Graphical Models for Processing Missing Data

Karthika Mohan*
Department of Computer Science, University of California Los Angeles
and
Judea Pearl
Department of Computer Science, University of California Los Angeles

January 6, 2018

Abstract

This paper reviews recent advances in missing data research using graphical models to represent multivariate dependencies. We first examine the limitations of traditional frameworks from three different perspectives: *transparency, estimability and testability*. We then show how procedures based on graphical models can overcome these limitations and provide meaningful performance guarantees even when data are Missing Not At Random (MNAR). In particular, we identify conditions that guarantee consistent estimation in broad categories of missing data problems, and derive procedures for implementing this estimation. Finally we derive testable implications for missing data models in both MAR (Missing At Random) and MNAR categories.

Keywords: Missing data, Graphical Models, Testability, Recoverability, Non-Ignorable, Missing Not At Random (MNAR)

*The authors gratefully acknowledge support of this work by grants from NSF IIS-1302448, IIS-1527490 and IIS-1704932; ONR N00014-17-1-2091; DARPA W911NF-16-1-0579.

1 Introduction

Missing data present a challenge in many branches of empirical sciences. Sensors do not always work reliably, respondents do not fill out every question in the questionnaire, and medical patients are often unable to recall episodes, treatments or outcomes. The statistical literature on this problem is rich and abundant and has resulted in powerful software packages such as MICE in R, Stata, SAS and SPSS which offer various ways of handling missingness. Most practices are based on the seminal work of Rubin (1976) who formulated procedures and conditions under which the damage due to missingness can be reduced. This theory has also resulted in a number of performance guarantees when data obey certain statistical conditions. However, these conditions are rather strong, and extremely hard to ascertain in real world problems. Little and Rubin (2002)(page 22), summarize the state of the art by observing: "*essentially all the literature on multivariate incomplete data assumes that the data are Missing At Random (MAR)*". Indeed, popular estimation methods for missing data such as Maximum Likelihood based techniques (Dempster et al., 1977) and Multiple Imputation (Rubin, 1978) require MAR assumption to guarantee convergence to consistent estimates. Futhermore, it is almost impossible for a practicing statistician to decide whether the MAR condition holds in a given problem.

Recent years have witnessed a growing interest in analysing missing data using graphical models to encode assumptions about the reasons for missingness. This development is natural since graphical models provide efficient representation of conditional independencies implied by modeling assumptions Earlier papers in this development are Daniel et al. (2012) who provided sufficient criteria under which consistent estimates can be computed exclusively from complete cases (i.e. samples in which all variables are fully observed). Thoemmes and Rose (2013) (and later on Thoemmes and Mohan (2015)) developed tech-

niques that guide the selection of auxiliary variables to improve estimability from incomplete data. In machine learning, particularly while estimating parameters of Bayesian Networks, graphical models have long been used as a tool when dealing with missing data (Darwiche (2009); Koller and Friedman (2009)).

In this paper we review the contributions of graphical models to missing data research and emphasize three main aspects: (1) Transparency (2) Recoverability (consistent estimation) and (3) Testability

Transparency Consider a practicing statistician who has acquired a statistical package that handles missing data and would like to know whether the problem at hand meets the requirements of the software. As noted by Little and Rubin (2002) (see appendix 7.1) and many others such as Rhoads (2012) and Balakrishnan (2010), almost all available software packages implicitly assume that data fall under two categories: MCAR (Missing Completely At Random) or MAR (formally defined in section 2.2). Failing this assumption, there is no guarantee that estimates produced by current software will be less biased than those produced by the raw data or some filtered version thereoff. Consequently, it is essential for the user to decide if the type of missingness present in the data is compatible with the requirements of MCAR or MAR.

Prior to the advent of graphical models, no tool was available to assist in this decision, since the independence conditions that define MCAR or MAR are neither visible in the data, nor in a mathematical model that a researcher can consult to verify those conditions. We will show how graphical models enable an efficient and transparent classification of the missingness mechanism. In particular, the question of whether the data fall into the MCAR or MAR categories can be answered by mere inspection of the graph structure. In addition, we will show how graphs facilitate a more refined, query-specific taxonomy of

missingness in MNAR (Missing Not At Random) problems.

The transparency associated with graphical models stems from three reasons. First, graphs excel in encoding and detecting conditional independence relations, far exceeding the capacity of human intuition. Second, all assumptions are encoded causally, mirroring the way researchers store qualitative scientific knowledge; direct judgments of conditional independencies are not required, since these can be read off the structure of the graph. Finally, the ultimate aim of all assumptions is to encode "the reasons for missingness" which is a causal, not a statistical concept. Thus, even when our target parameter is purely statistical, say a regression coefficient, causal modeling is still needed for encoding the "process that causes missing data" (Rubin (1976)).

Recoverability (Consistent Estimation) Recoverability (to be defined formally in Section 3) refers to the task of determining, from an assumed model, whether any method exists that produces a consistent estimate of a desired parameter and, if so, how. If the answer is negative, then an inconsistent estimate should be expected even with large samples, and no algorithm, however smart, can yield a consistent estimate. On the other hand, if the answer is affirmative then there exists a procedure that can exploit the features of the problem and produces consistent estimates. If the problem is MAR or MCAR, standard missing data software can be used to obtain consistent estimates. But if a recoverable problem is MNAR, the user would do well to discard standard software and resort to an estimator derived by graphical analysis. In Section 3 of this paper we present several methods of deriving consistent estimators for both statistical and causal parameters.

The general question of recoverability, to the best of our knowledge, has not received due attention in the literature. The notion that some parameters cannot be estimated by any method whatsoever while others can, still resides in an unchartered territory. We will

show in Section 3 that most MNAR problems exhibit this dichotomy. That is, problems for which it is impossible to properly impute all missing values in the data, would still permit the consistent estimation of some parameters of interest. More importantly, the estimable parameters can often be identified directly from the structure of the graph.

Testability Testability asks whether it is possible to tell if any of the model's assumptions is incompatible with the available data (corrupted by missingness). Such compatibility tests under missingness are hard to come by and the few tests reported in the literature are mostly limited to MCAR (Little, 1988). "*Worse still, there is no empirical way to discriminate one nonignorable model from another (or from the ignorable model).*" (Allison (2003)). In section 4 we will show that remarkably, discrimination is feasible; MAR problems do have a simple set of testable implications and MNAR problems can often be tested depending on their graph structures.

In summary, although mainstream statistical analysis of missing data problems has made impressive progress in the past few decades, it left several problem areas relatively unexplored, especially those touching on transparency, estimability and testability. This paper casts missing data problems in the language of causal graphs and shows how this representation facilitates solutions to several pending problems. In particular, we show how the MCAR, MAR, MNAR taxonomy becomes transparent in the graphical language, how the estimability of a needed parameter can be determined from the graph structure, what estimators would guarantee consistent estimates, and what modeling assumptions lend themselves to empirical scrutiny.

2 Graphical Models for Missing Data: Missingness Graphs (m-graphs)

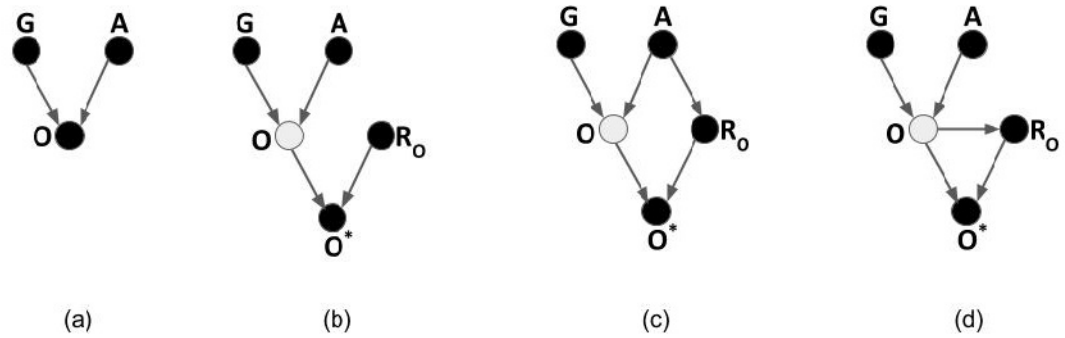

Figure 1: (a)causal graph under no missingness (b), (c) & (d) m-graphs modeling distinct missingness processes.

The following example, inspired by Little and Rubin (2002) (example-1.6, page 8), describes how graphical models can be used to explicitly model the missingness process and encode the underlying causal and statistical assumptions. Consider a study conducted in a school that measured three (discrete) variables: Age (A), Gender (G) and Obesity (O).

