Graphical Models for Recovering Probabilistic and Causal Queries from Missing Data,

with K. Mohan

Mowhebat Bazargani

31 May 2021

Introduction

- Mohan, K., and Pearl, J. (2014a), “Graphical Models for Recovering Probabilistic and Causal Queries From Missing Data,” in Advances in Neural Information Processing Systems (Vol. 27), eds. Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Weinberger, Curran Associates, Inc., pp. 1520–1528.

- They extend the results of Mohan et al. [2013] by presenting more general conditions for recovering probabilistic queries of the form P(y|x) and P(y, x) as well as causal queries of the form P(y|do(x)).
  - *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.*

Introduction (cont'd)

- Missing data in all branches of experimental science
- Improper handling of missing data can
  - bias outcomes and
  - potentially distort the conclusions drawn from a study.
- Therefore, accurate diagnosis of ***the causes of missingness*** is crucial for the success of any research.

Table of Contents

- An overview of missingness graphs and reviews the notion of recoverability.
- The sequential factorization theorem presented in Mohan et al. [2013] and extends its applicability to a wider range of problems in which missingness mechanisms may influence each other.
- We present general algorithms to recover joint distributions from the class of problems for which sequential factorization theorem fails.
- We discuss recoverability of causal queries and show that unlike probabilistic queries, $P(y|do(x))$ may be recovered even when Y and its missingness mechanism (R_y) are not d-separable.
- We demonstrate how we can apply our results to problems of attrition in which missingness is a severe obstacle to sound inferences.

Missingness Graphs

- Let G($\mathbb{V}$, E) be the causal DAG where $\mathbb{V} = V \cup U \cup V^* \cup \mathbb{R}$.
- E is the set of edges in the DAG.
- V is the set of observable nodes.
- Nodes in the graph correspond to variables in the data set.
- V is partitioned into V_o and V_m such that
  - $V_o \subseteq \mathbb{V}$ is the set of variables that are observed in all records in the population
  - $V_m \subseteq \mathbb{V}$ is the set of variables that are missing in at least one record.
- Variable X is termed as
  - *fully observed* if $X \in V_o$,
  - *partially observed* if $X \in V_m$ and
  - *substantive* if $X \in V_o \cup V_m$

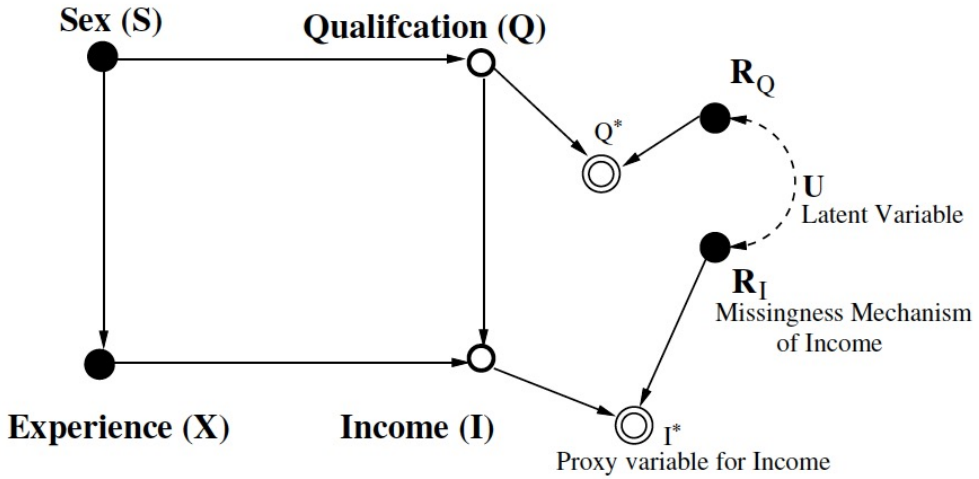
Missingness Graphs (cont'd)

- $\mathbb{V} = V \cup U \cup V^* \cup \mathbb{R}$
- U is the set of unobserved nodes (also called latent variables).
- Associated with every partially observed variable $V_i \in V_m$ are two other variables R_{v_i} and V_i^*, where
  - V_i^* is a proxy variable that is observed, and
  - R_{v_i} represents the status of the causal mechanism responsible for the missingness of V_i (missingness mechanism); formally,

