Doubly robust inference when combining probability and non-probability samples with high dimensional data

Ma Ting Tin

Department of Statistics, Chinese University of Hong Kong

2021 Summer Lab Meeting

August 17, 2021

Introduction

Introduction and motivation

Definition 1.1 (Probability and Non-probability sampling)

In general, we want to understand the DGP or some statistical quantities based on the available sample. There are two type of sampling

- **[Probability sampling]** The probability samples are selected under known sampling designs and thus are representative of target distribution.
- **[Non-probability sampling]** Involves non-random selection based on convenience or other criteria, allowing you to easily collect data.
  Examples: remote sensing data and web-based volunteer samples

1. Non-probability sample provide rich information about the DGP and might help finite population inference.
2. **AIM:** Propose data integration methods that take advantages of both probability and non-probability sampling.

Existing data integrating methods

Definition 1.2 (Auxiliary variable)

Auxiliary variable is a variable of the Sampling frame which is neither the contact variable, nor the identifier.
💡: An Auxiliary variable is any Variable about which information is available prior to data collection and this information is known for all units of the population.

There are three types of existing methods for data integration.

1. **Propensity score adjustment (Rosenbaum and Rubin, 1983)**
   Model the probability of aunit being selected into non-probability sample followed by estimation and selection biases adjustment (EG: by propensity score weighting or Stratification).
   👎: Highly biased and variable under model mis-specification.
2. **Calibration weighting (Deville and Sarndal, 1992)**
   IDEA: Forces the moments or the empirical distribution of auxiliary variable between probability and non-probability sample to be the same
3. **Mass imputation:** Regard non-probability sample as training data set, and apply imputation to all units in probability sample.

Basic Set-up

Notation: Two Samples

Consider a population with finite size N (N is known)

- Index set of N units $\mathcal{U} \coloneqq \{1, \cdots, N\}$
- Population consist of $\mathscr{F}_N \coloneqq \{(X_i, Y_i) : i \in \mathcal{U}\}$
- **Parameter of interest:** $\mu \coloneqq N^{-1} \sum_{i=1}^{N} Y_i$.

Denote the two group of sample as follow:

1. Sample A: probability sample with size n_A.
2. Sample B: non-probability sample with size n_B.

There are some more notation:

- $\pi_{A,i} \coloneqq \mathrm{P}(i \in A)$, i.e. probability of the i-th individual being included in the probability sample.
- $\mathbb{1}_{\#,i} \coloneqq \mathbb{1}(i\text{-th individual is included in sample } \#)$.

Data Structure

Sample	Size	Sampling Weigting π^{-1}	Covariate X	Study Variable Y
A	n_A	✓	✓	✗
B	n_B	✗	✓	✓

To summarize, our dataset is as follow

- Sample A: $\mathcal{O}_A = \{(d_{A,i}, X_i) : i \in A\}$, where $d_{A,i} := \pi_{A,i}^{-1}$.
- Sample B: $\mathcal{O}_B = \{(X_i, Y_i) : i \in B\}$.
- $\mathbb{1}_{A,i}$ and $\mathbb{1}_{B,i}$ for $i = 1, \cdots, N$

Remark 0.1

As sampling mechanism for sample B in unknown, Horvitz-Thompson estimation cannot be applied.

Identification Assumption

The following identification assumption is made:

1. $\pi_B(X) \coloneqq \mathrm{P}(\mathbb{1}_B = 1|X, Y) = \mathrm{P}(\mathbb{1}_B = 1|X)$.
2. $\exists \delta_B > 0$ and $\gamma \in (2/3, 1]$ s.t. $\forall X, \pi_B(X) > N^{\gamma-1}\delta_B > 0$.

We then interpret the meaning behind the assumption

1. (1) implies that the selection indicator $\mathbb{1}_B$ and Y are $m(X) \coloneqq \mathrm{E}(Y|X) = \mathrm{E}(Y|X, \mathbb{1}_B = 1)$ can be estimated solely on basis of sample B.
2. (2) implies $n_B^{-1} = O(N^{-\gamma})$.

Remark 0.2

Notice that assumption (1) is NOT verifiable from the observed data. One should consider more possible predictors for the selection indicator $\mathbb{1}_B$ or outcome Y, resulting in a rich set of variables in X.

