CS 476/676 (Spring 2021): Homework 2

Due: Feb 17, 2021 at 11:59pm EST

Name: Stefano Tusa Lavieri

Instructions: This homework requires answering some open-ended questions, short proofs, and programming. This is an individual assignment, not group work. Though you may discuss the problems with your classmates, you must solve the problems and write the solutions independently. As stated in the syllabus, copying code from a classmate or the internet (even with minor changes) constitutes plagiarism. You are required to submit your answers in pdf form (use LaTeX) in a file called `<your-JHED>-hw2.pdf` to Gradescope under "HW2". Code should be submitted also to Gradescope under "HW2 Programming". Note that the autograder setup is intended only to register submission of the code, and will not provide any feedback. Code grading will be performed manually. Late submissions will be penalized, except in extenuating circumstances such as medical or family emergency. Submissions submitted 0-24 hours late will be penalized 10%, 24-48 hours late by 20%, 48-72 hours late by 30%, and later than 72 hours by 100%. Late days may be used (if available) to avoid these penalties.

The total assignment is worth 85 points.

Problem 1 (5 points)

Briefly explain (for both senses of completeness) the statement – "d-separation is *sound* and *complete* for detecting conditional independences in statistical models of a DAG where all variables are observed." 2-3 paragraphs of explanation are sufficient.

Soundness: If a d-separation exists in a graph G, that is, $(X \perp\!\!\!\perp Y \mid Z)_{\text{d-sep}}$, soundness says that that conditional independence exists in any distribution P(v) that factorizes w.r.t G.
Completeness: Suppose that two variables X, Y are not d-separated given a third set of variables Z. Then, depending on our criteria for selecting a factorizing distribution P(Z), we could either have
- Weak completeness: whereby $\exists$ at least one distribution p(z) that factorizes w.r.t G where $X \not\perp\!\!\!\perp Y \mid Z$, or , if we restrict ourselves to faithful distributions for our choice of p(Z) we could get...
- Strong completeness, and every distribution p(Z) would have $X \perp\!\!\!\perp Y \mid Z$.

Problem 2 (15 points)

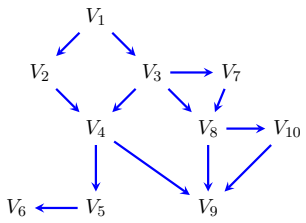


Figure 1

Refer to Figure 1 for this problem.

1) Assuming $V_1, \ldots, V_{10}$ are binary random variables, how many parameters are required to specify the full joint distribution $p(V)$ without a graphical model? (3 points)

For k binary variables $\rightarrow 2^k - 1$
$2^{10} - 1 = 1023$ parameters.

2) Assuming binary random variables, how many parameters are required to specify the joint distribution using the DAG factorization for Figure 1? Show your work. (5 points)

$p(v_1) \times p(v_2 \mid v_1) \times p(v_3 \mid v_1) \times p(v_4 \mid v_2, v_3) \times p(v_7 \mid v_3) \times p(v_8 \mid v_3, v_7) \times p(v_5 \mid v_4) \times p(v_6 \mid v_5) \times p(v_9 \mid v_4, v_8, v_{10}) \times p(v_{10} \mid v_8)$
$1 + 2 + 2 + 4 + 2 + 4 + 2 + 2 + 8 + 2 = 29$

3) Answer the following d-separation queries and for any queries where they are not d-separated, provide at least one active/unblocked path. Is

a) $V_2 \perp\!\!\!\perp_{\text{d-sep}} V_9 \mid V_4$?
b) $V_7 \perp\!\!\!\perp_{\text{d-sep}} V_5 \mid V_3, V_8$?
c) $V_2, V_4 \perp\!\!\!\perp_{\text{d-sep}} V_7 \mid V_6, V_9, V_{10}$?

