Pattern graph: a graphical approach to nonmonotone missing data

Yen-Chi Chen

Department of Statistics
University of Washington
◦ Supported by NSF DMS - 1810960 and DMS - 195278, NIH U01 - AG0169761

A typical missing data

ID	L_1	L_2	L_3
1	15	20	NA
2	12	NA	NA
3	NA	43	35
4	11	25	NA
5	NA	37	NA
6	15	23	32
7	NA	27	35

Probability model under missingness

- The variable of interest (also called a study variable) is a random variable $L = (L_1, \cdots, L_d) \in \mathbb{R}^d$.
- We represent the missingness of L using a binary response vector $R \in \{0, 1\}^d$.
- $R_j = 1$ if L_j is observed.

Probability model under missingness

- The variable of interest (also called a study variable) is a random variable $L = (L_1, \cdots, L_d) \in \mathbb{R}^d$.
- We represent the missingness of L using a binary response vector $R \in \{0, 1\}^d$.
- $R_j = 1$ if L_j is observed.
- We denote $L_R = (L_j : R_j = 1)$ be the observed variable when the response pattern is R.

Probability model under missingness

- The variable of interest (also called a study variable) is a random variable $L = (L_1, \cdots, L_d) \in \mathbb{R}^d$.
- We represent the missingness of L using a binary response vector $R \in \{0,1\}^d$.
- $R_j = 1$ if L_j is observed.
- We denote $L_R = (L_j : R_j = 1)$ be the observed variable when the response pattern is R.
- The joint distribution of (L, R), $F(\ell, r)$, is called the full-data distribution. The corresponding density $p(\ell, r)$ is called full-data density.

Probability model under missingness

- The variable of interest (also called a study variable) is a random variable $L = (L_1, \cdots, L_d) \in \mathbb{R}^d$.
- We represent the missingness of L using a binary response vector $R \in \{0, 1\}^d$.
- $R_j = 1$ if L_j is observed.
- We denote $L_R = (L_j : R_j = 1)$ be the observed variable when the response pattern is R.
- The joint distribution of (L, R), $F(\ell, r)$, is called the full-data distribution. The corresponding density $p(\ell, r)$ is called full-data density.
- We are often interested in some characteristic of $F(\ell)$, the distribution of L.

Response indicator

ID	L_1	L_2	L_3	R
1	15	20	NA	110
2	12	NA	NA	100
3	NA	43	35	011
4	11	25	NA	110
5	NA	37	NA	101
6	15	23	32	111
7	NA	27	35	011

Challenge of missing data

- A challenge in missing data is that $F(\ell)$ or $F(\ell, r)$ or $p(\ell, r)$ are often unidentifiable.
- Namely, even with infinite amount of data, we still cannot accurately estimate them.

Challenge of missing data

- A challenge in missing data is that $F(\ell)$ or $F(\ell, r)$ or $p(\ell, r)$ are often unidentifiable.
- Namely, even with infinite amount of data, we still cannot accurately estimate them.
- To see this, let $1_d = (1, 1, \cdots, 1)$ be the special response pattern that corresponds to fully observed case.

Challenge of missing data

- A challenge in missing data is that $F(\ell)$ or $F(\ell, r)$ or $p(\ell, r)$ are often unidentifiable.
- Namely, even with infinite amount of data, we still cannot accurately estimate them.
- To see this, let $1_d = (1, 1, \cdots, 1)$ be the special response pattern that corresponds to fully observed case.
- We denote $\bar{R} = 1_d - R$ as flipping 0 and 1 in R. Then, $L_{\bar{R}}$ is the unobserved variables under pattern R.

Challenge of missing data

- A challenge in missing data is that $F(\ell)$ or $F(\ell, r)$ or $p(\ell, r)$ are often unidentifiable.
- Namely, even with infinite amount of data, we still cannot accurately estimate them.
- To see this, let $1_d = (1, 1, \cdots, 1)$ be the special response pattern that corresponds to fully observed case.
- We denote $\bar{R} = 1_d - R$ as flipping 0 and 1 in R. Then, $L_{\bar{R}}$ is the unobserved variables under pattern R.
- The challenge of missing data comes from the fact that the PDF

$$p(\ell|R = r) = p(\ell_{\bar{r}}, \ell_r|R = r) = p(\ell_{\bar{r}}|\ell_r, R = r)p(\ell_r|R = r)$$

involves unobserved part $p(\ell_{\bar{r}}|\ell_r, R = r)$.

Response indicator, again

ID	L_1	L_2	L_3	R
1	15	20	NA	110
2	12	NA	NA	100
3	NA	43	35	011
4	11	25	NA	110
5	NA	37	NA	101
6	15	23	32	111
7	NA	27	35	011

Strategy in missing data: selection models

- **Selection models (SMs)**: attempt to identify the selection probability (missing data mechanism)

$$P(R = r|L)$$

from making assumptions (known as identifying restrictions).

Strategy in missing data: selection models

- **Selection models (SMs)**: attempt to identify the selection probability (missing data mechanism)

$$P(R = r|L)$$

from making assumptions (known as identifying restrictions).

- Motivation: consider the problem of estimating a mean

$$\theta_0 = \mathbb{E}\left(\theta(L)\right) = \mathbb{E}\left(\frac{\theta(L)I(R = 1_d)}{\pi(L)}\right), \pi(L) = P(R = 1_d|L).$$

Strategy in missing data: selection models

- **Selection models (SMs)**: attempt to identify the selection probability (missing data mechanism)

$$P(R = r|L)$$

  from making assumptions (known as identifying restrictions).
- Motivation: consider the problem of estimating a mean

$$\theta_0 = \mathbb{E}\left(\theta(L)\right) = \mathbb{E}\left(\frac{\theta(L)I(R = 1_d)}{\pi(L)}\right), \pi(L) = P(R = 1_d|L).$$

- If we can identify $\pi(L) = P(R = 1_d|L)$, we can construct an inverse probability weighting (IPW) estimator.

Strategy in missing data: missing at random

- Missing completely at random (MCAR):

$$P(R = r|L) = P(R = r).$$

- Missing at random (MAR; Little and Rubin 2002):

$$P(R = r|L) = P(R = r|L_r).$$

- Missing not at random (MNAR) is the case where the probability $P(R = r|L)$ may depends on the unobserved $L_{\bar{r}}$.
- In this talk, the identifying restrictions we construct are mostly MNAR.

Strategy in missing data: pattern mixture models

- **Pattern mixture models (PMMs)**: decompose the full-data density via

$$p(\ell, r) = p(\ell_{\bar{r}} | \ell_r, R = r) p(\ell_r | R = r) P(R = r).$$

- $p(\ell_{\bar{r}} | \ell_r, R = r)$: the extrapolation density (unidentifiable).
- $p(\ell_r | R = r) P(R = r)$: the observed-data density (identifiable).

Strategy in missing data: pattern mixture models

- **Pattern mixture models (PMMs)**: decompose the full-data density via

$$p(\ell, r) = p(\ell_{\bar{r}} | \ell_r, R = r) p(\ell_r | R = r) P(R = r).$$

- $p(\ell_{\bar{r}} | \ell_r, R = r)$: the extrapolation density (unidentifiable).
- $p(\ell_r | R = r) P(R = r)$: the observed-data density (identifiable).
- We attempt to identify $p(\ell_{\bar{r}} | \ell_r, R = r)$ by making assumptions.

Strategy in missing data: pattern mixture models

- **Pattern mixture models (PMMs)**: decompose the full-data density via

$$p(\ell, r) = p(\ell_{\bar{r}} | \ell_r, R = r) p(\ell_r | R = r) P(R = r).$$

- $p(\ell_{\bar{r}} | \ell_r, R = r)$: the extrapolation density (unidentifiable).
- $p(\ell_r | R = r) P(R = r)$: the observed-data density (identifiable).
- We attempt to identify $p(\ell_{\bar{r}} | \ell_r, R = r)$ by making assumptions.
- Complete-case missing value (CCMV; Little 1993 and Tchetgen et al. 2016) restriction:

$$p(\ell_{\bar{r}} | \ell_r, R = r) = p(\ell_{\bar{r}} | \ell_r, R = 1_d).$$

Pattern graphs and identification

Regular pattern graph

- Let $\mathscr{R} \subset \{0,1\}^d$ be the response set that $P(R \in \mathscr{R}) = 1$.
- A pattern graph is a directed graph G with vertex set being $\mathscr{R}$.
- Namely, a pattern graph is a graph representing the relationship between different patterns.

Regular pattern graph

- Let $\mathscr{R} \subset \{0, 1\}^d$ be the response set that $P(R \in \mathscr{R}) = 1$.
- A pattern graph is a directed graph G with vertex set being $\mathscr{R}$.
- Namely, a pattern graph is a graph representing the relationship between different patterns.
- If there is an arrow from $s \to r$, then s is a parent of r and r is a child of s. We denote PA_r as the parents of r.

Regular pattern graph

- Let $\mathscr{R} \subset \{0,1\}^d$ be the response set that $P(R \in \mathscr{R}) = 1$.
- A pattern graph is a directed graph G with vertex set being $\mathscr{R}$.
- Namely, a pattern graph is a graph representing the relationship between different patterns.
- If there is an arrow from $s \to r$, then s is a parent of r and r is a child of s. We denote PA_r as the parents of r.
- A node/vertex of a graph is called a *source* if it has no parents.

Regular pattern graph

- Let $\mathcal{R} \subset \{0,1\}^d$ be the response set that $P(R \in \mathcal{R}) = 1$.
- A pattern graph is a directed graph G with vertex set being $\mathcal{R}$.
- Namely, a pattern graph is a graph representing the relationship between different patterns.
- If there is an arrow from $s \to r$, then s is a parent of r and r is a child of s. We denote PA_r as the parents of r.
- A node/vertex of a graph is called a *source* if it has no parents.
- For two patterns $s, r \in \{0,1\}^d$, we write $r > s$ if $r_j \geq s_j$ for all j and there is at least one coordinate j^* such that $r_{j^*} > s_{j^*}$.