Further notation

Definition 2.1 (Sampling score)

Sampling (or propensity) score is probability of a unit being selected into the non-probability sample B. Let $\pi_B(X)$ be the sampling score function given X.

Let $\pi_B(X^T\alpha)$ and $m(X^T\beta)$ be the postulated models for $\pi_B(X)$ and $m(X)$ respectively, where α, β are unknown parameters.

- $\pi_B(X)$ and $m(X)$ are unknown in practice and need to be estimated.
- Various estimator for μ have been proposed, each requiring differen model assumption and estimation strategies.

Existing estimators

Definition 2.2 (Inverse Probability Weighting Estimator (IPW))

Given estimator of α, $\widehat{\alpha}$, IPW is defined as

$$\widehat{\mu}_{IPW} = \widehat{\mu}_{IPW}(\widehat{\alpha}) = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{1}_{B,i}}{\pi_B(X_i^T\widehat{\alpha})} Y_i.$$

Remark 0.3

Justification for $\widehat{\mu}$ requires $\pi_B(X) = \pi_B(X_i^T\alpha)$ and $\widehat{\alpha} \xrightarrow{pr} \alpha$.

The estimator $\widehat{\alpha}$ can be obtained through following estimating equation:

$$\sum_{i=1}^{N} \left\{ \frac{\mathbb{1}_{B,i}}{\pi(X_i^T\alpha)} - \frac{\mathbb{1}_{A,i}}{\pi_{A,i}} \right\} h(X_i;\alpha) = 0,$$

where possible choices of $h(X_i;\alpha)$ could be $h(X;\alpha) = X$ or $h(X;\alpha) = \pi(X^T;\alpha)X$.

Existing estimators

Definition 2.3 (Outcome regression based on sample A)

The outcome regression estimator is given by

$$\widehat{\mu}_{reg} = \widehat{\mu}_{reg}(\widehat{\beta}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{A,i} d_{A,i} m(X_i^T \widehat{\beta}).,$$

where $\widehat{\beta}$ is obtained by fitting the outcome model based solely on $\mathcal{O}_B$.

Remark 0.4

- *Justification for $\widehat{\mu}_{reg}$ requires $m(X) = m(X_i^T \beta)$ and $\widehat{\alpha} \xrightarrow{pr} \alpha$. If one of them are not satisfied, the estimator could be biased.*
- *Similar form with IPW, but in IPW, the sampling score has to be estimated and the response value is known, which is exactly opposite with this estimator.*

Existing estimators

Definition 2.4 (Calibration weigting)

The calibration weigting estimator is given by

$$\widehat{\mu}_{cal} = \frac{1}{N}\sum_{i=1}^{N} \omega_i \mathbb{1}_{B,i} Y_i, ,$$

where $\{\omega_i : i \in \mathcal{U}\}$ satisfies either

- $\sum_{i\in A} d_{A,i} X_i = \sum_{i\in B} \omega_i X_i$.
- $\sum_{i\in A} d_{A,i} m(X_i;\widehat{\beta}) = \sum_{i\in B} \omega_i m(X_i;\widehat{\beta})$.

Remark 0.5

Justification for $\widehat{\mu}_{cal}$ by constraint (1) requires linearity of

- *Outcome model, i.e. $m(X) = X^t\beta^*$ for some β^*; or*
- *Inverse probability of sampling weight, i.e. $\pi_B(X)^{-1} = X^T\alpha^*$ for some α^**

while it requires well-specification of $m(X;\beta)$ for constraint (2).

Existing estimators

Definition 2.5 (Doubly Robust Estimator)

The doubly robust estimator is given by

$$\widehat{\mu}_{dr} = \widehat{\mu}_{dr}(\widehat{\alpha}, \widehat{\beta}) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{\mathbb{1}_{B,i}}{\pi_B(X_i^T\widehat{\alpha})}\left\{Y_i - m(X_i^T\widehat{\beta})\right\} + \mathbb{1}_{A,i}d_{A,i}m(X_i^T\widehat{\beta})\right],$$

Remark 0.6

The estimator above is doubly robust in the following sense: $\widehat{\mu}_{dr} \xrightarrow{pr} \mu$ *if* **either**

- $\pi_B(X^T\alpha)$ *is correctly specified; or*
- $m(X^T\beta)$ *is correctly specified.*

Note that the consistency hold without requiring both of them to be well-specified.

