On Testability and Goodness of Fit Tests in Missing Data Models

Razieh Nabi[1] **Rohit Bhattacharya**[2]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, Georgia, USA
[2]Department of Computer Science, Williams College, Williamstown, Massachusetts, USA

Abstract

Significant progress has been made in developing identification and estimation techniques for missing data problems where modeling assumptions can be described via a directed acyclic graph. The validity of results using such techniques rely on the assumptions encoded by the graph holding true; however, verification of these assumptions has not received sufficient attention in prior work. In this paper, we provide new insights on the testable implications of three broad classes of missing data graphical models, and design goodness-of-fit tests for them. The classes of models explored are: sequential missing-at-random and missing-not-at-random models which can be used for modeling longitudinal studies with dropout/censoring, and a no self-censoring model which can be applied to cross-sectional studies and surveys.

1 INTRODUCTION

Missing data is a common issue in applied problems. To infer a parameter of interest under missingness, often a statistical model is posed that encodes a set of assumptions on the missingness mechanisms. These assumptions are commonly divided into three main types: missing-completely-at-random (MCAR) where missingness does not have a cause and hence complete-case analysis is justifiable, missing-at-random (MAR) where all causes of missingness are assumed to be fully observed, and missing-not-at-random (MNAR) where causes of missingness are either only partially observed and/or fully unobserved [Little and Rubin, 2019].

MNAR models are perhaps the most common form of missingness in practice, and the most challenging since observations are systematically missing; yet such models are underused due to the complexity of the identification and estimation procedures needed to recover parameters of interest as functions of observed data law. A recent line of research has proposed to use causal graphical models as a representation of the statistical models for missing data [Mohan et al., 2013, Thoemmes and Rose, 2014, Shpitser, 2016, Saadati and Tian, 2019]. A causal graph not only encodes conditional independence relations between variables but also depicts the causal mechanisms responsible for missingness, making it a useful tool for interpretation of the underlying assumptions [Glymour, 2006, Daniel et al., 2012, Martel García, 2013, Scharfstein et al., 2021]. Further, just as in causal inference, graphical representations of missingness allow for the design of algorithms that automate certain steps of identification and estimation schemes; see Mohan and Pearl [2021], Nabi et al. [2022] for detailed reviews.

While advances in graphical models of missing data have yielded useful insights into identifying and estimating parameters of interest, the validity of any result relies on the substantive assumptions encoded by the graph holding true. In order to confirm testability of a restriction in missing data models, we have to examine its implications on the observed data distribution; this enables the design of empirical testing procedures from finite (but partially unobserved) samples. Unfortunately, we may not always be able to test all the encoded restrictions. The permutation model, proposed by Robins [1997], is an example of a graphical MNAR model that is untestable. Mohan and Pearl [2014] provided examples of other impediments for testability in graphical missing data models. For instance, in many cases, the assumption that no variable influences its own missingness (a.k.a. lack of self-censoring causes) is untestable.

Nonetheless, there are MAR and MNAR models that entail empirically testable restrictions [Mohan and Pearl, 2014, Tian, 2015, Gain and Shpitser, 2018, Tu et al., 2019]. The contributions of this paper in this regard are two-fold: (i) We expand on testable implications of missing data models that resemble ordinary conditional independencies in the underlying full law, but manifest as generalized a.k.a. Verma independencies in the observed law; (ii) We design empirical tests for restrictions in three broad classes of missing data

Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023), PMLR 216:1467–1477.

models that use ideas from weighted likelihood-ratio tests and odds-ratio parameterizations of joint distributions. The model classes are:

1. Sequential MAR models where missingness at each time step depends only on past observed values,
2. Sequential MNAR models where missingness at each time step may depend on past observed variables as well as future unmeasured/missing values, and
3. MNAR models where missingness of each variable may depend on missing values of any other variable except itself.

The first two classes are particularly applicable for modeling missingness mechanisms in longitudinal studies with censoring, while the third class is more suitable for survey studies and situations where there is no inherent time ordering in the data collection process. With these insights, we can use partially observed data to gain information about the underlying missingness mechanisms and assess the adequacy of our chosen missingness models in our analyses. Additionally, we explore the extension of our results to scenarios where some variables are completely unmeasured or latent, further broadening the scope of our framework. Our results are also relevant for discovery and model selection tasks where the goal is not only to uncover the substantive relationships between the variables of interest but also to identify the processes that drive their missingness.
All proofs are deferred to the supplementary materials.

2 NOTATIONS AND PRELIMINARIES

Let $X = \{X_1, \ldots, X_K\}$ be a set of K random variables with probability distribution $p(X)$. We denote the values of $X_k \in X$ by lower case letter $x_k \in \mathfrak{X}_k$, where $\mathfrak{X}_k$ denotes the state space of X_k. We assume a sample of n i.i.d observations with missing values. To locate the missing cases, we consider a set of binary missingness indicators $R = \{R_1, \ldots, R_K\}$, where $R_k = 0$ when x_k is missing and $R_k = 1$ when x_k is observed. Let $X^* = \{X_1^*, \ldots, X_K^*\}$ denote the set of proxy random variables that represent the values of variables in X that we actually observe. Each $X_k^* \in X^*$ is deterministically defined in terms of $R_k \in R$ and $X_k \in X$ as follows: if $R_k = 1, X_k^* = X_k$, otherwise $X_k^* =$ "?". We refer to $p(X)$ as the *target law*, $p(R \mid X)$ as the *missingness mechanism*, $p(X, R)$ as the *full law*, and $p(R, X^*)$ as the *observed data law.*

A missing data model is a set of distributions defined over variables in $\{X, R, X^*\}$.[1] Following the conventions in Mohan et al. [2013], we represent the missing data model via a directed acyclic graph (DAG) $\mathcal{G}(V)$, where vertices V correspond to random variables in $\{X, R, X^*\}$. In addition to acyclicity, a missing data DAG (or m-DAG[2] for short) imposes certain restrictions on the edges: variables in R cannot point to variables in X, and each $X_k^* \in X^*$ has only two parents: X_k and R_k (due to deterministic relations.) Similar to Bhattacharya et al. [2019], we also allow for $X_i^* \to R_j$ edges. A few examples of m-DAGs are illustrated in Sections 3 and 4; edges corresponding to deterministic relations are drawn in gray. A full law $p(X, R, X^*)$ that is Markov relative to $\mathcal{G}(V)$ factorizes as follows:

$$\prod_{V_i \in X \cup R} p(V_i \mid \mathrm{pa}_{\mathcal{G}}(V_i)) \times \prod_{X_k^* \in X^*} p(X_k^* \mid R_k, X_k), \quad (1)$$

where $\mathrm{pa}_{\mathcal{G}}(V_i)$ denotes the parents of V_i in $\mathcal{G}(V)$. For convenience, we drop the deterministic terms, $p(X_k^* | R_k, X_k)$, when discussing the factorization of the full law.

The full law $p(X, R)$ is identified (can be expressed as a function of the observed data) if and only if the missingness mechanism $p(R|X)$ is identified; the target law $p(X)$ is identified if and only if $p(R = 1|X)$ is identified. Thus, identification of the full law implies that the target law (and any function of the full law) is identified, but the reverse is not true. The missingness mechanism in an m-DAG factorizes as $\prod_k p(R_k | \mathrm{pa}_{\mathcal{G}}(R_k))$, where the conditional density $p(R_k | \mathrm{pa}_{\mathcal{G}}(R_k))$ is referred to as the *propensity score* of R_k.