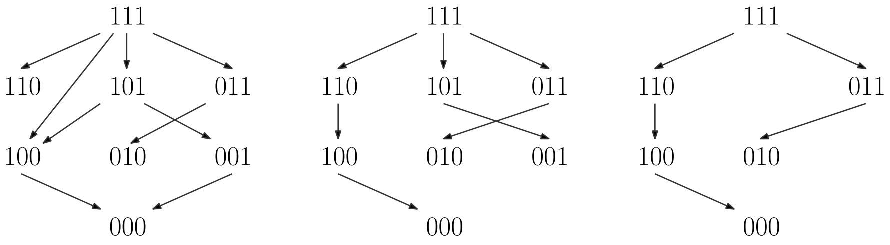
Regular pattern graph

- Let $\mathscr{R} \subset \{0,1\}^d$ be the response set that $P(R \in \mathscr{R}) = 1$.
- A pattern graph is a directed graph G with vertex set being $\mathscr{R}$.
- Namely, a pattern graph is a graph representing the relationship between different patterns.
- If there is an arrow from $s \to r$, then s is a parent of r and r is a child of s. We denote PA_r as the parents of r.
- A node/vertex of a graph is called a *source* if it has no parents.
- For two patterns $s, r \in \{0,1\}^d$, we write $r > s$ if $r_j \geq s_j$ for all j and there is at least one coordinate j^* such that $r_{j^*} > s_{j^*}$.
- A pattern graph is called a **regular pattern graph** if
  - (G1) pattern $1_d = (1, 1, \cdots, 1)$ is the only source.
  - (G2) if $s \to r$, then $s > r$.

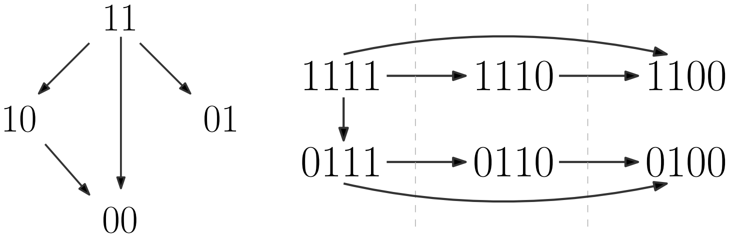
Regular pattern graph

- Let $\mathcal{R} \subset \{0,1\}^d$ be the response set that $P(R \in \mathcal{R}) = 1$.
- A pattern graph is a directed graph G with vertex set being $\mathcal{R}$.
- Namely, a pattern graph is a graph representing the relationship between different patterns.
- If there is an arrow from $s \to r$, then s is a parent of r and r is a child of s. We denote PA_r as the parents of r.
- A node/vertex of a graph is called a *source* if it has no parents.
- For two patterns $s, r \in \{0,1\}^d$, we write $r > s$ if $r_j \geq s_j$ for all j and there is at least one coordinate j^* such that $r_{j^*} > s_{j^*}$.
- A pattern graph is called a **regular pattern graph** if
  (G1) pattern $1_d = (1, 1, \cdots, 1)$ is the only source.
  (G2) if $s \to r$, then $s > r$.
- (G2) implies that the resulting graph is a directed acyclic graph (DAG).

Examples of regular pattern graphs



Examples of regular pattern graphs



Pattern graph and selection odds model

- We say that **the selection odds model of** (L, R) **factorizes with respect to** G if

$$\frac{P(R = r|L)}{P(R \in \mathsf{PA}_r|L)} = \frac{P(R = r|L_r)}{P(R \in \mathsf{PA}_r|L_r)} \equiv O_r(L_r).$$

Pattern graph and selection odds model

- We say that **the selection odds model of** (L, R) **factorizes with respect to** G if

$$\frac{P(R = r|L)}{P(R \in \mathsf{PA}_r|L)} = \frac{P(R = r|L_r)}{P(R \in \mathsf{PA}_r|L_r)} \equiv O_r(L_r).$$

- Namely, the graph factorization requires that the selection odds (against parents) is observable.

Pattern graph and selection odds model

- We say that **the selection odds model of** (L, R) **factorizes with respect to** G if

$$\frac{P(R = r|L)}{P(R \in \mathsf{PA}_r|L)} = \frac{P(R = r|L_r)}{P(R \in \mathsf{PA}_r|L_r)} \equiv O_r(L_r).$$

- Namely, the graph factorization requires that the selection odds (against parents) is observable.
- The selection odds model implies

$$P(R = r|L) = \sum_{s \in \mathsf{PA}_r} P(R = s|L) O_r(L_r).$$

The chance of observing a particular pattern equals the summation of all its parents' probability multiplied by an observed factor $O_r(L_r)$.

Pattern graph and selection odds model

- We say that **the selection odds model of** (L, R) **factorizes with respect to** G if

$$\frac{P(R = r|L)}{P(R \in \mathsf{PA}_r|L)} = \frac{P(R = r|L_r)}{P(R \in \mathsf{PA}_r|L_r)} \equiv O_r(L_r).$$

- Namely, the graph factorization requires that the selection odds (against parents) is observable.
- The selection odds model implies

$$P(R = r|L) = \sum_{s \in \mathsf{PA}_r} P(R = s|L) O_r(L_r).$$

  The chance of observing a particular pattern equals the summation of all its parents' probability multiplied by an observed factor $O_r(L_r)$.
- Note: pattern graphs are not the conventional graphical model!

Selection odds model and identifications

Theorem

Suppose that the selection odds model of (L, R) factorizes with respect to a regular pattern graph G. Let $Q_r(L) = \frac{P(R=r|L)}{P(R=1_d|L)}$ and $Q_{1_d}(L) = 1$. Then $\pi(L) \equiv P(R = 1_d|L)$ is identifiable and is defined via

$$\pi(L) = \frac{1}{\sum_r Q_r(L)}, \quad Q_r(L) = O_r(L_r) \sum_{s \in \mathsf{PA}_r} Q_s(L).$$

- This provides a recursive approach to identify $\pi(L)$.

Path identification interpretation -1

- A (directed) path Ξ in a graph G is a set of vertices $r_1, r_2, r_3, \cdots, r_k$ such that the edge $r_t \to r_{t+1}$ exists in G.

Path identification interpretation -1

- A (directed) path Ξ in a graph G is a set of vertices $r_1, r_2, r_3, \cdots, r_k$ such that the edge $r_t \to r_{t+1}$ exists in G.
- We define Π_r to be the collection of all paths from 1_d to r and let $\Pi_{1_d} = \{1_d, 1_d\}$. We also define $\Pi = \cup_r \Pi_r$ to be the collection of all paths.

Path identification interpretation -1

- A (directed) path Ξ in a graph G is a set of vertices $r_1, r_2, r_3, \cdots, r_k$ such that the edge $r_t \to r_{t+1}$ exists in G.
- We define Π_r to be the collection of all paths from 1_d to r and let $\Pi_{1_d} = \{1_d, 1_d\}$. We also define $\Pi = \cup_r \Pi_r$ to be the collection of all paths.
- The selection odds model implies an interesting representation of the selection probability:

Path identification interpretation -1

- A (directed) path Ξ in a graph G is a set of vertices $r_1, r_2, r_3, \cdots, r_k$ such that the edge $r_t \to r_{t+1}$ exists in G.
- We define Π_r to be the collection of all paths from 1_d to r and let $\Pi_{1_d} = \{1_d, 1_d\}$. We also define $\Pi = \cup_r \Pi_r$ to be the collection of all paths.
- The selection odds model implies an interesting representation of the selection probability:

Theorem

Suppose that the selection odds model of (L, R) factorizes with respect to a regular pattern graph G. Then

$$1 = \sum_{\Xi \in \Pi} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$

$$P(R = r|L) = \sum_{\Xi \in \Pi_r} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$

Path identification interpretation -2

- The two equations

$$1 = \sum_{\Xi \in \Pi} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$
$$P(R = r|L) = \sum_{\Xi \in \Pi_r} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$

show an interesting interpretation.

Path identification interpretation -2

- The two equations

$$1 = \sum_{\Xi \in \Pi} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$
$$P(R = r|L) = \sum_{\Xi \in \Pi_r} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$

  show an interesting interpretation.

- First, for $\Xi \in \Pi$, we can interpret

$$\pi(L) \prod_{s \in \Xi} O_s(L_s) = \kappa(\Xi|L)$$

  as the probability of selecting Ξ from Π.

Path identification interpretation -2

- The two equations

$$1 = \sum_{\Xi \in \Pi} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$

$$P(R = r|L) = \sum_{\Xi \in \Pi_r} \pi(L) \prod_{s \in \Xi} O_s(L_s)$$

  show an interesting interpretation.

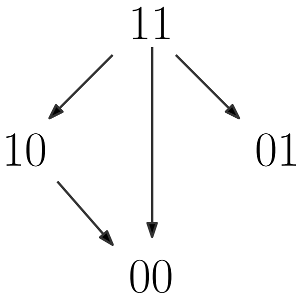
- First, for $\Xi \in \Pi$, we can interpret

$$\pi(L) \prod_{s \in \Xi} O_s(L_s) = \kappa(\Xi|L)$$

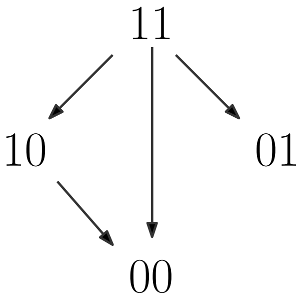
  as the probability of selecting Ξ from Π.

- Then the second equation implies $P(R = r|L) = \sum_{\Xi \in \Pi_r} \kappa(\Xi|L)$, i.e., $P(R = r|L)$ is the summation of contributions from all paths ending at r.

Path identification: an example



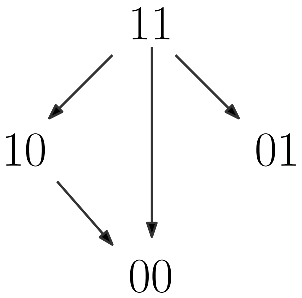
Path identification: an example



- There are 5 paths and each corresponds to probability:

$$\kappa(11 \to 11|L) = \pi(L)$$
$$\kappa(11 \to 10|L) = \pi(L)O_{10}(L_{10})$$
$$\kappa(11 \to 01|L) = \pi(L)O_{01}(L_{01})$$
$$\kappa(11 \to 00|L) = \pi(L)O_{00}(L_{00})$$
$$\kappa(11 \to 10 \to 00|L) = \pi(L)O_{10}(L_{10})O_{00}(L_{00})$$

Path identification: an example



- There are 5 paths and each corresponds to probability:

$$\kappa(11 \to 11|L) = \pi(L)$$
$$\kappa(11 \to 10|L) = \pi(L)O_{10}(L_{10})$$
$$\kappa(11 \to 01|L) = \pi(L)O_{01}(L_{01})$$
$$\kappa(11 \to 00|L) = \pi(L)O_{00}(L_{00})$$
$$\kappa(11 \to 10 \to 00|L) = \pi(L)O_{10}(L_{10})O_{00}(L_{00})$$

- So the chance of observing each pattern is

$$P(R = 11|L) = \pi(L), \quad P(R = 10|L) = \pi(L)O_{10}(L_{10}),$$
$$P(R = 01|L) = \pi(L)O_{01}(L_{01})$$
$$P(R = 00|L) = \pi(L)O_{00}(L_{00}) + \pi(L)O_{10}(L_{10})O_{00}(L_{00}).$$

Pattern graph and pattern mixture model - 1

- Recall that PMMs decompose the joint density via

$$p(\ell, r) = p(\ell_{\bar{r}} | \ell_r, R = r) p(\ell_r | R = r) P(R = r).$$

- $p(\ell_{\bar{r}} | \ell_r, R = r)$: the extrapolation density (unidentifiable).
- $p(\ell_r | R = r) P(R = r)$: the observed-data density (identifiable).
- Strategy of PMMs: try to identify the extrapolation density.