No Missingness If all three variables are completely recorded, then there is no missingness. The causal graph[1] depicting the interrelations between variables is shown in Figure 1 (a). Nodes correspond to variables and edges indicate the existence of a causal relationship between pairs of nodes they connect. The value of a child node is a (stochastic) function

[1]For a gentle introduction to causal graphical models see Elwert (2013); Lauritzen (2001), sections 1.2 and 11.1.2 in Pearl (2009).

of the values of its parent nodes. i.e. Obesity is a (stochastic) function of Age and Gender. The absence of an edge between Age and Gender indicates that A and G are independent, denoted by $A \perp\!\!\!\perp G$.

Table 1: Missing dataset in which Age and Gender are fully observed and Obesity is partially observed.

#	Age	Gender	Obesity*	R_O
1	16	F	Obese	0
2	15	F	m	1
3	15	M	m	1
4	14	F	Not Obese	0
5	13	M	Not Obese	0
6	15	M	Obese	0
7	14	F	Obese	0

Representing Missingness Assume that Age and Gender are are fully observed since they can be obtained from school records. Obesity however is corrupted by missing values due to some students not revealing their weight. When the value of O is missing we get an empty measurement which we designate by m. Table 1 exemplifies a missing dataset. The missingness process can be modelled using a proxy variable Obesity*(O^*) whose values are determined by Obesity and its missingness mechanism R_O.

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

R_O governs the masking and unmasking of Obesity. When $R_O = 1$ the value of obesity is concealed i.e. O^* assumes the values m as shown in samples 2 and 3 in table 1. When

$R_O = 0$, the true value of obesity is revealed i.e. O^* assumes the underlying value of Obesity as shown in samples 1, 4, 5, 6 and 7 in table 1.

Missingness can be caused by random processes or can depend on other variables in the dataset. An example of random missingness is students *forgetting* to return their questionnaires. This is depicted in figure 1 (b) by the absence of parent nodes for R_O. Teenagers rebelling and not reporting their weight is an example of missingness caused by a fully observed variable. This is depicted in figure 1 (c) by an edge between A and R_O. Partially observed variables can be causes of missingness as well. For instance consider obese students who are embarrassed of their obesity and hence reluctant to reveal their weight. This is depicted in figure 1 (d) by an edge between O and R_O indicating the O is the cause of its own missingness.

The following subsection formally introduces missingness graphs (m-graphs) as discussed in Mohan et al. (2013).

2.1 Missingness Graphs: Notations and Terminology

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 $X <\!\!-\!\!> Y$ denoting the existence of a U variable that is a parent of both X and Y.

We call this graphical representation a **Missingness Graph** (or m-graph). Figure 1 exemplifies three m-graphs in which $V_o = \{A, G\}$, $V_m = \{O\}$, $V^* = \{O^*\}$, $U = \emptyset$ and $R = \{R_O\}$. Proxy variables may not always be explicitly shown in m-graphs in order to keep the figures simple and clear. The missing data distribution, $P(V^*, V_o, R)$ is referred to as the *manifest distribution* and the distribution that we would have obtained had there been no missingness, $P(V_o, V_m, R)$ is called as the *underlying distribution*. Conditional Independencies are read off the graph using the d-separation[2] criterion (Pearl, 2009). For example, Figure 1 (c) depicts the independence $R_O \perp\!\!\!\perp O|A$ but not $R_O \perp\!\!\!\perp G|O$.

2.2 Classification of Missing Data Problems based on Missingness Mechanism

Rubin (1976) classified missing data into three categories: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR) based on

[2]For a quick introduction to d-separation see, http://www.dagitty.net/learn/dsep/index.html

the statistical dependencies between the missingness mechanisms (R variables) and the variables in the dataset (V_m, V_o). We capture the essence of this categorization in graphical terms below.

1. Data are MCAR if $V_m \cup V_o \cup U \perp\!\!\!\perp R$ holds in the m-graph. In words, missingness occurs at random and is entirely independent of both the observed and the partially observed variables. This condition can be easily identified in an m-graph by the absence of edges between the R variables and variables in $V_o \cup V_m$.

2. Data are MAR if $V_m \cup U \perp\!\!\!\perp R|V_o$ holds in the m-graph. In words, conditional on the fully observed variables V_o, missingness occurs at random. In graphical terms, MAR holds if (i) no edges exist between an R variable and any partially observed variable and (ii) no bidirected edge exists between an R variable and a fully observed variable. MCAR implies MAR, ergo all estimation techniques applicable to MAR can be safely applied to MCAR.

3. Data that are not MAR or MCAR fall under the MNAR category.

m-graphs in figure 1 (b), (c) and (d) are typical examples of MCAR, MAR and MNAR categories, respectively. Notice the ease with which the three categories can be identified. Once the user lays out the interrelationships between the variables in the problem, the classification is purely mechanical.

2.2.1 Missing At Random: A Brief Discussion

The original classification used in Rubin (1976) is very similar to the one defined in the preceding paragraphs; it is expressed however in terms of event-level conditional independencies as opposed to variable-level independencies. We will clarify the distinction between

the former (which we call *Rubin-MAR*) and the latter (referred to as MAR) with an example. Consider a dataset with three variables such that two variables, A and B are partially observed and the third one C is fully observed. For data to be MAR, we require $(A, B) \perp\!\!\!\perp (R_A, R_B)|C$ to hold. On the other hand, Rubin-MAR requires that, " *missingness depends only on the components* Y_{obs} *of* Y *that are observed and not on the components that are missing*"(Little and Rubin (2002)), where Y denotes the dataset. We exemplify Rubin-MAR in table 2. The primary difference between the two definitions is that MAR is a succinct statement comprising of a single conditional independence: $V_m \perp\!\!\!\perp R|V_o$, where as Rubin-MAR is a set of distinct conditional independencies of the form: $Y_{mis} \perp\!\!\!\perp R|Y_{obs}$, one for each subpopulation as described by the pattern of missingness[3]. Observe that both definitions coincide when $|V_m| = 1$.

Over the years the classification proposed in Rubin (1976) has been criticized both for its nomenclature and its opacity. Several authors noted that **MAR is a misnomer** (Scheffer (2002); Peters and Enders (2002); Meyers et al. (2006); Graham (2009)). What is currently defined as MCAR should have been called Missing At Random and as pointed out by Grace-Martin[4], what is currently defined as Missing At Random should have been called Missing Conditionally At Random.

However, the **opacity of the assumptions** underlying Rubin's MAR presents a more serious problem. The number of conditional independence relations that need be verified is exponential in the number of the partially observed variables. This is shown in Table 2, which displays the conditional independencies claimed by the *Rubin-MAR* condition: $Y_{mis} \perp\!\!\!\perp R|Y_{obs}$. Clearly, a researcher would find it cognitively taxing, if not impossible to

[3]Each instantiation of R variables corresponds to a pattern of missingness. In the case of the ongoing example with $V_m = \{A, B\}$ and $V_o = \{C\}$, there are 4 patterns of missingness: $(R_A = 0, R_B = 0), (R_A = 0, R_B = 1), (R_A = 1, R_B = 0)$ and $(R_A = 1, R_B = 1)$ (De Leeuw et al. (2008); Chen and Wilson (2015)).

[4]http://www.theanalysisfactor.com/mar-and-mcar-missing-data/

Table 2: Rubin-MAR detailed for the dataset in which A and B are partially observed variables and C is a fully observed variable.

Missing Components Y_{mis}	Observed Components Y_{obs}	Rubin-MAR Conditions $Y_{mis} \perp\!\!\!\perp R \mid Y_{obs}$	Description of Samples
A	B, C	$A \perp\!\!\!\perp R \mid B, C$	Samples in which A is missing and B is observed
B	A, C	$B \perp\!\!\!\perp R \mid A, C$	Samples in which B is missing and A is observed
A, B	C	$(A, B) \perp\!\!\!\perp R \mid C$	Samples in which both A and B are missing
–	A, B, C	–	Samples in which all variables are observed.

even decide if any of these assumptions is reasonable. This, together with the fact that Rubin-MAR is untestable (Allison (2002)) motivates the variable-based taxonomy presented above.

Nonetheless, Rubin-MAR has an interesting theoretical property: It is the weakest simple condition under which the process that causes missingness can be ignored while still making correct inferences about the data (Rubin, 1976). It was probably this theoretical result that changed missing data practices in the 1970's. The popular practice prior to 1976 was to assume that missingness was caused totally at random (Gleason and Staelin (1975); Haitovsky (1968)). With Rubin's identification of the MAR condition as sufficient for drawing correct inferences, MAR became the main focus of attention in the statistical literature.

