SCRI 2023 at Academia Sinica

Selective Inference in Propensity Score Analysis

Yoshiyuki Ninomiya

Department of Statistical Inference and Mathematics, The Institute of Statistical Mathematics

December 12th, 2023

1. Preparation (causal model, propensity score analysis, selective inference)
2. Our contribution (model selection method, non-asymptotic polyhedral lemma, asymptotic polyhedral lemma, asymptotic control of false coverage rate)
3. Simulation study, real data analysis, challenge

This is a joint work with Y. Umezu (Nagasaki Univ.) & I. Takeuchi (Nagoya Univ.)

Preparation: causal inference model (1)

Model in which $Y^{(h)}$ $(\in \mathbb{R})$ is observed if h-th treatment is assigned $(1 \le h \le H)$:

$$Y \equiv \sum_{h=1}^{H} T^{(h)} Y^{(h)} = \sum_{h=1}^{H} T^{(h)} \left\{ \mu^{(h)}(\boldsymbol{X}) + f(\boldsymbol{X}) + \epsilon^{(h)} \right\}$$

- $T^{(h)}$ $(\sum_{h=1}^{H} T^{(h)} = 1)$ is assignment variable that is 1 if h-th treatment is assigned and 0 otherwise, $\boldsymbol{X}$ $(\in \mathbb{R}^p)$ is confounding variable vector, $f : \mathbb{R}^p \to \mathbb{R}$ is nonlinear function, $\epsilon^{(h)}$ $(\sim \mathrm{N}(0, \sigma^2))$ is unobservable variable
- Y is observable outcome variable ($Y^{(h)}$ is potential outcome variable)
- $(c^{(1)}, \dots, c^{(H)})'$ is a contrast with $\sum_{h=1}^{H} c^{(h)} = 0$, and our estimand is CATE (conditional average treatment effect) $\sum_{h=1}^{H} c^{(h)} \mu^{(h)}(\boldsymbol{x})$ for $\boldsymbol{X} = \boldsymbol{x}$
  - Naively estimating by e.g. least squares method will lead to bias
  - The propensity score method is a standard one that does not yield such a bias by not estimating $f(\boldsymbol{X})$, which is difficult to model
  - Selective inference in this setting is also treated by Zhang et al. ('22 JRSSB), but it is eventually based on nonparametric estimation of $f(\boldsymbol{X})$

Preparation: causal inference model (2)

Notation: subscript $_i$ is put on variables of i-th sample

- Letting $\tilde{\boldsymbol{Y}} = (Y_1, \ldots, Y_n)'$, $\tilde{\boldsymbol{T}}^{(h)} = \mathrm{diag}(T_1^{(h)}, \ldots, T_n^{(h)})$, $\tilde{\boldsymbol{Y}}^{(h)} = (Y_1^{(h)}, \ldots, Y_n^{(h)})'$, $\tilde{\boldsymbol{X}} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)'$, $\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) = (\mu^{(h)}(\boldsymbol{X}_1), \ldots, \mu^{(h)}(\boldsymbol{X}_n))'$, $\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}}) = (f(\boldsymbol{X}_1), \ldots, f(\boldsymbol{X}_n))'$ and $\tilde{\boldsymbol{\epsilon}}^{(h)} = (\epsilon_1^{(h)}, \ldots, \epsilon_n^{(h)})'$, the model is expressed as

$$\tilde{\boldsymbol{Y}} = \sum_{h=1}^{H} \tilde{\boldsymbol{T}}^{(h)} \tilde{\boldsymbol{Y}}^{(h)} = \sum_{h=1}^{H} \tilde{\boldsymbol{T}}^{(h)} \left\{ \tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) + \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}}) + \tilde{\boldsymbol{\epsilon}}^{(h)} \right\}$$

Assumptions:

- Weak ignorability: $\forall h = \{1, 2, \ldots, H\},\ Y_i^{(h)} \perp\!\!\!\perp T_i^{(h)} \mid \boldsymbol{X}_i$
- Positivity: $0 < \mathrm{P}(T_i^{(h)} = 1 \mid \boldsymbol{X}_i) < 1$
- Independency: if $i \neq j$, $(T_i^{(h)}, \boldsymbol{X}_i, \epsilon_i^{(h)}) \perp\!\!\!\perp (T_j^{(h)}, \boldsymbol{X}_j, \epsilon_j^{(h)})$

Preparation: propensity score analysis

IPW estimation using propensity score $e^{(h)}(\boldsymbol{X}_i) \equiv \mathrm{P}(T_i^{(h)} = 1 \mid \boldsymbol{X}_i)$:

