CSCI 23: Research in Computing – Missing Data

Jacob M. Chen and Rohit Bhattacharya JMC8@WILLIAMS.EDU
Williams College Computer Science

1. Introduction

Missing data is a common problem encountered in many empirical sciences. It occurs when researchers are unable to completely observe all of the data for certain variables; when analyses are conducted without accounting for its potential missingness, it may lead to biased results. Fortunately, Saadati and Tian at Iowa State University have developed "a covariate adjustment formulation for controlling confounding bias in the presence of missing not at random (MNAR) data" (Saadati and Tian, 2019). Their m-adjustment criterion for missingness graphs, or m-graphs, is implemented in this project. M-graphs are a variant of causal directed acyclic graphs (DAGs) in which missingness mechanisms for each partially observed variable is augmented onto the graph. In the process of implementing the m-adjustment criterion, I find that the criterion is simple and easy to understand, but it is somewhat limiting. As a part of this project I also implemented data generation processes for different m-graphs that evaluate the ability of different adjustment sets to recover causal effects in missing data.

2. M-Adjustment Criterion

The basic formulation of the M-adjustment criterion is as follows. Let us assume that X is the treatment variable and that Y is the outcome variable. Z is a proposed adjustment set, and R_W is the set of all missingness mechanisms for variables in X, Y, and Z that are partially observed.[1]

1. No element of Z is either X, Y, an element on the proper causal path from X to Y, or a descendant of a variable on the proper causal path from X to Y.

2. Y and X are d-separated given Z and R_W in the graph where the first edge of every proper causal path from X to Y is removed.

3. Y and R_W are d-separated given X in the graph where all incoming edges to X are deleted.

4. X is either not an ancestor of R_W or is d-separated from Y in the graph where all outgoing edges from X are deleted.

[1] For formal definitions of proper causal paths and d-separation used below readers may refer to Saadati and Tian (2019) and Pearl (2009) respectively.

3. Implementation

I implemented Algorithm 1 from Saadati and Tian, which lists all possible m-adjustment sets that fulfill the m-adjustment criterion (Saadati and Tian, 2019). However, I implemented it differently than how they formulated it. Instead of using recursion, I iteratively enumerate all possible subsets of variables in the m-graph and test if each fulfills the m-adjustment criterion (Chen and Bhattacharya, 2022). My implementation also returns the m-adjustment set with the least amount of variables – the minimum adjustment set, which mirrors the output of Algorithm 2 in Saadati and Tian (2019). The running time of my implementation is $O(2^N)$ where N is the total amount of vertices in the m-graph. Any algorithm that finds all valid m-adjustment sets must be exponential because there can exist an exponential number of valid sets. However, Saadati and Tian (2019) formulate a more efficient algorithm to find a minimum m-adjustment set.

4. Testing Recoverability of Causal Effects Under Missing Data

In addition to implementing the m-adjustment criterion, I also created data generation processes in which I simulate missing data that follow data distributions implied by different m-graphs. I simulated and tested the m-graph in Figure 3 of Saadati and Tian (2019), which is reproduced below in Figure 1.

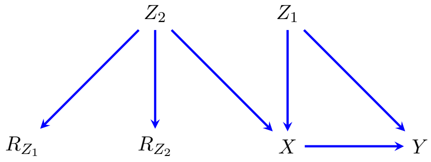

Figure 1: Simple M-graph

R_{Z_1} and R_{Z_2} are missingness mechanisms for the partially observed variables Z_1 and Z_2. According to the m-adjustment criterion, conditioning on Z_1 even when it is partially observed should still yield unbiased results albeit with higher variance. The point estimates and bootstrap distributions of estimates for causal effects are summarized in Figure 2. The true causal effect in the data generating process was 2.

I also tested an m-graph in which there are no valid m-adjustment sets in order to verify that I would obtain biased estimates for the causal effect in the absence of valid m-adjustment sets. The graph I tested for is shown in Figure 3. When there is no missingness, $\{Z_1\}$ is a sufficient adjustment set since the path through Z_2 is a collider. However, when Z_1 is partially observed, we are forced to condition on the missingness indicator, which is a descendant of a collider, and this biases our estimate for the causal effect. The point estimates and bootstrap distributions of estimates for causal effects are summarized in Figure 2. The true causal effect in the data generating process was 2.

