Information projection approach to propensity score estimation for correcting selection bias

Jae-kwang Kim[1]
Iowa State University

May 23, 2022
BIRS workshop on

[1] Joint work with Hengfang Wang

Motivating Example (Kim et al., 2019)

- Korean Workplace Panel Surveys (sponsored by Korean Labor Institute)
- They are interested in fitting a regression from the sample:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

  where
  - Y: log(Sale)/Person
  - X_1: Size of company (= number of employees)
  - X_2: Type of company
- (X_1, X_2) are always observed
- Y: subject to missingness

Motivating Example

- In addition to (X_1, X_2, Y), the survey company collected a paradata variable Z regarding the respondents' reaction

$$Z = \begin{cases} 1 & \text{friendly response} \\ 2 & \text{moderate response} \\ 3 & \text{negative response} \end{cases}$$

- The response rate is significantly low for units with $Z = 3$.
- The response rates are 0.71, 0.67, and 0.45 for $Z = 1$, $Z = 2$, and $Z = 3$, respectively.

Motivating Example

- The variable Z is a strong predictor for the response mechanism but it is not a good predictor for Y.
- In fact, the regression coefficient for Z in the regression model

$$Y = X\beta + Z\gamma + e$$

  is not significant (p-value $= 0.70$)
- Question: Should we include Z into the nonresponse adjustment weighting?

Introduction

- (X, Y): a vector of random variables satisfying

$$\mathbb{E}\left\{U(\theta_0; X, Y)\right\} = 0$$

for some function $U(\cdot; x, y)$ with unknown parameter $\theta_0 \in \Theta \in \mathbb{R}^p$.

- That is, the model with distribution function P should satisfy

$$\mathbb{E}\left\{U(\theta; X, Y)\right\} \equiv \int U(\theta; x, y) dP(x, y) = 0 \quad (1)$$

for all θ, where P is completely unspecified other than the restriction in (1). Thus, it is a semiparametric model.

- There are infinitely many P satisfying (1) for given θ. The model space $\mathcal{L}(\theta) = \{P; \int U(\theta; x, y) dP(x, y) = 0\}$ depends on θ.

Dual problem

- The Kullback-Leibler (KL) divergence of P with respect to Q is

$$D(P \parallel Q) = \int \log \left\{ \frac{dP(x,y)}{dQ(x,y)} \right\} dP(x,y).$$

- We are interested in finding P^* that minimizes $D(P \parallel \hat{P})$ among $P \in \mathcal{L}(\theta)$, where $\hat{P}$ is the empirical distribution in the sample.
- Note that

$$D(P \parallel \hat{P}) = \int P(x,y) \log \left\{ \frac{P(x,y)}{\hat{P}(x,y)} \right\} d\mu(x,y). \tag{2}$$

  Thus, to avoid $D(P \parallel \hat{P}) = \infty$, we set $P^*(x,y) = 0$ for any point with $\hat{P}(x,y) = 0$.
- The problem is equivalent to finding the minimizer of $D(\mathbf{p}) = \sum_{i=1}^N p_i \log(p_i)$ subject to $\sum_{i=1}^N p_i = 1$ and $\sum_{i=1}^N p_i U(\theta; y_i) = 0$.

ETEL estimation (Schennach, 2007)

Two-step estimation

1. ET step: Finding the minimizer of $D(P \parallel \hat{P})$ among $P \in \mathcal{L}(\theta)$ to get

$$p_i^*(\theta) = \frac{\exp\{\hat{\lambda}_\theta' U(\theta; x_i, y_i)\}}{\sum_{i=1}^N \exp\{\hat{\lambda}_\theta' U(\theta; x_i, y_i)\}}, \quad (3)$$

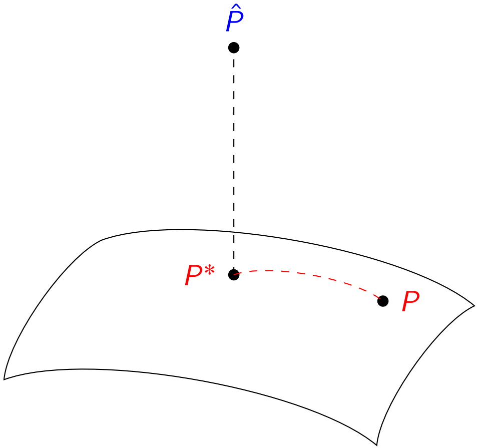
where $\hat{\lambda}_\theta$ satisfies $\sum_{i=1}^N p_i^*(\theta) U(\theta; x_i, y_i) = 0$.