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

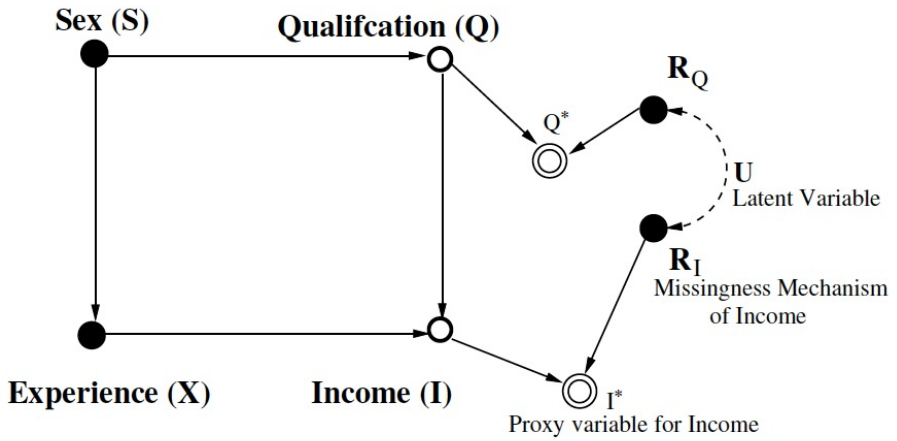
An example of an m-graph

- Missingness Graph depicting the missingness process in a hypothetical (job-specific) gender wage gap study that measured the variables: sex (S), work experience(X), qualification(Q) and income(I).
- Typical m-graph where $V_o = \{S, X\}$, $V_m = \{I, Q\}$, $V^* = \{I^*, Q^*\}$, $R = \{R_i, R_q\}$ and U is the latent common cause.



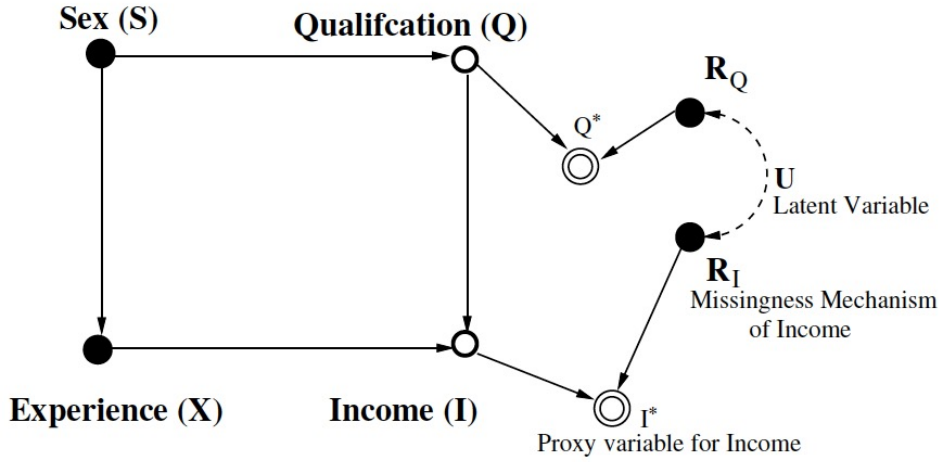
An example of an m-graph (cont'd)

- Members of V_o and V_m are represented by full and hollow circles respectively.
- We use bi-directed edges as a shorthand notation to denote the existence of a U variable as common parent of two variables in $V \cup \mathbb{R}$.



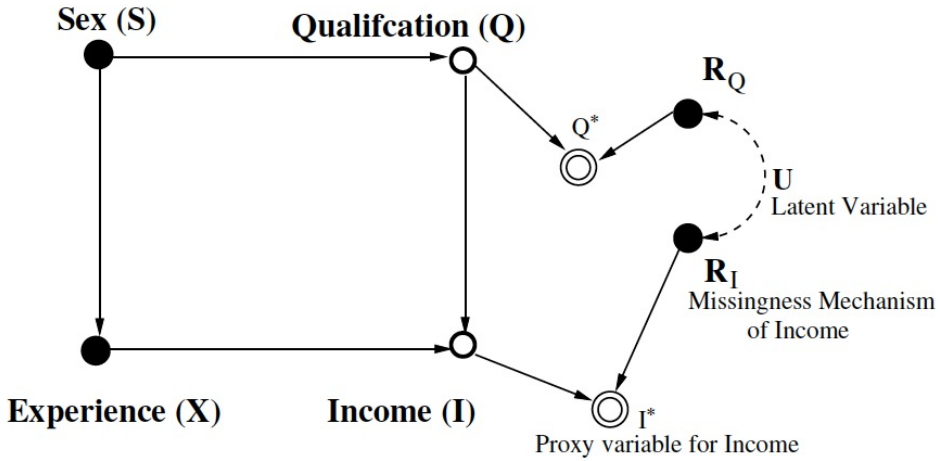
An example of an m-graph (cont'd)

- The assumptions in the model are:
  1) women are likely to be less qualified and experienced than men,
  2) income is determined by qualification and job experience of the candidate,
  3) missingness in Q and I are correlated, caused by unobserved common factors such as laziness or resistance to respond.



