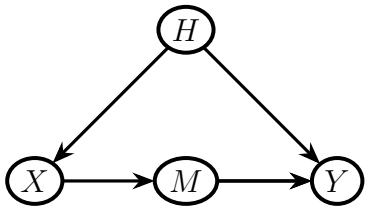
EXERCISES FOR RANDOMIZATION AND CAUSATION (MATH-336)

Exercise Sheet 10

Exercise 1 (Identification in another graph)**.** (From Technical Point 7.4 in [1]) Assume that variables X, M, Y satisfy the causal model $\mathcal{G}$ below, where we let H be an unmeasured variable. Furthermore, you can assume that all variables are discrete (and that Y is binary).



(a) Investigator 1 suggests the following identification formula (g-formula) for $E[Y^x]$:

$$E[Y^x] = E[Y \mid X = x] \ .$$

Show whether this identification formula holds or fails.

(b) Investigator 2 suggests another identification formula (not a g-formula) for a causal effect:

$$P(Y^x = 1) = \sum_m p(m \mid x) \sum_{x'} p(y \mid x', m) p(x') \ .$$

Show whether the identification formula holds or fails. You can assume that interventions on M are well-defined.

Hint: Draw several SWIGs corresponding on interventions on X, M, and both X and M. Next, remark that $Y^m = Y^x$ when $M^x = m$.

(c) State the positivity condition which is required for the identification formula in (b) to be well-defined.

(d) Prove that

$$E[Y^x] = \mathbb{E}\left[\frac{\pi(M \mid X = x)}{\pi(M \mid X)} Y\right] \ ,$$

where we defined π in the usual way as $\pi(\bullet \mid \circ) = P(M = \bullet \mid X = \circ)$. This is an IPW representation of the identification formula in part (b).

Solution:

In this exercise, we derive the front door formula.

(a) The identification formula fails because exchangeability on X fails, i.e. $Y^x \not\perp\!\!\!\perp X$.

(b)

$$P(Y^x = 1) \overset{\text{LOTP}}{=} \sum_m P(Y^x = 1 \mid M^x = m) P(M^x = m) \ .$$

The second factor can be written as

$$P(M^x = m) \overset{(M^x \perp\!\!\!\perp X)_{\mathcal{G}(x)}}{=} P(M^x = m \mid X = x) \overset{\text{consistency}}{=} P(M = m \mid X = x) \ .$$

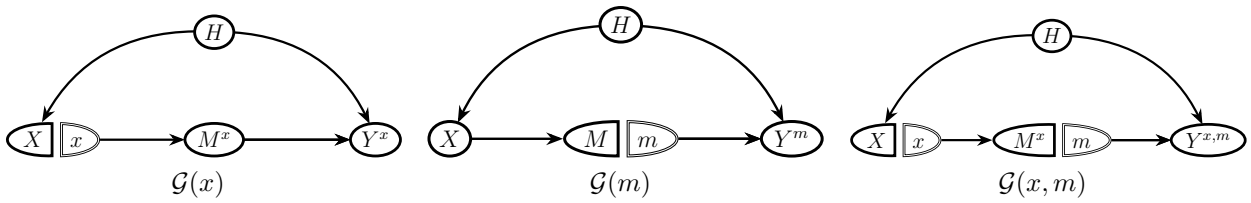
The first factor can be written as

$$P(Y^x = 1 \mid M^x = m) = P(Y^m = 1 \mid M^x = m)$$

because $Y^x = Y^m$ when $M^x = m$, because X only affects Y through the mediator M. Furthermore,

$$\begin{aligned} P(Y^m = 1 \mid M^x = m) &\overset{(Y^m \perp\!\!\!\perp M^x)_{\mathcal{G}(x,m)}}{=} P(Y^m = m) \\ &\overset{\text{LOTP}}{=} \sum_{x'} P(Y^m = 1 \mid X = x')P(X = x') \\ &\overset{(Y^m \perp\!\!\!\perp M \mid X)_{\mathcal{G}(m)}}{=} \sum_{x'} P(Y^m = 1 \mid M = m, X = x')P(X = x') \ . \end{aligned}$$

Using consistency, this shows that the identification formula is valid. Because identification failed in part (a), we thus deduce that exchangeability is not always necessary for identification. In the above argument, we have used independencies in the SWIGs $\mathcal{G}(x)$, $\mathcal{G}(m)$ and $\mathcal{G}(x, m)$.



