Data Integration Using Covariate Summaries from External Sources

Facheng Yu
Department of Statistics, University of Washington

Joint work with Yuqian Zhang
Institute of Statistics and Big Data, Renmin University of China

December 2, 2024

Outline

- A brief overview of data integration
- The use of summary statistics
- Our contribution
  - Efficient estimation using external covariate summaries
  - Efficient estimation for large external datasets
- Methodology and asymptotic theory
  - Mean estimation under data homogeneity
  - Correcting selection bias under data heterogeneity

A brief overview of data integration

Example 1. Semi-supervised learning

- A labeled dataset (primary data) and an unlabeled dataset (external data).
- Applications in linear regression, mean estimation, etc.
- Data is homogeneous (primary and external sources share the same marginal distribution).

Example 2. Causal inference

- Integration of randomized controlled trials and observational studies.
- Data could be heterogeneous (primary and external sources could preserve different marginal distributions).

Question: Popular integration methods usually require individualized covariates from external sources, which might be unavailable due to concerns over accessibility, privacy, storage, or cost.

The use of summary statistics

The use of summary statistics has been explored in various causal inference and data integration studies. For instance,

- Mendelian randomization in genome-wide association studies (GWAS).
- Data fusion in meta-analysis.
- Require access to outcome-related information from external sources, such as $\widehat{\beta}_{\text{least-squared}}$.

Question: How to conduct efficient estimation in scenarios where

- only summary statistics of the external covariates are available,
- and the outcome of external sources has not yet been generated?

Our contribution

Throughout the paper, we focus on mean estimation problems. Let (Γ, X, Y) be an independent copy of the complete data, where

- $\Gamma \in \{0, 1\}$: the group indicator, 1 for primary source and 0 for external source,
- $X \in \mathbb{R}^d$: the covariate vector,
- $Y \in \mathbb{R}$: the outcome.

The goal is to conduct the point estimates for

$$\theta_g = \mathbb{E}[Y] \text{ (Generalizability)} \quad \text{and} \quad \theta_t = \mathbb{E}[Y \mid \Gamma = 0] \text{ (Transportability)}.$$

Note.

- For the entire dataset with sample size n, the primary data can be written as $\mathcal{D}_n := (\Gamma_i, \Gamma_i X_i, \Gamma_i Y_i)_{i=1}^n$, where $(\Gamma_i, X_i, Y_i)_{i=1}^n$ contains i.i.d. samples.
- Since the outcome of external sources has not yet been generated, estimation on θ_t is meaningful.

Our contribution

With the sample mean $\bar{X}_0$ and sample gram matrix $\bar{\Xi}_0$ from the external source:

$$\bar{X}_0 = \frac{\sum_{i=1}^{n}(1-\Gamma_i)X_i}{\sum_{i=1}^{n}(1-\Gamma_i)}, \quad \bar{\Xi}_0 = \frac{\sum_{i=1}^{n}(1-\Gamma_i)X_iX_i^\top}{\sum_{i=1}^{n}(1-\Gamma_i)},$$

our results demonstrate that

- Knowing $\bar{X}_0$ is sufficient to construct consistent (doubly robust) estimators for $\theta_g = \mathbb{E}[Y]$ and $\theta_t = \mathbb{E}[Y \mid \Gamma = 0]$ under homogeneity and heterogeneity.
- If $\bar{\Xi}_0$ is accessible additionally, asymptotic inference can be conducted.
- Our estimations remain efficient for particularly large external data, i.e., $\gamma_n = \mathbb{P}(\Gamma = 1) \to 0$ as $n \to \infty$.

Note. Comment on $\gamma_n = \mathbb{P}(\Gamma = 1) \to 0$:

- Our measure $\mathbb{P}$ is defined on the space of $(\Gamma_i, X_i, Y_i)_{i=1}^n$.
- As $n \to \infty$, the measure could change while $(\Gamma_i, X_i, Y_i)_{i=1}^n$ remains independent.

Inspiration from linear model under MCAR

We start with the scenario of data homogeneity, which leads to the missing completely at random (MCAR) assumption:

Assumption 1 (MCAR)

$\Gamma \perp (X, Y)$.