An example of an m-graph (cont'd)

- Sex and work experience were found to be fully observed in all records i.e. $V_o = \{S, X\}$.
- Qualification and income were found to be missing in some of the records i.e. $V_m = \{I, Q\}$.
- R_Q and R_I denote the causes of missingness of Q and I respectively and are assumed to be independent of S, Q, I and X.



Distributions in m-graph

- A manifest distribution $P(V_o, V^*, \mathbb{R})$ is the distribution that governs the available dataset.

- An underlying distribution $P(V_o, V_m, \mathbb{R})$ is said to be ***compatible*** with a given manifest distribution $P(V_o, V^*, \mathbb{R})$ if the latter can be obtained from the former using this equation.

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

Distributions in m-graph (cont'd)

- Manifest distribution → $P(V_o, V^*, \mathbb{R})$
- Underlying distribution → $P(V_o, V_m, \mathbb{R})$

- Manifest distribution P_m is compatible with a given underlying distribution P_u if $\forall X$, $X \subseteq V_m$ and $Y = V_m \setminus X$, the following equality holds true.

$$P_m(V_o, X^*, Y^*, R'_x, R_y) = P_u(V_o, X, R'_x, R_y)$$

- where R'_x denotes $R_x = 0$ and R_y denotes $R_y = 1$.

Testing compatibility between underlying and manifest distributions

- Let the incomplete dataset contain two partially observed variables, Z and W. The tests for compatibility between manifest distribution: $P_m(Z^*, W^*, R_z, R_w)$ and the underlying distribution: $P_u(Z, W, R_z, R_w)$ are:

  - Case-1: Let X = {Z, W}, then Y = $V_m \setminus X = \{\}$

    $P_m(Z^* = z, W^* = w, R_z = 0, R_w = 0) = P_u(Z = z, W = w, R_z = 0, R_w = 0)\ \forall z, w$
  - Case-2: Let X = {Z}, then Y = $V_m \setminus X = \{W\}$

    $P_m(Z^* = z, W^* = m, R_z = 0, R_w = 1) = \sum_w P_u(Z = z, w, Rz = 0, Rw = 1)\ \forall z$
  - Case-3: Let X = {W}, then Y = $V_m \setminus X = \{Z\}$

    $P_m(Z^* = m, W^* = w, R_z = 1, R_w = 0) = \sum_z P_u(z, W = w, Rz = 1, Rw = 0)\ \forall w$
  - Case-4: Let X = {}, then Y = $V_m \setminus X = \{Z, W\}$

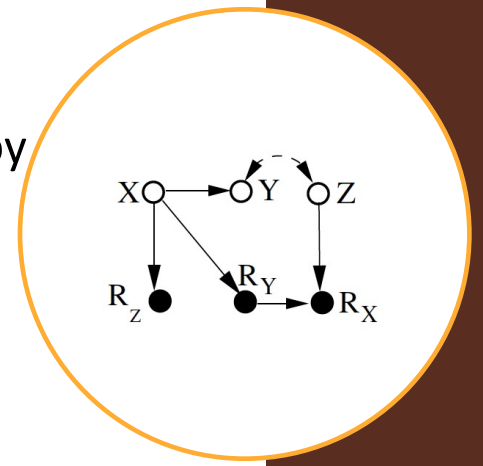
    $P_m(Z^* = m, W^* = m, R_z = 1, R_w = 1) = \sum_{z,w} P_u(z, w, Rz = 1, Rw = 1)$

Recoverability

- Given a manifest distribution $P(V_o, V^*, \mathbb{R})$ and an m-graph G that depicts the missingness process, query Q is **recoverable** if we can compute *a consistent estimate of Q* as if no data were missing.
- Definition 1 (Recoverability (Mohan et al. [2013])). Given an m-graph G, and a target relation Q defined on the variables in V, Q is said to be recoverable in G if there exists an algorithm that produces *a consistent estimate of Q* for every dataset D such that P(D) is
  - (1) compatible with G
  - (2) strictly positive i.e. $P(V_o, V^*, \mathbb{R}) > 0$.

