Estimation with Incomplete Data: The Linear Case

Karthika Mohan[1], **Felix Thoemmes**[2], **Judea Pearl**[3]
[1] University of California, Berkeley
[2] Cornell University
[3] University of California, Los Angeles
karthika@eecs.berkeley.edu, felix.thoemmes@cornell.edu, judea@cs.ucla.edu

Abstract

Traditional methods for handling incomplete data, including Multiple Imputation and Maximum Likelihood, require that the data be Missing At Random (MAR). In most cases, however, missingness in a variable depends on the underlying value of that variable. In this work, we devise model-based methods to consistently estimate mean, variance and covariance given data that are Missing Not At Random (MNAR). While previous work on MNAR data require variables to be discrete, we extend the analysis to continuous variables drawn from Gaussian distributions. We demonstrate the merits of our techniques by comparing it empirically to state of the art software packages.

1 Introduction

Incomplete (or missing) data are data in which values of one or more variables are missing. Almost all existing techniques for handling incomplete data employ maximum likelihood or multiple imputation methods to fill in the missing values and estimate parameters of interest. To guarantee convergence and consistency, these techniques require that the missing data mechanism be ignorable (Rubin [1976]) i.e. the causes of missingness be either random or fully observed.

In practice, however, missingness is almost always caused by variables that are themselves afflicted by missingness (Osborne [2012, 2014]; Sverdlov [2015]; Adams [2007]; van Stein and Kowalczyk [2016]). Such data are called Missing Not At Random (MNAR). Among all MNAR problems, the most frequently encountered case is that of a variable causing its own missingness which we call *self-masking* MNAR. It is discussed in almost all literature on missing data including Koller and Friedman [2009] (chapter 19) and Darwiche [2009] (chapter 17). Examples include smokers not answering questions about their smoking behavior in insurance applications, longitudinal studies with attrition (Little [1995]), people with high income not revealing their incomes and a general reluctance to answer questions about sensitive topics such as religion, sexual preference and abortion (Greenland and Finkle [1995]).

While there has been some recent work (Daniel *et al.* [2012]; Mohan *et al.* [2013]; Mohan and Pearl [2014a, 2018]; Thoemmes and Mohan [2015]; Shpitser *et al.* [2015]) on estimation in MNAR data with discrete variables, to the best of our knowledge there exists no theoretically sound and empirically efficient graph based procedure for handling MNAR missingness in datasets with continuous variables.

In this paper we focus on MNAR problems in linear systems. As in Mohan and Pearl [2014b] and Mohan and Pearl [2018], we treat missing data as a causal inference problem. We present sound, graph-based procedures for recovering (i.e. consistently estimating) parameters such as mean, variance and covariance. In the following section we review missingness graphs and structural equation models.

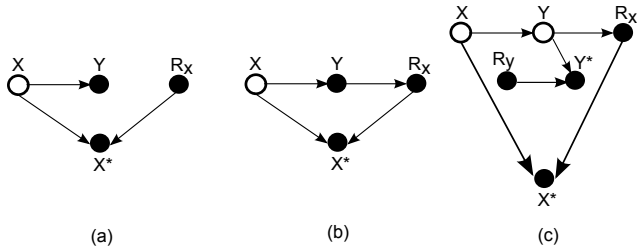


Figure 1: Examples of (a) MCAR, (b) MAR and (c) MNAR models

2 Preliminaries

2.1 Missingness Graphs: Notations and Terminology

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

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

V^* is the set of all proxy variables and $\mathbf{R}$ is the set of all missingness mechanisms. U is the set of unobserved nodes

(or latent variables). We use bi-directed edges as a shorthand notation to denote the existence of a U variable as common parent of two variables in $V_o \cup V_m \cup \mathbf{R}$. For any $W \subseteq V_o$, $R_w = \emptyset$. Unless stated otherwise it is assumed that no variable in $V \cup U$ is a child of an R variable. This graphical representation is called **Missingness Graph** (or m-graph) (Mohan *et al.* [2013]). An m-graph portrays Missing Completely At Random (MCAR) missingness if $(V_m, V_o, U) \perp\!\!\!\perp R$, Missing At Random (MAR) if $(V_m, U) \perp\!\!\!\perp R | V_o$ and Missing Not At Random (MNAR) if neither MAR nor MCAR hold. Figure 1 (a), (b) and (c) exemplify MCAR, MAR and MNAR missingness respectively.

Proxy variables may not always be explicitly shown in m-graphs in order to keep the figures simple and clear. Conditional Independencies are read off the graph using the d-separation[1] criterion [Pearl, 2009]. Before formally defining the linear missingness model, we shall briefly review Structural Equation Models. For a detailed discussion see Pearl [2009] (chapter 5) and Brito [2004].

