few slides give a few more details on each of these steps.

Estimation of the linear working model

- Let the spectral decomposition of the covariance function $G(t, t') = \text{cov}\{Z(t), Z(t')\}$ be $G(t, t') = \sum_{k \geq 1} \lambda_k \phi_k(t) \phi_k(t')$, where $\{\lambda_k, \phi_k(\cdot)\}_{k \geq 1}$ are the pairs of eigenvalues/eigenfunctions and $\lambda_1 > \lambda_2 > \cdots \geq 0$.
- Using the Karhunen-Loéve expansion, $Z(\cdot)$ can be written as

$$Z(t) = \mu(t) + \sum_{k \geq 1} \xi_k \phi_k(t)$$

where $\xi \equiv \int_0^1 \{Z(t) - \mu(t)\} \phi_k(t) dt$ are functional principle component scores which have mean zero, variance λ_k, and are mutually uncorrelated.

Estimation of the linear working model

- Using the eigenfunctions, we can represent $\beta^*(\cdot)$ and $\gamma^*(\cdot)$ as

$$\beta^*(t) = \sum_{k \geq 1} \beta_k^* \phi_k(t)$$

$$\gamma^*(t) = \sum_{k \geq 1} \gamma_k^* \phi_k(t)$$

where $\beta_k^* \equiv \int_0^1 \beta(t)\phi_k(t)dt$ and $\gamma_k^* \equiv \int_0^1 \gamma(t)\phi_k(t)dt$ are unknown basis coefficients.

- Now we can rewrite the linear working model (6) as

$$\begin{aligned} &\mathbb{E}\{Y|\mathbf{X} = \mathbf{x}, Z(\cdot) = z(\cdot), A = a\} \\ &= \mathbf{x}^\intercal \alpha^* + \sum_{k \geq 1} \xi_k \beta_k^* + a\left(\mathbf{x}^\intercal \delta^* + \sum_{k \geq 1} \xi_k \gamma_k^*\right) \end{aligned}$$

Estimation of the linear working model

- Under some mild assumptions we can obtain an expression for $Q\{\mathbf{x}, w(\mathbf{t}), a\}$:

$$Q\{\mathbf{x}, w(\mathbf{t}), a\} = \mathbf{x}^{\mathsf{T}}\alpha^* + \sum_{k\geq 1}\beta_k^*\mathbb{E}\{\xi_k|W(\mathbf{T}) = w(\mathbf{t})\}$$
$$a\left[\mathbf{x}^{\mathsf{T}}\delta^* + \sum_{k\geq 1}\gamma_k^*\mathbb{E}\{\xi_k|W(\mathbf{T}) = w(\mathbf{t})\}\right]. \quad (7)$$

- Let $\ell_k\{w(\mathbf{t})\} \equiv \mathbb{E}\{\xi_k|W(\mathbf{T}) = w(\mathbf{t})\}$
- Notice that this looks linear (i.e. we could use regression to estimate parameters). The problem is that we do not know $\ell_k\{w(\mathbf{t})\}$, $k \geq 1$.

Estimation of the linear working model

- K be the number of terms in a finite truncation of the infinite summations on the previous slide.
- Because of the assumption of joint normality of the functional principle components and measurement error,

$$\ell_k\{w(\mathbf{t})\} = \lambda_k \phi_k(\mathbf{t})\{\Phi(\mathbf{t})\Lambda\Phi(\mathbf{t})^\intercal + \sigma^2 I_m\}^{-1}\Phi(\mathbf{t})w(\mathbf{t})$$

  where $\Phi(\mathbf{t})$ is the $m \times K$ matrix whose columns are $\phi_1(t_1), \ldots, \phi_K(t_m)$ and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_K\}$.
- K corresponds to the number of leading functional principle components and can be selected using the percentage variance explained, cross-validation, or information criteria.
- It may or may not be clear at first glance, but the result is that we can use functional principle component analysis to estimate $\ell_k\{w(\mathbf{t})\}$ for $k = 1, \ldots, K$.