Recovering Probabilistic Queries by Sequential Factorization

- Mohan et al. [2013] (theorem-4) presented a sufficient condition for recovering probabilistic queries such as joint and conditional distributions by using ordered factorizations.
- However, the theorem is not applicable to certain classes of problems such as those in which edges exist between R variables.
- General ordered factorization defined in this work broadens the concept of ordered factorization (Mohan et al. [2013]) to include the set of R variables.
- Subsequently, the modified theorem (theorem 1) will permit us to handle cases in which R variables are contained in separating sets that d-separate partially observed variables from their respective missingness mechanisms (example: $X \perp\!\!\!\perp R_x \mid R_y$ in this figure).



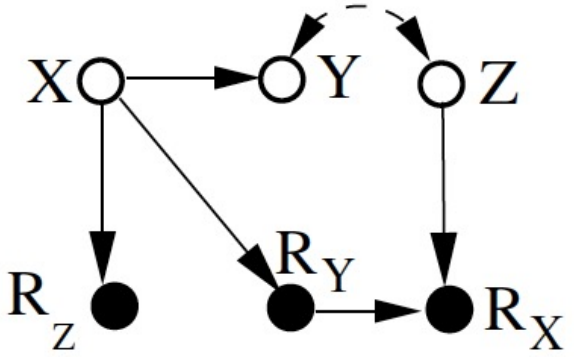
Definition 2 (General Ordered factorization)

- Given a graph G and a set O of ordered $V \cup R$ variables $Y_1 < Y_2 < \ldots < Y_k$,
- a general ordered factorization relative to G, denoted by f(O), is a product of conditional probabilities $f(O) = \prod_i P(Y_i | X_i)$
  - where $X_i \subseteq \{Y_{i+1}, \ldots, Y_n\}$ is a *minimal set* such that $Y_i \perp\!\!\!\perp (\{Y_{i+1}, \ldots, Y_n\} \setminus X_i) | X_i$ holds in G.

Example 1 (definition 2)

- $f(O) = \prod_i P(Y_i|X_i)$
  - where $X_i \subseteq \{Y_{i+1}, \ldots, Y_n\}$ is a *minimal set* such that $Y_i \perp\!\!\!\perp (\{Y_{i+1}, \ldots, Y_n\} \setminus X_i)|X_i$ holds in G.
- We are interested in recovering P(X, Y, Z) given the m-graph in this figure.
- There is no ordered factorization over the substantive variables {X, Y, Z} that will permit recoverability of P(X, Y, Z) .
  - We need to introduce R_y into the order.
- We discern from the graph that definition 2 is satisfied because:
  1) $P(Y|X, Z, R_y) = P(Y|X, Z)$ and (X, Z) is a *minimal set* such that $Y \perp\!\!\!\perp (\{X, Z, R_y\} \setminus (X, Z)) \mid (X,Z) \rightarrow Y \perp\!\!\!\perp R_y \mid (X,Z)$
  2) $P(X|R_y, Z) = P(X|R_y)$ and R_y is the *minimal set* such that $X \perp\!\!\!\perp (\{R_y, Z\} \setminus R_y) \mid (R_y) \rightarrow X \perp\!\!\!\perp Z \mid R_y$
  3) $P(Z|R_y) = P(Z)$ and $\emptyset$ is the *minimal set* such that $Z \perp\!\!\!\perp R_y \mid \emptyset$.
- Therefore, the order $Y < X < Z < R_y$ induces a general ordered factorization, $f(O) = \prod_i P(Y_i|X_i)$:

$$P(X, Y, Z, R_y) = P(Y \mid X, Z)P(X \mid R_y)P(Z)P(R_y)$$



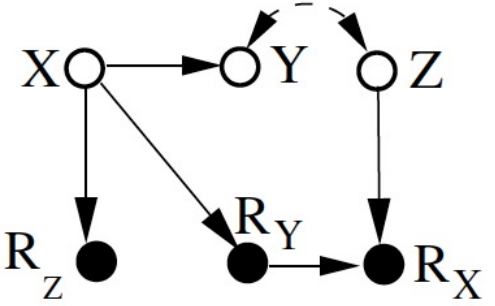
Theorem 1 (Sequential Factorization)

- A sufficient condition for recoverability of a relation Q defined over substantive variables is that Q be decomposable into a general ordered factorization, or a sum of such factorizations, such that every factor $Q_i = P(Y_i|X_i)$ satisfies
  1) $Y_i \perp\!\!\!\perp (R_{x_i}, R_{y_i}) | X_i \setminus \{R_{x_i}, R_{y_i}\}$, if $Y_i \in (V_o \cup V_m)$
  2) $R_z \perp\!\!\!\perp R_{x_i} | X_i$, if $Y_i = R_z$ for any $Z \in V_m$, $Z \notin X_i$, and $X_i \cap R_{x_i} = \emptyset$.