Estimation procedures such as Multiple Imputation and Maximum Likelihood were developed and implemented with MAR assumptions in mind, and popular textbooks were authored exclusively on MAR and its simplified versions (Graham, 2012). These developments have engendered a culture with a tendency to blindly assume MAR, with the consequence that the more commonly occurring MNAR class of problems remains relatively unexplored (Resseguier et al., 2011; Adams, 2007; Osborne, 2012, 2014; Sverdlov, 2015; van Stein and Kowalczyk, 2016).

To overcome these limitations Rubin (1976) recommended that researchers explicitly model the missingness process:

> **The inescapable conclusion seems to be that when dealing with real data, the practising statistician should explicitly consider the process that causes missing data far more often than he does. However, to do so, he needs models for this process and these have not received much attention in the statistical literature.**

Figure 2: Quote from Rubin (1976)

This recommendation invites in fact the graphical tools described in this paper, for they encourage investigators to model the details of the missingness process rather than blindly assume MAR. These tools have further enabled researchers to extend the analysis of estimation to the vast class of MNAR problems.

In the next section we discuss how graphical models accomplish these tasks.

3 Recoverability

Recoverability addresses the basic question of whether a quantity/parameter of interest can be estimated from incomplete data *as if* no missingness took place, that is, the desired quantity can be estimated consistently from the available (incomplete) data. This amounts

to expressing the target quantity Q in terms of the manifest distribution $P(V^*, V_O, R)$. Typical target quantities that shall be considered are conditional/joint distributions and conditional causal effects.

Definition 1 (Recoverability of target quantity Q) *Let A denote the set of assumptions about the data generation process and let Q be any functional of the underlying distribution $P(V_m, V_O, R)$. Q is recoverable if there exists an algorithm that computes a consistent estimate of Q for all strictly positive manifest distributions $P(V^*, V_o, R)$ that may be generated under A.*

Since we encode all assumptions in the structure of the m-graph G, recoverability becomes a property of the pair $\{Q, G\}$, and not of the data. We restrict the definition above to strictly positive manifest distributions, $P(V^*, V_o, R)$ except for instances of zero probabilities as specified in equation 1. The reason for this restriction can be understood as the need for observing some unmasked cases for all combinations of variables, otherwise, masked cases can be arbitrary. We note however that recoverability is sometimes feasible even when strict positivity does not hold.

We now demonstrate how a joint distribution is recovered given MAR data.

Example 1 *Consider the problem of recovering the joint distribution given the m-graph in Fig. 1 (c) and dataset in table 3. Let it be the case that 15-18 year olds were reluctant to reveal their weight, thereby making O a partially observed variable i.e. $V_m = \{O\}$ and $V_o = \{G, A\}$. This is a typical case of MAR missingness, since the cause of missingness is the fully observed variable: Age. The following three steps detail the recovery procedure.*

1. Factorization: The joint distribution may be factorized as:

$$P(G, O, A) = P(G, O|A)P(A)$$

2. Transformation into observables: G implies the conditional independence $(G, O) \perp\!\!\!\perp R_O | A$ *since A d-separates* (G, O) *from* R_O*. Thus,*

$$P(G, O, A) = P(G, O | A, R_O = 0) P(A)$$

3. Conversion of partially observed variables into proxy variables: $R_O = 0$ *implies* $O^* = O$ *(by eq 1). Therefore,*

$$P(G, O, A) = P(G, O^* | A, R_O = 0) P(A) \tag{2}$$

The RHS of Eq. (2) is expressed in terms of variables in the manifest distribution. Therefore, $P(G, A, O)$ *can be consistently estimated (i.e. recovered) from the available data. The recovered joint distribution is shown in table 4.*

Note that samples in which obesity is missing are not discarded but are used instead to update the weights $p_1, ..p_{12}$ of the cells in which obesity is has a definite value. This can be seen by the presence of probabilities $p_{13}, ...p_{18}$ in table 4 and the fact that samples with missing values have been utilized to estimate prior probability $P(A)$ in equation 2. Note also that the joint distribution permits an alternative decomposition:

$$\begin{aligned} P(O, A, G) &= P(O | A, G) P(A, G) \\ &= P(O^* | A, G, R_O = 0) P(A, G) \end{aligned}$$

The equation above licenses a different estimation procedure whereby $P(A, G)$ is estimated from all samples, including those in which obesity is missing, and only the estimation of $P(O^* | A, G, R_O = 0)$ is restricted to the complete samples. The efficiency of various decompositions are analysed in Van den Broeck et al. (2015); Mohan et al. (2014).

Finally we observe that for the MCAR m-graph in figure 1 (b), a wider spectrum of

Table 3: Manifest Distribution $P(G, A, O^*, R_O)$ where Gender (G) and Age (A) are fully observed, Obesity O is corrupted by missing values and Obesity's proxy (O^*) is observed in its place. Age is partitioned into three groups: $[10-13), [13-15), [15-18)$. Gender and Obesity are binary variables and can take values Male (M) and Female (F), and Yes (Y) and No (N), respectively. The probabilities $p_1, p_2, ..p_{18}$ stand for the (asymptotic) frequencies of the samples falling in the 18 cells (G, A, O^*, R_O).

G	A	O^*	R_O	$P(G, A, O^*, R_O)$
M	$10-13$	Y	0	p_1
M	$13-15$	Y	0	p_2
M	$15-18$	Y	0	p_3
M	$10-13$	N	0	p_4
M	$13-15$	N	0	p_5
M	$15-18$	N	0	p_6
F	$10-13$	Y	0	p_7
F	$13-15$	Y	0	p_8
F	$15-18$	Y	0	p_9
F	$10-13$	N	0	p_{10}
F	$13-15$	N	0	p_{11}
F	$15-18$	N	0	p_{12}
M	$10-13$	m	1	p_{13}
M	$13-15$	m	1	p_{14}
M	$15-18$	m	1	p_{15}
F	$10-13$	m	1	p_{16}
F	$13-15$	m	1	p_{17}
F	$15-18$	m	1	p_{18}

decompositions is applicable, including:

$$\begin{aligned} P(O, A, G) &= P(O, A, G | R_O = 0) \\ &= P(O^*, A, G | R_O = 0) \end{aligned}$$

The equation above licenses the estimation of the joint distribution using only those samples in which obesity is observed. This estimation procedure, called listwise deletion or complete-case analysis (Little and Rubin, 2002), would result of course in wastage of data and lower quality of estimate, especially when the number of samples corrupted by miss-

Table 4: Recovered joint distribution corresponding to dataset in table 3 and m-graph in figure 1(c)

G	A	O	$P(G,A,O)$
M	$10-13$	Y	$\frac{p_1*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
M	$13-15$	Y	$\frac{p_2*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
M	$15-18$	Y	$\frac{p_3*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$
M	$10-13$	N	$\frac{p_4*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
M	$13-15$	N	$\frac{p_5*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
M	$15-18$	N	$\frac{p_6*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$

G	A	O	$P(G,A,O)$
F	$10-13$	Y	$\frac{p_7*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
F	$13-15$	Y	$\frac{p_8*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
F	$15-18$	Y	$\frac{p_9*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$
F	$10-13$	N	$\frac{p_{10}*(p_1+p_4+p_7+p_{10}+p_{13}+p_{16})}{p_1+p_4+p_7+p_{10}}$
F	$13-15$	N	$\frac{p_{11}*(p_2+p_5+p_8+p_{11}+p_{14}+p_{17})}{p_2+p_5+p_8+p_{11}}$
F	$15-18$	N	$\frac{p_{12}*(p_3+p_6+p_9+p_{12}+p_{15}+p_{18})}{p_3+p_6+p_9+p_{12}}$

ingness is high. Considerations of estimation efficiency should therefore be applied once we explicate the spectrum of options licensed by the m-graph.

A completely different behavior will be encountered in the model of 1 (d) which, as we have noted, belong to the MNAR category. Here, the arrow $O \rightarrow R_O$ would prevent us from executing step 2 of the estimation procedure, that is, transforming $P(G,O,A)$ into an expression involving solely observed variables. We can in fact show that in this example the joint distribution is nonrecoverable. That is, regardless of how large the sample or how clever the imputation, no algorithm exists that produces consistent estimate of P(O,A,G).

