TECHNICAL REPORT
R-415
March 2014

On the Testability of Models with Missing Data

Karthika Mohan and Judea Pearl
Department of Computer Science
University of California, Los Angeles
Los Angeles, CA-90025
{karthika, judea}@cs.ucla.edu

Abstract

Graphical models that depict the process by which data are lost are helpful in recovering information from missing data. We address the question of whether any such model can be submitted to a statistical test given that the data available are corrupted by missingness. We present sufficient conditions for testability in missing data applications and note the impediments for testability when data are contaminated by missing entries. Our results strengthen the available tests for MCAR and MAR and further provide tests in the category of MNAR. Furthermore, we provide sufficient conditions to detect the existence of dependence between a variable and its missingness mechanism. We use our results to show that model sensitivity persists in almost all models typically categorized as MNAR.

1 Introduction

Missing data has been traditionally formulated in statistical terms to determine under what conditions an unbiased estimate of parameters of interest can be obtained despite missingness. Recently, several proposals have been made to use graphical models as carriers of both conditional independence (CI) relations and the causal mechanism responsible for the missingness process. These proposals have successfully identified conditions under which consistent inferences can be drawn in the presence of missing data (Daniel et al., 2012; Garcia, 2013; Thoemmes & Rose, 2013; Mohan et al., 2013). In this paper, we ask whether it is possible to detect misspecifications of the missingness model, we demonstrate this possibility, and identify conditions that permit such detection.

Mohan et al. (2013) encoded missingness process using graphical models called m-graphs and derived conditions under which a joint distribution or a property thereof can be estimated consistently given two inputs: an m-graph G and a dataset D with partially observed variables. Not surprisingly, the results of this investigation reveal substantial sensitivity to the structure of the m-graph. In other words, some properties of P (called "queries") that are recoverable in one graph are not recoverable in another. Moreover, this sensitivity persists even when the two graphs are statistically indistinguishable and the natural question to ask is whether the structure of the m-graph lends itself to statistical tests, given that we are not in possession of the underlying distribution but a distortion thereof in the form of a dataset with missing values. We will show that such tests are indeed available albeit weaker than misspecification tests under complete data.

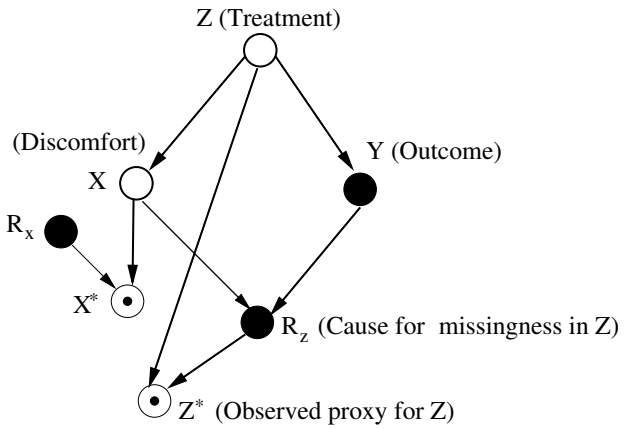

Figure 1: Example of an m-graph. Solid circles and hollow circles represent fully observed and partially observed variables respectively.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

$V_o = \{Y\}$	Let $W = \{Z, Y, R_x\}$
$V_m = \{X, Z\}$	$W_o = \{Y\}$
$R = \{R_x, R_z\}$	$W_m = \{Z\}$
$V^* = \{X^*, Z^*\}$	$W_r = \{R_x\}$
	$R_w = \{R_z\}$

Table 1: Notation relative to variables in Figure 1

In this paper we will limit our discussion to testable implications in the form of conditional independence claims entailed by the model. In Figure 1 for example, the model claims $X \perp\!\!\!\perp Y|Z$, $Z \perp\!\!\!\perp R_z|(X, Y)$ and $(X, Y, Z, R_z) \perp\!\!\!\perp R_x$. Such claims constitute the totality of testable implications if the underlying model is Markovian i.e. recursive and with independent error terms (Pearl, 2009). For constraints induced by latent variables, see Tian & Pearl (2002) and Shpitser & Pearl (2008).

This paper is organized as follows. In Section 2 we discuss the notion of m-graphs and recoverability. Section 3 defines testability of CIs portrayed by the m-graph and develops sufficient conditions under which a specific CI is testable given missing data. In Section 4 we call attention to an impediment which prevents testability of certain conditional independencies even when the distribution that carries these CIs is fully recoverable. We then present sufficient conditions for non-testability of CIs. Section 5 deals with testability of CIs comprising of substantive variables and presents sufficient conditions for such dependence to exist. In Section 6 we apply these theoretical results to classes of models which have been analysed in traditional missing data literature and show that (extending the results of Potthoff et al. (2006)) a large class of models traditionally thought of as non-testable are in fact testable. Finally, we use the results developed so far to show that model sensitivity persists in many models typically categorized as MNAR.