2.2 Structural Equation Models

A structural equation model (SEM) is a system of equations defined over a set of variables, such that each variable appears on the left hand side of at most one equation. Each equation describes the dependence of one variable in terms of the others and contains an error term to account for the influence of unobserved factors. Example: $X = \epsilon_x$ and $Y = \alpha X + \epsilon_y$. As in Pearl [2013], we interpret structural equations as an assignment process whose directionality is captured by a path diagram (see Figure 2). In this work all substantive variables ($\{V_m \cup V_o \cup U\}$) and error terms are assumed to be drawn from a Gaussian distribution. Linear Structural Equation Modeling is widely used for estimating parameters of interest given data that are missing at random (Allison [2003]; Graham [2003]; Ullman and Bentler [2003]; Enders [2006]; Schlomer *et al.* [2010]).

2.3 Recoverability

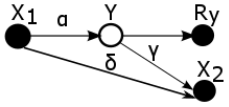
Definition 1 (Recoverability of target quantity Q (Mohan *et al.* [2013]))**.** *Let A denote the set of assumptions about the data generation process and let Q be any functional of the underlying distribution $P(V_m, V_O, R)$. Q is recoverable if there exists an algorithm that computes a consistent estimate of Q for all strictly positive manifest distributions $P(V^*, V_o, R)$ that may be generated under A.*

We present an example of recoverability in linear models in section 3.1. For examples of recoverability in non-parametric models with discrete variables, please see Mohan *et al.* [2013].

3 Quasi-linear Missingness Model

The causal missingness mechanism is a binary variable and as such the function generating it cannot be linear. The quasi-linear model defined below captures the missingness process.

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



$$
\begin{aligned}
&(i)\ X_1 = \epsilon_{x_1} \\
&(ii)\ Y = \alpha X_1 + \epsilon_y \\
&(iii)\ X_2 = \gamma Y + \delta X_1 + \epsilon_{x_2} \\
&(iv)\ R_y = f(Y, \epsilon_{r_y}) \\
&(v)\ Y^* = (1 - R_y)Y + mR_y
\end{aligned}
$$

Figure 2: Quasi-linear missingness model and equations constituting the SEM corresponding to it.

Definition 2. *A Quasi-linear Missingness Model is a Structural Equation Model such that:*
1. every substantive variable X is a linear function of its causes ,Y, and a random error term ϵ_x

$$X = \alpha_1 Y_1 + \alpha_2 Y_2 + ... + \alpha_n Y_n + \epsilon_x$$

The coefficient α's are called path coefficients or structural parameters.
2. For every $R_x \in R$, $R_x = f(Z, \epsilon_{R_x})$ where Z is the set of causes and f is a non-linear function. No R variable is a parent of any substantive variable.
3. Every proxy variable X^* is generated by the non-linear function: $X^* = (1 - R_x)X + mR_x$

Figure 2 and equations (i) to (v) exemplify an m-graph and its corresponding quasi-linear missingness model. Path coefficients can be identified in linear models by applying criteria such as single door and back door (see Pearl [2009] (chapter 5)). Appendix 8.1 lists basic formulae used in this paper. In the following lemma we rephrase and state a basic result in linear path analysis that links covariance with path coefficients (Pearl [2013, 2009]; Brito [2004]; Wright [1921]).

Lemma 1. *Let G be an m-graph with k unblocked paths $p_1, ..p_k$ between X and Y. Let A_{p_i} be the ancestor of all nodes on path p_i. Let the number of nodes on p_i be n_{p_i}.*

$$cov(X, Y) = \sum_{i=1}^{k} var(A_{p_i}) * \prod_{j=1}^{n_{p_i}-1} \alpha_j^{p_i} \qquad (2)$$

where $\prod_{j=1}^{n_{p_i}-1} \alpha_j^{p_i}$ is the product of all causal parameters on path p_i.

For example, in figure 2, there exist two paths, $X_1 \to Y \to X_2$ and $X_1 \to X_2$, between X_1 and X_2. On applying lemma 1 we get, $cov(X_1, X_2) = \alpha\gamma var(X_1) + \delta var(X_1)$.

In the following subsection we will exemplify a novel path analytic procedure for consistent estimation of the covariance matrix given MAR data.

3.1 Recoverability of Covariance Matrix: An Example

For any target quantity Q we use $Q_{\|_{R_X=0}}$ to denote: compute Q from samples in which all variables in X are observed.

Example 1. *Consider the problem of estimating the **covariance matrix** given the MAR model of Figure 1 (b). Y is fully observed and hence $var(Y)$ is trivially recoverable. In order*

to recover $cov(X,Y)$, we will first recover β_{XY}, the regression coefficient of X on Y. Since $X \perp\!\!\!\perp R_x|Y$ we have the license to compute β_{XY} (using OLS) from samples in which X is observed i.e. $\beta_{XY} = \beta_{XY\|_{R_x=0}}$. $cov(X,Y)$ can now be recovered as:

$$cov(X,Y) = \beta_{XY\|_{R_x=0}}\ var(Y) \tag{3}$$

To recover $var(X)$, we apply the law of total variance:

$$var(X) = E(var(X|Y)) + var(E(X|Y))$$

- *Recovering $var(E(X|Y))$: $E(X|Y) = \beta_{XY}Y + c_x$ where β_{XY} and c_x denote the slope and intercept of the regression line. Since $X \perp\!\!\!\perp R_x|Y$, $E(X|Y) = E(X|Y)_{\|R_X=0}$. Therefore,*

$$\begin{aligned} var(E(X|Y)) &= var\big(\beta_{XY\|_{R_x=0}}Y + c_{y\|_{R_x=0}}\big) \\ &= (\beta_{XY\|_{R_x=0}})^2 var(Y) \end{aligned} \tag{4}$$

