Assisted Policy Search

Xi Lu and Susan A. Murphy, Department of Statistics, University of Michigan

1. Abstract

In recent years, many Sequential Multiple Assignment Randomized Trials (SMART) have been conducted. The data collected from this type of trials is extremely useful to answer questions such as "what treatment policy (or, dynamic treatment regime) will yield highest mean outcome if followed by patients from the same population". In this research project, we propose novel approach to estimating the value of treatment policies and policy search method based on the policy value estimators. The proposed assisted estimator for policy value is based on Structural Nested Mean Model (SNMM) that was developed by Robins (1994) in causal inference research. Moreover, we propose some computationally more efficient variation of the original policy search problem.

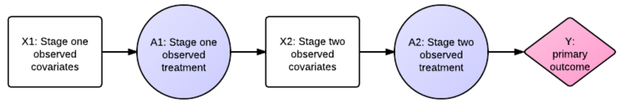
2. Framework and Data Structure

Treatment policies (dynamic treatment regimes, or adaptive treatment strategies): a sequence of decision rules:

- Take measurements of patients' time-varying covariates as inputs, output recommended treatments. Characteristics: Dynamic, Personalized.
- A policy class: A class of policies parameterized by a finite number of parameters.

Goal of this project: Identify the optimal treatment policy within a given policy class, using data from two-stage Sequential Multiple Assignment Randomized Trials.

Data structure: five components for each individual as in the following flowchart



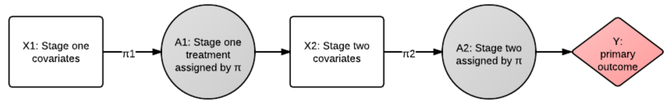
Example: In the alcohol dependence study, X_1, X_2 can include time-varying measurements of alcohol craving scores, number of heavy drinking days per week, number of drinking days per week. X_1 typically includes baseline demographics (gender, race, etc.); A_1, A_2 can be medication or behavioral therapy; interesting Y can be (1) end of study percentage of heavy drinking days; (2) end of study percentage of heavy drinking days; (3) end of study craving scores, etc.

A two-stage treatment policy consists of two mappings $\pi = (\pi_1, \pi_2)$:

$$\pi_1 : \mathcal{H}_1 \to \mathcal{A}_1 \qquad \pi_2 : \mathcal{H}_2 \to \mathcal{A}_2$$

$\mathcal{H}_1$ is space of $H_1 \equiv X_1$; $\mathcal{H}_2$ is space of $H_2 \equiv (X_1, A_1, X_2)$

The value of π is $V_\pi = E_\pi[Y]$. E_π is the expectation w.r.t. distribution of (X_1, A_1, X_2, A_2, Y) with A_1 and A_2 assigned according to rules in π (the other components of the joint distribution the same as population in the trial). The regret of π is $\rho_\pi = E_\pi[Y] - E_{(0,0)}[Y]$.



Given a policy class $\{\pi(b) : b \in \Omega \subset \mathbb{R}^r\}$ want to find

$$b^{opt} = \arg\max_b \mathbb{E}_{\pi(b)}[Y]$$

Basic scheme:

1. Construct policy value estimator $\hat{V}_{\pi(b)}$ for $V_{\pi(b)}$ (or policy regret estimator $\hat{\rho}_{\pi(b)}$ for $\rho_{\pi(b)}$);
2. Optimize $\hat{V}_{\pi(b)}$ over b (or optimize $\hat{\rho}_{\pi(b)}$ over b).

State-of-art policy search method: Marginal Structural Mean Model (MSM)

- Model $\mathbb{E}_{\pi(b)}[Y]$ as $g(b;\beta)$, $g(\cdot)$ known function (polynomial); estimate β.
- Drawback: weak scientific reasoning for parametric model $g(b,\beta)$; model complexity for multi-dimensional b.

3. Assisted Estimator for Policy Value / Regret

Define the treatment effect functions μ_1, μ_2:

$$\mu_2(H_2, A_2) = \mathbb{E}[Y|H_2, A_2] - \mathbb{E}[Y|H_2, A_2 = 0]$$
$$\mu_1(H_1, A_1) = \mathbb{E}[\mathbb{E}[Y|H_2, A_2 = 0]|H_1, A_1] - \mathbb{E}[\mathbb{E}[Y|H_2, A_2 = 0]|H_1, A_1 = 0]$$

- μ_1 represents the difference in the expected outcome of Y induced by following A_1 other than treatment coded by zero at stage one;
- μ_2 represents the difference in the expected outcome of Y induced by following A_2 other than treatment coded by zero at stage two.