2. EL step: To estimate the model parameter, we find the minimizer of $D(\hat{P} \parallel P^*)$. That is, find the maximizer of

$$\ell_p(\theta) = \frac{1}{N} \sum_{i=1}^N \log\{p_i^*(\theta)\}$$

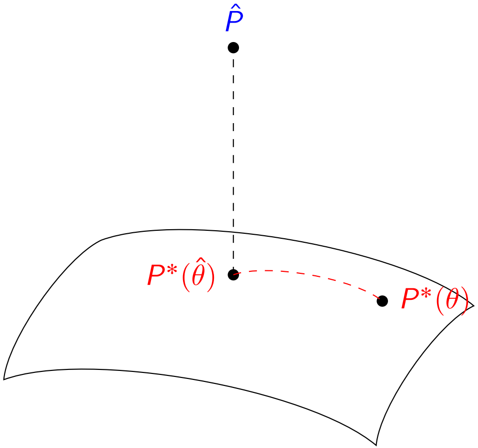
where $p_i^*(\theta)$ is defined in (3).

Graphical Illustration (for ET step)



KL divergence $D(P \parallel \hat{P})$ among $P \in \mathcal{L}(\theta)$ is minimized at $P^*(\theta)$ in (3).

Graphical Illustration (for EL step)



The KL divergence $D(\hat{P} \parallel P^*(\theta))$ among $\theta \in \Theta$ is minimized at $\theta = \hat{\theta}$.

Remark

- The first step is a modeling step: Use I-projection to obtain a dual expression of the model. The dual model is an exponential tilting form.
- The second step is an estimation step: Use maximum likelihood estimation of the parameters in the exponential tilting model.

Non-probability sample

- Two-phase sampling structure:
  1. Phase 1: A finite population of (x_i, y_i) follows a distribution P satisfying the semiparametric model (1).
  2. Phase 2: From the finite population, we obtain a sample S by an unknown sampling mechanism and observe (x_i, y_i) in the sample.
- Assume that x_i are observed throughout the finite population with index set $\{1, \cdots, N\}$.
- It is essentially a missing data setup where the sampling mechanism corresponds to the response mechanism.

Density ratio (DR) function

- P_k: probability distribution of (X, Y) conditional on $\delta = k$ for $k = 0, 1$, where $\delta_i = 1$ if $i \in S$ and $\delta_i = 0$ otherwise.
- $P_k \ll \mu$, with density $f_k = dP_k/d\mu$.
- The ratio of two density functions

$$\frac{f_0(x, y)}{f_1(x, y)} := r(x, y)$$

  is called the density ratio function.
- Using the density ratio (DR) function, the probability of an event B at P_0 can be expressed as an integration evaluated at P_1:

$$\mathbb{P}_0\{(X, Y) \in B\} = \int \mathbb{I}\{(x, y) \in B\} r(x, y) dP_1(x, y).$$

Alternative expression for the model assumption

- Recall that the model space that we are interested in is

$$\mathcal{L}(\theta) = \{P; \mathbb{E}\{U(\theta; X, Y)\} = 0\}.$$

- Using the DR function $r(x, y)$, we can express

$$\begin{aligned} & \mathbb{E}\{U(\theta; X, Y)\} \\ = \;& p \int U(\theta; x, y) dP_1(x, y) + (1-p) \int U(\theta; x, y) dP_0(x, y) \\ = \;& p \int U(\theta; x, y) dP_1(x, y) + (1-p) \int U(\theta; x, y) r(x, y) dP_1(x, y) \\ = \;& \int \{p + (1-p) r(x, y)\} U(\theta; x, y) dP_1(x, y) \end{aligned}$$

where $p = P(\delta = 1)$ is the proportion of sample in the finite population.