- An ordered factorization that satisfies the conditions in Theorem 1 is called an admissible sequence.

Example 1 (theorem 1)

- First condition: $Y_i \not\perp\!\!\!\perp (R_{x_i}, R_{y_i}) \mid X_i \setminus \{R_{x_i}, R_{y_i}\}$, *if* $Y_i \in (V_o \cup V_m)$
  - $P(Y \mid X, Z) \rightarrow Y \perp\!\!\!\perp \{R_y, R_x, R_z\} \mid (X,Z)$
  - $P(X \mid R_y) \rightarrow X \perp\!\!\!\perp R_x \mid R_y$
  - $P(Z) \rightarrow Z \perp\!\!\!\perp R_z$
- Second condition: $R_z \not\perp\!\!\!\perp R_{x_i} \mid X_i$, *if* $Y_i = R_z$ *for any* $Z \in V_m$, $Z \notin X_i$, *and* $X_i \cap R_{x_i} = \emptyset$.
  - $P(R_y) \rightarrow Y_i = R_y, X_i = \emptyset$



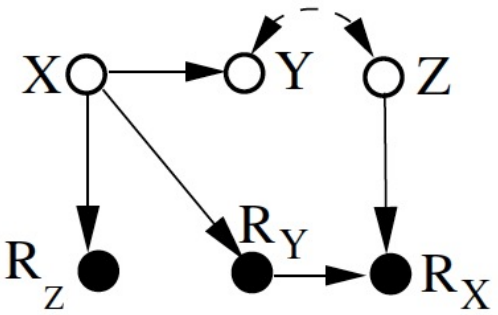
- Theorem 1 holds true!
  - $Y < X < Z < R_y$ is an admissible sequence.
  - It's recoverable!

Example 1 (cont'd)

- $P(X, Y, Z, R_y) = P(Y \mid X, Z)P(X \mid R_y)P(Z)P(R_y)$
- We now rewrite P(X, Y, Z) as follows:

$$P(X, Y, Z) = \sum_{R_y} P(X, Y, Z, Ry) = P(Y \mid X, Z)P(Z)\sum_{R_y} P(X|Ry)P(Ry)$$

- $P(X, Y, Z) = P(Y|X, Z, R'_x, R'_y, R'_z)P(Z|R'_z) \sum_{R_y} P(X|R'_x, R_y)P(Ry)$
  - $Y \perp\!\!\!\perp \{R_y, R_x, R_z\} \mid (X,Z)$
  - $X \perp\!\!\!\perp R_x \mid R_y$
  - $Z \perp\!\!\!\perp R_z$



Example 1 (cont'd)

- P(X, Y, Z) = P(Y|X, Z, R'_x , R'_y, R'_z)P(Z|R'_z) $\sum_{R_y}$ P(X|R'_x , R_y)P(Ry)
- Equation (1)

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

- Indeed, equation 1 permits us to rewrite it as:

P(X, Y, Z) = P(Y^*|X^*, Z^*, R'_x , R'_y, R'_z)P(Z^*|R'_z) $\sum_{R_y}$ P(X^*|R'_x , R_y)P(Ry)

- P(X, Y, Z) is recoverable because every term in the right hand side is consistently estimable from the available dataset.

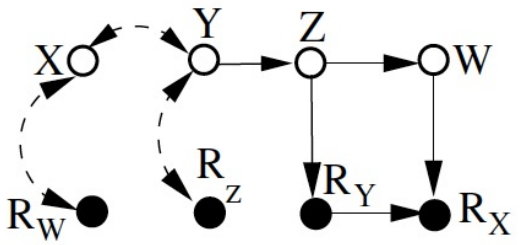
Recoverability in the Absence of an Admissible Sequence

- Mohan et al. [2013] presented a theorem that stated the necessary and sufficient condition for recovering the joint distribution for the class of problems in which the parent set of every R variable is a subset of $V_o \cup V_m$.

- In contrast to Theorem 1, their theorem can handle problems for which no admissible sequence exists.