(7 points)

a) $V_2 \perp\!\!\!\perp_{\text{d-sep}} V_9 \mid V_4$?
$v_2 \to v_4 \to v_9$ (closed), $v_2 \to v_4 \leftarrow v_3 \to v_8 \to v_9$ (closed)
$v_2 \to v_4 \leftarrow v_3 \to v_7 \to v_8 \to v_{10} \to v_9$ (closed)
d-separation holds

b) $V_7 \perp\!\!\!\perp_{\text{d-sep}} V_5 \mid V_3, V_8$?
$v_7 \leftarrow v_3 \to v_4 \to v_5$ (closed), $v_7 \to v_8 \to v_9 \leftarrow v_4 \to v_5$ (closed)
$v_7 \to v_8 \to v_{10} \to v_9 v_4$ (closed) $v_7 \leftarrow v_3 \leftarrow v_1 \to v_2 \to v_4 \to v_5$ (closed)
d-separation holds

c) $v_2, v_4 \perp\!\!\!\perp v_7 \mid v_6, v_9, v_{10}$?
$v_2 \leftarrow v_1 \to v_3 \to v_7$ (open!)
d-separation does NOT hold

Problem 3 (30 points)

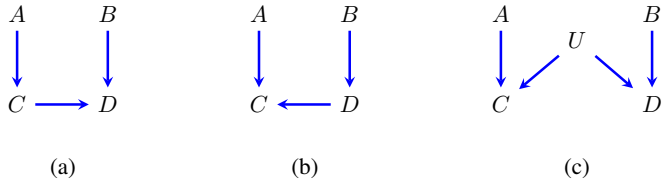


Figure 2

Refer to Figure 2 for this problem. Say we are given data on four variables $A, B, C,$ and D (observed variables) and are unable to obtain data on the variable U. Answer the following questions:

1) Describe qualitatively how the causal interpretations for (a), (b), and (c) are different from each other. A few sentences/a paragraph is sufficient. (3 points)

In all three scenarios, A and B are posited as directly causing C and D respectively. The interpretations differ as to what the relationship between C and D is. a) and b) say that either C directly causes D or that D directly causes C, again respectively. c) instead understands the causal relationship between C and D as being that they are both mutually caused by the unobserved variable U.

2) List all independences that comprise the local Markov property for the DAGs in (a) and (b). Based on these two lists, identify independences present in (a) but not in (b), and vice versa. (5 points)

a) $A \perp\!\!\!\perp B, D \perp\!\!\!\perp A(*)$
b) $A \perp\!\!\!\perp B, C \perp\!\!\!\perp B(*)$
Asterisks denote CI relations unique to each DAG.

3) Use d-separation to identify a **set** of conditional independences implied over the **observed** variables in (c) that is unique to (c), i.e., this set should not be implied in (a) and (b). (5 points)

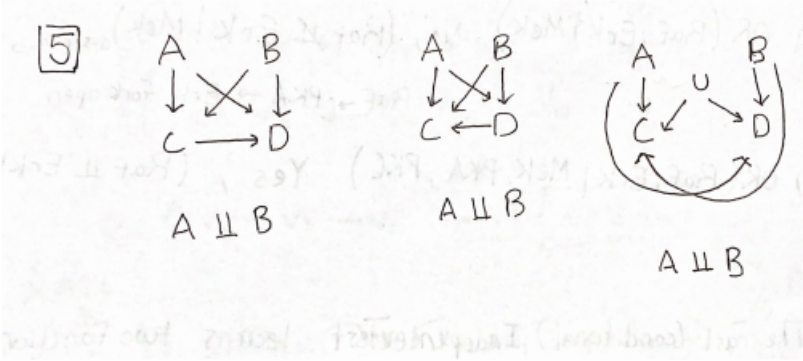
$\{A \perp\!\!\!\perp B \mid D,$
$C \perp\!\!\!\perp B \mid D$
$A \perp\!\!\!\perp D \mid C$
$A \perp\!\!\!\perp B \mid C\}$

4) Based on your answers to 2) and 3), and assuming that the true data generating process for A, B, C, D, and U corresponds to an NPSEM-IE defined over one of the DAGs shown in Figure 1, describe how you would construct a series of statistical hypothesis tests to arrive at the true causal DAG.[1] What is the underlying assumption used to ensure that these tests allow us to choose between these structures that imply different d-separation statements? (7 points)

[1] Ignore, for now, issues related to multiple testing (if you are aware of what that is.)

