Selecting and ranking individualized treatment rules with unmeasured confounding

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

January 14, 2021 @ Selective Inference Webinar

Based on joint work with Bo Zhang, Jordan Weiss, and Dylan Small to be published in a special *JASA* issue on precision medicine. More information can be found on `http://www.statslab.cam.ac.uk/~qz280/`.

What is an individualized treatment rule (ITR)?

As the name suggests, treatment is individualized according to the subject's characteristics.

A recent example: WHO interim guideline on dexamethasone

- "WHO **strongly recommends** that corticosteroids (i.e. dexamethasone, hydrocortisone or prednisone) be given orally or intravenously **for the treatment of patients with severe and critical COVID-19**."
- "WHO **advises against** the use of corticosteroids **in the treatment of patients with non-severe COVID-19**, unless the patient is already taking this medication for another condition." [a]

[a] https://www.who.int/news-room/q-a-detail/coronavirus-disease-covid-19-dexamethasone

Optimal treatment regimes with observational data

- **Optimal treatment regime** = ITR with the best *value*.
- **Dynamic treatment regimes** = extension to multiple decision points.
- This is central to the "new" initiative of **precision medicine** and has been widely studied.
- Existing methods usually assume **(sequentially) randomized** experiments, or observational studies that satisfy **(sequential) ignorability**. This allows us to estimate the value of any ITR.
- This talk: **realistic decisions** about **static ITRs** with **unmeasured confounders**.

Sensitivity analysis for observational studies

- If we acknowledge the possibility of unmeasured confounders, how will them change the conclusions of an observational study?
- Cornfield *et al.* (1959): In order for **a confounder genotype** to **fully explain the association** between smoking and lung cancer, it must **increase the propensity of smoking** by at least **nine fold**!

We will use the following model:

Rosenbaum's sensitivity model

In words, this model assumes that the **odds ratio of receiving the treatment** for any **two individuals with the same observed covariates** is bounded between $1/\Gamma$ and Γ (Rosenbaum 1987).

- $\Gamma \geq 1$; $\Gamma = 1$ corresponds to no unmeasured confounders.

This talk

Problem

How do we select and rank ITRs under Rosenbaum's sensitivity model?

Motivation: Effect modification and the power of sensitivity analysis

Hsu *et al.* (2013) found that **subgroups with a larger treatment effect may be more robust/less sensitivity to unmeasured confounders**.

- Important consequence: ITR with a **larger value** (estimated from observational data assuming ignorability) could be **more sensitive to unmeasured confounders**.

Value $\neq$ Robustness

The estimated **value** from some observational data **assuming ignorability** is a **poor indicator for robustness**.

A counter-intuitive example

Let $r_2 \succ_\Gamma r_1$ or simply $r_2 \succ r_1$ denote that the value of r_2 is *always* greater than r_1 under the Γ-sensitivity model.
Then, it is possible that

- Under $\Gamma = 1$, $r_2 \succ r_1 \succ r_0$ (so $r_2 \succ r_0$);
- Under some $\Gamma > 1$, $r_1 \succ r_0$ but $r_2 \not\succ r_0$.

Why? Value is only **partially identified** in Rosenbaum's sensitivity model and induces a **partial order** between ITRs.

Related work

- Maximize the estimated value assuming ignorability (Qian and Murphy 2011; Y. Zhao *et al.* 2012; Dudík *et al.* 2014; Athey and Wager 2017).
- Selecting and ordering subpopulations—an old and well studied topic involves many interesting objectives but assumes a total order (Gibbons *et al.* 1999).
- Screening hypotheses in sensitivity analysis (Heller *et al.* 2009; Q. Zhao *et al.* 2018).
- Kallus and Zhou (2018) consider a similar problem but with a different sensitivity model.

Notation

Running example: Malaria in West Africa

Dataset from Hsu *et al.* (2013): 1560 matched pairs of Nigerians.