Pattern graph and pattern mixture model - 2

- We say that **the pattern mixture model of** (L, R) **factorizes with respect to** G if

$$p(\ell_{\bar{r}} | \ell_r, R = r) = p(\ell_{\bar{r}} | \ell_r, R \in \mathsf{PA}_r).$$

- Namely, the extrapolation density is the same as the same variables' conditional density in the parent patterns.

Pattern graph and pattern mixture model - 2

- We say that **the pattern mixture model of** (L, R) **factorizes with respect to** G if

$$p(\ell_{\bar{r}}|\ell_r, R = r) = p(\ell_{\bar{r}}|\ell_r, R \in \mathsf{PA}_r).$$

- Namely, the extrapolation density is the same as the same variables' conditional density in the parent patterns.

Theorem

Suppose that the pattern mixture model of (L, R) *factorizes with respect to a regular pattern graph* G. *Then the full-data density* $p(\ell, r)$ *is (nonparametrically) identifiable.*

Pattern graph and pattern mixture model - 2

- We say that **the pattern mixture model of** (L, R) **factorizes with respect to** G if

$$p(\ell_{\bar{r}}|\ell_r, R = r) = p(\ell_{\bar{r}}|\ell_r, R \in \mathsf{PA}_r).$$

- Namely, the extrapolation density is the same as the same variables' conditional density in the parent patterns.

Theorem

Suppose that the pattern mixture model of (L, R) *factorizes with respect to a regular pattern graph* G*. Then the full-data density* $p(\ell, r)$ *is (nonparametrically) identifiable.*

- Namely, we can estimate the joint distribution of (L, R) and the resulting distribution will agree with the observed data (nonparametrically identifiable).

Equivalence of the graph factorizations

Theorem

If G is a regular pattern graph and $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$, then the following two statements are equivalent:

- *the selection odds model of (L, R) factorizes with respect to G.*
- *the pattern mixture model of (L, R) factorizes with respect to G.*

Equivalence of the graph factorizations

Theorem

If G is a regular pattern graph and $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$, then the following two statements are equivalent:

- *the selection odds model of (L, R) factorizes with respect to G.*
- *the pattern mixture model of (L, R) factorizes with respect to G.*

- The condition, $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$, can be viewed as a positivity condition.

Equivalence of the graph factorizations

Theorem

If G is a regular pattern graph and $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$, then the following two statements are equivalent:

- *the selection odds model of (L, R) factorizes with respect to G.*
- *the pattern mixture model of (L, R) factorizes with respect to G.*

- The condition, $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$, can be viewed as a positivity condition.
- Therefore, we can interpret the result using either a selection odds model perspective or a pattern mixture model perspective.
- Note that Robins et al. (2000) had shown that certain selection odds models and pattern mixture models are equivalent.

Estimation with pattern graphs

IPW estimator and graph factorization - 1

- With a slight abuse of notation, the observations are denoted as

$$(L_{1,R_1}, R_1), \cdots, (L_{n,R_n}, R_n).$$

- Recall that the IPW estimator is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\theta(L_i)I(R_i = 1_d)}{\pi(L_i)}.$$

- Often $\pi(\ell)$ is unknown and has to be estimated.

IPW estimator and graph factorization - 1

- With a slight abuse of notation, the observations are denoted as

$$(L_{1,R_1}, R_1), \cdots, (L_{n,R_n}, R_n).$$

- Recall that the IPW estimator is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\theta(L_i) I(R_i = 1_d)}{\pi(L_i)}.$$

- Often $\pi(\ell)$ is unknown and has to be estimated.
- With the selection odds model, we can estimate it by estimating each odds $O_r(L_r) = \frac{P(R=r|L_r)}{P(R\in \mathsf{PA}_r|L_r)}$.

IPW estimator and graph factorization - 1

- With a slight abuse of notation, the observations are denoted as

$$(L_{1,R_1}, R_1), \cdots, (L_{n,R_n}, R_n).$$

- Recall that the IPW estimator is

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\theta(L_i) I(R_i = 1_d)}{\pi(L_i)}.$$

- Often $\pi(\ell)$ is unknown and has to be estimated.
- With the selection odds model, we can estimate it by estimating each odds $O_r(L_r) = \frac{P(R=r|L_r)}{P(R\in \mathsf{PA}_r|L_r)}$.
- This can be done by applying a generative classifier or a regression model comparing two classes

$$R = r \text{ v.s. } R \in \mathsf{PA}_r$$

using only the variables L_r.

IPW estimator and graph factorization - 2

- Let $\widehat{O}(L_r) = O(L_r; \widehat{\eta}_r)$ be the estimated odds and $\widehat{\eta}_r \in \Theta_r$ is the corresponding parameter.
- Note: logistic regression leads to $O(L_r; \widehat{\eta}_r) = \exp(L_r^T \widehat{\eta}_r)$.

IPW estimator and graph factorization - 2

- Let $\widehat{O}(L_r) = O(L_r; \widehat{\eta}_r)$ be the estimated odds and $\widehat{\eta}_r \in \Theta_r$ is the corresponding parameter.
- Note: logistic regression leads to $O(L_r; \widehat{\eta}_r) = \exp(L_r^T \widehat{\eta}_r)$.
- This implies an estimator $\widehat{\pi}(L) = \pi(L; \widehat{\eta})$, where $\widehat{\eta} = (\widehat{\eta}_r : r \in \mathscr{R})$.

IPW estimator and graph factorization - 2

- Let $\widehat{O}(L_r) = O(L_r; \widehat{\eta}_r)$ be the estimated odds and $\widehat{\eta}_r \in \Theta_r$ is the corresponding parameter.
- Note: logistic regression leads to $O(L_r; \widehat{\eta}_r) = \exp(L_r^T \widehat{\eta}_r)$.
- This implies an estimator $\widehat{\pi}(L) = \pi(L; \widehat{\eta})$, where $\widehat{\eta} = (\widehat{\eta}_r : r \in \mathcal{R})$.
- The IPW estimator is

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\theta(L_i) I(R_i = 1_d)}{\pi(L_i; \widehat{\eta})}.$$

IPW estimator and graph factorization - 2

- Let $\widehat{O}(L_r) = O(L_r; \widehat{\eta}_r)$ be the estimated odds and $\widehat{\eta}_r \in \Theta_r$ is the corresponding parameter.
- Note: logistic regression leads to $O(L_r; \widehat{\eta}_r) = \exp(L_r^T \widehat{\eta}_r)$.
- This implies an estimator $\widehat{\pi}(L) = \pi(L; \widehat{\eta})$, where $\widehat{\eta} = (\widehat{\eta}_r : r \in \mathscr{R})$.
- The IPW estimator is

$$\widehat{\theta}_{\mathsf{IPW}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\theta(L_i) I(R_i = 1_d)}{\pi(L_i; \widehat{\eta})}.$$

Theorem

Suppose that parametric models are correctly specified. Then under regularity conditions,

$$\sqrt{n}(\widehat{\theta}_{\mathsf{IPW}} - \theta_0) \xrightarrow{D} N(0, \sigma^2_{\mathsf{IPW}}).$$

Recursive computation of $\widehat{\pi}(L)$

- Here is a simple approach to compute $\widehat{\pi}(L)$ from estimators $\widehat{O}_r(L_r)$ (not limited to parametric models).
- Recall that

$$\widehat{\pi}(L) = \frac{1}{\sum_r \widehat{Q}_r(L)}, \quad \widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$$

and $\widehat{Q}_{1_d}(L) = 1$.

Recursive computation of $\widehat{\pi}(L)$

- Here is a simple approach to compute $\widehat{\pi}(L)$ from estimators $\widehat{O}_r(L_r)$ (not limited to parametric models).
- Recall that

$$\widehat{\pi}(L) = \frac{1}{\sum_r \widehat{Q}_r(L)}, \quad \widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$$

  and $\widehat{Q}_{1_d}(L) = 1$.
- We first evaluate $\widehat{O}_r(L_r)$ for each r.

Recursive computation of $\widehat{\pi}(L)$

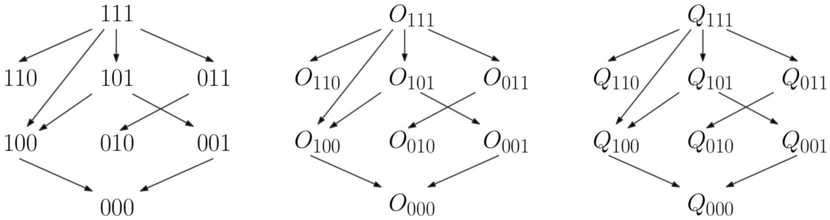
- Here is a simple approach to compute $\widehat{\pi}(L)$ from estimators $\widehat{O}_r(L_r)$ (not limited to parametric models).
- Recall that

$$\widehat{\pi}(L) = \frac{1}{\sum_r \widehat{Q}_r(L)}, \quad \widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$$

  and $\widehat{Q}_{1_d}(L) = 1$.
- We first evaluate $\widehat{O}_r(L_r)$ for each r.
- Then we sequentially compute $\widehat{Q}_r(L)$ for $|r| = d-1, d-2, \cdots, 1$ using the recursive relation where $|r| = \sum_j r_j$ is the number of observed patterns.