Since the NPSEM-IE induces a probability distribution w.r.t whatever DAG it is defined upon, the series of statistical hypothesis tests could be used to determine how likely the DAG is to accurately reflect the true data generation distribution. The underlying assumption at play here is that conditional independences present in this true data generation distribution p(z) will manifest themselves in the structure of our DAGs via d-separation of the appropriate vertices (faithfulness).

5) Now imagine that the edges $A \rightarrow D$ and $B \rightarrow C$ were also present in the DAGs shown in (a), (b), and (c). Show that these new DAGs are **not** statistically distinguishable using observed data. (5 points)



Since they all have the same set of independences, it would be impossible to distinguish these DAGS using only observed data.

6) Summarize, in a few sentences, what the above exercise tells you about the use of statistical tests in the statistical model of a DAG to learn information regarding causal models of a DAG. (5 points)

The ability of statistical tests to learn information regarding causal models of a DAG depends on the set of independences (conditional or otherwise) present within the graph. If we have several plausible DAGS, we can tests the validity of each ones' conditional independences using observed data, but if these DAGS contain identical sets of independences (like in problem 5), distinguishing between them becomes impossible.

Problem 4 (10 points)

Given a DAG $\mathcal{G}$ over a set of variables V. Define the Markov blanket of a variable V_i as,

$$\mathrm{mb}_{\mathcal{G}}(V_i) \equiv \mathrm{pa}_{\mathcal{G}}(V_i) \cup \mathrm{ch}_{\mathcal{G}}(V_i) \cup \mathrm{pa}_{\mathcal{G}}(\mathrm{ch}_{\mathcal{G}}(V_i)) \setminus \{V_i\},$$

i.e., the parents of V_i, the children of V_i, and the parents of the children of V_i, and excluding V_i itself.

Use d-separation to prove that the Markov blanket of V_i satisfies $V_i \perp\!\!\!\perp V \setminus \{\mathrm{mb}_{\mathcal{G}}(V_i), V_i\} \mid \mathrm{mb}_{\mathcal{G}}(V_i)$. In words, prove that the Markov blanket of V_i shields (or covers) V_i from all other variables in $\mathcal{G}$ in the following sense: V_i is independent of all other variables in $\mathcal{G}$ given its Markov blanket.

Three cases:
- Siblings of v_i: conditioning on parents of v_i breaks flow of association through fork triplet.
- Nieces of v_i: conditioning on children of v_i breaks flow of association through chain triplet.
- Step-siblings of v_i: conditioning on parents of children of v_i breaks flow of association through fork triplet.

Problem 5 (5 points)

In the case study on the link between estrogens and endometrial cancer we saw that $\text{OR}(\text{Estrogens}, \text{Diagnosis} \mid \text{Test})$ was not a valid test statistic for the causal null. What rule of d-separation came into play when

detecting this bias? Would you say that this is also a kind of Berksonian bias?

This was an instance of collider bias, which also goes by the name Berkson's bias–the problem arising because we were attempting to condition on a collider. Conditioning on Test, a descendant of Estrogens, opened a path from Estrogens to Diagnosis. Attempting to resolve this by conditioning on Bleeding would've led to the opening of multiple paths from Estrogens to Diagnosis.

Problem 6 (20 points)

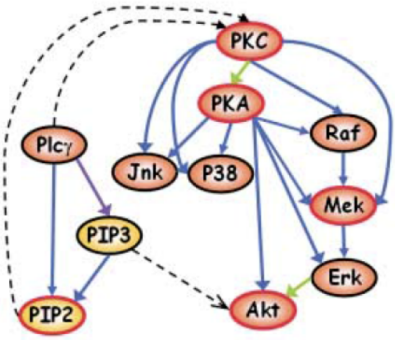


Figure 3: A DAG learned from phospho-protein/phospho-lipid expression data

This problem contains an analysis component and a programming component.