2 Preliminaries: m-graphs and Recoverability

We adopt the notations used in Mohan et al. (2013). Let $G(\mathbb{V}, E)$ be the causal DAG where $\mathbb{V} = V \cup U \cup V^* \cup \mathbb{R}$. V is the set of observable nodes. Nodes in the graph correspond to variables in the data set. U is the set of unobserved nodes (also called latent variables). E is the set of edges in the DAG. Oftentimes we use bi-directed edges as a shorthand notation to denote the existence of a U variable as common parent of two variables in $V_o \cup V_m \cup \mathbb{R}$. V is partitioned into V_o and V_m such that $V_o \subseteq V$ is the set of variables that are observed in all records in the population and $V_m \subseteq V$ is the set of variables that are missing in at least one record. Variable X is termed as *fully observed* if $X \in V_o$, *partially observed* if $X \in V_m$ and *substantive* if $X \in V_o \cup V_m$. Associated with every partially observed variable $V_i \in V_m$ are two other variables R_{v_i} and V_i^*, 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 $\mathbb{R}$ is the set of all causal mechanisms that are responsible for missingness. R variables may not be parents of variables in $V \cup U$. We call this graphical representation **Missingness Graph** or m-graph for short. An example of a m-graph is given in Figure 1 and the notations with respect to Figure 1 are explained in Table 1.

A *manifest distribution* $P(V_o, V^*, R)$ is the distribution that governs the available dataset. An *underlying distribution* $P(V_o, V_m, R)$ is said to be compatible with a given manifest distribution $P(V_o, V^*, R)$ if the latter can be obtained from the former using equation 1. In this paper we assume that all manifest distributions are strictly positive over complete cases and use the following shorthand. For any variable X, let X' be a shorthand for $X = 0$. For any set $W \subseteq V_m \cup V_o \cup R$, let W_r, W_o and W_m be the shorthand for $W \cap R$, $W \cap V_o$ and $W \cap V_m$ respectively. Let R_w be a shorthand for $R_{V_m \cap W}$ i.e. R_w is the set containing missingness mechanisms of all partially observed variables in W. Note that R_w and W_r are not the same. Table 1 offers an example.

2.1 Recoverability

Definition 1 (Recoverability (Mohan et al., 2013))**.** *Given a m-graph G, and a query Q defined on the variables in V, Q is said to be recoverable in G if there exists an algorithm that produces a consistent estimate of Q for every dataset D such that $P(D)$ is (1) compatible with G and (2) strictly positive over complete cases i.e. $P(V_o, V_m, \mathbb{R} = 0) > 0$.*

In layman terms, a given query Q is termed *recoverable* if in the limit of large samples a consistent estimate of Q can be computed from $P(D)$ and G, *as if* no data were missing.

The following example will demonstrate the sensitivity of recoverability to the structure of the graph.

Example 1. *Let G_1 (Figure 2(a)) and G_2 (Figure 2(b)) be the graphs hypothesized by the researcher for a given manifest distribution $P(X^*, Y^*, R_x, R_y)$. Let $P(X|Y)$ be the query to be recovered.*

(1) G_1 embeds the CI: $X \perp\!\!\!\perp R_x, R_y|Y$).

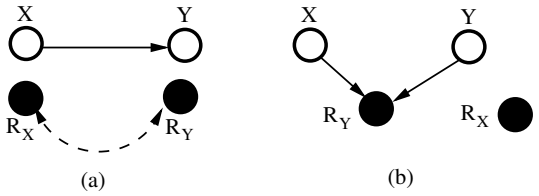

Figure 2: m-graphs that yield different estimands for the query $P(X|Y)$

Hence, $P(X|Y) = P(X|Y, R_x = 0, R_y = 0)$

On applying Equation-(1) we get:

$P(X|Y) = P(X^*|Y^*, R_x = 0, R_y = 0)$.

(2) G_2 *embeds the CI:* $X \perp\!\!\!\perp Y$. *Therefore,*

$P(X|Y) = P(X)$

G_2 *also embeds the CI:* $X \perp\!\!\!\perp R_x$. *Therefore,*

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

On applying Equation-(1) we get:

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

We observe that G_1 *and* G_2 *dictate different estimands which yield different results depending on the missingness process that each portrays. Therefore it is imperative to test whether the manifest distribution and hypothesized model are compatible.*

3 Testability of CI (d-separations) in m-graphs

Definition 2 (Testable d-separation)**.** *Let* $X \cup Y \cup Z \subseteq V_o \cup V_m \cup R$ *and* $X \cap Y \cap Z = \emptyset$. $X \perp\!\!\!\perp Y|Z$ *is testable if there exists a dataset* D *governed by a distribution* $P(V_o, V^*, R)$ *such that* $X \perp\!\!\!\perp Y|Z$ *is refuted in all underlying distributions* $P(V_o, V_m, R)$ *compatible with the distribution* $P(V_o, V^*, R)$.

If X and Y are singletons, $X \perp\!\!\!\perp Y|Z$ is termed as **singleton d-separation** and if not, $X \perp\!\!\!\perp Y|Z$ is termed as **compound d-separation**. Let us look at examples of singleton and compound d-separations that are testable.