Graphical representation of the recursive computation

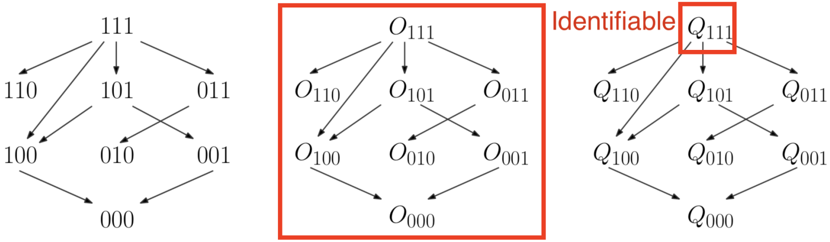
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- Consider the above graph and the corresponding O_r, Q_r.

Graphical representation of the recursive computation

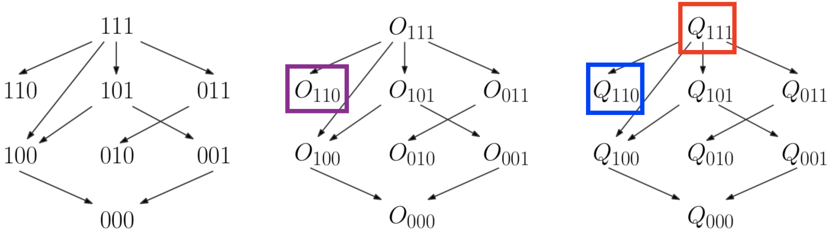
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- All these quantities are identifiable/computable ($Q_{111}(L) = 1$).

Graphical representation of the recursive computation

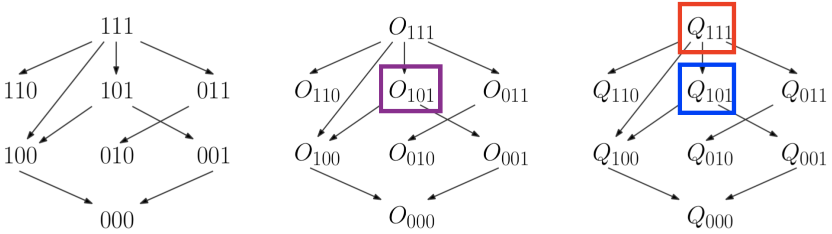
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- We compute Q_r using the parent(s) and the selection odds.

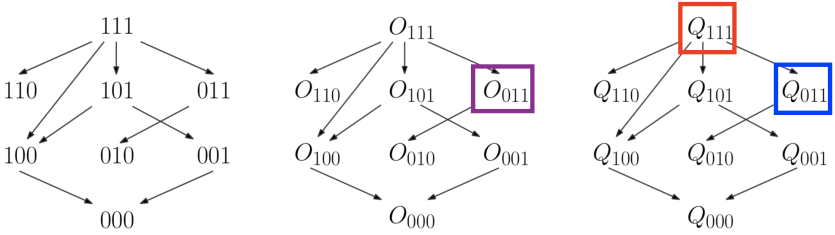
Graphical representation of the recursive computation

$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- We compute Q_r using the parent(s) and the selection odds.

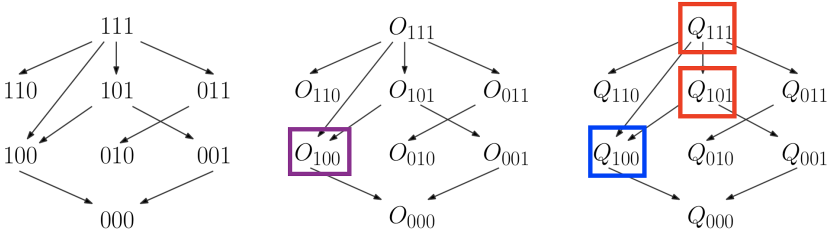
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- We compute Q_r using the parent(s) and the selection odds.

Graphical representation of the recursive computation

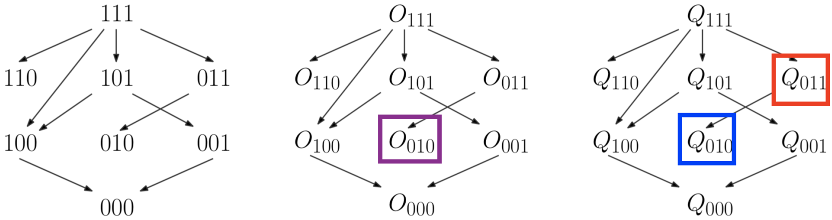
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- We compute Q_r using the parent(s) and the selection odds.

Graphical representation of the recursive computation

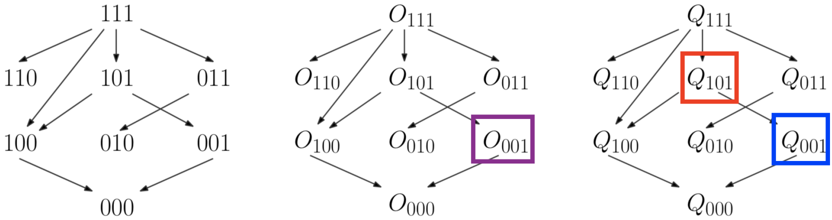
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- We compute Q_r using the parent(s) and the selection odds.

Graphical representation of the recursive computation

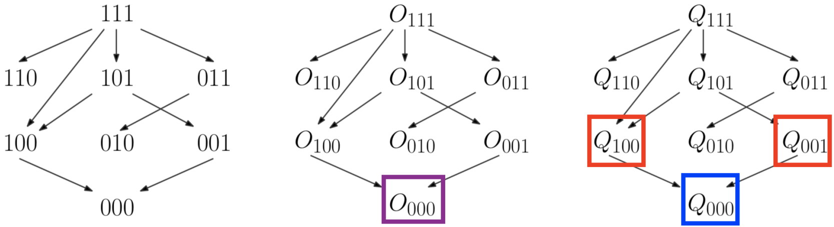
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- We compute Q_r using the parent(s) and the selection odds.

Graphical representation of the recursive computation

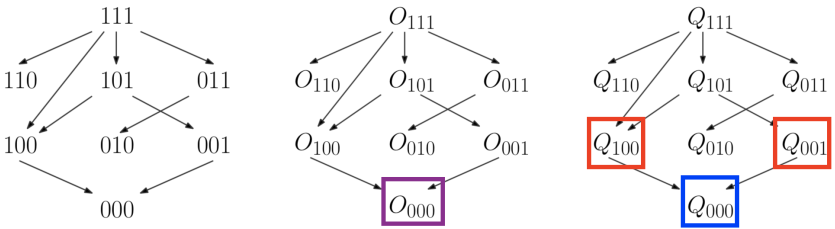
$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- We compute Q_r using the parent(s) and the selection odds.

Graphical representation of the recursive computation

$\widehat{Q}_r(L) = \widehat{O}_r(L_r) \sum_{s \in \mathsf{PA}_r} \widehat{Q}_s(L)$



- Having computed all Q_r, we can compute $\pi(L) = \frac{1}{\sum_r Q_r}$.

Regression adjustment and PMMs

- In addition to the IPW, we can rewrite

$$\theta_0 = \mathbb{E}(\theta(L)) = \mathbb{E}(\mathbb{E}(\theta(L)|L_R, R)) = \mathbb{E}(m(L_R, R)),$$

where $m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R)$ is the regression function under pattern R.

Regression adjustment and PMMs

- In addition to the IPW, we can rewrite

$$\theta_0 = \mathbb{E}(\theta(L)) = \mathbb{E}(\mathbb{E}(\theta(L)|L_R, R)) = \mathbb{E}(m(L_R, R)),$$

  where $m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R)$ is the regression function under pattern R.
- If we have estimator $\widehat{m}(L_R, R)$, then we can estimate the parameter of interest via

$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(L_{i,R_i}, R_i).$$

Regression adjustment and PMMs

- In addition to the IPW, we can rewrite

$$\theta_0 = \mathbb{E}(\theta(L)) = \mathbb{E}(\mathbb{E}(\theta(L)|L_R, R)) = \mathbb{E}(m(L_R, R)),$$

  where $m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R)$ is the regression function under pattern R.
- If we have estimator $\widehat{m}(L_R, R)$, then we can estimate the parameter of interest via

$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(L_{i,R_i}, R_i).$$

- We estimate $m(L_R, R)$ by the pattern graph and PMMs formulation.

Regression adjustment and PMMs

- In addition to the IPW, we can rewrite

$$\theta_0 = \mathbb{E}(\theta(L)) = \mathbb{E}(\mathbb{E}(\theta(L)|L_R, R)) = \mathbb{E}(m(L_R, R)),$$

  where $m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R)$ is the regression function under pattern R.
- If we have estimator $\widehat{m}(L_R, R)$, then we can estimate the parameter of interest via

$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(L_{i,R_i}, R_i).$$

- We estimate $m(L_R, R)$ by the pattern graph and PMMs formulation.
- You can show that if we use a Monte Carlo approximation to $\widehat{m}(L_{i,R_i}, R_i)$, this is identical to the multiple imputation method.

Semi-parametric theory

- In our on-going research, we are able to derive the efficient influence function (EIF) under a pattern graph.
- This leads to the following linear function:

$$\begin{aligned}\mathscr{L}_{\mathsf{semi}}(L,R) &= \frac{\theta(L)I(R=1_d)}{\pi(L)} + \mathsf{EIF}(L,R)\\ &= \frac{\theta(L)I(R=1_d)}{\pi(L)} + \sum_{r\neq 1_d}\sum_{\Xi\in\Pi_r}\sum_{s\in\Xi}\mathsf{EIF}_{\Xi,s}(L,R),\end{aligned}$$

  where

$$\mathsf{EIF}_{\Xi,s}(L,R) = \mu_{\Xi,s}(L_s)\left(I(R=s) - O_s(L_s)I(R\in\mathsf{PA}_s)\right)\prod_{w\in\Xi, w<s} O_w(L_w)$$

  is a 'pathwise' efficient influence function of a pattern s on a descending path Ξ.
- We prove that $\mathbb{E}(\mathscr{L}_{\mathsf{semi}}(L,R)) = \theta$ and the resulting estimator has a multiply-robust property.

Generalized pattern graphs and equivalence classes

Generalized pattern graphs

- A pattern graph is called a **generalized pattern graph** if

(G1) pattern $1_d = (1, 1, \cdots, 1)$ is the only source.

(DAG) the graph is a DAG.

Generalized pattern graphs

- A pattern graph is called a **generalized pattern graph** if

(G1) pattern $1_d = (1, 1, \cdots, 1)$ is the only source.

(DAG) the graph is a DAG.

