Testing compatibility of causal models with (historical) data under missingness not at random

Nick Gardner

Stanford University

Primer on missingness graphs and MCAR/MAR/MNAR

Missingness graphs (m-graphs) are a type of causal DAG developed to model missing data problems [6, 5]. **Nodes** (variables) in an m-graph come in five different types:

- $\boldsymbol{V_o}$ – the set of variables that are observed for every item in the sample, which we call **fully observed variables**;
- $\boldsymbol{V_m}$ – the set of variables that have a missing value for at least one item in the sample, which we call **partially observed variables**;
- $\boldsymbol{U}$ – the set of completely **unobserved variables** (also called latent or hidden variables);
- $\boldsymbol{R}$ – a set of indicator variables representing the **missingness mechanisms** that cause values in our data to be missing. For every variable V_i in the set of partially observed variables $\boldsymbol{V_m}$, there is a corresponding $R_{V_i} \in \boldsymbol{R}$ that represents the missingness mechanism acting on V_i;
- $\boldsymbol{V^*}$ – a set of **proxy variables** that represent what we actually observe of the partially observed variables. There is one such variable V_i^* associated with every V_i in the set of partially observed variables $\boldsymbol{V_m}$. The value of each proxy variable is (non-stochastically) determined by the values of its associated missingness mechanism R_{V_i} and underlying variable V_i

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

We refer to variables in $\boldsymbol{V_o} \cup \boldsymbol{V_m}$ as **substantive variables** (i.e. the variables we are primarily interested in modeling).

The **edges** (causal influences) in an m-graph are usually subject to the following additional restriction: missingness indicators causally influence only proxy variables and other missingness indicators, not substantive or unobserved variables. Graphically, an edge from $\boldsymbol{R}$ can only go to $\boldsymbol{V^*}$ or stay inside $\boldsymbol{R}$ (without creating a cycle).

Using m-graphs we can easily define **Rubin's missing data categories** [10] in terms of the conditional independence relations that hold between the missingness mechanisms $\boldsymbol{R}$ and the substantive variables (see fig. 1):

MCAR: Data are "missing completely at random" or MCAR if $\boldsymbol{V_o} \cup \boldsymbol{V_m} \perp\!\!\!\perp \boldsymbol{R}$ holds in the m-graph. In words, the missingness mechanisms are *unconditionally independent* of all substantive variables (i.e. both fully and partially observed). Values in partially observed variables are "missing completely at random" with respect to everything measured about the sample items. Graphically, MCAR holds if no arrows exist from any partially or fully observed variable to a missingness mechanism ($\boldsymbol{R}$ variable)

MAR: Data are "missing at random" or MAR if $\boldsymbol{V_m} \perp\!\!\!\perp \boldsymbol{R} | \boldsymbol{V_o}$ holds in the m-graph. In words, the missingness mechanisms are *independent* of the partially observed variables *conditional on* some collection of fully observed variables. Values in partially observed variables are "missing at random" conditional on the observed variables. Graphically, MAR holds if both (i) no arrows exist from any partially observed variable to a missingness mechanism ($\boldsymbol{R}$ variable) and (ii) no fully observed variable shares an unobserved common cause U with a missingness mechanism. (Note that MCAR is a special case of MAR.)

MNAR: Data that are not MCAR or MAR are "missing not at random" or MNAR. This means some missingness mechanism depends causally on a partially observed variable, or shares an unobserved common cause U with a fully observed variable (or both). In words, some data are systematically missing in the following strong sense: there is no collection of fully observed variables which we can stratify on to render (within each stratum/level) the missingness mechanisms independent of the partially observed variables. The missingness is "not at random (conditionally or unconditionally)." Graphically, MNAR holds if either or both of the graphical conditions for MAR are violated.