Estimation algorithm

Having now established the rationale for this approach, we can describe linear functional Q-learning:

1. Use functional principle components to construct estimators $\phi_{n,k}(\cdot)$ of $\phi_k(\cdot)$ and $\hat{\ell}_{n,k}\{w(\mathbf{t})\}$ of $\ell_k\{w(\mathbf{t})\}$.
2. Let $\theta \equiv (\alpha^\intercal, \delta^\intercal, \beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_K)^\intercal$, define

$$\hat{Q}_n^K\{\mathbf{x}, w(\mathbf{t}), a; \theta\} \equiv \mathbf{x}^\intercal\alpha + \sum_{k=1}^{K} \beta_k \hat{\ell}_k(w(\mathbf{t})) + a\{\mathbf{x}^\intercal\delta + \sum_{k=1}^{K} \gamma_k \hat{\ell}_k(w(\mathbf{t}))\}$$

   and compute $\hat{\theta}_n = \text{argmin}_\theta \mathbb{P}_n[Y - \hat{Q}_n^K\{\mathbf{X}, W(\mathbf{T}), A; \theta\}]^2$
3. The estimated optimal decision rule is $\hat{\pi}_n(\mathbf{x}, w(\mathbf{t})) = \text{argmax}_a \hat{Q}_n^K\{\mathbf{x}, w(\mathbf{t}), a; \hat{\theta}_n\}$.

Convergence rates for the functional Q-learning estimator

Under standard assumptions for sparse functional data analysis, regression, and scalar-on-function regression models,

- Let K_n be an increasing sequence of integers such that $K_n \to \infty$ and $K_n/n^{2\Delta} \to 0$ as $n \to \infty$, then

$$\mathbb{E}\left\{ |\hat{Q}_n^{K_n}\{\mathbf{X}, W(\mathbf{T}), A; \hat{\theta}_n\} - Q\{\mathbf{X}, W(\mathbf{T}), A\}| \mathcal{D}_n \right\}$$
$$= O_p(K_n n^{-1/2} + K_n^{-1/2} n^{-\Delta})$$

- We also have that

$$\mathbb{E}Y^*(\pi^{\text{opt}}) - \mathbb{E}\{Y^*(\hat{\pi}_n)\} = O_p\{(K_n n^{-1/2} + K_n^{1/2} n^{-\Delta})\}.$$

- If we further assume that $P[Q\{\mathbf{X}, W(\mathbf{T}), 1\} - Q\{\mathbf{X}, W(\mathbf{T}), -1\}] = 0$, then

$$\mathbb{E}\{|\hat{\pi}_n\{\mathbf{X}, W(\mathbf{T})\} - \pi^{\text{opt}}\{\mathbf{X}, W(\mathbf{T})\}| \mathcal{D}_n\} \to_p 0$$

Discussion

- Incorporating sparse, irregularly spaced patient information into the search for ITRs is not a trivial task.
- We have seen that modeling longitudinal measures as a stochastic process with error is a promising approach.
- Estimation depends critically on assumptions on the error; it allows us to do functional PCA, after which we can do standard Q-learning.

Outline

Personalized dose finding

Outcome weighted learning with censored data

Feature construction for dynamic treatment regimes

Balanced policy evaluation and learning

Estimating optimal treatment regimes using lists

Motivation

- With observational data, there are often data for more than 2 available treatments
- Many of the existing methods rely on a reduction to weighted classification via rejection and importance sampling
- Rejection and importance sampling discard a significant amount of observations effectively leading to smaller datasets
  - The reduction in sample size leads to high variance
  - The variance can be reduced using methods that introduce an unknown amount of bias
- Balanced policy evaluation and learning is an approach that directly optimizes the balance between the bias and variance

Data and notation

- Patient-level covariates $X = (X_1, X_2, \ldots, X_d)^T \in \mathcal{X}$
- Treatment assignment $T \in \mathcal{T} = \{1, \ldots, m\}$
- The cost $Y(t)$ is the outcome that would be observed if treatment t were given
- In this setting we assume that smaller outcomes are preferable
- A policy is a map $\pi : \mathcal{X} \mapsto \Delta^m$ from observed covariates to a probability vector in the m-simplex $\Delta^m = \{p \in \mathbb{R}_+^m : \sum_{t=1}^m p_t = 1\}$
- Given an observation of covariates x, the policy π specifies that treatment t should be applied with probability $\pi_t(x)$
- In policy evaluation, we wish to evaluate the performance of a given policy based on historical data