Another m-graph I tested for, shown in Figure 4, analyzes the tradeoff between using a minimal adjustment set and an "efficient adjustment set." An efficient adjustment set is defined to be one that produces the lowest asymptotic variance in the effect estimates

Figure	Dataset Used	Adjustment Set	Point Estimate	Confidence Interval
Fig. 1	Fully Observed	$\{Z_1\}$	2.00699	(1.91267, 2.10880)
Fig. 1	Partially Observed	$\{Z_1\}$	2.03828	(1.90371, 2.17447)
Fig. 3	Fully Observed	$\{Z_1\}$	2.02303	(1.91864, 2.10653)
Fig. 3	Partially Observed	$\{Z_1\}$	1.79606	(1.65187, 1.93385)
Fig. 4	Fully Observed	$\{Z_3\}$	3.72593	(3.31904, 4.12357)
Fig. 4	Fully Observed	$\{Z_1, Z_3\}$	4.04051	(3.83324, 4.23238)
Fig. 4	Partially Observed	$\{Z_3\}$	3.72593	(3.30343, 4.07593)
Fig. 4	Partially Observed	$\{Z_1, Z_3\}$	4.05742	(3.83952, 4.28270)

Figure 2: Estimates for Causal Effects in Figures 1, 3, and 4

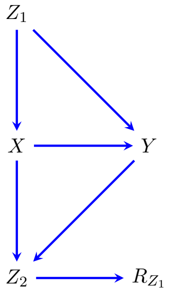

Figure 3: M-graph with a Collider

as given by the criterion described in Rotnitzky and Smucler (2020) for problems where variables are fully observed. When some of the variables in the efficient adjustment set are partially observed, it is currently unknown whether the same criterion still yields better estimates than using a smaller set of covariates. The minimal adjustment set in Figure 4 is $\{Z_3\}$; however, the efficient adjustment set is $\{Z_1, Z_3\}$. In the m-graph, Z_1 is a partially observed variable with its missingness being caused by Z_2. According to the m-adjustment criterion, $\{Z_1, Z_3\}$, the efficient adjustment set, is still valid. It turns out that, even when only Z_1 is partially observed, the adjustment set $\{Z_1, Z_3\}$ still yields more accurate estimates of the causal effect than Z_3 by itself. This result raises some open questions on whether the minimal or efficient adjustment set should be preferred when recovering causal effects from missing data. The point estimates and bootstrap distributions of estimates for causal effects are summarized in Figure 2. The true causal effect in the data generating process for this graph was 4.

5. Discussion and Conclusion

The simplicity of the m-adjustment criterion allows it to be implemented in such a manner that those who use the function do not need to anything about how it is implemented.

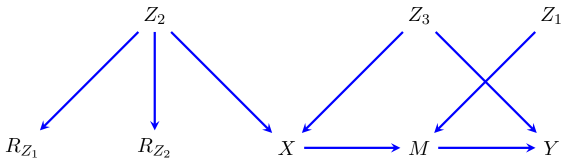

Figure 4: M-graph with Efficient Adjustment Set

Researchers that already have m-graphs drawn can find proper m-adjustment sets simply by calling the function listMAdj() in Chen and Bhattacharya (2022) without much preliminary knowledge of missing data or missingness mechanisms.

This m-adjustment criterion, however, trades its simplicity for a fairly limiting set of admissible adjustment sets. Because of the third requirement in the m-adjustment criterion, it would appear that any m-graph that has missingness indicators that are descendants of an element on a backdoor or proper causal path between the treatment and outcome will not fulfill the criterion. However, just because an adjustment set does not fulfill the m-adjustment criterion does not mean that the causal effect is not recoverable in that m-graph. For example, the following graph in Figure 5 is recoverable with the inverse probability weighting (IPW) method in which we reweight the distribution of the missing dataset with propensity scores for each missingness indicator (Bhattacharya et al., 2019). The new kernel we obtain can be treated as if the edges from the partially observed variables to each missingness indicator were deleted. Perhaps a combination of the m-adjustment criterion and the IPW method could lead to a method or algorithm that is capable of recovering causal effects for more general cases of missing data.

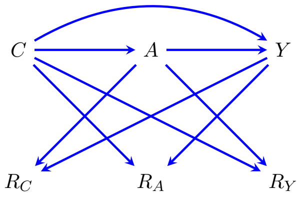

Figure 5: M-graph Recoverable with the IPW Method

It is easy to forget to take into account missing data during data analysis; however, doing so can lead to biased results. Considering how results from data analyses are often used to advise public policy, it is important that researchers use everything in their analytical toolkits to find estimates for unbiased results. Making missingness a more common consideration would contribute to that mission. Although the m-adjustment criterion discussed and implemented in this project is somewhat limiting, it is a step in the right direction in making accounting for missing data easier and more accessible.

References

Rohit Bhattacharya, Razieh Nabi, Ilya Shpitser, and James M. Robins. Identification in missing data models represented by directed acyclic graphs. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2019.

Jacob Chen and Rohit Bhattacharya. Implementation of the M-adjustment Criterion for Recovering Causal Effects from Missing Data, January 2022. URL `https://github.com/jacobchen01/m-adjustment`.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21:188–1, 2020.

Mojdeh Saadati and Jin Tian. Adjustment criteria for recovering causal effects from missing data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 561–577. Springer, 2019.