Define the population slope $\beta^* = \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}[(Y - X^\top \beta)^2]$.

Observe that under MCAR,

- By KKT conditions, $\theta = \mathbb{E}[Y] = \mathbb{E}[X]^\top \beta^*$.
- Replace $\mathbb{E}[X]$ by the sample mean $\bar{X}_{\text{all}}$ such that

$$\bar{X}_{\text{all}} = n^{-1} \sum_{i=1}^{n} X_i = (1 - n^{-1} \sum_{i=1}^{n} \Gamma_i) \bar{X}_0 + n^{-1} \sum_{i=1}^{n} \Gamma_i X_i.$$

Note. When β^* is estimated via least squares $\widehat{\beta}_{\text{least-squares}}$, this approach aligns with the semi-supervised least squares (SSLS) estimator $\widehat{\theta}_{SSLS}$ introduced by Zhang et al. (2019).

- In low dimensions, $\widehat{\theta}_{SSLS}$ is CAN and at least as efficient as $\bar{Y}_1$,
- require the covariate dimension d satisfying $d = o(\sqrt{n_P})$, where $n_P = \sum_{i=1}^{n} \Gamma_i$ is the primary sample size.

Inspiration from linear model and AIPW under MCAR

Let

$$m(X) = \mathbb{E}[Y \mid X] \text{ and } \gamma_n(X) = \mathbb{P}(\Gamma = 1 \mid X).$$

Let $m^*(\cdot)$ and $\gamma_n^*(\cdot)$ represent the working models for the outcome regression and propensity score.

- Under MCAR, $\gamma_n(X) = \gamma_n = \mathbb{P}(\Gamma = 1)$ and $\theta_g = \theta_t =: \theta$.
- Use the naive estimator $\widehat{\gamma}_n = n^{-1} \sum_{i=1}^n \Gamma_i$, then $\gamma_n^* = \gamma_n$.
- Thus, with $m^*(X) = X^\top \beta^*$, AIPW always gives the consistent estimator for θ:

$$\begin{aligned}\theta &= \mathbb{E}\left[X^\top \beta^* + \frac{\Gamma}{\gamma_n^*}\left(Y - X^\top \beta^*\right)\right] \\ &\approx \bar{X}_{\text{all}}^\top \beta^* + \mathbb{E}_n\left[\frac{\Gamma}{\gamma_n^*}\left(Y - X^\top \beta^*\right)\right].\end{aligned}$$

Note.

- To apply to high-dimensional setting, we can estimate β^* by Lasso.
- To relax the sparsity assumption, we can use cross-fitting.

.

Procedure under MCAR

Step 1: Divide the index set $[n]$ into K disjoint subsets $\mathcal{I}_1, \ldots, \mathcal{I}_K$ with equal sizes such that $n_k := |\mathcal{I}_k| = n/K$ for $k \in [K]$. Let $\widehat{\gamma}_k = n_k^{-1} \sum_{i \in \mathcal{I}_k} \Gamma_i$ and $\mathcal{I}_{-k} = [n] \setminus \mathcal{I}_k$.

Step 2: For each $k \in [K]$, compute the Lasso estimator $\widehat{\beta}^{(-k)}$: with some $\lambda_n \geq 0$,

$$\widehat{\beta}^{(-k)} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \frac{\sum_{i \in \mathcal{I}_{-k}} \Gamma_i \left(Y_i - X_i^\top \beta \right)^2}{\sum_{i \in \mathcal{I}_{-k}} \Gamma_i} + \lambda_n ||\beta||_1 \right\}.$$

Step 3: The mean estimator is proposed as:

$$\widehat{\theta} = n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (\Gamma_i X_i + (1 - \Gamma_i) \bar{X}_0)^\top \widehat{\beta}^{(-k)} + n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \frac{\Gamma_i}{\widehat{\gamma}_k} \left(Y_i - X_i^\top \widehat{\beta}^{(-k)} \right). \quad (0.1)$$