Methodology in high dimensional data

Motivation and Notation

Including unnecessary covariates in model increases the bias and variance. Variable selection is important to handle high dimensional covariates.

Let $\alpha \in \mathbb{R}^p$ and $\mathcal{J} \subseteq \{1, \cdots, p\}$ Then we define the following

- $||\alpha||_0 := \sum_{j=1}^p \mathbb{1}(\alpha_j \neq 0)$
- L_1-norm: $||\alpha||_1 := \sum_{j=1}^p |\alpha_j|$
- L_2-norm: $||\alpha||_2 := \sum_{j=1}^p \sqrt{\alpha_j^2}$
- L_∞-norm: $||\alpha||_\infty := \max_{1 \leq j \leq p} |\alpha_j|$

Define $\alpha_{\mathcal{J}}$ as subvector of α with indices are in $\mathcal{J}$ and $\mathcal{J}^c = \{1, \cdots, p\} \backslash \mathcal{J}$.
Let $\Sigma \in \mathbb{R}^{p \times p}$ and $\Sigma_{\mathcal{J}_1, \mathcal{J}_2}$ be submatrix of Σ formed by rows in $\mathcal{J}_1$ and columns in $\mathcal{J}_2$.

Remark 0.7

Covariates are standardized in prior to the inference so that they have variances approximately equal 1, which stabilize the variable selection procedure.

Model assumption and choice of parameters

Assumption 2 (sampling score model):
Assume that $\pi_B(X)$ follows logistic regression model, i.e. $\text{logit}\{\pi_B(X^T\alpha)\} = X^T\alpha$.

The true parameter α, α^* is the parameter that minimize Kullback-Leibler distance between $\pi_B(X)$ and $\pi_B(X^T\alpha)$, i.e.

$$\alpha^* := \underset{\alpha\in\mathbb{R}^p}{\arg\min}\, \mathrm{E}\left[\pi_B(X)\log\left\{\frac{\pi_B(X)}{\pi_B(X^T\alpha)}\right\} + \{1-\pi_B(X)\}\log\left\{\frac{1-\pi_B(X)}{1-\pi_B(X^T\alpha)}\right\}\right].$$

Remark 0.8

One can extend the proposed framework to other models such as probit model.

Assumption 3 (outcome model): Assume that $m(X)$ follows a generalized linear regression model, i.e. $m(X) = m(X^T\beta)$ and the true parameter β^* is defined as

$$\beta^* := \underset{\beta\in\mathbb{R}^p}{\arg\min}\, \mathrm{E}\left[\{Y - m(X^T\beta)\}^2\right].$$

Step 1: Variable Selection

The following are estimating functions for α and β

$$U_1(\alpha) = \sum_{i=1}^{N} \left\{ \frac{\mathbb{1}_{B,i}}{\pi(X_i^T\alpha)} - \frac{\mathbb{1}_{A,i}}{\pi_{A,i}} \right\} X_i \quad \text{and} \quad U_2(\beta) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{B,i} \{ Y_i - m(X_i^T\beta) \} X_i,$$

denote $U(\theta) = (U_1(\alpha)^T, U_2(\beta^T))^T$ as the joint estimating function for $\theta := (\alpha^T, \beta^T)^T$

For large p, consider penalized estimating function as

$$U^p(\alpha, \beta) = U(\alpha, \beta) - \begin{pmatrix} q_{\lambda_\alpha}(|\alpha|)\text{sgn}(\alpha) \\ q_{\lambda_\beta}(|\beta|)\text{sgn}(\beta) \end{pmatrix},$$

where

$$q_\lambda(|\theta|) := \lambda \left\{ \mathbb{1}(|\theta| < \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} \mathbb{1}(|\theta| \geq \lambda) \right\}$$

Step 1: Variable Selection

In order to have better understanding on the penalty function, we rewrite

$$q_\lambda(|\theta|) = \begin{cases} \lambda & , \text{ if } |\theta| < \lambda \\ \frac{a\lambda - |\theta|}{(a-1)\lambda} & , \text{ if } \lambda \leq |\theta| < a\lambda \\ 0 & , \text{ if } |\theta| \geq a\lambda \end{cases}$$

Therefore, it is not hard to see that

- λ represent the penalty size for small $|\theta|$ and it defines the term "small".
- a defines the term "large" so that no penalty given to $|\theta| \geq a\lambda$.