- $A \in \mathcal{A} = \{0, 1\}$ is a binary **treatment**. $A = 1$: receives treatment (insecticide spray + drug).
- $X \in \mathcal{X}$ is a vector of pre-treatment **covariates** (gender and age);
- $Y \in \mathbb{R}$ is the **outcome** (amount of malaria-causing parasites in blood).
- $r : \mathcal{X} \to \mathcal{A}$ is an **individualized treatment rule (ITR)**. We will consider six rules: $r_0, r_1, \ldots, r_5$, where r_i assigns treatment to the youngest $i \times 20\%$.
- Let $Y(0)$ and $Y(1)$ be the **potential outcomes** under control and treatment. This induces the definition: $Y(r) = Y(0)1_{\{r(X)=0\}} + Y(1)1_{\{r(X)=1\}}$.
- The **value function** is defined as $V(d) = \mathbb{E}[Y(d)]$.

Comparing two ITRs: No unmeasured confounders

- The **value difference** is $V(r_2) - V(r_1) = \mathbb{E}[Y(r_2) - Y(r_1) \,|\, r_2 \neq r_1] \cdot \mathbb{P}(r_2 \neq r_1)$.
- In our example (nested ITRs), $V(r_2) - V(r_1) = \mathbb{E}[Y(1) - Y(0) \,|\, \text{Age} \in [7, 20)] \cdot \mathbb{P}(\text{Age} \in [7, 20))$.

Standard assumptions for identifying $V(r)$ from observational data

1. **Positivity**: $\pi(a, x) = \mathbb{P}(A = a \,|\, X = x) > 0$ for all a and x;
2. **Consistency/SUTVA**: $Y = Y(A)$;
3. **Ignorability/no unmeasured confounders**: $Y(a) \perp\!\!\!\perp A \,|\, X$ for all a.

Under these assumptions, $V(r) = \mathbb{E}\Big[\dfrac{Y 1_{\{A = r(X)\}}}{\pi(A, X)}\Big]$ defines a total order.

Comparing two ITRs: Unmeasured confounders

Rosenbaum's sensitivity model

Suppose $Y(a) \perp\!\!\!\perp A \,|\, X, U$. Then we assume

$$\Gamma^{-1} \leq \mathrm{OR}\Big(\mathbb{P}(A=1 \,|\, X=x, U=u_1), \mathbb{P}(A=1 \,|\, X=x, U=u_2)\Big) \leq \Gamma, \ \forall x, u_1, u_2,$$

where $\mathrm{OR}(p_1, p_2) = \{p_1/(1-p_1)\}/\{p_2/(1-p_2)\}$ is the odds ratio.

- Definition: $r_1 \prec_{\Gamma,\delta} r_2$ (omit Γ if $\Gamma = 1$ and δ if $\delta = 0$) if $V(r_2) - V(r_1) > \delta$ **for all distributions in the Γ-sensitivity model**.
- Can verify that $\prec_\Gamma$ satisfies **irreflexivity** ($r_1 \not\prec_\Gamma r_1$), **transitivity** ($r_1 \prec_\Gamma r_2$ and $r_2 \prec_\Gamma r_3$ imply $r_1 \prec_\Gamma r_3$), and **asymmetry** ($r_1 \prec_\Gamma r_2$ implies $r_2 \not\prec_\Gamma r_1$). So it is a partial order.
- Fogarty (2020) has proposed a studentized test for Neyman's null hypothesis that the average treatment effect is zero, $(2n)^{-1} \sum Y_{ij}(1) - Y_{ij}(0) = 0$.
- This test can be adapted to test $r_1 \prec_{\Gamma,\delta} r_2$ (see our paper for detail).