When the sample gram matrix $\bar{\Xi}_0$ is also observable, the corresponding asymptotic variance estimator is defined as:

$$\begin{aligned} \widehat{\sigma}^2 =& n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (1 - \Gamma_i) \widehat{\beta}^{(-k),\top} \bar{\Xi}_0 \widehat{\beta}^{(-k)} \\ &+ n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \left\{ \Gamma_i X_i^\top \widehat{\beta}^{(-k)} + \frac{\Gamma_i}{\widehat{\gamma}_k} \left(Y_i - X_i^\top \widehat{\beta}^{(-k)} \right) \right\}^2 - \widehat{\theta}^2. \end{aligned} \quad (0.2)$$

Asymptotic theory

Assumption 2

Let the following conditions hold with constants $\kappa_l, \sigma, \sigma_\omega, \delta_\omega > 0$*: (a)* X *is a sub-Gaussian random vector with* $\left|\left|X^\top v\right|\right|_{\psi_2} \leq \sigma \left|\left|v\right|\right|_2, \forall v \in \mathbb{R}^d$*. In addition,* $\left|\left|X^\top \beta^*\right|\right|_{\psi_2} \leq \sigma$ *and* $\inf_{v \in \mathbb{R}^d, ||v||_2 = 1} \mathbb{E}[(X^\top v)^2] \geq \kappa_l$*. (b)* $w = Y - X^\top \beta^*$ *is a sub-Gaussian random variable with* $||w||_{\psi_2} \leq \sigma_w$ *and* $\mathbb{E}\left[w^2\right] \geq \delta_w$*.*

The following theorem characterizes asymptotic properties of the mean estimator $\widehat{\theta}$.

Theorem 3

Suppose that Assumptions 1 and 2 hold. Choose $\lambda_n \asymp \sqrt{\log d / (n\gamma_n)}$*. If* $n\gamma_n \gg (\log n)^2 \log d$ *and the sparsity level satisfies* $s = \|\beta^*\|_0 = o(n\gamma_n / \log d)$*, then as* $n, d \to \infty$*,* $\widehat{\theta} - \theta = O_p\left((n\gamma_n)^{-1/2}\right)$*,* $\widehat{\sigma}^2 = \sigma_n^2 \{1 + o_p(1)\}$*, and* $\widehat{\sigma}^{-1}\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 1)$*, where* $\sigma_n^2 = \mathrm{Var}\left(X^\top \beta^* + \frac{\mathsf{r}}{\gamma_n}\left(Y - X^\top \beta^*\right)\right) = \gamma_n^{-1}\mathrm{Var}(Y) + \left(1 - \gamma_n^{-1}\right)\mathrm{Var}\left(X^\top \beta^*\right)$*.*

AIPW under MAR

In the following, we consider the possibility of data heterogeneity between primary and external data and assume the following missing at random (MAR) condition instead

Assumption 4 (MAR)

$\Gamma \perp Y \mid X$ *and* $\gamma_n(X) := \mathbb{P}(\Gamma = 1 \mid X) > 0$ *almost surely.*

We are interested in

$$\theta_g = \mathbb{E}[Y] \text{ (generalizability)} \quad \text{and} \quad \theta_t = \mathbb{E}[Y \mid \Gamma = 0] \text{ (transportability)}.$$

AIPW under MAR

For AIPW, as long as either $m^*(\cdot) = m(\cdot)$ or $\gamma_n^*(\cdot) = \gamma_n(\cdot)$,

$$\theta_g = \mathbb{E}\left[m^*(X) + \frac{\Gamma}{\gamma_n^*(X)}\left(Y - m^*(X)\right)\right],$$
$$\theta_t = \mathbb{E}\left[\frac{1-\Gamma}{1-\gamma_n} m^*(X) + \frac{\Gamma\left(1-\gamma_n^*(X)\right)}{\gamma_n^*(X)\left(1-\gamma_n\right)}\left(Y - m^*(X)\right)\right].$$

