Meeting Notes

$\LaTeX\,2_\varepsilon$

January 24, 2020

Contents

Minutes 1

Those present Stephen Malina, Daniel Cizin

Date January 21, 2020

List of topics

1.1 Current status

1.1.1 What I (Stephen) did over break

- Run MR in the opposite direction (accessibility on TF binding).
- Get Kipoi working and do MR on top of data generated by it.
- Think about whether the single instrument framework makes sense.

1.2 Stuff we should discuss

Task:

Minutes 2

Meeting with Claudia

Those present Stephen Malina, Claudia Shi

Date January 21, 2020

List of topics

2.1 Initial approach problems and assumptions

We started out with me describing the problem with our initial approach. I'd summarize this as, "we didn't account for the fact that mutation impacts binding and accessibility through the sequence mediator."

Then, as I was describing the problem, we realized that, if we have the following graph, then our original approach reflects an implicit assumption of a uniform distribution over sequences.

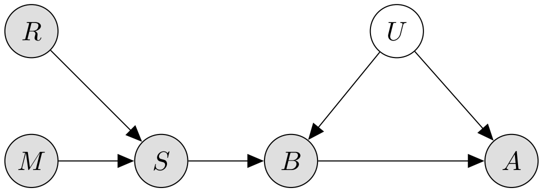

Figure 2.1: Graph of the causal DAG we get if we assume that mutation is upstream of sequence. Note that M represents mutation, S sequence, B binding, R randomization, and A accessibility.

Our data generation process using our neural net (in the binding case) gives us samples of the conditional distribution $\Pr(b \mid \mathrm{do}(m), s)$ and we want to estimate $\Pr(B \mid \mathrm{do}(M))$. Basic probability algebra gives us

$$\Pr(B \mid \mathrm{do}(M)) = \int \Pr(B \mid \mathrm{do}(M), S = s) \Pr(S = s) ds.$$

My understanding is that if we assume S is uniformly distributed, we can ignore the prior probability of S piece and then we get

$$\Pr(B \mid \mathrm{do}(M)) = \int \Pr(B \mid \mathrm{do}(M), S = s) ds,$$

which we can approximate by sampling from the distribution of $\Pr(B \mid S)$.

Assuming a uniform distribution over S seems like a strong assumption though.

2.2 Multi-DAG aggregation approach

Then we discussed an alternate that David had hinted at in which we treat each sequence as instantiating its own DAG and then try to aggregate the causal effects from each. The benefit of this approach is that it's conceptually clearer. The downside is that it's still not clear to me that it makes to only have one sample from each DAG. On the other hand, if we try to get multiple samples from each DAG, we run into the problem of there being no obvious way to impose a dimension on the space of possible mutations but it seems like IV methods typically assume there's one.

Task: o (L)ok into the methods Claudia mentioned may allow us to aggregate DAGs: TLME, IPTW, AIPTW (or AIPPW, I'm not sure).

2.3 Deep IV approach

Finally, we discussed the Deep IV approach and Claudia mentioned that a bunch of people in Dave's lab are working on things related to Deep IV, including Victor.