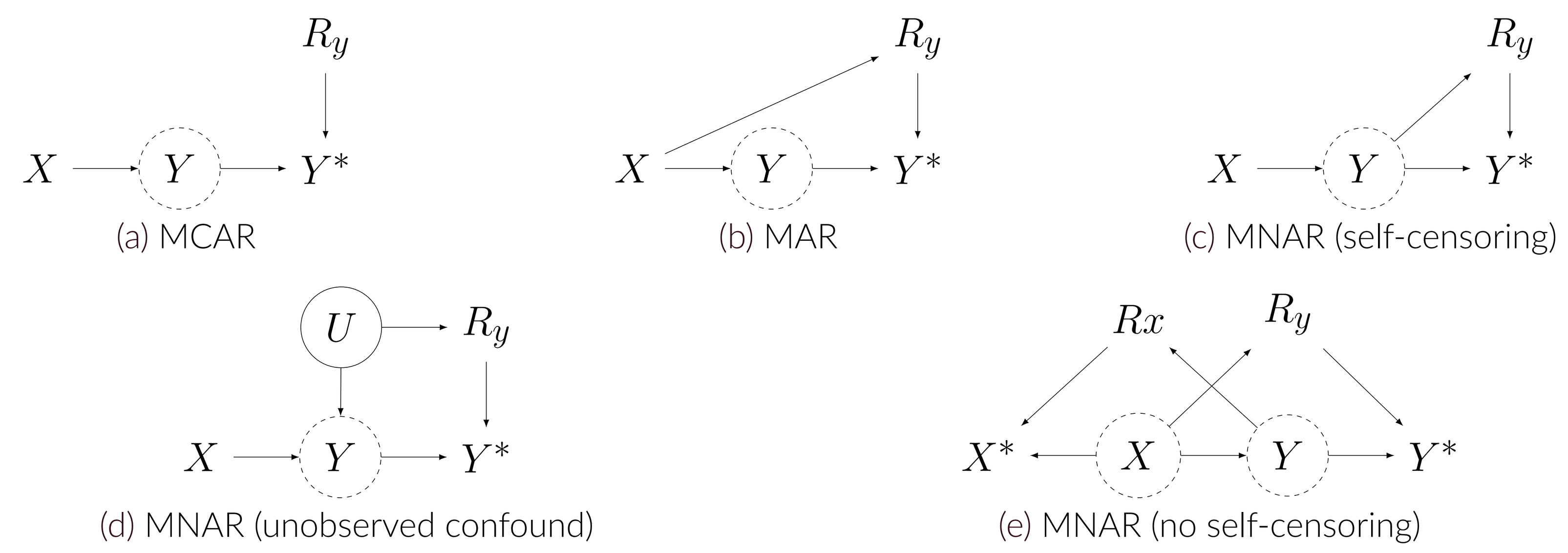

Figure 1. A family of simple m-graphs illustrating MCAR, MAR, and three types of MNAR models. Notation: variables surrounded by a solid circle are fully unobserved (elements of $\boldsymbol{U}$), variables surrounded by a dashed circle are partially observed (elements of $\boldsymbol{V_m}$), and uncircled variables are fully observed (elements of $\boldsymbol{V_o} \cup \boldsymbol{V^*} \cup \boldsymbol{R}$).

Recoverability and Estimation vs. Testability and Testing

- **Recoverability:** Assuming a causal model is correct (in the weak sense of being statistically compatible with data it is meant to model), which of its parameters can in principle be estimated given unlimited sample data (i.e. which parameters have consistent estimators)?
- **Estimation:** What are efficient methods for estimating recoverable parameters from finite samples?
- **Testability:** Can the overall assumption that a given model is correct be tested, in principle given unlimited sample data? Can the specific qualitative assumptions that make up a model be tested individually (so that if a model is incorrect, we could identify and repair the assumptions that made it incorrect)?
- **Testing:** What are efficient methods for testing compatibility of models with finite samples of data?