Sample average policy effect

- Define the (unknown) mean-outcome function $\mu_t(x) = E[Y(t)|X = x]$
- In policy evaluation, given a policy π, we wish to estimate its *sample-average policy effect* (SAPE)

$$\text{SAPE}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{m} \pi_t(X_i) \mu_t(X_i),$$

- Specifically, we want to estimate the SAPE with an estimator $\widehat{\tau}(\pi) = \widehat{\tau}(\pi; X_{1:n}, T_{1:n}, Y_{1:n})$ that depends only on the observed data and the policy π
- The SAPE quantifies the average outcome that a policy π would induce in the sample and hence measures its risk

Weighted estimators

- Weighting-based approaches seek weights that make the reweighted outcome data look as though it were generated by the policy being evaluated
- For any weights $W(\pi) = W(\pi; X_{1:n}, T_{1:n})$ the weighted estimator is of the form

$$\widehat{\tau}_W = \frac{1}{n} \sum_{i=1}^{n} W_i Y_i \tag{8}$$

- A common weighted estimator uses inverse propensity weighting (IPW) with weights

$$W_i^{\text{IPW}}(\pi) = \pi_{T_i}(X_i) / \widehat{\phi}_{T_i}(X_i)$$

where $\widehat{\phi}$ is the estimated propensity score

Regression based and DR estimators

- With an estimate of the mean model, $\widehat{\mu}$, we can define a regression-based estimator

$$\widehat{\tau}^{\text{reg}}_{\widehat{\mu}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}\pi_t(X_i)\widehat{\mu}_t(X_i)$$

- The regression-based estimator can be combined with a weighting-based approach to define a doubly robust (DR) estimator

$$\widehat{\tau}^{\text{DR}}_{W,\widehat{\mu}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m}\pi_t(X_i)\widehat{\mu}_t(X_i) + \frac{1}{n}\sum_{i=1}^{n} W_i(Y_i - \widehat{\mu}_{T_i}(X_i)) \quad (9)$$

- The DR estimator is consistent if either the mean model ($\widehat{\mu}$) or weights (e.g. propensity score model) are correctly specified

Conditional mean squared error

- For the simple weighted and DR estimators in equations (8) and (9), we can measure the risk of the estimators using the conditional mean square error (CMSE)

$$\text{CMSE}(\widehat{\tau}, \pi) = \mathbb{E}\left[(\widehat{\tau} - \text{SAPE}(\pi))^2 | X_{1:n}, T_{1:n}\right]$$

- The CMSE quantifies the squared bias and variance of the estimator $\widehat{\tau}$ for a policy π
- IPW weights minimize the bias at the expense of increased variance
- What if we instead choose the weighs to minimize the CMSE?
  - i.e. balance the bias variance trade-off

Conditional mean squared error, cont.

- We can decompose Y into mean + error by letting $Y_i = \mu_{T_i}(X_i) + \epsilon_i$ with $\Sigma = \text{diag}(\mathbb{E}[\epsilon_i^2 | X_i, T_i])$
- For any function f define

$$B(W, \pi; f) = \frac{1}{n} \sum_{t=1}^{m} \sum_{i=1}^{n} \left(W_i \delta_{T_i t} - \pi_t(X_i) \right) f_t(X_i)$$

  where $\delta_{T_i t} = 1$ if $T_i = t$ and 0 otherwise
- For the simple weighted and DR estimators, we can write

$$\widehat{\tau}_W - \text{SAPE}(\pi) = B(W, \pi; \mu) + \sum_{i=1}^{n} W_i \epsilon_i$$

$$\widehat{\tau}^{\text{DR}}_{W,\widehat{\mu}} - \text{SAPE}(\pi) = B(W, \pi; \mu - \widehat{\mu}) + \sum_{i=1}^{n} W_i \epsilon_i$$

