New Propensity Weighting For Data Integration Combining Auxiliary Information From a Probability Sample

Jae-kwang Kim
Iowa State University

September 24, 2020
Workshop on Blended Data

Joint work with Hangfang Wang at ISU
1. Introduction
1 Introduction

- We are now interested in combining a probability sample with a non-probability sample.
- We observe $X$ from the probability sample and observe $(X, Y)$ from the non-probability sample.

Table: Data Structure

<table>
<thead>
<tr>
<th>Data</th>
<th>$X$</th>
<th>$Y$</th>
<th>Representativeness</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>✓</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>B</td>
<td>✓</td>
<td>✓</td>
<td>No</td>
</tr>
</tbody>
</table>

- Sample B can be a big data.
Mass Imputation approach

- Rivers (2007) idea
  1. Use $X$ to create nearest neighbor imputation for each unit $i \in A$.
  2. Compute
     \[ \hat{\theta} = \sum_{i \in A} w_i y_i^* \]
     where $y_i^*$ is the imputed value of $y_i$ in $i \in A$.

- Assumptions
  1. Missing At Random (MAR) assumption:
     \[ f(Y_i \mid X_i, i \in B) = f(Y_i \mid X_i) \]
  2. Positivity:
     \[ P(i \in B \mid X) > 0 \]
     for all $X$ in the support of $X$.
  3. $E(Y \mid X)$ is a smooth function of $X$.

- Not really applicable for vector $X$. The bias is not asymptotically negligible for vector $X$ (Yang and Kim, 2020).
Mass Imputation approach

- Idea: Use sample B as a train sample to build a regression model for \( E(Y \mid x) \) and obtain \( \hat{y}_i \) for \( i \in A \).
- Mass imputation estimator

\[
\hat{Y}_{MIE} = \sum_{i \in A} w_i \hat{y}_i
\]

- Assumptions: MAR + Positivity + Model
- Parametric regression model (Kim et al, 2019): \( E(Y \mid x) = x' \beta \)
- Nonparametric regression model (Chen et al, 2020): For example, GAM (Generalized Additive Model) assumes

\[
h^{-1}\{ E(Y \mid x) \} = \sum_{k=1}^{p} f_k(x_k),
\]

where \( h(\cdot) \) is the link function (known) and \( f_k(x_k) \) is a smooth function of \( x_k \) (unknown).
Propensity weighting approach 1

- Wish to use \( \{ \hat{P}(I_B = 1 \mid X) \}^{-1} \) as the propensity weight for sample B
- Elliott and Valliant (2017) idea

1. Note that

\[
P(I_B = 1 \mid X) \propto P(I_A = 1 \mid X) \cdot \frac{f(X \mid I_B = 1)}{f(X \mid I_A = 1)}
\]

\[= 1/R(X)\]

2. Estimation of \( \hat{w}(X) = \{ P(I_A = 1 \mid X) \}^{-1} \): Regression of \( w_i \) on \( X_i \) from sample A.

3. Estimation of \( R(X) \): Use

\[
R(X) \equiv \frac{f(X \mid I_A = 1)}{f(X \mid I_B = 1)} \propto \frac{P(I_A = 1 \mid X, I_A + I_B = 1)}{P(I_B = 1 \mid X, I_A + I_B = 1)}
\]

and the apply a suitable classification method from the combined sample to obtain \( \hat{R}(x) \).

4. The final weight for sample B is

\[
\hat{w}_i = \tilde{w}_i \cdot \hat{R}(x_i).
\]
Chen, Li, and Wu (2019) idea

1. Assume a parametric model for $P(I_B = 1 \mid x)$, say $\pi(x; \phi)$.
2. Parameter $\phi$ is estimated by solving

$$\hat{S}(\phi) \equiv \sum_{i \in B} h(x_i; \phi) - \sum_{i \in A} w_i \pi(x_i; \phi)h(x_i; \phi) = 0,$$

where

$$h(x_i; \phi) = \frac{1}{\pi(x_i; \phi)\{1 - \pi(x_i; \phi)\}} \frac{\partial}{\partial \phi} \pi(x_i; \phi)$$

