On Handling Self-masking and Other Hard Missing Data Problems

Karthika Mohan
Computer Science Department
University of California, Berkeley
karthika@berkeley.edu

Abstract

This paper tackles hard incomplete (missing) data problems such as those in which missingness in a variable is caused by itself. To address these problems we develop a new technique that jointly harnesses model and data as opposed to existing methods that exploit properties of the model alone. We present necessary and sufficient conditions under which consistent estimates of target quantities can be computed. In sharp contrast to other techniques used for dealing with similar problems, we do not make any parametric assumptions.

Introduction

Analysing and drawing inferences from missing data can be extremely challenging when the dataset contains variables that are themselves causes of their missingness; this type of missingness known as **self-masking** missingness, is believed to be the most commonly encountered type in practice [Osborne, 2012; Sverdlov, 2015; Adams, 2007; Mohan *et al.*, 2018]. Examples include smokers not answering questions pertaining to their smoking behavior in insurance applications, people with very high and very low income not disclosing their income and people of certain age groups not revealing their age.

Recent years have witnessed a growing interest in handling missing data using graphical models that encode assumptions about the underlying missingness process. [Daniel *et al.*, 2012; Mohan *et al.*, 2013; Shpitser *et al.*, 2015; Mohan and Pearl, 2018]. Given a target quantity Q and a graph G, Q is **recoverable** from G if there exists an algorithm that can consistently estimate Q for all data generated by G, else Q is **non-recoverable**. Such non-recoverable (Q, G) pairs which we call *hard missing data problems* (or hard problems) are the focus of this paper. Examples include $Q = P(O|do(t))$ and the self-masking model shown in figure 1 and $Q = P(X)$ and the self-masking model shown in figure 2 (b). While previous work treated recoverability as a property of graph alone, in this paper we develop general techniques to solve hard missing data problems by harnessing the properties of *both graph and data*.

In the following section we review missingness graphs i.e. graphical models for handling missing data [Mohan *et al.*, 2013].

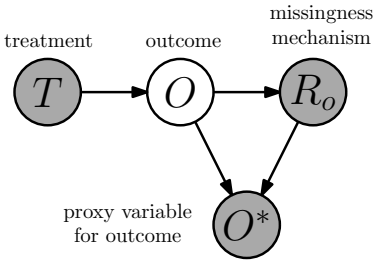


Figure 1: Missingness Graph in which outcome causes its own missingness

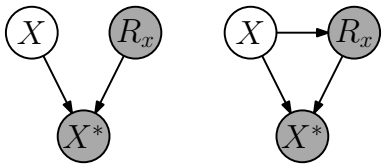
Missingness Graphs

Let $G(\mathbf{V}, E)$ be the causal DAG where $\mathbf{V}$ is the set of nodes and E is the set of edges. Nodes in the graph correspond to variables in the data set and are partitioned into five categories, i.e. $\mathbf{V} = V_o \cup V_m \cup U \cup V^* \cup R$.
V_o is the set of variables that are observed in all records in the population and V_m is the set of variables that are missing in at least one record. Variable X is termed as *fully observed* if $X \in V_o$ and *partially observed* if $X \in V_m$. R_{v_i} and V_i^* are two variables associated with every partially observed variable, where V_i^* is a proxy variable that is actually observed, and R_{v_i} represents the status of the causal mechanism responsible for the missingness of V_i^*; formally,

$$v_i^* = f(r_{v_i}, v_i) = \begin{cases} v_i & \text{if } r_{v_i} = 0 \\ m & \text{if } r_{v_i} = 1 \end{cases} \quad (1)$$

V^* is the set of all proxy variables and $\mathbf{R}$ is the set of all causal mechanisms that are responsible for missingness. Unless stated otherwise it is assumed that no variable in $V_o \cup V_m \cup U$ is a child of an R variable. U is the set of unobserved nodes, also called latent variables. Two nodes X and Y can be connected by a directed edge i.e. $X \rightarrow Y$, indicating that X is a cause of Y, or by a bi-directed edge $X \dashleftarrow\dashrightarrow Y$ denoting the existence of a U variable that is a parent of both X and Y. This graphical representation is called a **Missingness Graph** (or m-graph) [Mohan *et al.*, 2013]. $P(V^*, V_o, R)$ is called the observed data distribution.

