Missing at Random in Graphical Models

Jin Tian
Iowa State University

Abstract

The notion of missing at random (MAR) plays a central role in the theory underlying current methods for handling missing data. However the standard definition of MAR is difficult to interpret in practice. In this paper, we assume the missing data model is represented as a directed acyclic graph that not only encodes the dependencies among the variables but also explicitly portrays the causal mechanisms responsible for the missingness process. We introduce an intuitively appealing notion of MAR in such graphical models, and establish its relation with the standard MAR and a few versions of MAR used in the literature. We address the question of whether MAR is testable, given that data are corrupted by missingness, by proposing a general method for identifying testable implications imposed by the graphical structure on the observed data.

1 Introduction

The missing data problem is ubiquitous in every experimental science. There is a vast literature on missing data in diverse fields such as social science, biology, statistics and machine learning. Most of the current methods for handling missing data are based on the seminal theoretical work of Rubin [Rubin, 1976, Little and Rubin, 2002]. Central to Rubin's missing data theory is the concept of *missing at random (MAR)*.[1] Under the MAR assumption, likelihood-based inference (as well as Bayesian inference) can be carried out while ignoring the mechanism that leads to missing data. For example, most work in machine learning assumes MAR in dealing with missing data and proceeds with maximum likelihood or Bayesian inference. For exceptions see [Jaeger, 2006a], and recent work on collaborative filtering that explicitly incorporate missing data mechanism into the model [Marlin *et al.*, 2007, Marlin and Zemel, 2009, Marlin *et al.*, 2011].

Many authors find Rubin's definition of MAR difficult to apply in practice because it invokes subtle event-level conditional independences (CIs) [Schafer and Graham, 2002, McKnight *et al.*, 2007, Graham, 2012]. Indeed the MAR definition is quite unnatural from the modelling point of view (see Example 2 in Section 2). In Practice many authors often work with random variable level CIs when interpreting MAR. Recently Mohan et al (2013) have proposed to use graphical models to encode the missing data model, called *m-graphs*, by representing both CI relations among variables and the causal mechanims responsible for the missingness process. Mohan et al. (2013) derived graphical conditions under which a probability quantity can be estimated consistently from data with missing values. Mohan and Pearl (2014) investigated whether such graphical models are subjected to statistical test, noting testability of CIs is impeded when data are contaminated by missing values.

In this paper we investigate several versions of definition of MAR used in the literature and propose a graphical version. We then address the question of whether a hypothesized graphical model is testable and whether MAR is testable in a given model. Our main contributions are:

- We introduce a version of MAR defined in terms of graphical structures, called G-MAR, and formally establish its relation with variable level MAR and with the standard MAR. G-MAR is intuitively appealing from the modelling point of view, and is much easier to interpret and to justify/falsify than the standard MAR.

- We propose a method for identifying testable implications in graphical missing data models, and for testing MAR assumption in those models, given that data are corrupted by missingness.

[1] Missing data is a special case of *coarse data*, and MAR is a special case of coarsening at random (CAR). We will not consider CAR in this paper. For theories on CAR we refer to [Heitjan and Rubin, 1991, Gill *et al.*, 1997, Jaeger, 2005b].

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

The paper is organized as follows. In Section 2, we review Rubin's missing data theory. Section 3 defines the notion of m-graphs as introduced in [Mohan *et al.*, 2013]. In Section 4 we investigate several versions of MAR definition, introduce our graphical version G-MAR, and establish their relations with each other and with the stardard MAR. In Section 5 we discuss how to identify testable implications implied by m-graphs on the observed data in the face of data corrupted with missing values, and how to test the G-MAR assumption. Section 6 concludes the paper.

2 Missing Data Theory

In this section we review the theory of missing data mainly due to Rubin [Rubin, 1976, Little and Rubin, 2002].

Let $V = \{V_1, \ldots, V_n\}$ be a set of random variables with probability distribution $P(V|\Theta)$ where Θ denotes model parameters. Assume that we observe an i.i.d set of data cases that may contain missing values. It is crucial in the analysis of data with missing values to understand the mechanisms that lead to missing data, in particular whether the fact that variables are missing is related to the underlying values of the variables in the data set. For this purpose we introduce a set of *missingness indicator* variables $R = \{R_{V_1}, \ldots, R_{V_n}\}$ such that $R_{V_i} = 1$ where the value of corresponding V_i is missing and $R_{V_i} = 0$ where V_i is observed. Let $V^* = \{V_1^*, \ldots, V_n^*\}$ be a set of random variables that represent the values of V we actually observe such that each V_i^* is a deterministic function of V_i and R_{V_i}: $V_i^* = f(V_i, R_{V_i})$, defined by

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

where the $*$ symbol represents a missing value.

Assume our goal is to estimate Θ parameters given an i.i.d set of observations $\{(v^*, r)\}$. With the presence of missing values, we must consider not only the data-generation model $P(V|\Theta)$, but also the mechanism for missingness. The joint distribution $P(V, R|\Theta, \Phi) = P(V|\Theta)P(R|V, \Phi)$ will be called the *missing data model* where Θ and Φ denote model parameters, and the *missingness mechanism* is characterized by the conditional distribution $P(R|V, \Phi)$.