The possibility of encountering non-recoverability is rarely discussed in mainstream missing data literature, partly because the MAR assumption is taken for granted, and partly because researchers presume that the maximum likelihood method can deliver a consistent estimate of any desired parameter. Unfortunately, the maximum likelihood loses its standard consistency guarantee in missing-data problems, because the desired parameters are defined in terms of the underlying distribution, and are not uniquely defined in terms of the manifest distribution. This non uniqueness is redeemable in MAR problem

but comes to haunts us in MNAR problems where the likelihood function may no longer have a unique maximum (see Koller and Friedman (2009)).

Remark: Observe that equation 2 yields an **estimand** for the query, $P(G, O, A)$, as opposed to an *estimator*. An estimand is a functional of the manifest distribution, $P(V^*, R, V_o)$, whereas an estimator is a rule detailing how to calculate the estimate from measurements in the sample. Our estimands naturally give rise to a closed form estimator, for instance, the estimator corresponding to the estimand in equation 2 is $\frac{\#(G=g,O^*=o,A=a,R_O=0)}{\#(A=a,R_O=0)} \frac{\#(A=a)}{N}$, where N is the total number of samples collected and $\#(X_1 = x_1, X_2 = x_2, ...X_j = x_j)$ is the frequency of the event $x_1, x_2, ...x_j$. Algorithms inspired by such closed form estimation techniques were shown in Van den Broeck et al. (2015), to outperform conventional methods such as EM in terms of speed and accuracy of estimates.

In the following subsection we define the notion of Ordered factorization which leads to a criterion for sequentially recovering conditional probability distributions (Mohan et al. (2013); Mohan and Pearl (2014a)).

3.1 Recovery by Sequential Factorization

Definition 2 (Ordered factorization of $P(Y|Z)$) *Let $Y_1 < Y_2 < \ldots < Y_k$ be an ordered set of all variables in Y, $1 \leq i \leq |Y| = k$ and $X_i \subseteq \{Y_{i+1}, \ldots, Y_n\} \cup Z$. Ordered factorization of $P(Y|Z)$ is the product of conditional probabilities i.e. $P(Y|Z) = \prod_i P(Y_i|X_i)$, such that X_i is a minimal set for which $Y_i \perp\!\!\!\perp (\{Y_{i+1}, \ldots, Y_n\} \setminus X_i) | X_i$ holds.*

The following theorem presents a sufficient condition for recovering conditional distributions of the form $P(Y|X)$ where $\{Y, X\} \subseteq V_m \cup V_o$.

Theorem 1 *Given an m-graph G and a manifest distribution $P(V^*, V_o, R)$, a target quantity Q is recoverable if Q can be decomposed into an ordered factorization, or a sum of such*

factorizations, such that every factor $Q_i = P(Y_i|X_i)$ *satisfies* $Y_i \perp\!\!\!\perp (R_{y_i}, R_{x_i})|X_i$. *Then, each* Q_i *may be recovered as* $P(Y_i^*|X_i^*, R_{Y_i} = 0, R_{X_i} = 0)$.

An ordered factorization that satisfies theorem 1 is called as an *admissible factorization.*

Example 2 *Consider the problem of recovering* $P(X, Y)$ *given* G, *the m-graph in figure 3 (a).* G *depicts an MNAR problem since missingness in* Y *is caused by the partially observed variable* X. *The factorization* $P(Y|X)P(X)$ *is admissible since both* $Y \perp\!\!\!\perp R_x, R_y|X$ *and* $X \perp\!\!\!\perp R_x$ *hold in* G. $P(X, Y)$ *can thus be recovered using theorem 1 as* $P(Y^*|X^*, R_x = 0, R_y = 0)P(X^*|R_x = 0)$. *Here, complete cases are used to estimate* $P(Y|X)$ *and all samples including those in which* Y *is missing are used to estimate* $P(X)$. *Note that the decomposition* $P(X|Y)P(Y)$ *is not admissible.*

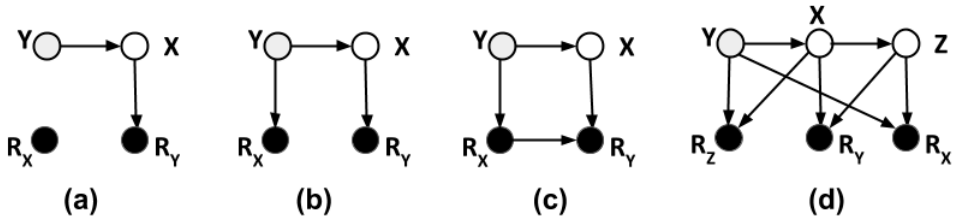

Figure 3: m-graphs from which joint and/or conditional distributions can be recovered using various factorizations.

3.2 R Factorization

Example 3 *Consider the problem of recovering* $Q = P(X, Y)$ *from the m-graph of Figure 3(b). Interestingly, no ordered factorization over variables* X *and* Y *would satisfy the conditions of Theorem 1. To witness we write* $P(X, Y) = P(Y|X)P(X)$ *and note that the*

graph does not permit us to augment any of the two terms with the necessary R_x or R_y terms; X is independent of R_x only if we condition on Y, which is partially observed, and Y is independent of R_y only if we condition on X which is also partially observed. This deadlock can be disentangled however using a non-conventional decomposition:

$$Q = P(X,Y) = P(X,Y)\frac{P(R_x=0,R_y=0|X,Y)}{P(R_x=0,R_y=0|X,Y)}$$
$$= \frac{P(R_x=0,R_y=0)P(X,Y|R_x=0,R_y=0)}{P(R_x=0|Y,R_y=0)P(R_y=0|X,R_x=0)}$$

where the denominator was obtained using the independencies $R_x \perp\!\!\!\perp (X,R_y)|Y$ and $R_y \perp\!\!\!\perp (Y,R_x)|X$ shown in the graph. The final expression below,

$$P(X,Y) = \frac{P(R_x=0,R_y=0)P(X^*,Y^*|R_x=0,R_y=0)}{P(R_x=0|Y^*,R_y=0)P(R_y=0|X^*,R_x=0)} \text{ (Using equation 1)} \quad (3)$$

which is in terms of variables in the manifest distribution, renders $P(X,Y)$ recoverable. This example again shows that recovery is feasible even when data are MNAR.

The following theorem (Mohan et al. (2013); Mohan and Pearl (2014a)) formalizes the recoverability scheme exemplified above.

Theorem 2 (Recoverability of the Joint $P(V)$) *Given a m-graph G with no edges between R variables the necessary and sufficient condition for recovering the joint distribution $P(V)$ is the absence of any variable $X \in V_m$ such that:*

1. X and R_x are neighbors

2. X and R_x are connected by a path in which all intermediate nodes are colliders[5] *and elements of $V_m \cup V_o$. When recoverable, $P(V)$ is given by*

$$P(v) = \frac{P(R=0,v)}{\prod_i P(R_i=0|Mb^o_{r_i}, Mb^m_{r_i}, R_{Mb^m_{r_i}}=0)}, \quad (4)$$

[5]A variable is a collider on the path if the path enters and leaves the variable via arrowheads (a term suggested by the collision of causal forces at the variable) (Greenland and Pearl, 2011).

where $Mb^o_{r_i} \subseteq V_o$ *and* $Mb^m_{r_i} \subseteq V_m$ *are the markov blanket of* R_i*.*

The preceding theorem can be applied to immediately yield an estimand for joint distribution. For instance, given the m-graphs in figure 3 (d), joint distribution can be recovered in one step yielding:

$$P(X,Y,Z) = \frac{P(X,Y,Z,R_x{=}0,R_y{=}0,R_z{=}0)}{P(R_x{=}0|Y,R_y{=}0,Z,R_z{=}0)P(R_y{=}0|X,R_x{=}0,Z,R_z{=}0)P(R_z{=}0|Y,R_y{=}0,X,R_x{=}0)}$$

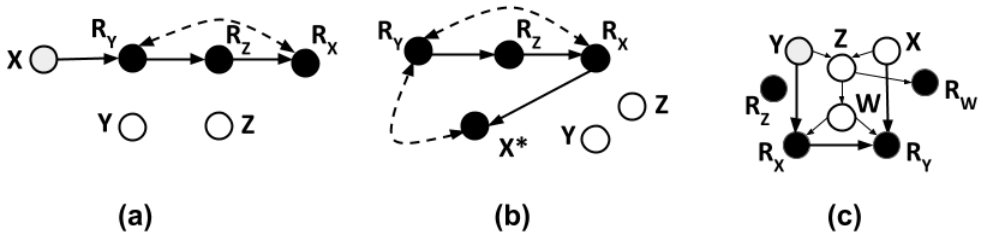

Figure 4: (a) & (c) m-graphs from which conditional distributions can be recovered aided by intervention, (b) latent structure (Pearl (2009), chapter 2) corresponding to m-graph in (a) when X is treated as a latent variable.

