Semi-Supervised Inference for Optimal Treatment Decision with Electronic Medical Record Data

Wenbin Lu

Department of Statistics
North Carolina State University

February 18-22, 2019

Background

- Traditional treatment regime: "one-size-fits-all"
- Individualized treatment regime: a set of treatment decision rules that aim to
  - account for individual heterogeneity in many aspects, such as clinical, genetic, social, environmental and behavior characteristics;
  - maximize long-term clinical outcomes;
  - reduce the risk of over- or under- treatment for individual patients.
- Develop statistical and machine learning tools for optimal treatment regime (OTR) have recently attracted much attention for complex diseases, such as cancer, AIDS and mental disorder.

Mathematical Framework

For a single treatment decision point:

- Y, the real-valued response;
- $A \in \mathcal{A}$, treatment received by patient, where $\mathcal{A}$ is the set of available treatment methods. e.g., $\mathcal{A} = \{0, 1\}$;
- $X \in \mathcal{X} \subset R^p$, p-vector baseline covariates;
- a treatment regime g: a mapping $\mathcal{X} \to \mathcal{A}$;
- $Y^*(a)$, a potential outcome that would result if a patient were assigned to the treatment $a \in \mathcal{A}$;
- optimal treatment regime: $g^{opt}(X) = \arg\max_{g \in \mathcal{G}} E[Y^*(g(X))]$, where $\mathcal{G}$ denote the set of all possible treatment regimes.

Optimal Treatment Decision using EMR data

Merits:

- Provide a wealth of de-identified clinical and phenotype data for large patient cohorts;
- Such large scale datasets give unique opportunities for addressing important questions in modern medical research, such as optimal treatment decision.

Challenges:

- Data are usually recorded not for research purpose;
- Phenotyping issues (measurement errors);
- Missing data (treatments and responses are not available for many patients);
- A variety of data types (structured or unstructured).

Our Considered Problem

Notation

- A - received treatment
- X - a vector of subject characteristics ascertained prior to treatment
- Y - response variable of interest; larger means a better outcome

Observed Data

- Complete data: (X_i, A_i, Y_i), $i = 1, ..., n$.
- Incomplete data: X_j, $j = n + 1, ..., N$.
- Here, $N >> n$ ($n/N \to 0$ as n, N goes to infinity).

Problem

- How to derive an optimal treatment regime (OTR) using both complete and incomplete data?
- How to make inference for the estimated treatment rule?

Assumptions and Models

- Consistency assumption:

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

- No unmeasured confounders assumption (strong ignorablility):

$$\{Y^*(1), Y^*(0)\} \perp\!\!\!\perp A \mid X$$

- Positivity assumption: $0 < P(A = 1|X) < 1$ for any X.
- Model: $Y = \mu(X) + A \cdot C(X) + \epsilon$.
- OTR: $g^{opt}(X) = I\{C(X) > 0\}$.
- Working model with a linear OTR: $Y = \mu(X) + A \cdot (\beta' \widetilde{X}) + \epsilon$, where $\widetilde{X} = (1, X')'$.

Estimation with Complete Data Only

- Define transformed response

$$\widetilde{Y} = \frac{Y\{A - \pi(X)\}}{\pi(X)\{1 - \pi(X)\}}.$$

- Note that $E(\widetilde{Y}|A, X) = C(X)$.
- Least squares estimation (Lu, Zhang and Zeng, 2013)

$$\widehat{\beta}_{TR} = \arg\min_{\beta} \sum_{i=1}^{n} (\widetilde{Y}_i - \beta' \widetilde{X}_i)^2.$$

- Estimated OTR: $\hat{g}_{TR}^{opt}(X) = I(\widehat{\beta}'_{TR} \widetilde{X} > 0)$.
- $\widehat{\beta}_{TR} \to \beta^*$ almost surely, where β^* is the least false parameters.

Proposed Semi-Supervised Learning with Kernel Imputation

- Define $Q(X, A) = E(Y|X, A)$.
- Kernel estimation of Q functions:

$$\hat{Q}(X, 1) = \frac{\sum_{i=1}^{n} W(\frac{X - X_i}{h}) A_i Y_i}{\sum_{i=1}^{n} W(\frac{X - X_i}{h}) A_i},$$

  and

$$\hat{Q}(X, 1) = \frac{\sum_{i=1}^{n} W(\frac{X - X_i}{h})(1 - A_i) Y_i}{\sum_{i=1}^{n} W(\frac{X - X_i}{h})(1 - A_i)},$$

  where W is a kernel function and h is the bandwidth.
- Define $\hat{C}(X) = \hat{Q}(X, 1) - \hat{Q}(X, 0)$.

