1

Discovering Causal Graph from a Longitudinal Education Study

Irfan Jamil ijamil1@jhu.edu
Computer Science and Applied Mathematics and Statistics
ijamil1

1. Introduction

The data I will be working with was gathered during the National Educational Longitudinal Study of 1988 (NELS). This was a survey of a nationally representative group of 8th graders; the sample consisted of about 2000 students. In this data set, there are somewhere between 250 and 300 attributes (all of which are numeric but the vast majority are discrete and a few are continuous). Attributes encode various pieces of information about the student's parents' education, marital status, student's gender, race, number of siblings, academic expectations/goals, academic achievement, extracurricular involvement, drug/alcohol use, number of hours spent watching TV, family income, and so forth.

The goal of my project is to discover the causal graph which represents the mechanism generating the data distribution from which this data was sampled. Specifically, I would like to discover what impacts students' academic performances. My hypotheses are that increased parental education leads to better academic performance in students, that parents' occupations with higher salaries will be a cause of the child having high expectations for themselves in regards to academics and profession, and that higher socioeconomic statuses are a cause of a student's academic success.

With that being said, the bigger picture is to understand the various interacting influences on a student's life and academics and then identify areas where intervention can be taken to improve the student's life and academic performance.

2. Choice of Graphical Model and Its Justification

For my analysis, I will be representing the distribution and causal mechanism from which my data arises with a chain graph. I will justify my use of chain graphs by providing reasoning as to why I chose not to represent the distribution with a different graphical model and explaining the benefits of using a chain graph.

Given the amount of information that the variables in my data represent, I could not justify precluding the possibility that there exists pairs of variables in the chain graph that exhibit a dynamic feedback relationship. This rules out a DAG.

On the other hand, I could represent the data distribution with an undirected graph (UG). If I were to represent the distribution in this way, I would be precluding the existence of directed edges. In regards to the causal model of the graph, this would mean ruling out the existence of a directed causal relationship between a pair of variables.

In the causal model of a chain graph, direct causal relationships between variables as well as dynamic feedback or contagion relationships can exist between variables. This flexibility made it the best option for the data I am working with. However, I was seriously contemplating the use of an ADMG instead. The edge set of an ADMG consists of both directed edges and bi-directed edges. In the causal model of an ADMG, directed edges represent direct causal relationships and bi-directed edges represent an unobserved confounder. I was unsure whether it was more appropriate for variables in my data to have a dynamic feedback or contagion causal relationship or a relationship defined by correlation resulting from unobserved confounding. In choosing to conduct my analysis with the use of chain graphs, I lost the power to represent the causal relationship between two variables with the unobserved confounder. The consequence of this decision is that all external influences on a variable can be captured by other variables that are within the dataset. That is, I'm making the assumption that all causal effects on a variable can be discerned from the data. Given that the variables I'm analyzing are diverse and collectively represent many different attributes and characteristics of a person, the assumption I am making is large and can be validly criticized.

3. Preliminaries

Chain graphs are a simple graph in which the edge set includes only directed edges and undirected edges. Furthermore, there exist no directed or partially directed cycles in the graph. The statistical interpretation of a chain graph is that distributions that can be represented by a chain graph, $\mathcal{G}$, can be factorized in to the following form:

$$p(V) = \prod_{B_i \in B} p(B_i \mid \text{pa}_{\mathcal{G}}(B_i)),$$

where the set B consists of all the blocks in the chain graph. A block refers to a set of undirected connected components of the CG. Blocks form a partition of the vertex set and since there does not exist any directed cycles or partially directed cycles in a CG, the above factorization makes sense and is analogous to the factorization of a DAG with the parents of a block simply being the union of the parents of variables in the block. Now, the conditional distribution of a block given its parents is defined as:

$$p(B_i \mid \text{pa}_{\mathcal{G}}(B_i)) = \frac{\prod_{c \in C((G_{bd_G(B)})^a) : c \not\subseteq \text{pa}(B)} \phi_C(c)}{Z}$$

This is saying that the conditional distribution of a block given its parents is defined as the normalized product of clique potentials in the augmented subgraph of the boundary of the block where the product excludes cliques that only involve parents of the block.

So, the statistical interpretation of a chain graph is one such that distributions represented by a CG can be represented by the above 2-level factorization.