3. Use the inverse of $\hat{\pi}_i = \pi(x_i; \hat{\phi})$ as the propensity weight for sample $B$.

Doubly robust estimator can also be derived using a “working” outcome regression model.
2. Proposed Method
Density ratio function

- Using Bayes formula, we can express

\[
P(I_B = 1 \mid x) = \frac{\pi_1 f(x \mid I_B = 1)}{\pi_1 f(x \mid I_B = 1) + \pi_0 f(x \mid I_B = 0)} = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_0 f_0(x)}
\]

where \(\pi_1 = P(I_B = 1)\) and \(\pi_0 = P(I_B = 0)\).

- Thus, writing

\[
r(x) = \frac{f_0(x)}{f_1(x)},
\]

we can express the propensity weight as

\[
w(x) \equiv \frac{1}{P(I_B = 1 \mid x)} = 1 + \frac{\pi_0}{\pi_1} r(x).
\]

- Since \(\pi_0 / \pi_1\) is known, there is an one-to-one correspondence between the model for \(P(I_B = 1 \mid x)\) and the model for \(r(x)\). The ratio \(r(x)\) is called the density ratio function.
Estimation of Density ratio function

- To estimate density ratio function, we consider the Kullback-Leibler divergence between $f_0$ and $f_1$ is defined by

$$D_{KL}(f_0, f_1) = \int \log \left( \frac{f_0}{f_1} \right) f_0 d\mu.$$  

We are interested in estimating density ratio function $r = f_0 / f_1$ from the sample.

- Note that we can express

$$D_{KL}(f_0, f_1) = \int \log \left( \frac{f_0}{f_1} \right) f_0 d\mu - \int \frac{f_0}{f_1} f_1 d\mu + 1 = \sup_{r>0} \left\{ \int \log (r) f_0 d\mu - \int rf_1 d\mu + 1 \right\}$$

so that the density ratio function can be understood as the maximizer of

$$Q(r) = \int \log (r) f_0 d\mu - \int rf_1 d\mu.$$  

(3)
The finite-population version of $Q(r)$ is

$$Q_p(r) = \frac{1}{N_0} \sum_{i=1}^{N} (1 - I_{B,i}) \log \{r(x_i)\} - \frac{1}{N_1} \sum_{i=1}^{N} I_{B,i} r(x_i),$$

where $N_1 = \sum_{i=1}^{N} I_{B,i}$ and $N_0 = N - N_1$.

The sample-version estimator of $Q(r)$ is

$$\hat{Q}(r) = \frac{1}{N_0} \sum_{i=1}^{N} (w_i I_{A,i} - I_{B,i}) \log \{r(x_i)\} - \frac{1}{N_1} \sum_{i=1}^{N} I_{B,i} r(x_i).$$

Writing $h(x) = \log \{r(x)\}$, the objective function is equivalent to

$$\hat{Q}(h) = \frac{1}{N_0} \sum_{i=1}^{N} (w_i I_{A,i} - I_{B,i}) h(x_i) - \frac{1}{N_1} \sum_{i=1}^{N} I_{B,i} \exp \{h(x_i)\}. \quad (4)$$
Thus, we can formulate the problem as a function estimation problem maximizing the sample entropy function $\hat{Q}(h)$ among $h \in \mathcal{H}$, which is the function space that we want to consider.

For the function space, we may use

$$\mathcal{H} = \text{span}\{B_1(x), \ldots, B_L(x)\}$$

with known basis functions $B_k(x)$.

Once $\hat{h}(x)$ is obtained, we can use

$$\hat{w}(x) = 1 + \frac{N_0}{N_1} \exp\{\hat{h}(x)\}$$

as the propensity weight for sample B.