Estimation using Incomplete Data

- Least square estimation with kernel imputation

$$\widehat{\beta}_{NP} = \arg\min_{\beta} \sum_{j=n+1}^{N} \left\{ \hat{C}(X_j) - \beta' \widetilde{X}_j \right\}^2 .$$

- Estimated OTR: $\hat{g}_{NP}^{opt}(X) = I(\widehat{\beta}'_{NP} \widetilde{X} > 0)$.
- Asymptotic distribution: under certain conditions, we have

$$n^{1/2}(\widehat{\beta}_{NP} - \beta^*) = n^{-1/2} \sum_{i=1}^{n} \Psi_{i,NP} + o_p(1),$$

where $\Psi_{i,NP} = \left\{ \frac{A_i}{\pi(X_i)} - \frac{1-A_i}{1-\pi(X_i)} \right\} \Lambda^{-1} \widetilde{X}_i \{ Y_i - Q(X_i, A_i) \}$ and $\Lambda = E(\widetilde{X}\widetilde{X}')$.

Semi-Supervised Learning with Bias Correction

- Divide the complete data into $\mathcal{K}$ folds: $\mathcal{O}_1, ..., \mathcal{O}_{\mathcal{K}}$.
- Let $\hat{Q}^{(-k)}(X, A)$ denote the kernel estimator of Q function based on the data excluding the kth fold.
- $\mathcal{K}$-fold cross-validation with linear refitting:

$$\widehat{\theta}_1 = \arg\min_{\theta_1} \sum_{k=1}^{\mathcal{K}} \sum_{i \in \mathcal{O}_k} \frac{A_i}{\hat{\pi}(X_i)} \left\{ Y_i - \hat{Q}^{(-k)}(X_i, 1) - \theta_1' \widetilde{X}_i \right\}^2$$

$$\widehat{\theta}_0 = \arg\min_{\theta_1} \sum_{k=1}^{\mathcal{K}} \sum_{i \in \mathcal{O}_k} \frac{1 - A_i}{1 - \hat{\pi}(X_i)} \left\{ Y_i - \hat{Q}^{(-k)}(X_i, 0) - \theta_0' \widetilde{X}_i \right\}^2$$

- Define $\hat{Q}_{SS}(X, A = a) = \frac{1}{\mathcal{K}} \sum_{k=1}^{\mathcal{K}} \hat{Q}^{(-k)}(X, a) + \widehat{\theta}_a' \widetilde{X}$.

Semi-Supervised Estimator

- Define $\hat{C}_{SS}(X) = \hat{Q}_{SS}(X,1) - \hat{Q}_{SS}(X,0)$.
- Least square estimation with bias correction

$$\widehat{\beta}_{SS} = \arg\min_{\beta} \sum_{j=n+1}^{N} \left\{ \hat{C}_{SS}(X_j) - \beta' \widetilde{X}_j \right\}^2 .$$

- Estimated OTR: $\hat{g}_{SS}^{opt}(X) = I(\widehat{\beta}'_{SS} \widetilde{X} > 0)$.
- Asymptotic distribution: under certain conditions, we have

$$n^{1/2}(\widehat{\beta}_{SS} - \beta^*) = n^{-1/2} \sum_{i=1}^{n} \Psi_{i,SS} + o_p(1).$$

Simulation Studies

- Consider two covariates ($p = 2$)
- Set $n = 500$ and $N = 5000$
- Consider the following three models:
  - *Model 1 (Linear)*: $Y = \mu(X) + A \cdot (\beta' \widetilde{X}) + \epsilon$
  - *Model 2 (Cubic)*: $Y = \mu(X) + A \cdot (\beta' \widetilde{X})^3 + \epsilon$
  - *Model 3 (Sine)*: $Y = \mu(X) + A \cdot \sin(\beta' \widetilde{X}) + \epsilon$
- Consider two baseline mean functions:
  - I: $\mu(X) = (\alpha' X)^3$
  - II: $\mu(X) = (\alpha' X)(1 + \theta' X)$
- Propensity score model:
  $\pi(X) = \exp(0.5X_1 - 0.5X_2)/\{1 + \exp(0.5X_1 - 0.5X_2)\}$
- The least false parameters β^* are calculated using Monte Carlo method based on data with size of 500,000.

Summary Statistics

- Relative efficiency in terms of MSE comparing $\widehat{\beta}_{SS}$ and $\widehat{\beta}_{TR}$;
- Percent of correct decisions (PCD), defined by

$$PCD = 1 - N^{-1} \sum_{i=1}^{N} |I(\widehat{\beta}' \widetilde{X}_i > 0) - I(\beta^{*\prime} \widetilde{X}_i > 0)|$$

- Value function of the estimated OTR, computed using Monte Carlo method by

$$V = M^{-1} \sum_{m=1}^{M} \left\{ \mu(X_m) + \hat{g}^{opt}(X_m) \cdot C(X_m) \right\},$$

where $M = 500,000$.