Rubin has classified missingness mechanisms into three types: *missing completely at random (MCAR)*, *missing at random (MAR)*, and *missing not at random (MNAR)*. The missing data model is MCAR if $P(r|v, \Phi) = P(r|\Phi)$, that is, missingness does not depend on the values of V, missing or observed. In other words, a missing data model is MCAR if $V \perp\!\!\!\perp R$.[2] One example of MCAR is when respondents decide to reveal their income levels based on coin-flips. If the missing data model is MCAR, then $P(V, R|\Theta, \Phi) = P(V|\Theta)P(R|\Phi)$. As a consequence, assuming that the parameters Θ and Φ are distinct, the likelihood function of Θ and Φ decouples, and we can maximize the likelihood of Θ parameters independently, ignoring the missingness mechanism $P(R|\Phi)$.

MCAR is a very strong requirement. It turns out there are weaker conditions under which the likelihood-based inference can proceed while ignoring the missingness mechanism. For any missingness pattern r, let v_{obs} denote the observed components of v, and v_{mis} the missing components. A missing data model is MAR if, for all $P(r, v) > 0$,

$$P(r|v_{obs}, v_{mis}, \Phi) = P(r|v_{obs}, \Phi) \quad \text{for all } v_{mis} \text{ values}, \quad (2)$$

where the values of r are given by $r_{V_i} = 0$ if $V_i \in V_{obs}$ and $r_{V_i} = 1$ if $V_i \in V_{mis}$. The MAR assumption requires that missingness is independent of the missing values v_{mis} given the observed values v_{obs}. Note that this is *event-level* conditional independence with specific r values, not conditional independence between random variables. Also it asks for different conditional independences for data cases with different missingness patterns.

To understand the requirements by the MAR assumption, we consider a couple of examples.

Example 1 *In the case that there is only one single variable X with missing values and all other variables S are fully observed, the MAR assumption requires $P(R_X = 1|x, s) = P(R_X = 1|s)$. This also leads to $P(R_X = 0|x, s) = P(R_X = 0|s)$. Therefore in the case that missingness only occurs for a single variable X, the MAR assumption requires that the random variable R_X is independent of the random variable X given observed variables.*

Example 2 *Now consider two variables X and Y both with missing values. The MAR assumption requires the following parameterization (Little and Rubin 2002, Example 1.13, page 18)*

$$\begin{aligned} &P(R_X = 1, R_Y = 1|x, y) = P(R_X = 1, R_Y = 1) = g_{11} \\ &P(R_X = 1, R_Y = 0|x, y) = P(R_X = 1, R_Y = 0|y) = g_{10}(y) \\ &P(R_X = 0, R_Y = 1|x, y) = P(R_X = 0, R_Y = 1|x) = g_{01}(x) \\ &P(R_X = 0, R_Y = 0|x, y) = g_{00}(x, y) = 1 - g_{11} - g_{10}(y) - g_{01}(x) \end{aligned} \quad (3)$$

where $g_{ij}(\cdot)$ represents some function. We see that the MAR assumption allows that missingness variable R_X sometimes depends on X and sometimes is independent of X depending on the values of R_X.

From a modelling point of view, the MAR assumption is quite unnatural and appears to be artificially made. The importance of the MAR assumption lies in that it is said to be the weakest general condition under which we can maximize the likelihood function in the parameters Θ of the data generation distribution $P(V|\Theta)$ based on

[2] We use $X \perp\!\!\!\perp Y|Z$ to denote that X is conditionally independent of Y given Z.

observed data while ignoring the missingness mechanism $P(R|V,\Phi)$ (assuming that the parameters Θ and Φ are distinct) [Rubin, 1976]. This is based on the observation that the distribution of the observed data can be decomposed as

$$P(v_{obs}, r|\Theta,\Phi) = \sum_{v_{mis}} P(v_{obs}, v_{mis}|\Theta)P(r|v_{obs}, v_{mis}, \Phi)$$
$$= P(v_{obs}|\Theta)P(r|v_{obs}, \Phi). \quad (4)$$

A missing data model is MNAR if it is not MAR. For example, online users were found to rate an item more likely either if they love the item or if they hate it [Marlin *et al.*, 2007]. In other words, the probability that a rating is missing is dependent on the user's underlying preferences. If the model is MNAR, then the full missing data model including the missingness mechanism is needed for likelihood-based inference for Θ.

In Section 4 we will investigate the MAR assumption articulated as conditional independences between random variables and then define a version of MAR when the missing data model can be represented as a directed acyclic graph (DAG).

3 Graphical Representation of Missing Data Model

Graphical models are widely used for representing data generation models [Pearl, 2000, Koller and Friedman, 2009]. Mohan et al. (2013) used DAGs, called *missingness graphs* or *m-graphs* for short, to represent both the data generation model and the causal mechanisms responsible for the missingness process. Next we define m-graphs, mostly following the notations used in [Mohan *et al.*, 2013].

Let G be a DAG over a set of variables $V \cup U \cup R$, where V is the set of observable variables, U is the set of unobserved latent variables, and R is the set of missingness indicator variables representing the causal mechanisms that are responsible for missingness.[3] We assume that V is partitioned into V_o and V_m such that V_o is the set of variables that are observed in all data cases and V_m is the set of variables that are missing in some data cases and observed in other cases.[4] Every variable $V_i \in V_m$ is associated with a variable $R_{V_i} \in R$ such that, in any observed data case, $R_{V_i} = 1$ if the value of corresponding V_i is missing and $R_{V_i} = 0$ if V_i is observed. We require that R variables may not be parents of variables in $V \cup U$.[5] For any set $S \subseteq V_m$, let R_S represent the set of R variables corresponding to variables in S.

