Title
Recoverability and testability of missing data: Introduction and summary of results

Permalink
https://escholarship.org/uc/item/4c4996s0

Authors
Pearl, Judea
Mohan, Karthika

Publication Date
2013-10-01

Recoverability and Testability of Missing data: Introduction and Summary of Results

(Based on slides presented by J. Pearl on August 6, 2013, at JSM-13, Montreal, Ca.)

Judea Pearl and Karthika Mohan

University of California, Los Angeles

Computer Science Department

Los Angeles, CA, 90095-1596, USA

(310) 825-3243

judea@cs.ucla.edu / karthika@cs.ucla.edu

October 29, 2013

Abstract

This is an expository paper, aimed to provide a gentle introduction to missing data problems as viewed from graphical modeling perspective. Aside from producing new theoretical results, the graphical perspective offers researchers a transparent language in which to understand, articulate and analyze missing data problems. Users can specify graphical features of their problems before choosing software or algorithms, and methodological researchers can use graph-based tools to develop software and algorithms that either exploit modeling assumptions or stand robust to such assumptions. The text of this paper is written around as set of 11 slides (marked 59-69) presented at the JSM-13 meeting, on August 6, 3013.

1

MISSING DATA:
A CAUSAL INFERENCE PERSPECTIVE
(Mohan, Pearl & Tian 2013)

- Pervasive in every experimental science.
- Huge literature, powerful software industry, deeply entrenched culture.
- Current practices are based on statistical characterization (Rubin, 1976) of a problem that is inherently causal.
- Needed: (1) theoretical guidance, (2) performance guarantees and (3) tests of assumptions.

Figure 1: Slide-59

Managing missing data is a problem in every experimental science. Sensors do not always work reliably, respondents do not fill out every question in the questionnaire, and medical patients are often unable to recall episodes, treatments or outcomes. The literature on this problem is huge and has resulted in a powerful software industry that makes missing data packages available through computer programs such as LISREL, M-plus and EQS. The availability of such software has engendered a culture that shares vocabulary, beliefs and expectations and uses common theoretical framework and default assumptions. Most practices are based on the seminal theoretical work of Rubin (Rubin, 1976; Little & Rubin, 2002) who have formulated procedures and conditions under which the damage of missingness can be minimized. This theory has also resulted in a number of performance guarantees when data obey certain statistical conditions. However, the theoretical guarantees provided by this theory are rather coarse, as will be shown in the discussion that follows.

Figure 2: Slide-60

Figure 2 explicates the kind of guidance and guarantees that are needed in missing data research(Pearl (2013)). Question Q1 refers to a researcher who has acquired a statistical package that handles missing data and would like to ask what the structure of his/her problem should be like for the procedure to produce an estimate that is consistent. (In this note we will be dealing only with the question of bias(consistency), and will assume therefore that infinitely large sample is available and that the user is concerned primarily with convergence to the right answer, rather than speed of convergence.)

Another question (Q2) that a user might ask is whether the problem at hand lends itself to solution by any method whatsoever. This is important because, if the answer is negative, then a biased result should definitely be expected with finite data and no software, however smart, can overcome this theoretical impediment. On the other hand, if the answer is affirmative, the user might next wish to ask whether another software can exploit the specific features of the problem so as to produce a consistent estimate. The third question (Q3) relates to testability, namely, once the user postulates a structure for the problem can the data tell us if the postulated structure is incorrect?

The first two questions address a problem we call "recoverability" and the third,

a problem called "testability." To answer these questions reliably the user must articulate features of the problem in some formal language, preferably in a model that captures both the inter-relationships among the variables of interest as well as the missingness process, i.e., explaining *why* some values are missing. While the theory of Rubin and Little (Little & Rubin, 2002) captures some of these relationships, the language they used is not sufficiently refined to capture details of the missingness process and to specify, for example, which variable is responsible for values missing in another. The characterization that emerged from this theory is likewise rather crude. Specifically, it divides problems into three categories: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR). In MAR for instance, missingness can only be explained by variables that are fully observed whereas those that are partially observed cannot be responsible for missingness in others; an unrealistic assumption in many cases.[1] Performance guarantees and some testability results are available for MCAR and MAR, while the vast space of MNAR problems has remained relatively unexplored.