- *Recovering $E(var(X|Y))$: The variance of a conditional gaussian distribution is constant. Therefore,*

$$E(var(X|Y)) = var(X|Y)$$

Variance of a conditional bivariate Gaussian distribution, $var(X|Y)$, is given by $var(X)(1-\rho^2)$, where $\rho = \frac{cov(X,Y)}{(var(X)var(Y))^{\frac{1}{2}}}$ denotes the correlation coefficient. Since $X \perp\!\!\!\perp R_x|Y$, $var(X|Y) = var(X|Y)_{\|R_X=0}$. Hence,

$$var(X|Y) = var(X)_{\|R_x=0}(1-(\rho_{\|R_x=0})^2) \tag{5}$$

Using 4 and 5 $var(X)$ is computed as,

$$\begin{aligned} var(X) =& var(X)_{\|R_x=0}(1-(\rho_{\|R_x=0})^2) \\ &+ (\beta_{XY\|_{R_x=0}})^2 var(Y) \end{aligned}$$

Note that while all factors in the preceding estimand, except $var(Y)$ are estimated from samples in which X is observed, $var(Y)$ is recovered from all samples in the dataset, including those in which X is missing. In other words, although a consistent estimate of $var(X)$ cannot be computed directly from fully observed data (i.e. $P(X^*, Y, R_x = 0)$), it can be recovered by a procedure in which each factor in the estimand is estimated from subsets of the available data.
We further note that as a consequence of recovering $var(X)$, β_{YX}, the causal effect of X on Y can be recovered as, $\frac{cov(X,Y)}{var(X)} = \frac{\beta_{XY\|_{R_x=0}} var(Y)}{var(X)_{\|R_x=0}(1-(\rho_{\|R_x=0})^2)+(\beta_{XY\|_{R_x=0}})^2 var(Y)}$.

The following section presents procedures for computing parameters of interest in quasi linear models.

4 Recovering mean, variance and covariance

Theorem 1 (Recovering Mean of Partially Observed Variables)**.** *Let $\{X,Z\} \subseteq V_m \cup V_o$, $X \cap Z = \emptyset$, $|X| = 1$. Given m-graph G, $E(X)$ is recoverable if there exists $Z = \{Z_1, Z_2, ...Z_n\}$ such that $X \perp\!\!\!\perp R_x R_z|Z$ and $E(Z_i)$ is recoverable for all $Z_i \in Z$. If recoverable, $E(X)$ is given by*

$$E(X) = c_{\|R_{xz}=0} + \sum_{i=1}^{n} \alpha_{i\|R_{xz}=0} E(Z_i) \tag{6}$$

where c and α_i's denote the intercept and coefficients of the regression line of X on Z.

Proof. Let $X = \alpha_1 Z_1 + \alpha_2 Z_2 + ... + \alpha_n Z_n + c$ be the regression line of X on Z. Since $X \perp\!\!\!\perp R_x R_z|Z$, $E(X|Z) = E(X|Z)_{\|R_{xz}=0}$. Hence,

$$E(X|Z) = \alpha_{1\|R_{xz}=0} Z_1 + ... + \alpha_{n\|R_{xz}=0} Z_n + c_{\|R_{xz}=0}$$

However since $E(X) = E(E(X|Z))$ we can write,

$$\begin{aligned} E(X) &= E\big(\alpha_{1\|R_{xz}=0} Z_1 + ... + \alpha_{n\|R_{xz}=0} Z_n + c_{\|R_{xz}=0}\big) \\ &= c_{\|R_{xz}=0} + \sum_{i=1}^{n} \alpha_{i\|R_{xz}=0} E(Z_i) \end{aligned}$$

□

Theorem 2 (Recovering covariance of X and Y)**.** *Let $\{X,Y,Z\} \subseteq V_m \cup V_o$, $X \cap Y \cap Z = \emptyset$, $|X| = 1, |Y| = 1$ and $|Z| = n$. Given m-graph G, $cov(Y,X)$ is recoverable if there exists $Z = \{Z_1, Z_2, ...Z_n\}$ such that (i) $Y \perp\!\!\!\perp R_x, R_y, R_z|X, Z$, and (ii) $E(X), E(Y), var(X)$, and $\forall i$ $E(Z_i)$ and $cov(X, Z_i)$ are recoverable. If recoverable, $cov(Y,X)$ is given by*

$$\begin{aligned} &cov(Y,X) = \alpha_{x\|R_{xyz}=0}(var(X) + E(X)^2) + c_{\|R_{xyz}=0} E(X) \\ &+ \sum_i \alpha_{i\|R_{xyz}=0}(cov(XZ_i) + E(X)E(Z_i)) - E(X)E(Y) \end{aligned}$$

where c and α's denote the intercept and coefficients of the regression line of Y on X, Z.