Proxy variables may not always be explicitly shown in m-graphs in order to keep the figures simple and clear. Con-



(a) (b)

Figure 2: m-graphs in which (a) $P(X)$ is recoverable, (b) $P(X)$ is non-recoverable.

ditional Independencies are read off the graph using the d-separation criterion Pearl [2009]. For any binary variable X, x' and x denote $X = 0$ and $X = 1$ respectively.

Example 1. *In the m-graph in figure 1, T denotes the treatment administered to patients and O denotes the outcome. While T is observed for all patients, O is observed only for some of them. The edge from O to R_o indicates that the missingness is of self-masking type. $V_o = \{T\}$, $V_m = \{O\}$, $V^* = \{O^*\}$, $U = \emptyset$ and $R = \{R_o\}$.*

Missingness Mechanisms

Based on Rubin [1976], missing data problems can be classified into Missing Completely At Random (MCAR) , Missing At Random (MAR) and Missing Not At Random (MNAR). In this paper we use the graph based definition of these mechanisms [Mohan *et al.*, 2013].

An m-graph G depicts an MCAR problem if $(V_m, V_o) \perp\!\!\!\perp R$ holds in G, an MAR problem if $V_m \perp\!\!\!\perp R|V_o$ holds in G and an MNAR problem otherwise. For example, figure 2 (a) depicts MCAR, figure 3 (c) depicts MAR and figure 1 depicts MNAR missingness problems. Among these, joint distribution ($Q = P(V_o, V_m)$) is always consistently estimable (i.e. recoverable) when missingness is either MCAR or MAR (Mohan *et al.* [2013]). However this is not true for MNAR missingness. As such all hard missing data problems discussed in this paper belong to the MNAR category.

Recoverability as a property of m-graph

In this section we exemplify the notions of recoverability and non-recoverability as a property of the m-graph.

X^*	R_x	$P(X^*, R_x)$
0	0	0.3
1	0	0.2
m	1	0.5

Table 1: Observed Data Distribution

Suppose X is a binary variable corrupted by missing values. The dataset with missing values is shown in table 1. Figures 2 (a) & (b) depict two distinct (but statistically indistinguishable[1]) processes that could have generated this data. In figure 2 (a) missingness is generated by a purely random process while figure 2 (b) depicts self-masking missingness.

Recoverability: Consider the problem of recovering $P(X)$ given the m-graph G in figure 2 (a).

Since $X \perp\!\!\!\perp R_x$ in G we have,

$$P(X) = P(X|R_x = 0)$$

Using equation 1 we can rewrite the above as,

$$P(X) = P(X^*|R_x = 0)$$

By showing that $P(X)$ is a quantity that can be computed from the observed data distribution, we have established its recoverability. To do this we used the assumption $X \perp\!\!\!\perp R_x$ embedded in the m-graph. Hence in this case recoverability is a property of the m-graph alone. The recovered distribution $P(X)$ is shown in table 2.

X	$P(X)$
0	3/5
1	2/5

Table 2: Recovered Distribution

Non-recoverability: Now consider the problem of recovering $P(X)$ given the m-graph in figure 2 (b). R_x is dependent on X and we have no additional information regarding this dependence. It could be that X is missing only when its value is 1 or it could be that X is missing only when its value is 0. In the former case $P(x') = 0.3$ where as in the latter case $P(x') = 0.8$. Using the available information in G it is not possible to find the (true) value of $P(X)$ even if we are given infinitely many samples i.e. $P(X)$ is non-recoverable. In fact, non-recoverability of $P(X)$ would persist even if G had more variables in it (formally proved in [Mohan *et al.*, 2013; Mohan and Pearl, 2014a]).

Inability to handle hard problems such as self-masking missingness is a major deficiency in the field of missing data. Recent papers such as Shpitser [2016] and Mohan *et al.* [2013], and missing data text books such as Enders [2010] have called attention to the problem of recoverability in self-masking models. Standard Bayesian network textbooks such as Darwiche [2009] (chapter 17) and Koller and Friedman [2009] (chapter 19) discuss models similar to that in figure 1 and have shown that none of the existing methods such as the EM algorithm can recover parameters in self-masking m-graphs. In the following section we develop techniques to recover queries in hard problems and thus eliminate this deficiency in the field.