The purpose of this note is to partition the space of MNAR problems along two orthogonal dimensions: (1) Recoverable vs non-recoverable and (2) Testable vs non-testable. Moreover, the partition will be query dependent. For example, in some problems we can obtain the consistent estimate of queries such as $P(Y|X)$ and $P(Y)$ but not of $P(X, Y)$. With the tools that we will develop, the user will be able to examine the features of his/her problem, determine whether a given query is recoverable (i.e., estimated bias-free from any given dataset (with missing values), that the model is capable of generating) and whether the assumptions that lead to such recoverability have testable implications.

---

[1] We are using a variable-based interpretation of MAR (Potthoff et al. (2006),Mohan et al. (2013)) which differs slightly from the event-based definition of Rubin (1976).

BASIC NOTIONS OF MISSING DATA: MISSINGNESS AND RECOVERABILITY

Estimate $E(X)$ from partially observed data:

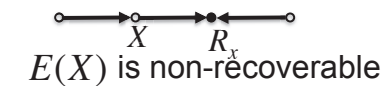$$X^* = [0,1,\ m,\ m,\ 1,\ 1,\ m,\ 0,\ m\ldots]\quad E(X)=?$$

True data:

$$X = [0,1,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 1\ldots]\quad E(X)=0.4$$

- The reasons for missingness can make a difference:
- Missingness mechanism: $R_x = 1$ if $X^* = m$, 0 otherwise

Model 1: (MCAR) $\quad X \perp\!\!\!\perp R_x \qquad \circ\!\!\longrightarrow\!\!\circ \quad \bullet\!\!\longleftarrow\!\!\circ$ coin
$$\underset{X}{}\quad\underset{R_x}{}$$
$$X^* = [0,1,m,m,1,1,m,0,m]\qquad \hat{X} = 3/5 \rightarrow \hat{E}(X)$$

Model 2: (MNAR) $\quad X \not\!\!\perp\!\!\!\perp R_x \qquad \circ\!\!\longrightarrow\!\!\circ\!\!\longrightarrow\!\!\bullet\!\!\bullet\!\!\longleftarrow\!\!\circ$
$$\underset{X}{}\quad\underset{R_x}{}$$
$$E(X) \text{ is non-recoverable}$$

Proof: Version (a): X is missing only if $X = 1$
$$X_a = [0,1,\ 1,\ 1,\ 1,\ 1,\ 1,\ 0,\ 1]\qquad \hat{X} = 7/9$$
$\qquad$ Version (b): X is missing only if $X = 0$
$$X_b = [0,1,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0]\qquad \hat{X} = 2/9$$

Figure 3: Slide-61

Figure 3 demonstrates the problem of determining recoverability and how missingness mechanisms affect that determination. Assume we observed the string X^* where each m stands for absence of a value and our task is to estimate $E(X)$ where X is a random variable whose actual values are shown in string X, below X^*. We will first demonstrate that different assumptions about the missingness process lead to different conclusions about $E(X)$. To articulate such assumptions we define a notion called "missingness mechanism" which stands for a binary variable R_x that acts like a switch. When the switch is activated ($R_x = 1$) we do not observe the value of X but rather the value $X^* = m$. When the switch is not activated ($R_x = 0$) we observe the correct value of X i.e. $X^* = X$. Each variable of interest will be assumed to have a missingness mechanism that, in general, may be activated by X itself, as well as by other variables in the model.

The simplest such mechanism is shown in Model-1 in which R_x is activated by a random coin and is shown to be unrelated to X. Formally, this independence is written: $X \perp\!\!\!\perp R_x$ and is implied by the absence of an arrow or any other graphical connection between X and R_x. Under this assumption, which falls under the category of MCAR, one can easily estimate the expected value of X. One need only examine the observed (unmasked) values of X^* and take their expectation. In our finite