Decomposing bias and variance

- Under the no unmeasured confounders assumption (i.e. $Y(t) \perp T|X$) we also have that

$$\text{CMSE}(\widehat{\tau}_W, \pi) = B^2(W, \pi; \mu) + \frac{1}{n^2} W^T \Sigma W$$

$$\text{CMSE}(\widehat{\tau}^{\text{DR}}_{W,\widehat{\mu}}, \pi) = B^2(W, \pi; \mu - \widehat{\mu}) + \frac{1}{n^2} W^T \Sigma W$$

- $B(W, \pi; \mu)$ and $B(W, \pi; \mu - \widehat{\mu})$ are the conditional bias in evaluating π for $\widehat{\tau}_W$ and $\widehat{\tau}^{\text{DR}}_{W,\widehat{\mu}}$
- $\frac{1}{n^2} W^T \Sigma W$ is the conditional variance for both $\widehat{\tau}_W$ and $\widehat{\tau}^{\text{DR}}_{W,\widehat{\mu}}$

Worst case bias

- $B(W, \pi; \mu)$ and $B(W, \pi; \mu - \widehat{\mu})$ both depend on the unknown mean process μ
- Rather than estimate the bias, we instead look at the worst-case bias under all functions, f, in a class of functions, $\mathcal{F}$, such that $\|f\| \leq 1$

$$\sup_{\|f\| \leq 1} B^2(W, \pi; f)$$

- The worst case bias is not well defined unless we assume that μ or $\mu - \widehat{\mu}$ are contained in a reproducing kernel Hilbert space (RKHS)
- Under this assumption, we can find $\sup_{\|f\| \leq 1} B^2(W, \pi; f)$ by applying Holder's inequality

Balanced policy weights

- For an estimate of the variance $\widehat{\Sigma}$, and with our measure of the worst-case bias, we can define the worst-case CMSE objective for policy evaluation

$$\mathfrak{E}^2(W, \pi; \|\cdot\|, \widehat{\Sigma}) = \sup_{\|f\| \leq 1} B^2(W, \pi; f) + \frac{1}{n^2} W^T \widehat{\Sigma} W \quad (10)$$

  where $\|\cdot\|$ is the norm induced by the RKHS

- The balanced policy weights are found as the minimizer of (10) over the space of all weights W that sum to n $\mathcal{W} = \{W \in \mathbb{R}_+^n : \sum_{i=1}^n W_i = n\}$

$$W^*(\pi; \|\cdot\|, \widehat{\Sigma}) \in \operatorname*{argmin}_{W \in \mathcal{W}} \mathfrak{E}^2(W, \pi; \|\cdot\|, \widehat{\Sigma})$$

Balanced policy evaluation

- The weights W^* can be found as the solution to a quadratic optimization problem
- For the balanced policy weights $W^* = W^*(\pi; \|\cdot\|, \widehat{\Sigma})$ the simple weighted balanced policy estimator is

$$\widehat{\tau}^*_W(\pi) = \frac{1}{n}\sum_{i=1}^{n} W_i^* Y_i$$

- Given the balanced policy weights $W^* = W^*(\pi; \|\cdot\|, \widehat{\Sigma})$ and an estimate of the mean process $\widehat{\mu}$, the DR balanced policy estimator is

$$\widehat{\tau}^{\mathrm{DR}}_{W^*,\widehat{\mu}}(\pi) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{m} \pi_t(X_i)\widehat{\mu}_t(X_i) + \frac{1}{n}\sum_{i=1}^{n} W_i^*(Y_i - \widehat{\mu}_{T_i}(X_i))$$

Balanced policy learning

- For a given policy class Π we let the balanced policy learner yield the policy $\pi \in \Pi$ which minimizes $\widehat{\tau}_W(\pi)$ or $\widehat{\tau}^{\text{DR}}_{W,\widehat{\mu}}(\pi)$:

$$\widehat{\pi}^{\text{balanced}} \in \underset{\pi}{\text{argmin}} \left\{ \widehat{\tau}_W : \pi \in \Pi, W^* \right\}$$

$$\widehat{\pi}^{\text{balanced-DR}} \in \underset{\pi}{\text{argmin}} \left\{ \widehat{\tau}^{\text{DR}}_{W,\widehat{\mu}} : \pi \in \Pi, W^* \right\}$$