Power of the sensitivity analysis

- A hallmark of Rosenbaum's sensitivity analysis—the tipping point or **sensitivity value**:

$$\Gamma^*_\alpha(r_1 \prec r_2) = \sup\{\Gamma \geq 1 \mid V(r_1) \geq V(r_2) \text{ is rejected at level } \alpha \text{ under the } \Gamma\text{-sensitivity model}\}.$$

- Asymptotic distribution of the sensitivity value (Q. Zhao 2019): Suppose $r_1(x) \leq r_2(x)$, $\forall x$, then

$$\sqrt{n}\left\{\Gamma^*_\alpha(r_1 \prec r_2) - \bar{\Gamma}\right\} \xrightarrow{d} \mathrm{N}(-z_\alpha \mu, \sigma^2),$$

where μ, σ^2 depends on the distribution of $D_i = (A_{i1} - A_{i2})(Y_{i1} - Y_{i2})$ and

$$\bar{\Gamma} = \frac{\mathbb{E}[|D_i| \mid r_1 < r_2] + \mathbb{E}[D_i \mid r_1 < r_2]}{\mathbb{E}[|D_i| \mid r_1 < r_2] - \mathbb{E}[D_i \mid r_1 < r_2]}$$

is called the **design sensitivity** (Rosenbaum 2004).

- Therefore, the power is determined by Γ with a phase transition at $\bar{\Gamma}$.
- This poses challenges to multiple hypothesis testing.

Objectives

Related problem: selecting subpopulations

- Suppose we observe $Y_i \overset{\text{ind.}}{\sim} \mathrm{N}(\mu_i, 1)$ for subpopulation i.
- Gibbons *et al.* (1999) has defined seven possible goals for ranking and selecting subpopulations.

Given $\mathcal{R} = \{r_0, r_1, \ldots, r_K\}$, three goals are relevant for comparing multiple ITRs:

1. What is the ordering of all the ITRs?
2. Which ITRs are among the best?
3. Which ITRs are better than the control rule r_0?

We cannot directly use existing methods because $\prec_\Gamma$ is not a total order.

Objectives

Some definitions

- The **maximal rules** are the ones not dominated by others,

$$\mathcal{R}_{\max,\Gamma} = \{r_i \mid r_i \not\prec_\Gamma r_j,\ \forall j\}.$$

- The **positive rules** are the ones which dominate the control. The **null rules** are the ones which don't dominate the control.

$$\mathcal{R}_{\text{pos},\Gamma} = \{r_i \mid r_0 \prec_\Gamma r_i\},\ \mathcal{R}_{\text{nul},\Gamma} = \mathcal{R} \setminus \mathcal{R}_{\text{pos},\Gamma}.$$

1. Construct a set of ordered ITR pairs, $\hat{\mathcal{O}}_\Gamma \subset \{(r_i, r_j),\ i, j = 0, \ldots, K,\ i \neq j\}$, such that

$$\mathbb{P}(r_i \prec_\Gamma r_j,\ \forall (r_i, r_j) \in \hat{\mathcal{O}}_\Gamma) \geq 1 - \alpha.$$

2. Construct $\hat{\mathcal{R}}_{\max,\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \geq 1 - \alpha$.
3. Construct $\hat{\mathcal{R}}_{\text{pos},\Gamma} \subseteq \mathcal{R}$ such that $\mathbb{P}(\hat{\mathcal{R}}_{\text{pos},\Gamma} \cap \mathcal{R}_{\text{null},\Gamma} = \emptyset) \geq 1 - \alpha$.

Objective 1: Ordering the ITRs

- Can apply Bonferroni's procedure to control the family-wise error rate, but this is very conservation because the sensitivity analysis considers the worst case scenario.
- Better alternative: reduce the number of tests using a planning sample (Heller *et al.* 2009; Q. Zhao *et al.* 2018).

Our proposal

Step 1: Split the data into two parts: one for planning and one for testing.

Step 2: For every pair of ITRs, use the planning sample to estimate the asymptotic distribution of Γ^* and the power of testing $H_{ij}: \ r_i \not\prec_\Gamma r_j$.

Step 3: Order the hypotheses by estimated power from the highest to the lowest.

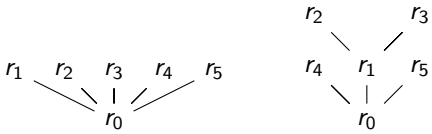
Step 4: Fixed sequence testing: sequentially test the ordered hypotheses using the testing sample at level α, until one hypothesis is rejected.