example the answer will be $\hat{X} = \frac{3}{5}$ but when the number of samples increases the sample estimate will converge to $E(X)$. Let us now examine Model-2, in which a direct arrow is drawn from X to R_x. In other words, whether or not missingness is activated depends on the value of X. Such dependence, written $X \not\!\perp\!\!\!\perp R_x$, may represent for example, a salary survey in which people with high (or low) income are reluctant to reveal their income. Naturally this model falls under the MNAR category. In this model, since we do not know the exact dependence of R_x on X, and since the arrow $X \to R_x$ permits any imaginable dependency to exist, we cannot estimate $E(X)$ without bias, and we say that $E(X)$ is *non-recoverable*. We can actually prove this assertion by showing two kinds of dependence between X and R_x each yielding a different answer. In version-a of the model we assume an extreme case where X is missing only if its value is 1, and in version-b we assume that X is missing only if its value is 0. As we see, substituting 1 for m in version-a results in $\hat{X} = \frac{7}{9}$ while substituting 0 for m (in version-b) results in $\hat{X} = \frac{2}{9}$. If two different versions, both permitted by the model, yield conflicting results, it must be that the query $E(X)$ is non-recoverable; no algorithm in the world, however smart, can produce the correct answer $E(X) = 0.4$ without further assumptions, even assuming the model is correct. We can further see why recoverability is achievable in Model-1 and not in Model-2. Model-1 is more constrained than Model-2. While Model-2 permits any dependency between X and R_x to be realized, Model-1 insists on total independence, $X \perp\!\!\!\perp R_x$, which severely constrains the values that we can substitute for m as we try to reconstruct the distribution of X.
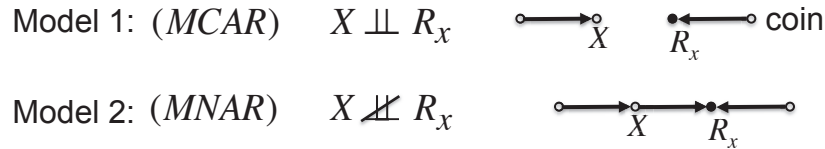
Figure 4: Slide-62

Figure 4 addresses the question that naturally surfaces in the mind of every reader; if we have two models, one permitting recoverability and the other not, why can't we test, from data itself, which model is more likely to be true? It turns out that, to the disappointment of many, Model-1 and Model-2 are *indistinguishable*. This might come as a surprise to readers familiar with graphical models because, after all, Model-1 has a missing arrow which stands for the independence $X \perp\!\!\!\perp R_x$, and independence is a property of the distribution that can be tested in data. Unfortunately, we are dealing with missing data; although R_x is observed unambiguously in the data, X is not, since it is contaminated with m's.

To show explicitly that Model-1 can emulate any data produced by the less constrained Model-2, let us go back to Figure-3 and examine X^*. It is always possible to replace the m's with one's and zero's stochastically, using the same distribution with which they appear in the non-missing values of X^*. The result would be a string that satisfies the independence claim advertised by the missing arrow in Model-1. This demonstrates our claim that Model-1, despite being constrained by the missing arrow, is not falsifiable by any data in which some values of X are missing. Any such data can be construed as coming from Model-1.

RECOVERABILITY AND TESTABILITY

Recoverability
Given a missingness model G and data D, when is a quantity Q estimable from D without bias?

Non-recoverability
Theoretical impediment to any estimation strategy

Testability
Given a model G, when does it have testable implications (refutable by some partially-observed data D')?

What is known about Recoverability and Testability?

$MCAR$	recoverable	testable (Little 1988)
MAR	recoverable	testable (Potthoff *et al* 2006)
$MNAR$	uncharted	uncharted

Figure 5: Slide-63

Figure 5 summarizes our discussion formally and emphasizes the fact that non-recoverability, if established in a given problem, constitutes an insurmountable impediment to any estimation strategy. Such finding should alert the user either to resign to the perils of biased results or to attempt to augment the model with auxiliary variables whose presence would alter the model structure and render the problem recoverable. Such strategy has indeed been proposed by many researchers (Collins et al. (2001); Graham (2003); Allison (2003); Enders (2010)). However, the question of which auxiliary variables are likely to increase or decrease bias is not well understood (Thoemmes & Rose, 2013) and is begging graphical analysis.