Theorem 2. (Explanation)

- In theorem 2:
- (i) collider path p between any two nodes X and Y is a path in which every intermediate node is a collider. Example, X -> Z <-> Y .
- (ii) $R^{part} = \{R^{(1)}, R^{(2)}, \ldots, R^{(N)}\}$ are partitions of R variables such that for every element R_x and R_y belonging to distinct partitions, the following conditions hold true:
  - (i) R_x and R_y are not neighbors and
  - (ii) R_x and R_y are not connected by a collider path.
  - In this figure: $R^{part} = \{R^{(1)}, R^{(2)}\}$ where $R^{(1)} = \{R_w, R_z\}$, $R^{(2)} = \{R_x, R_y\}$
- (iii) $Mb(R^{(i)})$ is the markov blanket of $R^{(i)}$ comprising of all substantive variables that are either neighbors or connected to variables in $R^{(i)}$ by a collider path (Richardson [2003]). In this figure: $Mb(R^{(1)}) = \{X, Y\}$ and $Mb(R^{(2)}) = \{Z, W\}$.



Theorem 2

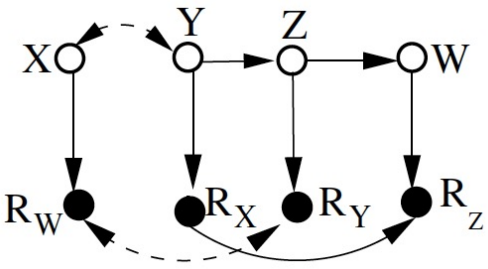
- This theorem relies on the notion of collider path and two new subsets:
  - $R^{(part)}$: the partitions of R variables and
  - $Mb(R^{(i)})$: substantive variables related to $R^{(i)}$.

- Theorem2. Given an m-graph G in which no element in V_m is either **a neighbor of its missingness mechanism** or **connected to its missingness mechanism by a collider path**, P(V) is recoverable if no $Mb(R^{(i)})$ contains a partially observed variable X such that $R_x \in R^{(i)}$ i.e. $\forall i$, $R^{(i)} \cap R_{Mb(R^{(i)})} = \emptyset$.
- Moreover, if recoverable, P(V) is given by,

$$P(v) = \frac{P(R=0, V)}{\prod_i P(R^{(i)}=0 | Mb(R^{(i)}), R_{Mb(R^{(i)})}=0)}$$

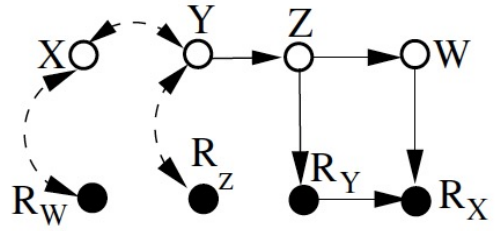
A subset of the class of problems in theorem 2

- The following corollary yields a sufficient condition for recovering the joint distribution from the class of problems in which no bi-directed edge exists between variables in sets R and $V_o \cup V_m$.



- These problems form a subset of the class of problems covered in theorem 2.

- Subset $Pa^{sub}(R^{(i)})$ used in the corollary is the set of all substantive variables that are parents of variables in $R^{(i)}$.
  - In this figure: $R^{(1)} = \{R_w, R_z\}$, $R^{(2)} = \{R_x, R_y\}$
  - $Pa^{sub}(R^{(1)}) = \emptyset$ and $Pa^{sub}(R^{(2)}) = \{Z, W\}$
  - $Mb(R^{(1)}) = \{X, Y\}$ and $Mb(R^{(2)}) = \{Z, W\}$



Corollary 1.

- Let G be an m-graph such that
  - (i) $\forall X \in V_o \cup V_m$, no latent variable is a common parent of X and any member of R, and
  - (ii) $\forall Y \in V_m$, Y is not a parent of R_y.
- If $\forall i$, $Pa^{sub}(R^{(i)})$ does not contain a partially observed variables whose missing mechanism is in $R^{(i)}$ i.e. $R^{(i)} \cap R_{pa^{sub}(R^{(i)})} = \emptyset$, then P(V) is recoverable and is given by,

$$P(v) = \frac{P(R=0, V)}{\prod_i P(R^{(i)}=0 | Pa^{sub}(R^{(i)}), R_{pa^{sub}(R^{(i)})}=0)}$$

Non-recoverability Criteria for Joint and Conditional Distributions

- Up until now, we dealt with sufficient conditions for recoverability.
- It is important however to supplement these results with criteria for non-recoverability in order to alert the user to the fact that the available assumptions are insufficient to produce a consistent estimate of the target query.
- Such criteria have not been treated formally in the literature thus far.
- We introduce two graphical conditions that preclude recoverability.