Numerous identification strategies in the field of graphical missing data literature focus on identifying each propensity score in a specific order, whether it is a total or partial order; notable works in this area include Shpitser et al. [2015] and Bhattacharya et al. [2019]. In essence, to identify the propensity score of R_k, one can verify if R_k, given parents, is independent of the corresponding missingness indicators of its parents that are counterfactuals. If this condition holds, the propensity score can be identified using a simple argument based on conditional independence (d-separation). However, if this condition is not satisfied, one needs to examine whether it holds in post-fixing distributions, which are obtained through the recursive application of the *fixing* operator. This operation involves inverse weighting the current distribution by the propensity score of the variable being fixed [Bhattacharya et al., 2022, Richardson et al., 2023]. We employ similar strategies in this work.

Similar to regular DAGs, absence of an edge in a missing data DAG $\mathcal{G}(V)$ entails conditional independence restrictions between the endpoint variables in the underlying distribution $p(V)$. These restrictions can be directly read off from the graph using Markov properties and d-separation rules [Pearl, 2009] – given disjoint sets $U, W, Z \subset V$, the global Markov property states that if $U \perp\!\!\!\perp_{\text{d-sep}} W \mid Z$ in $\mathcal{G}(V)$, then $U \perp\!\!\!\perp W \mid Z$ in $p(V)$. In this work, we focus on

[1] For simplicity of notations, we assume all variables have missing values. All discussions however, can be easily generalized to scenarios where a subset of variables are fully observed.

[2] The term "mDAG" has also been used by [Evans, 2016] to denote marginalized DAGs.

restrictions where all variables are at least partially observed, which allows us to narrow our focus to testability of ordinary conditional independence restrictions in the full law. However, as we will see, even ordinary restrictions in the full law may manifest as generalized equality restrictions in the observed law. Testability of generalized equality restrictions induced by latent variables in m-DAGs is a challenging problem left for future work; see Tian and Pearl [2002], Shpitser and Pearl [2006], Bhattacharya et al. [2022] for more details on such restrictions when there is no missingness.

Unlike regular DAGs where the independence constraints can be tested using observed samples from the joint distribution, a conditional independence restriction in an m-DAG might be empirically untestable, or may manifest as more complex restrictions on the observed data law. If all the restrictions encoded in an m-DAG are provably untestable (i.e., no restriction on the observed data law), the full law Markov relative to the m-DAG is said to be *nonparametric saturated* (as defined by Robins [1997]). Nonetheless, submodels of saturated missing data models may still be testable. In this paper, we discuss testability of assumptions in the three aforementioned classes of missing data models as submodels of two known saturated models: the *permutation* model and the *no self-censoring* model.

Robins [1997] introduced the permutation model as follows: given an ordering on variables in X, indexed by $k \in \{1, \ldots, K\}$, each missingness indicator R_k is independent of the current and past variables in X given the past observed variables in R, X^* and future variables in X. Formally, the model is defined via the following set of conditional independence restrictions:

$$R_k \perp\!\!\!\perp X_{\prec k+1} \mid R_{\prec k}, X^*_{\prec k}, X_{\succ k}, \forall k \quad \textit{(permutation)} \quad (2)$$

where $V_{\prec k} = \{V_1, \ldots, V_{k-1}\}, V_{\succ k} = \{V_{k+1}, \ldots, V_K\}$. Robins [1997] showed that the full law in this model is identified and is nonparametrically saturated. An m-DAG representation of the permutation model with $K = 2$ variables is shown in Fig. 1(b). For a discussion on the substantive distinctions between $X_1^* \to R_2$ and $X_1 \to R_2$ edges, refer to Appendix A.3.

The no self-censoring model was introduced by Shpitser [2016], Sadinle and Reiter [2017].[3] The central assumption in this model is that no variable directly causes its own missingness status. Formally, the model is defined by the following set of conditional independence restrictions:

$$R_k \perp\!\!\!\perp X_k \mid R_{-k}, X_{-k}, \forall k \quad \textit{(no self-censoring)} \quad (3)$$

where $V_{-k} = V \setminus V_k$. Malinsky et al. [2021] showed that this model is nonparametrically saturated and identified via an odds-ratio parameterization of the missingness mechanism; a description of this parameterization, which was proposed by Chen [2007], is provided in Appendix A.1. The graphical representation of this model relies on a generalization of m-DAGs to allow for undirected edges between all pairs of R vertices – a graph with both directed and undirected edges is called a chain graph [Lauritzen, 1996, Shpitser, 2016]. An example of this model with $K = 2$ variables is shown in Fig. 4(b). The assumptions of the no self-censoring model are encoded in this chain graph by the following local Markov property: each missingness indicator R_i is independent of all other variables on the graph given its neighboring missingness indicators (joined via an undirected edge $R_i - R_j$) and its parents ($X_j \to R_i$).

We now explore how one can select an appropriate m-DAG representation based on intuitive explanations of the missing data generation process. To illustrate this, imagine an investigator who is analyzing a large observational database that contains information on smoking habits and diagnostic test results for bronchitis among individuals in a city. In this scenario, let's consider X_1 as the true smoking status of an individual and X_2 as their bronchitis diagnosis. However, the investigator notices that there are missing entries in the database, which are indicated by the missingness indicators R_1 and R_2. If we choose the permutation model depicted in Figure 1(b) to explain the missingness mechanism, it implies the following two processes: (i) $R_1 \leftarrow X_2$ suggesting that the measurement of an individual's smoking status depends on the counterfactual value of their bronchitis status; this may occur for example when a patient's smoking status is inquired on a suspected diagnosis of bronchitis before administering the test., and (ii) $R_1 \to R_2 \leftarrow X_1^*$ suggesting that whether the true bronchitis status is measured via a diagnostic test depends on the doctor's awareness of the individual's smoking status (R_1) and their observed value of smoking (X_1^*). On the other hand, the no self-censoring model shown in Fig. 4(b) explains the missingness mechanism via the following three processes: (i) $R_1 \leftarrow X_2$ suggesting that a suspected diagnosis of bronchitis is likely to lead to an inquiry about the smoking status of the patient, (ii) $R_2 \leftarrow X_1$ suggesting that smokers are more likely to get tested for bronchitis, and (iii) $R_1 - R_2$ suggesting that ordering a diagnostic test for bronchitis increases the likelihood of ordering a test for bronchitis, and vice versa.

3 NEW INSIGHTS INTO TESTABLE IMPLICATIONS

Although the restrictions we study in this paper can be phrased in terms of ordinary independence restrictions in the full law of a missing data DAG model, they may only manifest in the observed data law via relatively complex functionals. In this section, we show that a d-separation statement between missingness indicators and substantive variables may correspond to generalized equality constraints, a.k.a. Verma constraints [Verma and Pearl, 1990], in the

[3] In Sadinle and Reiter [2017], the model is referred to as *itemwise conditionally independent nonresponse* model.

