Double Robust Mass-Imputation with Matching Estimators

Ali Furkan Kalay

School of Economics, The University of Queensland

Introduction

This paper proposes to use the **Double Score Matching (DSM)** method to do mass-imputation and presents an application to make inferences with a non-representative sample. DSM is a nearest neighbor algorithm that uses two balancing scores instead of covariates to reduce the dimension of the distance metric. As long as one of two balance score models is correctly specified, DSM mass-imputation is a consistent estimator. This property is known as "double robustness." Moreover, DSM performs better than recently proposed double robust estimators when the data generating process has nonlinear confounders.

Mass-Imputation

Mass-imputation is the process of imputing a variable with predicted values. The predictions require auxiliary data/sample. The auxiliary sample is assumed to be a **nonprobability (non-representative) sample**. The paper considers two specific types of samples that are used in estimation:

Sample Type	d	Y	$\boldsymbol{X}$	Population Representative
Probability Sample (B)	✓	✗	✓	✓
Nonprobability Sample (A)	✗	✓	✓	✗

Note: d denotes the sample weights, Y is variable of interest, and $\boldsymbol{X}$ is a $k-$dimensional covariate.

Objectives

1. Consistent imputation of $Y \in B$, i.e., consistent estimation of
$$\mu_B = \sum_{i \in B} E[Y_i | \boldsymbol{X}_i].$$

2. Making inference with the nonprobability sample, i.e., consistent estimation of
$$\mu = \sum_{i \in F} E[Y_i | \boldsymbol{X}_i],$$
where F denotes the population.

3. Constructing valid confidence intervals for μ_B and μ estimates. It is non-trivial to establish asymptotic results because matching estimators make non-smooth predictions. Also, It is known that the naive bootstrapping fails with the matching estimators [2].

Dimension Reduction with Double Score Matching

Conventionally, nearest neighbors matching estimators are used to match observations based on Euclidean distances between covariates X_i. The prediction/imputation is the average of the best M matches. *The curse of dimensionality* follows matching estimators when dimension of $\boldsymbol{X}$, i.e., k is large. The convergence rate of the matching estimator is $O_p(N^{-1/2})$ when $k \leq 2$ and $O_p(N^{-1/k})$ when $k > 2$ [1].

Double Score Matching solves this problem by replacing $\boldsymbol{X}$ with estimated balance scores to reduce dimension.

1. Prognostic Score: $E[Y_i | \boldsymbol{X}_i]$,
2. Propensity Score: $Prob(I(i \in A) | \boldsymbol{X}_i)$.

The use of estimated two balance scores ($\hat{\boldsymbol{Z}}$) as a distance metric in matching ensures that the estimator converges sufficiently fast, i.e., $O_p(N^{-1/2})$. Therefore, we require functional form assumptions both for prognostic score and propensity score models.

Nonprobability (Non-representative) Samples

While mass-imputation could be implemented in different context, the paper focuses on nonprobability samples because nonprobability samples are becoming the dominant type of data, e.g., big data, internet surveys, and samples variables with large proportion of missing values. Also, the non-representative is a major concern in many economic studies.

An Example: Selection Mechanisms in COVID-19 Samples

Sample Selection Mechanism: People with COVID-19 symptoms are more likely to get a test (and hence more likely to be in the sample).
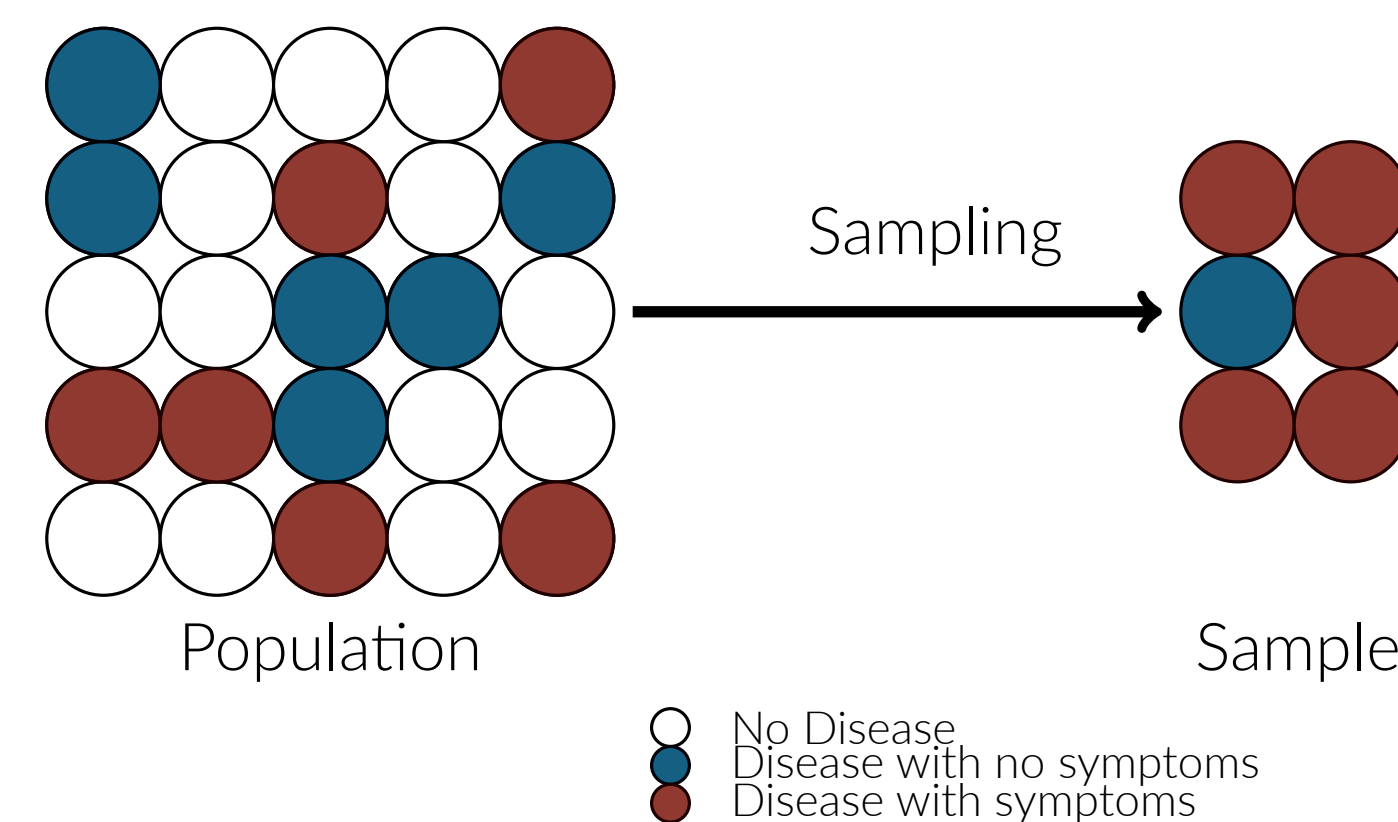
We cannot directly infer, for instance, the mortality rate from the available sample.