- To aid in the optimization, we often restrict the policy to come from a parameterized policy class such as

$$\Pi_{\text{logit}} = \{\pi_t(x; \beta_t) \propto \exp(\beta_{t0} + \beta_t^T x)\}$$

which constrains the decision boundaries to be linear

Optimization

- This is a bilevel optimization problem, meaning that one optimization problem is embedded (nested) within another optimization problem
- This approach requires solving a quadratic program for each objective gradient evaluation
- Solving for the optimal policy is computationally intensive, especially since the bilevel problem is nonconvex
- We can use a gradient based optimization algorithm (such as BFGS) with random starts to ensure that we find the optimal policy

Discussion

- Previous policy evaluation methods have several shortcomings such as near-zero propensities, too few positive weights, and an awkward two-stage procedure
- By controlling both the bias and variance of the estimators, balanced policy evaluation and learning is able to more accurately estimate the SAPE and find policies with better expected outcomes
- The new learning method is more computationally intensive than existing approaches, solving a QP at each gradient step, but there are specialized methods for bilevel optimization which could improve the computational efficiency

Outline

Motivation

- The application of flexible supervised learning methods to estimate optimal treatment regimes mitigates the risk of model misspecification but potentially at the price of interpretability.
- This is problematic when the role of the DTR is to generate new scientific hypotheses or inform future research.
- Decision lists are a way to to simplify DTRs. They are sequences of decision rules, each represented as a sequence of if-then statements mapping logical clauses to treatment recommendations.
- It can be shown that despite the structure imposed by decision lists, they are sufficiently expressive to provide high-quality regimes.

Notation

- Observe n iid observations from a clinical trial: $\{(\mathbf{S}_i, A_i, Y_i)\}_{i=1}^n$.
  - $A \in \mathcal{A}$ is the treatment actually received during the tth stage.
  - $Y \in \mathbb{R}$ is a scalar outcome.
  - $\mathbf{X}$ denotes the information available to the decision maker at baseline.
- Assume that larger values of Y are better.
- We will present the idea of decision lists using the single stage setting, but it can be extended to the multistage setting.

Notation

- A treatment regime $\boldsymbol{\pi}$ is a function $\pi : \mathcal{X} \to \mathcal{A}$ so that under $\boldsymbol{\pi}$ a patient presenting with $\mathbf{X} = \mathbf{x}$ is recommended treatment $\pi(x)$.
- For ease of exposition, assume that all treatments are feasible for all patients.
- Let $\mathbb{E}^{\boldsymbol{\pi}}$ denote expectation with respect to the distribution induced by assigning treatments according to $\boldsymbol{\pi}$.
- Given a class of regimes Π, an optimal treatment regime satisfies $\boldsymbol{\pi}^{\text{opt}} \in \Pi$ and $\mathbb{E}^{\boldsymbol{\pi}^{\text{opt}}} Y \geq \mathbb{E}^{\boldsymbol{\pi}} Y$ for all $\boldsymbol{\pi} \in \Pi$.

The goal

The goal is to construct an estimator of $\boldsymbol{\pi}^{\text{opt}}$ when Π is the class of list-based regimes. The decision rule π in a list-based regime has the form

$$\begin{aligned} &\text{If } \mathbf{x} \in R_1 \text{ then } a_1; \\ &\text{else if } \mathbf{x} \in R_2 \text{ then } a_2; \\ &\ldots \\ &\text{else if } \mathbf{x} \in R_L, \text{ then } a_L, \end{aligned} \tag{11}$$

where:

- Each R_ℓ is a subset of $\mathcal{X}$, and
  - $a_\ell \in \mathcal{A}$,
  - $\ell = 1, \ldots, L$, and
  - L is the length of π.

Thus, a compact representation of π_t is $\{(R_\ell, a_\ell)\}_{\ell=1}^{L}$.