The penalty of this model is different from lasso/ridge regression. They add penalty s.t. large value of coefficient is not favourable, while it is the opposite in our case.

- For large $|\alpha_j|$, $q_{\lambda_\alpha}(|\alpha_j|) = 0$ and no penalty added.
- For small $|\alpha_j|$, the penalty forces $\widehat{\alpha}_j$ to be zero and excludes j-th element of X from final selected set of variables.

The similar discussion applies to that of $U_2(\beta)$ and q_{λ_β}.

Step 2: Doubly Robust Estimation

After obtaining the coefficient estimate in Step 1, denote $\mathscr{C}$ as the index set s.t. those **variable are associated with non-zero coefficient estimate**.

AIM: Re-estimate parameters $\widehat{\alpha}, \widehat{\beta}$ in the doubly robust estimator $\widehat{\mu}_{dr} = \widehat{\mu}_{dr}(\widehat{\alpha}, \widehat{\beta})$ by using only $X_{\mathscr{C}}$ to minimize the asymptotic squared bias of $\widehat{\mu}_{dr}$, denoted as $a.bias(\alpha, \beta)$.

The $a.bias(\alpha^*, \beta^*)$ is in the following form

$$\mathrm{E}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\mathbb{1}_{B,i}}{\pi_B(X_i^T\alpha^*)}-1\right\}\left\{Y_i - m(X_i^T\beta^*)\right\}\right] + \mathrm{E}\left\{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}_{A,i}d_{A,i}-1\right)m(X_i^T\beta^*)\right\}$$

So we can compute $\nabla a.bias(\alpha^*, \beta^*)$ and regard its empirical version as the estimating function, i.e.

$$J(\alpha,\beta) := \begin{pmatrix} \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_{B,i}\left\{\frac{1}{\pi_B(X_i^T\alpha)}-1\right\}\left\{Y_i - m(X_i^T\beta)\right\}X_{i\mathscr{C}} \\ \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\mathbb{1}_{B,i}}{\pi_B(X_i^T\alpha)} - d_{A,i}\mathbb{1}_{A,i}\right\}\frac{\partial}{\partial\beta_{\mathscr{C}}}m(X_i^T\beta) \end{pmatrix}$$

constrainted on $\{(\alpha^T, \beta^T)^T \in \mathbb{R}^{2p} : \alpha_{\mathscr{C}^c} = \beta_{\mathscr{C}^c} = 0\}$.

Comments and remarks on 2 step procedure

Remark 0.9

1. *Estimating function in Step (1) seperates selection for α and β, thus it stabilizes the selection procedure if either one of model is misspeicified.*
2. *Estimating function in Step (2) leads to the consistency of $\widehat{\mu}_{dr}$ when either one of the model is well-specified.*

In step (2), we consider union of covariates $X_{\mathscr{C}}$, where $\mathscr{C} = \widehat{\mathscr{M}_\alpha} \cup \widehat{\mathscr{M}_\beta}$. While there are two other common choices in the literature, which may NOT be robust to model misspeicification.

Computation

Computation: Step 1 (MM Algorithm)

Definition 4.1 (Marjorize and Minorize)

A function $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is said to be majorize (resp. minorize) the function $f(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^{(t)}$ if

$$g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = f(\boldsymbol{\theta}^{(t)}) \quad \text{and} \quad g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geq f(\boldsymbol{\theta}) \text{ (resp. } g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \leq f(\boldsymbol{\theta})) \ \ \forall \boldsymbol{\theta}$$

Suppose we want to minimize $f(\boldsymbol{\theta})$. Define $\boldsymbol{\theta}^{(t+1)} = \arg\min_{\boldsymbol{\theta}} g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. Then we have

$$f(\boldsymbol{\theta}^{(t+1)}) \leq g(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \leq g(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = f(\boldsymbol{\theta}^{(t)}).$$

Remark 0.10

- *g is commonly chosen as lower/upper bound of f which is easier to be handled.*
- *EM algorithm is a particular case of MM algorithm by setting*

$$g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \coloneqq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - \sum_{i=1}^{N} \int_{\mathcal{Z}} f(\boldsymbol{z}|\boldsymbol{y}_n; \boldsymbol{\theta}^{(t)}) \ln f(\boldsymbol{z}|\boldsymbol{y}_n; \boldsymbol{\theta}^{(t)}) dz.$$

Computation: Step 1 (MM Algorithm)

AIM: Solving the penalized estimating equation.