The last three lines of Figure 5 tabulate what is currently known in the missing data literature about recoverability and testability. We know that MAR and MCAR are recoverable for all probabilistic queries. Little (1988); Lin & Bentler (2012)(2012) has devised tests for refuting MCAR while Potthoff et al. (2006) devised one for MAR.[2] The latter can be strengthened by methods based on graphical methods (Mo-

---

[2]The test devised by Potthoff et al. (2006) applies to a subset of MAR which they call MAR^+, defined by conditional independencies among variables. We adopt Potthoff et al. (2006) in our analysis as do most authors of modern books on missing data because of the simplicity of characterizing this class. All but an exceptionally small fraction of data labeled MAR by Rubin are in MAR^+

han & Pearl (2013) (in preparation)). Remarkably, the vast territory of MNAR problems remains largely uncharted along both the recoverability and testability dimensions.
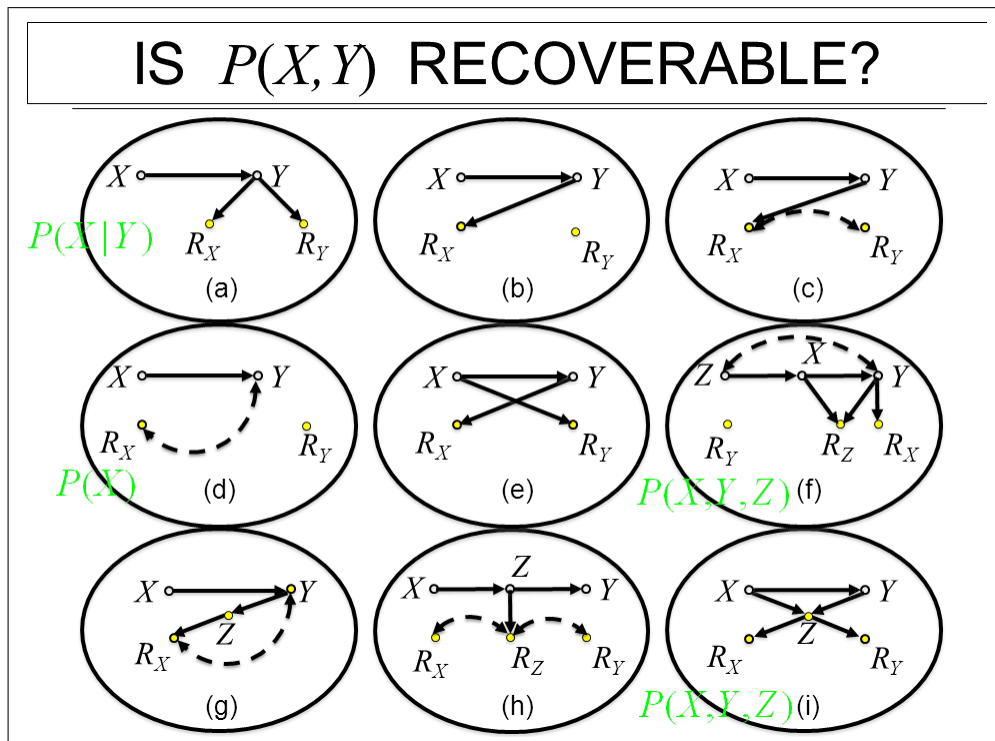

Figure 6: Slide-64

The following slide illustrates what we mean by charted territories. It contains nine missingness models and a query $P(X,Y)$ and the task is to determine, from each given structures whether $P(X,Y)$ is recoverable. (Here, solid circles represent fully observed variables and hollow circles represent partially observed variables, which are always accompanied by their respective R variables in the graph).

The technique that we have developed (Mohan et al., 2013), permits one to inspect the graphs and decide whether recoverability holds or not, for any given query. According to our criteria, one can categorically state that $P(X,Y)$ is recoverable in models (b), (c), (e), (f), (h) and (i), and non-recoverable in models (a),(d) and (g). A heuristic explanation follows:

In model (a), we see an arrow between Y and R_y which presents the same ambiguity we encountered in Figure 3 Model-2, hence it is non-recoverable.

In model (i) the connection between Y and R_y is intercepted by a fully observed variable Z. The pair X,Y becomes independent of the pair R_x, R_y which essentially turns the problem into MAR. Specifically, in every stratum of $Z = z$, the missingness of X,Y occurs completely at random.

category. The vast majority of probabilities generated by a model which we classify as MNAR would also be classified as MNAR according to Rubin(1976).