The Markov Blanket of a variable, x, is the set of variables such that once x is conditioned upon this set, it is independent from the rest of the variables in the distribution. In the statistical model of a chain graph, the Markov Blanket of a variable is the variables' neighbors and parents. The causal model of a chain graph involves an interpretation using a topological ordering of the variables in the distribution and Gibbs sampling from the distribution.

In a topological ordering of a distribution represented by a chain graph, the variables are listed in such a way so that any directed edge in the chain graph would be have its head to the left of its tail if we were to draw the edge between the two variables. A valid topological ordering of a distribution represented by a chain graph is guaranteed since the only directed edges occur between one block and another block and partially directed cycles are not present in the chain graph. Thus, a valid topological ordering would be to first topologically sort the blocks and within the blocks the order does not matter since there is no directed edges within a block.

Gibbs sampling is simply an easy method to sample from a complex probability distribution. In Gibbs sampling, each variable in the distribution is sampled from its full conditional one at a time. After an initial burn in period, the equilibrium distribution that the process converges to is the original distribution. However, instead of sampling each variable one at a time from its full conditional distribution, one can instead sample from the variable's distribution conditioned on its Markov Blanket. Since we know that there exists a valid topological ordering for the distribution represented by a chain graph, when running a Gibbs sampling algorithm, one can order the individual variables in the same as order as the topological ordering, and, then, one can sample from all of the variables' conditional distribution given its Markov Blanket in that order. Suppose, we are in xth iteration of the burn in period for sampling from some distribution represented by a chain graph. If the variable has been sampled from its conditional distribution at the xth iteration already, then the value it takes at that iteration is substituted in for wherever the variable is later conditioned upon in a conditional distribution.

This Gibbs sampling interpretation lends itself to a causal model of a chain graph. This interpretation provides a generative model for the data and from which a joint distribution arises. This interpretation also makes it easier to see how altering certain variables can affect other variables. It's important to recognize that in the generative story and causal model, only a variable's parents and neighbors can have a causal effect on that variable.

The causal interpretation of a chain graph incorporates the causal implications of directed edges in other graphical models as well as the causal implications of undirected edges in other graphical models. In other words, a directed edge can be interpreted as one variable possibly being a direct cause of the other. Whereas, an undirected edge can be interpreted as a pair of variables possibly coexisting in an equilibrium or dynamic feedback system.

4. Methods

4.1 Data Manipulation

After reading in the data csv file, there were 255 attributes. I reduced the number of variables in the data set by dropping columns from my dataframe. I did for a few reasons; the most pertinent reason I dropped many columns has to do with the correct tier ordering constraint in the structure learning algorithm I used to discover a chain graph from the data. I will elaborate upon this in the "Convergence to the true model" section. I have included a picture of a list of the variables that I analyzed and performed structure learning and statistical inference on.

```
Out[779]: ['Race',
 'FathersEducation',
 'MothersEducation',
 'FemGaurdianOccupation',
 'MaleGuardianOccupation',
 'AnnualFamilyIncome',
 'SocioeconomicStatusComposite',
 'Urbanicity',
 'Region',
 'TimeSpentAtHomeAfterSchoolWithoutAdult',
 'AskedIfHomeworkDone',
 'AskedToDoChores',
 'NumTimesDiscussedCourseMaterialWithGuardians',
 'GuardiansAttendedSchoolMeeting',
 'GuardiansSpokenToSchool',
 'LimitOnTvTime',
 'percentMinorityin8thgrade',
 'Theteachingisgood',
 'percentofSchoolwithDiscountedLunch',
 'Prolbemofstudentalcuse',
 'problemofstudentdruguse',
 'LimitOnTimeWithFriends',
 'numtimesyouvedrankalcinlifetime',
 'numcigsyousmokeaday',
 'MathAbilityGroup',
 'ScienceAbilityGroup',
 'EnglishClassAbilityGroup',
 'HistoryAbilityGroup',
 'PastScienceGrades',
 'PastHistoryGrades',
 'HistoryStandardizedScore',
 'ReadingStandardizedScore',
 'Relativereadingproficiencylevel',
 'Whatkindofworkdoyouexpecttobedoingwhenyouare30yearsold',
 'AggregateMeasureofControl1',
 'Asthingsstandnowhowfarinschooldoyouthinkyouwillget',
 'Howsureareyouthatyouwillgraduatefromhighschool']
```

All of these variables were discrete variables except for socioeconomic status. For brevity and because most are described well by their name, I will not explain each one and rather elaborate upon the variables of interest when they come up later in the 'Results' and Discussion' sections.