3.3 Constraint Based Recoverability

The recoverability procedures presented thus far relied entirely on conditional independencies that are read off the m-graph using d-separation criterion. Interestingly, recoverability can sometimes be accomplished by graphical patterns other than conditional independencies. These patterns represent distributional constraints which can be detected using mutilated versions of the m-graph. We describe below an example of constraint based recovery.

Example 4 *Let G be the m-graph in figure 4 (a) and let the query of interest be $P(X)$. The absence of a set that d-separates X from R_x, makes it impossible to apply any of the techniques discussed previously. While it may be tempting to conclude that $P(X)$ is not recoverable, we prove otherwise by using the fact that $X \perp\!\!\!\perp R_x$ holds in the ratio distribution $\frac{P(X,R_y,R_z,R_x)}{P(R_z|R_y)}$. Such ratios are called interventional distributions and the resulting constraints are called Verma Constraints (Verma and Pearl (1991); Tian and Pearl (2002)). The proof presented below employs the rules of do-calculus[6], to extract these constraints.*

$$\begin{aligned}
P(X) &= P(X|do(R_z = 0)) \text{ (Rule-3 of do-calculus)} \\
&= P(X|do(R_z = 0), R_x = 0) \text{ (Rule-1 of do-calculus)} \\
&= P(X^*|do(R_z = 0), R_x = 0) \text{ (using equation 1)} \\
&= \sum_{R_Y} P(X^*, R_Y|do(R_z = 0), R_x = 0) \qquad (5)
\end{aligned}$$

Note that the query of interest is now a function of X^* and not X. Therefore the problem now amounts to identifying a conditional interventional distribution using the m-graph in figure 4(b). A complete analysis of such problems is available in Shpitser and Pearl (2006) which identifies the causal effect in eq 5 as:

$$P(X) = \sum_{R_Y} P(X^*|R_Y, R_x = 0, R_z = 0) \frac{P(R_x = 0|R_y, R_z = 0)P(R_y)}{\sum_{R_Y} P(R_x = 0|R_y, R_z = 0)P(R_y)} \qquad (6)$$

In addition to $P(X)$, this graph also allows recovery of joint distribution as shown below.

$$P(X, Y, Z) = P(X)P(Y)P(Z)$$

$$P(X, Y, Z) = \Big(\sum_{R_Y} P(X^*|R_Y, R_x = 0, R_z = 0) \frac{P(R_x=0|R_y,R_z=0)P(R_y)}{\sum_{R_Y} P(R_x=0|R_y,R_z=0)P(R_y)} \Big) P(Y^* = Y|R_y = 0)P(Z^*|R_z = 0)$$

[6]For an introduction to do-calculus see, Pearl and Bareinboim (2014), section 2.5 and Koller and Friedman (2009)

The decomposition in the first line uses $(X, Y) \perp\!\!\!\perp Z$ *and* $X \perp\!\!\!\perp Y$*. Recoverability of* $P(X)$ *in the second line follows from equation 6. Theorem 1 can be applied to recover,* $P(Y)$ *and* $P(Z)$*, since* $Y \perp\!\!\!\perp R_Y$ *and* $Z \perp\!\!\!\perp R_Z$*.*

Remark 1 *In the preceding example we were able to recover a joint distribution despite the fact that the distribution* $P(X, R_Y, R_x)$ *is void of independencies. The ability to exploit such cases further underscores the need for graph based analysis.*

A more complex example detailing recoverability of joint distribution from the m-graph in figure 4 (c) is presented in Appendix 7.2. A general algorithm for recovering joint distributions in models with edges between R variables is presented in Shpitser et al. (2015).

Thus far, we dealt with recovering statistical properties and parameters. Similar results for recovering causal effects are available in Mohan and Pearl (2014a) and Shpitser (2016).

3.4 Overcoming Impediments to Recoverability

This section focuses on MNAR problems that are not recoverable. One such problem is elucidated in the following example.

Example 5 *Consider a missing dataset comprising of a single variable, Income (*I*), obtained from a population in which the very rich and the very poor were reluctant to reveal their income. The underlying process can be described as a variable causing its own missingness. The m-graph depicting this process is* $I \rightarrow R_I$*. Obviously, under these circumstances the true distribution over income,* $P(I)$*, cannot be computed error-free even if we were given infinitely many samples.*

The following theorem identifies graphical conditions that forbid recoverability of conditional probability distributions (Mohan and Pearl (2014a)).

Theorem 3 *Let $X \cup Y \subseteq V_m \cup V_o$ and $|X| = 1$. $P(X|Y)$ is not recoverable if either, X and R_X are neighbors or there exists a path from X to R_x such that all intermediate nodes are colliders and elements of Y.*

Quite surprisingly, it is sometimes possible to recover joint distributions given m-graphs with graphical structures stated in theorem 3 by jointly harnessing features of the data and m-graph. We exemplify such recovery with an example.

Example 6 *Consider the problem of recovering $P(Y, I)$ given the m-graph $G : Y \rightarrow I \rightarrow R_I$, where Y is a binary variable that denotes whether candidate has sufficient years of relevant work experience and I indicates income. I is also a binary variable and takes values high and low. $P(Y)$ is implicitly recoverable since Y is fully observed. $P(Y|I)$ may be recovered as shown below:*

$$\begin{aligned} P(Y|I) &= P(Y|I, r'_I) \text{ (using } Y \perp\!\!\!\perp R_I|I) \\ &= P(Y^* = Y|I^* = I, , r'_I) \text{ (using equation 1)} \end{aligned}$$

Expressing $P(Y) = \sum_y P(Y|I)P(I)$ in matrix form, we get:

$$\begin{pmatrix} P(y') \\ P(y) \end{pmatrix} = \begin{pmatrix} P(y'|i') & P(y'|i) \\ P(y|i') & P(y|i) \end{pmatrix} \begin{pmatrix} P(i') \\ P(i) \end{pmatrix}$$

Assuming that the square matrix on R.H.S is invertible, $P(I)$ can be estimated as:

$$\begin{pmatrix} P(y'|i') & P(y'|i) \\ P(y|i') & P(y|i) \end{pmatrix}^{-1} \begin{pmatrix} P(y') \\ P(y) \end{pmatrix}$$

Having recovered $P(I)$, the query $P(I, Y)$ may be recovered as $P(Y|I)P(I)$.

Theorem 4 formalizes the recovery procedure exemplified above. Let $M_{WY} = P(W|Y)$ denote a square matrix with non-negative entries such that entries in each column sum to

one. For example, for binary variables W and Y,
$M_{WY} = \begin{pmatrix} P(w=0|y=0) & P(w=0|y=1) \\ P(w=1|y=0) & P(w=1|y=1) \end{pmatrix}$. For any set W, $|W|$ denotes the sum of the cardinality of each element in W. For example if $W = \{W_1, W_2\}$, W_1 is binary and W_2 is ternary, then $|W| = 5$.

Theorem 4 *Let G be an m-graph, $V = V_o \cup V_m$ and W be a set of variables not in G such that $P(W)$ and $P(W|V)$ are recoverable and $|W| = |V|$. Given G and W, $P(V)$ and hence $P(WV)$, are recoverable if M_{WY} is invertible.*

Clearly, theorem 4 can be applied to any problem that satisfies its conditions; it makes no restrictions on the input graph structure. General procedures for handling non-recoverable cases is discussed in Mohan and Pearl (2016).

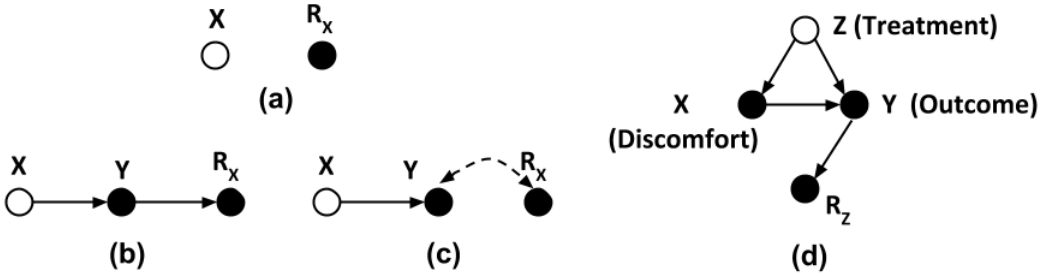

Figure 5: (a) m-graph with an untestable claim: $Z \perp\!\!\!\perp R_z | X, Y$, (b) & (c) Two statistically indistinguishable models, (d) m-graph depicting MCAR.