The DAG G provides a compact representation of the missing data model $P(V,U,R) = P(V,U)P(R|V,U)$, and will be called a m-graph of the model. We will often say the model P is *compatible* with the DAG G. The m-graph depicts both the dependency relationships among variables in $V \cup U$ and the missingness mechanisms, and it encodes conditional independence relationships that can be read off the graph by d-separation criterion [Pearl, 1988] such that every d-separation in the graph G implies conditional independence in the distribution P. On the other hand, if every conditional independence relation true in P is also captured by the d-separation in G, then we say P is *faithful* to G. See Figures 1(a) and 2(a) for examples of m-graphs. We use solid circles to represent always observed variables in V_o and R, and hollow circles to represent partially observed variables in V_m.

For every variable $V_i \in V_m$ we introduce a proxy variable V_i^* to represent the values of V_i that are actually observed such that V_i^* is a deterministic function of V_i and R_{V_i}: $V_i^* = f(V_i, R_{V_i})$, as defined by Eq. (1). We may expand a m-graph G with V_i^* variables by adding each V_i^* as a common child of V_i and R_{V_i}, and will denote the resulting DAG G^*, called *enriched* m-graph. See Figures 1(b) and 2(b) for examples of enriched m-graphs. Note that Mohan et al (2013) defined the graph G^* as the m-graph, and did not introduce the notion of enriched graph. However it is important for us in this paper to distinguish G with G^* as demonstrated in Section 5.

We remark that we do not require the m-graphs to be "causal". However we would like to emphasize that the power of DAG models often stems from its causal interpretation, and we find it hard to imagine how a m-graph involving missingness mechanism would be formed without causal thinking. Although Rubin's MAR does not include causal assumption, it suffers from the cognitive difficulty of whether researchers are capable of judging the plausibility of those assumptions.

4 MAR in M-graphs

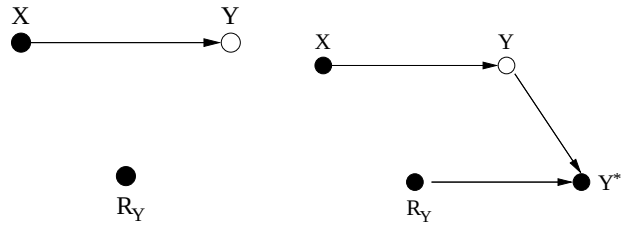
In this section, we consider the MAR assumption in m-graphs. But we will first investigate the MAR assumption articulated as conditional independences between random variables as often (implicitly) used in the literature.

The missing data model is MCAR if $R \perp\!\!\!\perp (V \cup U)$. In m-

[3] Note that we do not allow selection variables as we do not consider selection bias issue in this paper.

[4] We assume we can partition the V variables into V_o and V_m based on domain knowledge (or modeling assumption). In many applications, we have the knowledge that some variables are always observed in all data cases. In others, the partition could be based on the modeling assumption whether a variable (that is not actually missing in the given data set) could be potentially missing. Note this is not some extra assumption of m-graphs, rather Rubin's MAR faces the same issue whether the MAR condition is required only for the realized missing patterns in given data or not.

[5] R variables are missingness indicator variables and we assume that the data generation process over V, U variables does not depend on the missingness mechanism.



(a) G (b) G^*

Figure 1: A m-graph G and its enriched m-graph G^* for a model that is MCAR. Here $V_o = \{X\}$, $V_m = \{Y\}$, and $R = \{R_Y\}$. We use solid circles to represent always observed variables, and hollow circles to represent partially observed variables in V_m.

graphs, this corresponds to that there will be no edges between R variables and the variables in $V \cup U$. The two conditions are equivalent if the distribution $P(V, U, R)$ is faithful to its m-graph G. For example the model shown in Figure 1 is MCAR.

4.1 Variable-level MAR

As we have seen in Section 2, the standard MAR assumption appears to be artificially made so that likelihood-based inference can be performed while ignoring the missingness mechanism. The assumption is quite unnatural from the modelling point of view. Many authors find the definition difficult to apply in practice and often work with random-variable-level independences [Schafer and Graham, 2002, McKnight *et al.*, 2007, Graham, 2012]. There are two subtleties with the MAR definition in Eq. (2): (a) it invokes event-level independences, and (b) it asks for independences between different sets of variables for different missingness patterns (i.e., different partitions of V into V_{obs} and V_{mis}). To address subtlety (a), we formally define MAR in terms of variable-level independences and name it MAR*.[6]

Definition 1 (MAR*) *A missing data model is called* MAR* *if, for all possible missingness pattern* $r = (r_{V_{mis}} = 1, r_{V_{obs}} = 0)$ *with* $P(r, v, u) > 0$, $R \perp\!\!\!\perp (V_{mis} \cup U) | V_{obs}$.

In general MAR* is a strictly stronger requirement than MAR since it asks for variable-level independences (i.e., it requires Eq. (2) holds for all possible values of r instead of a specific value). MAR* may still be difficult to apply in practice because it asks for a number of independences between different sets of variables. For example, assume we have three variables X, Y, and Z, such that in each data case the value of one and only one variable is revealed to us (based on some random process). Then MAR* asks for $R \perp\!\!\!\perp (X, Y) | Z$, $R \perp\!\!\!\perp (X, Z) | Y$, and $R \perp\!\!\!\perp (Y, Z) | X$.