Testability under missingness: open questions, future directions

- Complete algorithm for finding **conditional independence constraints** (i.e. *testable* conditional independence claims) entailed by an m-graph about the observed data distributions it models [4, 5].
- Complete algorithm for finding **Verma or other equality constraints** [12, 11, 9] in missingness graphs. See recent work [7].
- (current work) Any algorithm for finding **instrumental or other inequality constraints** ("Bell inequality style" constraints) in missingness graphs, moving beyond the latent variable models where they were first discovered/studied [8, 2, 13].
- More **efficient tests** for the above constraints. Recent work explores tests based on weighted likelihood ratios and odds-ratio parameterizations of joint distributions [7].
- Algorithms for **causal structure learning under missingness**: given data with missing values, what is the set of all (or at least many of) the m-graphs compatible with the data? Is there a parsimonious/meaningful representation of this set that provides a unified characterization of the observed data? On structure learning under measurement error see [16].
- (speculative) Relate m-graph/causal inference work on recoverability, testability, and structure learning, to the **methods of information theory and coding theory** developed to study analogous problems: error correction and/or detection algorithms for variety of noise and erasure channels (classical, quantum, hybrid); algorithms for learning channels from observed input-output data.

Example application: m-graphs for ancient history research

Positive result (inferring more from existing evidence)

I use m-graphs and statistical inference in my dissertation to (approximately) answer for the first time many basic quantitative questions about the political and occupational demography of Ancient Athens (Greek city-state, focus on period 5th-3rd c. BCE). (Details and m-graphs in talk slides.)

Negative result (clarifying what can/cannot be inferred from existing evidence)

- An innovative historical paper published in Nature was retracted [14, 1, 15] because authors mishandled missingness in their dataset (fig. 2 plots the relevant missingness).
- Causal modelling using m-graphs both (1) clarifies the limits imposed missingness in this data and (2) suggests valid ways to proceed.
- Upshot: any historically plausible m-graph for the data shows that evidence about moralizing gods is missing not at random in a way that fatally underdetermines inferences the authors made. Conceptually the problem is one of **testability (under missingness)**: m-graphs which avoid being testably incompatible with this data are not historically plausible.
- Use of m-graphs to frame questions in terms of testability, both heuristically and (where possible) formally, can significantly advance historical research.