Alternative expression for the model assumption

- Thus, when $r(x, y)$ is known, the model space $\mathcal{L}$ has an one-to-one correspondence with

$$\mathcal{L}_1(\theta) = \left\{ P_1 : \int \{1 + (N_0/N_1) r(x, y)\} U(\theta; x, y) dP_1(x, y) = 0 \right\},$$

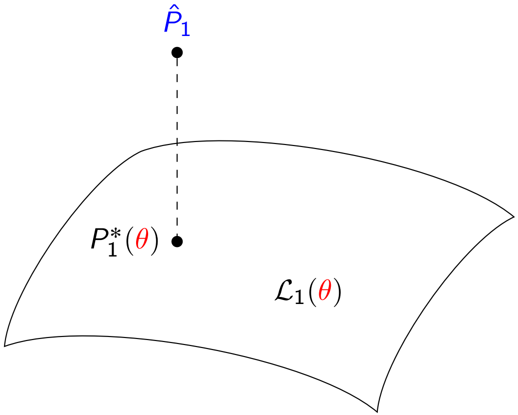
where $N_k = \sum_{i=1}^{N} \mathbb{I}(\delta_i = k)$ for $k = 0, 1$.

- We can apply the I-projection on $\mathcal{L}_1(\theta)$ to obtain $p^*(\theta)$. That is, use

$$\hat{P}_1(x, y) = \frac{1}{N_1} \sum_{i=1}^{N} \delta_i \mathbb{I}\{(x, y) = (x_i, y_i)\}$$

to find the minimizer of $D(P_1 \parallel \hat{P}_1)$ among $P_1 \in \mathcal{L}(\theta)$.

Graphical Illustration (Only $\hat{P}_1$ is observed)



The KL divergence $D(P_1 \parallel \hat{P}_1)$ among $P_1 \in \mathcal{L}_1(\theta)$ is minimized at P_1^*.

- Thus, the problem reduces to finding the maximizer of

$$\ell(\mathbf{p}) = \sum_{i \in S} p_i \log(p_i)$$

subject to $\sum_{i \in S} p_i = 1$ and

$$\sum_{i \in S} p_i \{1 + (N_0/N_1) r(x_i, y_i)\} U(\theta; x_i, y_i) = 0. \tag{4}$$

- If the dimension of θ is equal to the rank of the estimating function $U(\theta; x, y)$, then it is just-identified and equation (4) does not contain any extra information. In this case, condition (4) can be safely ignored in the optimization for $\mathbf{p}$.
- Using $\hat{p}_i = 1/N_1$ in (4) leads to a weighted estimating equation with weight

$$\omega(x, y) = 1 + \frac{N_0}{N_1} \cdot r(x, y).$$

Propensity score (PS) weight function

- Propensity score weight function is computed from the DR function:

$$\omega(x, y) = 1 + \frac{N_0}{N_1} \cdot r(x, y) = \frac{1}{\mathbb{P}(\delta = 1 \mid x, y)}.$$

- Propensity score weight function is used to estimate parameters from the sample with selection bias:

$$\hat{U}_{PS}(\theta) \equiv \sum_{i \in S} \omega(x_i, y_i) U(\theta; x_i, y_i) = 0.$$

- Two problems
  1. In practice, $r(x, y)$ is unknown.
  2. Even if $r(x, y)$ is known, it does not necessarily lead to efficient estimation for θ.

Simplifying assumption

- To avoid any issues on model identifiability, we consider MAR (missing at random) assumption of Rubin (1976):

$$Y \perp \delta \mid X.$$

- Under MAR,

$$r(x,y) = \frac{f_0(x,y)}{f_1(x,y)} = \frac{f_0(x)}{f_1(x)} \cdot \frac{f_0(y \mid x)}{f_1(y \mid x)} = \frac{f_0(x)}{f_1(x)} = r(x)$$

and

$$\omega(x) = 1 + \frac{N_0}{N_1} \cdot r(x).$$

Weight smoothing: Idea

- Instead of using

$$\hat{U}_{PS}(\theta) \equiv \sum_{i=1}^{N} \delta_i \omega(x_i) U(\theta; x_i, y_i) = 0,$$

we may use

$$\hat{U}_{SPS}(\theta) \equiv \sum_{i=1}^{N} \delta_i \omega^*(x_i) U(\theta; x_i, y_i) = 0,$$

where

$$\omega^*(x) = \mathbb{E}_1 \{\omega(x) \mid U(\theta; x, y)\} \tag{5}$$

and $\mathbb{E}_1(\cdot) = \mathbb{E}(\cdot \mid \delta = 1)$.