An easier to interpret definition called MAR+ has been suggested by Potthoff et al. (2006) and is used as definition for MAR in [Mohan *et al.*, 2013], which has addressed subtlety (b) . Recall V_o is the set of variables that are always observed and V_m is the set of variables that are missing in some cases.

Definition 2 (MAR+) *A missing data model is called* MAR+ *if* $R \perp\!\!\!\perp (V_m \cup U) | V_o$.

How is MAR+ related to MAR*? If the possible missingness patterns include $r = (r_{V_m} = 1, r_{V_o} = 0)$, then the MAR+ condition $R \perp\!\!\!\perp (V_m \cup U) | V_o$ is also required by MAR*. However, in general MAR+ may be a slightly stronger requirement than MAR* as shown in the following proposition.

Proposition 1

1. *If a missing data model is MAR+, then it is MAR*.*
2. *If the data model $P(V)$ is strictly positive, then MAR+ and MAR* are equivalent.*
3. *If $P(R_{V_m} = 1, R_{V_o} = 0) > 0$, then MAR+ and MAR* are equivalent.*

Proof: The proofs are based on the graphoid axioms [Pearl, 1988] shown in the Appendix.

1. Consider any missingness pattern (V_{obs}, V_{mis}). Let $C = V_{obs} \setminus V_o = V_{obs} \cap V_m$ be the set of variables that are observed in this case but missing in some other cases. Then we have $V_m = V_{mis} \cup C$. Given $R \perp\!\!\!\perp (V_{mis} \cup C \cup U) | V_o$, by the weak union axiom, we obtain $R \perp\!\!\!\perp (V_{mis} \cup U) | (V_o \cup C)$ where $V_o \cup C = V_{obs}$.
2. We show MAR* implies $R \perp\!\!\!\perp (V_m \cup U) | V_o$ if $P(V) > 0$. If the missing data model is MAR*, then for any two missingness patterns we have $R \perp\!\!\!\perp (V^1_{mis} \cup U) | V^1_{obs}$ and $R \perp\!\!\!\perp (V^2_{mis} \cup U) | V^2_{obs}$. We obtain $R \perp\!\!\!\perp (V^1_{mis} \cup V^2_{mis} \cup U) | (V^1_{obs} \cap V^2_{obs})$ by the generalized intersection rule given in Lemma 1 in the Appendix (with $Z = V^1_{obs} \cap V^2_{obs}$, $S = (V^1_{mis} \cap V^2_{mis}) \cup U$, $Y = V^1_{mis} \cap V^2_{obs}$, and $W = V^2_{mis} \cap V^1_{obs}$). Keep using the generalized intersection rule to all missingness patterns and we obtain $R \perp\!\!\!\perp (V_m \cup U) | V_o$ since V_o is the set of variables that are always observed and V_m is the set of variables that are missing in at least one missingness pattern.
3. With $(R_{V_m} = 1, R_{V_o} = 0)$ being an allowed missingness pattern, the MAR+ condition $R \perp\!\!\!\perp (V_m \cup U) | V_o$ is also required by MAR*.

□

[6] Although many authors work implicitly with variable-level independences in using the MAR assumption, we have not seen such a formal definition in the literature.

4.2 MAR in M-graphs

Next we assume that the missing data model P is compatible with a m-graph G, and we look for graphical conditions under which *MAR* will hold such that likelihood-based inference can be correctly performed while ignoring the missingness mechanism.

For any missingness pattern (V_{obs}, V_{mis}), the distribution of the observed data can be computed as

$$\begin{aligned} & P(v_{obs}, r|\Theta, \Phi) \\ &= \sum_{v_{mis},u} P(v_{obs}, v_{mis}, u|\Theta) P(r|v_{obs}, v_{mis}, u, \Phi) \\ &= \sum_{v_{mis},u} P(v_{obs}, v_{mis}, u|\Theta) \prod_{\{i:V_i \in V_m\}} P(r_{V_i}|pa^{obs}_{r_i}, pa^{mis}_{r_i}, pa^{u}_{r_i}, pa^{r}_{r_i}, \Phi), \end{aligned} \tag{5}$$

where $pa^{obs}_{r_i}$, $pa^{mis}_{r_i}$, $pa^{u}_{r_i}$, and $pa^{r}_{r_i}$ represent the parents of R_{V_i} in G that are observed V variable, unobserved V variables, latent variables, and R variables respectively. The MAR assumption would hold if each term in the product in Eq. (5) is independent of $pa^{mis}_{r_i}$ and $pa^{u}_{r_i}$ values, and we will call this condition the local MAR.