Non-recoverability Criteria for Joint and Conditional Distributions (cont'd)

- Theorem 3 (Non-recoverability of P(V)). Given a semi-markovian model G, the following conditions are necessary for recoverability of the joint distribution:
  i. $\forall X \in V_m$, X and R_x are not neighbors and
  ii. $\forall X \in V_m$, there does not exist a path from X to R_x in which every intermediate node is both a collider and a substantive variable.

- Corollary 2. [Non-recoverability of P(Y|X)] Let X and Y be disjoint subsets of substantive variables. P(Y|X) is non-recoverable in m-graph G if one of the following conditions is true:
  1) Y and R_y are neighbors
  2) G contains a collider path p connecting Y and R_y such that all intermediate nodes in p are in X.

Recovering Causal Queries

- Given a causal query and a causal bayesian network a complete algorithm exists for deciding whether the query is identifiable or not.

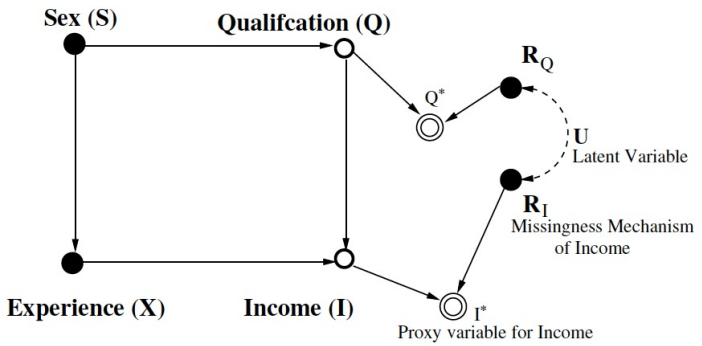
- Obviously, a query that is not identifiable in the substantive model is not recoverable from missing data.

- Therefore, a necessary condition for recoverability of a causal query is its identifiability which we will assume in the rest of our discussion.

Definition 3 (Trivially Recoverable Query).

- A causal query Q is said to be trivially recoverable given an m-graph G if it has an estimand (in terms of substantive variables) in which every factor is recoverable.

- Example 2. In the gender wage-gap study example, the effect of sex on income, P(I|do(S)), is identifiable and is given by P(I|S).
  - By theorem 2, P(S, X, Q, I) is recoverable.
  - Hence P(I|do(S)) is recoverable.



Recovering P(y|do(z)) when Y and R_y are inseparable

- The recoverability of P(V) hinges on the separability of a partially observed variable from its missingness mechanism (a condition established in theorem 3).

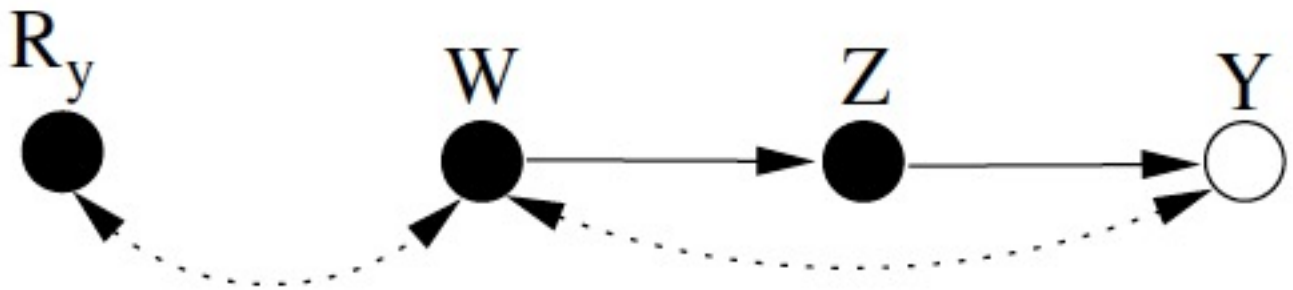
- Remarkably, causal queries may circumvent this requirement.

- The following example demonstrates that P(y|do(z)) is recoverable even when Y and R_y are not separable.

Example 3.

- By backdoor criterion, $P(y|do(z)) = \sum P(y|z, w)P(w)$.

- One might be tempted to conclude that the causal relation is non-recoverable because P(w, z, y) is non-recoverable (by theorem 2) and P(y|z, w) is not recoverable (by corollary 2).