(c) We require the following positivity conditions:
(1) $p(m \mid x') > 0$ for all m and for all x' such that $p(x') > 0$.
(2) $p(x) > 0$.
Condition (1) ensures that $p(y \mid x', m)p(x')$ is well-defined for all x' such that $p(x') > 0$, by ensuring that the conditioning set has non-zero probability. Condition (2) ensures that $p(m \mid x)$ is well-defined.

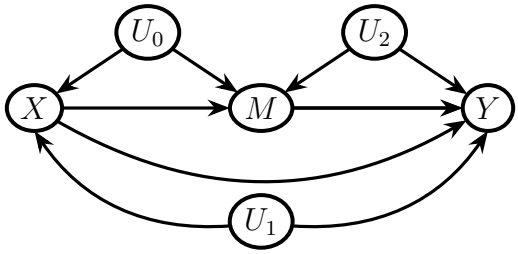
(d) We show that this is equal to the identification formula in part (b):

$$\begin{aligned} E\left[\frac{\pi(M \mid X = x)}{\pi(M \mid X)} Y\right] &= \sum_{myx'} p(x', m, y) \cdot y \cdot \frac{p(m \mid x)}{p(m \mid x')} \\ &= \sum_{myx'} y \cdot p(x', m, y) p(m \mid x) \frac{p(x')}{p(m, x')}. \\ &= \sum_{myx'} y \cdot p(y \mid m, x') p(m \mid x) p(x') \ . \end{aligned}$$

Exercise 2 (Mendelian randomization)**.** (Based on [2]) Consider a prospective Mendelian randomization study whose goal is to determine whether obesity is a cause of depression. Data are obtained on obesity ($M = 1$ indicates obese, $M = 0$ indicates non-obese), on

incident depression ($Y = 1$ indicates depressed, $Y = 0$ otherwise), and on genetic variants in the FTO gene.[1]

For simplicity, we define $X = 1$ if both of a subject's genetic variants (more specifically, FTO alleles) are the minor variants (alleles); $X = 0$ otherwise (i.e. if the subject is heterozygous or homozygous for the major allele.) Consider the DAG $\mathcal{G}$:



$\mathcal{G}$

We assume that this is the causal DAG generating the data, except some of the arrows may not actually be present. Furthermore, we assume all counterfactuals are well-defined and the consistency assumption holds. Finally we assume we have a near infinite study population so sampling variability can be ignored.

(a) (i) What arrows would have to be missing in order to have

$$Y^x \perp\!\!\!\perp X \ ?$$

Justify your answer by creating an appropriate SWIG.

(ii) If these arrows are missing give the identifying formula for $E[Y^{x=1} - Y^{x=0}]$ in terms of the distribution of the observed data on (X, M, Y).

(b) (i) What arrows would have to be missing to have

$$Y^m \perp\!\!\!\perp M \mid X \ ,$$
$$Y^m \not\perp\!\!\!\perp M \ ?$$

Justify your answer using an appropriate SWIG.

(ii) If these arrows are missing give the identifying formula for $E[Y^{m=1} - Y^{m=0} \mid X = x]$ in terms of the distribution of the observed data on (X, M, Y). Also give the identifying formula for the unconditional effect $E[Y^{m=1} - Y^{m=0}]$ of M on Y.

(c) (i) What arrows would have to be missing in order for the following independence statements to hold:

$$Y^m \not\perp\!\!\!\perp M \mid X \ ,$$
$$Y^m \perp\!\!\!\perp M \ ?$$

Justify your answer using an appropriate SWIG.

(ii) If these arrows are missing, give the identification formula for the unconditional effect $E[Y^{m=1} - Y^{m=0}]$ of M on Y.

(d) (i) What arrows would have to be missing in order for the joint effect $E[Y^{x,m} - Y^{x=0,m=0}]$ to be unconfounded, i.e. for

$$Y^{x,m} \perp\!\!\!\perp M \mid X = x \ ,$$

[1] FTO is a gene which is associated with obesity.

$$Y^{x,m} \perp\!\!\!\perp X \ ?$$

Justify your answer using an appropriate SWIG.

(ii) If these arrows are missing, give the identification formula for $E[Y^{x,m}]$ in terms of the distribution of the observed data on (X, M, Y).

(e) What arrows would have to be missing for the exclusion restriction

$$Y^{x=1,m} = Y^{x=0,m} \quad \text{for} \quad m = 0, 1$$

to hold for all subjects?