Example 2. *Let* $X \in V_o$ *and* $Y \in V_m$. $X \perp\!\!\!\perp R_y$ *is testable since* X *and* R_y *are fully observed variables and we can always find a dataset that refutes* $P(X|R_y = 0) = P(X|R_y = 1)$. *Similarly when* $\{X, Y\} \subseteq V_m$, $R_x \perp\!\!\!\perp R_y$ *and* $X \perp\!\!\!\perp Y|(R_x = 0, R_y = 0)$ *are testable.* $R_x \perp\!\!\!\perp R_y$ *is testable since* R_x *and* R_y *are fully observed variables.* $X \perp\!\!\!\perp Y|(R_x = 0, R_y = 0)$ *is testable since given* $R_x = 0$ *and* $R_y = 0$ *we can apply Equation 1 and equivalently write the CI as* $X^* \perp\!\!\!\perp Y^*|(R_x = 0, R_y = 0)$ *i.e. CI can be expressed equivalently in terms of observed variables and hence it can be refuted.*

Example 3. *Following are two examples of compound d-separations that are testable.*

a. CI: $(X, R_x) \perp\!\!\!\perp (Y, R_y)|(Z, R_z)$ *implies*
$P(X^*, R_x^{'}|Z^*, R_z^{'}) = P(X^*, R_x^{'}|Y^*, R_y^{'}, Z^*, R_z^{'})$

b. CI: $(X, R_x) \perp\!\!\!\perp (R_w, R_y)|Y$ *implies*
$P(X^*, R_x^{'}|Y^*, R_y^{'}, R_w) = P(X^*, R_x^{'}|Y^*, R_y^{'}, R_w^{'})$

Since both CIs imply CIs that can be expressed in terms of observed variables, the CIs can be refuted. Hence they are testable.

We would like to remark that there exist non-testable CI claims and they are discussed in Section 4, Example 5. From definition-2, we conclude that a d-separation is termed testable when it has at least one implication that is testable. Example:4 demonstrates that, in some cases, it might be necessary to examine all implications of a compound d-separation before labeling it as testable.

Example 4. *Consider the d-separation* $S : (X, R_y, R_{z1}) \perp\!\!\!\perp (Y, R_x, R_{z2})|(Z_1, Z_2)$. *This d-separation translates to* $\frac{P(X,R_y^{'},R_{z1}^{'},Y=0,R_x^{'},R_{z2}^{'})}{P(Y=0,R_x^{'},R_{z2}^{'},Z_1,Z_2)} = \frac{P(X,R_y^{'},R_{z1}^{'},Y=1,R_x^{'},R_{z2}^{'})}{P(Y=1,R_x^{'},R_{z2}^{'},Z_1,Z_2)}$. *Observe that the denominators cannot be directly expressed in terms of observed variables. To affirm testability of* S, *we have to examine its implications until we find an implication that is testable. For example,* $S' : X \perp\!\!\!\perp Y|(Z_1, Z_2, R_y, R_{z1}, R_x, R_{z2})$, *obtained by applying weak union graphoid axiom to* S *is testable since it translates into* $\frac{P(P(X,R_y^{'},R_{z1}^{'},Y=0,R_x^{'},R_{z2}^{'}))}{P(R_y^{'},R_{z1}^{'},Y=0,R_x^{'},R_{z2}^{'})} = \frac{P(P(X,R_y^{'},R_{z1}^{'},Y=1,R_x^{'},R_{z2}^{'}))}{P(R_y^{'},R_{z1}^{'},Y=1,R_x^{'},R_{z2}^{'})}$. *Since* S' *is testable we can conclude that* S *is testable.*

Clearly enumerating and testing the set of all implied d-separations is hard since the number of implications is exponential in the sizes of sets X and Y. The next subsection provides a rule to circumvent this enumeration for certain types of d-separations.

3.1 Directly testable d-separations

Testability of certain d-separations (such as the compound d-separations in Example:3) can be affirmed in one shot i.e. without explicitly examining all their implications. In other words, testability can be certified by looking at the placement of a mechanism R_X

relative to its partially observed variable X in the d-separation statement. We call such d-separations **directly testable**. The following is a syntactic criterion for determining direct testability of d-separations.

Theorem 1. *Let $X, Y, Z \subset V_o \cup V_m \cup \mathbb{R}$ and $X \cap Y \cap Z = \emptyset$. The conditional independence statement S: $X \perp\!\!\!\perp Y | Z$ is directly testable if all the following conditions hold:*

1. $Y \not\subseteq (R_{X_m} \cup R_{Z_m})$
   In words, Y should contain at least one element that is not in $R_{X_m} \cup R_{Z_m}$.
2. $R_{X_m} \subseteq X \cup Y \cup Z$
   In words, the missingness mechanisms of all partially observed variables in X are contained in $X \cup Y \cup Z$.
3. $R_{Z_m} \cup R_{Y_m} \subseteq Z \cup Y$
   In words, the missingness mechanisms of all partially observed variables in Y and Z are contained in $Y \cup Z$.