4 Testability Under Missingness

In this section we seek ways to detect mis-specifications of the missingness model. While discussing testability, one must note a phenomenon that recurs in missing data analysis: *Not all that looks testable is testable.* Specifically, although every d-separation in the graph implies conditional independence in the recovered distribution, some of those independencies are imposed by construction, in order to satisfy the model's claims, and these do not provide means of refuting the model. We exemplify this peculiarity below.

Example 7 *Consider the m-graph in figure 5(a). It is evident that the problem is MCAR (definition in section 4.2). Hence $P(X, R_x)$ is recoverable. The only conditional independence embodied in the graph is $X \perp\!\!\!\perp R_x$. At first glance it might seem as if $X \perp\!\!\!\perp R_x$ is testable since we can go to the recovered distribution and check whether it satisfies this conditional independence. However, $X \perp\!\!\!\perp R_x$ will always be satisfied in the recovered distribution, because it was recovered so as to satisfy $X \perp\!\!\!\perp R_x$. This can be shown explicitly as follows:*

$$\begin{aligned} P(X, R_x) &= P(X|R_x)P(R_x) \\ &= P(X|R_x = 0)P(R_x) \text{ (Using } X \perp\!\!\!\perp R_x) \\ &= P(X^*|R_x = 0)P(R_x) \text{(Using Equation 1)} \end{aligned}$$

Likewise,

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

Therefore, the claim, $X \perp\!\!\!\perp R_x$, cannot be refuted by any recovered distribution, regardless of what process actually generated the data. In other words, any data whatsoever with X partially observed can be made compatible with the model postulated.

The following theorem characterizes a more general class of untestable claims.

Theorem 5 (Mohan and Pearl (2014b)) *Let $\{Z, X\} \subseteq V_m$ and $W \subseteq V_o$. Conditional independencies of the form $X \perp\!\!\!\perp R_x | Z, W, R_z$ are untestable.*

The preceding example demonstrates this theorem as a special case, with $Z = W = R_z = \emptyset$. The next section provides criteria for testable claims.

4.1 Graphical Criteria for Testability

The criterion for detecting testable implications reads as follows: *A d-separation condition displayed in the graph is testable if the R variables associated with all the partially observed variables in it are either present in the separator set or can be added to the separator without spoiling the separation.* The following theorem formally states this criterion using three syntactic rules (Mohan and Pearl (2014b)).

Theorem 6 *A sufficient condition for an m-graph to be testable is that it encodes one of the following types of independences:*

$$X \perp\!\!\!\perp Y | Z, R_x, R_y, R_z \tag{7}$$

$$X \perp\!\!\!\perp R_y | Z, R_x, R_z \tag{8}$$

$$R_x \perp\!\!\!\perp R_y | Z, R_z \tag{9}$$

In words, any d-separation that can be expressed in the format stated above is testable. It is understood that, if X or Y or Z are fully observed, the corresponding R variables may be removed from the conditioning set. Clearly, any conditional independence comprised exclusively of fully observed variables is testable. To search for such refutable claims,

one needs to only examine the missing edges in the graph and check whether any of its associated set of separators satisfy the syntatctic format above.

To illustrate the power of the criterion we present the following example.

Example 8 *Examine the m-graph in figure 5 (d). The missing edges between Z and R_z, and X and R_z correspond to the conditional independencies: $Z \perp\!\!\!\perp R_z|(X,Y)$ and $X \perp\!\!\!\perp R_z|Y$, respectively. The former is untestable (following theorem 5) while the latter is testable, since it complies with (8) in theorem 6.*

4.1.1 Tests Corresponding to the Independence Statements in Theorem 6

A testable claim needs to be expressed in terms of proxy variables before it can be operationalized. For example, a specific instance of the claim $X \perp\!\!\!\perp Y|Z, R_x, R_y, R_z$, when $R_x = 0, R_y = 0, R_z = 0$ gives $X \perp\!\!\!\perp Y|Z, R_x = 0, R_y = 0, R_z = 0$. On rewriting this claim as an equation and applying equation 1 we get,

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

This equation exclusively comprises of observed quantities and can be directly tested given the input distribution: $P(X^*, Y^*, Z^*, R_x, R_y, R_z)$. Finite sample techniques for testing conditional independencies are cited in the next section. In a similar manner we can devise tests for the remaining two statements in theorem 6.

The tests corresponding to the three independence statements in theorem 6 are:

- $P(X^*|Z^*, R_x = 0, R_y = 0, R_z = 0) = P(X^*|Y^*, Z^*, R_x = 0, R_y = 0, R_z = 0)$,
- $P(X^*|Z^*, R_x = 0, R_z = 0) = P(X^*|R_y, Z^*, R_x = 0, R_z = 0)$
- $P(R_x|Z^*, R_z = 0) = P(R_x|R_y, Z^*, R_z = 0)$

The next section specializes these results to the classes of MAR and MCAR problems which have been given some attention in the existing literature.

4.2 Testability of MCAR and MAR

A chi square based test for MCAR was proposed by Little (1988) in which a high value falsified MCAR(Rubin, 1976). Rubin-MAR is known to be untestable (Allison, 2002). Potthoff et al. (2006) defined MAR at the variable-level (identical to that in section 2.2) and showed that it can be tested. Theorem 7, given below presents stronger conditions under which a given MAR model is testable (Mohan and Pearl (2014b)). Moreover, it provides diagnostic insight in case the test is violated. We further note that these conditional independence tests may be implemented in practice using different techniques such as G-test, chi square test, testing for zero partial correlations or by tests such as those described in Székely et al. (2007); Gretton et al. (2012); Sriperumbudur et al. (2010).

Theorem 7 (MAR is Testable) *Given that* $|V_m| > 0$, $V_m \perp\!\!\!\perp R | V_o$ *is testable if and only if* $|V_m| > 1$ *i.e.* $|V_m|$ *is not a singleton set.*

In words, given a dataset with two or more partially observed variables, it is always possible to test whether MAR holds. We exemplify such tests below.

Example 9 (Tests for MAR) *Given a dataset where* $V_m = \{A, B\}$ *and* $V_o = \{C\}$*, the MAR condition states that* $(A, B) \perp\!\!\!\perp (R_A, R_B) | C$*. This statement implies the following two statements which match syntactic criteria in 8 and hence are testable.*

1. $A \perp\!\!\!\perp R_B | C, R_A$
2. $B \perp\!\!\!\perp R_A | C, R_B$

While refutation of these tests immediately imply that the data are not MAR, we can never *verify* the MAR condition. However if MAR is refuted, it is possible to pinpoint and locate the source of error in the model. For instance, if claim (1) is refuted then one should consider adding an edge between A and R_B.

4.3 On the Causal Nature of the Missing Data Problem

Examine the m-graphs in Figure 5(b) and (c). $X \perp\!\!\!\perp R_x | Y$ and $X \perp\!\!\!\perp R_x$ are the conditional independence statements embodied in models 5(b) and (c), respectively. Neither of these statements are testable. Therefore they are statistically indistinguishable. However, notice that $P(XY)$ is recoverable in figure 5(b) but not in figure 5(c) implying that,

- No universal algorithm exists that can decide if a query is recoverable or not without looking at the model.

Further notice that $P(X)$ is recoverable in both models albeit using two different methods. In model 5(b) we have $P(X) = \sum_Y P(X^*|Y, R_x = 0)P(y)$ and in model 5(c) we have $P(X) = P(X^*|R_x = 0)$. This leads to the conclusion that,

- No universal algorithm exists that can produce a consistent estimate whenever such exists.

The impossibility of determining from statistical assumptions alone, (i) whether a query is recoverable and (ii) how the query is to be recovered, if it is recoverable, attests to the causal nature of the missing data problem. Although Rubin (1976) alludes to the causal aspect of this problem, subsequent research has treated missing data mostly as a statistical problem. A closer examination of the testability and recovery conditions shows however that a more appropriate perspective would be to treat missing data as a causal inference problem.

5 Related Work

For detailed discussion of missing data theory and practice we direct readers to the books (Allison, 2002; Enders, 2010; Little and Rubin, 2002; McKnight et al., 2007). Among all methods used for handling missing data, listwise deletion and pairwise deletion are the easiest to implement and have been found to be popular among practitioners (Peugh and Enders (2004)) even though estimates produced by these methods are guaranteed to converge only under MCAR.

Listwise deletion or (complete case analysis) refers to the simple technique in which samples with missing values are deleted (Buhi et al., 2008). Unless data are missing completely at random, listwise deletion can bias the outcome (Wothke, 2000). Evidently this technique results in wastage of data.