(f) (i) What arrow would have to be missing to have

$$Y^{x,m} \perp\!\!\!\perp X \ ?$$

Justify your answer using an appropriate SWIG.

(ii) If these arrows are missing, is $E[Y^{x,m}]$ point identified and, if so, what is the identifying formula in terms of the distribution of the observed data on (X, M, Y)?

(g) (i) What arrows would have to be missing for both

$$Y^{x,m} \perp\!\!\!\perp X$$

and exclusion restriction

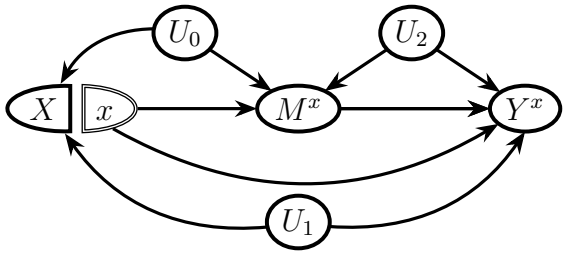
$$Y^{x=1,m} = Y^{x=0,m} \quad \text{for} \quad m = 0, 1$$

to hold?

(ii) If these arrows are missing, is $E[Y^{x,m}]$ point identified and what is the identifying formula in terms of the observed distribution on (X, M, Y)?

Solution:

(a) (i) The independence statement $Y^x \perp\!\!\!\perp X$ can be evaluated in the SWIG $\mathcal{G}(x)$. The



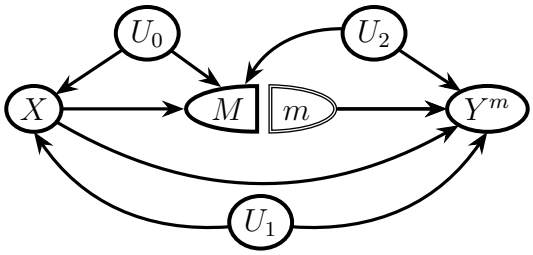
$\mathcal{G}(x)$

independence holds if we remove at least one arrow on each of the following paths: $X \leftarrow U_1 \rightarrow Y^x$ and $X \leftarrow U_0 \rightarrow M \rightarrow Y^x$.

(ii) The identification formula is given by

$$E[Y^{x=1}] - E[Y^{x=0}] = E[Y \mid X = 1] - E[Y \mid X = 0] \ .$$

(b) (i) The independencies can be evaluated using the SWIG $\mathcal{G}(m)$. We see that both independence statements hold if we remove (1) either arrow out of U_2 and 2a) either arrow out of U_0 or 2b) either arrow out of U_1. Note (2) is needed to prevent opening a confounding path when we conditon on X.



$\mathcal{G}(m)$

(ii) The identification formulas are

$$E[Y \mid M = 1, X = x] - E[Y \mid M = 0, X = x]$$

and

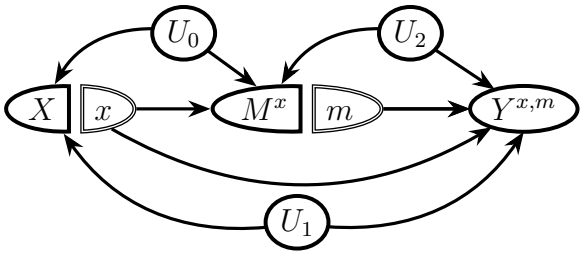
$$\sum_x P(X = x)(E[Y \mid M = 1, X = x] - E[Y \mid M = 0, X = x]),$$

respectively.

(c) (i) From the SWIG in part (b) (i), we see that the independence statements hold if we remove (1) either arrow out of U_2 and (2) $X \to M$ arrow, and $X \to Y$ arrow.

(ii) The effect is identified by $E[Y \mid M = 1] - E[Y \mid M = 0]$.

(d) (i) From the SWIG $\mathcal{G}(x, m)$, we see that the independence statements hold if we remove (1) an arrow out of U_1 and (2) an arrow out of U_2.



$\mathcal{G}(x, m)$

(ii) The effect is identified by $E[Y \mid X = x, M = m]$.

(e) The arrow $X \to Y$.

(f) (i) Either arrow out of U_1.

(ii) Not point identified.

(g) (i) Either arrow out of U_1 and arrow $X \to Y$.

(ii) Not point identified.

References

[1] Miguel Hernan and James M. Robins. *Causal Inference*. Chapman & Hall, 2018.

[2] J. M. Robins. EPI 207 (Harvard T.H. Chan School of Public Health).