- Missing values are pseudo-recovered by multiplying observed values by the inverse of the propensity score as weights; then usual estimation is conducted
- Letting $\tilde{\boldsymbol{W}}^{(h)}(\tilde{\boldsymbol{T}}^{(h)}, \tilde{\boldsymbol{X}}) \equiv \mathrm{diag}\{T_1^{(h)}/e^{(h)}(\boldsymbol{X}_1), \ldots, T_n^{(h)}/e^{(h)}(\boldsymbol{X}_n)\}$ and $\tilde{\boldsymbol{T}} = (\tilde{\boldsymbol{T}}^{(1)}, \ldots, \tilde{\boldsymbol{T}}^{(H)})$, and defining $\tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \equiv \sum_{h=1}^{H} c^{(h)} \tilde{\boldsymbol{W}}^{(h)}(\tilde{\boldsymbol{T}}^{(h)}, \tilde{\boldsymbol{X}})$ as weight matrix, IPW (inverse probability weighted) estimator is given by minimizing the following weighted squared loss

$$\left\| \tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})\tilde{\boldsymbol{Y}} - \sum_{h=1}^{H} c^{(h)} \tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) \right\|_2^2$$

- It has consistency under assumptions such as weak ignorability

Preparation: selective inference (1)

Explanation using non-causal model:

- Superscript $^{(h)}$ is omitted as $H = 1$ (i.e. $\tilde{\boldsymbol{T}}$ is $n \times n$ identity matrix $\boldsymbol{I}_n$), let $\tilde{\boldsymbol{X}}$ be a non-random matrix $\tilde{\boldsymbol{x}}$, and let $\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{x}})$ be n-dimensional zero vector $\mathbf{0}_n$
- After variable selection, to measure the extent to which the selected variables have an impact on causal effect, tests are performed or confidence intervals are constructed; however, p-values used in this process are no longer reliable
  - It is because the selected variables are likely to be significant, or in other words, the model using the selected variables likely overfits the data
- A linear function of $\tilde{\boldsymbol{x}}$ is supposed as a model for $\tilde{\boldsymbol{\mu}}(\tilde{\boldsymbol{x}})$; then, for each subset $\boldsymbol{M} \subset \{1, \ldots, p\}$, we consider estimand as follows:

$$\boldsymbol{\beta}^{\ddagger \boldsymbol{M}} \equiv \operatorname*{argmin}_{\boldsymbol{b}^{\ddagger \boldsymbol{M}}} \mathrm{E}\left(\left\| \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{x}}_{\boldsymbol{M}} \boldsymbol{b}^{\ddagger \boldsymbol{M}} \right\|_2^2 \right) = \left(\tilde{\boldsymbol{x}}'_{\boldsymbol{M}} \tilde{\boldsymbol{x}}_{\boldsymbol{M}} \right)^{-1} \tilde{\boldsymbol{x}}'_{\boldsymbol{M}} \tilde{\boldsymbol{\mu}}(\tilde{\boldsymbol{x}})$$

  - $\tilde{\boldsymbol{x}}_{\boldsymbol{M}} = (\tilde{x}_{ij})_{i \in \{1,\ldots,n\}, j \in \boldsymbol{M}}$
- If models $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are different, then $\beta_j^{\ddagger \boldsymbol{M}_1}$ and $\beta_j^{\ddagger \boldsymbol{M}_2}$ are generally different, i.e., the target of inference is different for each model selected; inference after model selection has some ambiguity

Preparation: selective inference (2)

Two methods resolving the ambiguity in confidence interval construction:

- Simultaneous inference is a method of creating an interval in which all the regression coefficients are included with probability $\geq 1-\alpha$ regardless of which model is selected
- Selective inference is a method of creating an interval in which the regression coefficients in the selected model are included with probability $\geq 1-\alpha$ under the condition that the model was selected; it is $C_j^{\ddagger\boldsymbol{M}}$ $(j \in \boldsymbol{M})$ such that

$$\mathrm{P}\left(\beta_j^{\ddagger\boldsymbol{M}} \in C_j^{\ddagger\boldsymbol{M}} \mid \hat{\boldsymbol{M}} = \boldsymbol{M}\right) \geq 1-\alpha \quad (\hat{\boldsymbol{M}} \text{ is selected model})$$

- After Lee & Taylor ('14 NeurIPS) and Lee et al. ('16 AS) presented beautiful inferences for marginal screening and LASSO, and Taylor & Tibshirani ('15 PNAS) explained its great potential, selective inference is rapidly developed
- Here we develop it for propensity score analysis based on Lee et al. ('16 AS)

Preparation: method in Lee et al. ('16 AS)

Selective inference in non-causal model:

- Let us denote usual LASSO estimator by $\hat{\boldsymbol{\beta}}^{\ddagger} = (\hat{\beta}_1^{\ddagger}, \ldots, \hat{\beta}_p^{\ddagger})'$, the collection of non-zero estimators by $\hat{\boldsymbol{\beta}}^{\ddagger\hat{\boldsymbol{M}}}$, and its sign by $\hat{\boldsymbol{s}}^{\ddagger\hat{\boldsymbol{M}}} = \mathrm{sign}(\hat{\boldsymbol{\beta}}^{\ddagger\hat{\boldsymbol{M}}})$; from Karuch-Kuhn-Tucker conditions, there exists an $n \times n$ matrix $\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})$ and an n dimensional vector $\boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s})$ such that

$$\forall \boldsymbol{s} \in \{-1, 1\}^{|\boldsymbol{M}|},\ \left\{\hat{\boldsymbol{M}}^{\ddagger} = \boldsymbol{M},\ \hat{\boldsymbol{s}}^{\ddagger\hat{\boldsymbol{M}}} = \boldsymbol{s}\right\} = \left\{\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\tilde{\boldsymbol{Y}} \leq \boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s})\right\}$$

- Using a unit vector $\boldsymbol{e}_j$ $(\in \mathbb{R}^{|\boldsymbol{M}|})$, we define $\tilde{\boldsymbol{\eta}}_j^{\ddagger} \equiv \tilde{\boldsymbol{x}}_{\boldsymbol{M}}(\tilde{\boldsymbol{x}}_{\boldsymbol{M}}'\tilde{\boldsymbol{x}}_{\boldsymbol{M}})^{-1}\boldsymbol{e}_j$; since the target parameter can be written as $\beta_j^{\ddagger\boldsymbol{M}} = \tilde{\boldsymbol{\eta}}_j^{\ddagger\prime}\tilde{\boldsymbol{\mu}}(\tilde{\boldsymbol{x}})$, we use $\tilde{\boldsymbol{\eta}}_j^{\ddagger\prime}\tilde{\boldsymbol{Y}}$ to create its confidence interval and obtain the following polyhedral lemma:

$$F^{[\mathcal{V}^-_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}}), \mathcal{V}^+_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}})]}_{\beta_j^{\ddagger\boldsymbol{M}}, \sigma^2\tilde{\boldsymbol{\eta}}_j^{\ddagger\prime}\tilde{\boldsymbol{\eta}}_j^{\ddagger}}\left(\tilde{\boldsymbol{\eta}}_j^{\ddagger\prime}\tilde{\boldsymbol{Y}}\right)\ \Big|\ \left\{\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\tilde{\boldsymbol{Y}} \leq \boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s})\right\} \sim \mathrm{Unif}(0, 1)$$

  - $F^{[a,b]}_{\mu,\sigma^2}$ denotes the c.d.f. of $\mathrm{N}(\mu, \sigma^2)$ truncated into $[a, b]$, $\boldsymbol{c}_j^{\ddagger} \equiv \tilde{\boldsymbol{\eta}}_j^{\ddagger}(\tilde{\boldsymbol{\eta}}_j^{\ddagger\prime}\tilde{\boldsymbol{\eta}}_j^{\ddagger})^{-1}$, $\mathcal{V}^-_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}}) = \max_{k:(\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\boldsymbol{c}_j^{\ddagger})_k < 0}\{b(\boldsymbol{M}, \boldsymbol{s})_k - (\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\tilde{\boldsymbol{Z}})_k\}/(\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\boldsymbol{c}_j^{\ddagger})_k$, $\mathcal{V}^+_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}}) = \min_{k:(\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\boldsymbol{c}_j^{\ddagger})_k > 0}\{b(\boldsymbol{M}, \boldsymbol{s})_k - (\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\tilde{\boldsymbol{Z}})_k\}/(\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\boldsymbol{c}_j^{\ddagger})_k$, $\tilde{\boldsymbol{Z}} = \{\boldsymbol{I}_n - \boldsymbol{c}_j^{\ddagger}\tilde{\boldsymbol{\eta}}_j^{\ddagger\prime}\}\tilde{\boldsymbol{Y}}$, $\mathrm{Unif}(0, 1)$ denotes uniform distribution on $[0, 1]$

Our purpose

Selective inference in causal inference model:

- As a model for $\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}})$, we consider linear sum of $\tilde{\boldsymbol{X}}_{\boldsymbol{M}}$ (confounding variables belonging to $\boldsymbol{M} \subset \{1, \dots, p\}$), and because the causal effect is $\sum_{h=1}^{H} c^{(h)} \mu^{(h)}(\boldsymbol{X})$, we define an estimand as

$$\boldsymbol{\beta}^{\boldsymbol{M}} \equiv \operatorname*{argmin}_{\boldsymbol{b}^{\boldsymbol{M}}} \mathrm{E}\left(\left\| \sum_{h=1}^{H} c^{(h)} \tilde{\boldsymbol{Y}}^{(h)} - \tilde{\boldsymbol{X}}_{\boldsymbol{M}} \boldsymbol{b}^{\boldsymbol{M}} \right\|_2^2 \,\middle|\, \tilde{\boldsymbol{X}} \right)$$

$$= \left(\tilde{\boldsymbol{X}}_{\boldsymbol{M}} \tilde{\boldsymbol{X}}'_{\boldsymbol{M}} \right)^{-1} \tilde{\boldsymbol{X}}_{\boldsymbol{M}} \sum_{h=1}^{H} c^{(h)} \mu^{(h)}(\tilde{\boldsymbol{X}})$$