[1]Although d-separations are testable implications of a graphical model, under missingness not all d-separations are testable. In particular no statement of independence between a variable and its missingness mechanism ($X \perp\!\!\!\perp R_x$) is testable [Mohan and Pearl, 2014b].

Recoverability as a property of both m-graph and missing data

We exemplify below a procedure that exploits the properties of both graph and data to recover the joint distribution in a self masking model.

Example 2. *Consider the problem of recovering $P(O,T)$ given the m-graph G in figure 1 and the missing data distribution, $P(T,O^*,R_o)$. Let both T and O be binary variables. We will first recover $P(T|O)$ and then use it for recovering $P(O)$.*

Using $T \perp\!\!\!\perp R_o|O$ and eq 1, $P(T|O)$ can be recovered as,

$$P(T|O) = P(T|O, R_o = 0) = P(T|O^*, R_o = 0)$$

Since the variables are binary, $P(T) = \sum_O P(T|O)P(O)$ yields the following equations:

$$P(t') = P(t'|o')P(o') + P(t'|o)P(o)$$
$$P(t) = P(t|o')P(o') + P(t|o)P(o)$$

On substituting $P(T|O)$ in the equations above with its recovered estimand we get,

$$P(t') = P(t'|O^* = 0, r'_o)P(o') + P(t'|O^* = 1, r'_o)P(o)$$
$$P(t) = P(t|O^* = 0, r'_o)P(o') + P(t|O^* = 1, r'_o)P(o)$$

The two preceding equations constitute a system of equations in two unknowns: $P(o')$ and $P(o)$. If the solution is unique then $P(O)$ is recoverable and is given by,

$$P(o') = \frac{P(t') - P(t'|O^* = 1, R_o = 0)}{P(t'|O^* = 0, r'_o) - P(t'|O^* = 1, r'_o)}$$
$$P(o) = 1 - \frac{P(t') - P(t'|O^* = 1, R_o = 0)}{P(t'|O^* = 0, r'_o) - P(t'|O^* = 1, r'_o)}$$

$P(T,O)$ can now be recovered as: $P(T|O)P(O)$.

However, if the system of equations has infinitely many solutions then $P(O)$ is non-recoverable. This can happen when $T \perp\!\!\!\perp O$. In this case T provides no information about O and hence, cannot be leveraged to recover $P(O)$. This is to be expected since we do not insist on faithfulness and hence it is possible that an independence relation exist between two variable even when they are connected by an edge.

We further note that it is impossible for the system to have no solutions since it contradicts our assumption that the graph and data are compatible (i.e. there exist parameterization(s) of the graph that generated the data as per compatibility assumption).

Finally we note that as a result of recovering $P(O,T)$, we can also recover another hard problem: $P(O|do(t))$, the causal effect of treatment on outcome. Since G is Markovian, $P(O|do(t)) = P(O|t)$. Recoverability of $P(O|do(t))$ thus implicitly follows from that of $P(O,T)$.

Necessary and Sufficient Conditions for Recoverability in Hard Missing Data Problems

Notations $(\mathbf{M_{Z|W}}$, $\mathbf{M_Z}$ & $\mathbf{Aug(M_{Z|W}, M_W)})$ $M_{Z|W} = P(Z|W)$ denotes a $|Z| \times |W|$ matrix in which the columns sum to one. For example if Z and W are binary, then $M_{Z|W} = \begin{pmatrix} P(z'|w') & P(z'|w) \\ P(z|w') & P(z|w) \end{pmatrix}$. $M_Z = P(Z)$ denotes a $|Z| \times 1$ column matrix. $Aug(M_{Z|W}, M_W)$ denotes a $|Z| \times (|W| + 1)$ augmented matrix obtained by appending the columns of matrices $M_{Z|W}$ and M_W.