Proof.

$$\begin{aligned} cov(Y,X) &= E(E(XY|Z)) - E(Y)E(X) \\ &= E(XE(Y|Z,X)) - E(Y)E(X) \end{aligned}$$

Using $Y \perp\!\!\!\perp R_x, R_y, R_z|X, Z$, we get
$E(Y|X,Z) = E(Y|X,Z)_{\|R_{xyz}=0}$. Therefore,

$$\begin{aligned} &E(XE(Y|X,Z)) = E(X\big(\alpha_{x\|R_{xyz}=0} X + \alpha_{1\|R_{xyz}=0} Z_1 \\ &+ \alpha_{2\|R_{xyz}=0} Z_2 + ... + \alpha_{k\|R_{xyz}=0} Z_k + c_{\|R_{xyz}=0}\big) \\ &= \alpha_{x\|R_{xyz}=0} E(X^2) + c_{\|R_{xyz}=0} E(X) + \sum_i \alpha_{i\|R_{xyz}=0} E(XZ_i) \\ &= \alpha_{x\|R_{xyz}=0}(var(X) + E(X)^2) + c_{\|R_{xyz}=0} E(X) \\ &+ \sum_i \alpha_{i\|R_{xyz}=0}(cov(XZ_i) + E(X)E(Z_i)) \end{aligned}$$

□

A well known and widely used property of Gaussian distributions is that their conditional variances are constant. Let $|X| = m$ and $|Y| = n$. Let M_{xx}, M_{yy} and M_{xy} denote the covariance matrix of X, Y and X and Y respectively. Variance of the conditional Gaussian distribution $f(Y|X)$ is given by,

$$Var(Y|X) = M_{yy} - M'_{xy} M_{xx}^{-1} M_{xy} \tag{7}$$

Theorem 3 (Recovering variance of a partially observed variable X)**.** *Let $\{X,Z\} \subseteq V_m \cup V_o$, $X \cap Z = \emptyset$, $|X| = 1$ and $|Z| = n$. Given m-graph G, $var(X)$ is recoverable if*

there exists $Z = \{Z_1, Z_2, ...Z_n\}$ such that $X \perp\!\!\!\perp R_x R_z | Z$ and $cov(Z_i, Z_j)$ is recoverable for all $Z_i, Z_j \in Z$. If recoverable, $var(X)$ is given by

$$var(X) = (M_{xx} - M'_{zx} M_{zz}^{-1} M_{zx})_{\|_{R_{xz}=0}} + \sum_i \alpha_i^2 {}_{\|_{R_{xz}=0}} var(Z_i) + \sum_{i \neq j} (\alpha_i \alpha_j)_{\|_{R_{xz}=0}} cov(Z_i, Z_j)$$

where α_i's denote the coefficients of the regression line of X on Z.

Proof. We first apply the law of total variance: $var(X) = E(var(X|Z)) + var(E(X|Z))$, and then prove recoverability of $var(X)$ by showing that both summands are recoverable.

Recovering $E(var(X|Z))$: Since X and $Z_i \in Z$, $\forall i$ are Gaussian variables, $var(X|Z)$ is constant; therefore,

$$\begin{aligned} E(var(X|Z)) &= var(X|Z) \\ &= var(X|Z)_{\|_{R_{xz}=0}} \text{ (since } X \perp\!\!\!\perp R_x, R_z | Z) \\ &= (M_{xx} - M'_{zx} M_{zz}^{-1} M_{zx})_{\|_{R_{xz}=0}} \text{ (using eqn 7)} \end{aligned}$$

Recovering $var(E(X|Z))$: Since $X \perp\!\!\!\perp R_x, R_z | Z$ we have,

$$E(X|Z) = E(X|Z)_{\|_{R_{xz}=0}} = c_{x\|_{R_{xz}=0}} + \sum_i \alpha_{i\|_{R_{xz}=0}} Z_i$$

$$\begin{aligned} var(E(X|Z)) &= var\big(c_{\|_{R_{xz}=0}} + \sum_i \alpha_{i\|_{R_{xz}=0}} Z_i\big) \\ &= \sum_i \alpha_i^2 {}_{\|_{R_{xz}=0}} var(Z_i) + 2 \sum_{i \neq j} (\alpha_i \alpha_j)_{\|_{R_{xz}=0}} cov(Z_i, Z_j) \end{aligned}$$

□

Lemma 2. *[Recovering Partial Regression Coefficients] Let $\{X, Y, Z\} \subseteq V_o \cup V_m$. Given m-graph G and missing data D, partial regression coefficient $\beta_{XY.Z}$ is recoverable if $X \perp\!\!\!\perp R_X, R_Y, R_Z | Y, Z$ and is given by $\beta_{XY.Z\|_{R_{xyz}=0}}$.*

Notice that recoverability of $E(X)$ using theorem 1 is contingent upon the recoverability of $E(Z_i)$, for all i. Clearly, to recover $E(X)$ we should recover all $E(Z_i)$'s first. Similar is the case with theorems 2 and 3. In the case of recoverability of conditional distributions in datasets with discrete variables, Mohan *et al.* [2013] defined the notion of *ordered factorization* to sequence the recoverability procedure. In the following theorem we adapt the idea of ordered factorization in Mohan *et al.* [2013] to formulate a sufficient condition for recovering mean.