- Denoting the selected model as $\hat{\boldsymbol{M}}$, we try to find $C_j^{\boldsymbol{M}}$ $(j \in \boldsymbol{M})$ such that

$$\mathrm{P}\left(\beta_j^{\boldsymbol{M}} \in C_j^{\boldsymbol{M}} \,\middle|\, \hat{\boldsymbol{M}} = \boldsymbol{M} \right) \geq 1 - \alpha$$

- We have $\beta_j^{\boldsymbol{M}} = \boldsymbol{e}'_j (\tilde{\boldsymbol{X}}_{\boldsymbol{M}} \tilde{\boldsymbol{X}}'_{\boldsymbol{M}})^{-1} \tilde{\boldsymbol{X}}_{\boldsymbol{M}} \sum_{h=1}^{H} c^{(h)} \mu^{(h)}(\tilde{\boldsymbol{X}})$ using an appropriate unit vector $\boldsymbol{e}_j$ $(\in \mathbb{R}^{|\boldsymbol{M}|})$; then defining $\tilde{\boldsymbol{\eta}}_j \equiv \tilde{\boldsymbol{X}}_{\boldsymbol{M}} (\tilde{\boldsymbol{X}}'_{\boldsymbol{M}} \tilde{\boldsymbol{X}}_{\boldsymbol{M}})^{-1} \boldsymbol{e}_j$, later we consider the conditional distribution of $\tilde{\boldsymbol{\eta}}'_j \tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \tilde{\boldsymbol{Y}}$

Our model selection method

LASSO for causal inference model:

- Considering that $\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}})$ and then $\sum_{h=1}^{H} c^{(h)} \tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}})$ are linear functions of $\tilde{\boldsymbol{X}}$ in our model, we propose the following:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ \left\| \tilde{\boldsymbol{W}} \left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}} \right) \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \boldsymbol{\beta} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}$$

- Let us denote $\hat{\boldsymbol{M}} = \{j : \ \hat{\beta}_j \neq 0\}$, $\hat{\boldsymbol{\beta}}^{\hat{\boldsymbol{M}}} = (\hat{\beta}_j)_{j \in \hat{\boldsymbol{M}}}$ and $\hat{\boldsymbol{s}}^{\hat{\boldsymbol{M}}} = \operatorname{sign}(\hat{\boldsymbol{\beta}}^{\hat{\boldsymbol{M}}})$; for any model $\boldsymbol{M}$ $(\subset \{1, \ldots, p\})$ and any sign $\boldsymbol{s}$ $(\in \{-1, 1\}^{|\boldsymbol{M}|})$, there exist $n \times n$ matrix $\boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})$ and n dimensional vector $\boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s})$ such that

$$\left\{ \hat{\boldsymbol{M}} = \boldsymbol{M}, \ \hat{\boldsymbol{s}}^{\hat{\boldsymbol{M}}} = \boldsymbol{s} \right\} = \left\{ \boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s}) \tilde{\boldsymbol{W}} \left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}} \right) \tilde{\boldsymbol{Y}} \leq \boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s}) \right\}$$

  from Karuch-Kuhn-Tucker conditions

The first main theorem

Problem unique to selective inference for causal model and its resolution:

- Unlike Lee et al. ('16 AS), $\tilde{\boldsymbol{Y}}$ is not Gaussian and $\tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}})$ exists; also, $\tilde{\boldsymbol{X}}$ is random, although it is a trivial difference
- Since this is conditional inference, we can in fact further condition $\tilde{\boldsymbol{T}} = \tilde{\boldsymbol{t}}$ and $\tilde{\boldsymbol{X}} = \tilde{\boldsymbol{x}}$ to obtain Polyhedral Lemma (this should be a key point)
    - It is unusual to initially condition $\tilde{\boldsymbol{T}} = \tilde{\boldsymbol{t}}$ in propensity score analyses

Theorem 1 (Causal inference model version of non-asymptotic Polyhedral Lemma)

Under notation on the next page, the conditional distribution below is $\mathrm{Unif}(0,1)$

$$F^{[\mathcal{V}^-_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\mathcal{V}^+_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})]}_{\kappa^{\boldsymbol{M}}_j(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\zeta^{\boldsymbol{M}}_j(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})}\left(\tilde{\boldsymbol{\eta}}'_j\tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}}\right)\;\Big|\;\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}}\le\boldsymbol{b}(\boldsymbol{M},\boldsymbol{s})$$