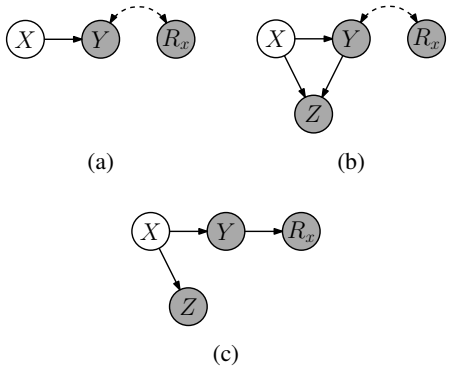


Figure 3: (a) $P(X,Y)$ is not recoverable, (b)$P(X,Y)$ may be recoverable using graph and data, (c) $P(X,Y)$ is recoverable.

The following theorem states the necessary and sufficient conditions for recoverability in hard missing data problems.

Theorem 1. *Let m-graph G and query $P(W)$ be such that the pair $(P(W), G)$ constitutes a hard missing data problem. Let $P(V^*, V_O, R)$ denote the distribution over missing data and $Z \subseteq \{V_m, V_o, R\} - \{W, R_w\}$.*
Given G and $P(V^*, V_O, R)$, $P(W)$ is recoverable if and only if $P(Z|W)$ and $P(Z)$ are recoverable and rank($M_{Z|W}$) = rank($Aug(M_{Z|W}, M_W)$) = $|W|$.

Proof: See Appendix.

The theorem makes no assumptions about the structure of m-graph G. It is *applicable to all hard MNAR problems* and not just to self-masking models. For example $P(X,Y)$ and the m-graph in figure 3 (a) and $P(X,Y)$ and the m-graph in figure 3 (b) constitute hard MNAR problems [Mohan and Pearl, 2014a]. However in the case of the latter, theorem 1 can be used to recover $P(X,Y)$ by leveraging Z.

Sufficient Conditions for Recoverability in Any Missing Data Problem

The following corollary states sufficient conditions for recovering *any* given query $P(W)$ using both graph and data.

Corollary 1. *Given m-graph G and missing data distribution $P(V^*, V_o, R)$, $P(W)$ is recoverable if for any $Z \subseteq \{V_m, V_o, R\} - \{W, R_w\}$, $P(Z|W)$ and $P(Z)$ are recoverable and rank($M_{Z|W}$) = rank($Aug(M_{Z|W}, M_W)$) = $|W|$.*

$P(X)$ and the m-graph in figure 3(c) do not constitute a hard missing data problem. In fact, the query is recoverable: $P(X) = \sum_Y P(X^*|Y, R_x = 0)P(Y)$ [Mohan *et al.*, 2013]. However, $P(X)$ can still be recovered using the preceding corollary by leveraging Z. X and Z being binary and Y having a high cardinality, say $|Y| = 50$ is an instance where applying the corollary is more convenient from a computational standpoint.

Remark 1. *Although theorem 1 and corollary 1 aim to recover $P(W)$, they can also be used to recover $P(Z, W)$ since $P(Z|W)$ is already known to be recoverable.*

Scope of Results: Given a query graph pair (Q, G), if Q is not recoverable using G then theorem 1 presents necessary and sufficient conditions for its recoverability. If Q is not recoverable using theorem 1 then we deem it as non-recoverable. Corollary 1 shows that the recoverability technique in theorem 1 is applicable to simple missing data problems such as MAR that are known to be recoverable using graphs. However, there exists problems that cannot be recovered using corollary 1 but can be recovered using graphs. For example $P(X, Y)$ cannot be recovered from $G : X \; Y \rightarrow R_X$ using corollary 1 since $X \perp\!\!\!\perp Y$. However $P(X, Y)$ is still recoverable as, $P(X^*|R_X = 0)P(Y)$ [Mohan *et al.*, 2013].

The preceding recoverability procedures are inspired by similar results in epidemiology (Rothman *et al.* [2008]), regression analysis (Carroll *et al.* [2006]) and causal inference (Pearl [2012]; Kuroki and Pearl [2014]). In contrast to Pearl [2012] that relied on external studies to compute causal effect in the presence of an unmeasured confounder, Kuroki and Pearl [2014] showed how the same could be effected without external studies. In missing data settings we have access to partial information that allows us to compute conditional distributions. This allows us to adapt the procedure in Pearl [2012] to compute consistent estimates as detailed above. We further note that to the best of our knowledge previous work on self-masking models relied on parametric assumptions (Mohan *et al.* [2018]; **?**]; **?**). In sharp contrast we present a complete and non-parametric solution to handle all hard problems.