9

Model (b) has similar features except that no fully observed variable is present. Instead, the missingness in Y occurs totally at random and, conditional on Y, X is independent of its missingness mechanism R_x. This combination allows us to first recover $P(Y)$ from tuples in which Y is observed and then recover $P(X|Y)$ from tuples in which both X and Y are observed.

Almost identical behavior is exhibited by model (c) since the bi directed arrow $R_x < --- > R_y$ does not interfere with the recovery scheme just outlined. All independencies needed for this recovery scheme are authorized by the graph.

In model (e) we do not have a variable that is totally independent of its missingness mechanism. Instead, conditioning on Y will render X and R_x independent, while conditioning on X will render Y and R_y independent. Since both X and Y are partially unobserved it is not clear whether this structure will lend itself to recoverability. The theory nevertheless confirms recoverability of $P(X,Y)$ in this case and in many other so called "entangled" cases (Mohan et al., 2013).

Model (d) prohibits recoverability because to recover $P(X,Y)$ amounts to recovering both $P(X)$ and $P(X|Y)$. However, conditioning on Y renders X and R_x dependent which prohibits recoverability. Note that, in this example, $P(X)$ and $P(Y)$ are recoverable.

Model (h) permits recovery. The graph informs us that X and Y are jointly independent of R_x and R_y. Indeed, if we marginalize over Z and R_z, the remaining problem becomes MCAR.

Model (f) can be shown to permit recoverability by going through a sequence of conditionalizations as we did in (b). First we notice that Y is MCAR. Next we notice that conditioning on Y renders X and R_x independent. This makes P(X,Y) recoverable. (We can continue and condition on X and Y, thus rendering Z and R_z independent which leads to the recovery of $P(X,Y,Z)$.)

Next examine model (g). Attempting the same scheme as in (f) we find the following obstacle. Recovering $P(Y,Z)$ is trivial but, now, we observe an inducing path between X and R_x. Thus, conditioning on either Y or Z (or both) renders X and R_x dependent. The result is that there is no sequence of nodes that separate X from R_x. Hence $P(X,Y,Z)$ is non recoverable in (g).

THE RECOVERABILITY PIE
(and what's in it for the user)

MCAR

MAR

Sequential-*MAR*

(**M**)

(**S**)

(**N**)

Markovian*

Others $\underset{=}{?}$ {∅}

Non-recoverable
(by graphs)
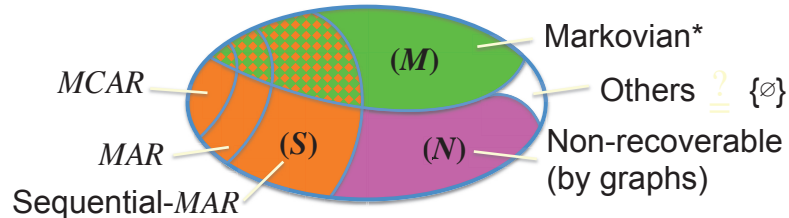
- Recoverability in *MAR* and *MCAR* models can be achieved by model-blind estimators (e.g., *MI* or *EM*)
- In areas (*M*) and (*S*), recoverability requires model-smart estimation
- Conjecture: *ML* on missing model recover areas (*S*) and (*M*)
- *ML* users receive no warning entering area (*N*) – They should!
- *MI* (Multiple Imputation) users navigating in areas (*S*) and (*M*) are likely to obtain biased results. Not to mention area (*N*)
- Testability – charted over the entire terrain

Figure 7: Slide-65

This slide shows MCAR and MAR as nested subsets of a class of problem instances which we call "Sequential-MAR" and which permit the recoverability of the joint distribution. Problems in this class are characterized by the following feature: There exists an ordering $X_1, X_2, \ldots, X_n$ of the variables such that for every i we have: $X_i \perp\!\!\!\perp (R_{X_i}, R_{Y_i}) | Y_i$ where $Y_i \subseteq \{X_{i+1}, .., X_n\}$. Such sequences yield the following estimand: $P(X) = \prod_i P(X_i | Y_i, R_{x_i} = 0, R_{y_i} = 0)$ and as we can see, every term in this product is estimable from the data.