Detailed notation

$F^{[a,b]}_{\mu,\sigma^2}$ is the c.d.f. of $\mathrm{N}(\mu,\sigma^2)$ truncated to the interval $[a,b]$, $\boldsymbol{X}_{\boldsymbol{M}} = (X_j)_{j\in\boldsymbol{M}}$

$$\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}) \equiv \tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})^2\tilde{\boldsymbol{\eta}}_j\{\tilde{\boldsymbol{\eta}}_j'\tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})^2\tilde{\boldsymbol{\eta}}_j\}^{-1}$$

$$\tilde{\boldsymbol{Z}} \equiv \{\boldsymbol{I}_n - \tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})\tilde{\boldsymbol{\eta}}_j'\}\tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})\tilde{\boldsymbol{Y}}$$

$$\mathcal{V}^-_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}) \equiv \max_{k:(\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}))_k<0} \frac{\boldsymbol{b}(\boldsymbol{M},\boldsymbol{s})_k - (\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\tilde{\boldsymbol{Z}})_k}{(\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}))_k}$$

$$\mathcal{V}^+_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}) \equiv \min_{k:(\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}))_k>0} \frac{\boldsymbol{b}(\boldsymbol{M},\boldsymbol{s})_k - (\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\tilde{\boldsymbol{Z}})_k}{(\boldsymbol{A}(\boldsymbol{M},\boldsymbol{s})\tilde{\boldsymbol{D}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}))_k}$$

$$\kappa_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}) \equiv \boldsymbol{e}_j'(\tilde{\boldsymbol{X}}_{\boldsymbol{M}}'\tilde{\boldsymbol{X}}_{\boldsymbol{M}})^{-1}\tilde{\boldsymbol{X}}_{\boldsymbol{M}}'\sum_{h=1}^{H} c^{(h)}\tilde{\boldsymbol{W}}^{(h)}(\tilde{\boldsymbol{T}}^{(h)},\tilde{\boldsymbol{X}})\{\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) + \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}})\}$$

$$\zeta_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}) \equiv \sigma^2\boldsymbol{e}_j'(\tilde{\boldsymbol{X}}_{\boldsymbol{M}}'\tilde{\boldsymbol{X}}_{\boldsymbol{M}})^{-1}\tilde{\boldsymbol{X}}_{\boldsymbol{M}}'\tilde{\boldsymbol{W}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})^2\tilde{\boldsymbol{X}}_{\boldsymbol{M}}(\tilde{\boldsymbol{X}}_{\boldsymbol{M}}'\tilde{\boldsymbol{X}}_{\boldsymbol{M}})^{-1}\boldsymbol{e}_j$$

$$\tau_j^{\boldsymbol{M}}(\tilde{\boldsymbol{Y}}^\dagger,\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}) \equiv \boldsymbol{e}_j'(\tilde{\boldsymbol{X}}_{\boldsymbol{M}}'\tilde{\boldsymbol{X}}_{\boldsymbol{M}})^{-1}\tilde{\boldsymbol{X}}_{\boldsymbol{M}}'\sum_{h=1}^{H} c^{(h)}\{\tilde{\boldsymbol{W}}^{(h)}(\tilde{\boldsymbol{T}}^{(h)},\tilde{\boldsymbol{X}}) - \boldsymbol{I}_n\}\tilde{\boldsymbol{Y}}^{(h)\dagger}$$

$$\rho_j^{\boldsymbol{M}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right) \equiv \zeta_j^{\boldsymbol{M}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right) + \tilde{\boldsymbol{\eta}}_j'\sum_{h=1}^{H} c^{(h)2}\mathrm{diag}\left\{\frac{\sigma^2}{|\mathcal{N}_i^{(h)}|}\frac{1-e^{(h)}(\boldsymbol{X}_i)}{e^{(h)}(\boldsymbol{X}_i)}\right\}\tilde{\boldsymbol{\eta}}_j$$

The second main theorem

Problem of non-asymptotic Polyhedral Lemma and its resolution:

- Theorem 1 cannot be used for the inference about $\beta_j^{\boldsymbol{M}}$ without further consideration, because $\beta_j^{\boldsymbol{M}}$ does not appear explicitly in the pivot statistic; we consider higher-order asymptotics and extract $\beta_j^{\boldsymbol{M}}$ from $\kappa_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})$
- From calculations frequently used under ignorability condition and the fact that the sum of contrasts $\sum_{h=1}^{H} c^{(h)}$ is 0, we obtain the following

Theorem 2 (Causal inference model version of asymptotic Polyhedral Lemma)