Pairwise deletion (or available case analysis) is a deletion method that drastically reduces data loss by operating on all samples in which the variables of interest are observed (Schlomer et al., 2010). For example, to compute the covariance of variables X and Y, all those cases or observations in which both X and Y are observed are used, regardless of whether other variables in the dataset have missing values.

Another approach to handling missing data is imputation: substituting a reasonable guess for each missing value (Allison, 2002). A simple example is *mean Substitution*, in which all missing observations of variable X are substituted with the mean of all observed values of X. Hot-deck imputation, cold-deck imputation (McKnight et al., 2007), regression imputation (Scheffer (2002)) and Multiple Imputation (Rubin, 1987, 1996) are examples of popular imputation procedures. Among these techniques, regression imputation guarantees consistent estimates for MAR data (Peugh and Enders (2004)). While many other imputation techniques are attractive in practice, performance guarantees (eg: convergence

and unbiasedness) are based primarily on simulation experiments.

Whenever data are Missing At Random, Maximum Likelihood (ML) based methods can be used for computing consistent estimates of parameters of interest (Little and Rubin, 2002). Recent increase in the popularity of ML based procedures can be attributed to its quick and easy availability in the form of software packages. The expectation-maximization (EM) algorithm (Dempster et al. (1977)) is a general technique for finding maximum likelihood (ML) estimates from MAR data.

Weighting procedures for missing data are based on creating weighted copies of complete cases and are succinctly summarized in Li et al. (2013). These procedures that are primarily based on Horvitz and Thompson (1952) and have been generalized to address missing data problems in Robins et al. (1994), Robins et al. (1995) and Robins et al. (2000).

The handling of MNAR data is more or less limited to performing sensitivity analysis (Resseguier et al., 2011). Methods of performing sensitivity analysis has been suggested in research publications such as Rotnitzky et al. (1998); Molenberghs et al. (2001) and Thijs et al. (2002). Special handling of MNAR problems based on use of selection models (Heckman (1977)) and pattern mixture models is discussed in Enders (2011).

Missing data discussed so far is a special case of *coarse data*, namely data that contains observations made in the power set rather than the sample space of variables of interest Heitjan and Rubin (1991). The notion of coarsening at random (CAR) was introduced in Heitjan and Rubin (1991) and identifies the condition under which coarsening mechanism can be ignored while drawing inferences on the distribution of variables of interest (Gill et al. (1997)). The notion of sequential CAR has been discussed in Gill and Robins (1997). Detailed discussions on coarsened data are available in Van der Laan and Robins (2003).

6 Conclusions

All methods of missing data analysis rely on assumptions regarding the reasons for missingness. Casting these assumptions in a graphical model, permits researchers to benefit from the inherent transparency of such models as well as their ability to explicate the statistical implication of the underlying assumptions in terms of conditional independence relations among observed and partially observed variables. We have shown that these features of graphical models can be harnessed to study unchartered territories of missing data research. In particular, we charted the estimability of statistical and causal parameters in broad classes of MNAR problems, and the testability of the model assumptions under missingness conditions.

The testability criteria derived in this paper can be used not only to rule out misspecified models but also to locate specific mis-specifications for the purpose of model updating and re-specification. More importantly, we have shown that it is possible to determine if and how a target quantity is recoverable, even in models where missingness is not ignorable. Finally, knowing which sub-structures in the graph prevent recoverability can guide data collection procedures by identifying auxiliary variables that need to be measured to ensure recovery, or problematic variables that may compromise recovery if measured imprecisely.

References

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

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

Allison, P. D. (2003). Missing data techniques for structural equation modeling. *Journal of abnormal psychology 112*(4), 545.

Balakrishnan, N. (2010). *Methods and applications of statistics in the life and health sciences*. John Wiley & Sons.

Buhi, E., P. Goodson, and T. Neilands (2008). Out of sight, not out of mind: strategies for handling missing data. *American journal of health behavior 32*, 83–92.

Chen, D.-G. and J. Wilson (2015). *Innovative Statistical Methods for Public Health Data*. Springer.

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

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

De Leeuw, J., E. Meijer, and H. Goldstein (2008). Handbook of multilevel analysis.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.

Elwert, F. (2013). Graphical causal models. In *Handbook of causal analysis for social research*, pp. 245–273. Springer.

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

Enders, C. K. (2011). Missing not at random models for latent growth curve analyses. *Psychological methods 16*(1), 1.

Gill, R. and J. Robins (1997). Sequential models for coarsening and missingness. In *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 295–305. Springer.

Gill, R., M. Van Der Laan, and J. Robins (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pp. 255–294. Springer.

Gleason, T. C. and R. Staelin (1975). A proposal for handling missing data. *Psychometrika 40*(2), 229–252.

Graham, J. (2012). *Missing Data: Analysis and Design (Statistics for Social and Behavioral Sciences)*. Springer.

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual review of psychology 60*, 549–576.

Greenland, S. and J. Pearl (2011). Causal diagrams. In *International encyclopedia of statistical science*, pp. 208–216. Springer.

Gretton, A., K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola (2012). A kernel two-sample test. *Journal of Machine Learning Research 13*(Mar), 723–773.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, 67–82.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions).

Heitjan, D. and D. Rubin (1991). Ignorability and coarse data. *The Annals of Statistics*, 2244–2253.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association 47*(260), 663–685.

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

Lauritzen, S. L. (2001). Causal inference from graphical models. *Complex stochastic systems*, 63–107.

Li, L., C. Shen, X. Li, and J. M. Robins (2013). On weighting approaches for missing data. *Statistical methods in medical research 22*(1), 14–30.

Little, R. and D. Rubin (2002). *Statistical analysis with missing data*. Wiley.

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

McKnight, P., K. McKnight, S. Sidani, and A. Figueredo (2007). *Missing data: A gentle introduction*. Guilford Press.

Meyers, L. S., G. Gamst, and A. J. Guarino (2006). *Applied multivariate research: Design and interpretation*. Sage.

Mohan, K. and J. Pearl (2014a). Graphical models for recovering probabilistic and causal queries from missing data. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27*, pp. 1520–1528. Curran Associates, Inc.

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

Mohan, K. and J. Pearl (2016). When data are missing not at random (mnar): Overcoming theoretical impediments. Technical Report R-465, UCLA. Currently Under Review.

Mohan, K., J. Pearl, and J. Tian (2013). Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems 26*, pp. 1277–1285.

Mohan, K., G. Van den Broeck, A. Choi, and J. Pearl (2014). An efficient method for bayesian network parameter learning from incomplete data. Technical report, UCLA. Presented at Causal Modeling and Machine learning Workshop, ICML-2014.

Molenberghs, G., M. G. Kenward, and E. Goetghebeur (2001). Sensitivity analysis for incomplete contingency tables: the slovenian plebiscite case. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 50*(1), 15–29.

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

Osborne, J. W. (2014). *Best practices in logistic regression*. SAGE Publications.

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

Pearl, J. and E. Bareinboim (2014). External validity: From do-calculus to transportability across populations. *Statistical Science 29*(4), 579–595.

Peters, C. L. O. and C. Enders (2002). A primer for the estimation of structural equation models in the presence of missing data: Maximum likelihood algorithms. *Journal of Targeting, Measurement and Analysis for Marketing 11*(1), 81–95.

Peugh, J. and C. Enders (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research 74*(4), 525–556.

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

Resseguier, N., R. Giorgi, and X. Paoletti (2011). Sensitivity analysis when data are missing not-at-random. *Epidemiology 22*(2), 282.

Rhoads, C. H. (2012). Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy 3*(1).

Robins, J. M., M. A. Hernan, and B. Brumback (2000). Marginal structural models and causal inference in epidemiology.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association 89*(427), 846–866.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association 90*(429), 106–121.

Rotnitzky, A., J. M. Robins, and D. O. Scharfstein (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the american statistical association 93*(444), 1321–1339.

Rubin, D. (1976). Inference and missing data. *Biometrika 63*, 581–592.

Rubin, D. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley Online Library.

Rubin, D. B. (1978). Multiple imputations in sample surveys-a phenomenological bayesian approach to nonresponse. In *Proceedings of the survey research methods section of the American Statistical Association*, Volume 1, pp. 20–34. American Statistical Association.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association 91*(434), 473–489.

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, 153–160.

Schlomer, G. L., S. Bauman, and N. A. Card (2010). Best practices for missing data management in counseling psychology. *Journal of Counseling psychology 57*(1), 1.

Shpitser, I. (2016). Consistent estimation of functions of data missing non-monotonically and not at random. In *Advances in Neural Information Processing Systems*, pp. 3144–3152.