Proof. Let $Y_1 \in R \cup V_o \cup V_m$ be an element in Y such that condition (1) is satisfied. $X \perp\!\!\!\perp Y | Z$ implies:
$\frac{P(X, Y_1=0, Y-Y_1, Z)}{P(Y_1=0, Y-Y_1, Z)} = \frac{P(X, Y_1=1, Y-Y_1, Z)}{P(Y_1=1, Y-Y_1, Z)} \qquad (a)$
From conditions (2) and (3), we know that the terms in the numerator of both fractions contain R_{X_m}, R_{Y_m} and R_{Z_m}. Similarly, from condition (3), we know that the terms in the denominator of both fractions contain R_{Y_m} and R_{Z_m}. When all R variables in $R_{X_m} \cup R_{Y_m} \cup R_{Z_m}$ are set to zero we can apply Equation 1 and express the numerators and denominators of equation-(a) in terms of observed variables, thereby making the claim testable. □

Corollary 1. *A given graphical model G is testable if it has one of the following directly testable singleton conditional independencies:*

1. $X \perp\!\!\!\perp Y | Z, R_x, R_y, R_z$
2. $X \perp\!\!\!\perp R_y | Z, R_x, R_z$
3. $R_x \perp\!\!\!\perp R_y | Z, R_z$

It is understood that, if X or Y or Z are fully observed, the corresponding missingness mechanism may be removed from the conditioning set. Clearly, any conditional independence comprised exclusively of fully observed variables is testable.

So far we have discussed testable CI statements. In the following section we shall discuss an impediment to testability when data are afflicted by missingness.

4 Impediments to Testability in Missing Data

Unlike testability under complete data, testability in missing data has an impediment to overcome. When data are complete we simply select a conditional independence statement in the model and test it against the data. Under missing data however, some conditional independencies in the model may not be testable even when the joint distribution is recoverable. An example demonstrating this impediment is discussed below.

Example 5. *Consider the missingness process described by the graph G in Figure 5 (a) that states the CI: $X \perp\!\!\!\perp R_x | Y$. Let $Q : P(X, Y, R_x)$ be the query to be recovered. We will show that although Q is recoverable, the CI statement $X \perp\!\!\!\perp R_x | Y$ is not testable.*

First we will prove that Q is recoverable.

$$P(X, Y, R_x = 1) = P(X | Y, R_x = 1) P(Y, R_x = 1)$$

Since G embeds $X \perp\!\!\!\perp R_x | Y$ we have,

$P(X | Y, R_x = 1) = P(X | Y, R_x = 0)$. *Therefore,*

$$P(X, Y, R_x = 1) = P(X | Y, R_x = 0) P(Y, R_x = 1)$$

Using Equation 1,

$P(X | Y, R_x = 0) = P(X^* | Y, R_x = 0)$. *Therefore,*

$$P(X, Y, R_x = 1) = P(X^* | Y, R_x = 0) P(Y, R_x = 1)$$

Hence $P(X, Y, R_x = 1)$ is recoverable.

Using Equation 1,
$P(X, Y, R_x = 0) = P(X^*, Y, R_x = 0)$. *Thus,* $P(X, Y, R_x = 0)$ *is also recoverable.*

We will now show that $X \perp\!\!\!\perp R_x | Y$ is not testable. $X \perp\!\!\!\perp R_x | Y$ *translates into,*
$P(X | Y, R_x = 1) = P(X | Y, R_x = 0)$
Hence, $P(X, Y, R_x = 1) = \frac{P(X, Y, R_x = 0)}{P(Y, R_x = 0)} P(Y, R_x = 1)$
In other words, for any manifest distribution $P^*(X^*, Y, R_X)$ in which $P^*(Y, R_X = 0) > 0$, we can always construct (as shown below) a compatible distribution $P(X, Y, R_X)$ in which the CI statement $X \perp\!\!\!\perp R_x | Y$ holds.

$\forall x, y$

$$P(X = x, Y = y, R_x = 0) = P^*(X^* = x, Y = y, R_x = 0)$$
$$P(X = x, Y = y, R_x = 1) = \frac{P^*(X^* = x, Y = y, R_x = 0)}{P^*(Y = y, R_x = 0)} * P^*(Y = y, R_x = 1)$$

Thus, $X \perp\!\!\!\perp R_x | Y$ is not refutable and hence we conclude that it is not testable.

This example showed that a probability distribution $P(v)$ can be perfectly recoverable from missingness, (i.e., it can be estimated consistently, *as if* no missingness occurred) and yet, $P(v)$ may have testable implications (eg, conditional independence (CI) statements) that are not testable for any data with the same manifest structure (i.e. the same sets of partially and fully observed variables).

The explanation of this impediment is as follows. When we say $P(V)$ has testable implications we refer to refutation by some distribution taken from the space of all distributions on V. In contrast, when we say 'testable under missingness' we demand refutation by a set of distributions with the same manifest structure. The refutation power of the latter set is weaker than the former.