Theorem 4. *Let $Y \cup \{Z\} \subseteq V_m \cup V_o$ and let O: $Y_1 < Y_2 < \ldots < Z = Y_k$ be an ordered set of all variables in $Y \cup \{Z\}$. Let $X_i \subseteq \{Y_1, \ldots, Y_{i-1}\}$, $1 \leq i \leq k$.*

Given an m-graph G, a sufficient condition for recovering $E(Z)$ is that there exist O and $X_i \forall i$ such that $Y_i \perp\!\!\!\perp (R_{y_i}, R_{x_i}) | X_i$.

Proof. Recoverability follows from theorem 1. We proceed by first recovering $E(Y_1)$ and then recovering $E(Y_2)$,..,$E(Y_k)$ sequentially in the order dictated by O. □

Example 2. *Consider the problem of recovering $E(X)$ and $E(Y)$ given the m-graph G in figure 1 (c) and ordering $O: Y < X$. Ordering O directs us to first recover $E(Y)$ and then recover $E(X)$. Since $Y \perp\!\!\!\perp R_y$ in G, we can apply theorem 1 to recover $E(Y)$ as $E(Y)_{\|_{R_y=0}}$. Since $X \perp\!\!\!\perp R_x, R_y | Y$ and $E(Y)$ is recoverable, we can apply theorem 1 to recover $E(X)$ as $c_{\|_{R_{xy}=0}} + \alpha_{1\|_{R_{xy}=0}} E(Y)_{\|_{R_y=0}}$.*

In a similar manner, ordered recoverability procedures can be extended to recover covariance and variance as well. In the case of MCAR data, all orderings of variables will guarantee recoverability, whereas in the case of MAR data, orderings in which all fully observed variables precede partially observed variables will guarantee recoverability. Finally in the case of MNAR data, the ordering is determined by graph structure, and heuristics for finding admissible orders are discussed in Mohan *et al.* [2013].

5 Extending recoverability to complex MNAR models

In this section we focus on m-graphs in which conditional independence between a variable and its missingness mechanism (such as $Y \perp\!\!\!\perp R_y$ and $Y_i \perp\!\!\!\perp (R_{y_i}, R_{x_i}) | X_i$) that are required by theorems 1, 2, 3 and 4 does **not** hold in the graph. In the following subsection we present a theorem for recovering $E(Y)$ for any $Y \in V_m$, by leveraging X, a neighbor of Y.

5.1 Recovering $E(Y)$ when Y and R_y are not separable

Theorem 5 (Recovering $E(Y)$ for any $Y \in V_m$)**.** *Let $X \in V_m \cup V_o$ be a neighbor of Y. Given m-graph G and missing data D, $E(Y)$ is recoverable if $E(X)$ is recoverable and there exists $Z \subseteq V_m \cup V_o - \{X, Y\}$ such that $X \perp\!\!\!\perp R_x, R_y, R_z | Y, Z$ and $E(Z_i)$ is recoverable for all i. If recoverable, $E(Y)$ is given by,*

$$E(Y) = \frac{E(X) - c_{x\|_{R_{xyz}=0}} - \sum_{i=1}^{k} \alpha_{i\|_{R_{xyz}=0}} E(Z_i)}{\alpha_{y\|_{R_{xyz}=0}}} \quad (8)$$

where α's, and c denote the coefficients and intercept of the regression line of X on Z and Y.

Proof: See Appendix 8.2

Example 3. *To recover $E(Y)$ given the m-graph in figure 2, we leverage X_1. $X_1 \perp\!\!\!\perp R_y | Y$ and $E(X_1)$ is recoverable since $X_1 \in V_O$. Hence mean of Y can be recovered using theorem 5 as, $\frac{E(X_1) - c_{\|_{R_y=0}}}{\beta_{X_1,Y\|_{R_y=0}}}$.*

5.2 Recovering $var(Y)$ when Y and R_y are not separable

Algorithm 1 recovers variance of a variable Y given an m-graph in which X_1 is a parent of Y and X_2 is a child of Y. It uses two subroutines (outlined in appendix 8.3) that compute path coefficient and covariance using lemma 1. If a quantity of interest Q is recoverable, then Q^* is used as a shorthand for the recovered estimand. For example if $var(Y)$ is recoverable

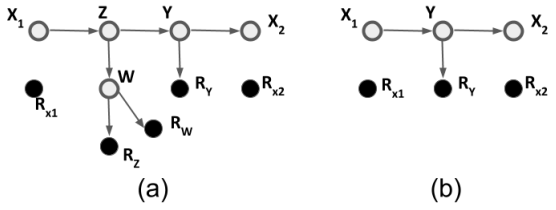


Figure 3: (a) Self masking model in which all variables are partially observed, yet $E(Y)$ and $var(Y)$ are recoverable. (b) m-graph constructed from (a) by treating Z, Y, R_z and R_w as latent.