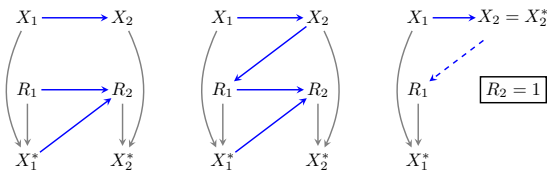

Figure 1: (a) Example of a MAR model; (b) Example of a saturated permutation model; (c) The absence of $X_2 \to R_1$ edge in (a) can be tested in the intervention distribution $p(V \setminus R_2|\text{do}(R_2 = 1))$ where the dashed edge indicates whether $p(V)$ is Markov wrt the MAR model in (a) or the permutation supermodel in (b).

observed data distribution. This observation extends the current state of the art on testability in missing data models.

Consider the m-DAG shown in Fig. 1(a): a MAR submodel of the permutation model in Fig. 1(b) where $X_2 \to R_1$ is removed. Though the permutation model itself is nonparametric saturated, it is natural to ask if this MAR submodel, which encodes an additional d-separation relation $R_1 \perp\!\!\!\perp X_2$, has a testable restriction on the observed data. To determine testability, we initially apply the criterion proposed by Mohan and Pearl [2014]. According to their criterion, a d-separation condition displayed in an m-DAG $\mathcal{G}$ is testable if the missingness indicators associated with all partially observed variables involved in the relation are either already present in the separating set, or can be added to the set without spoiling the separation. By applying this criterion, we note that the relation $R_1 \perp\!\!\!\perp_{\text{d-sep}} X_2|R_2$ does *not* hold in $\mathcal{G}$ due to the open collider R_2 on the path $R_1 \to R_2 \leftarrow X_1^* \leftarrow X_1 \to X_2$. Therefore, one might conclude that $R_1 \perp\!\!\!\perp X_2$ is not testable.

Let us momentarily assume that X in Fig. 1(a) consists of binary variables. We can compare number of parameters in the full law using (1) against the saturated observed data law using pattern-mixture factorization Rubin [1976] given by the marginal distribution of R and conditional distribution of X^* given R. The full law in Fig. 1(a) requires 7 parameters (3 for $p(X)$, 1 for $p(R_1)$, and 3 for $p(R_2|R_1, X_1^*)$) which is less than the number of parameters in the saturated observed law which is 8 (3 for $p(R)$ and 5 for $p(X^*|R)$.) Hence, it can be inferred that the additional restriction $R_1 \perp\!\!\!\perp X_2$ (implied by the absence of $X_2 \to R_1$ edge) imposes constraints on the observed data law, at least in the discrete case.[4] Interestingly, this contradicts the conclusion from the previous paragraph stating that $R_1 \perp\!\!\!\perp X_2$ is not testable. Such contradictions are expected, however, as Mohan and Pearl's criterion is sufficient but not necessary for testability.

Now, we aim to demonstrate that $R_1 \perp\!\!\!\perp X_2$ is indeed testable, not only in discrete cases but also in more general nonparametric settings. In Fig. 1(a), conditioning on R_2 opened up the collider $R_1 \to R_2 \leftarrow X_1^*$ on the path from R_1 to X_2. From a causal perspective, removal of these edges corresponds to an intervention in which R_2 is set to a specific value.[5] This results in the m-*conditional* DAG (m-CDAG), as shown in Fig. 1(c). Due to determinism (alternatively consistency), once R_2 is set to 1, then $X_2 = X_2^*$ forming a single (observed) node. Following standard notation in Pearl [2009] we denote the intervention where we set R_2 to 1 as $\text{do}(R_2 = 1)$ and the corresponding intervention distribution as $p(X, R \setminus R_2, X^*|\text{do}(R_2 = 1))$, or $p(.|\text{do}(R_2 = 1))$ for short. This intervention distribution can be obtained via truncation of the full law factorization where the propensity score of R_2, $p(R_2|\,\text{pa}_{\mathcal{G}}(R_2))$, is dropped. That is, $p(.|\text{do}(R_2 = 1)) = p(X, R, X^*)/p(R_2|R_1, X_1^*)|_{R_2=1}$ and it factorizes according to the m-CDAG shown in Fig. 1(c). The relation $R_1 \perp\!\!\!\perp_{\text{d-sep}} X_2$ holds in the resulting m-CDAG, and X_2 is now fully observed. Further, the propensity score of R_2 that takes us to the intervention distribution is a function of observed data. These facts combined imply that $R_1 \perp\!\!\!\perp X_2$ imposes a restriction on the observed data in the form of a Verma constraint; i.e., a d-separation statement in an identified intervention distribution.

The above example illustrates the core idea for essential extensions of the previous testability criterion [Mohan and Pearl, 2014, 2021]. We state that a d-separation condition displayed in an m-DAG is also testable if the missingness indicators associated with all partially observed variables involved in the relation can be intervened on (or, in other words, their corresponding propensity scores are identified) without spoiling the separation. We formalize this extension of testable restrictions in the next section (and partly in the appendix), where we also consider testability of independence statements between proxy variables and missingness indicators, and among missingness indicators themselves.

4 TESTABLE IMPLICATIONS AND GOODNESS-OF-FIT TESTS

In this section, we investigate independence assumptions in the full law and their implications on the observed data law in three broad classes of missing data models, and provide ways of empirically evaluating these constraints. We formulate the testability criteria and goodness-of-fit tests for the general case of a missing data model with K variables and illustrate the steps via examples. We consider likelihood-ratio tests for evaluating the independence $A \perp\!\!\!\perp B \mid C$, which is typically performed by fitting $p(A \mid C)$ and $p(A \mid B, C)$ and comparing their goodness-of-fit. Under the null hypothesis of independence, both models should fit the data equally well. In addition to likelihood-ratio tests, we consider evaluating $A \perp\!\!\!\perp B \mid C$ by computing the odds ratio of A and B

[4] Appendix A.2 contains the parameter counting arguments.

[5] We borrow the notion of intervention from causal inference by viewing each missingness indicator as a "treatment variable"; see [Nabi et al., 2022] for details.

conditioned on C – the independence relation holds *if and only if* the odds ratio equals one for all values of A, B, C. Therefore, if under the alternative hypothesis of dependence, the odds ratio still equals one (with statistical significance-level α), then the data agrees with the independence relation. In the following discussion, we let $\mathcal{M}_o$ denote the statistical model where the independence relation holds (the *null* hypothesis), and let $\mathcal{M}_a$ denote the statistical supermodel where the independence relation does not hold (the *alternative* hypothesis).

4.1 SEQUENTIAL MAR MODELS

We call a missing data model a *sequential MAR* model if under an ordering $\prec$ that indexes variables by $k = 1, \ldots, K$, the following set of independence restrictions hold:

$$R_k \perp\!\!\!\perp X \mid R_{\prec k}, X^*_{\prec k}, \forall k \qquad \textit{(sequential-MAR)} \quad (4)$$

Examples of this model are shown in Fig. 1(a) and Fig. 2(a) (without the dashed edges). In addition to restrictions of a permutation model described in (2), the sequential MAR model assumes $R_k \perp\!\!\!\perp X_{\succ k} | R_{\prec k}, X^*_{\prec k}, \forall k$ (it is straightforward to see this using graphoid axioms; see e.g., [Lauritzen, 1996] for description of the axioms). Thus, we can view the sequential MAR model as a submodel of the permutation model. Since assumptions imposed by the permutation model alone are untestable, we focus on testable implications of these extra assumptions and propose ways to empirically evaluate them.