The next theorem characterizes a set of CI that are not testable from missing data.

Theorem 2. *Given that $Y \subseteq V_o \cup R$, the singleton d-separation $X \perp\!\!\!\perp R_x | Y$ is not testable.*

Proof. We can always compute $P(X, R_x = 1, Y)$ as $P(X, R_x = 1, Y) = P(R_x = 1, Y)P(X^*|R_x = 0, Y)$ such that $X \perp\!\!\!\perp R_x|Y$ is always true. Hence $X \perp\!\!\!\perp R_x|Y$ is not refutable given any manifest distribution that is strictly positive over complete cases. Hence $X \perp\!\!\!\perp R_x|Y$ is not testable. □

Corollary 2. *Given that Y contains at least one partially observed variable and $R_{y_m} \subset Y$, singleton conditional independence $X \perp\!\!\!\perp R_x | Y_r = 0, Y - Y_r$ is not testable.*

Corollary 3. *Direct Testability of a conditional independence statement does not imply testability of all its implications.*

Proof. Consider the CI statement $X \perp\!\!\!\perp (Y, R_y, R_x)$. On applying decomposition graphoid axiom, we get the non-testable CI: $X \perp\!\!\!\perp R_x$. □

However, there exist directly testable d-separations whose implications obtained by weak union and decomposition graphoid axioms are always testable.

Example 6. *Let $X \perp\!\!\!\perp Y | Z, R_z, R_x, R_y$ be a compound d-separation such that $X \cup Y \cup Z \subseteq V_o \cup V_m$. In this case it can be easily seen that all implications obtained by applying decomposition and weak union graphoid axioms comply with conditions for direct testability given in Theorem-1. Hence they are all testable.*

5 Testability of CIs comprising of only substantive variables

Let us examine the testability of singleton CI: $X \perp\!\!\!\perp Y$. Clearly, when $X, Y \in V_o$, $X \perp\!\!\!\perp Y$ is testable. However, testability of $X \perp\!\!\!\perp Y$ when $X \in V_o$ and $Y \in V_m$ is not obvious. In the following theorem we prove that $X \perp\!\!\!\perp Y$ is testable when $X \in V_o$ and $Y \in V_m$ and, X and Y are *binary*. We further specify necessary conditions that the manifest distribution must satisfy for $X \perp\!\!\!\perp Y$ to hold true in the underlying distribution.

Theorem 3. *Given that $X \in V_o$ and $Y \in V_m$, the conditional independence statement $X \perp\!\!\!\perp Y$ is testable. Moreover, a graph depicting $X \perp\!\!\!\perp Y$ should be summarily rejected if none of the following conditions hold:*

$$0 \leq \frac{-k}{P(x)} \leq P(x', r_y) \tag{2}$$

$$0 \leq \frac{k}{P(x')} \leq P(x, r_y) \tag{3}$$

$$0 \leq \frac{k + P(x)P(x', r_y)}{P(x')} \leq P(x, r_y) \tag{4}$$

$$0 \leq \frac{P(x')P(x, r_y) - k}{P(x)} \leq P(x', r_y) \tag{5}$$

where $k = P(x)(P(x', y, r'_y) + P(x, y, r'_y)) - P(x, y, r'_y)$.

Proof. We first show that violation of all conditions from 2 to 5 is sufficient to rule out $X \perp\!\!\!\perp Y$. Then by constructing an example that violates conditions 2 to 5, we confirm the testability of $X \perp\!\!\!\perp Y$.

$X \perp\!\!\!\perp Y$ may be equivalently written as,
$P(x, y) = P(x)P(y)$
The equation above is equivalent to,
$P(x, y, r_y) - P(x)(P(x', y, r_y) + P(x, y, r_y)) = P(x)(P(x', y, r'_y) + P(x, y, r'_y)) - P(x, y, r'_y)$

Let the constant terms in RHS evaluate to k. Then we can rewrite the equation as:

$$P(x', y, r_y) = \frac{P(x')}{P(x)} P(x, y, r_y) + \frac{-k}{P(x)} \tag{6}$$

Equation 6 is linear, the variables are $P(x', y, r_y)$ and $P(x, y, r_y)$ and it resembles the general equation of a line:y=mx+c. Equation 6 should also satisfy:
(a) $0 \leq P(x, y, r_y) \leq P(x, r_y)$
(b) $0 \leq P(x', y, r_y) \leq P(x', r_y)$
The constraints (a) and (b) above delineate a rectangular region $\Re$ in the first quadrant of the Cartesian plane. Equation 6 can be solved subject to constraints (a) and (b) only if the line described in Equation 6 intersects the boundary lines enclosing $\Re$ (i.e. at least one intersection point should satisfy (a) and (b)).