4.2 Missing Data and Data Cleaning

As mentioned briefly in the introduction, the dataset I am analyzing was constructed from a survey and the questions on the survey correspond to variables in the dataset that I renamed accordingly. There were 258 rows out of 2001 total rows which had missing data. Since the number of rows with missing data was quite small relative to the total size of the dataset, I decided to just drop these rows with missing data. After some data exploration, I could not discern any pattern in the occurrences of missing data so I believe that the missingness was random. Next, I noticed that the study designers often encoded non-valid responses into specific numbers and often also included an "I don't know" answer for many questions. Thus, for many of the variables, there would often be two numbers corresponding to some type of invalid response or the respondent not knowing that would lie outside the domain for valid responses as outlined in the survey "codebook". To deal with invalid responses or "I don't know" responses, I repeated the following procedure for each variable that I was analyzing. I would reference the codebook for the domain of the valid responses. I'd take the mean of the valid responses for that variable. If the variable was discrete, I'd simply take the floor of the mean. Then, I'd assign this value to the non-valid responses and other non-informative responses that I could identify since they would lie outside the domain given. After this process, I had data which had no missing rows and all of the data was representative of a valid response to a survey question. Of course, this method may not been correct in the sense that it could change the likelihood that the data arose from its true distribution and thus bias the causal discovery procedure.

4.3 Structure Learning Algorithm

To do structure learning, I used a structure learning algorithm written by Professor Bhattacharya based on theory presented in Bhattacharya et al. (2019) and Javidian et al. (2020).

4.4 Convergence to the true model

For this structure learning algorithm to asymptotically converge to the correct model, there exist certain assumptions and requirements that must be met. A necessary element for this structure learning algorithm to learn the true model is a correct tier ordering for the variables. Below are the tiers pertaining to the variables I am analyzing.

```
tier1 = ['Race','FathersEducation','MothersEducation']

tier2 = ['FemGaurdianOccupation','MaleGuardianOccupation']

tier3 = ['AnnualFamilyIncome']

tier4 = ['SocioeconomicStatusComposite','Urbanicity','Region']

tier5 = ['TimeSpentAtHomeAfterSchoolWithoutAdult','AskedIfHomeworkDone','AskedToDoChores',
         'NumTimesDiscussedCourseMaterialWithGuardians','GuardiansAttendedSchoolMeeting',
         'GuardiansSpokenToSchool','LimitOnTvTime','percentMinorityin8thgrade',
         'Theteachingisgood','percentofSchoolwithDiscountedLunch',
         'Prolbemofstudentalcuse','problemofstudentdruguse','LimitOnTimeWithFriends']

tier6 = ['numtimesyouvedrankalcinlifetime','numcigsyousmokeaday']

tier7 = ['MathAbilityGroup','ScienceAbilityGroup','EnglishClassAbilityGroup',
         'HistoryAbilityGroup','PastScienceGrades','PastHistoryGrades', 'HistoryStandardizedScore',
         'ReadingStandardizedScore','Relativereadingproficiencylevel']

tier8 = [ 'Whatkindofworkdoyouexpecttobedoingwhenyouare30yearsold','AggregateMeasureofControl1',
         'Asthingsstandnowhowfarinschooldoyouthinkyouwillget','Howsureareyouthatyouwillgraduatefromhighschool']
```

In addition, since the algorithm uses the BIC score to determine what is in and not in the Markov Blanket of a variable, for the algorithm to asymptotically arrive at the correct Markov Blanket for a variable, it is assumed that the variable's conditional distribution given its Markov Blanket is generated from a distribution belonging to the curved exponential family. Furthermore, convergence to the true model is dependent on the analyst specifying the correct member of the exponential family which defines the distribution of the target variable given its Markov Blanket. In my analysis, for the continuous variables, I assume that their conditional distribution given each of its Markov Blankets is in the form of a Gaussian distribution. For the discrete variables, I assume that the conditional distributions given each of its Markov Blankets is in the form of a Poisson distribution. Of course, incorrect specifications of these distributions will introduce error into my analysis.

So, assuming that the distribution of each variable given its Markov Blanket arises from a distribution that is a member of the curved exponential family, that the distribution can be faithfully represented by some chain graph, that the tier ordering is correct, and that the analyst chose the correct member of the exponential family to define the conditional distribution, then the structure learning algorithm described above will asymptotically converge to the true chain graph.