We model treatment effects parametrically (first proposed by Robins [1994] as structural nested mean model):

$$\mu_1(H_1, A_1) = h_1(H_1, A_1)^T \beta_1$$
$$\mu_2(H_2, A_2) = h_2(H_2, A_2)^T \beta_2$$

h_1, h_2 are known functions with constraints $h_1(H_1, 0) \equiv 0, h_2(H_2, 0) \equiv 0$: they should incorporate prior knowledge on the interaction of the assessed treatments with past covariates.

Example: Again in alcohol dependence study, one can model $h_1(H_1, A_1) = A_1 \cdot (1, CS_1, PH_1)^T$ and $h_2(H_2, A_2) = A_2 \cdot (1, CS_2, PH_2)^T$ if it is believed at each stage treatment has interactive effect with the recent alcohol craving score (CS_i) and the recent percentage of heavy drinking days (PH_i).

Suppose we already have consistent and asymptotically normal estimator $(\hat{\beta}_1, \hat{\beta}_2)$ for (β_1, β_2). Then the following assisted estimator for policy value is consistent and asymptotically normal for V_π under regularity conditions:

$$\hat{V}_\pi(\hat{\beta}_1, \hat{\beta}_2) = \mathbb{P}_n\big\{Y - h_2(H_2, A_2)^T \hat{\beta}_2 - h_1(H_1, A_1)^T \hat{\beta}_1 + h_1\left(H_1, \pi_1(H_1)\right)^T \hat{\beta}_1 + \frac{I\{A_1 = \pi_1(H_1)\}}{f_1(A_1|H_1)} h_2\left(H_2, \pi_2(H_2)\right)^T \hat{\beta}_2\big\}, \quad (1)$$

where $f_1(A_1|H_1)$ is the known treatment assignment probability in the randomized trial.

Intuition: Construct a *pseudo-outcome* by subtracting the treatment effects of observed treatment sequence, and adding in the treatment effects of the sequence specified by π.

Similarly we propose the assisted estimator for policy regret:

$$\hat{\rho}_\pi(\hat{\beta}_1, \hat{\beta}_2) = \mathbb{P}_n\big\{h_1\left(H_1, \pi_1(H_1)\right)^T \hat{\beta}_1 + \frac{I\{A_1 = \pi_1(H_1)\}}{f_1(A_1|H_1)} h_2\left(H_2, \pi_2(H_2)\right)^T \hat{\beta}_2\big\} \quad (2)$$

3. Efficiency Correction on Assisted Estimators

Motivated by augmented estimator from missing data theory, consider the following augmented-assisted estimator for policy regret:

$$\hat{\rho}_\pi(\hat{\beta}_1, \hat{\beta}_2, \hat{\alpha}_m) = \mathbb{P}_n\big\{h_1\left(H_1, \pi_1(H_1)\right)^T \hat{\beta}_1 + \frac{I\{A_1 = \pi_1(H_1)\}}{f_1(A_1|H_1)} \left(h_2\left(H_2, \pi_2(H_2)\right) - m(H_1, A_1; \hat{\alpha}_m)\right)^T \hat{\beta}_2 + m\left(H_1, \pi_1(H_1); \hat{\alpha}_m\right)^T \hat{\beta}_2\big\} \quad (3)$$

where $m(H_1, A_1; \alpha_m)$ is a model for $\mathbb{E}[h_2(H_2, \pi_2(H_2))|H_1, A_1]$.

$\hat{\rho}_\pi(\hat{\beta}_1, \hat{\beta}_2, \hat{\alpha}_m)$ is consistent and asymptotically normal for ρ_π as long as treatment effect model is correctly specified, and does not rely on correct modeling of $m(H_1, A_1; \alpha_m)$.