The independence $R_k \perp\!\!\!\perp X_{\succ k} | R_{\prec k}, X^*_{\prec k}$ would be easily testable using observed data if we could add $R_{\succ k} = 1$ in the conditioning set and thus evaluate the restriction using only observed cases of $X_{\succ k}$. Unfortunately, the independence no longer holds if we condition on $R_{\succ k}$ (this is easily confirmed from discussion in the previous section and Fig. 2(a).) However, we can instead intervene on $R_{\succ k}$ and check if the independence holds in the intervention distribution. The following theorem formalizes that restrictions in sequential MAR models defined above can always be tested as Verma constraints, i.e., (i) the independence holds in the corresponding m-CDAG, and (ii) the required intervention distributions are identified from observed data.

Theorem 1. *The independence $R_k \perp\!\!\!\perp X_{\succ k} | R_{\prec k}, X^*_{\prec k}$ has a testable implication on the observed data distribution in the form of a Verma constraint: $R_k \perp\!\!\!\perp X_{\succ k} | R_{\prec k}, X^*_{\prec k}, do(R_{\succ k} = 1)$, where the intervention distribution $p(X, R \setminus R_{\succ k}, X^* | do(R_{\succ k} = 1))$ is identified.*

The intuition for this result will become clear as we discuss testing such constraints using n (finite) i.i.d samples (denoted by $\mathcal{D}_n$). One possibility is to use a likelihood-ratio test and compare goodness-of-fits between $p(R_k | R_{\prec k}, X^*_{\prec k})$ and $p(R_k | R_{\prec k}, X^*_{\prec k}, X_{\succ k})$ **but** with respect to a distribution where $R_{\succ k}$ are intervened on and set to 1. This intervention distribution is a truncated factorization of the full law where propensity scores of $R_{\succ k}$ are dropped, i.e.,

$$p(. \mid \text{do}(R_{\succ k} = 1)) = \frac{p(V)}{\prod_{j \succ k} p(R_j \mid \text{pa}_{\mathcal{G}}(R_j))}\Big|_{R_{\succ k}=1}.$$

Let $W_k(\beta^o_k) := p(R_k | R_{\prec k}, X^*_{\prec k}; \beta^o_k)$ and $W_k(\beta^a_k) := p(R_k | R_{\prec k}, X^*_{\prec k}, X_{\succ k}; \beta^a_k)$ (the null and alternative respectively.) Estimating β^o_k is relatively straightforward as $W_k(\beta^o_k)$ is a direct function of observed data, but estimating β^a_k is more involved. We propose to estimate β^a_k, wrt the truncated/weighted distribution above. This entails using a weighted estimating equation where propensity scores of $R_{\succ k}$ are used as inverse weights to fit β^a_k. It is important to note however, that a propensity score $p(R_j \mid R_{\prec j}, X^*_{\prec j}, X_{\succ j})$ for any $R_j \in R_{\succ k}$, may itself need to be fitted via a weighted estimating equation, since $X_{\succ j}$ appears in the conditioning set and $R_j \not\perp\!\!\!\perp R_{\succ j} | R_{\prec j}, X^*_{\prec j}, X_{\succ j}$.

As an example, consider the sequential MAR model in Fig. 2(a) (without the dashed edges). The null hypothesis $\mathcal{M}_o$ is the statistical model of this m-DAG and the alternative $\mathcal{M}_a$ is the permutation supermodel with the dashed edges. We are interested in evaluating the independencies $R_1 \perp\!\!\!\perp X_2, X_3$ and $R_2 \perp\!\!\!\perp X_3 | R_1, X^*_1$, which given Theorem 1 translates into independence restrictions in $p(.|\text{do}(R_2 = 1, R_3 = 1))$ and $p(.|\text{do}(R_3 = 1))$, respectively. Testing $R_1 \perp\!\!\!\perp X_2, X_3$ entails fitting $W_{r_1}(\beta^a_{r_1}) := p(R_1 | X_2, X_3; \beta^a_{r_1})$ wrt the truncated/weighted factorization Markov relative to Fig. 2(b) where R_2 and R_3 are intervened on and set to 1. Thus, we can use propensity scores of R_2 and R_3 as inverse weights to estimate $\beta^a_{r_1}$. Let $\mathbb{P}_n[U(\beta^a_{r_1})] = 0$ be an unbiased estimating equation for $\beta^a_{r_1}$ wrt the full law ($\mathbb{P}_n[.] = \frac{1}{n}\sum_{i=1}^n (.)$). In other words, $\mathbb{P}_n[U(\beta^a_{r_1})]$ is any estimating equation that is unbiased for $\beta^a_{r_1}$ had there been no missingness. The following *weighted* estimating equation then yields an unbiased estimator for $\beta^a_{r_1}$ wrt the observed data law:

$$\mathbb{P}_n\left[\frac{R_2 \times R_3}{p(R_2 \mid \text{pa}_{\mathcal{G}}(R_2)) \times p(R_3 \mid \text{pa}_{\mathcal{G}}(R_3))} \times U(\beta^a_{r_1})\right] = 0,$$

where propensity score of R_3, $p(R_3 | R_1, R_2, X^*_1, X^*_2)$, can be fit using just observed data, denote it with $W_{r_3}(\widehat{\beta}_{r_3})$. However, fitting the propensity score of R_2, $p(R_2 | R_1, X^*_1, X_3)$, requires an intermediate step involving the intervention distribution where R_3 is intervened on and set to 1, i.e., $p(X, R, X^*)/p(R_3 | \text{pa}_{\mathcal{G}}(R_3))$ evaluated at $R_3 = 1$. Similar to the above logic, this entails a weighted estimating equation using $W_{r_3}(\widehat{\beta}_{r_3})$ as inverse weights to fit the propensity score of R_2, denoted by $W_{r_2}(\widehat{\beta}_{r_2})$. Now that we have a way of estimating $\beta^a_{r_1}$, we can test $R_1 \perp\!\!\!\perp X_2, X_3$ using a weighted likelihood-ratio by computing

$$\rho = n\mathbb{P}_n\left[\frac{R_2 \times R_3}{W_{r_2}(\widehat{\beta}_{r_2}) \times W_{r_3}(\widehat{\beta}_{r_3})} \times \log\Big(\frac{W_{r_1}(\widehat{\beta}^a_{r_1})}{W_{r_1}(\widehat{\beta}^o_{r_1})}\Big)\right],$$

where $W_{r_1}(\beta^o_{r_1}) := p(R_1; \beta^o_{r_1})$ and $\beta^o_{r_1}$ is simply the proportion of complete cases of X_1, we can use likelihood chi-

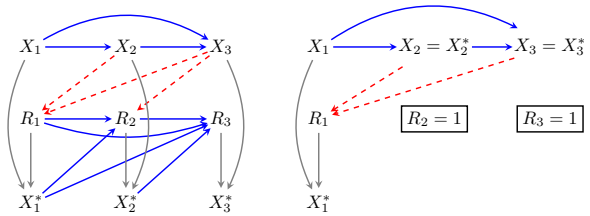

Figure 2: (a) Example of a sequential MAR model (without the dashed edges) along with its permutation supermodel (with the dashed edges); (b) The graph Markov wrt the intervention distribution $p(.|\text{do}(R_2 = 1, R_3 = 1))$.

square or Wald tests to compare goodness-of-fits [Robins and Wasserman, 1997, Agostinelli and Markatou, 2001].