Theorem

For a graph G that satisfies (G1) and (DAG) and $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$, then

1. *selection odds model and pattern mixture model factorizations are equivalent.*
2. *it leads to an (nonparametrically) identifiable full-data distribution.*

Generalized pattern graphs

- A pattern graph is called a **generalized pattern graph** if

(G1) pattern $1_d = (1, 1, \cdots, 1)$ is the only source.

(DAG) the graph is a DAG.

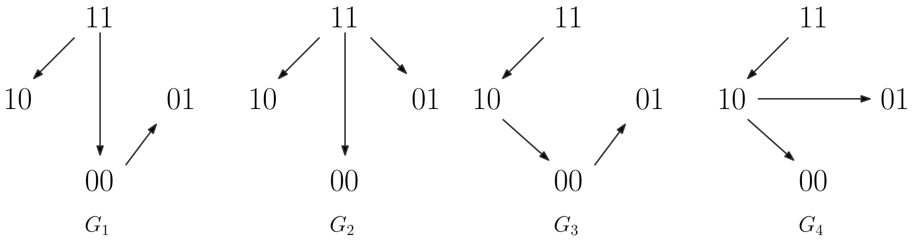
Theorem

For a graph G that satisfies (G1) and (DAG) and $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$, then

1. *selection odds model and pattern mixture model factorizations are equivalent.*
2. *it leads to an (nonparametrically) identifiable full-data distribution.*

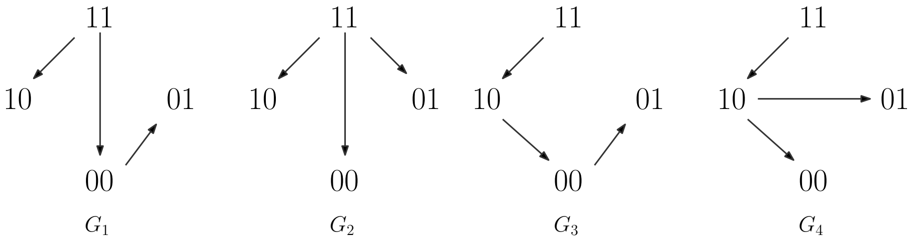
- The above theorem shows a powerful result–as long as the pattern graph has unique source 1_d and is a DAG, it can be used to represent an identifying restriction.

Example: equivalence classes



- These are generalized pattern graphs and each of them represent an identifying restriction.

Example: equivalence classes



- These are generalized pattern graphs and each of them represent an identifying restriction.
- Interestingly, G_1 and G_2 represent the same restriction; G_3 and G_4 represent the same restriction.
- Namely, G_1 and G_2 belong to the same equivalence class and G_3 and G_4 belong to another class.

A characterization of equivalence classes - 1

Theorem

Let G be a generalized pattern graph. For a pattern r and another pattern s such that $s \neq \mathsf{PA}_r$. This graph is equivalent to the graph G' such that

$$G' = G \oplus e_{s \to r} \ominus \{e_{\tau \to r} : \tau \in \mathsf{PA}_r\}$$

if the following conditions holds

1. **(blocking)** *all paths from 1_d to r intersects s.*
2. **(uninformative)** *for any pattern q that is on a path from s to r, $q < r$.*

A characterization of equivalence classes - 2

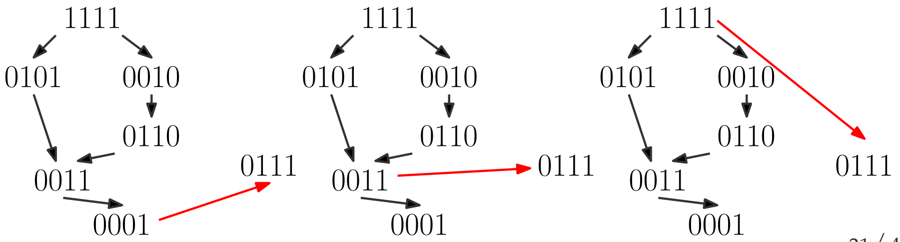
Theorem

Let G be a generalized pattern graph. For a pattern r and another pattern s such that $s \neq \mathsf{PA}_r$. *This graph is equivalent to the graph* G' *such that*

$$G' = G \oplus e_{s \to r} \ominus \{e_{\tau \to r} : \tau \in \mathsf{PA}_r\}$$

if the following conditions holds

1. **(blocking)** *all paths from* 1_d *to* r *intersects* s.
2. **(uninformative)** *for any pattern* q *that is on a path from* s *to* r, $q < r$.



Choice of pattern graph and PISA data - 1

- The choice of pattern graph reflects our knowledge on how the missingness is generated. Here is a possible way to choose a reasonable graph.
- We use the Programme for International Student Assessment (PISA) data at year 2009 as an example.
- It is a survey on students' ability on math, science, and literature from different countries.

Choice of pattern graph and PISA data - 1

- The choice of pattern graph reflects our knowledge on how the missingness is generated. Here is a possible way to choose a reasonable graph.
- We use the Programme for International Student Assessment (PISA) data at year 2009 as an example.
- It is a survey on students' ability on math, science, and literature from different countries.
- We focus on Germany and focus on three variables:
  - `MATH`: the math score (always observed).
  - `FA`: father's education level (H/L; may be missing).
  - `MA`: mother's education label (H/L; may be missing).

Choice of pattern graph and PISA data - 1

- The choice of pattern graph reflects our knowledge on how the missingness is generated. Here is a possible way to choose a reasonable graph.
- We use the Programme for International Student Assessment (PISA) data at year 2009 as an example.
- It is a survey on students' ability on math, science, and literature from different countries.
- We focus on Germany and focus on three variables:
  - MATH: the math score (always observed).
  - FA: father's education level (H/L; may be missing).
  - MA: mother's education label (H/L; may be missing).
- Here is the table of the response pattern $(R_{\mathtt{FA}}, R_{\mathtt{MA}})$:

$(R_{\mathtt{FA}}, R_{\mathtt{MA}}) =$	11	10	01	00
$n =$	3282	230	340	1126
Proportion=	65.9%	4.6%	6.8%	22.6%

Choice of pattern graph and PISA data - 2

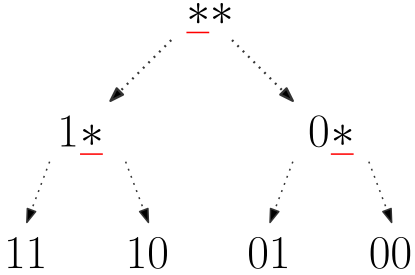
- Variables FA and MA are collected by questionnaire before a student took the exam.
- Suppose that a participant is asked about father's education first and then mother's education.

Choice of pattern graph and PISA data - 2

- Variables `FA` and `MA` are collected by questionnaire before a student took the exam.
- Suppose that a participant is asked about father's education first and then mother's education.
- Before asking any questions, every individual is expected to answer all questions so every one start with a response pattern $(1, 1)$. Then when asked a question, the participant will decide answer it or not.

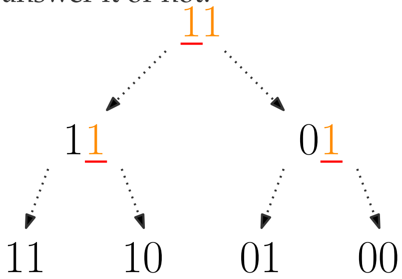
Choice of pattern graph and PISA data - 2

- Variables FA and MA are collected by questionnaire before a student took the exam.
- Suppose that a participant is asked about father's education first and then mother's education.
- Before asking any questions, every individual is expected to answer all questions so every one start with a response pattern $(1, 1)$. Then when asked a question, the participant will decide answer it or not.

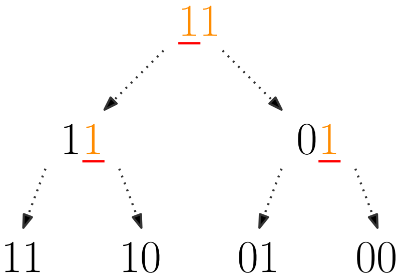


Choice of pattern graph and PISA data - 2

- Variables FA and MA are collected by questionnaire before a student took the exam.
- Suppose that a participant is asked about father's education first and then mother's education.
- Before asking any questions, every individual is expected to answer all questions so every one start with a response pattern $(1, 1)$. Then when asked a question, the participant will decide answer it or not.



Choice of pattern graph and PISA data - 3



- There will be 4 possible scenarios that an individual respond:

Answer FA and then answer MA $\Rightarrow$ 11 ▹ 11 ▹ 11

Answer FA and then not answer MA $\Rightarrow$ 11 ▹ 11 ▹ 10

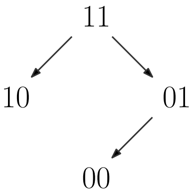
Not answer FA but then answer MA $\Rightarrow$ 11 ▹ 01 ▹ 01

Not answer FA and then not answer MA $\Rightarrow$ 11 ▹ 01 ▹ 00

Choice of pattern graph and PISA data - 4

$$\begin{aligned}
\text{Answer } \texttt{FA} \text{ and then answer } \texttt{MA} &\Rightarrow 11 \triangleright 11 \triangleright 11 \\
&\Rightarrow \text{path} = 11 \rightarrow 11 \\
\text{Answer } \texttt{FA} \text{ and then not answer } \texttt{MA} &\Rightarrow 11 \triangleright 11 \triangleright 10 \\
&\Rightarrow \text{path} = 11 \rightarrow 10 \\
\text{Not answer } \texttt{FA} \text{ but then answer } \texttt{MA} &\Rightarrow 11 \triangleright 01 \triangleright 01 \\
&\Rightarrow \text{path} = 11 \rightarrow 01 \\
\text{Not answer } \texttt{FA} \text{ and then not answer } \texttt{MA} &\Rightarrow 11 \triangleright 01 \triangleright 00 \\
&\Rightarrow \text{path} = 11 \rightarrow 01 \rightarrow 00.
\end{aligned}$$

- The notation $\triangleright$ denotes the decision of answering one question or not.
- $r_1 \triangleright r_2$ will becomes an arrow in a DAG when $r_1 \neq r_2$.
- The only exception is the scenario that $1_d \triangleright 1_d \triangleright \cdots \triangleright 1_d$; in this case we denote it as $1_d \rightarrow 1_d$.