Solution: Transferring variable of interest from biased sample to a representative sample (Mass-imputation).

Results

I show that Double Score Matching mass-imputation is **double robust**, i.e., mass-imputation and population inferences are asymptotically consistent as long as prognostic score or propensity score models are correctly specified. The asymptotic results show that:

$$\hat{\mu}_B - \mu_B = O_p(N^{-1/2}) \quad \text{and} \quad \hat{\mu} - \mu = O_p(N^{-1/2})$$

where $\hat{\mu}_B$ and $\hat{\mu}$ denote DSM estimates respectively for sample B mean and population. The confidence intervals are constructed using a wild bootstrapping approach [5].

Beyond Double Robustness: The Violation of The Consistency Assumptions

When the consistency assumption fails, DSM mass-imputation is robust under certain scenarios, unlike other double robust inference methods. The consistency assumptions could fail, for example, if the data generating process is nonlinear. The robustness of DSM follows from the fact that the matching estimators are nonparametric, and hence DSM is a semi-parametric estimator. Consequently, DSM is less sensitive to model misspecifications because model predictions are merely used for matching observations.

Simulations

We simulate a super-population (F) and draw probability sample (B) and nonprobability sample (A) from F. The first simulation shows the **"double robustness"** of mass-imputation and population inferences. The second simulations shows that DSM performs well when the consistency assumptions fail due to nonlinear confounders. The simulations compare models in terms of Relative Bias (**RB**) and Mean Squared Error (**MSE**).

Results: Double Robustness

There are four possible scenarios for the prognostic and propensity score models' specifications, e.g., (**TF**) implies prognostic score model is a **TRUE (T)** specification and propensity score model is a **FALSE (F)** specification.

Estimator	Mean	RB (%)	MSE
Population Mean	9.278	-	-
Sample A Mean	11.906	28.331	6.949
Sample B Mean	10.120	-	-
DSM Mass-imputation (TT)	10.100	-0.193	0.399
DSM Mass-imputation (FT)	10.189	0.688	0.414
DSM Mass-imputation (TF)	10.111	-0.086	0.393
DSM Mass-imputation (FF)	12.386	22.403	5.542
DSM Population Inference (TT)	9.268	-0.105	0.407
DSM Population Inference (FT)	9.330	0.564	0.397
DSM Population Inference (TF)	9.282	0.044	0.380
DSM Population Inference (FF)	11.577	24.777	5.642

Results: Beyond Double Robustness

This simulation compares DSM with a recently proposed Double Robust Estimator (DRE) [4] when the DGP contains **nonlinear confounders**, i.e., the consistency assumptions fail. All models are misspecified due to nonlinearity, but some models, in addition, are misspecified due to omitted variable bias (denoted by **F**).

Estimator	Mean	RB (%)	MSE
DRE Population Inference (TT)	9.600	3.466	0.395
DRE Population Inference (FT)	9.819	5.827	0.621
DRE Population Inference (TF)	9.926	6.983	0.695
DRE Population Inference (FF)	11.665	25.724	5.955
DSM Population Inference (TT)	9.392	1.229	0.409
DSM Population Inference (FT)	9.400	1.315	0.416
DSM Population Inference (TF)	9.401	1.321	0.415
DSM Population Inference (FF)	11.559	24.586	5.567

Concluding Remarks

- The proposed method achieves good robustness properties and prevents bias from inflating when the consistency assumptions fail.
- It has been shown that the proposed Wild Bootstrapping approach produces valid confidence intervals. Furthermore, the constructed confidence intervals are double robust as the estimator itself.
- The theoretical findings could be extended into different context such as Average Treatment Effect Estimation [3] and Missing Value Imputation. (The latter is currently work-in-progress.)

References

[1] Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

[2] Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.

[3] Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018.

[4] Yilin Chen, Pengfei Li, and Changbao Wu. Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, 115(532):2011–2021, 2020.

[5] Taisuke Otsu and Yoshiyasu Rai. Bootstrap inference of matching estimators for average treatment effects. *Journal of the American Statistical Association*, 112(520):1720–1732, 2017.