Interpretable clauses

To increase interpretability, we will restrict R_ℓ to clauses involving thresholding with at most two covariates. Hence, R_ℓ is an element of

$$\begin{aligned}\mathcal{R} =& \{\mathcal{X}, \{\mathbf{x} \in \mathcal{X} : x_{j_1} \leq \tau_{j_1}\}, \{\mathbf{x} \in \mathcal{X} : x_{j_1} > \tau_1\}, \\ & \{\mathbf{x} \in \mathcal{X} : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} \leq \tau_2\}, \\ & \{\mathbf{x} \in \mathcal{X} : x_{j_1} \leq \tau_1 \text{ and } x_{j_2} > \tau_2\}, \\ & \{\mathbf{x} \in \mathcal{X} : x_{j_1} > \tau_1 \text{ and } x_{j_2} \leq \tau_2\}, \\ & \{\mathbf{x} \in \mathcal{X} : x_{j_1} > \tau_1 \text{ and } x_{j_2} > \tau_2\} : \\ & 1 \leq j_1 < j_2 \leq d, \tau_1, \tau_2 \in \mathbb{R}\}, \end{aligned} \tag{12}$$

where j_1, j_2 are indices and τ_1, τ_2 are thresholds.
Thus the class of regimes of interest is Π, where
$\Pi = \{\{R_\ell, a_\ell\}_{\ell=1}^{L} : R_\ell \in \mathcal{R}, a_\ell \in \mathcal{A}, L \leq L_{\text{max}}\}$ and L_{max} is an upper bound on list length L.

General idea: Q-learning with policy search

How can we learn the optimal treatment regime π^{opt}?

- Combine non-parametric Q-learning with policy-search.
- We haven't covered Q-learning (yet), but in the single stage setting, it's just regression.

To understand this method,

- First we'll define the Q-functions.
- Then we'll see the steps for learning π^{opt}.
- Finally, we'll walk through each step.

Defining the Q-functions

- Define

$$Q(\mathbf{x}, a) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x}, A = a).$$

- It can be shown that

$$\pi^{\text{opt}} = \underset{\pi \in \Pi}{\text{argmax}}\, \mathbb{E}\{Y|\mathbf{X}, \pi(\mathbf{X})\}.$$

Overview of how to find the optimal policy

Let $\mathcal{Q}$ denote a postulated class of models for Q. We have the following schematic:

- Construct an estimator of $Q \in \mathcal{Q}$. E.g., using penalized least squares

$$\hat{Q} = \underset{Q \in \mathcal{Q}}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - Q(\mathbf{X}_i, A_i)\}^2 + \mathcal{P}(Q)$$

  where $\mathcal{P}(Q)$ is a penalty on the complexity of Q.
- Define $\hat{\pi} = \operatorname{argmax}_{\pi \in \Pi} \sum_{i=1}^{n} \hat{Q}\{\mathbf{X}_i, \pi(\mathbf{X}_i)\}$.

Implementation overview

To implement the schematic on the previous slide, we need to do the following:

1. Choose a class of models for the Q-function.
2. Construct an estimator within that class.
3. Compute $\text{argmax}_{\pi \in \Pi} \sum_{i=1}^{n} \hat{Q}\{\mathbf{X}_i, \pi(\mathbf{X}_i)\}$.

We will look at each step in detail in the following slides.

Step 1: Estimation of the Q-function

In Zhang et al. (2018) the basic idea for estimating the Q-function is to use kernel ridge regression with a Gaussian kernel. This is a very general approach to modeling a function. The key modeling choice is the kernel; we recall that the kernel relates the "closeness" of points in the predictor space to how the function surface changes. Specifically:

- Let $K(\cdot, \cdot)$ be a symmetric and positive-definite function from $\mathbb{R}^d \times \mathbb{R}^d$, and let $\mathbb{H}$ be the corresponding reproducing kernel Hilbert space (RKHS).
- The Gaussian kernel is given by,

$$K(\mathbf{x}, \mathbf{z}) = \exp\left\{ -\sum_{j=1}^{d} \gamma_j (x_j - z_j)^2 \right\},$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)^\intercal$ is a tuning parameter and $\gamma_j > 0$ for all j.