- We can show that

$$\mathbb{E}\{\hat{U}_{PS}(\theta)\} = \mathbb{E}\{\hat{U}_{SPS}(\theta)\} \text{ and } \mathbb{V}\{\hat{U}_{PS}(\theta)\} \geqslant \mathbb{V}\{\hat{U}_{SPS}(\theta)\}.$$

How to compute (5) in practice?

- First, we can show that

$$\mathbb{E}_1\left\{\omega(x) \mid U(\theta; x, y)\right\} = \mathbb{E}_1\left\{\omega(x) \mid \bar{U}(\theta; x)\right\}$$

  where $\bar{U}(\theta; \mathbf{x}) = \mathbb{E}\{U(\theta; X, Y) \mid \mathbf{x}\}$.

- Next, find the linear space $\mathcal{H}$ such that

$$\bar{U}(\theta; \mathbf{x}) \in \text{span}\{b_1(\mathbf{x}), \cdots, b_L(\mathbf{x})\} := \mathcal{H} \tag{6}$$

  holds.

- Thus, the smoothed propensity score weight in (5) reduces to

$$\omega^*(x) = \mathbb{E}_1\left\{\omega(x) \mid \mathcal{H}\right\}. \tag{7}$$

How to compute the smoothed weight function in (7)?

- We wish to minimize

$$D(f_0 \parallel f_1) = \int \log\left(f_0/f_1\right) f_0 d\mu, \tag{8}$$

w.r.t. f_0 such that $\int f_0 d\mu = 1$, and some moment constraints

- The linear space that we are projecting on is

$$\frac{N_1}{N} \int \boldsymbol{b}(x) f_1(x) d\mu + \frac{N_0}{N} \int \boldsymbol{b}(x) f_0(x) d\mu = \mathbb{E}\{\boldsymbol{b}(X)\}, \tag{9}$$

where $\boldsymbol{b}(x)$ is the basis functions in $\mathcal{H}$.

- The I-projection solution is

$$f_0^*(x) = f_1(x) \times \frac{\exp\{\phi_1' \boldsymbol{b}(x)\}}{\mathbb{E}_1\left[\exp\{\phi_1' \boldsymbol{b}(x)\}\right]}, \tag{10}$$

where ϕ_1 is the Lagrange multiplier satisfying (9).

- Expression (10) leads to a parametric density ratio model:

$$\log\{r^*(x)\} = \phi_0 + \phi_1 b_1(x) + \cdots + \phi_L b_L(x). \tag{11}$$

Model (11) can be called the log-linear density ratio model.

- The model parameters should satisfy the original constraint in (9). Thus,

$$\frac{N_1}{N} \int \boldsymbol{b}(x) \left[1 + \frac{N_0}{N_1} \cdot \exp\{\phi_0 + \phi_1' \boldsymbol{b}(x)\} \right] f_1(x) d\mu = \mathbb{E}\{\boldsymbol{b}(X)\}, \tag{12}$$

where ϕ_0 satisfies

$$\int \exp\{\phi_0 + \phi_1' \boldsymbol{b}(x)\} f_1(x) d\mu = 1. \tag{13}$$

Parameter estimation using ETEL

- Now, we use the empirical distribution $\hat{P}_1$ to find the minimizer $D(P_1 \parallel \hat{P}_1)$ on the model space satisfying (12) and (13).
- Thus, we maximize

$$\ell(\boldsymbol{p}) = \sum_{i=1}^{N} \delta_i p_i \log(p_i)$$

subject to

$$\sum_{i=1}^{N} \delta_i p_i = 1, \tag{14}$$

$$\frac{N_1}{N} \sum_{i=1}^{N} \delta_i \boldsymbol{b}(x_i) \left[1 + \frac{N_0}{N_1} \cdot \exp\{\phi_0 + \phi_1' \boldsymbol{b}(x_i)\} \right] p_i = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{b}(x_i), \tag{15}$$

and

$$\sum_{i=1}^{N} \delta_i \exp\{\phi_0 + \phi_1' \boldsymbol{b}(x_i)\} p_i = 1. \tag{16}$$