Shpitser, I., K. Mohan, and J. Pearl (2015). Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*.

Shpitser, I. and J. Pearl (2006). Identification of conditional interventional distributions. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 437–444.

Sriperumbudur, B. K., A. Gretton, K. Fukumizu, B. Schölkopf, and G. R. Lanckriet (2010).

Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research 11*(Apr), 1517–1561.

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

Székely, G. J., M. L. Rizzo, N. K. Bakirov, et al. (2007). Measuring and testing dependence by correlation of distances. *The annals of statistics 35*(6), 2769–2794.

Thijs, H., G. Molenberghs, B. Michiels, G. Verbeke, and D. Curran (2002). Strategies to fit pattern-mixture models. *Biostatistics 3*(2), 245–265.

Thoemmes, F. and K. Mohan (2015). Graphical representation of missing data problems. *Structural Equation Modeling: A Multidisciplinary Journal*.

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

Tian, J. and J. Pearl (2002). 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.

Van den Broeck, G., K. Mohan, A. Choi, A. Darwiche, and J. Pearl (2015). Efficient algorithms for bayesian network parameter learning from incomplete data. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 161–170.

Van der Laan, M. and J. Robins (2003). *Unified methods for censored longitudinal data and causality*. Springer Verlag.

van Stein, B. and W. Kowalczyk (2016). An incremental algorithm for repairing training sets with missing values. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 175–186. Springer.

Verma, T. and J. Pearl (1991). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Conference in Artificial Intelligence*, pp. 220–227. Association for Uncertainty in AI.

Wothke, W. (2000). *Longitudinal and multigroup modeling with missing data*. Lawrence Erlbaum Associates Publishers.

7 Appendix

7.1 Estimation when the Data May not be Missing at Random. (Little and Rubin (2002), page-22)

Essentially all the literature on multivariate incomplete data assumes that the data are MAR, and much of it also assumes that the data are MCAR. Chapter 15 deals explicitly with the case when the data are not MAR, and models are needed for the missing-data mechanism. Since it is rarely feasible to estimate the mechanism with any degree of confidence, the main thrust of these methods is to conduct sensitivity analyses to assess the effect of alternative assumptions about the missing-data mechanism.

7.2 A Complex Example of Recoverability

We use $R = 0$ as a shorthand for the event where all variables are observed i.e. $R_{V_m} = 0$.

Example 10 *Given the m-graph in figure 4 (c), we will now recover the joint distribution.*

$$P(W,X,Y,Z) = P(W,X,Y,Z)\frac{P(W,X,Y,Z,R=0)}{P(W,X,Y,Z,R=0)} = \frac{P(W,X,Y,Z,R=0)}{P(R=0|W,X,Y,Z)}$$

Factorization of the denominator based on topological ordering of R variables yields,

$$P(W,X,Y,Z) = \frac{P(W,X,Y,Z,R=0)}{P(R_y=0|W,X,Y,Z,R_x=0,R_w=0,R_z=0)P(R_x=0|W,X,Y,Z,R_w=0,R_z=0)} \\ \frac{1}{P(R_w=0|W,X,Y,Z,R_z=0)P(R_z=0|W,X,Y,Z)}$$

On simplifying each factor of the form: $P(R_a = 0|B)$, by removing from it all $B_1 \in B$ such that $R_a \perp\!\!\!\perp B_1|B - B_1$, we get:

$$P(W,X,Y,Z) = \frac{P(W,X,Y,Z,R=0)}{P(R_z=0)P(R_w=0|Z)P(R_y=0|X,W,R_x=0)P(R_x=0|Y,W)} \qquad (10)$$

$P(WXYZ)$ is recoverable if all factors in the preceding equation is recoverable. Examining each factor one by one we get:

- *$P(W, X, Y, Z, R = 0)$: Recoverable as $P(W^*, X^*, Y^*, Z^*, R = 0)$ using equation 1.*
- *$P(R_z = 0)$: Directly estimable from the manifest distribution.*
- *$P(R_w = 0|Z)$: Recoverable as $P(R_w = 0|Z^*, R_z = 0)$, using $R_w \perp\!\!\!\perp R_z|Z$ and equation 1.*
- *$P(R_y = 0|X, W, R_x = 0)$: Recoverable as $P(R_y = 0|X^*, W^*, R_x = 0, R_w = 0)$, using $R_y \perp\!\!\!\perp R_w|X, W, R_x$ and equation 1.*
- *$P(R_x = 0|Y, W)$: The procedure for recovering $P(R_x = 0|Y, W)$ is rather involved and requires converting the probabilistic sub-query to a causal one as detailed below.*

$$\begin{aligned} P(R_x = 0|Y, W = w) &= P(R_x = 0|Y, do(W = w)) \textit{(Rule-2 of do calculus)} \\ &= \frac{P(R_x = 0|Y, R_y = 0, do(w))}{P(R_x = 0|Y, R_y = 0, do(w))} P(R_x = 0|Y, do(W = w)) \\ &= P(R_x = 0|Y, R_y = 0, do(w)) \frac{P(R_y = 0|Y, do(w))}{P(R_y = 0|Y, do(w), R_x = 0)} \end{aligned} \quad (11)$$

To prove recoverability of $P(R_x = 0|Y, W = w)$, we have to show that all factors in equation 11 are recoverable.

Recovering $\mathbf{P(R_y = 0|Y, do(w), R_x = 0)}$: *Observe that $P(R_y = 0|Y, do(w), R_x = 0) = P(R_y = 0|do(w), R_x = 0)$ by Rule-1 of do calculus. To recover $P(R_y = 0|do(w), R_x = 0)$ it is sufficient to show that $P(X^*, Y^*, R_x, R_y, Z|do(w))$ is recoverable in G', the latent*

structure corresponding to G in which X and Y are treated as latent variables.

$$\begin{aligned}
P(X^*, Y^*, R_x, R_y, Z|do(w)) &= P(X^*, Y^*, R_x, R_y|Z, do(w))P(Z|do(w)) \\
&= P(X^*, Y^*, R_x, R_y|Z, w)P(Z|do(w)) \text{ (Rule-2 of do-calculus)} \\
&= P(X^*, Y^*, R_x, R_y|Z, w)P(Z) \text{ (Rule-3 of do-calculus)}
\end{aligned}$$

Using $(X^*, Y^*, R_x, R_y) \perp\!\!\!\perp (R_z, R_w)|(Z, W)$, equation 1 and $Z \perp\!\!\!\perp R_z$ we show that the causal effect is recoverable as:

$$P(X^*, Y^*, R_x, R_y, Z|do(w)) = P(X^*, Y^*, R_x, R_y|Z^*, w^*, R_w = 0, R_z = 0)P(Z^*|R_z = 0) \tag{12}$$

Recovering $\mathbf{P(R_x = 0|Y, do(w), R_y = 0)}$: *Using equation 1, we can rewrite $P(R_x = 0|Y, do(w), R_y = 0)$ as $P(R_x = 0|Y^*, do(w), R_y = 0)$. Its recoverability follows from equation 12.*

Recovering $\mathbf{P(R_y = 0|Y, do(w))}$:

$$\begin{aligned}
P(R_y = 0|Y, do(w)) &= \frac{P(R_y = 0, Y|do(w))}{\sum_{R_x} P(R_y = 0, Y, R_x|do(w)) + P(R_y = 1, Y, R_x|do(w))} \\
&= \frac{P(R_y = 0, Y^*|do(w))}{\sum_{R_x} P(R_y = 0, Y^*, R_x|do(w)) + P(R_y = 1, Y, R_x|do(w))} \text{(using eq 1)}
\end{aligned}$$

$P(R_y = 0, Y^*|do(w))$ and $P(R_y = 0, Y^*, R_x|do(w))$ are recoverable from equation 12. We will now show that $P(R_y = 1, Y^*, R_x|do(w))$ is recoverable as well.

$$P(R_y = 1, Y, R_x|do(w)) = \frac{P(R_y = 0, Y, R_x|do(w))}{P(R_y = 0|R_x, Y|do(w))} - P(R_y = 0, R_x, Y|do(w))$$

Using equation 1 and Rule-1 of do-calculus we get,

$$= \frac{P(R_y = 0, Y^*, R_x|do(w))}{P(R_y = 0|R_x, do(w))} - P(R_y = 0, R_x, Y^*|do(w))$$

Each factor in the preceding equation is estimable from equation 12. Hence $P(R_y = 1, Y, R_x, do(w))$ *and therefore,* $P(R_y = 0|Y, do(w))$ *is recoverable.*

Since all factors in equation 11 are recoverable, joint distribution is recoverable.