Step 1: By Minorization-maximization (MM) Algorithm, the estimator $\widetilde{\theta} = (\widetilde{\alpha}, \widetilde{\beta})$ satisfies

$$U^p(\widetilde{\theta}) = U(\widetilde{\theta}) - \begin{pmatrix} q_{\lambda_{\widetilde{\alpha}}}(|\widetilde{\alpha}|)\text{sgn}(\widetilde{\alpha})\frac{|\widetilde{\alpha}|}{\epsilon+|\widetilde{\alpha}|} \\ q_{\lambda_{\widetilde{\beta}}}(|\widetilde{\beta}|)\text{sgn}(\widetilde{\beta})\frac{|\widetilde{\beta}|}{\epsilon+|\widetilde{\beta}|} \end{pmatrix} = 0$$

Computation: Step 2 (Coordinate Descent Algorithm)

Definition 4.2 (Coordinate Descent Algorithm)

AIM: Solve $F(\boldsymbol{x}) = 0$ for multivariate function F, where $\boldsymbol{x} \in \mathbb{R}^n$.

1. Initialize $\boldsymbol{x}^0 = (x_1^{(0)}, \cdots, x_n^{(0)})$ and step size α (could be non-fixed).
2. There are two possible choice in this step

- Update $x_i^{(k+1)} = x_i^{(k)} - \alpha \frac{\partial F}{\partial x_i^{(k)}}(\boldsymbol{x}^{(k)})$. (Similar to the idea of gradient descent).
- Set $x_i^{(k+1)} = \arg\min_{y \in \mathbb{R}} f(x_1^{(k+1)}, \cdots, x_{i-1}^{(k+1)}, y, x_{i+1}^{(k)}, \cdots, x_n^{(k)})$.

Remark 0.11

- *Coordinate Descent algorithm prevent the evaluation of large matrix inverse in the Newton-Raphson Algorithm.*
- *Within the two possible choice, the first one is only available for continously differentiable function F, while the second one work in general scenario*

Computation: Step 2 (Coordinate Descent Algorithm)

Step 2: Solve $U^p(\widetilde{\theta}) = 0$ by co-ordinate decent algorithm (instead of Newton-Raphson algorithm to lower computational burden): For $k \geq 1$, we write

$$\nabla U(\theta) = \text{diag}\left\{\frac{\partial U_1(\alpha)}{\partial \alpha^T}, \frac{\partial U_2(\beta)}{\partial \beta^T}\right\}$$

where

$$\frac{\partial U_1(\alpha)}{\partial \alpha^T} = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{B,i} \frac{1-\pi_B(X_i^T\alpha)}{\pi_B(X_i^T\alpha)} X_i X_i^T \text{ and } \frac{\partial U_2(\beta)}{\partial \beta^T} = -\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}_{B,i} m^{(1)}(X_i^T\beta)^2 X_i X_i^T.$$

With the other coordinates fixed, the coordinate decent algorithm (k-th Newton-Raphson update) for θ_j is

$$\widetilde{\theta}_j^{[k]} = \widetilde{\theta}_j^{[k-1]} + \frac{U_j(\widetilde{\theta}_j^{[k-1]}) - Nq_{\lambda_j}(|\theta_j^{[k-1]}|)\widetilde{\theta}_j^{[k-1]}}{\nabla_{jj}(\widetilde{\theta}_j^{[k-1]}) + Nq_{\lambda_j}(|\theta_j^{[k-1]}|)}$$

Computation: Step 3 (K-fold Cross Validation)

Step 3: Apply K-fold CV to select the hyperparameter $(\lambda_\alpha, \lambda_\beta)$.