Consider a linear outcome model $m^*(X) = X^\top \beta_{OR}^*$. The above representations can also be expressed as:

$$\theta_g = \left(1-\gamma_n\right)\mathbb{E}\left[X \mid \Gamma = 0\right]^\top \beta_{OR}^* + \mathbb{E}\left[\Gamma X^\top \beta_{OR}^* + \frac{\Gamma}{\gamma_n^*(X)}\left(Y - X^\top \beta_{OR}^*\right)\right],$$
$$\theta_t = \mathbb{E}\left[X \mid \Gamma = 0\right]^\top \beta_{OR}^* + \mathbb{E}\left[\frac{\Gamma\left(1-\gamma_n^*(X)\right)}{\gamma_n^*(X)\left(1-\gamma_n\right)}\left(Y - X^\top \beta_{OR}^*\right)\right].$$

Note. With linear model, we can still construct consistent estimator for θ_g and θ_t.

Estimation for nuisances under MAR

Consider a cross-fitted estimation. For each fold $\mathcal{I}_k$, and separate $\mathcal{I}_{-k} = [n] \setminus \mathcal{I}_k$ into two disjoint subsets $\mathcal{I}_{-k,\alpha}, \mathcal{I}_{-k,\beta}$ with same sizes M.

> Construct a logistic propensity estimator for $\gamma_n^*(X) = \text{expit}(X^\top \alpha_{PS}^*)$ as

$$\widehat{\alpha}_{PS}^{(-k)} = \underset{\alpha \in \mathbb{R}^d}{\arg\min} \left\{ M^{-1} \sum_{i \in \mathcal{I}_{-k,\alpha}} \left\{ (1 - \Gamma_i) \bar{X}_0^\top \alpha + \Gamma_i \exp\left(-X_i^\top \alpha\right) \right\} + \lambda_\alpha ||\alpha||_1 \right\}.$$

> Construct a linear outcome estimator

$$\widehat{\beta}_{OR}^{(-k)} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ M^{-1} \sum_{i \in \mathcal{I}_{-k,\beta}} \Gamma_i \exp\left(-X_i^\top \widehat{\alpha}_{PS}^{(-k)}\right) \left(Y_i - X_i^\top \beta\right)^2 + \lambda_\beta ||\beta||_1 \right\}.$$

To leveraging conditional independence, we also define a oracle nuisance estimator assuming α_{PS}^* is known.

$$\widetilde{\beta}_{OR}^{(-k)} = \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ M^{-1} \sum_{i \in \mathcal{I}_{-k,\beta}} \Gamma_i \exp\left(-X_i^\top \alpha_{PS}^*\right) \left(Y_i - X_i^\top \beta\right)^2 + \lambda_\beta ||\beta||_1 \right\}.$$

Asymptotic theory

When $\bar{X}_0$ is observed, $\theta_g = \mathbb{E}[Y]$ can be estimated by

$$\widehat{\theta}_g = n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \left\{ \left(\Gamma_i X_i + (1 - \Gamma_i) \bar{X}_0 \right)^\top \widehat{\beta}_{OR}^{(-k)} + \frac{\Gamma_i}{g\left(X_i^\top \widehat{\alpha}_{PS}^{(-k)} \right)} \left(Y_i - X_i^\top \widehat{\beta}_{OR}^{(-k)} \right) \right\}.$$

When $\bar{\bar{\Xi}}_0$ is further observed, we can construct the asymptotic variance estimators as:

$$\begin{aligned} \widehat{\sigma}_g^2 = {} & n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} (1 - \Gamma_i) \widehat{\beta}_{OR}^{(-k),\top} \bar{\bar{\Xi}}_0 \widehat{\beta}_{OR}^{(-k)} - \widehat{\theta}_g^2 \\ & + n^{-1} \sum_{k=1}^{K} \sum_{i \in \mathcal{I}_k} \left\{ \Gamma_i X_i^\top \widehat{\beta}_{OR}^{(-k)} + \frac{\Gamma_i}{g\left(X_i^\top \widehat{\alpha}_{PS}^{(-k)} \right)} \left(Y_i - X_i^\top \widehat{\beta}_{OR}^{(-k)} \right) \right\}^2 \end{aligned}$$

Assumptions

Let $g(x) = \text{expit(x)}$. For propensity score:

- $k_0(1-\gamma_n)/\gamma_n \leq (1-g(X^\top \alpha_{PS}^*))/g(X^\top \alpha_{PS}^*) \leq k_0^{-1}(1-\gamma_n)/\gamma_n$ almost surely.
- $\mathbb{P}(\Gamma = 0) \geq c_0$.
- $\mathbb{E}[\gamma_n^q(X)] \leq \nu \gamma_n^q$ for some $q > 1$ and $\nu > 0$.