Conclusions

In this work we eliminated a major deficiency in the field of missing data. We developed a sound, complete and non-parametric technique to handle *hard* missing data problems. Furthermore we showed that this technique is also applicable to queries that are known to be recoverable using graphs.

Appendix

Proof of theorem 1.

Proof of theorem 1 relies on the following lemma that states the conditions under which a system of linear equations is consistent. [Cramer, 1750; Strang, 1993].

Lemma 1. *The system of equations $\mathbf{Ax} = \mathbf{b}$ with m equations and n unknowns has (i) a unique solution if and only if $rank(A) = rank(Aug(A, b)) = n$ and (ii) infinite solutions if and only if $rank(A) = rank(Aug(A, b)) < n$.*

(Proof of sufficiency) When the conditions in the theorem are met the constraint $P(Z) = \sum_W P(Z|W)P(W)$ yields a unique solution as per lemma 1, thus establishing the recoverability of $P(W)$.

(Proof of necessity) We need to show that for every element Z_i in the power set of $\{V_m, V_o, R\} - \{W, R_w\}$, $P(W)$ is non-recoverable using G and data if any of the following hold: (i) $P(Z_i)$ is non-recoverable, (ii) $P(Z_i|W)$ is non-recoverable, (iii) rank($M_{Z_i|W}$) = rank($Aug(M_{Z_i|W}, M_W)$) = $|W|$ does not hold. Non-recoverability of $P(Z_i)$ implies that its value is not unique i.e there exists at least two distinct distributions $P_1(Z_i)$ and $P_2(Z_i)$. For each of them we can construct distinct distributions of $P(W)$ using $P(Z_i) = \sum_W P(Z_i|W)P(W)$, thereby proving that $P(W)$ is non-recoverable. Similarly, we can show that $P(W)$ is non-recoverable when $P(Z_i|W)$ is non-recoverable. In the case of condition (iii) non-recoverability of $P(W)$ follows from lemma 1.

References

J Adams. *Researching complementary and alternative medicine*. Routledge, 2007.

R J Carroll, D Ruppert, L A Stefanski, and C M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

G Cramer. *Introduction a l'analyse des lignes courbes algebriques par Gabriel Cramer...* chez les freres Cramer & Cl. Philibert, 1750.

R M Daniel, M G Kenward, S N Cousens, and B L De Stavola. Using causal diagrams to guide analysis in missing data problems. *Statistical Methods in Medical Research*, 21(3):243–256, 2012.

A Darwiche. *Modeling and reasoning with Bayesian networks*. Cambridge University Press, 2009.

C.K. Enders. *Applied Missing Data Analysis*. Guilford Press, 2010.

D Koller and N Friedman. *Probabilistic graphical models: principles and techniques*. 2009.

M Kuroki and J Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

K Mohan and J Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in NIPS 27*, pages 1520–1528. 2014.

K Mohan and J Pearl. On the testability of models with missing data. *Proceedings of AISTAT*, 2014.

K Mohan and J Pearl. Graphical models for processing missing data. Technical report, Department of Computer Science, University of California, Los Angeles, CA, 2018.

K Mohan, J Pearl, and J Tian. Graphical models for inference with missing data. In *Advances in NIPS 26*, pages 1277–1285. 2013.

K Mohan, F Thoemmes, and J Pearl. Estimation with incomplete data: The linear case. In *Proceedings of IJCAI-2018*, pages 5082–5088, 2018.

J W Osborne. *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage Publications, 2012.

J Pearl. *Causality: models, reasoning and inference*. Cambridge Univ Press, New York, 2009.

J Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

K J Rothman, S Greenland, and T L Lash. *Modern epidemiology*. Lippincott Williams & Wilkins, 2008.

D.B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

I Shpitser, K Mohan, and J Pearl. Missing data as a causal and probabilistic problem. In *Proceedings of UAI*, 2015.

I Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. In *Advances in NIPS*, pages 3144–3152, 2016.

G Strang. *Introduction to linear algebra*, volume 3. Wellesley-Cambridge Press Wellesley, MA, 1993.

O Sverdlov. *Modern adaptive randomized clinical trials: statistical and practical aspects*. Chapman and Hall/CRC, 2015.