Algorithm 1 RecoverVariance(Y, G, X_1, X_2)

Input: Y: variable whose variance is to be recovered. G: Markovian m-graph in which X_1 is a parent of Y and X_2 is a child of Y

Output: $var(Y)^*$ if $var(Y)$ is recoverable
$NULL$ if $var(Y)$ is not recoverable

1: **if** $var(Y)$ is recoverable using theorem 3 **then**
2: Recover estimand $var(Y)^*$
3: **return** $var(Y)^*$
4: **if** $\beta_{X_1,Y}$ is recoverable by lemma 2 **then**
5: Recover estimand ${\beta_{X_1,Y}}^*$
6: **else return** NULL
7: ${\alpha_y}^* \leftarrow$ Recover_$\alpha_y(G, Y, X_1, X_2)$
8: **if** ${\alpha_y}^*$== NULL **then return** NULL
9: $cov(X_1, Y)^* \leftarrow$ Recover_cov$(G, Y, X_1, {\alpha_y}^*)$
10: **if** $cov(X_1, Y)^* ==$ NULL **then return** NULL
11: **return** $\frac{cov(X_1,Y)^*}{{\beta_{X_1,Y}}^*}$

then $var(Y)^*$ denotes the recovered estimand. We exemplify below the recovery procedure using algorithm 1.

Example 4. *Consider the problem of recovering mean and variance of Y given the m-graph in figure 2. We will show that $var(Y)$ is recoverable using algorithm 1.*
Steps 1-3: Since Y and R_y are neighbors theorem 3 is not applicable. Hence we proceed to the next step.
Steps 4-6: Since$X_1 \perp\!\!\!\perp R_y|Y$, $\beta_{X_1,Y}$ is recoverable using lemma 2 i.e. ${\beta_{X_1,Y}}^* = \beta_{X_1,Y\|_{R_y=0}}$.
Steps 7-8: We invoke subroutine Recover_α_y to recover the path coefficient α in figure 2. Since X_1 and X_2 are fully observed, $cov(X_1, X_2)$ is recoverable. Path coefficients δ and γ may be recovered as:

$$\delta = \beta_{X_2X_1.Y} \text{ (by single door criterion, Pearl [2009])}$$
$$= \beta_{X_2X_1.Y\|_{R_y=0}} \text{ (using lemma 2)}$$
$$\gamma = \beta_{X_2Y.X_1} \text{(by back door criterion, Pearl [2009])}$$
$$= \beta_{X_2Y.X_1\|_{R_y=0}} \text{ (using lemma 2)}$$

Since X_1 is fully observed, $var(X_1)$ is recoverable. On applying lemma 1 we get,

$$cov(X_2, X_1) = \gamma\alpha var(X_1) + \delta var(X_1)$$

$$\text{Therefore, } \alpha = \frac{1}{\beta_{X_2Y.X_1\|_{R_y=0}}}\Big(\frac{cov(X_2, X_1)}{var(X_1)} - \beta_{X_2X_1.Y\|_{R_y=0}}\Big)$$

Steps 9-10: We invoke subroutine Recover_cov to recover $cov(X_1, Y)$. On applying lemma 1 we get: $cov(X_1, Y) = \beta_{X_1Y} var(X_1) = (\beta_{X_1Y})_{\|_{R_y=0}} var(X_1)$.
Step 11: $var(Y)$ is recovered as:

$$var(Y) = \frac{cov(X_1, Y)}{\beta_{X_1Y}} = \frac{\alpha * var(X_1)}{(\beta_{X_1Y})_{\|_{R_y=0}}}$$
$$= \frac{var(X_1)}{(\beta_{X_1Y})_{\|_{R_y=0}}} \frac{1}{\beta_{X_2Y.X_1\|_{R_y=0}}}$$
$$* \Big(\frac{cov(X_2, X_1)}{var(X_1)} - \beta_{X_2X_1.Y\|_{R_y=0}}\Big)$$

While algorithm 1 currently handles only Markovian graphs, we note that it is possible to extend the algorithm to Semi-Markovian models provided we make additional assumptions (pertaining to variances of latent variables). We further note that suitable candidates for X_1 and X_2 are the non-descendants and descendants of Y, respectively. Latent projection (Pearl [2009], chapter 2) of the input graph constructed by treating all intermediate nodes on unblocked paths between X_1 and Y, and X_2 and Y, as latent can yield graphs compatible with the requirements of algorithm 1. When latent projection does not introduce bi-directed edges, recovery is straight forward. We exemplify this in section 6.