The choice of the basis function may use the “working” outcome model. We will cover this shortly.
Log-linear DR model

\[ h(x) = \phi_0 + \sum_{k=1}^{L} \phi_k B_k(x) \]

The sample entropy function is

\[ \hat{Q}(\phi) = \frac{1}{N_0} \left\{ \sum_{i=1}^{N} (I_{A,i} w_i - I_{B,i}) \sum_{k=1}^{L} \phi_k B_k(x_i) \right\} - \frac{1}{N_1} \sum_{i=1}^{N} I_{B,i} \exp \left\{ \sum_{k=1}^{L} \phi_k B_k(x_i) \right\} \]

The estimating equation for \( \phi \) is

\[ \frac{\partial}{\partial \phi} \hat{Q}(\phi) = 0 \] (5)

and \( \phi_0 \) is determined by

\[ \frac{1}{N_1} \sum_{i \in B} \exp \{ \phi_0 + \sum_{k=1}^{L} \phi_k B_k(x) \} = 1. \] (6)
Estimating equation (5) reduces to

$$\frac{1}{N_1} \sum_{i=1}^{N} I_{B,i} \exp \left\{ \sum_{k=1}^{L} \phi_k B_k(x_i) \right\} B_k(x_i) = \frac{1}{N_0} \left\{ \sum_{i=1}^{N} (I_{A,i} w_i - I_{B,i}) B_k(x_i) \right\}$$

for $k = 1, 2, \ldots, L$.

Writing $b_i = (1, B_1(x_i), \ldots, B_L(x_i))'$, the final propensity weight

$$\hat{p}_i^{-1} = 1 + \frac{N_0}{N_1} \exp \left\{ \phi_0 + \sum_{k=1}^{L} \hat{\phi}_k B_k(x_i) \right\}$$

satisfies

$$\sum_{i \in B} \hat{p}_i^{-1} b_i = \sum_{i \in A} w_i b_i.$$

Thus, it satisfies the covariate-balancing property (which is quite close to calibration property).
Remark

- For the choice of the basis functions in $\mathcal{H}$, we may use $Y$-variable information in sample B.
- That is, since we observe $(x_i, y_i)$ in sample B, we can use the standard regression techniques (such as LASSO) to fit a regression model for $Y$ from sample B:

$$y_i = \beta_0 + \sum_{k=1}^{L} \beta_k B_k(x_i) + e_i \quad (7)$$

with $E(e_i) = 0$. Model (7) can be regarded as a “working” outcome model.
- Therefore, the basis functions obtained in (7) can be used to construct the propensity score (PS) weights using the proposed maximum entropy method. The resulting PS estimator is efficient if the working model is good.
3. Simulation Study
Simulation Study: Simulation Setup

- Finite population size $N = 100,000$.
- 10 auxiliary variables $X = (X_1, \ldots, X_p)$ with each $X_k \sim \mathcal{N}(2, 1)$.
- The study variable $y_i$ are constructed from
  1. Model A:
     \[
     y_i = 1 + x_{1i} + x_{2i} + x_{3i} + e_i.
     \]
  2. Model B:
     \[
     y_i = 1 + 0.5 * x_{1i}x_{2i} + 0.5 * x_{1i}x_{3i} + e_i.
     \]
- Here, $e_i \sim \mathcal{N}(0, 1)$.
- Two independent samples are generated from each finite population.
  - Sample A: Simple random sample of size $n_A = 500$.
  - Sample B: Stratified random sample of size $n_B \in \{500, 1000\}$ with two strata based on $x_{1i}$ (Stratum 1 if $x_{1i} \leq 2$). Within each stratum, we select $n_h$ elements by simple random sampling, independent between the two strata, where $n_1 = 0.7n_B$ and $n_2 = 0.3n_B$. We assume that the stratum information is unavailable at the time of data analysis.
1. The simple mean (Simple Mean) from sample $B$: $\hat{\theta}_B = \frac{1}{n_B} \sum_{i \in B} y_i$.


3. CLW: The PS estimator proposed by Chen, Li, and Wu (2019) using
   \[ \pi(x_i; \phi) = \left\{ 1 + \exp \left( -\phi_0 - \sum_{k=1}^{P} \phi_k x_{ik} \right) \right\}^{-1} \] as the working propensity score model.