Definition 3 (Local MAR) *A missing data model $P(V, U, R)$ compatible with m-graph G is called* local MAR *if, for every $R_{V_i} \in R$, for all $pa^{mis}_{r_i}$ and $pa^{u}_{r_i}$ values,*

$$P(r_{V_i}|pa^{obs}_{r_i}, pa^{mis}_{r_i}, pa^{u}_{r_i}, pa^{r}_{r_i}, \Phi) = P(r_{V_i}|pa^{obs}_{r_i}, pa^{r}_{r_i}, \Phi). \tag{6}$$

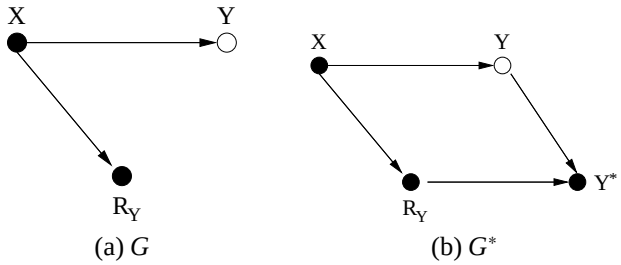
Note Eq. (6) represents context-specific independences in the fixed $pa^{r}_{r_i}$ values since it holds for both values (0 or 1) of R_{V_i}, that is, it requires that R_{V_i} is independent of $Pa^{mis}_{r_i}$ and $Pa^{u}_{r_i}$ given $Pa^{obs}_{r_i}$ in the *context* $pa^{r}_{r_i}$. Local MAR is somewhat easier to interpret than *MAR* by taking advantage of the conditional independences encoded in the m-graph, however it is also a stronger requirement than MAR.

Proposition 2 *If a missing data model P compatible with m-graph G is local MAR then it is MAR.*

Proof: If Eq. (6) holds, then $P(r|v_{obs}, v_{mis}, u, \Phi) = \prod_{\{i:V_i \in V_m\}} P(r_{V_i}|pa^{obs}_{r_i}, pa^{mis}_{r_i}, pa^{u}_{r_i}, pa^{r}_{r_i}, \Phi)$ is independent of v_{mis} and u values. Therefore the model is MAR. □

In general local MAR is a stronger requirement than MAR because that $P(r|v_{obs}, v_{mis}, u, \Phi)$ is independent of v_{mis} values does not necessarily mean each term $P(r_{V_i}|pa^{obs}_{r_i}, pa^{mis}_{r_i}, pa^{u}_{r_i}, pa^{r}_{r_i}, \Phi)$ in the product is independent of v_{mis} values. We have confirmed this by constructing a counter example.

A general graphical condition to guarantee that Eq. (6) always holds for all possible missingness patterns is that none of the unobserved V variables or latent variables are parents of R_{V_i} (i.e., $Pa^{mis}_{r_i} = Pa^{u}_{r_i} = \emptyset$). We then have



(a) G (b) G^*

Figure 2: A m-graph and its enriched m-graph for a model that is G-MAR.

$$\begin{aligned} P(v_{obs}, r|\Theta, \Phi)) &= \sum_{v_{mis},u} P(v_{obs}, v_{mis}, u|\Theta) \prod_{i} P(r_i|pa^{obs}_{r_i}, pa^{r}_{r_i}, \Phi) \\ &= P(v_{obs}|\Theta) P(r|v_{obs}, \Phi). \end{aligned} \tag{7}$$

The presence of any edge from some latent variable or a variable in V_m to some R_{V_i} will not guarantee the decouple of the likelihood function without further more subtle assumptions. We therefore use this requirement as the MAR condition in m-graphs and call it G-MAR (*graphical MAR*).

Definition 4 (G-MAR) *A missing data model compatible with a m-graph is called* G-MAR *if none of the variables in $V_m \cup U$ are parents of R variables.*

In other words, this definition says the model is G-MAR if missingness is not (directly) caused by any variables with missing values. This assumption is intuitively appealing and is much easier to understand than the standard MAR assumption. As an example, the model shown in Figure 2 is G-MAR. Note that if $V_o = \emptyset$, that is, every variable is missing in some data cases, then G-MAR is reduced to MCAR.

It can be expected that G-MAR is a stronger requirement than MAR. In fact, G-MAR is more closely related with random-variable-level MAR* and MAR+, and we have the following results.

Theorem 1
Let P be a missing data model compatible with a m-graph G.

1. *If P is G-MAR, then it is MAR+.*
2. *If P is MAR+, then it is MAR*.*
3. *If P is faithful to G, then the three conditions G-MAR, MAR+, and MAR* are equivalent.*
4. *If P is faithful to G, and there are no edges between R variables in G, then G-MAR is equivalent to local MAR.*

Proof:

1. Any path from any variable $R_{V_i} \in R$ to any variable in V_m or U is blocked by R's parents which are variables in V_o (recall we do not allow R variables to have V variables or U variables as children in the m-graph).

2. This has been proved in Proposition 1.

3. (a) We show MAR* implies MAR+. If P is faithful to G then all the conditional independences in P are captured by the d-separation criterion which is known to satisfy the intersection axiom [Pearl, 1988]. Then the generalized intersection rule holds (see Lemma 1 in the Appendix). Therefore MAR* implies MAR+ by the same proof for Proposition 1.2.

   (b)We show MAR+ implies G-MAR. If P is faithful to G then all the conditional independences in P are captured by the d-separation. If $R \perp\!\!\!\perp (V_m \cup U)|V_o$ in P, then there cannot exist edges between any R variable and any V_m or U variable. Therefore the model is G-MAR.

4. First it is obvious that if P is G-MAR then it is local MAR. Now if $Pa^r_{r_i} = \emptyset$, then Eq. (6) becomes a conditional independence requirement $R_{V_i} \perp\!\!\!\perp (Pa^{mis}_{r_i}, Pa^u_{r_i})|Pa^{obs}_{r_i}$ with the disappearance of the context $pa^r_{r_i}$. Since P is faithful to G, this means none of the variables in $V_m \cup U$ are parents of R variables.