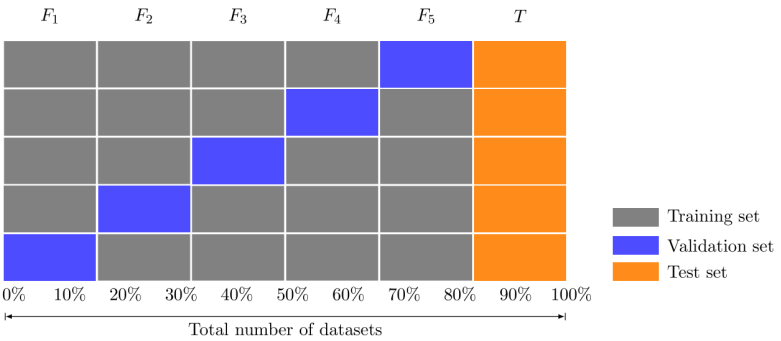

Figure: Idea of K-fold cross validation (Picture extracted from course of STAT4012)

Computation: Step 3 (K-fold Cross Validation)

In this paper, the author suggest to tune λ_α and λ_β one by one. First, as we want to search for the λ_α s.t. the "behaviour" of two sample are similar, so the loss function for λ_α is chosen as

$$L(\lambda_\alpha) := \sum_{j=1}^{p} \left(\sum_{i=1}^{N} \left[\frac{\mathbb{1}_{B,i}}{\pi_B\{X_i^T \widetilde{\alpha}(\lambda_\alpha)\}} - \frac{\mathbb{1}_{A,i}}{\pi_{A,i}} \right] X_{i,j} \right)^2$$

and we use thre prediction error loss function for selection of λ_β, i.e.

$$L(\lambda_\beta) := \sum_{i=1}^{N} \mathbb{1}_{B,i} \left[Y_i - m\left\{ X_i^T \widetilde{\beta}(\lambda_\beta) \right\} \right]^2$$

Asymptotic results for variable selection and estimation

Assumption and regularity conditions

Consider sample A with high entropy sampling designs (Berger, 1998). Let $\mathscr{M}_\alpha = \{1 \leq j \leq p : \alpha_j^* \neq 0\}, \mathscr{M}_\beta = \{1 \leq j \leq p : \beta_j^* \neq 0\}$ and $\mathscr{M}_\theta := \mathscr{M}_\alpha \cup \{p + \mathscr{M}_\beta\}$. Define $s_\alpha = ||\alpha^*||_0, s_\beta = ||\beta^*||_0, s_\theta = s_\alpha + s_\beta$ and $\lambda_\theta = \min(\lambda_\alpha, \lambda_\beta)$. Assume the following:

1. $\exists \delta_A$ and $\gamma \in (2/3, 1]$ s.t. for all $1 \leq i \leq N$, $\pi_{A,i} \geq N^{\gamma-1}\delta_A > 0$.
2. θ belongs to a compact subset in $\mathbb{R}^{2p}$ and θ^* lies in the interior of that set.
3. $\{X_i : i \in \mathscr{U}\}$ are fixed and uniformly bounded.
4. $\exists c_1, c_2$ s.t. $0 < c_1 \leq \lambda_{\min}(\frac{1}{N}\sum_{i=1}^N X_i^T X_i) \leq \lambda_{\max}(\frac{1}{N}\sum_{i=1}^N X_i^T X_i) \leq c_2 < \infty$, where $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ refers to the min/max eigenvalue of a matrix $(\cdot)$ respectively.
5. Let $\epsilon_i(\beta) := Y_i - m(X_i^T\beta)$. $\exists c_3, c_4, c_5$ and $\delta > 0$ s.t. $\mathrm{E}|\epsilon_i(\beta^{2+\delta})| \leq c_3$ and $\mathrm{E}[\exp\{c_4|\epsilon_i(\beta^*)\}|X_i] \leq c_5$ for $1 \leq i \leq N$.
6. Let $n := \min(n_A, n_B)$. Assi,e $m^{(k)}(X_i^T\beta)$ being uniformly bounded away from ∞ on $\mathscr{N}_{\theta,\tau} = \{\theta \in \mathbb{R}^{2p} : ||\theta_\theta - \theta_\theta^*|| \leq \tau\sqrt{s_\theta/n}, \theta_{\theta^c} = 0\}$ for some $\tau > 0$ and $k = 1, 2, 3$.
7. $\min_{j \in \mathscr{M}_\alpha} |\alpha_j^*|/\lambda_\alpha \to \infty$ and $\min_{k \in \mathscr{M}_\beta} |\beta_k^*|/\lambda_\beta \to \infty$ as $n \to \infty$.
8. $s_\theta = o(n^{1/3}), \lambda_\alpha, \lambda_\beta \to 0, \log(n)^2 = o(n\lambda_\theta^2), \log(p) = o(n\lambda_\theta^2/\log(n)^2), ps_\theta^4 \log(n)^6 = o(n^3\lambda_\theta^2)$ and $ps_\theta^4 \log(n)^8 = o(n^4\lambda_\theta^4)$ as $n \to \infty$.