The pie chart also represents a set of problems labeled as $Markov^*$ which permits recoverability. These problems are characterized by graphs with no hidden variables in which no variable X is a parent of its missingness mechanism R_x.

The area labeled N consists of models which are recognizable as non-recoverable. Examples are models in which X and R_x are connected by an edge or by an induced path (Verma & Pearl, 1991).

The area labeled "others" (which we conjecture to be empty) consists of problem structures that are non-Markovian and are devoid of features that enable us to prove non-recoverability with our current methods.

We now address the implications of this partition for a typical user. Most users prefer of course to ignore the structure of the problem and use a "model-blind"

estimator, namely, a universal estimator that is driven by data alone and does not change its strategy as we move from structure to structure. If data are MAR or MCAR, this requirement is satisfied, which explains the popular use of model-blind estimators such as MI or EM under the assumption of MAR; the specific features of the problem at hand need not be attended to and the user is spared the effort of choosing and defending a specific graph structure. In areas (M) and (S) however, such effort is unavoidable; recoverability requires model smart estimators. We will actually prove this requirement in a following slide.

The popular estimator known as ML suffers from the same deficiency unless it is applied to the full model, namely, the substantive part and the missingness part. We conjecture that ML would recover the joint probability in areas (S) and (M) if conducted on the full missingness model. This is rarely done in practice. Users are rarely requested to specify the missingness part of the model. Consequently, ML users receive no warning when facing a problem in area N. Given our criterion for N, such a warning can easily be produced by mere inspection of the model structure. A more stringent warning should await MI users, since the danger may also be present in the minefield of areas (M) and (S). Even problems that are recoverable require knowledge of the model structure, without which biased results are likely to be produced.

The final result we report in this summary is that testability has been charted over the entire terrain of missing data problems. In other words, given a model structure, one can discern whether the model yields testable implications. Some of these tests are powerful enough to rule out MAR data and placing the problem in the MNAR category.

Figure 8: Slide-66
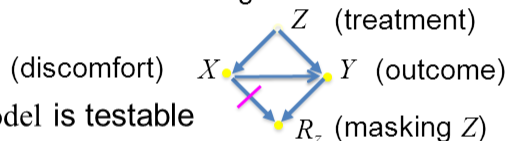
Speaking about testability, it is important to note a peculiar phenomenon that takes place in missing data analysis. Consider the example illustrated in Figure 8. It is easily shown that the problem is MAR, hence $P(X, Y, Z, R_Z)$ is recoverable. This is so because the fully observed variables X and Y "explain" the missingness in the third variable Z. Moreover, the recovered probability distribution embeds the conditional independence claim $Z \perp\!\!\!\perp R_z | (X, Y)$, which would be testable had Z been fully observed. Unfortunately, the fact that Z is only partially observed prevents this independence claim from being testable in the available data. In other words, any data whatsoever with Z partially observed is compatible with the model above and, so, no such data can falsify the conditional independence $Z \perp\!\!\!\perp R_z | (X, Y)$.

We shall now give a very simple syntactic criteria to determine if a conditional independence $A \perp\!\!\!\perp B | C$ is testable with $|A| = |B| = 1$ where A, B and C are allowed to include not only variables but also missingness mechanisms. The syntactic rule states that a conditional independence is testable if it has one of the following forms:

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

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

3. $R_x \perp\!\!\!\perp R_y | Z, R_z$

(It is understood that, if X or Y or Z are fully observed, the corresponding missingness mechanism may be removed from the conditioning set. Clearly, any conditional

independence comprised exclusively of fully observed variables is testable.) This rule, combined with the fact that all conditional independencies claimed by a graph can be reduced to pair-wise independencies permits us to determine whether the model, as a whole, is testable.

To illustrate the power of this criterion note that if we remove the $X \rightarrow R_z$ edge from the diagram in Figure 8, the model becomes testable, because $X \perp\!\!\!\perp R_z | Y$ complies with (2) above.

AN IMPOSSIBILITY THEOREM FOR MISSING DATA

(a) Accident   Injury

(b) Injury   Treatment

Education

X Y R_x (latent)

Missing (X)   Missing (X)

- Two statistically indistinguishable models, yet $P(X,Y)$ is recoverable in (a) and not in (b).