□

In conclusion, G-MAR is a strictly stronger requirement than MAR, with the main difference being the former requires variable-level independences while the latter only asks for event-level independences. From modelling point of view, if a model is judged as MNAR by G-MAR, then most likely it will not be MAR, unless certain subtle independences between events are satisfied which are not captured by graphical structures and are normally difficult to judge or justify. An example of subtle independences that are difficult to justify and capture by graphical structure is the model specified by Eq. (3) in Example 2. An m-graph compatible with this model would have R_X (and R_Y) depending on X and Y, and the model would be considered MNAR by G-MAR.

5 Testable Implications in M-graphs

Does G-MAR assumption lend itself to statistical tests from data? To answer this question, we first address the more general question of whether a m-graph is testable given that data are corrupted by missing values. The problem of identifying constraints implied by graphical models has been well studied. Constraints in the form of conditional independences can be read off the graph through the d-separation criterion, and they constitute all the testable implications if there is no latent variables in the model [Pearl, 1988]. In the presence of latent variables, algorithms for identifying equality constraints (often called Verma constraints) are given in [Tian and Pearl, 2002, Shpitser and Pearl, 2008], and methods for identifying inequality constraints are given in [Pearl, 1995, Geiger and Meek, 1999, Bonet, 2001, Kang and Tian, 2006]. When data are complete, these constraints can be tested against data. The testability, however, is impeded when the data available are corrupted by missing values.

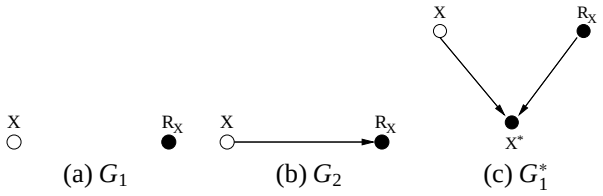


Figure 3: The two models G_1 and G_2 are indistinguishable, despite of the independence $X \perp\!\!\!\perp R_X$ encoded by G_1.

5.1 Peculiarity of Testability in Missing Data

Mohan and Pearl (2014) noted a peculiar phenomenon that in the presence of missing data, some conditional independences conveyed by the m-graph may not be testable even when the full joint distribution is estimable unbiasedly, as shown in the following example.

Example 3 *Consider a model P_1 compatible with the m-graph G_1 shown in Figure 3(a). P_1 is MCAR, and encodes independence $X \perp\!\!\!\perp R_X$. The joint distribution can be estimated from the observed data as*

$$P_1(X, R_X) = P_1(X|R_X = 0)P_1(R_X) = P_1(X^*|R_X = 0)P_1(R_X). \quad (8)$$

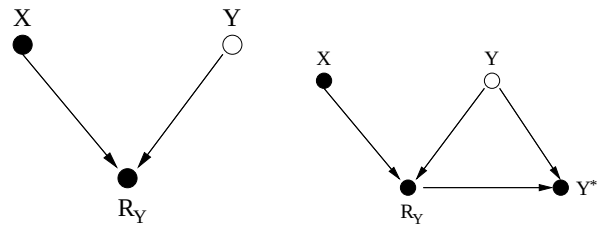
Can we then test the independence claim $X \perp\!\!\!\perp R_X$ in data and therefore distinguish this model with the model G_2 in Figure 3(b)? Perhaps surprisingly, the independence $X \perp\!\!\!\perp R_X$ is not testable, and the two models are indistinguishable. Next we show explicitly that any observed distribution $P(X^*, R_X)$ (produced by G_2) can be emulated by G_1. Formally, for any observed distribution $P(X^*, R_X)$, we can construct a model (shown in the following) satisfying $X \perp\!\!\!\perp R_X$ that produces $P(X^*, R_X)$:

$$P_1(X, R_X) = P_1(X)P_1(R_X), \quad (9)$$

where

$$P_1(X = x) = P(X^* = x|R_X = 0) \quad (10)$$

$$P_1(R_X) = P(R_X). \quad (11)$$



(a) G (b) G^*

Figure 4: The m-graph G encodes $X \perp\!\!\!\perp Y$, however G^* encodes no CIs but inequality constraints.

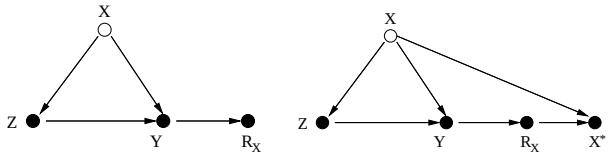
Mohan and Pearl (2014) addressed this issue by investigating the following question: what conditional independences (CI) $X \perp\!\!\!\perp Y|Z$ are (syntactically) testable, where X, Y, and Z may include V_m and R variables? Note they studied the syntactic testability of a given CI, not with respect to a given model or m-graph. Whether $X \perp\!\!\!\perp Y|Z$ is testable or not depends only on the syntax of the CI sentence, that is, the type (V_o, V_m, R) of variables that appear in the CI. For example, if $X, Y, Z \subset V_o$, then $X \perp\!\!\!\perp Y|Z$ is said to be testable. They have developed a number of sufficient conditions under which CI claims involving $V_o \cup V_m$ and R variables can be expressed in terms of the observed variables V_o, V^*, and R.