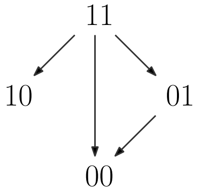


G_1

- The above plot is the pattern graph that corresponds to these scenarios:

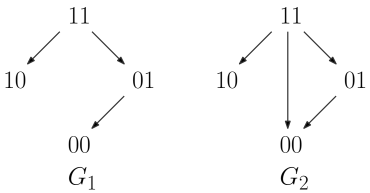
$$\text{Report } \texttt{FA} \text{ and then report } \texttt{MA} \Rightarrow \text{path} = 11 \rightarrow 11$$
$$\text{Report } \texttt{FA} \text{ and then not report } \texttt{MA} \Rightarrow \text{path} = 11 \rightarrow 10$$
$$\text{Not report } \texttt{FA} \text{ but then report } \texttt{MA} \Rightarrow \text{path} = 11 \rightarrow 01$$
$$\text{Not report } \texttt{FA} \text{ and then not report } \texttt{MA} \Rightarrow \text{path} = 11 \rightarrow 01 \rightarrow 00.$$

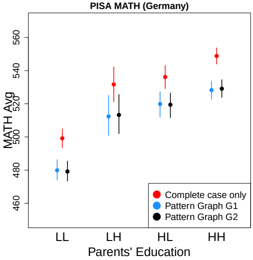
Choice of pattern graph and PISA data - 6



G_2

- Suppose that there are some individuals who would skip any questions relating to parent's education level.
- This can be represented by a path $11 \to 00$.
- Then the above graph will be a better description.





- The left two panels show the two possible pattern graphs.
- The right panel displays the average score of mathematics, separated by different parents' education level.
- The estimator is obtained by the IPW with logistic regression; uncertainty is obtained by the bootstrap.

Conclusion

Conclusion

- Pattern graph provides a theoretical framework for missing data.
- Identification, interpretation, estimation, efficiency, computation, sensitivity analysis all depend on the underlying pattern graph.
- It is a new graph-based model for data analysis.
- And it opens several new research directions.
- Note again: the pattern graph is not a conventional graphical model.

Future work

- **Pattern separation and missing data:** if a set of patterns A separates B and C, what does this mean?
- **Semi-parametric inference:** how to find the underlying efficient estimator with graph-based augmentation?
- **Merging patterns to avoid small sample size:** what should we do when some patten only has a few observations.
- **Deeper understanding on the equivalence class:** given a pattern graph, how to find other patterns in the same class?
- **Inference with multiple graphs:** what should we do if we have many identifying restrictions?

Thank You!

More details can be found in `https://arxiv.org/abs/2004.00744`.

References

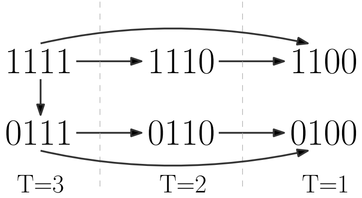
1. Chen, Y. C. (2020). Pattern graphs: a graphical approach to nonmonotone missing data. arXiv preprint arXiv:2004.00744.
2. Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. J. Am. Statist. Assoc., 88(421), 125-134.
3. Little, R. J. A., & Rubin, D. B. (2002). Statistical Analysis with Missing Data. Hoboken, New Jersey: Wiley, 2nd ed.
4. Little, R. (1995). Modeling the drop-out mechanism in longitudinal studies. Journal of the American Statistical Association, 90(1), 1.
5. Tchetgen, E. J. T., Wang, L., & Sun, B. (2018). Discrete choice models for nonmonotone nonignorable missing data: Identification and inference. Statistica Sinica, 28(4), 2069-2088.
6. Thijs, H., Molenberghs, G., Michiels, B., Verbeke, G., & Curran, D. (2002). Strategies to fit pattern-mixture models. Biostatistics, 3(2), 245-265.
7. Molenberghs, G., Michiels, B., Kenward, M. G., & Diggle, P. J. (1998). Monotone missing data and pattern-mixture models. Statistica Neerlandica, 52(2), 153-161.
8. Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley & Sons.
9. Robins, J. M. (1997). Non-response models for the analysis of non-monotone non-ignorable missing data. Statistics in Medicine, 16(1), 21-37.
10. Efron, B. (1994). Missing data, imputation, and the bootstrap. Journal of the American Statistical Association, 89(426), 463-475.
11. Liu, J. S. (2008). Monte Carlo strategies in scientific computing. Springer Science & Business Media.
12. Kim, J. K., & Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data. Journal of the American Statistical Association, 106(493), 157-165. 2011.

Example: Conditional MAR

- Let $L = (Z, Y_1, Y_2, Y_3)$ where Z is a covariate and Y_t is measured at different time points. Also, we define $R_z = R_1$ and $T = R_2 + R_3 + R_4$.
- Both Z and Y_t are subject to missing and the missingness of Y_t is monotone.

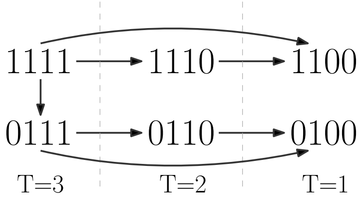
Example: Conditional MAR

- Let $L = (Z, Y_1, Y_2, Y_3)$ where Z is a covariate and Y_t is measured at different time points. Also, we define $R_z = R_1$ and $T = R_2 + R_3 + R_4$.
- Both Z and Y_t are subject to missing and the missingness of Y_t is monotone.



Example: Conditional MAR

- Let $L = (Z, Y_1, Y_2, Y_3)$ where Z is a covariate and Y_t is measured at different time points. Also, we define $R_z = R_1$ and $T = R_2 + R_3 + R_4$.
- Both Z and Y_t are subject to missing and the missingness of Y_t is monotone.



- Then the above pattern graph implies the following conditional MAR:

$$P(T = t | R_z = 1, L) = P(T = t | R_z = 1, Z, Y_1, \cdots, Y_t), \qquad t = 1, 2, 3$$

$$P(T = t | R_z = 0, L) = P(T = t | R_z = 0, Y_1, \cdots, Y_t), \qquad t = 1, 2, 3$$

$$P(R_z = 0 | T = 3, L) = P(R_z = 1 | T = 3, L) \cdot \frac{P(R_z = 0 | T = 3, Y_1, Y_2, Y_3)}{P(R_z = 1 | T = 3, Y_1, Y_2, Y_3)}$$

Assumptions on the IPW estimator

Let $\eta = (\eta_r : r \in \mathscr{R}) \in \Theta$ be any parameter value, where Θ is the total parameter space. We assume the following conditions:

(L1) there exists $\underline{O}, \overline{O}$ such that

$$0 < \underline{O} \leq O_r(\ell_r; \eta) \leq \overline{O} < \infty$$

for all $\ell_r \in \mathbb{S}_r$ and $r \in \mathscr{R}$ and $\eta \in \Theta$.

(L2) there exists $\eta^* = (\eta_r^* : r \in \mathscr{R})$ in the interior of Θ such that $O_r(\ell_r; \eta^*) = \frac{P(R=r|\ell_r)}{P(R\in \mathsf{PA}_r|\ell_r)}$ and

$$\sqrt{n}(\widehat{\eta}_r - \eta_r^*) \to N(0, \sigma_r^2), \quad \int \theta^2(\ell)(O_r(\ell_r; \widehat{\eta}) - O_r(\ell_r; \eta^*))^2 F(d\ell) = o_P(1),$$

for some $\sigma_r^2 > 0$ for all r.

(L3) for every r, the class $\{f_{\eta_r}(\ell_r) = O_r(\ell_r; \eta_r) : \eta_r \in \Theta_r\}$ is a Donsker class.

(L4) for every r, the differentiation of $O_r(\ell_r; \eta_r)$ with respect to η_r, $O_r'(\ell_r; \eta_r) = \nabla_{\eta_r} O_r(\ell_r; \eta_r)$, exists and $\int \|O_r'(\ell_r; \eta_r)\| F(d\ell_r) < \infty$ for a ball $B(\eta^*, \tau_0)$ for some $\tau_0 > 0$.

Assumptions on the regression adjustments

The regression adjustment estimator has asymptotic normality under the following conditions:

(R1) There exists $\lambda_r^* \in \Lambda_r$ such that the true conditional density $p(\ell_r|R=r) = p(\ell_r|R=r;\lambda_r^*)$ for every r.

(R2) For every r, the class

$$\{f_\lambda(\ell_r) = m(\ell_r, r; \lambda) : \lambda \in \Lambda\}$$

is a Donsker class.

(R3) For every r, $q_r(\lambda) = \mathbb{E}(m(L_r, r; \lambda)I(R=r))$ is bounded twice-differentiable and

$$\int (m(\ell_r, r; \widehat{\lambda}) - m(\ell_r, r; \lambda))^2 F(d\ell_r, r) = o_P(1)$$

$$\sqrt{n}(\widehat{\lambda}_r - \lambda_r^*) \to N(0, \sigma_r^2).$$

Regression adjustment and PMMs - 1

- In addition to the IPW, we can rewrite

$$\theta_0 = \mathbb{E}(\theta(L)) = \mathbb{E}(\mathbb{E}(\theta(L)|L_R, R)) = \mathbb{E}(m(L_R, R)),$$

where $m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R)$ is the regression function under pattern R.

Regression adjustment and PMMs - 1

- In addition to the IPW, we can rewrite

$$\theta_0 = \mathbb{E}(\theta(L)) = \mathbb{E}(\mathbb{E}(\theta(L)|L_R, R)) = \mathbb{E}(m(L_R, R)),$$

  where $m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R)$ is the regression function under pattern R.

- If we have estimator $\widehat{m}(L_R, R)$, then we can estimate the parameter of interest via

$$\widehat{\theta}_{\mathsf{RA}} = \frac{1}{n}\sum_{i=1}^{n} \widehat{m}(L_{i,R_i}, R_i).$$

Regression adjustment and PMMs - 2

- The regression function

$$m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R) = \int \theta(\ell_{\bar{R}}, L_R) p(\ell_{\bar{R}}|L_R, R) d\ell_{\bar{R}}$$

  is essentially the integral of $\theta(L)$ with respect to the extrapolation density.

Regression adjustment and PMMs - 2

- The regression function

$$m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R) = \int \theta(\ell_{\bar{R}}, L_R) p(\ell_{\bar{R}}|L_R, R) d\ell_{\bar{R}}$$

  is essentially the integral of $\theta(L)$ with respect to the extrapolation density.
- With a PMM, we can identify this regression function using an estimator $\widehat{p}(\ell_{\bar{R}}|L_R, R)$.