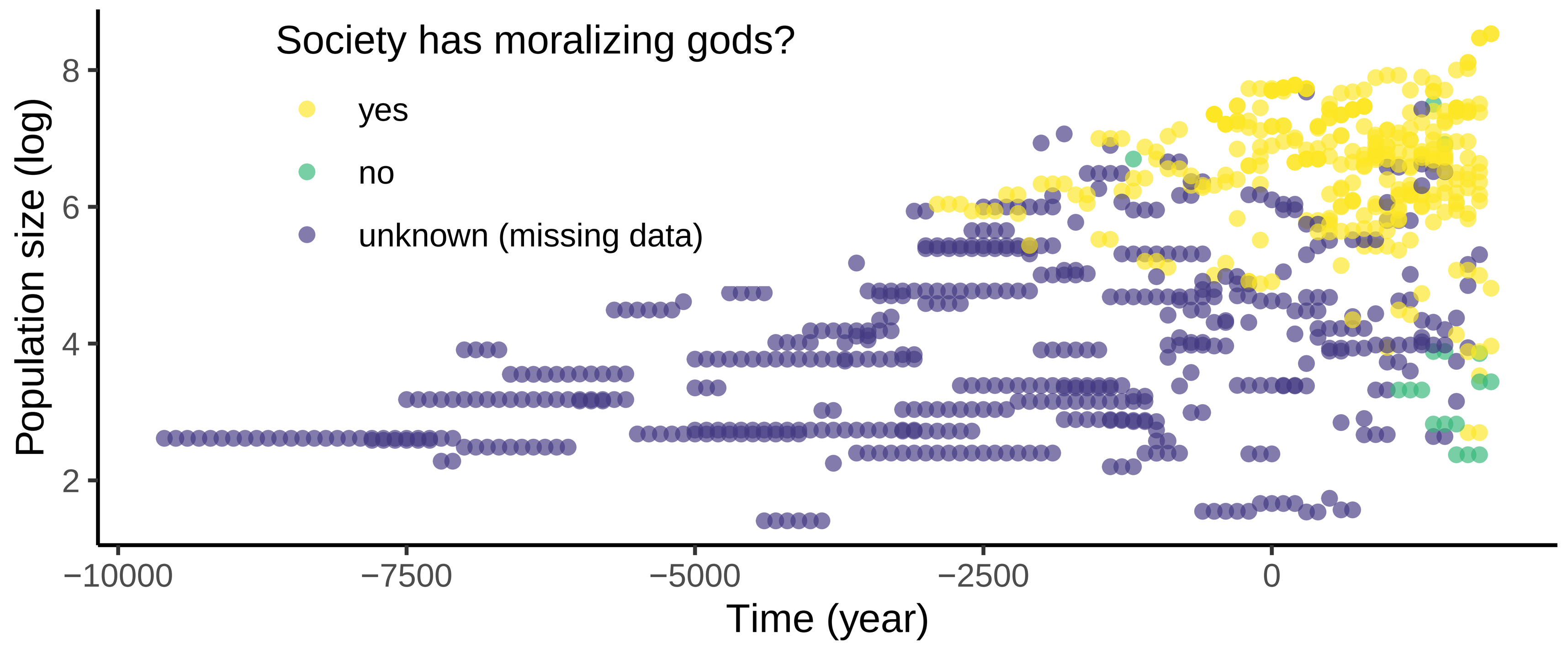

Figure 2. Societies plotted by estimated year and estimated population (on logarithmic scale), and colored according to the status of evidence about moralizing gods. Note the strong associations between missingness and both population size and time period. Plot adapted from [3], which has an excellent high-level overview of the retracted paper's missing data problems.

References

[1] Bret Beheim, Quentin D Atkinson, Joseph Bulbulia, Will Gervais, Russell D Gray, Joseph Henrich, Martin Lang, M Willis Monroe, Michael Muthukrishna, Ara Norenzayan, et al. Treatment of missing data determined conclusions regarding moralizing gods. *Nature*, 595(7866):E29–E34, 2021.

[2] Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 48–55, 2001.

[3] Richard McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*. Chapman and Hall/CRC Texts in Statistical Science. CRC Press LLC, Milton, 2020. ISBN 036713991X.

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

[5] Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

[6] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. *Advances in neural information processing systems*, 26, 2013.

[7] Razieh Nabi and Rohit Bhattacharya. On testability and goodness of fit tests in missing data models. *arXiv preprint arXiv:2203.00132*, 2022.

[8] Judea Pearl. On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 435–443, 1995.

[9] Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

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

[11] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*, pages 519–527, 2002.

[12] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270, 1990.

[13] Linbo Wang, James M Robins, and Thomas S Richardson. On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236, 2017.

[14] Harvey Whitehouse, Pieter Francois, Patrick E Savage, Thomas E Currie, Kevin C Feeney, Enrico Cioni, Rosalind Purcell, Robert M Ross, Jennifer Larson, John Baines, et al. Retracted article: Complex societies precede moralizing gods throughout world history. *Nature*, 568(7751):226–229, 2019.

[15] Harvey Whitehouse, Pieter François, Patrick E Savage, Thomas E Currie, Kevin C Feeney, Enrico Cioni, Rosalind Purcell, Robert M Ross, Jennifer Larson, John Baines, et al. Retraction note: Complex societies precede moralizing gods throughout world history. *Nature*, 595(7866):320–320, 2021.

[16] Yuqin Yang, AmirEmad Ghassami, Mohamed Nafea, Negar Kiyavash, Kun Zhang, and Ilya Shpitser. Causal discovery in linear latent variable models subject to measurement error. *arXiv preprint arXiv:2211.03984*, 2022.

ngardner@stanford.edu