Intersection of Eq 6 and left boundary of $\Re$ yields: $0 \leq \frac{-k}{P(x)} \leq P(x', r_y)$

Intersection of Eq 6 and bottom boundary of $\Re$ yields: $0 \leq \frac{k}{P(x')} \leq P(x, r_y)$

Intersection of Eq 6 and top boundary of $\Re$ yields: $0 \leq \frac{k+P(x)P(x',r_y)}{P(x')} \leq P(x, r_y)$

Intersection of Eq 6 and right boundary of $\Re$ yields: $0 \leq \frac{P(x')P(x,r_y)-k}{P(x)} \leq P(x', r_y)$

We prove testability of $X \perp\!\!\!\perp Y$ by presenting manifest distribution P_3 in Table 2 that violates conditions 2 to 5 and thus refutes the claim: $X \perp\!\!\!\perp Y$. □

R_y	X	Y^*	P_1	P_2	P_3
0	0	0	$\frac{8}{81}$	$\frac{13}{41}$	$\frac{100}{125}$
0	0	1	$\frac{6}{81}$	$\frac{11}{41}$	$\frac{5}{125}$
0	1	0	$\frac{4}{81}$	$\frac{7}{41}$	$\frac{7}{125}$
0	1	1	$\frac{3}{81}$	$\frac{5}{41}$	$\frac{3}{125}$
1	0	m	$\frac{20}{81}$	$\frac{3}{41}$	$\frac{5}{125}$
1	1	m	$\frac{40}{81}$	$\frac{2}{41}$	$\frac{5}{125}$

Table 2: $X \perp\!\!\!\perp Y$ can hold in manifest distributions P_1 and P_2 but cannot hold in manifest distribution P_3

Example 7. *Table 2 describes three distributions; P_1 and P_2 in which $X \perp\!\!\!\perp Y$ could possibly hold and P_3 in which $X \perp\!\!\!\perp Y$ cannot hold. $X \perp\!\!\!\perp Y$ can possibly hold in P_1 and P_2 because both the distributions satisfy condition 3. P_3 does not satisfy any of the conditions from 2 to 5; hence $X \perp\!\!\!\perp Y$ cannot hold in P_3.*

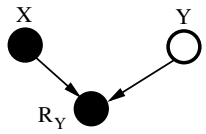

Figure 3: m-graph in which recoverability of $P(X|Y)$ depends only on $X \perp\!\!\!\perp Y$.

The following example demonstrates an application of Theorem 3. It describes an instance where recoverability of a given query hinges exclusively on the independence between X and Y.

Example 8. *Let G_1 in Figure 3 be the hypothesized graph and $Q = P(X|Y)$ be the query to be recovered. $P(X, Y)$ is not recoverable from G_1 since Y itself is the cause of its missingness(R_y). G_1 embeds the CI statement: $X \perp\!\!\!\perp Y$ and if we assume G_1 is the true graph then $P(X|Y)$ can be recovered as follows:*

$$P(X|Y) = P(X)$$

Recoverability however depends critically on the independence $X \perp\!\!\!\perp Y$ embedded in G_1. Our question is whether or not the CI statement $X \perp\!\!\!\perp Y$ holds in any underlying distribution compatible with the data available. Theorem 3 answers this question immediately by providing us with four conditions, one of which ought to be satisfied by the manifest distribution for $X \perp\!\!\!\perp Y$ to hold. For example, given P_3 in Table 2 and G_1, we can immediately conclude that G_1 and P_3 are not compatible.

It is interesting to note that though recoverability is generally facilitated by (usually non-testable) CI between a variable and its missingness mechanism such as $Y \perp\!\!\!\perp R_y$ or $Y \perp\!\!\!\perp R_y|X$, in Example 8 recoverability of Q facilitated by the independence between substantive variables X and Y.

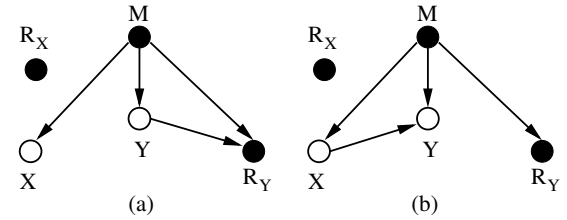

Figure 4: (a) m-graph depicting MNAR (b) m-graph depicting MAR^+

6 Testability of MCAR and MAR

Missingness mechanisms are traditionally classified into three categories (Rubin, 1976): Missing Completely At Random(MCAR), Missing At Random(MAR) and Missing Not At Random (MNAR). A chi square based test for MCAR was proposed by Little (1988) in which a higher value falsified MCAR(Rubin, 1976). MAR (Rubin, 1976) is not testable. Given below is a quote from Allison (2002),pg:4.

"It is impossible to test whether the MAR condition is satisfied, and the reason should be intuitively clear. Because we do not know the values of the missing data, we can not compare the values of those with and without missing data to see if they differ systematically on that variable."

Potthoff et al. (2006) defined MAR at the variable-level and named it MAR^+ and showed that it can be tested. A given dataset is MAR^+ if $V_m \perp\!\!\!\perp R|V_o$ (Potthoff et al., 2006). Theorem 4, given below presents stronger conditions under which a given MAR^+ model is testable. Furthermore, it provides diagnostic insight in case the test is violated.