Step 1: Estimation of the Q-function (continued)

- For each $a \in \mathcal{A}$, estimate $Q(\cdot, a)$ via penalized least squares

$$\hat{Q}(\cdot, a) = \underset{f \in \mathbb{H}}{\text{argmin}} \frac{1}{n_a} \sum_{i \in \mathcal{I}_a} \{Y_i - f(\mathbf{X}_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2$$

  where $\mathcal{I}_a = \{i : A_i = a\}$, $n_a = |\mathcal{I}_a|$, and $\lambda > 0$.
- By the representer theorem

$$\hat{Q}(\mathbf{x}, a) = \sum_{i \in \mathcal{I}_a} K(\mathbf{x}, \mathbf{X}_i) \hat{\beta}_{ia},$$

  where $\mathbf{Y}_a = (Y_i)_{i \in \mathcal{I}_a}$, $\mathbf{K}_a = \{K(\mathbf{X}_i, \mathbf{X}_j)\}_{i,j \in \mathcal{I}_a}$, and $\hat{\boldsymbol{\beta}}_a = (\hat{\beta}_{ia})_{i \in \mathcal{I}_a}$ satisfy $\hat{\boldsymbol{\beta}}_a = \text{argmin}_{\boldsymbol{\beta}} \|\mathbf{Y}_a - \mathbf{K}_a \boldsymbol{\beta}\|^2 + n_a \lambda \boldsymbol{\beta}^{\mathsf{T}} \mathbf{K}_a \boldsymbol{\beta}$.
- This looks hard, but it's actually intuitive. The left hand side is the conditional mean surface. The right hand side relates $\mathbf{x}$ to the observed data via the kernel. This is then scaled by $\hat{\beta}_{ia}$.
- As usual, define $\hat{\pi} = \text{argmax}_{\pi \in \Pi} \sum_{i=1}^{n} \hat{Q}\{\mathbf{X}_i, \pi(\mathbf{X}_i)\}$.

Step 2: Estimating the policy

- Beyond estimating the Q-function, we also need to compute $\hat{\pi} = \text{argmax}_{\pi \in \Pi} \sum_{i=1}^{n} \hat{Q}\{\mathbf{X}_i, \pi(\mathbf{X}_i)\}$ where Π is the space of list-based decision rules we previously defined.
- Any element in Π can be expressed as $\{(R_\ell, a_\ell)\}_{\ell=1}^{L}$, but optimizing over all regions and treatments is not computationally feasible (except for small problems).
- Zhang et al. (2018) propose a greedy algorithm for constructing $\hat{\pi}$. This translates into optimizing "clause-by-clause" rather than optimizing over all clauses at once.
- In the next slides, we detail this algorithm.

Step 2: Estimating the first clause

- Define $\hat{\pi}^Q$ to be the map $\mathbf{x} \mapsto \text{argmax}_{a \in \mathcal{A}} \hat{Q}(\mathbf{x}, a)$.
- To estimate the first clause (R_1, a_1) in π, consider the following decision-list parameterized by R and a:

$$\begin{aligned} &\text{If } \mathbf{x} \in R \text{ then } a; \\ &\text{else if } \mathbf{x} \in \mathcal{X} - R \text{ then } \hat{\pi}^Q(\mathbf{x}). \end{aligned} \tag{13}$$

- If all subjects follow (13), the estimated mean outcome is

$$\frac{1}{n} \sum_{i=1}^{n} [I(\mathbf{X}_i \in R) \hat{Q}(\mathbf{X}_i, a) + I(\mathbf{X}_i \notin R) \hat{Q}(\mathbf{X}_i, \hat{\pi}^Q(\mathbf{X}_i))]. \tag{14}$$

  We want the maximizer of (14) to be the estimator of (R_1, a_1).
- Note the difference between the estimated mean under $\hat{\pi}^Q$ and (13) is

$$\frac{1}{n} \sum_{i=1}^{n} I(\mathbf{X}_i \in R)[\hat{Q}\{\mathbf{X}_i, \hat{\pi}^Q(\mathbf{X}_i)\} - \hat{Q}(\mathbf{X}_i, a)].$$

  This is the price of interpretability.

Step 2: Estimating the first clause (cont.)

Adding a complexity penalty to (14) encourages parsimonious lists and ensures a unique minimizer.