If we start our tests for the sequential MAR model by testing the restriction $R_2 \perp\!\!\!\perp X_3 | R_1, X_1^*$, there are two possibilities: (i) The null might be rejected which immediately implies that the missing data model is not sequential MAR; (ii) The null is accepted which means R_2 does not have X_3 as a cause. In future tests, say for $R_1 \perp\!\!\!\perp X_2, X_3$ in this case, this justifies fitting a simplified propensity score $p(R_2|R_1, X_1^*, X_3^*)$ that makes use of all the observed data. This simplified propensity score also corresponds to the same model we would have already fit for the previous null hypothesis $\beta^o_{r_2}$. This example reveals that there is a natural way to order the tests. For a model with K variables, we would proceed backwards by first testing restrictions involving R_{K-1}, moving to R_{K-2}, and so on. If the current test succeeds, the corresponding model for the null can be re-used to produce weights for future estimating equations; if the test fails, then the assumptions of sequential MAR does not hold. Following such a sequence may help improve the power of each test by using all of the observed samples to estimate the weights in each step. We formalize this sequence of goodness-of-fit tests based on weighted likelihood-ratios in Algorithm 1, which takes an ordering $\prec$ on the missingness indicators, null and alternative models as a tuple $\mathcal{M}$, and data samples $\mathcal{D}_n$ as input. The k^{th} iteration of the for loop concerns testing the independence $R_k \perp\!\!\!\perp X_{\succ k} \mid R_{\prec k}, X^*_{\prec k}$. Note however, that as we proceed with the tests, we are restricted to fewer and fewer samples which impacts the power of our tests. Although weighting approaches are common in missing data models [Li et al., 2013], an interesting direction for future work is to develop semiparametric methods to use data more efficiently.

4.2 SEQUENTIAL MNAR MODELS

We call a missing data model a *sequential MNAR* model if under an ordering $\prec$ that indexes variables by $k = 1, \ldots, K$, the following set of independence restrictions hold:

$$R_k \perp\!\!\!\perp X_{\prec k+1}, X^*_{\prec k} \mid R_{\prec k}, X_{\succ k}, \forall k \quad \textit{(sequential-MNAR)} \quad (5)$$

Algorithm 1 TESTING SEQUENTIAL MAR $(\prec, \mathcal{M}, \mathcal{D}_n)$

1: Let $\prec$ index variables by $k = 1, \ldots, K$.
2: Let $W_K(\beta^o_K) \coloneqq p(R_K|R_{\prec K}, X^*_{\prec K}; \beta^o_K)$.
3: Estimate β^o_K (denote it by $\widehat{\beta}^o_K$).
4: **for** $k \in \{K-1, \ldots, 1\}$ **do**
5: Let $W_k(\beta^o_k) \coloneqq p(R_k|R_{\prec k}, X^*_{\prec k}; \beta^o_k)$ and $W_k(\beta^a_k) \coloneqq p(R_k|R_{\prec k}, X^*_{\prec k}, X_{\succ k}; \beta^a_k)$.
6: Estimate β^o_k (denote it by $\widehat{\beta}^o_k$).
7: Estimate β^a_k via the weighted estimating equation:

$$\mathbb{P}_n\left[\frac{\mathbb{I}(R_{\succ k} = 1)}{\prod_{j \succ k}^K W_j(\widehat{\beta}^o_j)} \times U(\beta^a_k)\right] = 0,$$

where $\mathbb{P}_n\big[U(\beta^a_k)\big] = 0$ is an unbiased estimating equation for β^a_k wrt the full law (denote it by $\widehat{\beta}^a_k$).
8: Compute a weighted likelihood-ratio as follows:

$$\rho = n\mathbb{P}_n\left[\frac{\mathbb{I}(R_{\succ k} = 1)}{\prod_{j \succ k}^K W_j(\widehat{\beta}^o_j)} \times \log\Big(\frac{W_k(\widehat{\beta}^a_k)}{W_k(\widehat{\beta}^o_k)}\Big)\right].$$

9: Test ρ with α significance level.
10: **if** $\mathcal{M}_o$ is rejected (i.e., $R_k \not\perp\!\!\!\perp X_{\succ k}|R_{\prec k}, X^*_{\prec k}$) **then**
11: **return** not sequential MAR
12: **return** sequential MAR

An example of this model is shown in Fig. 3(a) (without the dashed edges.) We can view the sequential MNAR model as a submodel of the permutation model since in addition to the restrictions in (2), it assumes $R_k \perp\!\!\!\perp X^*_{\prec k} \mid R_{\prec k}, X_{\succ k}, \forall k$. Thus, we focus on testable implications of these extra assumptions and propose ways to empirically evaluate them.

Unlike sequential MAR models, the d-separation statements being tested in sequential MNAR models are between missingness indicators and proxy variables, which can be viewed as context-specific restrictions. Due to determinism, when $R_j = 0, X^*_j =$ "?", an independence restriction such as $R_k \perp\!\!\!\perp X^*_j | R_j = 0$ becomes a statement of independence between a random variable R_k and some constant, which is trivially true. Hence, the set $R_k \perp\!\!\!\perp X^*_{\prec k}|R_{\prec k}, X_{\succ k}, \forall k$ is equivalent to context-specific restrictions $R_k \perp\!\!\!\perp X_{\prec k}|R_{\prec k} = 1, X_{\succ k}, \forall k$; note that $R_{\prec k}$ is evaluated at one. Even though, these independences restrict us to rows where $X_{\prec k}$ is fully observed, we still need enough assumptions to plug in $R_{\succ k} = 1$ in the conditioning set, since $X_{\succ k}$ is in the conditioning set. Unfortunately, the independence between R_k and $X^*_{\prec k}$ no longer holds if we condition on $R_{\succ k}$. However, the following theorem formalizes that restrictions in sequential MNAR models defined above can still be tested as Verma constraints in identified intervention distributions where $R_{\succ k}$ are intervened and $X_{\succ k}$ are fully observed.