5. Results and Discussion

I have included 3 chain graphs in the "Pictures" folder of this project. "cgPoisson.gv.pdf" represents the original chain graph that was learned from the data when I used a Poisson distribution to represent the conditional distribution of a discrete variable given its Markov Blanket.

"cgBS.gv.pdf" represents the chain graph learned from the data using bootstrap sampling and the Poisson conditional distribution for discrete variables. I ran a 'for loop' 26 times within which I re-sampled the original data with replacement and the same number of rows and then called the causal discovery algorithm to learn a chain graph. So, during each iteration, I arrived at a new chain graph learned that was learned with respect to re-sampled data with the same numbers of examples and attributes.

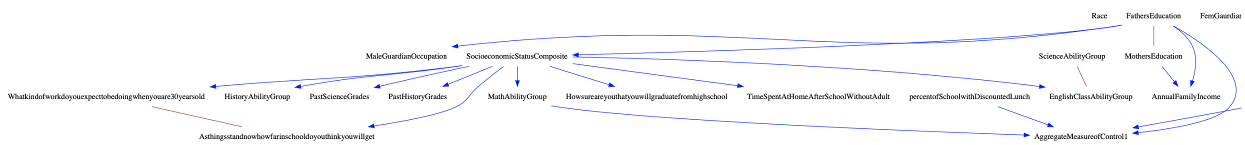
Prior to this 'for loop', I created two dictionaries. One dictionary was a dictionary mapping each directed edge that appeared in the original non-bootstrapped chain graph learned to the number of times that same edge appeared in a chain graph that was learned within the 'for loop'. The second dictionary was the same as the first except that it was a dictionary for the undirected edges in the original graph learned instead.

After the 'for loop' terminated, I dropped from the original chain graph the edges that appeared in fewer than 17 of the 26 'bootstrapped chain graphs'. I'd like to note that during the 'for loop', I kept the assumption of the Poisson conditional distribution for the discrete variable consistent.

The following is a list of the directed edges in the chain graph in 'cgBS.gv.pdf'. It is here for reference as I will be discussing the causal implications of some of these edges.

```
{('FathersEducation', 'AggregateMeasureofControl1'),
 ('FathersEducation', 'AnnualFamilyIncome'),
 ('FathersEducation', 'MaleGuardianOccupation'),
 ('FathersEducation', 'SocioeconomicStatusComposite'),
 ('MathAbilityGroup', 'AggregateMeasureofControl1'),
 ('MothersEducation', 'AnnualFamilyIncome'),
 ('NumTimesDiscussedCourseMaterialWithGuardians',
  'Relativereadingproficiencylevel'),
 ('Prolbemofstudentalcuse', 'AggregateMeasureofControl1'),
 ('SocioeconomicStatusComposite',
  'Asthingsstandnowhowfarinschooldoyouthinkyouwillget'),
 ('SocioeconomicStatusComposite', 'EnglishClassAbilityGroup'),
 ('SocioeconomicStatusComposite', 'HistoryAbilityGroup'),
 ('SocioeconomicStatusComposite',
  'Howsureareyouthatyouwillgraduatefromhighschool'),
 ('SocioeconomicStatusComposite', 'MathAbilityGroup'),
 ('SocioeconomicStatusComposite', 'PastHistoryGrades'),
 ('SocioeconomicStatusComposite', 'PastScienceGrades'),
 ('SocioeconomicStatusComposite', 'TimeSpentAtHomeAfterSchoolWithoutAdult'),
 ('SocioeconomicStatusComposite',
  'Whatkindofworkdoyouexpecttobedoingwhenyouare30yearsold'),
 ('numcigsyousmokeaday', 'ReadingStandardizedScore'),
 ('percentofSchoolwithDiscountedLunch', 'AggregateMeasureofControl1')}
```

And, below is a portion of the chain graph learned after completion of the for loop pertaining to bootstrap re-sampling.



The most striking finding is number of directed edges originating from the variable, SocioeconomicStatusComposite. This variable is a continuous variable that is essentially an aggregated average of different socioeconomic-related variables in the survey corresponding to the dataset I'm analyzing and a sister survey corresponding to a survey of the respondents' parents. The multitude of edges originating from this variable implies that it is a direct cause of many other variables including the student's 'Math Ability Group', 'History Ability Group', confidence in graduating from high school, and the student's prediction of what kind of work he or she will be doing when they are 30. My hypothesis was that higher and better socioeconomic statuses would be a cause of a student's success in academics. Thus, my initial conclusion was that edges emanating from 'SocioeconomicStatusComposite' supported my hypothesis.