In addition,

- For each $j \in \{0,1\}$, and conditional on $\Gamma = j$, X is a sub-Gaussian random vector and $X^\top \beta_{OR}^*$ be a sub-Gaussian random variable, both with parameter σ.
- $\inf_{v \in \mathbb{R}^d, ||v||_2 = 1} \mathbb{E}\left[\left(X^\top v\right)^2 \mid \Gamma = 1\right] \geq \kappa_l$.
- The residual $w_{OR} = Y - X^\top \beta_{OR}^*$ is a sub-Gaussian random variable with parameter σ_w, $\mathbb{E}\left[w_{OR}^8 \mid \Gamma = 1\right] \leq \sigma_w^8$, and $\mathbb{E}\left[w_{OR}^2 \mid \Gamma = 1\right] \geq \delta_w$.

Asymptotic theory

Theorem 5

Under MAR and regular assumptions. Choose $\lambda_\alpha \asymp \lambda_\beta \asymp \sqrt{\log d/(n\gamma_n)}$. *Let* $n\gamma_n \gg (\log n)^2 \log d$, $\|\alpha^*_{PS}\|_0\|\beta^*_{OR}\|_0 = o(n\gamma_n/(\log n(\log d)^2))$, *and either of the following conditions hold: (1) (Correct OR model)* $\mu(X) = X^\top \beta^*_{OR}$ *or (2) (Correct PS model)* $\gamma_n(X) = g\left(X^\top \alpha^*_{PS}\right)$ *and* $\|\alpha^*_{PS}\|_0 = o(\sqrt{n\gamma_n}/\log d)$. *Then as* $n, d \to \infty$,

(a) $\widehat{\theta}_g - \theta_g = O_p\left((n\gamma_n)^{-1/2}\right)$, $\widehat{\sigma}^2_g = \sigma^2_g\{1 + o_p(1)\}$, *and* $\widehat{\sigma}^{-1}\sqrt{n}(\widehat{\theta}_g - \theta_g) \xrightarrow{d} \mathcal{N}(0,1)$,

where $\sigma^2_g = \mathrm{Var}\left(X^\top \beta^*_{OR} + \frac{\Gamma}{g\left(X^\top \alpha^*_{PS}\right)}\left(Y - X^\top \beta^*_{OR}\right)\right)$.

(b) $\widehat{\theta}_t - \theta_t = O_p\left((n\gamma_n)^{-1/2}\right)$, $\widehat{\sigma}^2_t = \sigma^2_t\{1 + o_p(1)\}$, *and* $\widehat{\sigma}_t^{-1}\sqrt{n}(\widehat{\theta}_t - \theta_t) \xrightarrow{d} \mathcal{N}(0,1)$,

where $\sigma^2_t = \mathbb{E}\left[\left\{\frac{1-\Gamma}{1-\gamma_n}\left(X^\top \beta^*_{OR} - \theta_t\right) + \frac{\Gamma\left(1-g\left(X^\top \alpha^*_{PS}\right)\right)}{(1-\gamma_n)g\left(X^\top \alpha^*_{PS}\right)}\left(Y - X^\top \beta^*_{OR}\right)\right\}^2\right]$.

Extensions

We also study some extensions of our method in

- Estimate ATE in causal inference
- Estimate ATE on the external population in causal inference.

References I

Zhang, A., Brown, L. D., and Cai, T. T. (2019). Semi-supervised inference: General theory and estimation of means. The Annals of Statistics, 47(5):2538–2566.