Theorem 4. *Given that $|V_m| > 0$, MAR^+ ($V_m \perp\!\!\!\perp R|V_o$) is testable if and only if $|V_m| > 1$ i.e. $|V_m|$ is not a singleton set.*

Proof. Let $|V_m| = k > 1$ and $X \subseteq V_m$ such that $|X| = k - 1$. By applying decomposition graphoid axiom to $V_m \perp\!\!\!\perp R | V_o$, we get $(V_m - X) \perp\!\!\!\perp R | V_o$ that is directly testable by Theorem 1. Therefore, $V_m \perp\!\!\!\perp R | V_o$ is testable if V_m is not a singleton. On the other hand if $|V_m| = 1$ then by Theorem-2, $V_m \perp\!\!\!\perp R | V_o$ is not testable. □

Example 9. *In the graph in Figure 4(b), MAR^+ holds because $(Y, X) \perp\!\!\!\perp (R_x, R_y) | M$. Therefore, the tests are:*
(i) $X^* \perp\!\!\!\perp R_y | M, R_x = 0$
(ii) $Y^* \perp\!\!\!\perp R_x | M, R_y = 0$
(iii) $X^* \perp\!\!\!\perp R_y = 0 | M, R_x = 0, Y$
(iv) $Y^* \perp\!\!\!\perp R_x = 0 | M, R_y = 0, X$
(i) and (ii) are tests obtained by applying weak union and decomposition graphoid axioms to $(Y, X) \perp\!\!\!\perp (R_x, R_y) | M$ and (iii) and (iv) are tests obtained by applying weak-union graphoid axiom to $(Y, X) \perp\!\!\!\perp (R_x, R_y) | M$. Note that the graph has more testable implications than those listed above. For example, the graph advertises the CI statement $R_x \perp\!\!\!\perp R_y$. However, the latter test is model specific, whereas (i)-(iv) are model-independent, applicable to any MAR^+ model with the same manifest structure.

The following corollary shows that MCAR is testable.

Corollary 4. *Given that $|V_m| > 0$, MCAR ($(V_m \cup V_o) \perp\!\!\!\perp R$) is testable if and only if $|V_o \cup V_m| \geq 2$.*

If the dataset contains only one variable(X) and $X \in V_m$, then $X \perp\!\!\!\perp R_x$ is not testable (by Theorem 2), even though the corresponding missingness mechanism is MCAR. If the dataset additionally contained at least another fully observed variable (Y) then $(X, Y) \perp\!\!\!\perp R_x$ is testable since its implication $Y \perp\!\!\!\perp R_x$ is testable. On the other hand, if the dataset additionally contained at least another partially observed variable (Z) then $(X, Z) \perp\!\!\!\perp (R_x, R_z)$ is testable since its implications such as $Z \perp\!\!\!\perp R_x | R_z = 0$ and $X \perp\!\!\!\perp R_z | R_x = 0$ are testable.

6.1 Detecting MNAR missingness mechanism

Consider the graph in Figure 4(a). The model is clearly MNAR since there is an edge between Y and R_y. However, Theorem 4 will not be able to falsify MAR^+. The following subsection will show that such falsification is nevertheless possible.

6.1.1 Graph based tests for detecting the edge between a variable an its missingness mechanism (eg: $X \rightarrow R_x$)

Ordinarily an edge E between a variable and its missingness mechanism is not testable. However, if the contentious edge is embedded in a structure that meets certain conditions we will show that a test exists to ascertain the existence of E. The following lemma gives the condition under which an edge $X \rightarrow R_x$ may be detected in a Markovian Model.

Lemma 1. *Given a Markovian model in which (1) there exists Z which is a parent of X and not a parent of R_x and (2) no R variable is a parent of another R variable, an edge $X \rightarrow R_x$ exists whenever $Z \not\perp\!\!\!\perp R_x | R_z = 0, (R \cup V) - \{X, Z, R_x, R_z\}$.*

Proof. Condition (2) prevents R_x from being a parent of any node in R and by definition of m-graph R_x cannot be a parent of variables in $V_o \cup V_m$. Hence no variable in $V_o \cup V_m \cup R$ is a child of any R variable. Moreover, the model is Markovian. Therefore the m-graph can only contain uni-directed edges that enter R_x and thus no parent of R_x can be a collider on any path that enters R_x. In the test, $Z \perp\!\!\!\perp R_x | R_z = 0, (R \cup V) - \{X, Z, R_x, R_z\}$ we condition on all variables except X. Therefore, if the test does not hold true then it is because there is an unblocked path from Z to R_x via X (by condition-1, $Z \rightarrow R_x$ does not exist). This is possible only if X is a parent of R_x i.e. there exists an edge between X and R_x.