Define $V(R) \in \{0, 1, 2\}$ to be the number of covariates needed to check inclusion in R. Define $\hat{R}_1$ and $\hat{a}_1$ as the maximizers over R and a in

$$\frac{1}{n}\sum_{i=1}^{n}[I(\mathbf{X}_i \in R)\hat{Q}(\mathbf{X}_i, a) + I(\mathbf{X}_i \notin R)\hat{Q}\{\mathbf{X}_i, \hat{\pi}^Q(\mathbf{X}_i)\}]$$

$$+ \underbrace{\zeta\left\{\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in R)\right\}}_{\substack{\text{Reward regions } R \text{ with} \\ \text{large mass relative} \\ \text{to the distribution of } \mathbf{X}_i}} + \underbrace{\eta\{2 - V(R)\}}_{\substack{\text{Rewards regions} \\ \text{that involve} \\ \text{fewer covariates}}} \qquad (15)$$

where ζ, $\eta > 0$ are tuning parameters. Also impose the constraint $\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in R) > 0$ to avoid searching over vacuous clauses.

Step 2: Estimating the second clause

- To estimate the second clause, we consider the following decision list parameterized by R and a

$$\begin{aligned} &\text{If } \mathbf{x} \in \hat{R}_1 \text{ then } \hat{a}_1; \\ &\text{else if } \mathbf{x} \in R \text{ then } a; \\ &\text{else if } \mathbf{x} \in \mathcal{X} \text{ then } \hat{\pi}^Q(\mathbf{x}). \end{aligned} \tag{16}$$

- If all subjects follow (16), the estimated outcome is

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \in \hat{R}_1)\hat{Q}(\mathbf{X}_i, \hat{a}_1)}_{\text{Independent of } R \text{ and } a} + \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \notin \hat{R}_1, \mathbf{X}_i \in R)\hat{Q}(\mathbf{X}_i, a) + \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \notin \hat{R}_1, \mathbf{X}_i \notin R)\hat{Q}\{\mathbf{X}_i, \hat{\pi}^Q(\mathbf{X}_i)\}. \tag{17}$$

Step 2: Estimating the second clause (cont.)

As with the first clause, we maximize the penalized criterion

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \notin \hat{R}_1, \mathbf{X}_i \in R)\hat{Q}(\mathbf{X}_i, a) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \notin \hat{R}_1, \mathbf{X}_i \notin R)\hat{Q}\{\mathbf{X}_i, \hat{\pi}^Q(\mathbf{X}_i)\} \\
&\quad + \zeta\left\{\frac{1}{n}\sum_{i=1}^{n} I(\mathbf{X}_i \notin \hat{R}_1, \mathbf{X}_i \in R)\right\} + \eta\{2 - V(R)\}. \qquad (18)
\end{aligned}$$

with respect to $R \in \mathcal{R}$, $a \in \mathcal{A}$ and subject to the constraint $n^{-1}\sum_{i=1}^{n} I(\mathbf{X}_i \notin \hat{R}_1, \mathbf{X}_i \in R) > 0$.

Continue this procedure until every subject gets a recommended treatment, i.e., $R_\ell = \mathcal{X}$ for some ℓ or the maximum length is reached, $\ell = L_{\max}$. If the max list length is reached, set $R_{L_{\max}} = \mathcal{X}$ and choose $\hat{a}_{L_{\max}}$ to be the estimated best single treatment for all remaining subjects.

Step 3: Computation

- Via brute-force search for $(\hat{R}_\ell, \hat{a}_\ell)$ has computational complexity $O(n^3 d^2 m)$
- Zhang et al. (2018) provides an algorithm such that for each ℓ, the estimator $(\hat{R}_\ell, \hat{a}_\ell)$ can be computed with $O(n \log n d^2 m)$ operations.

Discussion

- Although the estimation strategy looks intimidating, it moves forward in a very logical way by going clause-by-clause.
- Via regularization, clauses with large mass relative to the distribution of the available information at the decision time are encouraged as are regions that involve fewer covariates (simple rules).
- Optimization over lists is in general computationally intensive. The algorithm proposed in Zhang et al. (2018) reduces this burden.