Theorem 2. *The independence* $R_k \perp\!\!\!\perp X^*_{\prec k}|R_{\prec k}, X_{\succ k}$

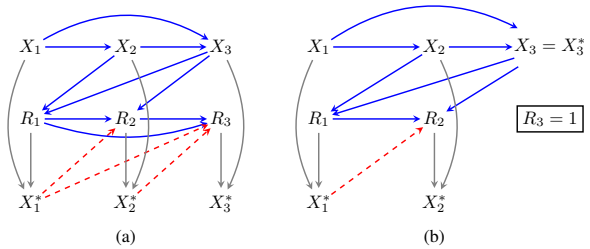

Figure 3: (a) Example of a sequential MNAR model (without the dashed edge) along with its permutation supermodel (with the dashed edge); (b) The graph Markov wrt the intervention distribution $p(.|\text{do}(R_3 = 1))$.

has a testable implication on the observed data distribution in the form of a Verma constraint $R_k \perp\!\!\!\perp X^*_{\prec k} | R_{\prec k}, X_{\succ k}, do(R_{\succ k} = 1)$, *where the intervention distribution* $p(X, R \setminus R_{\succ k}, X^* | do(R_{\succ k} = 1))$ *is identified.*

To evaluate these Verma constraints, we use weighted likelihood-ratio tests again. We explain this via a sequential MNAR example in Fig. 3(a) (without the dashed edges.) $\mathcal{M}_o$ is the statistical model of this DAG and $\mathcal{M}_a$ is the permutation supermodel with the dashed edges. We are interested in testing absence of the dashed edges which imply $R_3 \perp\!\!\!\perp X_1^*, X_2^* \mid R_1, R_2$ and $R_2 \perp\!\!\!\perp X_1^* \mid R_1, X_3$. To empirically evaluate the first restriction, we need to compare $p(R_3 \mid R_1, R_2, X_1^*, X_2^*)$ and $p(R_3 \mid R_1, R_2)$, which is straightforward since these two models are direct functions of observed data. To evaluate the second restriction however, we need to compare $p(R_2|X_1^*, R_1, X_3)$ and $p(R_2|R_1, X_3)$ wrt the intervention distribution $p(.|\text{do}(R_3 = 1)$, which corresponds to the truncated factorization $p(X, R, X^*)/p(R_3|R_1, R_2, X_1^*, X_3^*)$ (evaluated at $R_3 = 1$) and is Markov relative to the graph in Fig. 3(b). Thus, we can use $p(R_3|R_1, R_2, X_1^*, X_2^*)$ as inverse weights to fit models wrt this truncated distribution. Let $W_{r_2}(\beta^a_{r_2}) := p(R_2|R_1, X_1^*, X_3; \beta^a_{r_2})$ and let $\mathbb{P}_n[U(\beta^a_{r_2})] = 0$ be an unbiased estimating equation for $\beta^a_{r_2}$ wrt the full law. We can estimate $\beta^a_{r_2}$ using observed data via this weighted estimating equation: $\mathbb{P}_n[\{R_3/p(R_3|R_1, R_2, X_1^*, X_2^*; \widehat{\eta})\} \times U(\beta^a_{r_2})] = 0$, where $\widehat{\eta}$ is the estimated parameters for $p(R_3|R_1, R_2, X_1^*, X_2^*)$. Following the same logic, we can also estimate $\beta^o_{r_2}$ in $W_{r_2}(\widehat{\beta}^o_{r_2}) := p(R_2|R_1, X_3; \beta^o_{r_2})$. Finally, we use the following statistic in a weighted likelihood-ratio to test the restriction $R_2 \perp\!\!\!\perp X_1^* \mid R_1, X_3$:

$$\rho = n\mathbb{P}_n\left[\frac{R_3}{p(R_3 \mid R_1, R_2, X_1^*, X_2^*; \widehat{\eta})} \times \log\left(\frac{W_{r_2}(\widehat{\beta}^a_{r_2})}{W_{r_2}(\widehat{\beta}^o_{r_2})}\right)\right].$$

If we test the restriction $R_3 \perp\!\!\!\perp X_1^*, X_2^* | R_1, R_2$ first and conclude that the independence holds, we can use the R_3 fitted propensity score under the accepted null, that is $p(R_3 \mid R_1, R_2; \widehat{\beta}^o_{r_3})$, in above (without conditioning on X_1^*, X_2^*). This implies that for testing $R_2 \perp\!\!\!\perp X_1^* \mid R_1, X_3$, we do not have to use the full permutation model as a supermodel. Instead, we can use the permutation model where the $\{X_1^*, X_2^*\} \to R_3$ edges are absent.

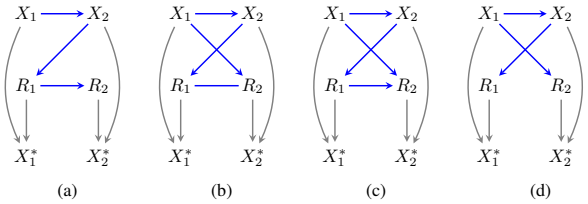

Figure 4: The sequential MNAR model in (a) can be tested as a submodel of the saturated no self-censoring model in (b); (c) A criss-cross supermodel of (a) where the test statistic is not identifiable; (d) Example of a block-parallel MNAR model which can be tested as a submodel of (b).

Algorithm 1 in Appendix B.1 provides an automated procedure for performing sequential goodness-of-fit tests based on weighted likelihood-ratios for $K > 3$ variables. The algorithm is similar to testing sequential MAR, but due to space limits, it is defferred to the supplements.

Remark 1. It is worth pointing out that the sequential MNAR model is a special case of models Markov relative to m-DAGs with no *colluders* studied in Nabi et al. [2020] – a colluder exists at R_j if there exists $X_i \in X \setminus X_j$ such that $X_i \to R_j \leftarrow R_i$. Nabi et al. [2020] showed that under the absence of colluder structures and self-censoring edges ($X_k \to R_k$), the full law Markov relative to such an m-DAG is identified. Further, they showed that such m-DAGs are a submodel of the saturated no self-censoring model defined in (3). Thus, it is possible to use the no self-censoring model as an alternative supermodel to test some of the restrictions in the sequential MNAR model. Namely, we can empirically evaluate this set of restrictions: $R_k \perp\!\!\!\perp X_{\prec k} \mid R_{-k}, X_{\succ k}$.

As an example, consider the absence of an edge between X_1 and R_2 in Fig. 4(a) which implies $R_2 \perp\!\!\!\perp X_1 | R_1$. The no self-censoring supermodel is drawn in Fig. 4(b) (with R_1, R_2 edge undirected). We can evaluate this independence by showing $p(R_2|R_1, X_1)$ is not a function of X_1. See Appendix B.2 for details on how to set up such a test.

Remark 2. The m-DAG in Fig. 4(c) is also a supermodel of Fig. 4(a). However, we cannot use it to evaluate the independence $R_2 \perp\!\!\!\perp X_1 \mid R_1$, because $p(R_2|R_1, X_1)$ is not fully identified under this supermodel, due to the colluder structure at R_2 as shown by Bhattacharya et al. [2019].

We call the m-DAG in Fig. 4(c), the *criss-cross* structure. In the following theorem, we show that unlike the permutation and no self-censoring models, the target law (and thus the full law) is not identified when such structures are present.

Theorem 3. *The target law* $p(X)$ *is not identified in an m-DAG model where there exists at least one criss-cross structure between a pair of variables.*

The above result characterizes a novel graphical structure that impedes target law identification; this may lead to fur-

ther insights on an open problem regarding the discovery of a *sound* and *complete* algorithm for target law identification.

Remark 3. As an alternative to the likelihood-ratio test, we can compute odds ratios to perform independence tests. For instance, in the MAR model of Fig. 1(a), $R_1 \perp\!\!\!\perp X_2$ translates into $\text{OR}(R_1, X_2) = 1$, and in the MNAR model of Fig. 4(a), $R_2 \perp\!\!\!\perp X_1^* | R_1$ translates into $\text{OR}(R_2, X_1 | R_1 = 1) = 1$, which both can be empirically evaluated. See Appendices C.2 and C.3 for a generalization of the idea of using odds ratio for goodness-of-fit tests in the sequential MAR and MNAR models with $K > 2$ variables.

