Meeting Notes

LaTeX 2_ε

January 29, 2020

Contents

Minutes 1

Status Update and Next Steps Discussion

Those present Daniel Cizin, David Knowles, Stephen Malina

Date January 29, 2020

We discussed a few topics:

1. Getting standard error estimates from NN predictions.
2. Inverse variance weighted methods for conducting meta-analyses.
3. How to do in-silico mutation such that we don't condition on the exposure and also get effects we can combine.

Summarizing our decisions in order of dependency rather than chronologically:

1. At least for now, we're going to try and get results from the "multiple DAG instantiations with mutations as IVs" framework.
2. High-level our strategy will be:
    - randomly sample sequences;
    - do saturation mutagenesis for each, getting a standard error for each mutation's prediction;
    - run Egger (or similar, eg̈; RAPS) on the result for each sequence, treating the difference between the binding probability for the mutated and wild type sequences as an effect size; and
    - combine the causal estimates for each sequence together using a meta-analysis protocol.
3. For getting standard errors for individual predictions, we'll using Yarin Gal's method of repeated dropout.

4. We hope that using a robust regression method will allow us to get an estimate of how much violation of exclusion-restriction there is in our model.

Given this, our follow-ups are:
Task: Verify that with classification, we don't care about the noise variables in the prediction variance formula. (Stephen)
Task: Figure out and implement the saturation mutagenesis with uncertainty estimates logic on top of Kipoi. (Stephen)
Task: Do some research into how likely it seems that sequence will influence accessibility directly. (Daniel)
Task: Try and understand the difference between Egger regression and RAPS. (Daniel)
Task: Keep thinking about the idea of using cell type, potentially in a deep IV framework, as an alternative IV.
Task: Keep thinking about whether there's something interesting in investigating the covariance betwen predictions from saturation mutagenesis (see figure 1.2).

On the next page, I put the two pictures I took of the whiteboard during our meeting.

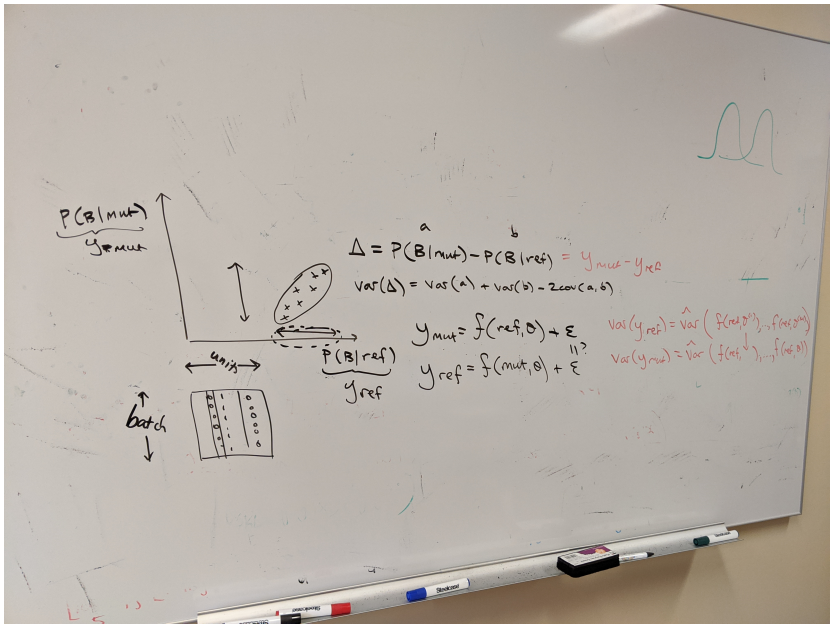

Figure 1.1: Shows two things: that we should take the std error of the diff between mut and wild type probabilities and that we want to use the same dropout mask for mut and wild type predictions so that we can estimate the covariance between the two.

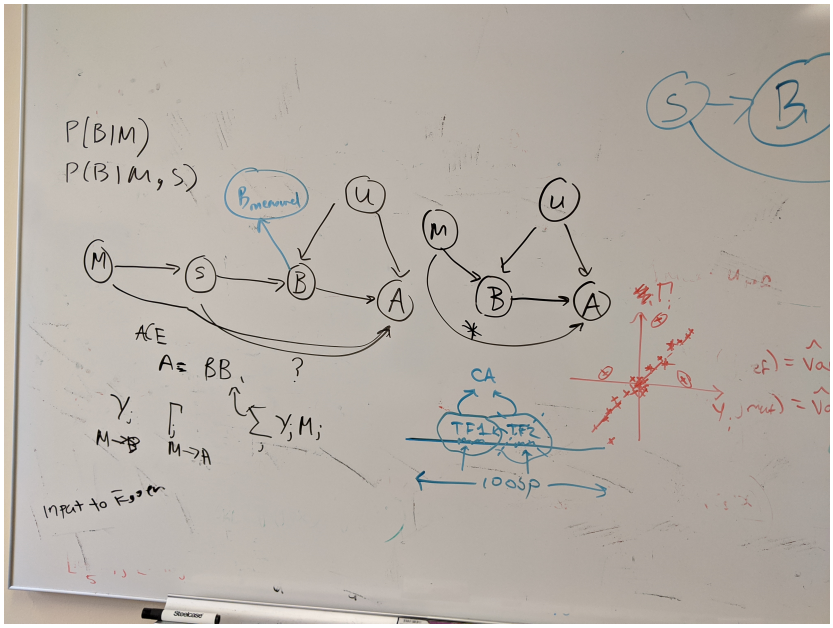

Figure 1.2: Captures two notable things. First, that we mostly came to the conclusion that because of potential exclusion restriction violations, it makes to analyze the causal effect separately for each sequence and then aggregate. Second, that there may be something interesting in the covariance between different mutation predictions for the same sequence.