To confirm this, I fit a Poisson regression with the endogenous variable being 'Math Ability Group' and the exogenous variable being 'SocioeconomicStatusComposite'. The coefficient for the latter was 0.117541 and can be seen in 'math ability group regression.png' in the Pictures folder of this project. This was an unexpected finding in two ways. First, since 'MathAbilityGroup' is a discrete variable in which the highest ability group was designated with a 1 and the lowest ability group was designated with a 4. So, the implication was that increased socioeconomic status is a cause of lower math ability group. However, it should be noted that range of the exogenous variable was 4; so, such a small coefficient and such a small range lead me to believe that the causal effect is negligible and may only be a result of having the lowest BIC score relative to other configurations of the variable's Markov Blanket. I fit another Poisson regression with the endogenous variable being 'History Ability Group' and the exogenous being 'SocioeconomicStatusComposite'. The coefficient for the latter was -0.102143. Unlike the previous regression, this coefficient is negative and thus supports my hypothesis. However, again, the coefficient is so small that it's likely just an artifact of the learning process and not indicative of any actual causal effect.

It's difficult to determine what magnitude confers significance upon the coefficients in a Poisson regression especially since the range of both the exogenous and endogenous variables in the above regressions are narrow. For simplicity, I will assume that the coefficients, regardless of magnitude, are significant and indicative of the effect of the causal relationship between the variables. This is obviously a large assumption but makes exposition simpler. Under this assumption, the nature of the causal relationship between 'Math Ability Group' being caused by 'SocioeconomicStatusComposite' is actually the reverse of what my hypothesis was. On the other hand, my hypothesis for 'History Ability Group' and 'SocioeconomicStatusComposite' is supported by the chain graph and the fitted regression. The negative coefficient implies that higher socioeconomic status is a cause of better grades. I want to explicitly state here that my hypotheses of higher income and higher socioeconomic statuses being a cause of better grades is meant to be interpreted as a causal mechanism as follows: higher incomes families spend more money on private services like tutoring and SAT prep; children of these families can dedicate more time to studies and not worry about working or taking care of a loved one and so forth. The consequence of these is that these students achieve better academic outcomes. I wanted to explicitly state this since I wanted to clarify that my hypotheses are NOT meant to imply that I believe rich kids are inherently smarter than poor kids.

Another hypothesis I stated earlier in the paper was that higher education in parents would be a cause of better academic performance for the student. Although in the chain graph learned from bootstrap re-sampling there does not exist a directed edge from 'FathersEducation' or 'MothersEducation' to a variable related to academic performance, there DOES exist a directed edge from 'FathersEducation' to 'SocioeconomicStatusComposite'. Intuitively, my interpretation of this edge led me to think of a causal mechanism in which higher education for a parent leads to a higher socioeconomic status for the family which leads to better academic performance for the student. I have addressed the second part of this causal mechanism above; from the chain graph and the fitted regressions, it is not clear whether or not my hypothesis was correct. Nonetheless, I will fit a Gaussian distribution with the exogenous variable being 'FathersEducation' and the endogenous variable being 'SocioeconomicStatusComposite'. This corresponds to the causal mechanism by which 'SocioeconomicStatusComposite' is purportedly generated.

```
: formula = 'SocioeconomicStatusComposite ~ 1 +' + '+'.join(levelsNames['FathersEducation'][:-1])
  model = sm.GLM.from_formula(formula, data=data, family=sm.families.Gaussian()).fit()
  model.params
```

```
: Intercept            0.265377
  FathersEducation1   -0.367179
  FathersEducation2   -0.279072
  FathersEducation3   -0.324707
  FathersEducation4   -0.172306
  FathersEducation5   -0.109155
  FathersEducation6   -0.034289
  dtype: float64
```