4.3 BLOCK PARALLEL MNAR MODELS

We call a missing data model a *block-parallel MNAR* model[6] if it satisfies the following set of independence restrictions:

$$R_k \perp\!\!\!\perp R_{-k}, X_k \mid X_{-k}, \forall k \quad \textit{(block-parallel MNAR)} \quad (6)$$

An example of this model is shown in Fig. 4(d). Using graphoid axioms, it is easy to show that the block-parallel model assumes $R_k \perp\!\!\!\perp R_j \mid X, \forall j \neq k$ on top of what the no self-censoring model, defined in (3), already assumes. Thus, we view the block-parallel model ($\mathcal{M}_o$) as a submodel of the saturated no self-censoring model ($\mathcal{M}_a$), and focus on testable implications and empirical evaluations of these extra assumptions. Unlike the sequential models, the independence statements here are between missingness indicators and there is no predefined ordering.

If we were to follow ideas from the previous two subsections, we would need to intervene on R_k and R_j to test the independence $R_k \perp\!\!\!\perp R_j \mid X$ as X_k, X_j appear in the conditioning set. Interventions on R_k and R_j fix them to constants, which prevent us from evaluating independence. One might then conclude that such constraints are untestable. However, we use odds-ratio parameterization of the missingness mechanism to argue that these restrictions are indeed testable. We formalize the results in the following theorem.

Theorem 4. *The independence $R_k \perp\!\!\!\perp R_j | X \; \forall j \neq k$ has a testable implication on observed data law which can be stated via OR$(R_k, R_j | X_{-kj}, R_{-kj} = 1) = 1$.*

As an example, consider the m-DAG in Fig. 4(d) and its supermodel in (b). The absence of an edge between R_1, R_2 implies $R_1 \perp\!\!\!\perp R_2 | X$. This is equivalent to stating that the odds ratio between R_1 and R_2 conditioned on X_1, X_2 is one, i.e., $\text{OR}(R_1 = 0, R_2 = 0 | X) = 1$. See Appendix A.1 for a description of the odds ratio parameterization. Let θ denote the odds ratio. Malinsky et al. [2021] proposed the following unbiased estimating equation to estimate θ:

$$\mathbb{P}_n\Big[R_1 R_2 \times \frac{p(R_1 = 0, R_2 = 0 \mid X)}{p(R_1 = 1, R_2 = 1 \mid X)} - (1 - R_1)(1 - R_2)\Big] = 0,$$

where θ appears in the density ratio since it equals:

$$\frac{p(R_1 = 0 \mid R_2 = 1, X_2)}{p(R_1 = 1 \mid R_2 = 1, X_2)} \times \frac{p(R_2 = 0 \mid R_1 = 1, X_1)}{p(R_2 = 1 \mid R_1 = 1, X_1)} \times \theta.$$

See Appendix A.1.1 for detailed derivations. For $K > 2$ variables, we can test the absence of edges between any two pairs of missingness indicators by computing pairwise odds ratios. We formalize the goodness-of-fit tests based on these calculations in Algorithm 2 outlined in Appendix C.4.

4.4 EXTENSIONS TO SETTINGS WITH UNMEASURED CONFOUNDERS

We can extend the applicability of our results to scenarios where not only variables are missing but some are completely unobserved, by considering hidden variable DAGs $\mathcal{G}(V \cup U)$, where $V = \{X, R, X^*\}$ and the variables in U are unobserved. In such cases, we can obtain a missing data acyclic directed mixed graph (m-ADMG) $\mathcal{G}(V)$, by applying the latent projection operator [Verma and Pearl, 1990] to the hidden variable DAG $\mathcal{G}(V \cup U)$. The full law then follows the nested Markov factorization [Richardson et al., 2023] with respect to the m-ADMG $\mathcal{G}(V)$. A m-ADMG obtained through the projection of a hidden variable m-DAG adheres to the same edge restrictions.

If there is a concern about latent confounding in the target law, our framework allows for arbitrary confounding among the X variables without any modifications to the proposed tests in the previous section. This means we can incorporate unmeasured confounders of the form $X_i \leftarrow U \rightarrow X_j$, often represented as a bidirected edge $X_i \leftrightarrow X_j$ between any pair of variables X_i and X_j. It is important to note that the missing data mechanism $p(R|X)$ is largely unaffected by the inclusion of these bidirected edges. For example, in Fig. 2(a), we can introduce unmeasured confounders between every pair of variables in X while maintaining the same testable implications and goodness-of-fit tests.

Unmeasured confounding is also possible in the missing data mechanism, as long as there are no "colluding paths" between any X_i and its corresponding missingness indicator R_i. A colluding path is a path where every node on the path is a collider. The presence of colluding paths leads to the non-identification of the full law, potentially affecting the validity of the proposed goodness-of-fit tests. For instance, in Fig. 3(a), we can have unmeasured confounders between X_3 and R_2, or between X_2 and R_1, or between all variables in X, and so on. For a formal definition of colluding paths and a detailed explanation of why they result in non-identification, refer to [Nabi et al., 2020].

5 SIMULATIONS

We conduct three sets of simulation analyses to illustrate the utility of our proposed methods in testing the miss-

[6] Block-parallel model was introduced in Mohan et al. [2013].

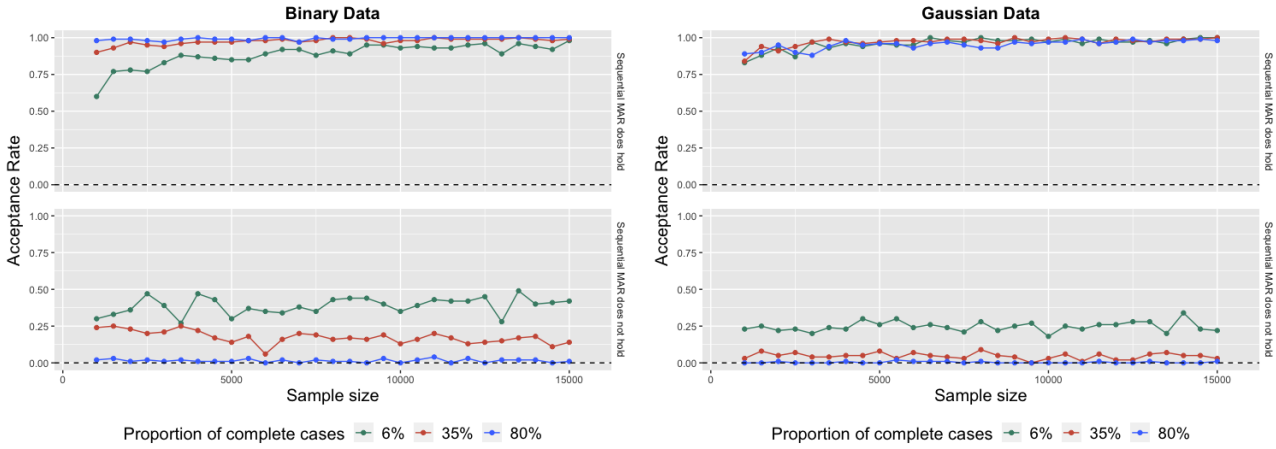

Figure 5: Results on testing **sequential MAR** models. *(top row)* The sequential MAR model captures the true underlying missingness mechanism. *(bottom row)* The assumptions of sequential MAR model are violated.

ing data restrictions using only partially observed samples. Each set focuses on a class of m-DAGs that were considered here. For each simulation, we generate four random variables from either a multivariate normal distribution or binomial distribution. We induce missing values in all four variables according to a missingness mechanism that follows restrictions of either sequential MAR, sequential MNAR, block-parallel, or supermodels of them. The exact data generating mechanism is described in Appendix E. R code can be found at `https://github.com/raziehna/missing-data-testability`.