Results I

			TR		SS	
$\mu(X)$	Model	V_0	V	PCD	V	PCD
I	Linear	0.56	0.54 (0.04)	0.92 (0.06)	0.56 (0.01)	0.96 (0.02)
	Cubic	0.24	0.22 (0.06)	0.87 (0.12)	0.24 (0.01)	0.93 (0.05)
	Sine	0.32	0.21 (0.12)	0.80 (0.17)	0.26 (0.06)	0.88 (0.09)
II	Linear	1.31	1.29 (0.05)	0.91 (0.06)	1.31 (0.01)	0.96 (0.02)
	Cubic	0.99	0.98 (0.05)	0.86 (0.11)	0.99 ($<$0.01)	0.94 (0.04)
	Sine	1.06	0.96 (0.11)	0.80 (0.16)	1.02 (0.04)	0.89 (0.06)

Results II: $\mu(X) = (\alpha' X)^3$

		TR				SS				
Model	β^*	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP	RE
	0	−0.010	0.220	0.209	0.94	−0.005	0.123	0.122	0.95	3.21
Linear	1	−0.021	0.329	0.347	0.98	−0.008	0.182	0.173	0.93	3.28
	1	−0.020	0.355	0.347	0.97	−0.004	0.198	0.175	0.92	3.22
	0	−0.008	0.206	0.205	0.94	0.001	0.123	0.132	0.95	2.80
Cubic	0.41	0.002	0.352	0.338	0.95	−0.002	0.212	0.193	0.95	2.74
	0.81	0.002	0.423	0.390	0.94	−0.010	0.239	0.212	0.91	3.14
	0	−0.006	0.171	0.168	0.96	0.004	0.118	0.116	0.95	2.13
Sine	0.37	−0.007	0.282	0.270	0.94	−0.011	0.170	0.158	0.91	2.74
	0.37	0.011	0.296	0.272	0.94	0.011	0.176	0.161	0.92	2.82

Results III: $\mu(X) = (\alpha' X)(1 + \theta' X)$

		TR				SS				
Model	β^*	Bias	ESE	ASE	CP	Bias	ESE	ASE	CP	RE
Linear	0	−0.017	0.242	0.230	0.94	−0.007	0.114	0.116	0.95	4.53
	1	−0.012	0.348	0.326	0.94	0.011	0.150	0.151	0.93	5.37
	1	−0.021	0.352	0.347	0.94	−0.011	0.163	0.154	0.91	4.66
Cubic	0	−0.005	0.236	0.223	0.93	−0.026	0.121	0.122	0.95	3.77
	0.41	−0.014	0.341	0.313	0.93	0.003	0.155	0.154	0.92	4.86
	0.81	−0.004	0.432	0.392	0.91	−0.008	0.193	0.186	0.92	4.99
Sine	0	0.003	0.210	0.202	0.94	0.009	0.117	0.113	0.95	3.14
	0.37	0.002	0.296	0.289	0.94	−0.004	0.146	0.146	0.94	4.12
	0.37	−0.005	0.300	0.290	0.93	0.001	0.136	0.140	0.95	4.90

Data Application

- Consider patients undergoing a hypotensive episode (HE) in the ICU within the MIMIC-III database (Johnson et al. 2016)
- Goal: to minimize end-organ damage (measured by an increase in serum creatinine (Lehman et al. 2010))
- Treatments: IV fluid resuscitation and vasopressors (Lee et al. 2012)
- Response: pre-HE serum creatinine - post-HE serum creatinine
- A total of 3,316 patients were included: 1,243 patients have complete treatment and response information; 2,073 patients have missing information in treatment and/or response
- Covariates included in the OTR: baseline serum creatinine and age
- Covariates included in the propensity score model: baseline serum creatinine, age, gender, service type, comorbidity score, total urine output, mean blood oxygen saturation, and average mean arterial pressure

Estimation Results

	TR			SS		
Covariates	$\hat{\beta}_{TR}$	SE	P-Value	$\hat{\beta}_{SS}$	SE	P-Value
Intercept	−0.102	0.111	0.355	−0.119	0.102	0.246
Baseline Creatinine	−0.371	0.256	0.147	−0.508	0.197	0.010
Age	0.196	0.146	0.178	0.114	0.133	0.392

Table : Treatment allocation

		SS	
	Treatment	IV Fluid	Vasopressors
TR	IV Fluid	1553	314
	Vasopressors	119	1330

Future Works

- Incorporate high-dimensional predictors
- Incorporate unstructured data, such as clinical notes
- Consider dynamic treatment regime
- Consider multiple disease outcomes of interest