Calibration equation

- Note that the last two constraints, (15) and (16), do not add any new information.
- Thus, the optimization may use the first constraint (14) only and obtain $\hat{p}_i = 1/N_1$.
- Thus, the estimating equation for model parameters reduces to

$$\sum_{i=1}^{N} \delta_i \underbrace{\left[1 + \frac{N_0}{N_1} \cdot \exp\{\hat{\phi}_0 + \hat{\phi}_1' \boldsymbol{b}(\mathbf{x}_i)\}\right]}_{=\hat{\omega}_i^*} [1, \boldsymbol{b}(x_i)] = \sum_{i=1}^{N} [1, \boldsymbol{b}(x_i)], \quad (17)$$

which is a calibration equation for $[1, \boldsymbol{b}(\mathbf{x})]$.
- We may use $\hat{\omega}_i^*$ in (17) to compute the (smoothed) PS estimator for θ.

Example: $\theta = \mathbb{E}(Y)$

- The smoothed PS estimator of θ is

$$\widehat{\theta}_{SPS} = \frac{1}{N}\sum_{i=1}^{N} \delta_i \hat{\omega}_i^* y_i,$$

where $\hat{\omega}_i^*$ is defined in (17).

- Writing $\hat{\theta}_N = N^{-1}\sum_{i=1}^{N} y_i$, we obtain

$$\widehat{\theta}_{SPS} - \widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} \left(\delta_i \hat{\omega}_i^* - 1\right) y_i = \frac{1}{N}\sum_{i=1}^{N} \left(\delta_i \hat{\omega}_i^* - 1\right)\left\{m(x_i) + e_i\right\}$$

- If $m(x) \in \mathcal{H} = \text{span}\{\mathbf{b}(x)\}$, then, by (17),

$$\widehat{\theta}_{SPS} - \widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} \left(\delta_i \hat{\omega}_i^* - 1\right) e_i,$$

which has zero expectation under MAR.

Remark

- The smoothed PS estimator of θ can be written as

$$\widehat{\theta}_{SPS} = \frac{1}{N}\sum_{i=1}^{N} \delta_i \widehat{\omega}_i^* y_i = \frac{1}{N}\sum_{i=1}^{N} \left[m_i(\boldsymbol{\beta}) + \delta_i \widehat{\omega}_i^* \left\{ y_i - m_i(\boldsymbol{\beta}) \right\} \right] \quad (18)$$

where $m_i(\boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{L} \beta_j b_j(\mathbf{x}_i)$ for any $\beta_0, \beta_1, \cdots, \beta_L$.

- Now, since $\widehat{\omega}_i^* = 1 + (N_0/N_1) \cdot \exp\{\widehat{\lambda}_0 + \widehat{\boldsymbol{\lambda}}_1^T \boldsymbol{b}(\mathbf{x}_i)\}$, the smoothed PS estimator in (18) is algebraically equivalent to

$$\begin{aligned} \widehat{\theta}_{SPS} &= \frac{1}{N}\sum_{i=1}^{N} \left\{ \delta_i y_i + (1-\delta_i) m_i(\boldsymbol{\beta}) \right\} \\ &+ \frac{1}{N} \cdot \frac{N_0}{N_1} \sum_{i=1}^{N} \delta_i \exp\{\widehat{\lambda}_0 + \widehat{\boldsymbol{\lambda}}_1^T \boldsymbol{b}(\mathbf{x}_i)\} \left\{ y_i - m_i(\boldsymbol{\beta}) \right\} \end{aligned}$$

for all $\boldsymbol{\beta}$.

- Thus, the equality also holds for a particular $\hat{\beta}$ that satisfies

$$\sum_{i=1}^{N} \delta_i \exp\{\widehat{\lambda}_0 + \widehat{\boldsymbol{\lambda}}_1^T \boldsymbol{b}(\mathbf{x}_i)\} \left\{ y_i - m_i(\hat{\boldsymbol{\beta}}) \right\} = 0,$$

which leads to

$$\frac{1}{N} \sum_{i=1}^{N} \delta_i \widehat{\omega}_i^* y_i = \frac{1}{N} \sum_{i=1}^{N} \left\{ \delta_i y_i + (1 - \delta_i) m_i(\hat{\boldsymbol{\beta}}) \right\}. \quad (19)$$

- Note that (19) takes the form of the regression imputation estimator under the regression model

$$\mathbb{E}(Y \mid \mathbf{x}) = \beta_0 + \sum_{j=1}^{L} \beta_j b_j(\mathbf{x}).$$

- The final calibration weight $\widehat{\omega}_i^*$ does not directly use the regression model for imputation, but it implements regression imputation indirectly.