If $\tilde{\boldsymbol{Y}}^{(h)\dagger} - \{\tilde{\boldsymbol{\mu}}^{(h)}(\tilde{\boldsymbol{X}}) + \tilde{\boldsymbol{f}}(\tilde{\boldsymbol{X}})\} = \mathrm{o}_{\mathrm{P}}(1)$ under the condition, the conditional distribution below is $\mathrm{Unif}(0,1)$ asymptotically ($\tilde{\boldsymbol{Y}}^{\dagger} = (\tilde{\boldsymbol{Y}}^{(1)\dagger}, \ldots, \tilde{\boldsymbol{Y}}^{(H)\dagger})$)

$$F^{[\mathcal{V}^{-}_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\mathcal{V}^{+}_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})]}_{\beta_j^{\boldsymbol{M}}+\tau_j^{\boldsymbol{M}}(\tilde{\boldsymbol{Y}}^{\dagger},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\zeta_j^{\boldsymbol{M}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})} \left(\tilde{\boldsymbol{\eta}}_j' \tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}}\right) \,\Big|\, \boldsymbol{A}(\boldsymbol{M}, \boldsymbol{s})\tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}} \le \boldsymbol{b}(\boldsymbol{M}, \boldsymbol{s})$$

- Note that $\tau_j^{\boldsymbol{M}}(\tilde{\boldsymbol{Y}}^{\dagger}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}) \xrightarrow{\mathrm{p}} 0$, but this higher order correction is necessary

Automatic development and accecible form

Development for increasing power using a union of intervals:

- Similarly to Lee et al. ('16 AS), if we want to condition $\hat{\boldsymbol{M}} = \boldsymbol{M}$, we have only to use $\mathrm{N}(\mu, \sigma^2)$ truncated into $S \equiv \bigcup_s [\mathcal{V}^-_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}}), \mathcal{V}^+_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}}, \tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}})]$; denoting the c.d.f. by F^S_{μ,σ^2}, we obtain

$$\left\{ F^{\bigcup_{\boldsymbol{s}}[\mathcal{V}^-_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\mathcal{V}^+_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})]}_{\beta^{\boldsymbol{M}}_j + \tau^{\boldsymbol{M}}_j(\tilde{\boldsymbol{Y}}^\dagger,\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\zeta^{\boldsymbol{M}}_j(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})} \left(\tilde{\boldsymbol{\eta}}'_j \tilde{\boldsymbol{W}} \left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}} \right) \tilde{\boldsymbol{Y}} \right) \;\middle|\; \hat{\boldsymbol{M}} = \boldsymbol{M} \right\} \xrightarrow{\mathrm{d}} \mathrm{Unif}(0,1)$$

Form of confidence interval:

- Since this pivot statistic is a monotonically decreasing function with respect to $\beta^{\boldsymbol{M}}_j$, if we set L or U to satisfy

$$F^{\bigcup_{\boldsymbol{s}}[\mathcal{V}^-_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\mathcal{V}^+_{\boldsymbol{s},j}(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})]}_{L \text{ or } U + \tau^{\boldsymbol{M}}_j(\tilde{\boldsymbol{Y}}^\dagger,\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\zeta^{\boldsymbol{M}}_j(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})} \left(\tilde{\boldsymbol{\eta}}'_j \tilde{\boldsymbol{W}} \left(\tilde{\boldsymbol{T}}, \tilde{\boldsymbol{X}} \right) \tilde{\boldsymbol{Y}} \right) = 1 - \frac{\alpha}{2} \text{ or } \frac{\alpha}{2},$$

  we get an asymptotically guaranteed conditional coverage as follows:

$$\mathrm{P} \left(\beta^{\boldsymbol{M}}_j \in [L, U] \;\middle|\; \hat{\boldsymbol{M}} = \boldsymbol{M} \right) \to 1 - \alpha$$

Corollary

Asymptotics a little faster than usual:

- We skipped to define $\tilde{\boldsymbol{Y}}^\dagger$ because of needless tedium, but we can define it appropriately and obtain the followings from ignorability conditions

$$\mathrm{E}\left\{\tilde{\boldsymbol{\eta}}_j'\tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}}-\tau_j^{\boldsymbol{M}}\left(\tilde{\boldsymbol{Y}}^\dagger,\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\;\middle|\;\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right\}=\beta_j^{\boldsymbol{M}}+\mathrm{o_P}(n^{-5/6})$$

$$\mathrm{V}\left\{\tilde{\boldsymbol{\eta}}_j'\tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}}-\tau_j^{\boldsymbol{M}}\left(\tilde{\boldsymbol{Y}}^\dagger,\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\;\middle|\;\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right\}=\rho_j^{\boldsymbol{M}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)+\mathrm{o_P}(n^{-5/3})$$

Corollary (Asymptotic control of false coverage rate)