Regression adjustment and PMMs - 2

- The regression function

$$m(L_R, R) = \mathbb{E}(\theta(L)|L_R, R) = \int \theta(\ell_{\bar{R}}, L_R)p(\ell_{\bar{R}}|L_R, R)d\ell_{\bar{R}}$$

  is essentially the integral of $\theta(L)$ with respect to the extrapolation density.
- With a PMM, we can identify this regression function using an estimator $\widehat{p}(\ell_{\bar{R}}|L_R, R)$.
- This leads to

$$\widehat{m}(L_R, R) = \int \theta(\ell_{\bar{R}}, L_R)\widehat{p}(\ell_{\bar{R}}|L_R, R)d\ell_{\bar{R}}.$$

Monte Carlo approximation and multiple imputation

- The integral

$$\widehat{m}(L_R, R) = \int \theta(\ell_{\bar{R}}, L_R)\widehat{p}(\ell_{\bar{R}}|L_R, R)d\ell_{\bar{R}}$$

is hard to compute in general.

Monte Carlo approximation and multiple imputation

- The integral

$$\widehat{m}(L_R, R) = \int \theta(\ell_{\bar{R}}, L_R)\widehat{p}(\ell_{\bar{R}} | L_R, R) d\ell_{\bar{R}}$$

  is hard to compute in general.
- But we can numerically approximate it with Monte Carlo method.

Monte Carlo approximation and multiple imputation

- The integral

$$\widehat{m}(L_R, R) = \int \theta(\ell_{\bar{R}}, L_R)\widehat{p}(\ell_{\bar{R}}|L_R, R)d\ell_{\bar{R}}$$

  is hard to compute in general.
- But we can numerically approximate it with Monte Carlo method.
- We generate

$$L^*_{\bar{R},1}, \cdots, L^*_{\bar{R},N} \sim \widehat{p}(\ell_{\bar{R}}|L_R, R).$$

Monte Carlo approximation and multiple imputation

- The integral

$$\widehat{m}(L_R, R) = \int \theta(\ell_{\bar{R}}, L_R)\widehat{p}(\ell_{\bar{R}}|L_R, R)d\ell_{\bar{R}}$$

is hard to compute in general.

- But we can numerically approximate it with Monte Carlo method.
- We generate

$$L^*_{\bar{R},1}, \cdots, L^*_{\bar{R},N} \sim \widehat{p}(\ell_{\bar{R}}|L_R, R).$$

- Then use the average

$$\frac{1}{N}\sum_{k=1}^{N}\theta(L^*_{\bar{R},k}, L_R) \approx \widehat{m}(L_R, R).$$

Monte Carlo approximation and multiple imputation

- The integral

$$\widehat{m}(L_R, R) = \int \theta(\ell_{\bar{R}}, L_R)\widehat{p}(\ell_{\bar{R}}|L_R, R)d\ell_{\bar{R}}$$

  is hard to compute in general.
- But we can numerically approximate it with Monte Carlo method.
- We generate

$$L^*_{\bar{R},1}, \cdots, L^*_{\bar{R},N} \sim \widehat{p}(\ell_{\bar{R}}|L_R, R).$$

- Then use the average

$$\frac{1}{N}\sum_{k=1}^{N} \theta(L^*_{\bar{R},k}, L_R) \approx \widehat{m}(L_R, R).$$

- You can show that this is identical to the multiple imputation method!

Sampling from PMMs

- The PMM factorization implies

$$\begin{aligned}\widehat{p}(\ell_{\bar{r}}|L_r, R = r) &= \widehat{p}(\ell_{\bar{r}}|L_r, R \in \mathsf{PA}_r) \\ &= \sum_{s \in \mathsf{PA}_r} P(R = s|R \in \mathsf{PA}_r, L_r) \cdot \widehat{p}(\ell_{\bar{r}}|L_r, R = s).\end{aligned}$$

Sampling from PMMs

- The PMM factorization implies

$$\begin{aligned}\widehat{p}(\ell_{\bar{r}}|L_r, R = r) &= \widehat{p}(\ell_{\bar{r}}|L_r, R \in \mathsf{PA}_r) \\ &= \sum_{s \in \mathsf{PA}_r} P(R = s|R \in \mathsf{PA}_r, L_r) \cdot \widehat{p}(\ell_{\bar{r}}|L_r, R = s).\end{aligned}$$

- Moreover,

$$\widehat{p}(\ell_{\bar{r}}|L_r, R = s) = \widehat{p}(\ell_{\bar{s}}|\ell_{s-r}, L_r, R = s)\widehat{p}(\ell_{s-r}|L_r, R = s).$$

Sampling from PMMs

- The PMM factorization implies

$$\begin{aligned}\widehat{p}(\ell_{\bar{r}}|L_r, R = r) &= \widehat{p}(\ell_{\bar{r}}|L_r, R \in \mathsf{PA}_r)\\ &= \sum_{s \in \mathsf{PA}_r} P(R = s|R \in \mathsf{PA}_r, L_r) \cdot \widehat{p}(\ell_{\bar{r}}|L_r, R = s).\end{aligned}$$

- Moreover,

$$\widehat{p}(\ell_{\bar{r}}|L_r, R = s) = \widehat{p}(\ell_{\bar{s}}|\ell_{s-r}, L_r, R = s)\widehat{p}(\ell_{s-r}|L_r, R = s).$$

- It implies that we first choose a parent pattern $s \in \mathsf{PA}_r$ with a probability of $P(R = s|R \in \mathsf{PA}_r, L_r)$.

Sampling from PMMs

- The PMM factorization implies

$$\widehat{p}(\ell_{\bar{r}}|L_r, R = r) = \widehat{p}(\ell_{\bar{r}}|L_r, R \in \mathsf{PA}_r)$$
$$= \sum_{s \in \mathsf{PA}_r} P(R = s|R \in \mathsf{PA}_r, L_r) \cdot \widehat{p}(\ell_{\bar{r}}|L_r, R = s).$$

- Moreover,

$$\widehat{p}(\ell_{\bar{r}}|L_r, R = s) = \widehat{p}(\ell_{\bar{s}}|\ell_{s-r}, L_r, R = s)\widehat{p}(\ell_{s-r}|L_r, R = s).$$

- It implies that we first choose a parent pattern $s \in \mathsf{PA}_r$ with a probability of $P(R = s|R \in \mathsf{PA}_r, L_r)$.
- Then we fill-in variable ℓ_{s-r} by sampling from $\widehat{p}(\ell_{s-r}|L_r, R = s)$.

Sampling from PMMs

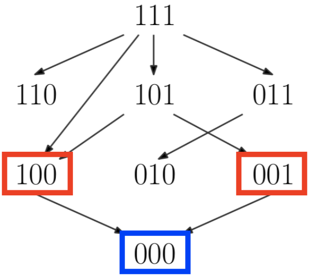
- The PMM factorization implies

$$\widehat{p}(\ell_{\bar{r}}|L_r, R = r) = \widehat{p}(\ell_{\bar{r}}|L_r, R \in \mathsf{PA}_r)$$
$$= \sum_{s \in \mathsf{PA}_r} P(R = s|R \in \mathsf{PA}_r, L_r) \cdot \widehat{p}(\ell_{\bar{r}}|L_r, R = s).$$

- Moreover,

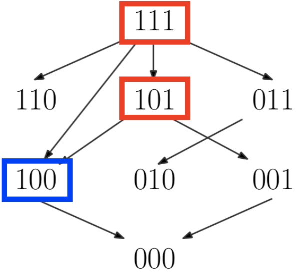
$$\widehat{p}(\ell_{\bar{r}}|L_r, R = s) = \widehat{p}(\ell_{\bar{s}}|\ell_{s-r}, L_r, R = s)\widehat{p}(\ell_{s-r}|L_r, R = s).$$

- It implies that we first choose a parent pattern $s \in \mathsf{PA}_r$ with a probability of $P(R = s|R \in \mathsf{PA}_r, L_r)$.
- Then we fill-in variable ℓ_{s-r} by sampling from $\widehat{p}(\ell_{s-r}|L_r, R = s)$.
- And treat this observation as the one with pattern $R = s$.



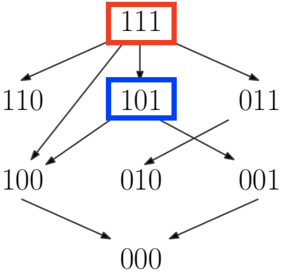
- Suppose we have an individual without any observed variables.
- It has two parents: 100 and 001 (red).
- We will randomly choose one parent as our next pattern.

Illustration: sampling from PMMs



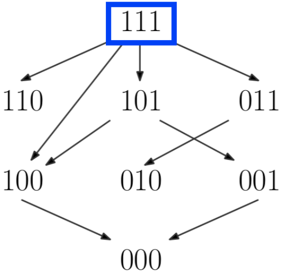
- Suppose that pattern 100 is chosen.
- We will generate variable L_{100} from $\widehat{p}(\ell_{100}|R = 100)$.
- Then we will treat this as an observation with pattern 100.
- Now we continue to randomly choose one pattern from the two parents (red).

Illustration: sampling from PMMs



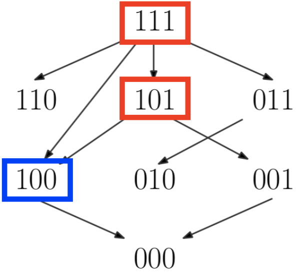
- Suppose that pattern 101 is chosen.
- We will generate variable L_{001} from $\widehat{p}(\ell_{001}|L_{100}, R = 100)$ because it is still missing.
- Then we will treat this as an observation with pattern 101.
- Now we continue to randomly choose one pattern from the parent set.

Illustration: sampling from PMMs



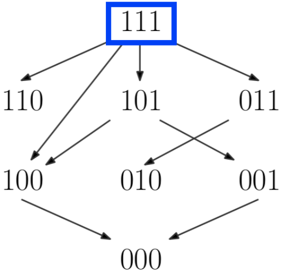
- Because there is only one parent 111, we will alway move to this node.
- We generate variable L_{010} from $\widehat{p}(\ell_{010}|L_{101}, R = 111)$.
- Now the pattern is 111 so we have finished the sampling/imputation.

Illustration: sampling from PMMs



- Note that at the pattern 100, it is possible to directly move to 111.

Illustration: sampling from PMMs



- In this case, we will generate L_{011} from $\widehat{p}(\ell_{011}|L_{100}, R = 111)$.
- And the sampling/imputation process is done.