- No universal algorithm exists that decides recoverability (or guarantees consistent estimates) without looking at the model.
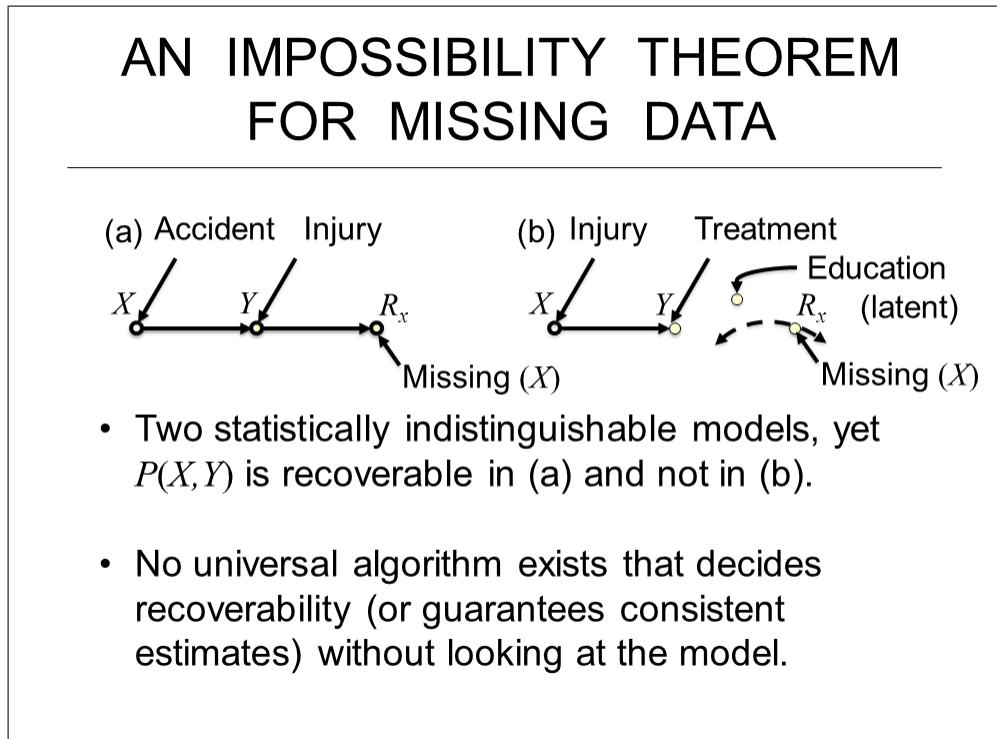
Figure 9: Slide-67

The indistinguishability of the two models presented in Figure 9 can be verified by our testability rules showing that none has testable implications. The fact that one of these models permits recoverability while the other does not, establishes the need for model specific algorithms in the MNAR domain.

Figure 10: Slide-68

This necessity is further strengthened in Figure 10, showing that, even if one settles on receiving a consistent estimate only when such exists, no universal algorithm exists that can offer such guarantee. Figure 10 shows a query that is recoverable by two indistinguishable models, yet each dictates a different procedure for recoverability, and a different estimand resulting from each procedure. If we run a procedure informed by M1 on M2 (or vice-versa), a biased estimate will ensue. This is precisely the case in region (S) of the pie-chart; models in this region permit recoverability but, at the same time, the estimates produced are sensitive to model structure.

15

Figure 11: Slide-69

References

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

Collins, Linda M, Schafer, Joseph L, and Kam, Chi-Ming. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 6(4):330, 2001.

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

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

Lin, J.C. and Bentler, P.M. A probability based test for missing completely at random data patterns, 2012. Paper presented at the meeting of the National Council on Measurement in Education, Vancouver, Canada, April.

Little, R.J.A. and Rubin, D.B. *Statistical analysis with missing data.* Wiley, 2002.

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

Mohan, K. and Pearl, J. On the testability of m-graphs. Technical Report R-415, UCLA, 2013. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r415.pdf.

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

Pearl, J. The mathematics of causal inference. Technical Report R-416, UCLA, 2013. Forthcoming, Proceedings of the Joint Statistical Meetings Conference, 2013. Available at http://ftp.cs.ucla.edu/pub/stat_ser/r416.pdf.

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

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

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

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