6 Empirical Evaluation

We denominate the graph based recovery procedure presented in this paper as Model Based Estimation (MBE) and evaluate MBE by simulating partially observed datasets from missingness graphs and estimating their parameters from the incomplete data. We compare our estimates against those yielded by state of the art packages for SEM that apply Multiple Imputation (MI) (using mice package in R Mic [2018]) and Maximum Likelihood (ML) (using lavaan in R Lav [2018]) techniques [Schminkey *et al.*, 2016; Enders, 2006]. Parameters are evaluated in terms of mean squared error and KL Divergence between original and learned distributions.
Missing At Random: We generate data according to the following model and evaluate the performance of MBE, MI and FIML in terms of Mean Squared Error (MSE) and time taken to compute mean of X.
$X = \epsilon_x$,
$R_y = f(Y_1, Y_2, ...Y_k, \epsilon_{r_y})$
$Y_i = \alpha_i X + \epsilon_y,\ 1 \leq i \leq k$
The first experiment measures the mean squared errors of the three estimators: MI, ML, MBE, and studies how it varies with increase in sample size, for the problem of recovering $mean(X)$ when $|Y| = k = 5$. The results of this experiment are plotted in figure 4. We further estimated the average time each procedure took in recovering and plotted time vs sample size in figure 5. Figure 4 shows that all three methods (MBE, MI, FIML) work almost identically so far as the quality of estimate of $E(X)$ is concerned. From figure 5 it is clear that MBE (a non-iterative procedure) takes less time in computing $E(X)$ as compared to MI and FIML. While all the reported numbers (MSE and time) are averaged over 50 repetitions with different random estimation problems, for a

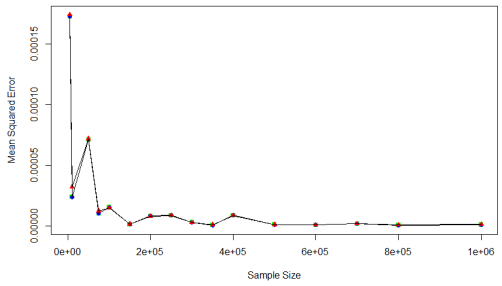


Figure 4: MSE of $E(X)$ vs Sample Size for MAR: $X \perp\!\!\!\perp R_x|Y$, when $|Y| = 5$

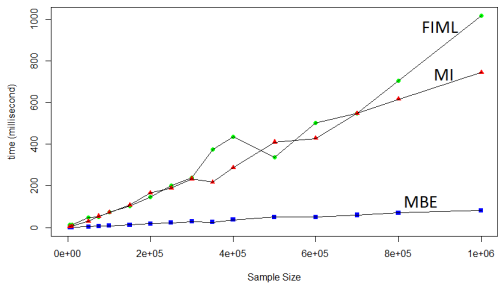


Figure 5: Time to recover $E(X)$ vs Sample Size for MAR: $X \perp\!\!\!\perp R_x|Y$, when $|Y| = 5$

given problem each individual MSE was computed from 500 simulations. In the next experiment we study the efficiency of these procedures as the complexity of the model increases. In MAR models, complexity of recoverability depends on the size of the separating set Y that d-separates X from R_x (and on the size of the dataset). This experiment was conducted by fixing sample size to 100,000. We observe in Figure 6 that as the separating set becomes larger, the time taken to recover estimates also increases. Note that in this case the time gain pertains to computing parameters of one partially observed variable X. Clearly, in a real world dataset with several partially observed variables, the time savings offered by MBE will be substantial.

Missing Not At Random: The goal here is to evaluate the

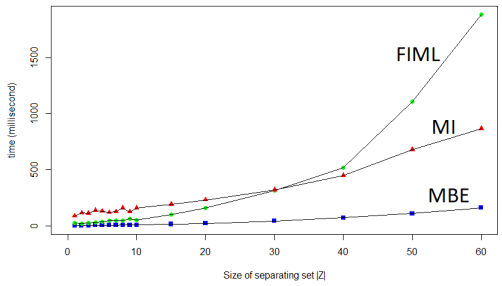


Figure 6: Time to recover $E(X)$ vs $|Y|$ for MAR: $X \perp\!\!\!\perp R_x|Y$

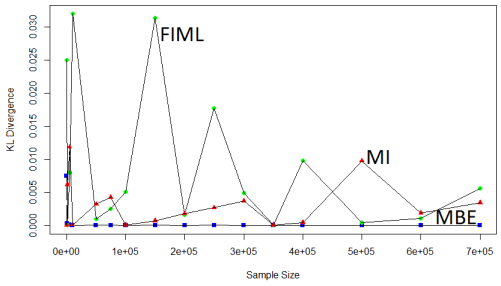


Figure 7: KL divergence vs Sample size for self masking model

effectiveness of algorithm 1 using data generated by the self masking model in 3. β_{wy} and βzy are not recoverable by lemma 2; hence Z and W cannot be used for recovering $var(Y)$. However, $\beta x_1 y$ is recoverable. We can create a latent projection by treating Z, W, R_z and R_w as latent variables as shown in figure 3 (b). Figure 7 depicts the KL divergence between true and estimated distribution of Y. As expected, FIML and MI behave unpredictably while MBE with minimum KL Divergence behaves ideally.