5.2 Resolving the Peculiarity

In this paper, we study the following question: what are the testable implications implied by a given model structure on the observed data given that data are corrupted by missing values? In the presence of missing data, the observed distribution is specified by $P(V_o, V^*, R)$. Our idea is that, to determine constraints implied by the m-graph G on $P(V_o, V^*, R)$, we could directly work with the enriched graph G^* which adds V^* variables to G. It is not necessary to first look for testable implications of G on $P(V, R)$ and then to try to figure out whether these implications are testable or not in terms of observed $P(V_o, V^*, R)$ (as done in [Mohan and Pearl, 2014]).

In other words, in order to identify testable implications implied by the m-graph G on the observed data, we look for testable implications implied by the enriched graph G^* on $P(V_o, V^*, R)$ assuming that the V_m variables are latent variables (as well as U). This latter problem has been well studied and many techniques have been developed (see the discussion in the beginning of Section 5). In conclusion, we have converted the peculiar testability problem in missing data into a problem well understood and studied.

Example 4 *To see whether the model in Figure 3(a) is testable, we consider the enriched graph shown in Figure 3(c), and conclude that the model imposes no constraints on $P(X^*, R_X)$. The peculiarity is resolved.*



(a) G (b) G^*

Figure 5: A m-graph in which whether there is an edge from X to R_X can be tested by independence $R_X \perp\!\!\!\perp Z$.

Example 5 *To see whether the model G in Figure 4(a) is testable (note this model is considered in [Mohan and Pearl, 2014, Example 8]), we consider the enriched graph G^* shown in Figure 4(b). Although G encodes $X \perp\!\!\!\perp Y$, there is no CIs among X, Y^*, and R_Y in G^*. However, for X and Y being discrete variables, it is known that G^* implies testable constraints in the form of inequalities called* instrumental inequality *[Pearl, 1995]. For a comprehensive treatment of this model using convex analysis please see [Bonet, 2001].*

5.3 Testability of G-MAR

In the literature, in general MCAR is considered testable [Little, 1988], while MAR is said to be not testable. However, when assumptions are made on the data generation model (such as compatible with a hypothesized m-graph G), MAR may become testable [Jaeger, 2005a, Jaeger, 2006b]. The testability of MAR+ has been studied in [Potthoff *et al.*, 2006, Mohan and Pearl, 2014].

The testability of G-MAR concerns with the question of whether there exist edges between a R variable and variables in $V_m \cup U$. In general, whether an edge is testable or not is sensible to the graphical structure, and can be judged by looking for testable implications in the enriched graph, for which many techniques have been developed (see the discussion in the beginning of Section 5).

Example 6 *The model in Figure 1(a) is MCAR. G^* encodes a testable independence $R_Y \perp\!\!\!\perp X$, which can be used to test whether there is an edge from X or Y to R_Y.*

Example 7 *The model in Figure 2(a) is G-MAR. G^* encodes no testable implications. Therefore G-MAR assumption is not testable.*

Example 8 *The model in Figure 5(a) is G-MAR. G^* encodes a testable independence $R_X \perp\!\!\!\perp Z$, which can be used to test whether there is an edge from X to R_X. G-MAR assumption is testable in this sense.*

6 Conclusions

MAR plays a central role in the theory underlying the current methods for handling missing data. Under the

MAR assumption, likelihood-based inference (as well as Bayesian inference) can be carried out while ignoring the missingness mechanism. However the standard difinition of MAR is difficult to apply in practice. We have proposed G-MAR as an alternative definition for MAR in graphical missing data models and established its relation with the standard MAR and a few versions of MAR used in the literature. G-MAR is intuitively appealing from the modeling point of the view and is easier to interpret and to apply in practice than the standard MAR. We have addressed the question of whether G-MAR is testable and whether the missing data model is testable, given that data are corrupted by missingness, by converting the problem into a well studied problem of identifying constraints in graphical models with latent variables. The results presented in this paper should be useful for practitioners in many fields since the missing data problem is common across many disciplines including artificial intelligence and machine learning, statistics, economics, and the health and social sciences.

Appendix - Graphoid Axioms

The probability distributions satisfy the following so-called *semi-graphoid axioms* [Pearl, 1988]:

- Symmetry
$$X \perp\!\!\!\perp Y|Z \iff Y \perp\!\!\!\perp X|Z$$
- Decomposition
$$X \perp\!\!\!\perp (Y \cup W)|Z \implies X \perp\!\!\!\perp Y|Z \quad \& \quad X \perp\!\!\!\perp W|Z$$
- Weak Union
$$X \perp\!\!\!\perp (Y \cup W)|Z \implies X \perp\!\!\!\perp Y|(Z \cup W)$$
- Contraction
$$X \perp\!\!\!\perp Y|Z \quad \& \quad X \perp\!\!\!\perp W|(Z \cup Y) \implies X \perp\!\!\!\perp (Y \cup W)|Z$$

The four axioms together with the following intersection axiom are known as the *graphoid axioms*.

- Intersection: if $P(Y, Z, W) > 0$,
$$X \perp\!\!\!\perp Y|(Z \cup W) \quad \& \quad X \perp\!\!\!\perp W|(Z \cup Y) \implies X \perp\!\!\!\perp (Y \cup W)|Z.$$

Note that in the literature the intersection axiom is said to hold only if the distribution P is strictly positive. However, if following its proof carefully, it can be shown that we only need the marginal distribution $P(Y, Z, W)$ to be strictly positive for intersection to hold (see, e.g., the proof of Proposition 3.1 in [Lauritzen, 1996]). We need this version of intersection in the proof of Proposition 1.