□

Example 10. *Consider the m-graph G_1 in Figure 1 that implies $X \perp\!\!\!\perp R_x$. Let it be the case that Z does not cause the missingness in X. Then, we can confirm dependence i.e. the existence of $X \rightarrow R_x$, if $Z \perp\!\!\!\perp R_x | Y, R_z = 0$ does not hold.*

7 Model Sensitivity of Estimation Procedures

An important consequence of identifying the testable implications of a given model is the ability to demonstrate the limits of model-blind algorithms, i.e. algorithms that attempt to handle missing-data problems on the basis of the data alone, without making any assumptions about the structure of the missingness process. A fundamental limitation of model-blind algorithms is unveiled in Example 11, which presents two statistically indistinguishable models such that a given query is recoverable in one and non-recoverable in the other.

Example 11. *The two graphs in Fig. 5 (a) and (b) cannot be distinguished by any statistical means, since Fig. 5(a) has no testable implications and Fig. 5(b) is a complete graph. However in Fig. 5 (a) $P(X, Y)$ is recoverable (refer Example 5)while in Fig. 5 (b)*

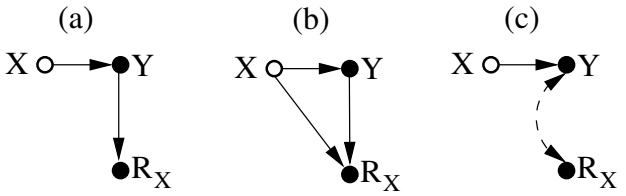

Figure 5: Statistically indistinguishable graphs. (a) $P(X, Y)$ is recoverable (b) $P(X, Y)$ is not recoverable (c) $P(X)$ is recoverable

$P(X, Y)$ is not recoverable (by Theorem-2 in Mohan et al. (2013)).

An even stronger limitation is demonstrated in Example 12; it shows that no model-blind algorithm exists[1] even in those cases where recoverability is feasible. We construct two statistically indistinguishable models, G_1 and G_2, dictating different estimation procedure S_1 and S_2 respectively; yet Q is not recoverable in G_1 by S_2 or in G_2 by S_1.

Example 12. *The graphs in Fig. 5 (a) and (c) are statistically indistinguishable; neither has testable implications. Let the target relation of interest be $Q = P(X)$. In Fig. 5 (a), Q may be estimated as $P(X) = \sum_y P(X|Y, R_x = 0)P(Y)$ since $X \perp\!\!\!\perp R_x|Y$ and in Fig. 5 (b), Q can be derived as $P(X) = P(X|R_x = 0)$ since $X \perp\!\!\!\perp R_x$.*

8 Conclusions

Researchers are typically uncertain about the model that accounts for loss of data while at the same time many procedures for recovering information from missing data rely on such models. These two facts motivate us to address the question of whether one can submit a given model to a test of compatibility with the data available, which of course is corrupted by missingness. In this paper we illuminated the boundary between testable and non-testable models with emphasis on models which are considered MNAR in the literature. We have provided syntactic rules for ensuring testability of a given conditional independence claim (CI) based on the type of variables (V_o, V_m, R) that appear in the CI. We further presented conditions for non-testability of a CI and discussed a general impediment to testability in missing data.

We have shown that there are singleton CI among substantive variables - not all of them fully observed - that can be tested, and we provided conditions on the dataset to falsify such CI claims. We refined the results of Potthoff et al. (2006) and showed that the class of models denominated as MAR^+ are testable whenever $|V_m| \geq 2$ and that the class of models denominated as MCAR are testable whenever $|V_o \cup V_m| \geq 2$. Additionally, we presented graphical and statistical conditions that confirm dependence between a variable and its missingness mechanism. Finally, we demonstrated sensitivity of missing data recovery procedures to hypothesized models and confirmed that this sensitivity is inevitable in datasets classified as MNAR.

[1] We leave open the unlikely possibility that there exists an estimation scheme, different from ours that could recover $Q = P(X)$ in both models. We propose this example as a litmus test for any such estimator.

References

Allison, P.D. Missing data series: Quantitative applications in the social sciences, 2002.

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

Garcia, F. M. Definition and diagnosis of problematic attrition in randomized controlled experiments. Working paper, April 2013. Available at SSRN: http://ssrn.com/abstract=2267120.

Little, Roderick JA. A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83 (404):1198–1202, 1988.

Mohan, K., Pearl, J., and Tian, J. Missing data as a causal inference problem. Technical Report R-410, UCLA, 2013. Forthcoming, Proceedings of NIPS, 2013. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r410.pdf.

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

Potthoff, R.F., Tudor, G.E., Pieper, K.S., and Hasselblad, V. Can one assess whether missing data are missing at random in medical studies? *Statistical methods in medical research*, 15(3):213–234, 2006.

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

Shpitser, Ilya and Pearl, Judea. Dormant independence. In *AAAI*, pp. 1081–1087, 2008.

Thoemmes, F. and Rose, N. Selection of auxiliary variables in missing data problems: Not all auxiliary variables are created equal. Technical Report Technical Report R-002, Cornell University, 2013.

Tian, Jin and Pearl, Judea. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pp. 519–527. Morgan Kaufmann Publishers Inc., 2002.