7 Conclusions

We presented novel graph-based procedures that are non-iterative and independent of likelihood function, for recovering parameters in quasi-linear missingness models. We further developed procedures for recovering parameters in self masking models. Finally we showed that given MAR data, our techniques are much faster than state of the art procedures and given MNAR data our techniques can recover parameters where existing techniques fail.

References

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

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

C Brito. *Graphical Models for Identification in Structural Equation Models*. PhD thesis, University of California Los Angeles, 2004.

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

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

C K Enders. Analyzing structural equation models with missing data. *Structural equation modeling: A second course*, pages 313–342, 2006.

J W Graham. Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling*, 10(1):80–100, 2003.

S Greenland and W D Finkle. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *American journal of epidemiology*, 142(12):1255–1264, 1995.

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

Lavaan: R package. `https://cran.r-project.org/web/packages/lavaan/README`, 2018. Accessed: 2018-01-26.

R J A Little. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121, 1995.

Mice: Multivariate imputation by chained equations. `https://cran.r-project.org/web/packages/mice/README.html`, 2018. Accessed: 2018-01-26.

K Mohan and J Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In *Advances in Neural Information Processing Systems 27*, pages 1520–1528. Curran Associates, Inc., 2014.

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

K Mohan and J Pearl. Graphical models for processing missing data. *arXiv preprint arXiv:1801.03583*, 2018.

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

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

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

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

J Pearl. Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, 1(1):155–170, 2013.

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

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

D L Schminkey, T von Oertzen, and L Bullock. Handling missing data with multilevel structural equation modeling and full information maximum likelihood techniques. *Research in nursing & health*, 39(4):286–297, 2016.

I Shpitser, K Mohan, and J Pearl. Missing data as a causal and probabilistic problem. In *Uncertainty in Artificial Intelligence (UAI)*. 2015.

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

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

J B Ullman and P M Bentler. *Structural equation modeling*. Wiley Online Library, 2003.

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

S Wright. Correlation and causation. *Journal of agricultural research*, 20(7):557–585, 1921.

8 Appendix

8.1 Basic Formulae Used in Linear Models

$$cov(X,Y) = E(XY) - E(X)E(Y)$$
$$var(X) = E(X^2) - E(X)^2$$
$$\beta_{YX} = \frac{cov(X,Y)}{var(X)} = \frac{d}{dx}E(Y|X=x)$$
$$\beta_{YX.Z} = \frac{d}{dx}E(Y|X=x,Z=z)$$

8.2 Proof of theorem 5

$E(X) = E(E(X|Z,Y)) = E(E(X|Z,Y)_{\|_{R_{xyz}=0}})$

$= \alpha_{y\|_{R_{xyz}=0}} E(Y) + \sum_{i=1}^{k} \alpha_{i\|_{R_{xyz}=0}} E(Z_i) + c_{x\|_{R_{xyz}=0}}$

Therefore, $E(Y) = \frac{E(X) - c_{x\|_{R_{xyz}=0}} - \sum_{i=1}^{k} \alpha_{i\|_{R_{xyz}=0}} E(Z_i)}{\alpha_{y\|_{R_{xyz}=0}}}$

8.3 Subroutines used in Algorithm 1

In this subsection we outline the subroutines $Recover\_\alpha$ and $Recove\_cov$ invoked by algorithm 1. Recoverability of queries in these subroutines are determined using results in section 4. Note that in this work we only deal with recoverability of statistical parameters. Path coefficients (which are causal parameters) are considered recoverable if (i) they are identifiable and (ii) all factors in the (identified) estimand are recoverable.

$\underline{Recover\_\alpha_y}$: **Input:** G: m-graph in which X_1 is a parent of Y and X_2 is a child of Y. α_y: Path coefficient of edge $X_1 \rightarrow Y$. **Output:** ${\alpha_y}^*$ if α_y is recoverable, $NULL$ otherwise.

This routine applies lemma 1 on X_1 and X_2 to obtain, $cov(X_1,X_2) = \sum_{i=1}^{k} var(A_{p_i}) * \prod_{j=1}^{n_{p_i}-1} \alpha_j^{p_i}$. If $cov(X_1,X_2)$, $var(A_{p_i})$'s and all α_j's (excluding α_y) are recoverable, their recovered estimands are substituted into the equation above to recover and return the estimand of α_y.

$\underline{Recover\_cov}$: **Input:** G: m-graph in which X_1 is a parent of Y and X_2 is a child of Y, ${\alpha_y}^*$: Recovered estimand of α_y, the path coefficient of edge $X_1 \rightarrow Y$. **Output:** Estimand for $cov(X_1,Y)$ if it is recoverable, $NULL$ otherwise.

This routine applies lemma 1 on X_1 and Y and returns the estimand for recovering $cov(X_1,Y)$ if all variances and path coeffients on the RHS of equation 2 are recoverable.