The intersection, decomposition, and contraction axioms imply the following rule that we will call "Generalized intersection rule".

Lemma 1 Generalized intersection rule*: if* $P(Y, Z, W) > 0$,
$$X \perp\!\!\!\perp (Y \cup S)|(Z \cup W) \quad \& \quad X \perp\!\!\!\perp (W \cup S)|(Z \cup Y) \implies X \perp\!\!\!\perp (Y \cup W \cup S)|Z.$$

Proof: $X \perp\!\!\!\perp (Y \cup S)|(Z \cup W)$ leads to $X \perp\!\!\!\perp Y|(Z \cup W)$, and $X \perp\!\!\!\perp (W \cup S)|(Z \cup Y)$ leads to $X \perp\!\!\!\perp W|(Z \cup Y)$ by decomposition. We then obtain $X \perp\!\!\!\perp (Y \cup W)|Z$ by intersection, and further $X \perp\!\!\!\perp Y|Z$ by decomposition. Now $X \perp\!\!\!\perp Y|Z$ and $X \perp\!\!\!\perp (W \cup S)|(Z \cup Y)$ lead to $X \perp\!\!\!\perp (Y \cup W \cup S)|Z$ by contraction. □

Acknowledgements

We thank the anonymous reviewers for valuable comments.

References

[Bonet, 2001] B. Bonet. Instrumentality tests revisited. In *Proc. 17th Conf. on Uncertainty in Artificial Intelligence*, pages 48–55, Seattle, WA, 2001. Morgan Kaufmann.

[Geiger and Meek, 1999] Dan Geiger and Christopher Meek. Quantifier elimination for statistical problems. In *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–99)*, pages 226–235, San Francisco, CA, 1999. Morgan Kaufmann Publishers.

[Gill *et al.*, 1997] Richard D Gill, Mark J Van Der Laan, and James M Robins. Coarsening at random: Characterizations, conjectures, counter-examples. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–294. Springer, 1997.

[Graham, 2012] John W Graham. *Missing data: Analysis and design*. Springer, 2012.

[Heitjan and Rubin, 1991] Daniel F Heitjan and Donald B Rubin. Ignorability and coarse data. *The Annals of Statistics*, pages 2244–2253, 1991.

[Jaeger, 2005a] Manfred Jaeger. Ignorability for categorical data. *Annals of statistics*, pages 1964–1981, 2005.

[Jaeger, 2005b] Manfred Jaeger. Ignorability in statistical and probabilistic inference. *Journal of Artificial Intelligence Research*, 24:889–917, 2005.

[Jaeger, 2006a] Manfred Jaeger. The ai&m procedure for learning from incomplete data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 225–232, 2006.

[Jaeger, 2006b] Manfred Jaeger. On testing the missing at random assumption. In *ECML*, pages 671–678. Springer, 2006.

[Kang and Tian, 2006] C. Kang and J. Tian. Inequality constraints in causal models with hidden variables. In *Proceedings of the Seventeenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 233–240, Arlington, Virginia, 2006. AUAI Press.

[Koller and Friedman, 2009] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[Lauritzen, 1996] S.L. Lauritzen. *Graphical Models*. Clarendon Press, Oxford, 1996.

[Little and Rubin, 2002] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. Wiley, 2002.

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

[Marlin and Zemel, 2009] Benjamin M Marlin and Richard S Zemel. Collaborative prediction and ranking with non-random missing data. In *Proceedings of the third ACM conference on Recommender systems*, pages 5–12. ACM, 2009.

[Marlin *et al.*, 2007] Benjamin Marlin, Richard S Zemel, Sam Roweis, and Malcolm Slaney. Collaborative filtering and the missing at random assumption. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[Marlin *et al.*, 2011] Benjamin M Marlin, Richard S Zemel, Sam T Roweis, and Malcolm Slaney. Recommender systems, missing data and statistical model estimation. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011.

[McKnight *et al.*, 2007] Patrick E McKnight, Katherine M McKnight, Souraya Sidani, and Aurelio Jose Figueredo. *Missing data: A gentle introduction*. Guilford Press, 2007.

[Mohan and Pearl, 2014] Karthika Mohan and Judea Pearl. On the testability of models with missing data. *Proceedings of AISTAT-2014*, 2014.

[Mohan *et al.*, 2013] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In *Advances in Neural Information Processing Systems (NIPS 2013)*, pages 1277–1285, 2013.

[Pearl, 1988] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[Pearl, 1995] J. Pearl. On the testability of causal models with latent and instrumental variables. In P. Besnard and S. Hanks, editors, *Uncertainty in Artificial Intelligence 11*, pages 435–443. Morgan Kaufmann, 1995.

[Pearl, 2000] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, NY, 2000.

[Potthoff *et al.*, 2006] Richard F Potthoff, Gail E Tudor, Karen S Pieper, and Vic Hasselblad. Can one assess whether missing data are missing at random in medical studies? *Statistical methods in medical research*, 15(3):213–234, 2006.

[Rubin, 1976] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[Schafer and Graham, 2002] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[Shpitser and Pearl, 2008] I. Shpitser and J. Pearl. Complete identification methods for the causal hierarchy. *Journal of Machine Learning Research*, 9:1941–1979, 2008.

[Tian and Pearl, 2002] J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2002.