In the main body, we present results on testing the sequential MAR model defined via the set of restrictions in (4). We follow Algorithm 1 to test the independence restrictions, which entails running a total of $K-1$ tests. Our test statistic is 2ρ and we use a chi-square distribution with $K-k$ degrees of freedom to evaluate the goodness-of-fits – the degree of freedom is chosen as the difference between number of parameters in $W_k(\beta_k^a)$ and $W_k(\beta_k^0)$, as defined in the algorithm. If the p-values are all greater than 0.05, we accept the sequential MAR model. Results on sequential MNAR and block-parallel MNAR models are provided in Appendix E.

For a fixed sample size, we simulate 100 different datasets and calculate the acceptance rate of a sequential MAR model. The acceptance rate is plotted as a function of sample size in Fig. 5. The sample size ranges from $1,000$ to $15,000$ with 500 increments. In each panel, there are three plots that vary in terms of the proportion of complete cases in the dataset, i.e., $6\%, 35\%, 80\%$ which is achieved by changing the range in the uniform distribution where the parameters are sampled from (the proportion of complete cases is taken as an average of complete cases over 100 iterations). The top row of Fig. 5 illustrates the results when the true underlying missingness mechanism satisfies the assumptions of the sequential MAR model, and the bottom row illustrates results for when the restrictions are no longer valid As seen in the figure, the acceptance rate is quite high when the sequential MAR model holds true and it is low when the model does not hold, even if we have only 6% complete cases which is impressive performance with small data. The plots at the bottom row also illustrate that the tests would perform better in terms of rejecting the sequential MAR model while the truth is not MAR when the missingness rate decreases; with 80% complete cases the acceptance rate vanishes.

6 CONCLUSIONS

Independence restrictions in a missing data model might be empirically untestable, or they might translate into more complex restrictions on the observed data law than ordinary d-separation statements. In this paper, we considered various graphical models of missing data and investigated testable implications of the underlying statistical assumptions on the observed data law. We have extended the notion of testability in prior literature by viewing ordinary conditional independence tests as Verma constraints in intervention distributions. We have proposed goodness-of-fit tests based on weighted likelihood-ratio tests and odds-ratio parameterizations. Our results are essential in validating the assumed statistical missing data models in practice and discovering the mechanisms that drive the missingness of variables. A potential future direction is to develop estimation methods that would complement our proposals by allowing a more efficient use of data in performing goodness-of-fit tests.

Acknowledgements

This work is partly supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378.

References

Claudio Agostinelli and Marianthi Markatou. Test of hypotheses based on the weighted likelihood methodology. *Statistica Sinica*, pages 499–514, 2001.

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

Rohit Bhattacharya, Razieh Nabi, and Ilya Shpitser. Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research*, 23:1–76, 2022.

Hua Yun Chen. A semiparametric odds ratio model for measuring association. *Biometrics*, 63:413–421, 2007.

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

Robin J Evans. Graphs for margins of Bayesian networks. *Scandinavian Journal of Statistics*, 43(3):625–648, 2016.

Alex Gain and Ilya Shpitser. Structure learning under missing data. In *Proceedings of the 9th International Conference on Probabilistic Graphical Models*, pages 121–132, 2018.

M Maria Glymour. Using causal diagrams to understand common problems in social epidemiology. *Methods in Social Epidemiology*, pages 393–428, 2006.

Steffen L Lauritzen. *Graphical Models*. Oxford, U.K.: Clarendon, 1996.

Lingling Li, Changyu Shen, Xiaochun Li, and James M Robins. On weighting approaches for missing data. *Statistical methods in medical research*, 22(1):14–30, 2013.

Roderick JA Little and Donald B Rubin. *Statistical Analysis With Missing Data*, volume 793. John Wiley & Sons, 2019.

Daniel Malinsky, Ilya Shpitser, and Eric J Tchetgen Tchetgen. Semiparametric inference for nonmonotone missing-not-at-random data: the no self-censoring model. *Journal of the American Statistical Association*, pages 1–9, 2021.

Fernando Martel García. Definition and diagnosis of problematic attrition in randomized controlled experiments. *Available at SSRN 2302735*, 2013.

Karthika Mohan and Judea Pearl. On the testability of models with missing data. In *Artificial Intelligence and Statistics*, pages 643–650. PMLR, 2014.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, pages 1–16, 2021.

Karthika Mohan, Judea Pearl, and Tian Jin. Missing data as a causal inference problem. In *Proceedings of the Neural Information Processing Systems Conference*, 2013.

Razieh Nabi, Rohit Bhattacharya, and Ilya Shpitser. Full law identification in graphical models of missing data: Completeness results. In *Proceedings of the Twenty Seventh International Conference on Machine Learning (ICML-20)*, 2020.

Razieh Nabi, Rohit Bhattacharya, Ilya Shpitser, and James M Robins. Causal and counterfactual views of missing data models. *arXiv preprint arXiv:2210.05558*, 2022.

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

Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1): 334–361, 2023.

James M Robins. Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16:21–37, 1997.

James M Robins and Larry Wasserman. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. In *Proceedings of the 13th Conference on Uncertainty in Artificial Intelligence*, pages 409–420, 1997.

Donald B Rubin. Causal inference and missing data (with discussion). *Biometrika*, 63:581–592, 1976.

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

Mauricio Sadinle and Jerome P Reiter. Itemwise conditionally independent nonresponse modeling for incomplete multivariate data. *Biometrika*, 104(1):207–220, 2017.

Daniel O Scharfstein, Jaron JR Lee, Aidan McDermott, Aimee Campbell, Edward Nunes, Abigail G Matthews, and Ilya Shpitser. Markov-restricted analysis of randomized trials with non-monotone missing binary outcomes: Sensitivity analysis and identification results. *arXiv preprint arXiv:2105.08868*, 2021.

Ilya Shpitser. Consistent estimation of functions of data missing non-monotonically and not at random. *Advances in Neural Information Processing Systems*, 29: 3144–3152, 2016.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the 21st National Conference on Artificial Intelligence*, 2006.

Ilya Shpitser, Karthika Mohan, and Judea Pearl. Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 802–811, 2015.

Felix Thoemmes and Norman Rose. A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research*, 49(5):443–459, 2014.

Jin Tian. Missing at random in graphical models. In *Artificial Intelligence and Statistics*, pages 977–985. PMLR, 2015.

Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pages 519–527, 2002.

Ruibo Tu, Cheng Zhang, Paul Ackermann, Karthika Mohan, Hedvig Kjellström, and Kun Zhang. Causal discovery in the presence of missing data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1762–1770. PMLR, 2019.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence*, 1990.