Remarks

Remark 0.12

1. *If sample A is collected through simple random sampling or poisson sampling, only condition (1) is required.*
2. *Condition (3) specifies a fixed design which is well suited under the finite population inference framework.*
3. *Condition (5) holds for Gaussian distribution, sub-Gaussian distribution (strong tail decay).*
4. *Condition (6) holds for common model.*
5. *Condition (8) specifies the restriction on dimension of covariates p and dimension of the rtue non-zero coefficients s_θ.*

Existence Theorem

Theorem 5.1

Under the assumption stated above, $\exists \widetilde{\theta}$ *s.t.*

- $\mathrm{P}(|U_j^p(\widetilde{\theta})| = 0, j \in \mathscr{M}_\theta) \to 1.$
- $\mathrm{P}\left(\left|U_j^p(\widetilde{\theta}) \leq \frac{\lambda_\theta}{\log(n)}\right|, j \in \mathscr{M}_\theta^c\right) \to 1.$

which implies that $U(\widetilde{\theta}) = o_p\{\lambda_\theta / \log(n)\}$. *Also,*

- $\mathrm{P}(\widetilde{\theta}_{\mathscr{M}_\theta^c} = 0) \to 1.$
- $\widetilde{\theta}_{\mathscr{M}_\theta} - \theta^*_{\mathscr{M}_\theta} = O_p\{\sqrt{s_\theta / n}\}.$

implying that for large n, the penalized estimating equation procedure would NOT overselect irrelevant variables and estimate the true non-zero coefficients at the $\sqrt{s_\theta / n}$ *convergence rate, which is the so-called oracle property of variable selection.*

Asymptotic Normality

Under the assumption stated before, if either $\pi_B(X^T\alpha)$ or $m(X^T\beta)$ is correctly specified, then we have

$$\sqrt{n}\,(\widehat{\mu}_{p-dr} - \mu) \xrightarrow{d} N(0, V),$$

where $V := \lim_{n\to\infty}(V_1 + V_2)$, and

$$V_1 := \mathrm{E}\left\{ \frac{n}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} (\pi_{A,ij} - \pi_{A,i}\pi_{A,j}) \frac{m(X_i^T\beta^*)m(X_j^T\beta^*)}{\pi_{A,i}\pi_{A,j}} \right\}$$

$$V_2 := \mathrm{E}\left\{ \frac{n}{N^2} \sum_{i=1}^{N} \left(\frac{\mathbb{1}_{B,i}}{\pi_{B,i}(X_i^T\alpha^*)} - 1 \right)^2 \left(Y_i - m(X_i^T\beta^*) \right)^2 \right\}$$

Asymptotic Normality

We can estimate V_1 by its empirical version $\widehat{V}_1$. For V_2, we estimate it by

$$\widehat{V}_2 = \frac{n}{N^2}\sum_{i=1}^{N}\left[\left\{\frac{\mathbb{1}_{B,i}}{\pi_B(X_i^T\widehat{\alpha})} - \frac{2I_{B,i}}{\pi_B(X_i^T\widehat{\alpha})}\left(Y_i - m(X_i^T\widehat{\beta})\right)^2 + \mathbb{1}_{A,i}d_{A,i}\widehat{\sigma}^2(X_i)\right\}\right],$$

where $\widehat{\sigma}^2$ is arbitrary consistent estimator of $\sigma^2(X_i^T\beta^*)$. Define $\widehat{V} := \widehat{V}_1 + \widehat{V}_2$.

Under assumption stated before, if either $\pi_B(X^T\alpha)$ or $m(X^T\beta)$ is well-specified, then

$$\widehat{V} \xrightarrow{pr} V.$$

Simulation

Brief Comment

- In all the simulation trials, it perform the best in terms of the magnitude of bias.
- However in some cases, even though it is nearly unbiased, it show larger variability (Wider confidence band. It might be the tradeoff of the doubly robustness.

Thank You!