Under notation on page 12, if we set $L_j^{\hat{\boldsymbol{M}}}$ and $U_j^{\hat{\boldsymbol{M}}}$ to satisfy

$$F_{L_j^{\hat{\boldsymbol{M}}}+\tau_j^{\hat{\boldsymbol{M}}}(\tilde{\boldsymbol{Y}}^\dagger,\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\rho_j^{\hat{\boldsymbol{M}}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})}^{\bigcup_{\boldsymbol{s}}[\mathcal{V}_{\boldsymbol{s},j}^-(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\mathcal{V}_{\boldsymbol{s},j}^+(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})]}\left(\tilde{\boldsymbol{\eta}}_j'\tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}}\right)=1-\frac{\alpha}{2}$$

$$\text{and}\quad F_{U_j^{\hat{\boldsymbol{M}}}+\tau_j^{\hat{\boldsymbol{M}}}(\tilde{\boldsymbol{Y}}^\dagger,\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\rho_j^{\hat{\boldsymbol{M}}}(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})}^{\bigcup_{\boldsymbol{s}}[\mathcal{V}_{\boldsymbol{s},j}^-(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}),\mathcal{V}_{\boldsymbol{s},j}^+(\tilde{\boldsymbol{Z}},\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}})]}\left(\tilde{\boldsymbol{\eta}}_j'\tilde{\boldsymbol{W}}\left(\tilde{\boldsymbol{T}},\tilde{\boldsymbol{X}}\right)\tilde{\boldsymbol{Y}}\right)=\frac{\alpha}{2},$$

we can control asymptotically the false coverage rate as follows:

$$\lim_{n\to\infty}\mathrm{E}\left(|\{j\in\hat{\boldsymbol{M}}:\beta_j^{\hat{\boldsymbol{M}}}\notin[L_j^{\hat{\boldsymbol{M}}},U_j^{\hat{\boldsymbol{M}}}]\}|/|\hat{\boldsymbol{M}}|;\ |\hat{\boldsymbol{M}}|>0\right)\leq\alpha$$

Simulation study: Evaluation of coverages

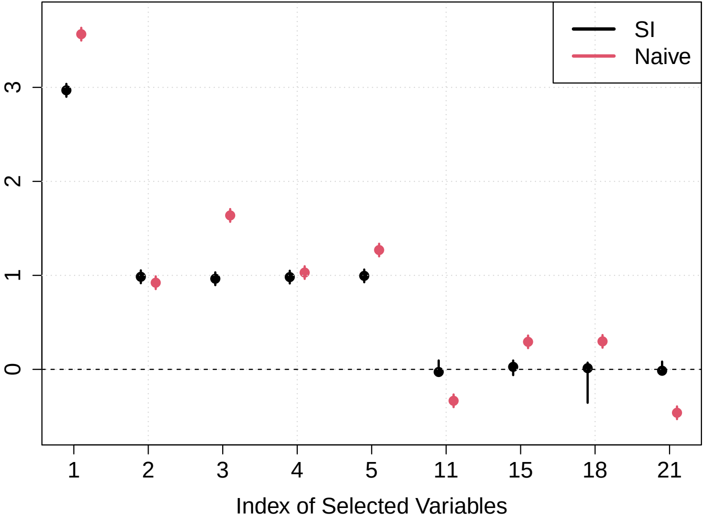
Comparison between SI (selective inference) and Na (naive inference)
f: 0 or L (linear) or nL (non-linear), μ: linear, e: logistic
X: discrete or continuous, n: 1000, p: 25 (20 zeros, 5 non-zeros)

		X: discrete				X: continuous			
f		$\lvert\hat{M}\rvert$	TP	FP	FCR	$\lvert\hat{M}\rvert$	TP	FP	FCR
0	SI	13.841 (2.322)	4.979 (0.211)	0.405 (0.729)	0.048 (0.080)	5.287 (0.539)	4.994 (0.089)	0.013 (0.113)	0.053 (0.098)
	Na		5.000 (0.000)	8.841 (2.322)	0.874 (0.089)		5.000 (0.000)	0.287 (0.539)	0.491 (0.229)
L	SI	20.380 (1.906)	4.938 (0.280)	0.834 (0.935)	0.053 (0.056)	11.844 (2.121)	4.780 (0.498)	0.391 (0.673)	0.054 (0.077)
	Na		4.954 (0.214)	15.425 (1.887)	0.968 (0.038)		4.818 (0.435)	7.026 (2.085)	0.904 (0.084)
nL	SI	17.238 (2.244)	4.963 (0.309)	0.647 (1.028)	0.052 (0.075)	7.050 (1.361)	4.974 (0.222)	0.101 (0.359)	0.054 (0.099)
	Na		5.000 (0.000)	12.238 (2.244)	0.933 (0.059)		4.997 (0.055)	2.053 (1.359)	0.724 (0.174)