Theorem 1 (for $\theta = E(Y)$)

Let

$$\widehat{\theta}_{SPS} = \frac{1}{N} \sum_{i=1}^{N} \delta_i \hat{\omega}_i^* y_i,$$

be the smoothed PS estimator of $\theta = \mathbb{E}(Y)$, where $\hat{\omega}_i^*$ is defined in (17). Under assumption $\mathbb{E}(Y \mid \mathbf{x}) \in \mathcal{H} = \text{span}\{\mathbf{b}(x)\}$ and other regularity conditions, we have

$$\sqrt{N}\left(\widehat{\theta}_{SPS} - \theta\right) \xrightarrow{\mathcal{L}} N(0, V_d),$$

as $N \to \infty$, where

$$V_d = \mathbb{V}\left\{\mathbb{E}(Y \mid \mathbf{X})\right\} + \mathbb{E}\left[\delta\{\omega^*(\mathbf{X})\}^2 \mathbb{V}(Y \mid \mathbf{X})\right], \tag{20}$$

and $\omega^*(\mathbf{x}) = \mathbb{E}_1\{\omega(\mathbf{x}) \mid \mathcal{H}\}$.

Remark 1

1. Because of
$$\mathbb{E}_1\{\omega(\mathbf{x}) \mid \mathcal{H}\} = \{\mathbb{P}(\delta = 1 \mid \mathcal{H})\}^{-1},$$
the asymptotic variance in (20) reduces to
$$V_d = \mathbb{V}\{\mathbb{E}(Y \mid \mathbf{X})\} + \mathbb{E}[\omega^*(\mathbf{X})\mathbb{V}(Y \mid \mathbf{X})],$$
which is the lower bound of the asymptotic variance of the $\sqrt{n}$-consistent estimator of θ (Robins et al., 1994).
2. If we can find $\mathcal{H}_0 \subset \mathcal{H}$ such that $\mathbb{E}(Y \mid \mathbf{x}) \in \mathcal{H}_0$. In this case, we can make V_d in (20) smaller and obtain a more efficient PS estimator using the basis functions in $\mathcal{H}_0$ only. Therefore, increasing the dimension of $\mathcal{H}$ may lose efficiency: penalization technique can be used.

Remark 2

- The proposed PS weighting method can be described as a calibration weighting problem: Minimize

$$Q_1(\omega) = \sum_{i \in S} (\omega_i - 1) \log (\omega_i - 1)$$

subject to

$$\sum_{i \in S} \omega_i [1, \boldsymbol{b}(\mathbf{x}_i)] = \sum_{i=1}^{N} [1, \boldsymbol{b}(\mathbf{x}_i)],$$

- On the other hand, Hainmueller (2012) used

$$Q_2(\omega) = \sum_{i \in S} \omega_i \log (\omega_i)$$

subject to the same calibration constraint. This method is called the entropy balancing method.

Back to the motivating example

- The outcome model is

$$Y = X\beta + Z\gamma + e$$

  and $\gamma = 0$.
- Response model
$$\pi(X, Z) = \mathbb{P}(\delta = 1 \mid X, Z)$$
- The conditional expectation of Y given (X, Z) does not depend on Z, the smoothed PS weight should be a function of X only.
- Thus, it is better not to use Z in constructing the PS weights.

Application: Multivariate Missingness

- The proposed method can be extended to multivarite missing data.
- The missingness pattern can be non-monotone.