Sensitivity analysis

Sensitivity analysis - 1

- Sensitivity analysis attempts to perturb our identifying restrictions a bit and study how the final estimator change with respect to the perturbation.

Sensitivity analysis - 1

- Sensitivity analysis attempts to perturb our identifying restrictions a bit and study how the final estimator change with respect to the perturbation.
- A common approach in sensitivity analysis is the exponential tilting (Kim and Yu 2011).

Sensitivity analysis - 1

- Sensitivity analysis attempts to perturb our identifying restrictions a bit and study how the final estimator change with respect to the perturbation.
- A common approach in sensitivity analysis is the exponential tilting (Kim and Yu 2011).
- For SMs, this can be written as

$$\frac{P(R=r|L)}{P(R \in \mathsf{PA}_r|L)} = \frac{P(R=r|L_r)}{P(R \in \mathsf{PA}_r|L_r)} \exp(L_{\bar{r}}^T \delta_{\bar{r}}),$$

where $\delta_{\bar{r}}$ is a given vector.

Sensitivity analysis - 1

- Sensitivity analysis attempts to perturb our identifying restrictions a bit and study how the final estimator change with respect to the perturbation.
- A common approach in sensitivity analysis is the exponential tilting (Kim and Yu 2011).
- For SMs, this can be written as

$$\frac{P(R = r|L)}{P(R \in \mathsf{PA}_r|L)} = \frac{P(R = r|L_r)}{P(R \in \mathsf{PA}_r|L_r)} \exp(L_{\bar{r}}^T \delta_{\bar{r}}),$$

  where $\delta_{\bar{r}}$ is a given vector.
- When $\delta_{\bar{r}} = 0$, we recover the original restriction.

Sensitivity analysis - 1

- Sensitivity analysis attempts to perturb our identifying restrictions a bit and study how the final estimator change with respect to the perturbation.
- A common approach in sensitivity analysis is the exponential tilting (Kim and Yu 2011).
- For SMs, this can be written as

$$\frac{P(R = r|L)}{P(R \in \mathsf{PA}_r|L)} = \frac{P(R = r|L_r)}{P(R \in \mathsf{PA}_r|L_r)} \exp(L_{\bar{r}}^T \delta_{\bar{r}}),$$

  where $\delta_{\bar{r}}$ is a given vector.
- When $\delta_{\bar{r}} = 0$, we recover the original restriction.
- For PMMs, we can use

$$p(\ell_{\bar{r}}|\ell_r, R = r) = p(\ell_{\bar{r}}|\ell_r, R \in \mathsf{PA}_r) \exp(L_{\bar{r}}^T \delta_{\bar{r}}).$$

Sensitivity analysis - 2

- Here is an interesting result–perturbing the selection odds and perturbing the pattern mixture models are equivalent.

Theorem

Let r be a response pattern and $g(\ell_{\bar{r}})$ be any function of the unobserved entries. Then the assumption

$$\frac{P(R = r|\ell)}{P(R \in \mathsf{PA}_r|\ell)} = \frac{P(R = r|\ell_r)}{P(R \in \mathsf{PA}_r|\ell_r)} \cdot g(\ell_{\bar{r}})$$

is equivalent to the assumption

$$p(\ell_{\bar{r}}|\ell_r, R = r) = p(\ell_{\bar{r}}|\ell_r, R \in \mathsf{PA}_r) \cdot g(\ell_{\bar{r}}).$$

Sensitivity analysis: perturbing graph - 1

- We can also perform sensitivity analysis via perturbing the graph.
- Before doing so, we first note that the number of identifying restrictions generated by regular pattern graphs is huge.

Sensitivity analysis: perturbing graph - 1

- We can also perform sensitivity analysis via perturbing the graph.
- Before doing so, we first note that the number of identifying restrictions generated by regular pattern graphs is huge.

Proposition

If all the study variable $L \in \mathbb{R}^d$ are subjected to missing, then there are

$$M = M_d = \prod_{k=0}^{d-1} (2^{2^{d-k}-1} - 1)^{\binom{d}{k}}$$

distinct graphs satisfying conditions (G1-2) .

Sensitivity analysis: perturbing graph - 1

- We can also perform sensitivity analysis via perturbing the graph.
- Before doing so, we first note that the number of identifying restrictions generated by regular pattern graphs is huge.

Proposition

If all the study variable $L \in \mathbb{R}^d$ are subjected to missing, then there are

$$M = M_d = \prod_{k=0}^{d-1} (2^{2^{d-k}-1} - 1)^{\binom{d}{k}}$$

distinct graphs satisfying conditions (G1-2) .

- Here are the first few values of $M = M_d$:

$$M_1 = 1, \quad M_2 = 7, \quad M_3 = 43561, \quad M_4 > 10^{18}.$$

Augmented IPW - 1

- It is known that we can improve the efficiency of the IPW via augmentation (Tsiatis 2007).

Augmented IPW - 1

- It is known that we can improve the efficiency of the IPW via augmentation (Tsiatis 2007).
- A general form of augmentation is

$$\frac{\theta(L)I(R = 1_d)}{\pi(L)} + \Psi(L_R, R),$$

where $\mathbb{E}(\Psi(L_R, R)) = 0$ and $\mathbb{E}(\Psi^2(L_R, R)) < \infty$.

Augmented IPW - 1

- It is known that we can improve the efficiency of the IPW via augmentation (Tsiatis 2007).
- A general form of augmentation is

$$\frac{\theta(L)I(R = 1_d)}{\pi(L)} + \Psi(L_R, R),$$

where $\mathbb{E}(\Psi(L_R, R)) = 0$ and $\mathbb{E}(\Psi^2(L_R, R)) < \infty$.

- Because we are modeling the selection odds to be observable, consider another augmentation:

$$\frac{\theta(L)I(R = 1_d)}{\pi(L)} + \sum_{r \neq 1_d} (I(R = r) - O_r(L_r)I(R \in \mathsf{PA}_r))\phi_r(L_r),$$

where $\mathbb{E}(\phi_r^2(L_r)) < \infty$ for each r.

Augmented IPW - 2

- All possible augmentations:

$$\mathcal{G} = \left\{ \frac{\theta(L)I(R = 1_d)}{\pi(L)} + \Psi(L_R, R) : \mathbb{E}(\Psi(L_R, R)) = 0, \mathbb{E}(\Psi^2(L_R, R)) < \infty \right\}.$$

- Augmentations using selection odds:

$$\mathcal{F} = \left\{ \frac{\theta(L)I(R = 1_d)}{\pi(L)} + \sum_{r \neq 1_d} (I(R = r) - O_r(L_r)I(R \in \mathsf{PA}_r))\phi_r(L_r) : \mathbb{E}(\phi_r^2(L_r)) < \infty \right\}.$$

Augmented IPW - 2

- All possible augmentations:

$$\mathcal{G} = \left\{ \frac{\theta(L) I(R = 1_d)}{\pi(L)} + \Psi(L_R, R) : \mathbb{E}(\Psi(L_R, R)) = 0, \mathbb{E}(\Psi^2(L_R, R)) < \infty \right\}.$$

- Augmentations using selection odds:

$$\mathcal{F} = \left\{ \frac{\theta(L) I(R = 1_d)}{\pi(L)} + \sum_{r \neq 1_d} (I(R = r) - O_r(L_r) I(R \in \mathsf{PA}_r)) \phi_r(L_r) : \mathbb{E}(\phi_r^2(L_r)) < \infty \right\}.$$

Theorem

Suppose that (L, R) factorizes with respect to a regular pattern graph and and $p(\ell_r, r) > 0$ for all ℓ_r and $r \in \mathcal{R}$. Then $\mathcal{G} = \mathcal{F}$.

Sensitivity analysis: perturbing graph - 2

- Because the regular pattern graphs span a large class of identifying restriction, we can perturb the graph to perform sensitivity analysis.
- Define

$$\Delta_1 G = \{G' : |G' - G| = 1, \text{condition (G1-2) holds for } G'\},$$

where $|G' - G| = 1$ means that the two graphs only differ by one edge (arrow).

Sensitivity analysis: perturbing graph - 2

- Because the regular pattern graphs span a large class of identifying restriction, we can perturb the graph to perform sensitivity analysis.
- Define

$$\Delta_1 G = \{G' : |G' - G| = 1, \text{condition (G1-2) holds for } G'\},$$

where $|G' - G| = 1$ means that the two graphs only differ by one edge (arrow).
- The class $\Delta_1 G$ can be decomposed into

$$\Delta_1 G = \Delta_{+1} G \cup \Delta_{-1} G,$$

where

$$\Delta_{+1} G = \{G' : |G' - G| = 1, G' \text{ is a regular pattern graph}, G \subset G'\},$$
$$\Delta_{-1} G = \{G' : |G' - G| = 1, G' \text{ is a regular pattern graph}, G' \subset G\}.$$

Sensitivity analysis: perturbing graph - 3

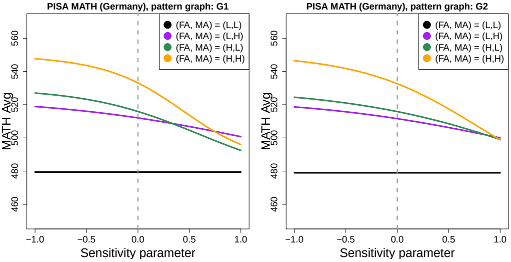
Proposition

Let s, r *be vertices of* G *and* $e_{s\to r}$ *be the edge/arrow from* s *to* r. *We define* $G \oplus e_{s\to r}$ *to be the graph where edge* $e_{s\to r}$ *is added and* $G \ominus e_{s\to r}$ *to be the graph where edge* $e_{s\to r}$ *is moved. Then*

$$\Delta_{+1}G = \{G \oplus e_{s\to r} : s > r, s \notin \mathsf{PA}_r\},$$
$$\Delta_{-1}G = \{G \ominus e_{s\to r} : s \in \mathsf{PA}_r, |\mathsf{PA}_r| > 1\}.$$

- This proposition provides a simple way to characterize the two perturbed classes of graphs.

Exponential tilting on the PISA data



- We use the same sensitivity parameter for all pattern and all values, i.e., every element of $\delta_{\bar{r}}$ is the same.
- Note that because only `FA` and `MA` are subject to missing, the sensitivity parameter only applies to these two variables.
- In both panels we see that the group (L, L) is unaffected by the sensitivity parameter. This is because when both `FA` and `MA` are L (the binary representation of L is 0 and H is 1).