5. New-2: The new PS estimator using the maximum entropy method using the selected covariates from the working regression model (using SCAD method).
Simulation Two: Result \((p = 10)\)

Table: Monte Carlo bias, Monte Carlo variance, and mean square error (MSE) of the four point estimators, based on 10,000 Monte Carlo samples

<table>
<thead>
<tr>
<th>Case</th>
<th>Estimator</th>
<th>Model I</th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias ((\times 10^3))</td>
<td>Var ((\times 10^3))</td>
</tr>
<tr>
<td>Case 1 ((n_B = 500))</td>
<td>Simple Mean</td>
<td>-0.318</td>
<td>6.898</td>
</tr>
<tr>
<td></td>
<td>EV</td>
<td>0.003</td>
<td>9.385</td>
</tr>
<tr>
<td></td>
<td>CLW</td>
<td>-0.007</td>
<td>11.390</td>
</tr>
<tr>
<td></td>
<td>New-1</td>
<td>0.001</td>
<td>8.338</td>
</tr>
<tr>
<td></td>
<td>New-2</td>
<td>0.001</td>
<td>8.299</td>
</tr>
<tr>
<td>Case 2 ((n_B = 1,000))</td>
<td>Simple Mean</td>
<td>-0.319</td>
<td>3.410</td>
</tr>
<tr>
<td></td>
<td>EV</td>
<td>0.006</td>
<td>7.702</td>
</tr>
<tr>
<td></td>
<td>CLW</td>
<td>-0.007</td>
<td>9.443</td>
</tr>
<tr>
<td></td>
<td>New-1</td>
<td>0.001</td>
<td>7.147</td>
</tr>
<tr>
<td></td>
<td>New-2</td>
<td>0.001</td>
<td>7.128</td>
</tr>
</tbody>
</table>
Simulation Two: Result ($p = 30$)

Table: Monte Carlo bias, Monte Carlo variance, and mean square error (MSE) of the four point estimators, based on 10,000 Monte Carlo samples

<table>
<thead>
<tr>
<th>Case</th>
<th>Estimator</th>
<th>Model I</th>
<th></th>
<th>Model II</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Bias</td>
<td>Var ($\times 10^3$)</td>
<td>MSE ($\times 10^3$)</td>
</tr>
<tr>
<td>Case 1</td>
<td>Simple Mean</td>
<td>-0.319</td>
<td>6.716</td>
<td>108.543</td>
</tr>
<tr>
<td></td>
<td>EV</td>
<td>0.002</td>
<td>9.972</td>
<td>9.976</td>
</tr>
<tr>
<td></td>
<td>CLW</td>
<td>0.011</td>
<td>14.664</td>
<td>14.790</td>
</tr>
<tr>
<td></td>
<td>New-1</td>
<td>0.001</td>
<td>8.477</td>
<td>8.477</td>
</tr>
<tr>
<td></td>
<td>New-2</td>
<td>0.001</td>
<td>8.232</td>
<td>8.233</td>
</tr>
<tr>
<td>Case 2</td>
<td>Simple Mean</td>
<td>-0.319</td>
<td>3.388</td>
<td>105.274</td>
</tr>
<tr>
<td></td>
<td>EV</td>
<td>0.006</td>
<td>7.811</td>
<td>7.849</td>
</tr>
<tr>
<td></td>
<td>CLW</td>
<td>0.011</td>
<td>11.424</td>
<td>11.543</td>
</tr>
<tr>
<td></td>
<td>New-1</td>
<td>0.001</td>
<td>7.153</td>
<td>7.154</td>
</tr>
<tr>
<td></td>
<td>New-2</td>
<td>0.001</td>
<td>7.070</td>
<td>7.071</td>
</tr>
</tbody>
</table>
4. Conclusion
Conclusion

- Density ratio function is the key component in computing the propensity score function.
- To estimate the density ratio function, we developed a sample entropy function for optimization. The function space may be constructed from a model for $Y$.
- We may impose additional restriction on the density ratio estimation. For example, we may impose boundedness ($r(x) \leq M$) to avoid extreme propensity weights.
- Extension to Non-ignorable sampling mechanism (Non-MAR) is under investigation.