Table: Missing Pattern Example

	y_1	y_2	y_3
S_1	✓	✓	✓
S_2	✓		✓
S_3	✓	✓	
S_4	✓		

Model

- Parameter of interest is defined through

$$\mathbb{E}\{U(\theta; \mathbf{y})\} = 0.$$

- We wish to construct an estimating function using all available information:

$$\begin{aligned}\bar{U}(\theta) &= \sum_{i \in S_1} U(\theta; \mathbf{y}_i) + \sum_{i \in S_2} \mathbb{E}\{U(\theta; \mathbf{y}_i) \mid y_{1i}, y_{3i}\} \\ &\quad + \sum_{i \in S_3} \mathbb{E}\{U(\theta; \mathbf{y}_i) \mid y_{1i}, y_{2i}\} + \sum_{i \in S_4} \mathbb{E}\{U(\theta; \mathbf{y}_i) \mid y_{1i}\} \\ &:= \sum_{t=1}^{4} \sum_{i \in S_t} \mathbb{E}\{U(\theta; \mathbf{y}_i) \mid \mathbf{y}_{i,obs(t)}\}\end{aligned}$$

where $\mathbf{y}_{i,obs(t)}$ is the observed part of $\mathbf{y}_i$ for $i \in S_t$.

- Instead of using a model for each conditional distribution, we can use the density ratio model such that

$$N_1^{-1} \sum_{i \in S_1} r_t^*(\mathbf{y}_{i,obs(t)}) U(\theta; \mathbf{y}_i) = N_t^{-1} \sum_{i \in S_t} \mathbb{E}\{U(\theta; \mathbf{y}_i) \mid \mathbf{y}_{i,obs(t)}\} \quad (21)$$

for $t = 2, 3, 4$.

- To construct the density ratio function satisfying (21), we first find $\mathcal{H}_t = \text{span}\{b_1^{(t)}(\mathbf{y}_{obs(t)}), \cdots, b_{L(t)}^{(t)}(\mathbf{y}_{obs(t)})\}$ such that $\mathbb{E}\{U(\theta; \mathbf{y}_i) \mid \mathbf{y}_{i,obs(t)}\} \in \mathcal{H}_t$.
- Thus, using the I-projection idea, we may assume

$$\log\{r_t^*(\mathbf{y}_{obs(t)}; \phi^{(t)})\} = \phi_0^{(t)} + \sum_{j=1}^{L(t)} \phi_j^{(t)} b_j^{(t)}(\mathbf{y}_{obs(t)}). \quad (22)$$

Estimation Method

- The model parameters can be estimated by calibration equation derived from (21) and model assumption (22):

$$N_1^{-1} \sum_{i \in S_1} r_t^*(\mathbf{y}_{i,obs(t)}; \phi^{(t)})(1, \mathbf{b}_i^{(t)}) = N_t^{-1} \sum_{i \in S_t} (1, \mathbf{b}_i^{(t)})$$

  with respect to $\phi^{(t)}$ for $t = 2, 3, 4$, where $\mathbf{b}_i^{(t)}$ is a vector of $b_j^{(t)}(\mathbf{y}_{i,obs(t)})$ for $j = 1, \cdots, L(t)$.

- Once the model parameters are estimated, we can use

$$\hat{\omega}_i^* = \sum_{t=1}^{4} \frac{N_t}{N_1} r^*(\mathbf{y}_{i,obs(t)}; \hat{\phi}^{(t)})$$

  as the final weights for PS estimation.

Simulation 1: MAR

- A 2×2 factorial structure with two factors: outcome regression (OR); response mechanism (RM). We generate δ and $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$ first based on the RM first. We have
  1. RM1 (Logistic regression model):

$$
\begin{aligned}
x_{ik} &\sim N(2,1), \text{for } k = 1, \ldots, 4, \\
\delta_i &\sim \text{Ber}(p_i), \\
\text{logit}(p_i) &= 1 - x_{i1} + 0.5x_{i2} + 0.5x_{i3} - 0.25x_{i4}.
\end{aligned}
$$

  2. RM2(Gaussian mixture model):

$$
\begin{aligned}
\delta_i &\sim \text{Bern}(0.6) \\
x_{ik} &\sim N(2,1), \text{for } k = 1, \ldots, 3, \\
x_{i4} &\sim \begin{cases} N(3,1), \text{if } \delta_i = 1 \\ N(1,1), \text{otherwise.} \end{cases}
\end{aligned}
$$

Simulation 1

- Generate y from
  1. OR1: $y_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + e_i$.
  2. OR2: $y_i = 1 + 0.5x_{i1}x_{i2} + 0.5x_{i3}^2x_{i4}^2 + e_i$.

  where $e_i \sim N(0, 1)$.
- The parameter of interest is $\theta = \mathbb{E}(Y)$.
- Sample size $n = 5,000$ (with 5,000 simulation sample).