TP: # of true positives, FP: # of false positives, FCR: false coverage rate

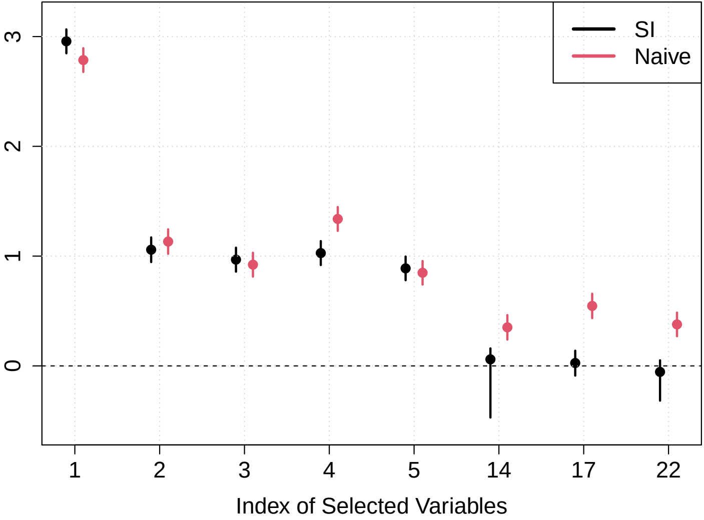
Simulation study: Confidence interval (X: discrete)

Comparison between SI (selective inference) and Naive (non selective inference)
f: non-linear, μ: linear, e: logistic, n: 1000, p: 25 (20 zeros, 5 non-zeros)



Simulation study: Confidence interval (X: continuous)

Comparison between SI (selective inference) and Naive (non selective inference)
f: non-linear, μ: linear, e: logistic, n: 1000, p: 25 (20 zeros, 5 non-zeros)



Real data analysis (Benchmark lalonde dataset)

- Causal effect is set as the difference in annual income in 1978 between the group that took the U.S. job training program in 1976 and the group that did not take it
- By LASSO, 5 variables were selected among 10 variables, and 95% confidence intervals by SI and Naive are compared below
- SI does not regard re74 as significant since its confidence interval contains 0, while Naive does; we do not know which is true, but anyway their results are different

	SI			Naive		
selected	estimates	lower	upper	estimates	lower	upper
age	49.353	−164.745	165.970	51.345	−66.302	168.991
educ	721.188	228.418	1196.124	862.309	390.500	1334.117
re74	0.240	−0.173	0.479	0.236	0.019	0.454
re75	0.275	−0.341	0.645	0.293	−0.040	0.625
u74	6459.088	3709.181	8956.314	6480.235	3997.591	8962.880

Notes not mentioned

Estimation of variance σ^2 of $\epsilon_i^{(h)}$:

- We only have to substitute a consistent estimator, but since the model contains $f(\cdot)$ and we are trying to develop a method that avoids its identification, its determination is more difficult than usual
- Let us choose an appropriate sequence of real numbers $\{\delta_n\}$ that converges to 0, and define $\mathcal{N}_i^{\dagger} = \{l \neq i : \ \|\boldsymbol{X}_l - \boldsymbol{X}_i\|_2 < \delta_n, \ \boldsymbol{T}_l = \boldsymbol{T}_i\}$ using $\boldsymbol{T}_i = (T_i^{(1)}, \ldots, T_i^{(H)})$; then we use

$$\frac{1}{n}\sum_{i=1}^{n} \frac{|\mathcal{N}_i^{\dagger}|}{1+|\mathcal{N}_i^{\dagger}|}\left(Y_i - \frac{1}{|\mathcal{N}_i^{\dagger}|}\sum_{l\in\mathcal{N}_i^{\dagger}} Y_l\right)^2$$

Estimation of propensity scores:

- We only have to use $\hat{e}^{(h)}(\boldsymbol{X}_i)$ such that $\hat{e}^{(h)}(\boldsymbol{X}_i) - e^{(h)}(\boldsymbol{X}_i) = \mathrm{O_P}(n^{-1/2})$
- Note that in our unique conditioning, $\hat{e}^{(h)}(\boldsymbol{X}_i)$ is non-random and thus does not require special care to obtain Theorems 1 and 2
- A special care, specifically changing some expressions of the asymptotic order, is necessary to obtain Corollary, but anyway it still holds

Summary and challenges

1. Conditioning on the assignment variables first (unusual in propensity score analysis) and using higher-order asymptotics (unusual in selective inference), we have developed asymptotically guaranteed post-selection inference

2. Numerical experiments showed that a method that ignores the quiet scandal of statistics results in significant deviations from the preset coverage of the confidence intervals, whereas our method maintains the coverage

3. Since this result cannot handle even GLM-based causal inference, we would like to develop the results of Charkhi & Claeskens ('18 Biometrika) into a causal inference version

4. The fact that the proposal is based on a formula that seems complicated hinders its widespread use among users (while the calculation cost is not high), so the development of package and manual is an urgent issue