Example 3. (cont'd)

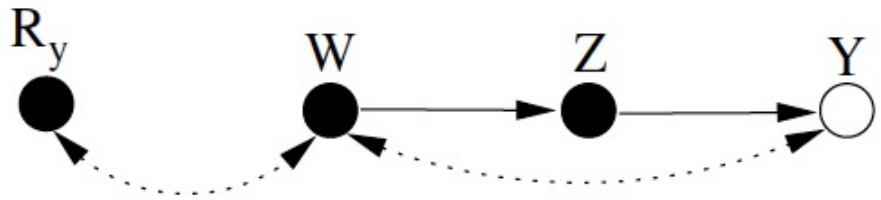
- However, P(y|do(z)) is recoverable as demonstrated below:

P(y|do(z)) = P(y|do(z), R'_y) = $\sum_w$ P(y|$do(z)$, w, R'_y)P(w| do(z), R'_y)

P(y|$do(z)$, w, R'_y) = P(y|z, w, R'_y) Rule-2 of do-calculus

P(w| do(z), R'_y)= P(w| R'_y) Rule-3 of do-calculus

P(y|do(z)) = $\sum_w$ P(y|z, w, R'_y)P(w| z, R'_y) = $\sum_w$ P(y^*|z, w, R'_y)P(w| z, R'_y)



Attrition

- Attrition: participants dropping out from a study/experiment.
- Here, we discuss a special case of attrition called ‘Simple Attrition’.
- In this problem, a researcher conducts a randomized trial, measures a set of variables (X,Y,Z) and obtains a dataset where outcome (Y) is corrupted by missing values (due to attrition). Clearly, due to randomization, the effect of treatment (X) on outcome (Y), P(y|do(x)), is identifiable and is given by P(Y|X).

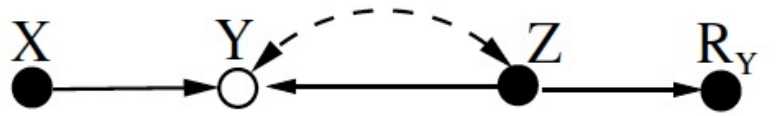
Typical attrition problems

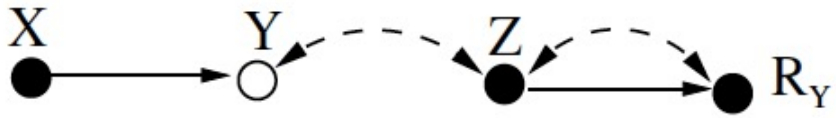
- We shall now demonstrate the usefulness of our previous discussion in recovering P(y|do(x)).

- In this figure, we can apply theorem 1 to recover P(y|do(x)) as given below:

$$P(Y|X) = \sum_z P(Y^*|X, Z, R'_y)P(Z|X)$$

- In this figure, we observe that Y and R_y are connected by a collider path. Therefore by corollary 2, P(Y|X) is not recoverable; hence P(y|do(x)) is also not recoverable.





Recovering Joint Distributions under simple attrition

- The following theorem yields the necessary and sufficient condition for recovering joint distributions from semi-markovian models with a single partially observed variable i.e. $|V_m| = 1$ which includes models afflicted by simple attrition.

- Theorem 5. Let $Y \in V_m$ and $|V_m| = 1$. $P(V)$ is recoverable in m-graph G if and only if Y and R_y are not neighbors and Y and R_y are not connected by a path in which all intermediate nodes are colliders.
- If both conditions are satisfied, then $P(V)$ is given by,

$$P(V) = P(Y|V_O, R_y = 0)P(V_O)$$

Recovering Causal Effects under Simple Attrition

- Theorem 6. P(y|do(x)) is recoverable in the simple attrition case (with one partially observed variable) if and only if Y and R_y are neither neighbors nor connected by an inducing path.

- Moreover, if recoverable,

$$P(Y|X) = \sum_z P(Y^*|X, Z, R'_y)P(Z|X)$$

- where Z is the separating set that d-separates Y from R_y.

Conclusion

- Graphical models play a critical role in portraying the missingness process, encoding and communicating assumptions about missingness and deciding recoverability given a dataset afflicted with missingness.
- We presented graphical conditions for recovering joint and conditional distributions and sufficient conditions for recovering causal queries.
- We exemplified the recoverability of causal queries of the form $P(y|do(x))$ despite the existence of an inseparable path between Y and R_y, which is an insurmountable obstacle to the recovery of P(Y).
- We applied our results to problems of attrition and presented necessary and sufficient graphical conditions for recovering causal effects in such problems.

Thanks!