Lemma 1. *If $(\hat{\beta}_1, \hat{\beta}_2)$ is semiparametric efficient g-estimator (Robins 1994) for treatment effect parameters, and the augmented term is correctly modeled (i.e. $m(H_1, A_1; \alpha_m)$ is a correct model for $\mathbb{E}[h_2(H_2, \pi_2(H_2))]$), it is then guaranteed that $\hat{\rho}_\pi(\hat{\beta}_1, \hat{\beta}_2, \hat{\alpha}_m)$ has lower variance than $\hat{\rho}_\pi(\hat{\beta}_1, \hat{\beta}_2)$ asymptotically.*

Remark: In general, since "semiparametric efficient g-estimator" is only a theoretical notion, and it is unlikely that $m(H_1, A_1; \alpha_m)$ is accurately modeled,we are not guaranteed lower variance by using augmented estimator. However, simulation does show that augmented estimator is behaving at least as well as the non-augmented estimator, if not better.

5. Policy Search

For simplicity, consider assisted policy search using non-augmented assisted estimator for policy regret.

Consider motivating example: two-stage setting with binary treatments. Consider policy class as follows (this type of policy also called "linear decision boundary"):

$$\pi_1(H_1; b_1) = I\{b_1^T X_1 > 1\}, \quad \pi_2(H_2; b_2) = I\{b_2^T X_2 > 1\}.$$

With binary treatments, models for treatment effects can only take the form:

$$\mu_1(H_1, A_1) = A_1 S_1^T \beta_1,$$

$$\mu_2(H_2, A_2) = A_2 S_2^T \beta_2,$$

S_1, S_2 are certain features of H_1, H_2 (could be some functions of covariates).

Optimize $\hat{\rho}(b_1, b_2)$ over b_1, b_2:

$$\max_{b_1 \in \mathcal{B}_1, b_2 \in \mathcal{B}_2} \mathbb{P}_n I\{b_1^T X_1 > 1\} S_1^T \hat{\beta}_1 + \frac{(2A_1 - 1) I\{b_1^T X_1 > 1\} + (1 - A_1)}{f_1(A_1|H_1)} I\{b_2^T X_2 > 1\} S_2^T \hat{\beta}_2 \quad (4)$$

Natural idea: iteratively optimize (4) over b_1 and b_2.

Fixing b_2, (4) equivalent to

$$\min_{b_1 \in \mathcal{B}_1} \mathbb{P}_n W_1 I\{b_1^T X_1 \le 1\} \quad (5)$$

where $W_1 = S_1^T \hat{\beta}_1 + \frac{2A_1 - 1}{f_1(A_1|H_1)} I\{b_2^T S_2 > 1\} S_2^T \hat{\beta}_2$. Other ways to decompose W_1 may have better property. furthermore equivalent to

$$\min_{b_1 \in \mathcal{B}_1} \mathbb{P}_n W_{1,+} I\{b_1^T X_1 \le 1\} - W_{1,-} I\{b_1^T X_1 \le 1\} \quad (6)$$

where $W_{1,+} = \max\{W_1, 0\}, W_{1,-} = \max\{-W_1, 0\}$.

One possible convex relaxation of (6):

$$\min_{b_1 \in \mathcal{B}_1} \mathbb{P}_n W_{1,+}(2 - b_1^T X_1)_+ + W_{1,-}(b_1^T X_1)_+$$

We are currently investigating properties of different ways of relaxing the optimization. The step of optimizing over b_2 when fixing b_1 can proceed similarly.

6. Simulation

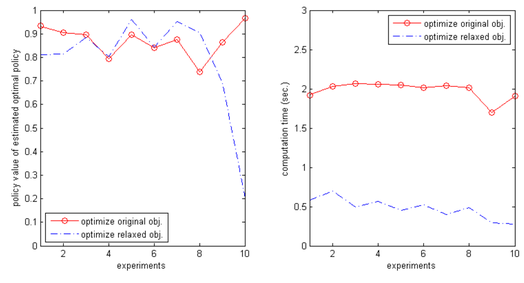


Figure 1: Comparison between optimizing the original objective function and iteratively optimizing the relaxed problem. In the left panel, y-axis is rescaled so that in each experiment 0 corresponds to the mean value of Y under randomized treatment, 1 corresponds to the policy value of true optimal policy.

We acknowledge NIMH RO1-MH-080015, NIDA grant P50-DA-010075 for support.
For References, Questions and comments please contact luxi@umich.edu