You are provided data that is drawn from the paper Sachs et al. (2005) in the file `data.txt`. In this paper, the authors infer a protein-signaling network over the set of proteins/lipids that they measured. The resulting DAG is shown in Figure 3. One of the remarkable results of the paper was the exact reconstruction of the protein kinase cascade Raf $\to$ Mek $\to$ Erk, including the absence of direct regulation between Raf and Erk, corresponding to the absence of Raf $\to$ Erk. Your goal here is to see if you can recreate the latter finding (absence of direct regulation) from the data they have provided. Assume all colors of edges (including dashed ones) in Figure 3 are present in the null DAG.

1) Pick (and justify) which of the following is a valid test statistic for the causal null (5 points)

$$H_0: \text{ Raf} \to \text{Erk is absent.}$$

a) OR(Raf, Erk | Mek)
b) OR(Raf, Erk | Mek, PKA, PKC)

OR(Raf, Erk | Mek) No, $\left(\text{Raf} \not\perp\!\!\!\perp \text{Erk} \mid \text{Mek}\right)_{\text{d-sep in G}}$, Raf $\leftarrow$ PKA $\rightarrow$ Erk fork open

OR(Raf, Erk | Mek, PKA, PKC) Yes; Raf $\perp\!\!\!\perp$ Erk | Mek, PKA, PKC

2) If you take a look at the data, you will notice that it consists of mostly continuous valued variables. This makes computing the odds ratio fairly difficult, unless we assume simple parametric models like linear Gaussian models. However, we expect biological relations to be non-linear. We will instead use a non-parametric conditional independence test called the "Fast Conditional Independence Test (FCIT)" described by Chalupka et al. (2018). Read the introduction of the paper till the end of Section 1.1 and Section 2 till the end of Section 2.1 to get a sense of what the test does. Write 2-3 sentences to provide your understanding (this does not need to be rigorous, just demonstrate you tried reading about the method.) (5 points)

The Fast (conditional) Independence Test learns two functions, $X, Z \rightarrow Y$ and $Z \rightarrow Y$. The intuition is that if $X \not\perp\!\!\!\perp Y \mid Z$, then prediction of Y using both X and Z as covariates would be more accurate than prediction using X alone. FCIT uses the Mean-Squared Error metric to compare the results of both learned functions, and after a number of iterations determines the p-value of the null hypothesis that the MSEs of the function fitted using X and Z contains on average higher values than the MSEs of the function only using X.

3) Instead of computing the OR, use an implementation of FCIT from the `fcit` package as a non-parametric test of conditional independence for the causal null based on your answer to 1). The skeleton for the output of your program is provided in `hw2.py`. Using a significance level of $\alpha = 0.05$, interpret the output of whichever test you decided was valid. Report the p-value and interpretation in your writeup. (10 points)

$0.010803750532695354 < \alpha = 0.05$
reject the null; Raf $\not\perp\!\!\!\perp$ Erk | Mek

$0.24451177158165738 > \alpha = 0.05$
fail to reject the null; suggests that Raf $\perp\!\!\!\perp$ Erk | Mek, PKA, PKC

Helpful notes:

- The capitalization of the variable names in the data file may not exactly match the capitalization of the variables shown in Figure 3. But it should be fairly easy to tell which one is which.
- The GitHub page for FCIT provides helpful usage and installation tips: `https://github.com/kjchalup/fcit`.
- The inputs to FCIT have to be `numpy` matrices. So make sure all your inputs are at least $n \times 1$ `numpy` matrices (not 1-d arrays) where n is the number of samples of data.
- Don't worry too much if the p-value from your implementation does not perfectly fit the story of the Raf $\rightarrow$ Mek $\rightarrow$ Erk cascade – that is the point of re-examining the actual data and determining the robustness of the conclusion (Sachs et al. (2005) used linear models.) The FCIT test can be quite sensitive to random number initialization. If you implement the method correctly, and interpret the number appropriately according to the given significance level, you will receive full points. You may also run FCIT a few times over to see how sensitive your final conclusion is, but this is not necessary.

References

Chalupka, K., Perona, P., and Eberhardt, F. (2018). Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*.

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P. (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529.