In this regression, we see coefficients for 6 levels of 'FathersEducation'. When fitting a regression with discrete exogenous variables, I used one hot encoding with a reference value corresponding to 0's for all levels in the regression model. First, I'd like to point out that the larger the number appended to 'FathersEducation' corresponds to a higher degree or more education. For instance, 'FathersEducation1' corresponds to not finishing high school, and 'FathersEducation7' corresponds to obtaining an M.D. or PhD. So, the intercept corresponds to the mean socioeconomic status value for families in which the father obtained an M.D. or PhD. The coefficient for the other 6 levels of 'FathersEducation' corresponds to the difference between the mean socioeconomic status value for families in which the father obtains the corresponding level of education and the mean socioeconomic status value for families in which the father obtains an M.D. or PhD. Taken together, my hypotheses would imply that higher education attainment in parents would lead to higher socioeconomic statuses. The parameters from the fitted regression support this hypothesis. The negative coefficients imply that the families in which the father obtains an M.D. or PhD typically have higher socioeconomic statuses than those that don't. Additionally, the magnitude of the coefficient increases as the education attainment of the father decreases; this supports the hypothesis that higher levels of education improves a family's socioeconomic status.

All this being said, it's interesting to note that there does not exist a directed edge from 'FathersEducation' to variables related to the student's grades or expectations about his or her future. The absence of these edges is welcome; as 'FathersEducation' is not something that can be intervened on directly, it'd be hard to find a tangible solution to improve a student's academic performance through his or her father's education. However, as I have discussed earlier in this section, there do exist many edges from 'SocioeconomicStatusComposite' to variables related to the student's academic performance. Although the regressions I have fit do not all support the hypothesis that higher socioeconomic statuses cause better academic performance, I believe that my hypothesis is still correct, and in the 'Conclusion' section, I discuss why I think my regressions fail to capture this. Also, I have already proposed a possible causal mechanism by which higher socioeconomic status can lead to better academic performance from the student. Now, I'd like to discuss possible interventions that can be taken to improve a student's academic performance and consequently his or her future socioeconomic status. Given it's not truly feasible to drastically improve a family's socioeconomic status, one must seek other ways in which to affect the causal mechanism in which academic performance can be improved through increased socioeconomic status. In my opinion, these other ways include subsidizing internet access, increasing the quality and access to free educational resources similar to Khan Academy, and generally just investing more money in the middle and lower classes. I believe that this combination of intervention can effectively replicate the effects that higher socioeconomic status has upon a student's life and academics.

6. Conclusion

I've touched on this at times during the 'Results and Discussion' section, but many of the regressions I fit where the exogenous variable are the Markov Blanket of a variable, x, and the endogenous variable is x gave me parameters that were quite small. Additionally, the difference in coefficients between different levels of a discrete exogenous variable was often small as well. This made it tough to discern what variables truly play a causal effect in the generation of another variable and what variables are in the Markov Blanket due to idiosyncrasies of the dataset and the learning process.

So, although the bootstrapped chain graph I learned did imply many of the mechanisms and causal effects I figured it would, upon further analysis it was difficult to really quantify the actual magnitude of this supposed causal effect. In a similar vein, during this process, I have learned of the difficulties of using a chain graph to conduct analysis of my data. Aside from needing to have a correct tier ordering which was tricky in its own right, the constraint that the conditional distributions of a variable given its Markov Blanket must be from the curved exponential family likely introduced quite a bit of error into my analysis. Since the majority of my variables were discrete, the regressions that I could choose from were limited primarily to the Poisson regression and multinomial logit regressions. The multinomial regressions were taking a very long time to run and were having convergence issues, so I settled on using the Poisson regression. If I were to have more time on this project, I would work to fix the convergence issues and opt to use multinomial logit regressions or do more research on discrete distributions in the curved exponential family. Assuming that the discrete variable are generated according to some Poisson regression is a large assumption and unlikely to be the actual causal mechanism at work. I believe it is this constraint that led me to arrive at some unusual and odd results. It is also possible that the data itself is inaccurate and noisy. I noticed some discrepancies between the 'codebook' explaining the variables and their values and the actual data itself.

Another possible route I would have liked to explore is conducting my analysis through the lens of an ADMG. As I mentioned earlier, I was initially contemplating using an ADMG to begin with. Aside from the confounder causal mechanism that could be introduced in my analysis, structure learning could potentially be easier and more accurate. If the structure learning algorithms were constraint-based rather than score-based, I'd be able to relax the constraint that the causal mechanisms are such that values are generated from a curved exponential distribution. This relaxation would give me more flexibility and likely allow for better causal discovery.

References

Rohit Bhattacharya, Daniel Malinsky, and Ilya Shpitser. Causal inference under interference and network uncertainty, 2019.

Mohammad Ali Javidian, Marco Valtorta, and Pooyan Jamshidi. Learning lwf chain graphs: A markov blanket discovery approach, 2020.