Step 5: Use transitivity of $\prec_\Gamma$ and output a Hasse diagram.

Objective 1: Malaria example

Ordered hypotheses after using the planning sample

- $\Gamma = 1$: $H_{01}, H_{02}, H_{03}, H_{04}, H_{05}, H_{13}, H_{12}, H_{14}, H_{15}, H_{23}, \ldots$.
- $\Gamma = 2$: $H_{02}, H_{01}, H_{03}, H_{04}, H_{05}, H_{12}, H_{13}, H_{14}, H_{15}, H_{45}, \ldots$.



$|\hat{\mathcal{O}}| = 5$ $\quad$ $|\hat{\mathcal{O}}| = 7$

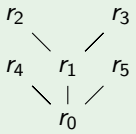
Hasse diagrams for $\Gamma = 2$: Bonferroni's correction (left) and our proposal (right).

Objective 2: Selecting the best ITRs

- Key observation: $\mathbb{P}(r_i \not\prec_\Gamma r_j \text{ is rejected} \mid r_i \in \mathcal{R}_{\max,\Gamma}) \leq \alpha$.
- This motivates us to use all the "leaves" in the Hasse diagram as the maximal elements:

$$\hat{\mathcal{R}}_{\max,\Gamma} = \{r_i \mid (r_i, r_j) \notin \hat{\mathcal{O}}_\Gamma,\ \forall j\}.$$

An example



$$\Longrightarrow \quad \hat{\mathcal{R}}_{\max} = \{r_2, r_3, r_4, r_5\}$$

- This satisfies $\mathbb{P}(\mathcal{R}_{\max,\Gamma} \not\subseteq \hat{\mathcal{R}}_{\max,\Gamma}) \leq \alpha$ if the FWER for $\hat{\mathcal{O}}_\Gamma$ is less than α.
- Can "trim" hypotheses using the following: $r_i \notin \hat{\mathcal{R}}_{\max,\Gamma}$ if $H_{ij} : r_i \not\prec_\Gamma r_j$ is rejected for **a single** r_j.

Objective 2: Malaria example

- Ordered and trimmed hypotheses for $\Gamma = 2$: $H_{02}, H_{12}, H_{45}, H_{35}, H_{53}, H_{21}$.

Table: $\hat{\mathcal{R}}_{\max,\Gamma}$ for different choices of Γ.

Γ	$\hat{\mathcal{R}}_{\max,\Gamma}$	Γ	$\hat{\mathcal{R}}_{\max,\Gamma}$
1.0	$\{r_3, r_4, r_5\}$	2.5	$\{r_2, r_3, r_4, r_5\}$
1.3	$\{r_3, r_4, r_5\}$	3.0	$\{r_1, r_2, r_3, r_4, r_5\}$
1.5	$\{r_2, r_3, r_4, r_5\}$	3.5	$\{r_1, r_2, r_3, r_4, r_5\}$
1.8	$\{r_2, r_3, r_4, r_5\}$	4.0	$\{r_1, r_2, r_3, r_4, r_5\}$
2.0	$\{r_2, r_3, r_4, r_5\}$	6.0	$\{r_0, r_1, r_2, r_3, r_4, r_5\}$

Objective 3: Selecting the positive ITRs

- Simply needs to test the hypotheses $H_{0i} : r_0 \not\prec_\Gamma r_i,\ i = 1, \ldots, K$.
- Can use the same multiple testing procedure above.