Simulation 1

Methods considered for computing the PS weights

1. The proposed information projection (IP) method using calibration variable $(1, x_1, x_2, x_3, x_4)$.
2. Entropy balancing propensity score (EBPS) method of Hainmueller (2012) using calibration variable $(1, x_1, x_2, x_3, x_4)$.
3. Covariate balancing propensity score method (CBPS) of Imai and Ratkovic (2014) using calibration variable $(1, x_1, x_2, x_3, x_4)$.
4. Maximum likelihood estimator (MLE) with Bernoulli distribution with parameter $\text{logit}(p_i) = \mathbf{x}_i^T \phi$.

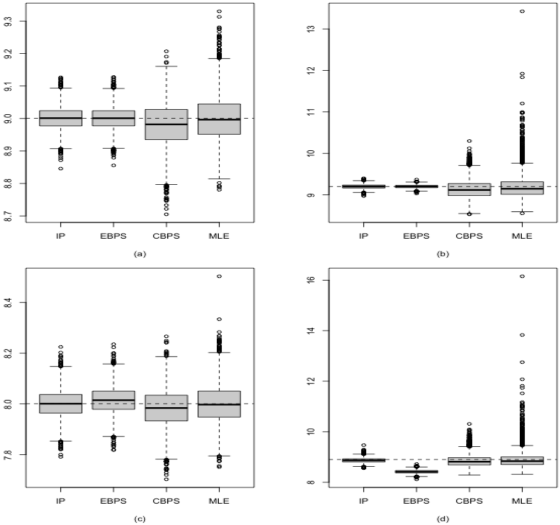


Figure: Boxplots with four estimators for four models under simulation study one: (a) for OR1RM1, (b) OR1RM2, (c) for OR2RM1 and (d) for OR2RM2.

Take-Home message

- Density ratio estimation is a key component for propensity score weighting:

$$\omega^*(\mathbf{x}) = 1 + c \cdot r^*(\mathbf{x})$$

  where $c = N_0/N_1$.
- Proposal
  1. Identify the linear function space $\mathcal{H}$ such that $E(U \mid \mathbf{x}) \in \mathcal{H}$.
  2. The I-projection justifies a parametric log-linear DR model

$$\log\{r^*(\mathbf{x})\} \in \mathcal{H}$$

  3. Model parameter can be used by calibration equation which means

$$r^*(\mathbf{x}) \in \mathcal{H}^{\perp},$$

     where $\mathcal{H}^{\perp}$ is the orthogonal complement space of $\mathcal{H}$.
- Increasing the dimension of $\mathcal{H}$ may lose efficiency: penalization technique can be used.

Future Research Topics

- Extension to non-MAR case.
- Instead of using Kullback-Leibler divergence, we may use Hellinger divergence to achieve some robustness (Antoine and Dovonon, 2021; Li et al., 2019).
- Can be applied to handle data integration combining a probability sample with a non-probability sample.
- The idea can be used to develop weight smoothing for probability samples.

Thank You

References I

Antoine, Bertille and Prosper Dovonon (2021), 'Robust estimation with exponentially tilted Hellinger distance', *Journal of Econometrics* **224**, 330–344.

Hainmueller, Jens (2012), 'Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies', *Political Analysis* **20**, 25–46.

Imai, K. and M. Ratkovic (2014), 'Covariate balancing propensity score', *Journal of the Royal Statistical Society: Series B* **76**, 243–263.

Kim, J.K., S. Park and K. Kim (2019), 'A note on propensity score weighting method using paradata in survey sampling', *Survey Methodology* **45**, 451–463.

Li, Lei, Anand Vidyashankar, Guoqing Diao and Ejaz Ahmed (2019), 'Robust inference after random projections via Hellinger distance for location-scale family', *Entropy* **21**, 348.

Robins, James M, Andrea Rotnitzky and Lue Ping Zhao (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association* **89**, 846–866.

References II

Rubin, D. B. (1976), 'Inference and missing data', *Biometrika* **63**, 581–592.

Schennach, S. M. (2007), 'Point estimation with exponentially tilted empirical likelihood', *The Annals of Statistics* **35**, 634–672.