Results for the malaria example

	$\Gamma = 1$	$\Gamma = 1.3$	$\Gamma = 1.5$	$\Gamma = 1.8$
$\delta = 0$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$
$\delta = 2$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$
$\delta = 6$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_2, r_3, r_4, r_5\}$	$\{r_2\}$
	$\Gamma = 2.0$	$\Gamma = 2.5$	$\Gamma = 3.0$	
$\delta = 0$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3, r_4, r_5\}$	
$\delta = 2$	$\{r_1, r_2, r_3, r_4, r_5\}$	$\{r_1, r_2, r_3\}$	$\{r_1, r_2\}$	
$\delta = 6$	$\emptyset$	$\emptyset$	$\emptyset$	
	$\Gamma = 3.5$	$\Gamma = 4.0$	$\Gamma = 6.0$	
$\delta = 0$	$\{r_1, r_2, r_3\}$	$\{r_1, r_2\}$	$\emptyset$	
$\delta = 2$	$\emptyset$	$\emptyset$	$\emptyset$	
$\delta = 6$	$\emptyset$	$\emptyset$	$\emptyset$	

Retirement timing on health outcome

Setup

- **Treatment**: late retirement (retire between 65 and 70).
- **Outcome**: self-reported health status at 70.
- **Covariates**: year of birth, gender, education, race, occupation, partnered, annual income, smoking.
- Optimal matching (exactly on year, gender, occupation, partnered): 1858 matched pairs.
- We considered **4 subgroups**: male, white-collar workers (G_1), female, white-collar workers (G_2), male, blue-collar workers (G_3), and female, blue-collar workers (G_4).
- **16 regimes** with binary coding. For example, $r_9 = r_{1001}$ treats G_1 and G_4.

Results

Γ	$\hat{\mathcal{R}}_{\max,\Gamma}$	$\hat{\mathcal{R}}_{\text{pos},\Gamma}$
1.0	$\{r_{11}, r_{13}, r_{15}\}$	$\{r_1, r_3, r_5, r_7, r_8, \ldots, r_{15}\}$
1.2	$\{r_9, r_{11}, r_{13}, r_{15}\}$	$\{r_1, r_3, r_5, r_7, r_8, \ldots, r_{11}, r_{13}, r_{14}, r_{15}\}$
1.35	$\{r_1, r_3, r_5, r_7, r_9, r_{11}, r_{13}, r_{15}\}$	$\{r_1, r_9\}$

Discussion

- Robustness to unmeasured confounders: another dimension in decision making.
- Best ITR (largest value assuming ignorability) is often not the most robust.
- Many possible objectives for selection and ranking.
- Selective inference for partially identified/ordered problems: a potentially new topic?
- Our method cannot handle too many ITRs. Better alternatives?

Take-home message: Precision medicine or Jenga?

References

S. Athey, S. Wager, *arXiv preprint arXiv:1702.02896* (2017).

J. Cornfield *et al.*, *Journal of the National Cancer Institute* **22**, 173–203 (1959).

M. Dudík, D. Erhan, J. Langford, L. Li, *et al.*, *Statistical Science* **29**, 485–511 (2014).

C. B. Fogarty, *Journal of the American Statistical Association* **115**, 1518–1530 (531 2020).

J. D. Gibbons, I. Olkin, M. Sobel, *Selecting and ordering populations: A new statistical methodology*, (SIAM, 2nd, 1999).

R. Heller, P. R. Rosenbaum, D. S. Small, *Journal of the American Statistical Association* **104**, 1090–1101 (2009).

J. Y. Hsu, D. S. Small, P. R. Rosenbaum, *Journal of the American Statistical Association* **108**, 135–148 (2013).

N. Kallus, A. Zhou, presented at the Advances in Neural Information Processing Systems 2018, pp. 9269–9279.

M. Qian, S. A. Murphy, *The Annals of Statistics* **39**, 1180–1210 (2011).

P. R. Rosenbaum, *Biometrika* **74**, 13–26 (1987).

P. R. Rosenbaum, *Biometrika* **91**, 153–164 (2004).

Q. Zhao, *Journal of the American Statistical Association* **114**, 713–722 (2019).

Q. Zhao, D. S. Small, P. R. Rosenbaum, *Journal of the American Statistical Association* **113**, 1070–1084 (2018).

Y. Zhao, D. Zeng, A. J. Rush, M. R. Kosorok, *Journal of the American Statistical Association* **107**, 1106–1118 (2012).